ΜΠΣ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

ΕΝΙΣ ΤΣΕΛΑΙ

Μελέτη διασποράς των HIV-1 ανασυνδυασμένων στελεχών (B/G και B/F) στην Ισπανία με μεθόδους μοριακής επιδημιολογίας

ΑΘΗΝΑ, 2018

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη

ΒΙΟΣΤΑΤΙΣΤΙΚΗ

που απονέμει η Ιατρική Σχολή και το Τμήμα Μαθηματικών του Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών

Εγκρίθηκε τον ΕΝΙΣ ΤΣΕΛΑΙ από την εξεταστική επιτροπή:

ΟΝΟΜΑΤΕΠΩΝΥΜΟ	***ΒΑΘΜΙΔΑ***	***ΥΠΟΓΡΑΦΗ***
Δ. ΠΑΡΑΣΚΕΥΗ(Επιβλέπων)	**Ε. ΚΑΘΗΓΗΤΗΣ**	**……**
Β. ΣΥΨΑ	**Ε. ΚΑΘΗΓΗΤΡΙΑ**	**…….**
ΓΚ. ΜΑΓΙΟΡΚΙΝΗΣ	**ΛΕΚΤΟΡΑΣ**	**……..**

Περιεχόμενα

Σκοπός

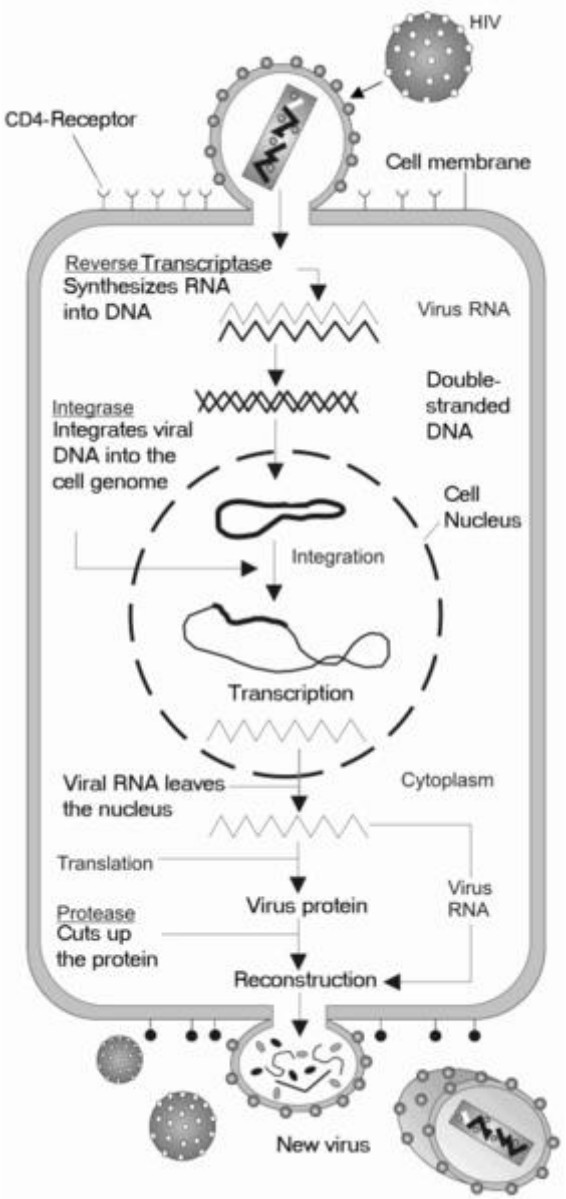
Σκοπός της μελέτης είναι η εύρεση του τρόπου διασποράς των ανασυνδυασμένων στελεχών B/G και B/F στην Ισπανία. Η ανάλυση πραγματοποιήθηκε με συνδυασμό μεθόδων μοριακής επιδημιολογίας (φυλογενετικής ανάλυσης και στατιστικής φυλογεωγραφίας).

1. Εισαγωγή

Η επιδημία HIV/AIDS παραμένει ένα σημαντικό πρόβλημα δημόσιας υγείας σε παγκόσμιο επίπεδο. Σύμφωνα με στοιχεία του Παγκόσμιου Οργανισμού Υγείας (1) (ΠΟΥ), εκτιμάται ότι ο συγκεκριμένος ιός είχε προσβάλει περίπου 36.9 εκατομμύρια ανθρώπους μέχρι τα τέλη του 2017 ανά την υφήλιο. Η γεωγραφική κατανομή και ο επιπολασμός των HIV-1 υπότυπων και ανασυνδυασμένων τύπων (circulating recombinant forms - CRFs) στην Ευρώπη χαρακτηρίζονται από υψηλή γενετική ετερογένεια (2). Παρόλο που ο υπότυπος Β αποτελεί τον κυρίαρχο υπότυπο σε δυτική και κεντρική Ευρώπη από την αρχή της επιδημίας, τα τελευταία 15 έτη η επιδημία έχει διαφοροποιηθεί σημαντικά. Στελέχη μη-Β υπότυπων και ανασυνδυασμένων τύπων έχουν εισαχθεί στην Ευρώπη μέσω μεταναστευτικών κυμάτων και έχουν διαδοθεί μεταξύ των Ευρωπαϊκών χωρών αυξάνοντας την πολυπλοκότητα της επιδημίας του HIV-1. Η Ισπανία αποτελεί μια από τις χώρες με τον υψηλότερο επιπολασμό HIV-1 στην Ευρώπη. Προηγούμενες μελέτες σε δείγμα 6.632 HIV-1 αλληλουχιών, έδειξαν ότι ο υπότυπος Β παρουσιάζει τον υψηλότερο επιπολασμό (83.7%) στην Ισπανία, ενώ εντοπίστηκαν και άλλοι υπότυποι (A1, F1, C, D, G), CRFs και unique recombinant forms (URFs) με επιπολασμό της τάξης του 5.5%, 10.7% και 0.1% αντίστοιχα (2).

1.1 Ο HIV ως ρετροϊός

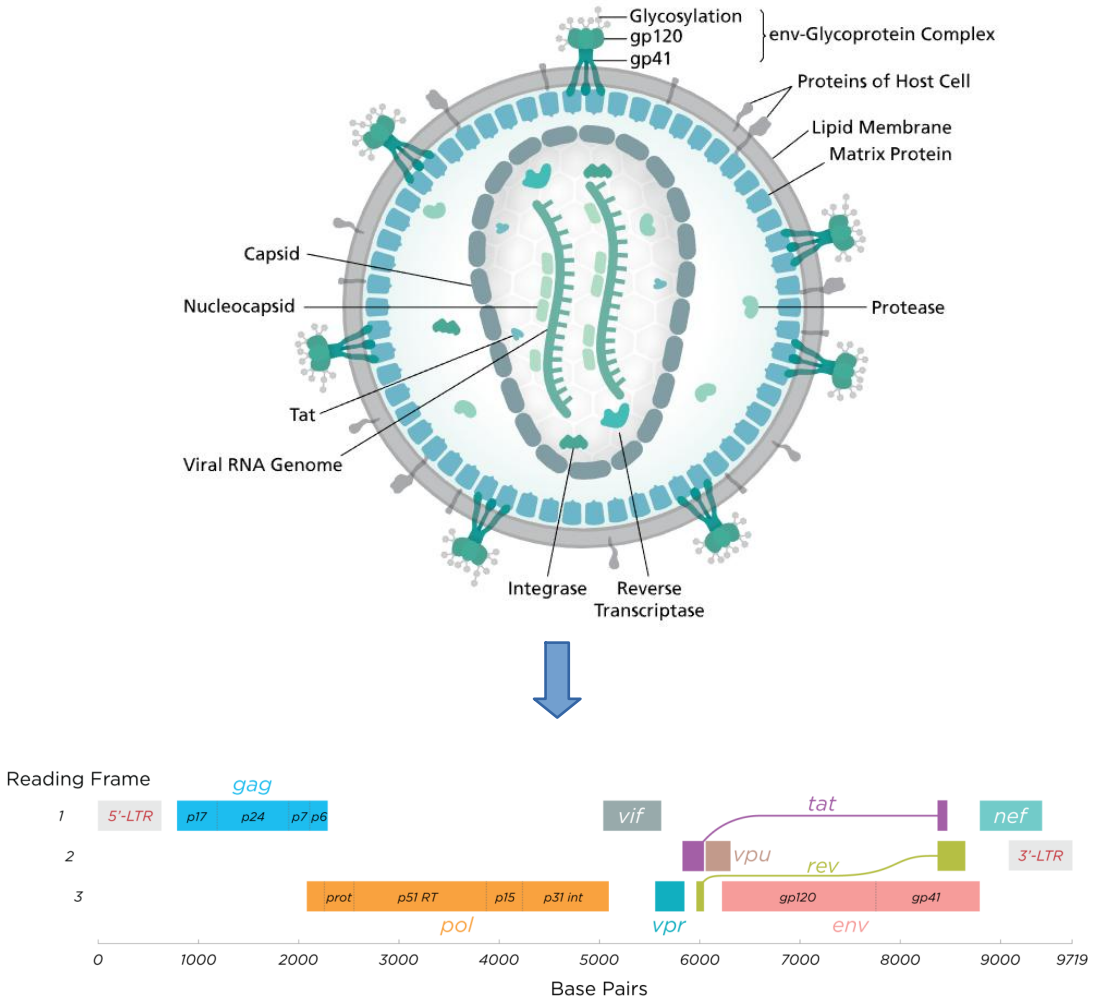
Οι ρετροϊοί είναι μια ετερογενής ομάδα ιών με γενετικό υλικό RNA. Όταν ένας ρετροϊός εισέρχεται στο κύτταρο-ξενιστή το RNA του μετατρέπεται σε DNA με τη διαδικασία της αντίστροφης μεταγραφής. Ύστερα το ιικό DNA ενσωματώνεται στο γονιδίωμα του ξενιστή όπου πολλαπλασιάζεται μαζί με αυτό. Έπειτα με τη διαδικασία της μεταγραφής από το ιικό DNA παράγονται RNA και mRNA που χρησιμοποιούνται για την παράγωγη νέων ιών και την παράγωγη πρωτεϊνών αντίστοιχα (Εικονα.1). Στην κατηγορία των ρετροϊών ανήκει και ο ιός της επίκτητης ανθρώπινης ανοσοποιητικής ανεπάρκειας (HIV), ο οποίος κατηγοριοποιείται σε δυο αρκετά διαφορετικά είδη τον HIV-1 και HIV-2. Ο HIV θεωρείται ένα από τα ταχύτερα μεταλλασσόμενα ανθρώπινα παθογόνα.
Ο HIV (Human Immunodeficiency Virus) δηλαδή, ο Ιός της Ανθρώπινης Ανοσοανεπάρκειας είναι ένας ρετροϊός που μολύνει τα CD4 λεμφοκύτταρα γνωστά και ως Τ λεμφοκύτταρα. Όταν ο αριθμός των κυττάρων αυτών πέσει κάτω από τα 200 χιλιοστά του λίτρου στο αίμα τότε θεωρούμε ότι το άτομο νοσεί. Ο ιός προέρχεται από των ιό SIV (simian immunodeficiency virus), ο οποίος μολύνει ένα είδος χιμπατζή στην κεντρική Αφρική και ο οποίος κάποια στιγμή μεταλλάχτηκε σε HIV(3-4) .



Εικόνα 1. Ο κύκλος ζωής του HIV: εισαγωγή στον ξενιστή, αντίστροφη μεταγραφή, μεταγραφή και ανασύνθεση.

1.1.1 Η δομή του ίου

Ο HIV έχει σχεδόν σφαιρικό σχήμα με διάμετρο γύρω στα 120 nm. Η γενετική οργάνωση του ίου είναι πολύπλοκη καθώς αποτελείται από δύο μονόκλωνα μόρια RNA και έχει μήκος 9,7 kb (9.700 βάσεις). Ο ιός αποτελείται από δυο αντίγραφα μονόκλωνου RNA (ssRNA) που περιβάλλονται από ένα καψίδιο λιπιδιακής φύσης (env), το οποίο προέρχεται από το κύτταρο-ξενιστή. Στην επιφάνεια του ιού υπάρχουν υποδοχείς που δημιουργούνται από τις γλυκοπρωτεΐνες gp120 και gp41. Η gp120 είναι μια γλυκοπρωτεΐνη πρόσδεσης που τοποθετείται προς το εξωκυττάριο περιβάλλον, ενώ η gp41 είναι μια διαμεμβρανική γλυκοπρωτεΐνη που βρίσκεται μέσα στη λιπιδιακή μεμβράνη του ιού (5). Η g120 επιτρέπει στον ιό να μολύνει τα ανθρώπινα άνοσο-κύτταρα με το να συνδέεται με έναν υποδοχέα κυτοσίνης του κυττάρου-ξενιστή, όπως είναι η CCR5 ή η CXCR4 ανάλογα το στέλεχος του ίου (6, 7). Στο εσωτερικό του ίου βρίσκεται η μήτρα (matrix) που περιλαμβάνει την πρωτεΐνη p17 ενώ το καψίδιο που την περιβάλλει συγκροτείται από την πρωτεΐνη p24. Μέσα στο καψίδιο βρίσκεται το νουκλεοκαψίδιο, το οποίο περιέχει το γενετικό υλικό του ιού, συνδεδεμένο με την πρωτεΐνη p7, τα ιικά ένζυμα πρωτεάση (PR), αντίστροφη μεταγραφάση (RT), ιντεγκράση (IN) καθώς και οι ικές πρωτεΐνες Vpu, Vif, Vpr και Nef (8).



Εικόνα 2. Το γονιδίωμα του HIV-1. Τα ανοιχτά πλαίσια ανάγνωσης εμφανίζονται ως ορθογώνια. Η έναρξη του γονιδίου, που υποδεικνύεται από τον μικρό αριθμό που βρίσκεται στην πάνω αριστερά γωνία κάθε ορθογωνίου και καταγράφει τη θέση του α στο ATG κωδικόνιο έναρξης για το γονίδιο αυτό, ενώ ο αριθμός στην κάτω δεξιά καταγράφει την τελευταία θέση του κωδικονίου λήξης. Για την *pol*, η αρχή θεωρείται ότι είναι η πρώτη T στην ακολουθία TTTTTTAG, η οποία αποτελεί μέρος του βλαστού που ενισχύει την ριβοσωμική ολίσθηση στο RNA και μια προκύπτουσα -1 μετατόπιση πλαισίου και την μετάφραση της Gag-Pol πολυπρωτεΐνης. Τα εξωγενώς των γονίδιων *tat* και *rev* εμφανίζονται ως σκιασμένα ορθογώνια. Στο HXB2, το 5772 σηματοδοτεί τη θέση μιας μετατόπισης πλαισίου στο γονίδιο *vpr* που προκαλείται από ένα "επιπλέον" T σε σχέση με τους περισσότερους άλλους ιούς υπότυπου B. Το 6062 υποδεικνύει ένα ελαττωματικό κωδικόνιο εκκίνησης ACG στο *vpu*; † 8424 και † 9168 σηματοδοτούν πρόωρα κωδικόνια διακοπής σε *tat* και *nef*. (9)

1.1.2 Η γενετική ετερογένεια του ίου HIV-1

Ο ιός του HIV χωρίζετε σε δυο είδη: τον HIV-1, ο όποιος ευθύνεται για το μεγαλύτερο μέρος των μολύνσεων παγκοσμίως (90%) και τον HIV-2 (10,11). Ένα από τα κύρια χαρακτηριστικά του ιού HIV-1 είναι η εκτενής γενετική του ετερογένεια, στην οποία οφείλεται και η ταξινόμηση του σε τέσσερις ομάδες: την ομάδα M (major), την ομάδα O (outlier), την ομάδα N (new) και την ομάδα P που απομονώθηκε πρόσφατα στην Αφρική. Αναλυτικότερα, η ομάδα M η οποία θεωρείται υπεύθυνη για την πανδημία του AIDS, ταξινομείται φυλογενετικά σε εννιά υπότυπους

(A-D, F-H, J και K), υπό-υπότυπους (A1, A2, F1, F2, κλπ) καθώς και σε ανασυνδυασμένους τύπους (Circulating Recombinant Forms, CRFs) (11 ,12).

1.1.3 Ο γενετικός ανασυνδυασμός του ίου (CRFs)

Ο ανασυνδυασμός γονιδίων, ή γενετικός ανασυνδυασμός, είναι η ανταλλαγή γενετικού υλικού μεταξύ των ομόλογων χρωματοσωμάτων. Κατά τον ανασυνδυασμό των γονιδίων, ένα τμήμα του πατρικού χρωμοσώματος αλλάζει θέση με το αντίστοιχό τμήμα του μητρικού. Ο γενετικός ανασυνδυασμός πραγματοποιείται μέσω μιας βιολογικής πορείας που ονομάζεται επιχιασμός και συμβαίνει στο στάδιο της ζυγοταινίας κατά την πρόφαση της πρώτης μειωτική διαίρεση. Συμβάλλει στην αύξηση της ποικιλομορφίας στην φύση και καθιστά αδύνατη την περίπτωση δύο άτομα του ίδιου είδους να είναι πανομοιότυπα, εκτός αν πρόκειται για μονοζυγωτικά δίδυμα (13).

Ο ανασυνδυασμός είναι ένας από τους κύριους μηχανισμούς στους οποίους βασίζεται η ποικιλομορφία του HIV-1. Επί του παρόντος αναγνωρίζονται 98 ανασυνδυασμένα στελέχη του HIV-1. Οι υπότυποι και οι υπό-υπότυποι προέκυψαν ως αποτελέσματα του φαινομένου του ιδρυτή σε διαφορετικές χρονικές στιγμές στο παρελθόν ενώ οι ανασυνδυασμένες μορφές των υποτύπων μπορούν να εμφανιστούν σε ασθενείς που έχουν μολυνθεί από δύο διαφορετικούς υπότυπους του ιού. Στην περίπτωση που αυτά τα πρόσφατα ανασυνδυασμένα στελέχη έχουν σημαντική επιδημική διάδοση, ονομάζονται ανασυνδυασμένες μορφές (CRFs ή Circulating Recombinant Forms) (14, 15).

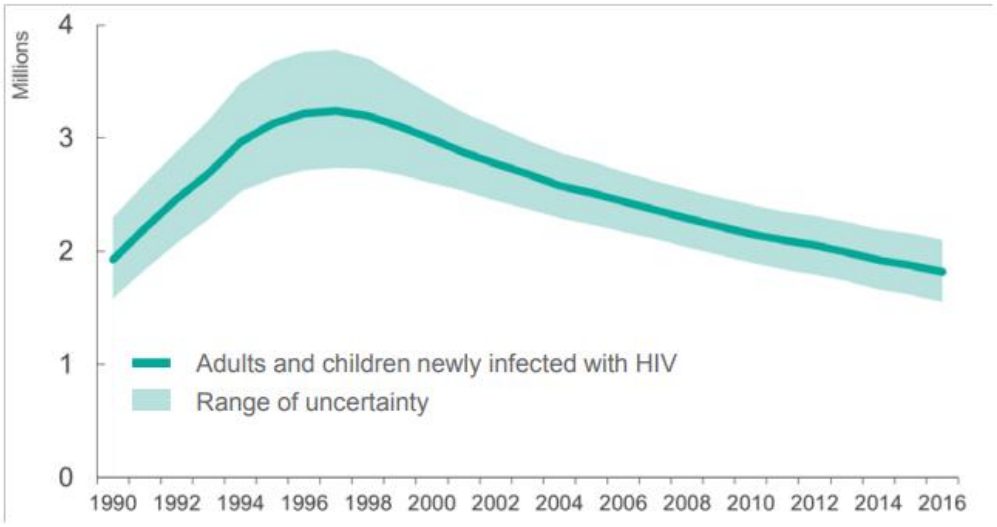
1.2 Ιστορική ανάδρομη

Το Σύνδρομο της Επίκτητης Ανοσοανεπάρκειας γνωστό και ως AIDS (Acquired Immunodeficiency Disorder Syndrome), το οποίο προκαλείται από τον ιό του HIV, εμφανίστηκε για πρώτη φορά το 1981. Τα πρώτα κρούσματα αφορούσαν χρήστες ενδοφλέβιων ναρκωτικών και ΑΣΑ (Άντρες που κάνουν σεξ με άντρες) στις Η.Π.Α (16). Οι ασθενείς παρουσίασαν συμπτώματα πνευμονίας από πνευμονοκύστη (PCP) που παρουσιάζεται σε άτομα με εξασθενημένο ανοσοποιητικό σύστημα. Επίσης, κάποιοι ανδρών που κάνουν σεξ με άνδρες, εμφάνισαν μια σπάνια μορφή καρκίνου του δέρματος. Ο HIV ανακαλύφθηκε το 1983 από δυο διαφορετικές ερευνητικές ομάδες, την ομάδα του Γκάλλο και την ομάδα του Montagnier (17).

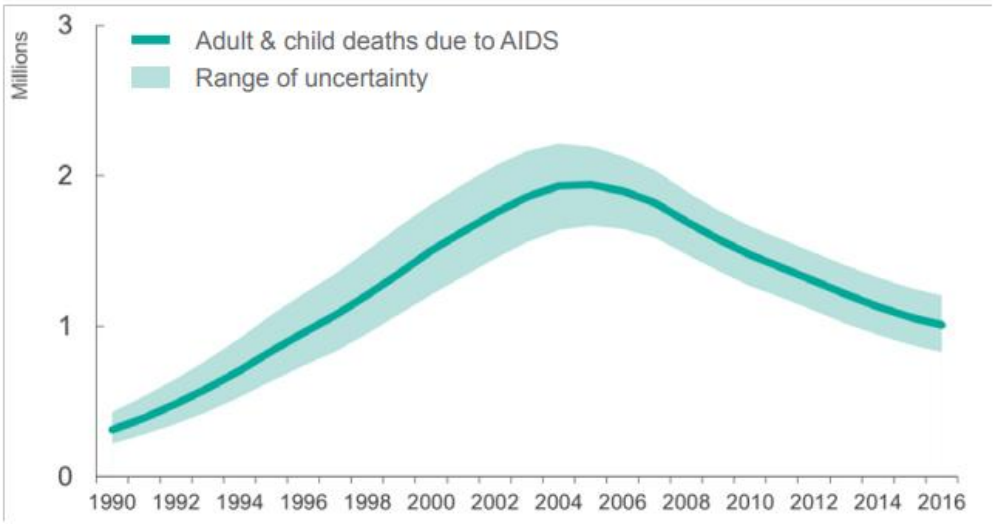
Ο ιός ονομάστηκε HIV το 1986 όταν αποδείχτηκε ότι ο ιός που είχαν απομονώσει οι δυο ομάδες ήταν ο ίδιος. Και οι δύο τύποι του ιού, HIV-1 και HIV-2, πιστεύεται ότι προέρχονται από διαφορετικό στέλεχος SIV. Συγκεκριμένα, ο HIV-1 θεωρείται ότι έχει προέλευση το Καμερούν όπου το στέλεχος του ίου (SIV-cpz) βρέθηκε να μολύνει άγριους χιμπατζήδες (Pan troglodytes troglodytes) (18) ενώ ο HIV-2 προέρχεται από τη Δυτική Αφρική, όπου βρέθηκε το στέλεχος (SIVsmm) να μολύνει το είδος του πιθήκου Cercosebus alys alys (19). Ο HIV-1 έχει μεταδοθεί μεταξύ διαφορετικών ειδών τουλάχιστον τρεις φορές, γεγονός που οδήγησε στην εμφάνιση των τριών διαφορετικών ομάδων του ίου: M, N και O (20). Από φυλοδυναμικές μελέτες προέκυψε ότι, ο πιο πρόσφατος κοινός πρόγονος των υποτύπων της ομάδας M υπολογίζετε περίπου το 1910.

1.3 Η παγκόσμια επιδημία του HIV-1

Τα πρώτα χρόνια της επιδημίας ο πληθυσμός που επηρεάστηκε πιο πολύ ήταν εκείνος των ανδρών που κάνουν σεξ με άνδρες (ΑΣΑ). Πλέον όμως τα δημογραφικά χαρακτηριστικά ατόμων που νοσούν φαίνεται να έχουν αλλάξει κατά πολύ. Συνολικά 76.100.000 [65.200.00-88.000.000] άνθρωποι έχουν μολυνθεί από τον HIV παγκοσμίως ενώ 35.000.000 [30.800.000-42.900.000] έχουν πεθάνει από την αρχή της επιδημίας από ασθένειες που σχετίζονται με το AIDS. Σε μελέτη που πραγματοποιήθηκε το 2016 υπολογίσθηκε ότι υπάρχουν παγκοσμίως 36.700.000 [30.800.000-42.9 00.000] εκατομμύρια άνθρωποι που ζουν με AIDS, ενώ καταγράφηκαν 1.800.000 [1.600.000-2.100.000] νέες μολύνσεις και επίσης υπολογίζεται ότι 1.000.000 άτομα έχουν πεθάνει από ασθένειες που σχετίζονται με το AIDS (Εικόνα 3, 4). Το έτος με τον μεγαλύτερο αριθμό μολύνσεων ήταν το 1997 με περίπου 3.300.000 περιστατικά. Η παγκόσμια θνησιμότητα έπεσε ραγδαία από το 1997 έως το 2005, σε περίπου 2.600.000 το χρόνο, και από το 2005 έως το 2015 παραμένει σταθερή (21). Η Νοτιοανατολική Αφρική είναι η περιοχή με τα περισσότερα κρούσματα. Το 2016 εκτιμάται ότι το 53% (19.400.000) όλων των HIV περιστατικών, το 42% (420.000) όλων των θανάτων και το 44% (790.000) των νέων μολύνσεων παγκοσμίως πραγματοποιήθηκαν σε αυτή την περιοχή (22). Παράλληλα, η χώρα που επλήγη περισσότερο είναι η Νότια Αφρική με 7.100.000 άτομα να ζουν με HIV, 270.000 νέες μολύνσεις και 110.000 θανάτους σχετιζόμενους με το AIDS (24).



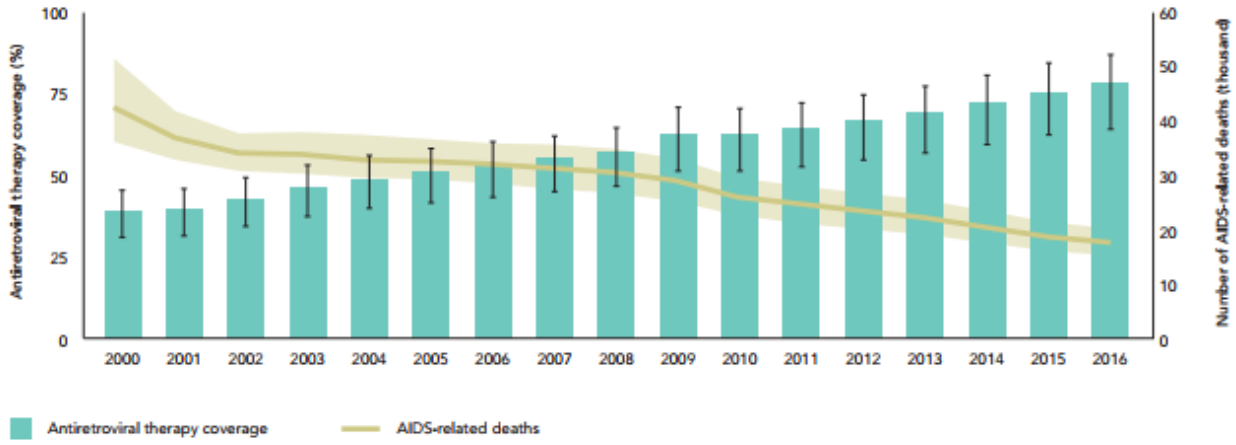
Εικόνα.3 : Αριθμός νέων μολύνσεων από HIV παγκοσμίως (24)



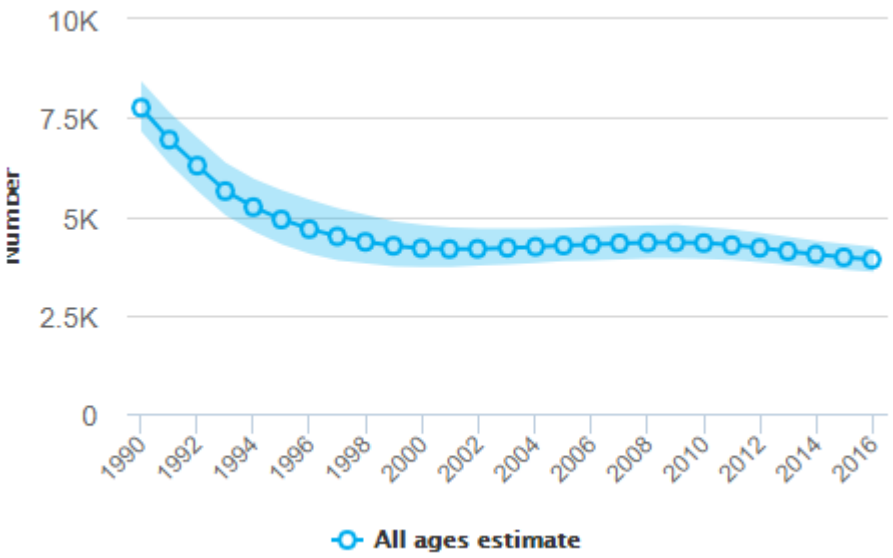
Εικόνα 4 : Σχετιζόμενοι με τον AIDS θάνατοι παγκοσμίως (24)

1.4 Η επιδημία του HIV στην Ισπανία

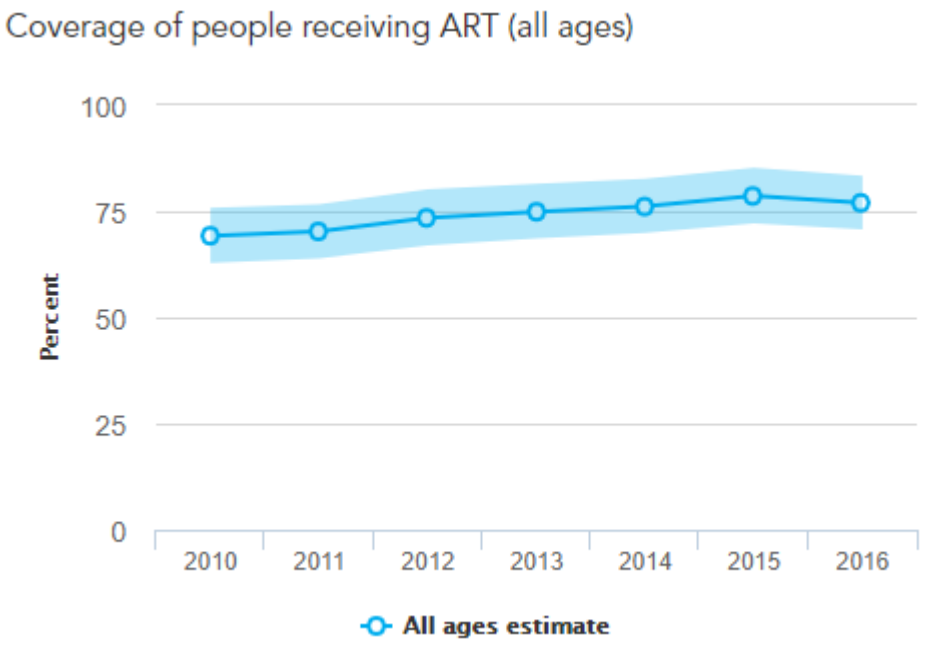
Στην Ευρώπη και τη Βόρεια Αμερική ο αριθμός των ατόμων που ζουν με HIV (PLHIV) καθώς και ο αριθμός των θανάτων (18.000 [15 000–20 000]) που σχετίζονται με το AIDS είναι πιο χαμηλός (το 2016) σε σχέση με τις υπόλοιπες ηπείρους. Αυτό οφείλεται κυρίως στην εύκολη πρόσβαση σε αντιρετροϊκή θεραπεία (Εικόνα 5). Στην Ισπανία υπολογίσθηκε ότι υπήρχαν 140.000 [130.000-160.000] άτομα που ζουν με HIV (PLHIV) το 2016 από τα οποία τα 32.000 [29.000 – 35.000] ήταν γυναίκες και τα 110.000 [100.000-120.000] ήταν άνδρες. Τα νέα περιστατικά μολύνσεων PLHIV) υπολογίσθηκε ότι είναι περίπου 3.900 [3.600-4.200] από τα οποία η πλειοψηφία ήταν άντρες με 3.200 νέα περιστατικά (Εικόνα 6). Από τους PLHIV εκτιμάται ότι το 77% [71-83] (110.000) λαμβάνουν αντιρετροϊκή θεραπεία (Εικόνα 7). Το μεγαλύτερο ποσοστό PLHIV φαίνεται να είναι ΑΣΑ οι όποιοι αγγίζουν το 11,3% (100.596 άτομα) του συνολικού πληθυσμού δηλαδή το 91% όλων των PLHIV και το 72% του πληθυσμού. Τα ποσοστά των PLHIV στην σεξουαλική εργασία είναι 2% και στους ναρκομανείς και φυλακισμένους είναι 2,3% και 5,4 % αντίστοιχα (23) .



Εικόνα 5: Ο αριθμός θανάτων σχετιζόμενων με το AIDS και ο αριθμός ατόμων που λαμβάνουν αντιρετροϊκή θεραπεία ανά έτος στην Ευρώπη και τη Βόρεια Αμερική.



Εικόνα 6: Νέα περιστατικά PLHIV στην Ισπανία ανά έτος (24)



Εικόνα 7: Ποσοστό ατόμων που λαμβάνουν αντιρετροϊκή θεραπεία ανά έτος στην Ισπανία (24)

2. Υλικό

Η μελέτη πραγματοποιήθηκε σε ανασυνδυασμένα στελέχη B/G (N=102) και B/F (N=98) του ιού HIV-1 που είχαν συλλεχθεί το χρονικό διάστημα 2000-2014 στην Ισπανία. Οι αλληλουχίες είχαν συλλεχθεί στα πλαίσια της εθνικής μελέτης CoRIS (2004-2013) (25) και της Eastern Andalucia Resistance Cohort (2000-2014). Παράλληλα χρησιμοποιήθηκαν όλες οι αλληλουχίες από τους αντίστοιχους ανασυνδυασμένους τύπους με παγκόσμια δειγματοληψία που ήταν διαθέσιμες σε παγκόσμιες βάσεις δεδομένων (https://www.hiv.lanl.gov/).

3. Μέθοδοι

3.1 Φυλογενετική και Φυλογεωγραφία

3.1.1 Μοριακή Επιδημιολογία

Η Μοριακή Επιδημιολογία μελετά τη συμβολή των γενετικών και περιβαλλοντικών παραγόντων, που ανιχνεύονται σε μοριακό επίπεδο, στην αιτιολογία, την κατανομή της συχνότητας και την πρόληψη των νοσημάτων στους ανθρώπινους πληθυσμούς. Αποτελεί τη σύζευξη της μοριακής βιολογίας με την επιδημιολογία (26), ενώ παράλληλα βοηθά στην κατανόηση της παθογένειας της νόσου, αναγνωρίζοντας συγκεκριμένες οδούς, μόρια και γονίδια που επηρεάζουν τον κίνδυνο εμφάνισης μιας ασθένειας (27). Σε γενικές γραμμές, επιδιώκει να κατανοήσει τον τρόπο με τον οποίο οι αλληλεπιδράσεις μεταξύ των γενετικών χαρακτηριστικών και των περιβαλλοντικών εκθέσεων οδηγούν στην εμφάνιση ασθενειών (28).

3.1.2 Μοριακή εξέλιξη

Η εξελικτική βιολογία είναι ένα πεδίο της βιολογίας που ασχολείται με τη μελέτη της εξελικτικής διαδικασίας, η οποία παρήγαγε την ποικιλομορφία της ζωής στη γη και προέρχεται από έναν κοινό πρόγονο. Η μοριακή εξέλιξη είναι η διαδικασία της αλλαγής των αλληλουχιών του γενετικού υλικού (DNA, RNA) ανά τις γενιές (στον χρόνο). Το αντικείμενο της μοριακής εξέλιξης είναι ο τομέας της εξελικτικής βιολογίας που μελετά τις εξελικτικές πληροφορίες που περιέχονται στο γενετικό υλικό των οργανισμών και τον τρόπο με τον οποίο η συγκεκριμένη πληροφορία μπορεί να επεξεργαστεί. Η μοριακή επιδημιολογία περιλαμβάνει δυο πεδία. Το πρώτο πεδίο αφορά την μελέτη των αλλαγών που συμβαίνουν στο γενετικό υλικό (DNA, RNA), σε συνδυασμό με την μελέτη του ρυθμού και του εξελικτικού μοντέλου που ακολουθούν αυτές οι αλλαγές στον χρόνο. Το δεύτερο πεδίο αφορά τη μοριακή φυλογένεια που μελετά την εξελικτική ιστορία των οργανισμών βασισμένη σε μοριακά δεδομένα, με τη βοήθεια φυλογενετικών δέντρων. Ο ιός του HIV παρουσιάζει ένα γρήγορο ρυθμό εξέλιξης, και για αυτό το λόγο θεωρείται κατάλληλος για μελέτες που αφορούν την μετάδοση μεταξύ ασθενών για τους οποίους έχουμε διαθέσιμο γενετικό υλικό.

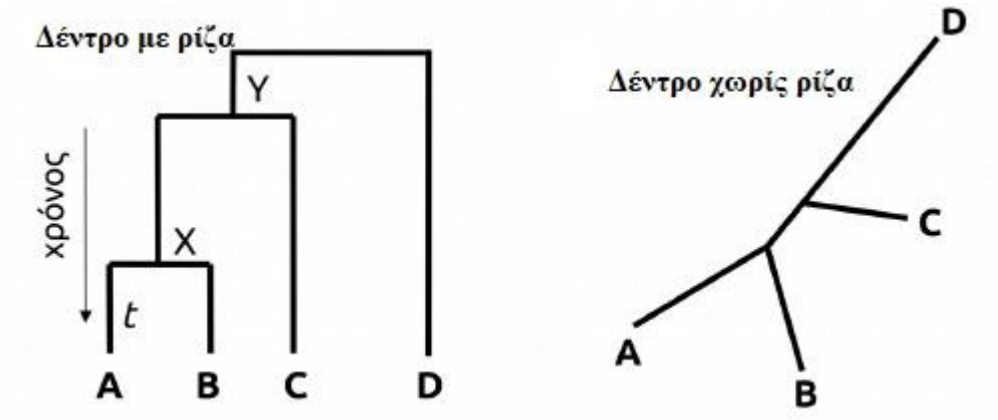
3.2 Φυλογενετική Μεθοδολογία

Στην φυλογενετική ανάλυση εκτιμάμε τις εξελικτικές σχέσεις οργανισμών ή διαφορετικών ατόμων του είδους χρησιμοποιώντας ως δεδομένα το γενετικό τους υλικό. Για να πραγματοποιήσουμε μια φυλογενετική ανάλυση πρέπει πρώτα να στοιχίσουμε της αλληλουχίες DNA ή πρωτεϊνών που μελετάμε. Μπορούμε να στοιχίσουμε της αλληλουχίες εισάγοντας κενά και αντικαταστάσεις στην συστοιχία. Ανάλογα με το πλήθος τον παραπάνω ανέρχεται και το κόστος της στοίχισης. Για να παρέχουμε όσο το δυνατόν μικρότερο κόστος χρησιμοποιούμε αλγορίθμους στοίχισης. Τέτοιοι αλγόριθμοι είναι εκείνοι που χρησιμοποιούν προοδευτική στοίχιση, κριμένα Μαρκοβιανά μοντέλα ή γενετικούς αλγορίθμους.

3.2.1 Φυλογενετικά δένδρα

Ένα φυλογενετικό δέντρο είναι ένα διάγραμμα διακλάδωσης ή “δέντρο” που δείχνει τις εξελικτικές σχέσεις μεταξύ διαφόρων βιολογικών ειδών που η φυλογένειά τους βασίζεται σε ομοιότητες και διαφορές στα φυσικά ή γενετικά τους χαρακτηριστικά. Το δέντρο αποτελείται από κλάδους (branches) και κόμβους (nodes). Η ρίζα (Root)

συμβολίζει τον κοινό πρόγονο των αλληλουχιών. Στα δέντρα με ρίζα (rooted) έχουμε ξεκάθαρη κατεύθυνση του χρόνου, και έτσι μπορούμε να προσδιορίσουμε τον αρχαίο κοινό προγονό. Επίσης μπορούμε να έχουμε δέντρα χωρίς ρίζα (unrooted), στα οποία δεν μπορούμε να προσδιορίσουμε την κατεύθυνση κατά την οποία έχει συντελεστεί η εξελικτική διαδικασία (Εικόνα 8). Τα δέντρα χωρίς ρίζα μπορούν να αναπαρίστανται από ένα δέντρο με ρίζα απλά παραλείποντας τη ρίζα. Για την αντίθετη περίπτωση (δέντρο με ρίζα)χρειαζόμαστε κάποια πληροφορία για τον κοινό πρόγονο. Αυτό συνήθως γίνετε με το να ανάγουμε κάποια εξωτερική ομάδα στις αλληλουχίες, έτσι ώστε η ρίζα να είναι ανάμεσα στην εξωτερική ομάδα και τις υπόλοιπες ταξινομικές μονάδες του δέντρου ή με την εισαγωγή πρόσθετων υποθέσεων σχετικά με τους ρυθμούς εξέλιξης σε κάθε κλαδί, όπως της υποθέσεως του μοριακού ρολογιού. Το μήκος των κλαδιών δείχνει των αριθμό των εξελικτικών αλλαγών στο χρονικό διάστημα που εξελίχθηκαν οι αλληλουχίες. Η τοπολογία των δέντρων υπολογίζεται με τη μέθοδο των αποστάσεων (Distance methods), τη μέθοδο μεγίστης φειδωλότητας (Maximum parsimony), τη μέθοδο μεγίστης πιθανοφάνειας (Maximum Likelihood), και της Μπεϋζιανής συμπερασματολογίας.



Εικόνα 8: Αριστερά έχουμε ένα δέντρο όπου τα μήκη των κλαδιών αντιστοιχούν σε χρόνους απόκλισης, ενώ στα δεξιά έχουμε ένα δέντρο χωρίς ρίζα.

3.2.2 Εξελικτικό Μοντέλο

Για να μπορέσουμε να συγκρίνουμε δυο αλληλουχίες μεταξύ τους πρέπει να γνωρίζουμε τη γενετική τους απόσταση (αναμενόμενος αριθμός γενετικών αλλαγών) η οποία εκτιμά το μέγεθος των εξελικτικών αλλαγών που συνέβησαν στο γενετικό υλικό των υπό μελέτη οργανισμών. Ο παρατηρούμενος αριθμός των αλλαγών μεταξύ δυο αλληλουχιών υποεκτιμά τον αριθμό των πραγματικών αλλαγών. Ο αριθμός των θέσεων (sites) που διαφέρουν μεταξύ δυο αλληλουχιών ονομάζετε ποσοστό ανομοιότητας. Για μια καλύτερη εκτίμηση του πραγματικού αριθμού των αλλαγών χρησιμοποιούμε τα εξελικτικά μοντέλα. Τα μοντέλα αυτά ανήκουν στην γενική κατηγορία των ομογενών (Time-homogeneous) και διαρκώς στάσιμων στον χρόνο (Time-continous stationary Markov models). Αυτά τα μοντέλα Markov δεν απεικονίζουν ρητά τον μηχανισμό της μετάλλαξης ούτε τη δράση της φυσικής επιλογής. Αντίθετα, περιγράφουν τα σχετικά ποσοστά διαφορετικών αλλαγών. Τα μοντέλα αυτά ουσιαστικά αποτελούν τετραγωνικούς 4x4 πίνακες, οι τιμές των οποίων αντιστοιχούν στο ρυθμό με τον οποίο κάθε νουκλεοτίδιο μπορεί να αντικατασταθεί από οποιοδήποτε άλλο. Το πιο απλό είναι ένα μοντέλο μίας παραμέτρου α (Εικόνα 8)

η οποία αντιστοιχεί στο ρυθμό όλων των πιθανών αντικαταστάσεων, οι οποίες είναι ισοπίθανες. Το μοντέλο που προτάθηκε από τους Jukes και Cantor το 1969 (29) αντιστοιχεί σε μια εξαιρετικά απλοποιημένη θεώρηση. Κάθε βάση έχει ίση πιθανότητα π_A →G = π_A →C = π_A →T = α να μετατραπεί σε καθεμία από τις άλλες τρεις βάσεις και πιθανότητα π_A → A = 1-3α να μην αντικατασταθεί καθόλου. Μεταγενέστερα μοντέλα έκαναν βελτιώσεις ως προς τον ρυθμό αντικατάστασης, τη συχνότητα των νουκλεοτιδίων και τον ρυθμό μεταβολής των θέσεων (sites) (Εικόνα 8). Το μοντέλο που χρησιμοποιήθηκε στην εργασία μας είναι το GTR (Generalised time-reversible).

Εκτός από τα μοντέλα που περιγράφουν τους ρυθμούς αντικατάστασης από το ένα νουκλεοτίδιο στο άλλο, υπάρχουν μοντέλα που περιγράφουν τη μεταβολή των επιπέδων μεταξύ των βάσεων σε μια αλληλουχία. Ο διαφορετικός ρυθμός νουκλεοτιδικής αντικατάστασης περιγράφεται με μια συνάρτηση γ-κατανομής, η οποία αντιπροσωπεύει τον ποσοστό των θέσεων που έχουν κοινό ρυθμό αντικατάστασης. Με αυτό τον τρόπο μπορούμε να προσεγγίσουμε καλύτερα τη πραγματική εξελικτική διαδικασία.

3.2.2.1 Ιδιότητες των Εξελικτικών μοντέλων

Τα εξελικτικά μοντέλα χρησιμοποιούν τρεις βασικές παραδοχές. Η πρώτη είναι ότι για οποιαδήποτε χρονική στιγμή, ο ρυθμός αντικατάστασης από ένα νουκλεοτίδιο i σε j είναι ανεξάρτητος από το νουκλεοτίδιο που υπήρχε σε αυτή την θέση πριν το i (ιδιότητα Markov). Η δεύτερη είναι ότι οι ρυθμοί αντικατάστασης δεν μεταβάλλονται με τον χρόνο (ομογένεια) και η τρίτη ότι οι σχετικές συχνότητες είναι σε ισορροπία (στασιμότητα).

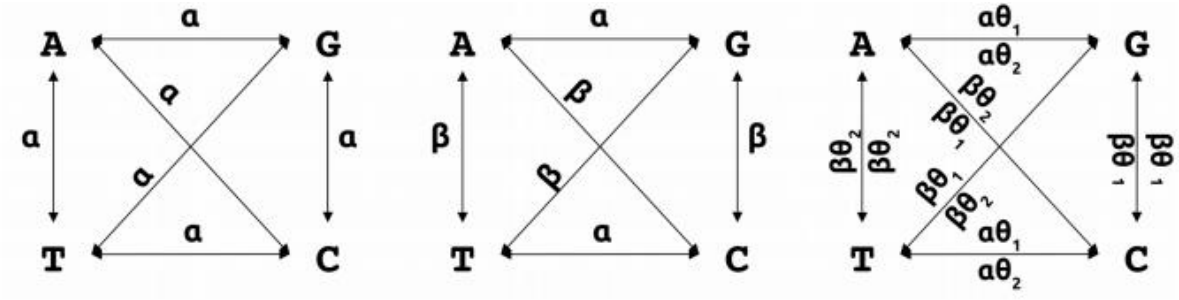
3.2.2.2 Γενικό Μοντέλο νουκλεοτιδικής αντικατάστασης

Τα μοντέλα που χρησιμοποιούμε θέλουμε να προσεγγίζουν όσο το δυνατόν καλύτερα την πραγματικότητα. Στην γενική περίπτωση θεωρούμε ότι η διαδικασία αντικατάστασης στο DNA περιγράφετε από έναν πίνακα Q, στον οποίο περιλαμβάνονται όλες οι παράμετροι του εξελικτικού μοντέλου (ρυθμός αντικατάστασης, συχνότητα των νουκλεοτιδίων κλπ.).

$$Q = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -(a\pi_C + b\pi_A + c\pi_G) & a\pi_C & b\pi_A & c\pi_G \\ g\pi_T & -(g\pi_T + d\pi_A + e\pi_G) & d\pi_A & e\pi_G \\ h\pi_T & i\pi_C & -(h\pi_T + i\pi_C + f\pi_G) & f\pi_G \\ j\pi_T & k\pi_C & l\pi_A & -(j\pi_T + k\pi_C + l\pi_A) \end{pmatrix} \end{array}$$

Οπού:

π→ συχνότητα εμφάνισης του κάθε νουκλεοτιδίου
a,b,⋯,l→ ρυθμός αντικατάστασης (πχ ο δεύτερος όρος της πρώτης γραμμής είναι ο ρυθμός αντικατάστασης από A σε C).
Τα στοιχεία της διατόνου ισούνται με 1- το άθροισμα των υπόλοιπων στοιχείων της γραμμής έτσι ώστε το άθροισμα σε κάθε γραμμή να ισούται με 1. Κάθε μη διαγώνιο στοιχείο αναπαριστά τον ρυθμό αντικατάστασης από το νουκλεοτίδιο i στο j.



Εικόνα 9: Διαφορετικά μοντέλα νουκλεοτιδικών αντικαταστάσεων από το απλούστερο μοντέλο μιας παραμέτρου (Jukes-Cantor), στο διπαραμετρικό του Kimura και στο σύνθετο μοντέλο των Tamura-Nei.

3.2.2.3 Μοντέλο GTR

Το GTR είναι το πιο γενικά ουδέτερο, ανεξάρτητο, πεπερασμένο, χρονικά αναστρέψιμο μοντέλο. Αρχικά περιγράφτηκε σε γενική μορφή από τον Simon Tavaré το 1986 (30). Οι παράμετροι του GTR αποτελούνται από ένα διάνυσμα συχνοτήτων βάσης ισορροπίας, $\Pi = (\pi_T, \pi_C, \pi_A, \pi_G)$, δίνοντας τη συχνότητα με την οποία κάθε βάση εμφανίζεται σε κάθε θέση και τον πίνακα του ρυθμού μεταβολής, και είναι:

$$Q = \begin{pmatrix} -(\alpha\pi_C + \beta\pi_A + \gamma\pi_G) & \alpha\pi_C & \beta\pi_A & \gamma\pi_G \\ \alpha\pi_T & -(\alpha\pi_T + \delta\pi_A + \varepsilon\pi_G) & \delta\pi_A & \varepsilon\pi_G \\ \beta\pi_T & \delta\pi_C & -(\beta\pi_T + \delta\pi_C + \eta\pi_G) & \eta\pi_G \\ \gamma\pi_T & \varepsilon\pi_C & \eta\pi_A & -(\gamma\pi_T + \varepsilon\pi_C + \eta\pi_A) \end{pmatrix}$$

,όπου:

$$\alpha = r(T \to C) = r(C \to T)$$

$$\beta = r(T \to A) = r(A \to T)$$

$$\gamma = r(T \to G) = r(G \to T)$$

$$\delta = r(C \to A) = r(A \to C)$$

$$\varepsilon = r(C \to G) = r(G \to C)$$

$$\eta = r(A \to G) = r(G \to A)$$

, είναι οι παράμετροι του ρυθμού της μεταβολής.

Επομένως, το GTR (για τέσσερις χαρακτήρες) απαιτεί έξι παραμέτρους ρυθμού αντικατάστασης, καθώς και τέσσερις παραμέτρους συχνότητας βάσης ισορροπίας. Ωστόσο, αυτό συνήθως εξαλείφεται σε εννιά παραμέτρους συν το μ, που είναι συνολικός χρόνος των αντικαταστάσεων για κάθε μονάδα χρόνου. Όταν μετράμε τον αριθμό των αντικαταστάσεων το μ είναι ίσο με 1. Γενικά, για να υπολογίσουμε τον αριθμό των παραμέτρων θα πρέπει να μετρήσουμε τον αριθμό των παραμέτρων του πίνακα κάτω από την διαγώνιο.
Οπότε έχουμε:

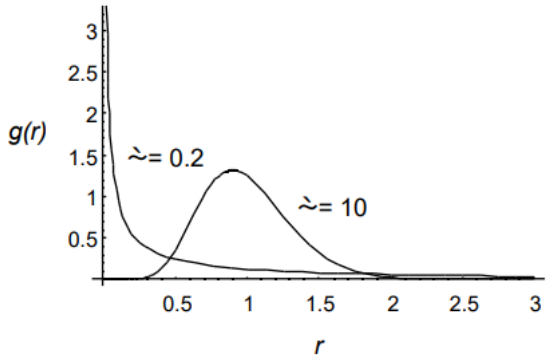
$$\frac{n^2-n}{2} + n - 1$$

3.2.2.4 Μοντέλο Γ με διαφορετικό ρυθμό εξέλιξης

Το μοντέλο Γ είναι το πιο διαδεδομένο μοντέλο με διαφορετικό ρυθμό εξέλιξης και χρησιμοποιεί μια Γ κατανομή με αναμενόμενη μέση τιμή 1 και διακύμανση 1/α για την εκτίμηση των σχετικών ρυθμών νουκλεοτιδικής αντικατάστασης στις θέσεις της αλληλουχίας.

$$g(r) = \frac{a^a r^{a-1}}{e^{ar}\Gamma(a)}$$

Αλλάζοντας της τιμές της παραμέτρου α παίρνουμε δύο διαφορετικά σενάρια (Εικονα.11). Για την περίπτωση που ο ρυθμός αντικατάστασης διαφέρει σημαντικά μεταξύ των θέσεων το α παίρνει τιμές μικρότερες του 1 (α<1), ενώ όταν το α είναι μεγαλύτερο του 1 (α>1) τότε ο ρυθμός νουκλεοτιδικής αντικατάστασης είναι παρόμοιος στην πλειοψηφία των θέσεων. Αυτό δείχνει ότι υπάρχουν θέσεις στις αλληλουχίες που έχουν πολύ μεγάλες σχετικές συχνότητες rs ενώ πολλές άλλες τοποθεσίες είναι σχεδόν αμετάβλητες (rs ≈ 0).



Εικόνα10: Συνάρτηση Γ κατανομής για διαφορετικές τιμές α. Για α=10 έχουμε το σχήμα παρουσιάζει μορφή κορδονιού, ενώ για α=0.2 είναι σε σχήμα Λ.

3.2.3 Μέθοδοι Απόστασης

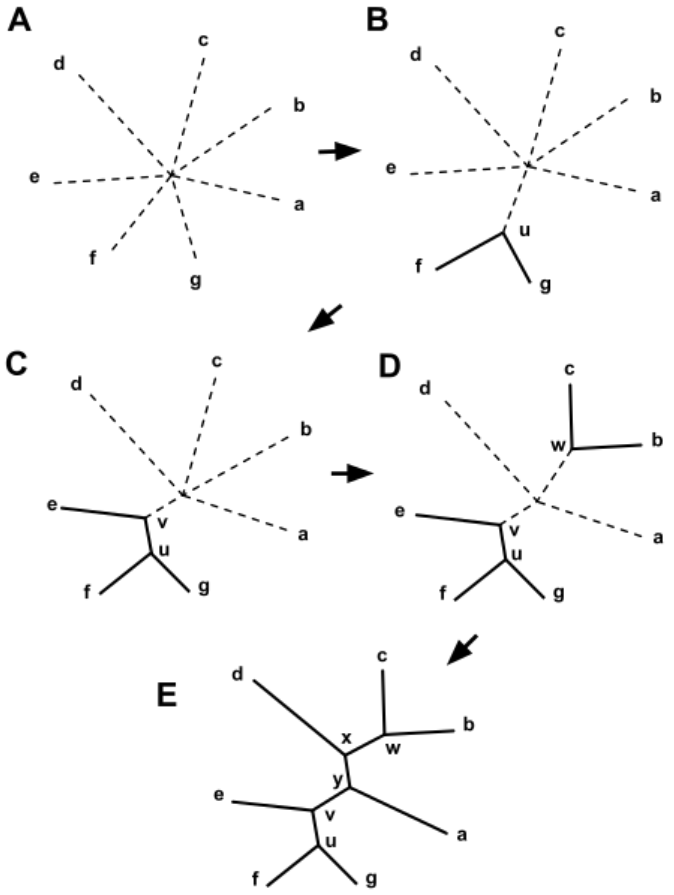
Οι μέθοδοι απόστασης όπως είναι η Neighbor-joining, η WPGMA ή η UPGMA, υπολογίζουν την γενετική απόσταση από τις πολλαπλές στοιχίσεις αλληλουχιών. Συγκρίνουν δυο συστοιχίες κάθε φορά δημιουργώντας έναν πίνακα όλων των πιθανών ζευγαριών των αλληλουχιών. Έπειτα κατά την διάρκεια κάθε σύγκρισης υπολογίζεται ο αριθμός των αλλαγών και παρουσιάζεται σαν την αναλόγια του συνολικού μήκους των αλληλουχιών. Αυτές οι εκτιμήσεις της διαφοράς μεταξύ όλων των πιθανών ζευγαριών αλληλουχιών είναι γνώστη ως απόσταση κατά ζεύγη (pairwise distance). Ακολούθως, οι αποστάσεις εισάγονται σε ένα δέντρο. Υπάρχουν πολλοί τρόποι για να "εισαχθούν" το είδος ή τα γονίδια ανάλογα με την απόστασή τους. Ένας τρόπος να συγκεντρωθούν ή να βελτιστοποιηθούν οι αποστάσεις είναι να ενωθούν είδη ή γονίδια μαζί βάσει των αυξανόμενων διαφορών τους σύμφωνα με τις αποστάσεις τους. Άλλοι τρόποι χρησιμοποιούν διάφορους συντελεστές για να μετρήσουν πόσο καλά τα μήκη των κλαδιών του δέντρου αντανακλούν τις αρχικές αποστάσεις ανά ζεύγος.(31)

3.2.3.1 UPGMA

Η μέθοδος UPGMA (Unweighted Pair Group Method with Arithmetic Mean) είναι μια απλή μέθοδος ιεραρχικού clustering (συστάδες) και είναι η απλούστερη μέθοδος κατασκευής φυλογενετικών δέντρων. Η μέθοδος αποδίδεται στους Sokal και Michener (32). Ο αλγόριθμος UPGMA κατασκευάζει ένα δέντρο με ρίζα (δενδρόγραμμα) που αντανακλά τη δομή που υπάρχει σε έναν pricewise πίνακα (ή έναν πίνακα ανομοιότητας). Σε κάθε βήμα, οι δυο πλησιέστερες συστάδες συνδυάζονται σε ένα σύμπλεγμα υψηλότερου επιπέδου. Η απόσταση μεταξύ οποιονδήποτε δυο συστάδων Α και Β, το καθένα μεγέθους |A| και |B|, είναι ο μέσος όλων των αποστάσεων d(x,y) μεταξύ ζευγών του αντικείμενου χ στο Α και y στο Β, το οποίο είναι η μέση απόσταση των στοιχειών κάθε συστάδας ($\frac{1}{|A|*B \vee \sum_{X \in A} \sum_{y \in B} d(x,y)}$). Ο αλγόριθμος μπορεί να εφαρμοστεί στην περίπτωση που ο ρυθμός εξέλιξης των ταξινομικών μονάδων είναι σταθερός και η γενετική του απόσταση υπολογίζεται ανάλογα με το εξελικτικό μοντέλο που επιλέγεται.

3.2.3.2 Neighbor-joining

Η μέθοδος αυτή είναι παρόμοια με την UPGMA χωρίς όμως να κάνει την παραδοχή ότι ο ρυθμός εξέλιξης των αλληλουχιών είναι σταθερός. Δημιουργήθηκε από τους Naruya Saitou και Masatoshi Nei το 1987 (33). Η μέθοδος χρησιμοποιεί το κριτήριο της ελαχίστης εξέλιξης και συνδυάζει ένα ζεύγος αλληλουχιών ελαχιστοποιώντας την τιμή του μήκους του κλαδιού σε κάθε βήμα βρίσκοντας το ζεύγος των γειτονικών ταξινομικών μονάδων (Εικόνα 11).



Εικόνα 11: Ο αλγόριθμος Neighbor-Joining, ξεκινά με ένα πλήρως ανεπίλυτο δέντρο, του οποίου η τοπολογία αντιστοιχεί σε αυτό ενός δικτύου αστέρων και ύστερα πραγματοποιεί κάποια βήματα έως ότου το δέντρο λυθεί και όλα τα μήκη των κλαδιών είναι γνωστά (34).

3.2.4 Μέθοδος μεγίστης Πιθανοφάνειας

Η μέγιστη πιθανοφάνεια είναι μια γενική στατιστική μέθοδος για την εκτίμηση άγνωστων παραμέτρων ενός μοντέλου πιθανότητας. Στη φυλογενετική η πιθανοφάνεια ορίζεται ως μια ποσότητα ανάλογη της πιθανότητας παρατήρησης των δεδομένων που δίδονται από το μοντέλο P (D | M ,T) ,οπού D είναι τα δεδομένα , M το μοντέλο και T τα δέντρα. Η μέθοδος μεγίστης πιθανοφάνειας χρησιμοποιεί εξελικτικά μοντέλα αλληλουχιών DNA με κοινούς η διαφορετικούς ρυθμούς εξέλιξης. Η μέθοδος αυτή εκτίμα όλες τις πιθανές σχέσεις για όλα τα δέντρα. Έτσι, αν έχουμε ένα μοντέλο μπορούμε να υπολογίσουμε την πιθανότητα οι παρατηρήσεις να είχαν πράγματι λάβει χώρα ως συνάρτηση του μοντέλου. Στη συνέχεια, εξετάζουμε αυτή την πιθανοφάνεια για να δούμε πού μεγιστοποιείται. Το σημείο οπού μεγιστοποιείται η παράμετρος που μας ενδιαφέρει (συνήθως το μήκος των κλαδιών ή το δέντρο) είναι ο εκτιμητής μέγιστης πιθανοφάνειας της παραμέτρου. Συνήθως οι πιθανότητες μετατρέπονται σε λογαριθμική μορφή, μέσω μιας συνάρτησης λογαριθμικού σκορ, προτού επιλεχτεί το δέντρο για την καλύτερη αναπαράσταση του.

3.2.4.1 Πιθανοφάνεια

Η πιθανοφάνεια στη φυλογενετική ορίζετε ως η πιθανότητα να παρατηρήσουμε τα δεδομένα (αλληλουχίες DNA) δοθέντος ενός μοντέλου.

$L(\tau,\theta)$= Prob(Δεδομένα | δέντρο, εξελικτικό μοντέλο)

Γενικότερα εάν έχουμε ένα δείγμα $\underline{X}$ το οποίο έχει συνάρτηση πυκνότητας $f(\underline{x},\theta), με \theta \in \Theta$, όπου θ είναι η άγνωστη παράμετρος, τότε η πιθανοφάνεια ορίζετε ως:

$$L(\theta|\underline{x}) = f(\underline{x},\theta), \theta \in \Theta$$

, με:

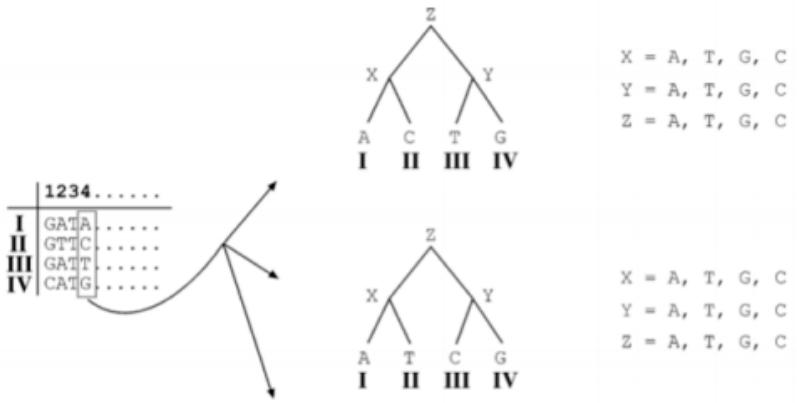
$$f(\underline{x},\theta) = f(x1, x2, \dots, x\nu|\theta) = f(x1|\theta) * f(x2|\theta) * \dots * f(x\nu|\theta)$$

Δηλαδή η συνάρτηση πιθανοφάνειας είναι η πυκνότητα του $\underline{X}$,$f(\underline{x},\theta)$ υπολογιζόμενη στην παραχωρηθείσα τινά $\underline{x}$ του $\underline{X}$ και θεωρούμενη ως συνάρτηση του θ (με σταθερό $\underline{x}$). Στο σημείο που το $L(\tau,\theta)$ μεγιστοποιείται βρίσκεται και ο εκτιμητής μεγίστης πιθανοφάνειας (ΕΜΠ). Εάν θεωρήσουμε το απλό παράδειγμα ενός κέρματος με δεδομένα (ρίψεις) : ΚΚΓΚΓΚ και με μοντέλα και πιθανότητα :

M1: τίμιο νόμισμα P(K)= 0,5 , P(Γ)= 0,5
M2: διπλή κορόνα P(K)= 1 , P(Γ)= 0
M3: δίπλα γράμματα P(K)= 0 , P(Γ)= 1
Τότε η πιθανοφάνεια για τα μοντέλα θα ήταν ,
L (Δεδομενα| M1) = P(K|M1)*P(K|M1)*P(Γ|M1)*P(K|M1)*P(Γ|M1)*P(K|M1)= 0.5*0.5*0.5*0.5*0.5*0.5=0.0156
L(Δεδομενα| M2)= 1*1*0*1*0*1= 0
L(Δεδομενα| M1) = 0*0*1*0*1*0=0
Άρα το μοντέλο που μεγιστοποιεί την πιθανοφάνεια των δεδομένων είναι το M1
Παρόμοια στην περίπτωση μιας αλληλουχίας: GGACGCCTGACGCCGCTCGG
όπου
M1: με ίσους ρυθμούς εξέλιξης για το A,C,G,T P=0.25
M2: υπερεκτίμα το G και το C P(G)=P(C)=0.4 , P(A)=P(T)=0.1
M3: υπερεκτίμα το A και το T P(G)=P(C)=0.1 , P(A)=P(T)=0.4
θα είχαμε
L(Δεδομενα| M1)=$P(G|M1) * P(G|M1) * P(A|M1) * \dots * P(G|M1) = 0{,}25^{20}$
L(Δεδομενα| M2)=$0.4^{16}0.1^4 = 4.3 * 10^{-11}$
L(Δεδομενα| M2)=$0.1^{16}0.4^4 = 2.6 * 10^{-18}$
Άρα το μοντέλο που μεγιστοποιεί την πιθανοφάνεια των δεδομένων είναι το M2 και άρα οι αλληλουχίες είναι πιθανό να προσήλθαν από το μοντέλο 2

Στην περίπτωση που θέλουμε να βρούμε ένα φυλογενετικό δέντρο με την μέθοδο της μέγιστης πιθανοφάνειας ας υποθέσουμε ότι έχουμε μια συστοιχία από 4 αλληλουχίες με N θέσεις. Για να υπολογίσουμε την πιθανοφάνεια στην θέση 4 (A, C, T , G) για παράδειγμα (Εικόνα 12) πρέπει να λάβουμε υπόψη όλα τα πιθανά σενάρια για το πώς εξελιχτήκαν οι αλληλουχίες. Αν υποθέσουμε ότι αναπαριστούμε τους προγονικούς

χαρακτήρες στους εσωτερικούς κόμβους με X και Y και την ρίζα με Z τότε για κάθε δυνατή τοπολογία θα πρέπει να εκτιμήσουμε την πιθανότητα να παρατηρήσουμε τα συγκεκριμένα νουκλεοτίδια της θέσης 4 λαμβάνοντας υπόψη όλα τα δυνατά σενάρια για τα νουκλεοτίδια στους εσωτερικούς κόμβους.



Εικόνα 12: Γραφική απεικόνιση της κατασκευής φυλογενετικών δέντρων με την μέθοδο της μέγιστης πιθανοφάνειας.

$$L_4 = P(Z \to X) * P(Z \to Y) * P(X \to A) * P(X \to C) * P(Y \to T) * P(Y \to G)$$
$$lnL_4 = lnP(Z \to X) + lnP(Z \to Y) + lnP(X \to A) + lnP(X \to C) + lnP(Y \to T) + lnP(Y \to G)$$

Το παράδειγμα δίνει μονό κάποιες από τις τοπολογίες από μια από τις θέσεις, η μέθοδος όμως χρησιμοποιεί όλες τις τοπολογίες για όλες τις θέσεις.

3.2.5 Μέθοδος μεγίστης Φειδωλότητας (Maximum parsimony)

Η μέθοδος μέγιστης φειδωλότητας είναι μια μέθοδος που χρησιμοποιείται για τον υπολογισμό της τοπολογίας ενός φυλογενετικού δένδρου. Σε αυτήν την προσέγγιση ο στόχος είναι να προσδιοριστεί η φυλογένεση που απαιτεί τις λιγότερες απαραίτητες αλλαγές για να εξηγήσει τις διαφορές μεταξύ των αλληλουχιών. Οι βασικές ιδέες πίσω από τη μέγιστη φειδωλότητα παρουσιάστηκαν από τον James S. Farris το 1970 και τον Walter M. Fitch το 1971 (35-36).

3.2.6 Αξιολόγηση φυλογενετικών δέντρων

Για τον στατιστικό έλεγχο των υποθέσεων φυλογενετικών δέντρων χρησιμοποιούμε κυρίως τη μέθοδο bootstrap (40). Στη μέθοδο αυτή ανακατατάσσουμε με τυχαίο τρόπο τις θέσεις στη στοίχιση. Για κάθε ανακαταταγμένη στοίχιση, δημιουργείται ένα δέντρο με την ίδια μεθοδολογία που ακολουθήθηκε κατά την κατασκευή του αρχικού δέντρου. Στο δέντρο που προκύπτει από κάθε ανακατάταξη καταμετρώνται οι κόμβοι που βρίσκονται στην ίδια θέση με το αρχικό δέντρο. Αν ένας κόμβος ενώνει τα ίδια δύο φύλλα, η τιμή bootstrap του αυξάνεται κατά 1. Η διαδικασία αυτή επαναλαμβάνεται για ένα μεγάλο αριθμό φορών (συνήθως N>100). Στο τέλος των επαναλήψεων, αποδίδεται σε κάθε κόμβο η τιμή bootstrap που έχει υπολογιστεί ως ποσοστό με βάσει τον αριθμό N των επαναλήψεων. Μεγάλες τιμές bootstrap στους

κόμβους δηλώνουν ότι ο κόμβος είναι στατιστικά σημαντικός και μεγάλες τιμές για όλους τους κόμβους δηλώνουν ότι το δέντρο είναι αξιόπιστο.

3.2.7 RAxML

Για την εκτίμηση των φυλογενετικών δέντρων χρησιμοποιήθηκε το πρόγραμμα RAxML στην διαδικτυακή πλατφόρμα CIPRES. Το σύνολο των bootstrap δένδρων για την εύρεση του τελικού δέντρου μέγιστης πιθανοφάνειας ήταν για το στέλεχος BF 402 δέντρα και για το BG 504 δέντρα. Στην διαδικασία χρησιμοποιήθηκαν το εξελικτικό μοντέλο GTR σε συνδυασμό με την συνάρτηση γ-κατανομής. Το RAXML–HPC BlackBox (Randomized Axelerated Maximum Likelihood) είναι ένα πρόγραμμα διαδοχικών και παράλληλων μέγιστων πιθανοφανειών που βασίζει τα συμπεράσματα του σε μεγάλα φυλογενετικά δέντρα. Το πρόγραμμα έχει σχεδιαστεί ειδικά για να παράγει αποτελεσματικά δέντρα για εξαιρετικά μεγάλα σύνολα δεδομένων, είτε από την άποψη του αριθμού των ταξινομικών μονάδων είτε του μήκους αλληλουχιών. Μπορεί επίσης να χρησιμοποιηθεί και για την εκ των υστέρων ανάλυση συνόλων φυλογενετικών δέντρων καθώς και για την ανάλυση αλληλουχιών. Το πρόγραμμα αυτό προήλθε από το fastDNAml, το οποίο με τη σειρά του προέρχεται από το dnaml του Joe Felsentein που αποτελεί μέρος του πακέτου PHYLIP (39).
Σε κάθε αναδιάταξη, κρατιέται μια λίστα με τα είκοσι καλύτερα δέντρα. Μετά από κάθε βήμα αναδιάταξης πραγματοποιείται η διαδικασία της βελτιστοποίησης του μήκους των κλαδιών για τις είκοσι αυτές τοπολογίες. Οι αναδιατάξεις και οι βελτιστοποιήσεις του μήκους των κλαδιών επαναλαμβάνονται μέχρι να καλυφθούν ορισμένα κριτήρια σύγκλισης, οπότε και σταματάει η διαδικασία. Τότε, τα είκοσι καλύτερα δέντρα που έχουν προκύψει συνδυάζονται μεταξύ τους χρησιμοποιώντας το πρόγραμμα Consense του πακέτου PHYLIP και προκύπτει το τελικό δέντρο με την μεγίστη πιθανοφάνεια. Το πρόγραμμα ξοδεύει τον περισσότερο χρόνο (95% περίπου) του στον υπολογισμό του βαθμού πιθανοφάνειας για χιλιάδες πιθανές τοπολογίες δέντρων. Το ίδιο πρόβλημα παρουσιάζουν και προγράμματα Μπεϋζιανής φυλογενετικής ανάλυσης καθώς βασίζονται στον υπολογισμό φυλογενετικής πιθανοφάνειας.

3.2.8 Fig Tree

Η γραφική αναπαράσταση των δέντρων που προέκυψαν από το RAxML αλλά και αργότερα στο BEAST έγινε με το πρόγραμμα FigTree (http://tree.bio.ed.ac.uk/software/figtree/). Το FigTree είναι ένα πρόγραμμα που αναπαριστά φυλογενετικά δέντρα από άλλα προγράμματα, τα επεξεργάζεται οπτικά και τα εξάγει σε μορφή εικόνων.

3.3 Αναζήτηση και στοίχιση

Αρχικά μέσω της ιστοσελίδας της παγκόσμιας βάσης δεδομένων για τον HIV (HIV Sequence Database, https://www.hiv.lanl.gov/content/index) έγινε αναζήτηση χρησιμοποιώντας τον αλγόριθμο HIV blast με σκοπό να βρεθούν οι δέκα πιο όμοιες αλληλουχίες για κάθε μια αλληλουχία BF και BG του πληθυσμού της μελέτης. Ύστερα ακολούθησε απαλοιφή των διπλότυπων αλληλουχιών. Τα τελικά αρχεία περιλάμβαναν 264 αλληλουχίες για το ανασυνδυασμένο στέλεχος BF και 265 για το BG.

Η στοίχιση των αλληλουχιών έγινε μέσω της μεθόδου muscle στην ιστοσελίδα CIPRES (http://www.phylo.org/). Η ιστοσελίδα αυτή είναι ένας δημόσιος ιστότοπος ο οποίος έχει σχεδιαστεί για να παρέχει πρόσβαση σε μεγάλη υπολογιστική ισχύ από το NSF XSEDE.

3.3.1 MUSCLE

Το MUSCLE είναι ένα πρόγραμμα που χρησιμοποιείται για την πολλαπλή στοίχιση αλληλουχιών. Τα στοιχεία του αλγόριθμου περιλαμβάνουν την εκτίμηση ταχείας απόστασης χρησιμοποιώντας την μέτρηση k mer (για μη στοιχισμένες αλληλουχίες) ή την Kimura (για στοιχισμένες αλληλουχίες), την προοδευτική στοίχιση χρησιμοποιώντας μια συνάρτηση που ονομάζουμε score log expectation (προκειμένου να γίνει στοίχιση κατά ζεύγη), και βελτιστοποίηση χρησιμοποιώντας εξαρτώμενη από τα δέντρα δεσμευμένη κατανομή. Το MUSCLE είναι το πιο γρήγορο και το πιο ακριβές από τα προγράμματα στοίχισης όπως το T-Coffee, το MAFFT και το CLUSTALW (36-37).

log- expectation (LE) score:

$$LE_{xy} = (1 - f_{xG})\,(1 - f_{yG})\log \Sigma\, i\, \Sigma\, jf_{x}if_{y}jpij \,/\, pipj$$

3.3.2 MEGA

Αφού πραγματοποιήθηκε η στοίχιση των αλληλουχιών, στη συνέχεια οι αλληλουχίες επεξεργαστήκαν και διαμορφώθηκαν στην τελική τους μορφή στο πρόγραμμα MEGA (Molecular Evolutionary Genetic Analysis). Το MEGA είναι ένα πρόγραμμα πολλαπλής στοίχισης και επεξεργασίας αλληλουχιών σε γραφικό περιβάλλον και χρησιμοποιεί μεθόδους αποστάσεων, φειδωλότητας και πιθανοφάνειας σε μοριακά δεδομένα (38). Οι τελικές μας αλληλουχίες είχαν μήκος 846 νουκλεοτίδια.

3.4 Φυλοδυναμική ανάλυση

Η φυλοδυναμική ορίζεται ως η μελέτη του τρόπου με τον οποίο οι επιδημιολογικές, ανοσολογικές και εξελικτικές διεργασίες δρουν και δυνητικά αλληλοεπιδρούν για να διαμορφώσουν τις ιογενείς φυλογένειες (41). Η έρευνα σχετικά με τη ιογενή φυλοδυναμική επικεντρώθηκε στη δυναμική της μετάδοσης σε μια προσπάθεια να αποκαλύψει τον τρόπο με τον οποίο η δυναμική αυτή επηρεάζει τις ιογενείς γενετικές παραλλαγές. Η δυναμική μετάδοσης μπορεί να εξεταστεί σε επίπεδο κυττάρων εντός μολυσμένου ξενιστή, μεμονωμένων ξενιστών εντός ενός πληθυσμού ή ολόκληρων πληθυσμών ξενιστών. Πολλοί ιοί, ιδιαίτερα οι ιοί RNA, συσσωρεύουν γρήγορα γενετικές παραλλαγές λόγω των μικρών χρόνων παραγωγής και των υψηλών ποσοστών μετάλλαξης. Επομένως, τα μοτίβα της γενετικής παραλλαγής του ιού επηρεάζονται σημαντικά από το πόσο γρήγορα συμβαίνει η μετάδοση και από ποιες οντότητες γίνεται η μετάδοση. Λόγω των επιπτώσεων που μπορεί να έχει η δυναμική και η επιλογή της μετάδοσης στη γενετική ποικιλομορφία του ιού, οι ιογενείς φυλογένειες μπορούν να χρησιμοποιηθούν για τη διερεύνηση σημαντικών επιδημιολογικών, ανοσολογικών και εξελικτικών διεργασιών, όπως η εξάπλωση της επιδημίας, η χωροχρονική δυναμική συμπεριλαμβανομένης της δυναμικής μεταπολλαπλασιασμού, η ζωονοτική μετάδοση, ο τροπισμός των ιστών και η αντιγονική μετατόπιση. Η ποσοτική διερεύνηση αυτών των διεργασιών μέσω της εξέτασης των φυλογενειών είναι ο κεντρικός στόχος της φυλοδυναμικής. Ο στόχος

των φυλοδυναμικών αναλύσεων είναι να βγάλουν συμπεράσματα επιδημιολογικού χαρακτήρα από ιικές φυλογένειες. Έτσι, οι περισσότερες φυλοδυναμικές αναλύσεις ξεκινούν με την ανακατασκευή ενός φυλογενετικού δέντρου. Οι γενετικές αλληλουχίες συχνά υποβάλλονται σε δειγματοληψία σε πολλαπλά χρονικά σημεία, γεγονός που επιτρέπει την εκτίμηση των ποσοστών υποκατάστασης χρησιμοποιώντας μοντέλο μοριακού ρολογιού. Για τους ιούς, οι μέθοδοι Bayes είναι δημοφιλείς λόγω της ικανότητας να ταιριάζουν με σύνθετα δημογραφικά σενάρια ενώ ενσωματώνουν τη φυλογενετική αβεβαιότητα.

3.4.1 BEAST

Για την φυλοδυναμική ανάλυση και την εύρεση των τοπικών επιδημιών για τα ανασυνδυασμένα στελέχη των υποτύπων BF και BG έγινε χρήση του προγράμματος BEAST.
Το BEAST είναι πρόγραμμα cross-platform για Μπεϋζιανή ανάλυση αλληλουχιών χρησιμοποιώντας τον αλγόριθμο MCMC. Το πρόγραμμα είναι προσανατολισμένο προς την ανάλυση μοριακού ρολογιού. Μπορεί να χρησιμοποιηθεί ως μέθοδος κατασκευής της φυλογένεσης, αλλά προορίζεται επίσης να δοκιμάσει εξελικτικές υποθέσεις χωρίς προϋποθέσεις σε μια ενιαία τοπολογία δέντρων. Το BEAST χρησιμοποιεί το MCMC για να υπολογίσει κατά μέσο όρο το διάστημα των δένδρων, έτσι ώστε κάθε δέντρο να είναι ανάλογο με την πιθανότητα του προηγούμενου. Χρησιμοποιεί μια μορφή εισόδου XML που επιτρέπει στο χρήστη να σχεδιάσει και να εκτελέσει ένα ευρύ φάσμα μοντέλων. Περιλαμβάνει επίσης μια γραφική εφαρμογή που παράγει αυτή τη μορφή υπό μια ευρεία ποικιλία μοντέλων (TRACER). Για τη δημιουργία των αρχείων εισόδου του BEAST χρησιμοποιούμε το πρόγραμμα BEAUti (Baysian Evolutionary Analysis Utility), στο οποίο ορίζονται τα δεδομένα και οι παράμετροι του μοντέλου. Το BEAUti είναι τμήμα του λογισμικού BEAST και δέχεται αρχεία σε μορφή Nexus, όπου οι αλληλουχίες του έχουν υποστεί πολλαπλή στοίχιση και κατάλληλη επεξεργασία. Στην ανάλυση μας χρησιμοποιήσαμε την έκδοση 1.8.0 του BEAST και του BEAUti (42). Για όλα τα δεδομένα χρησιμοποιήθηκε η ίδια παραμετροποίηση στο BEAUti. Για την ανεύρεση των τοπικών επιδημιών του BG έγινε ανάλυση για δυο διαφορετικά ανασυνδυασμένα στελέχη, το CRF14 και το CRF20, ενώ για το BF έγινε ανάλυση για το CFR47.

Το BEAUti αποτελείται από 11 καρτέλες (43).

1. Partitions: Μας δίνει μια συνολική περιγραφή των δεδομένων. Για το BF_CRF47 είχαμε 32 αλληλουχίες ενώ είχαμε 14 αλληλουχίες για το BG_CRF14 και 38 για το BG_CRF20 με μήκος 846 νουκλεοτιδίων.

2. Taxa: Στην καρτέλα αυτή μπορούμε να δημιουργήσουμε υποομάδες αλληλουχιών και στην συνέχεια να δώσουμε prior κατανομή για την εκτίμηση του χρόνου προέλευσης της κάθε υποομάδας. Στην ανάλυση μας δεν χρειάστηκε να δημιουργήσουμε κάποια υποομάδα.

3. Tips: Σε αυτή τη καρτέλα ορίζουμε τον τρόπο που έχει δοθεί η παράμετρος του χρόνου στην ονομασία των αλληλουχιών μας. Λειτουργεί προσπαθώντας να βρει ένα αριθμητικό πεδίο σε κάθε όνομα. Εάν τα ονόματα των στοιχείων περιέχουν περισσότερα από ένα αριθμητικά πεδία τότε μπορεί ο ερευνητής να

καθορίσει τον τρόπο ανίχνευσης αυτού που αντιστοιχεί στην ημερομηνία της δειγματοληψίας.

4. Traits: Εδώ ορίζουμε της παραμέτρους της Μπεϋζιανής φυτογεωγραφικής ανάλυσης. Στην ανάλυση μας δεν χρειάστηκε να πραγματοποιηθεί Μπεϋζιανή φυλογεωγραφία.

5. Sites: Σε αυτή τη καρτέλα ορίζουμε το εξελικτικό μοντέλο και της παραμέτρους του. Στην ανάλυση μας επιλέξαμε το μοντέλο GTR (substitution model) και για την κατανομή Γ (Site Heterogeneity Model) για τους διαφορετικούς ρυθμούς εξέλιξης.

6. Clock: Εδώ ορίζουμε τις παραμέτρους του μοντέλου για την εκτίμηση του μοριακού ρολογιού. Επιλέξαμε το μοντέλο log-normal Relaxed Clock.

7. Trees: Σε αυτό το σημείο γίνεται η επιλογή του μοντέλου βάσει του οποίου θα εκτιμήσουμε τα δημογραφικά χαρακτηριστικά. Επιλέξαμε το μοντέλο Byaesian Skyline Coalescent, που αποτελεί μια μη παραμετρική προσέγγιση για την εκτίμηση των δημογραφικών δεδομένων. Το μοντέλο αυτό διαιρεί τον χρόνο μεταξύ του παρόντος και της ρίζας του δέντρου σε διαστήματα. Κάθε διάστημα θα έχει διαφορετικό αποτελεσματικό μέγεθος πληθυσμού. Το μοντέλο θα υπολογίζει τον αριθμό των γεγονότων (coalescent events) σε κάθε διάστημα, το οποίο δίνεται από την παράμετρο group size (αριθμός ομάδων), καθώς και το αποτελεσματικό μέγεθος πληθυσμού για το διάστημα αυτό. Ο αριθμός των ομάδων που θα ορίζεται εξαρτάται από των αριθμό των αλληλουχιών. Για το BF_CRF47 και το BG_CRF20 ορίσαμε οκτώ ομάδες και για το BG_CRF14 τέσσερις ομάδες. Τέλος επιλέξαμε το αρχικό δέντρο να εκτιμηθεί μέσω της μεθόδου UPGMA.

8. States: Εδώ ορίζουμε τις παραμέτρους του μοντέλου της φυλογεωγραφικής ανάλυσης. Μπορούμε να επιλέξουμε την εκτίμηση στους προγονικούς χαρακτήρες σε κάθε κόμβο της χρονολογικής τοπολογίας ή μόνο τον προγονικό χαρακτήρα μιας συγκριμένης υποομάδας. Στην ανάλυση μας δεν χρειάστηκε τίποτα από τα δυο.

9. Priors: Στην καρτέλα αυτή ορίζουμε τις κατανομές για τις παραμέτρους των μοντέλων που βασίζονται σε προγενέστερη γνώση. Για την ανάλυση μας ορίσαμε ως κατανομή στο ucld.mean, που αντιπροσωπεύει τον εξελικτικό ρυθμό, την ομοιόμορφη κατανομή (Uniform) με αρχική τιμή 0,001 και όρια από 0 μέχρι 0,1. Η ίδια κατανομή χρησιμοποιήθηκε και για την παράμετρο treeModel.rootHeight, που αντιπροσώπευε τον χρόνο προέλευσης της ρίζας του δέντρου, με αρχική τιμή 6 και όρια από 1 μέχρι 40.

10. Operators: Εδώ καθορίζεται ο υπολογιστικός χρόνος που επενδύει το πρόγραμμα για την εκτίμηση των παραμέτρων των μοντέλων καθώς εκτελείται το MCMC.

11. MCMC: Εδώ ορίζουμε τις επιλογές της MCMC αλυσίδας ως προς το μήκος και το βήμα δειγματοληψίας. Το μήκος (Length of chain) είναι ο αριθμός των βημάτων που θα κάνει η MCMC στην αλυσίδα πριν τελειώσει. Το πόσο

χρονικό διάστημα θα τρέχει η αλυσίδα θα εξαρτάται από το μέγεθος του συνόλου των δεδομένων, την πολυπλοκότητα του μοντέλου και την ακρίβεια της απαιτούμενης απάντησης. Το βήμα δειγματοληψίας, το οποίο αποτελείται από δύο παραμέτρους echo state to screen every και log parameters every, καθορίζει το πόσο συχνά θα εμφανίζονται στην οθόνη οι τρέχουσες τιμές παραμέτρων και θα καταγράφονται στο αρχείο καταγραφής. Το βήμα είναι σημαντικό γιατί επηρεάζει τον χρόνο που θα τρέχει η αλυσίδα. Στην ανάλυση μας ορίσαμε ως μήκος 30.000.000 και ως βήμα δειγματοληψίας 3.000.

Αφού θέσαμε όλες τις παραμέτρους, στη συνέχεια εξάγουμε τα δεδομένα σε ένα αρχείο XML το οποίο τρέξαμε με το BEAST για να διεξαχθεί η ανάλυση. Για την αναπαράσταση των αποτελεσμάτων της ανάλυσης χρησιμοποιήθηκε το πρόγραμμα Tracer. Το Tracer συνοψίζει όλες τις εκτιμήσεις που παράγονται από τις αλυσίδες MCMC. Μπορεί να χρησιμοποιηθεί για να αναπαραστήσει τα αποτελέσματα των BEAST, MrBayes, LAMARC, και πιθανώς άλλων προγραμμάτων MCMC (44). Με το Tracer βλέπουμε όλες τις εκτιμήσεις της ανάλυσης και το χρησιμοποιούμε μεταξύ άλλων και για να δούμε αν έχει συγκλίνει η αλυσίδα MCMC. Οι παράμετροι που μας ενδιαφέρουν στην συγκριμένη ανάλυση είναι η prior, η posterior, η likelihood, η treemodel.rootHeight και η ucld.mean. Τέλος για την εύρεση του δέντρου μεγίστης αξιοπιστίας χρησιμοποιούμε το πρόγραμμα TreeAnnotator. Το TreeAnnotator είναι ένα πρόγραμμα που συνοψίζει τις πληροφορίες από ένα δείγμα δέντρων που παράγεται από το BEAST σε ένα μόνο δέντρο "στόχο". Οι συνοπτικές πληροφορίες περιλαμβάνουν τις posterior πιθανότητες των κόμβων στο δέντρο στόχο, τις posterior εκτιμήσεις και τα όρια HPD (Highest Posterior Density) των επιπέδων των κόμβων και (στην περίπτωση μοντέλου relaxed μοριακού ρολογιού) τα ποσοστά (45). Η μόνη παράμετρος που ορίστηκε διαφορετικά στην ανάλυση μας είναι η burn in η οποία πήρε τη τιμή 1000. Αυτή η επιλογή μας επιτρέπει να επιλέξουμε τον αριθμό των δειγμάτων στην αρχή της ανάλυσης όπου η αλυσίδα MCMC δεν έχει προλάβει να συγκλίνει, τα οποία θα απορριφθούν, ώστε να αναλύεται μόνο το τμήμα του ίχνους που βρίσκεται σε ισορροπία.

3.4.2 Μπεϋζιανή συμπερασματολογία στην Φυλογενετική

3.4.2.1 Θεώρημα Bayes

Το θεώρημα του Bayes περιγράφει την πιθανότητα ενός συμβάντος, βάσει προηγούμενης γνώσης των συνθηκών που μπορεί να σχετίζονται με το συμβάν. Για παράδειγμα, αν ο καρκίνος σχετίζεται με την ηλικία, χρησιμοποιώντας το θεώρημα του Bayes, η ηλικία ενός ατόμου μπορεί να χρησιμοποιηθεί για να εκτιμηθεί με μεγαλύτερη ακρίβεια η πιθανότητα καρκίνου, σε σύγκριση με την εκτίμηση της πιθανότητας εμφάνισης καρκίνου χωρίς τη γνώση της ηλικίας. Μία από τις πολλές εφαρμογές του θεωρήματος του Bayes είναι η Μπεϋζιανή συμπερασματολογία, μια ιδιαίτερη προσέγγιση στο στατιστικό συμπέρασμα. Όταν εφαρμόζονται, οι πιθανότητες που εμπλέκονται στο θεώρημα του Bayes μπορεί να έχουν διαφορετικές ερμηνείες πιθανότητας. Με την ερμηνεία της Μπεϋζιανής πιθανότητας το θεώρημα εκφράζει τον τρόπο με τον οποίο ένας υποκειμενικός βαθμός πεποίθησης πρέπει να αλλάξει λογικά ανάλογα με τη διαθεσιμότητα σχετικών στοιχείων.
Το θεώρημα του Bayes εκφράζετε μαθηματικά ως εξής :

$$P(A \vee B) = \frac{P(B|A)P(A)}{P(B)}$$

όπου Α και Β είναι ενδεχόμενα και:

- P(B) ≠ 0
- P(A|B) είναι μια δεσμευμένη πιθανότητα δηλαδή η πιθανότητα να συμβεί το ενδεχόμενο Α δοθέντος ότι το Β αληθεύει.
- P(B | A) είναι η πιθανότητα να συμβεί το ενδεχόμενο Β δοθέντος ότι το Α αληθεύει. Η ποσότητα αυτή είναι γνωστή και ως πιθανοφάνεια.
- P(A) και P(B) είναι οι πιθανότητες να παρατηρήσουμε τα ενδεχόμενα Α και Β ανεξαρτήτως του ενός από το άλλο. Αυτές οι πιθανότητες είναι γνώστες και ως περιθώριες.

Εάν αντικαταστήσουμε τα ενδεχόμενα με τυχαίες μεταβλητές τότε το θεώρημα του Bayes παίρνει την μορφή:

$$f(\theta|\chi = \frac{f(\theta)f(x \vee \theta)}{\int f(\theta)\, f(x \vee \theta)}, για\ συνεχείς\ τυχαίες\ μεταβλητές$$

και

$$f(\theta|\chi = \frac{f(\theta)f(x \vee \theta)}{\sum f(\theta)\, f(x \vee \theta)}, για\ διακριτές\ τυχαίες\ μεταβλητές$$

Όπου:

- $f(\theta|\chi$ είναι η εκ των υστερών κατανομή (posterior).
- $f(\theta)$ είναι η εκ των πρότερων κατανομή (prior).
- $f(x \vee \theta)$ είναι η συνάρτηση πιθανοφάνειας.
- $\sum f(\theta)\, f(x \vee \theta)$ ή $\int f(\theta)\, f(x \vee \theta)$ ονομάζεται σταθερά κανονικοποίησης και για συγκριμένες παρατηρήσεις χ είναι σταθερός αριθμός.

Αρά η εκ των υστέρων κατανομή είναι ανάλογη της πιθανοφάνειας και της εκ των προτέρων κατανομής:

$$f(\theta|\chi \propto f(\theta)f(x \vee \theta)$$

3.4.2.2 Μπεϋζιανή συμπερασματολογία

Η Μπεϋζιανή συμπερασματολογία είναι μια μέθοδος στατιστικής συμπερασματολογίας στην οποία χρησιμοποιείται το θεώρημα του Bayes για την επικαιροποίηση της πιθανότητας μιας υποθέσεως καθώς περισσότερες πληροφορίες ή στοιχεία γίνονται διαθέσιμα. Η Μπεϋζιανή συμπερασματολογία είναι μια σημαντική τεχνική στην στατιστική, και ιδιαίτερα στην μαθηματική στατιστική. Είναι ιδιαίτερα σημαντική στη δυναμική ανάλυση μιας ακολουθίας δεδομένων και βρίσκει εφαρμογές σε ένα ευρύ φάσμα δραστηριοτήτων, συμπεριλαμβανομένων της επιστήμης, της μηχανικής, της φιλοσοφίας, της ιατρικής, του αθλητισμού και του νόμου. Στην Μπεϋζιανή συμπερασματολογία ο τύπος του Bayes διαμορφώνετε ως εξής :

$$P(\theta|X, \alpha) = \frac{p(X|\theta)p(\theta|\alpha)}{P(X|a)} \propto P(X|\theta)p(\theta|\alpha)$$

οπού,

- χ είναι μια παρατήρηση.
- θ είναι η παράμετρος της κατανομής που ακόλουθη η παρατήρηση χ ~ p(x|θ).
- α είναι η υπερπαράμετρος της κατανομής των παράμερων θ ~ p(θ|α).
- Χ είναι το δείγμα που αποτελείται από n παρατηρήσεις $x_1, \dots x_n$,.

Και,

- p(θ|α) είναι η εκ των πρότερων κατανομή των παραμέτρων προτού παρατηρηθεί οποιαδήποτε παρατήρηση.
- $p(X|\theta)$ είναι η πιθανοφάνεια ή κατανομή δειγματοληψίας των παρατηρούμενων δεδομένων δοθέντος των παραμέτρων.
- $P(X|a)$ είναι η περιθώρια κατανομή.
- $P(\theta|X,\alpha)$ είναι η εκ των υστέρων κατανομή των παραμέτρων δοθέντος των παρατηρήσεων.

Η Μπεϋζιανή προσέγγιση στην φυλογενετική συνδυάζει την εκ των προτέρων πιθανότητα ενός δέντρου (φυλογένειας) P (Tree) με την πιθανοφάνεια P (Data(δεδομένα)/Tree) για να εκτιμήσει μια εκ των υστέρων πιθανότητα στα δέντρα P (Tree/Data). Η σταθερά κανονικοποίησης P (Data) είναι ισοδύναμη με τον υπολογισμό όλων των δυνατών τοπολογιών. Η εκ των υστέρων πιθανότητα είναι η πιθανότητα το δέντρο να είναι σωστό. Το δέντρο με την μεγαλύτερη ύστερη πιθανότητα είναι εκείνο που αντιπροσωπεύει καλύτερα την φυλογένεια. Ο υπολογισμός της ύστερης πιθανότητας είναι πολύ δύσκολος και χρονοβόρος λόγω της δυσκολίας υπολογισμού της σταθεράς κανονικοποίησης. Με την ανάπτυξη των MCMC (46) μεθόδων και άλλων ευρετικών αλγορίθμων διευκολύνεται ο υπολογισμός της σταθεράς. Λόγω αυτών των αλγορίθμων και της αυξανόμενης υπολογιστικής ισχύος, η Μπεϋζιανή συμπερασματολογία έγινε μια ευρέως διαδεδομένη και αξιόπιστη μέθοδος στην φυλογενετική.

3.4.3 Markov Cain Monte Carlo

Οι μέθοδοι αλυσίδων Markov Monte Carlo (MCMC) είναι μια κλάση αλγορίθμων για δειγματοληψία από μια κατανομή πιθανότητας που βασίζεται στην κατασκευή μιας αλυσίδας Markov που έχει την επιθυμητή κατανομή ως κατανομή ισορροπίας. Η κατάσταση της αλυσίδας μετά από μια σειρά βημάτων χρησιμοποιείται στη συνέχεια ως δείγμα της επιθυμητής κατανομής. Η ποιότητα του δείγματος βελτιώνεται ανάλογα με τον αριθμό των βημάτων της αλυσίδας.

3.4.3.1 Αλυσίδα Markov

Μια αλυσίδα Markov είναι ένα στοχαστικό μοντέλο που περιγράφει μια σειρά πιθανών γεγονότων στα οποία η πιθανότητα κάθε συμβάντος εξαρτάται μόνο από την κατάσταση που επιτεύχθηκε στο προηγούμενο συμβάν. Μια ακολουθία $X_1, X_2, \dots, X_n$ από τυχαία στοιχεία ενός συνόλου είναι μια Markov αλυσίδα εάν η δεσμευμένη κατανομή του X_{n+1} με δεδομένο το $X_1, X_2, \dots, X_n$ εξαρτάται μόνο από το X_n. Το σύνολο στο οποίο οι τιμές Xi παίρνουν τις τιμές είναι ο χώρος καταστάσεων της αλυσίδας Markov. Μια αλυσίδα Markov έχει σταθερά μεταβατικές πιθανότητες εάν η

εξαρτώμενη κατανομή του X_{n+1}με δεδομένο το X_n δεν εξαρτάται από το n. Αυτός είναι ο κύριος τύπος αλυσίδας Markov που ενδιαφέρει στην MCMC. Η από κοινού κατανομή μια αλυσίδας Markov ορίζεται από την περιθώρια κατανομή του X_1, η οποία ονομάζεται αρχική κατανομή, και την δεσμευμένη κατανομή του X_{n+1} δεδομένου X_n , που ονομάζεται κατανομή μεταβατικής πιθανότητας (λόγω της παραδοχής των σταθερών μεταβατικών πιθανοτήτων που κάναμε). Αν ο χώρος καταστάσεων είναι πεπερασμένος $\{x_1, x_2, \ldots, x_n\}$, τότε η αρχική κατανομή μπορεί να συσχετιστεί με ένα διάνυσμα λ = $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ που ορίζεται από το:

$$P(X_1 = x_i) = \lambda_i, i = 1, \ldots, n$$

και οι πιθανότητες μετάβασης μπορούν να συσχετιστούν με έναν πίνακα P που έχει στοιχεία που ορίζονται από το:

$$P\left(X_{n+1} = x_j \middle| X_n = x_i\right) = p_{ij}, i = 1, \ldots, n \kappa \alpha \iota j = 1, \ldots, n$$

Το p_{ij}=$P(i \to j)$ είναι η πιθανότητα μετάβασης της κατάσταση x_i στην κατάσταση x_j σε ένα βήμα. Η τυχαία μεταβλητή θεωρείται αλυσίδα Markov αν οι πιθανότητες μετάβασης των τιμών του χώρου καταστάσεων εξαρτώνται μονό από την παρούσα κατάσταση της μεταβλητής.

3.4.3.2 Monte Carlo

Οι μέθοδοι του Monte Carlo είναι μια ευρεία κατηγορία υπολογιστικών αλγορίθμων που βασίζονται σε επαναλαμβανόμενες τυχαίες δειγματοληψίες για την επίτευξη αριθμητικών αποτελεσμάτων. Η βασική τους ιδέα είναι να χρησιμοποιηθούν τυχαίες λύσεις για την επίλυση προβλημάτων που μπορεί να είναι προκαθορισμένα. Χρησιμοποιούνται συχνά σε φυσικά και μαθηματικά προβλήματα και είναι πολύ χρήσιμα όταν είναι δύσκολο ή αδύνατο να χρησιμοποιηθούν άλλες προσεγγίσεις. Όταν η κατανομή πιθανοτήτων της μεταβλητής παραγοντοποιείται, μπορούμε να χρησιμοποιήσουμε ένα δειγματολήπτη MCMC ώστε να σχεδιάσουμε ένα μοντέλο αλυσίδας Markov με μια καθορισμένη σταθερή κατανομή πιθανότητας. Δηλαδή, στο όριο, τα δείγματα που παράγονται από τη μέθοδο MCMC θα είναι δείγματα από την επιθυμητή κατανομή. Με το νόμο μεγάλων αριθμών, οι ακέραιοι που περιγράφονται από την αναμενόμενη τιμή κάποιας τυχαίας μεταβλητής μπορούν να προσεγγιστούν λαμβάνοντας τον εμπειρικό μέσο ανεξάρτητων δειγμάτων της μεταβλητής. Γενικά, οι μέθοδοι Monte Carlo χρησιμοποιούνται για την επίλυση διαφόρων προβλημάτων, δημιουργώντας κατάλληλους τυχαίους αριθμούς και παρατηρώντας το κλάσμα των αριθμών που υπακούει σε κάποια ιδιότητα ή ιδιότητες. Η μέθοδος είναι χρήσιμη για τη λήψη αριθμητικών λύσεων σε προβλήματα που είναι πολύ περίπλοκα για να επιλυθούν αναλυτικά.
Στην απλή περίπτωση ενός MCMC όπου τα $X_1, X_2, \ldots$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές (που σημαίνει ότι η αλυσίδα Markov είναι στάσιμη και αναστρέψιμη), έστω ότι θέλουμε να υπολογίσουμε μια αναμενόμενη τιμή:

$$\mu = E\{g(X)\}$$

όπου το g είναι μια συνάρτηση πραγματικών τιμών στον χώρο καταστάσεων, που δεν μπορούμε να την υπολογίσουμε με ακρίβεια. Αν υποθέσουμε ότι μπορούμε να

θεωρήσουμε ότι $X_1, X_2, \ldots$ είναι ανεξάρτητες και ισόνομες με την ίδια κατανομή X, τότε:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \cong \int_{\alpha}^{b} h(x)dx$$

Ο συγκεκριμένος δειγματικός μέσος είναι γνωστός ως προσέγγιση Monte Carlo.

Εάν έχουμε $Y_i = g(X_i)$, τότε τα Y_i είναι ανεξάρτητα και ισόνομα με μέσο μ και διασπορά

$\sigma^2 = var\{g(X)\}$,

Το $\hat{\mu}_n$ είναι ο δειγματικός μέσος των Y_i και από το Κεντρικό Οριακό Θεώρημα προσεγγίζει την κανονική κατανομή με μέσο μ και διακύμανση σ^2/n

$$\hat{\mu}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

και η δειγματική διακύμανση του είναι

$$\hat{\sigma}_n = \frac{1}{n}\sum_{i=1}^{n} (g(X_i) - \hat{\mu}_n)^2$$

Στην περίπτωση της δειγματοληψίας από πιο περίπλοκες κατανομές η προσέγγιση αυτή παρουσιάζει προβλήματα.

3.4.4 Αλγόριθμος Metropolis-Hastings

Ο αλγόριθμος Metropolis-Hastings είναι μια μέθοδος MCMC για τη λήψη μιας σειράς τυχαίων δειγμάτων από μια κατανομή για την οποία είναι δύσκολη η άμεση δειγματοληψία. Οι αλγόριθμοι Metropolis-Hastings και άλλοι αλγόριθμοι MCMC χρησιμοποιούνται γενικά για την δειγματοληψία από πολυδιάστατες κατανομές, ειδικά όταν ο αριθμός των διαστάσεων είναι μεγάλος.
Οι αλγόριθμοι Metropolis-Hastings μπορούν να λαμβάνουν δείγματα από κάθε κατανομή P(x), αρκεί να μπορεί να υπολογίσει την τιμή της συνάρτησης f(x) η οποία είναι ανάλογη με την πυκνότητα του P. Το γεγονός αυτό καθιστά τους αλγόριθμους Metropolis-Hastings ιδιαίτερα χρήσιμους, διότι ο υπολογισμός της σταθεράς κανονικοποίησης είναι συχνά εξαιρετικά δύσκολος στην πράξη.
Ο αλγόριθμος Metropolis-Hastings λειτουργεί έτσι ώστε με τη δημιουργία μιας αλληλουχίας τιμών του δείγματος όλο και περισσότερες τιμές του δείγματος που παράγονται, να έχουν ως αποτέλεσμα η κατανομή των τιμών να προσεγγίζει στενότερα την επιθυμητή κατανομή P(x). Αυτές οι τιμές του δείγματος παράγονται διαδοχικά, με την κατανομή του επόμενου δείγματος να εξαρτάται μόνο από την τρέχουσα τιμή του δείγματος (μετατρέποντας έτσι την αλληλουχία των δειγμάτων σε

μια αλυσίδα Markov). Συγκεκριμένα, σε κάθε επανάληψη ο αλγόριθμος επιλέγει έναν υποψήφιο για την επόμενη τιμή δείγματος βάσει της τρέχουσας τιμής δείγματος. Στη συνέχεια, με κάποια πιθανότητα, ο υποψήφιος είτε γίνεται αποδεκτός (στην οποία περίπτωση η υποψήφια τιμή χρησιμοποιείται στην επόμενη επανάληψη) είτε απορρίπτεται (όπου η υποψήφια τιμή απορρίπτεται, και η τρέχουσα αξία επαναχρησιμοποιείται στην επόμενη επανάληψη).
Η πιθανότητα της αποδοχής προσδιορίζεται συγκρίνοντας τις τιμές της συνάρτησης $f(x)$ των τιμών του τρέχοντος και του υποψήφιου δείγματος σε σχέση με την επιθυμητή κατανομή $P(x)$.
Έστω ότι $f(x)$ είναι μια συνάρτηση που είναι ανάλογη με την επιθυμητή κατανομή πιθανότητας $P(x)$:

1. Αρχικά επιλέγεται ένα αυθαίρετο σημείο x_0 ως το πρώτο δείγμα και μια αυθαίρετη πυκνότητα πιθανότητας $g(x \vee y)$ που υποδηλώνει μια επόμενη υποψήφια τιμή δείγματος x, λαμβάνοντας υπόψη την προηγούμενη τιμή y. Για τον αλγόριθμο Metropolis, το g πρέπει να είναι συμμετρικό, δηλαδή να ικανοποιεί την συνθήκη $g(x \vee y) = g(y \vee x)$. Συνήθως ως κατανομή του $g(x \vee y)$ επιλέγουμε την κανονική κατανομή. Η κατανομή αυτή είναι γνώστη ως κατανομή πρότασης.

2. Ύστερα για κάθε επανάληψη t:

- Δημιουργείται έναν υποψήφιο x για το επόμενο δείγμα επιλέγοντας από τη κατανομή $g(x' \vee x_t)$.
- Υπολογίζεται ο λόγος αποδοχής $\alpha = \frac{f(\chi')}{f(x_t)}$ για να αποφασίσει εάν θα αποδεχθεί ή θα απορρίψει τον υποψήφιο. Επειδή το f είναι ανάλογο με την πυκνότητα του P, έχουμε ότι
$$\alpha = \frac{f(\chi')}{f(x_t)} = \frac{P(\chi')}{P(x_t)}$$
- Αποδοχή ή απόρριψη: Δημιουργείται ένας ομοιόμορφος τυχαίος αριθμός (από την ομοιόμορφη κατανομή) u στο [0,1]. Αν ο $u \leq a$ αποδεχόμαστε τον υποψήφιο, ορίζοντας $x_{t+1} = x'$, ενώ αν ο $u > a$ απορρίπτουμε τον υποψήφιο και ορίζουμε ως $x_{t+1} = x_t$.

Αυτός ο αλγόριθμος προχωρά τυχαία προσπαθώντας να κινηθεί γύρω από το χώρο δειγματοληψίας, όπου μερικές φορές δέχεται τις κινήσεις και μερικές φορές παραμένει στη θέση του. Ο λόγος αποδοχής α υποδεικνύει πόσο πιθανό είναι το νέο προτεινόμενο δείγμα σε σχέση με το τρέχον δείγμα, σύμφωνα με την κατανομή $P(x)$. Αν προσπαθήσουμε να μετακινηθούμε σε ένα σημείο που είναι πιο πιθανό από το υπάρχον σημείο (δηλαδή ένα σημείο σε μια περιοχή υψηλότερης πυκνότητας $P(x)$), πάντα θα αποδεχόμαστε την κίνηση. Ωστόσο, αν προσπαθήσουμε να μετακινηθούμε σε ένα λιγότερο πιθανό σημείο, θα απορρίπτουμε μερικές φορές την κίνηση, και όσο μεγαλύτερη η σχετική πτώση της πιθανότητας, τόσο πιο πιθανό είναι να απορρίψουμε το νέο σημείο. Επομένως, θα έχουμε την τάση να παραμείνουμε σε περιοχές υψηλής πυκνότητας $P(x)$, ενώ μόνο περιστασιακά θα πηδάμε σε περιοχές χαμηλής πυκνότητας.
Οι Metropolis-Hastings αλγόριθμοι έχουν δυο κύρια μειονεκτήματα. Πρώτον, τα δείγματα μπορεί να συσχετίζονται. Αν και μακροπρόθεσμα ακολουθούν σωστά το P(x), ένα σύνολο δειγμάτων που βρίσκονται κοντά θα συσχετιστούν μεταξύ τους και

δεν θα αντικατοπτρίζουν σωστά την κατανομή. Αυτό σημαίνει ότι αν θέλουμε ένα σύνολο ανεξάρτητων δειγμάτων, πρέπει να πετάξουμε την πλειονότητα των δειγμάτων και να πάρουμε μόνο κάθε $\nu - οστο$ δείγμα, για κάποια τιμή ν (που συνήθως προσδιορίζεται εξετάζοντας την αυτοσυσχέτιση μεταξύ μεταθετών δειγμάτων). Η αυτοσυσχέτιση μπορεί να μειωθεί αυξάνοντας το πλάτος μετάθεσης, αλλά αυτό θα αυξήσει επίσης την πιθανότητα απόρριψης του προτεινόμενου άλματος. Πολύ μεγάλο ή πολύ μικρό μέγεθος μετάθεσης θα οδηγήσει σε μια πιο αργή αλυσίδα Markov. Δεύτερον, αν και η αλυσίδα Markov συγκλίνει τελικά στην επιθυμητή κατανομή, τα αρχικά δείγματα μπορεί να ακολουθούν μια πολύ διαφορετική κατανομή, ειδικά εάν το σημείο εκκίνησης βρίσκεται σε περιοχή χαμηλής πυκνότητας. Ως αποτέλεσμα, μια περίοδος προθέρμανσης (Burn-in) είναι συνήθως απαραίτητη, όπου ένας αρχικός αριθμός δειγμάτων (π.χ. οι 1.000 πρώτες) αφαιρούνται από το τελικό δείγμα.

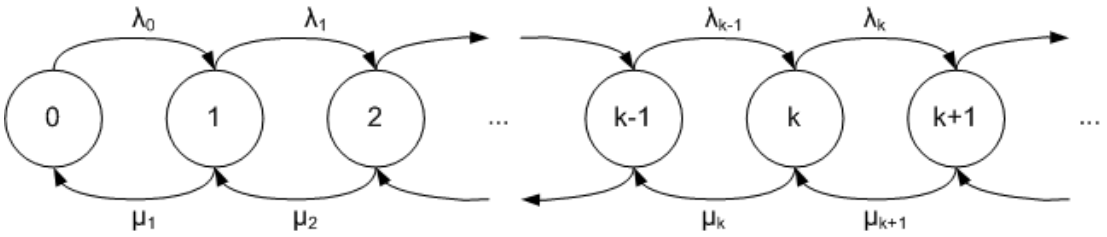
3.4.5 Μοριακό Ρολόι

Το μοριακό ρολόι παρουσιάζει ένα μέσο εκτίμησης των εξελικτικών ρυθμών και χρονοδιαγραμμάτων χρησιμοποιώντας γενετικά δεδομένα. Τα γενετικά δεδομένα που χρησιμοποιούνται για τέτοιους υπολογισμούς είναι συνήθως αλληλουχίες νουκλεοτιδίων για αλληλουχίες DNA ή αμινοξέων για πρωτεΐνες. Η κατανόηση των εξελικτικών ρυθμών και των χρονικών πλαισίων μπορεί να προσφέρει χρήσιμες γνώσεις για τις βιολογικές διεργασίες και τους μηχανισμούς, καθώς και να αποτελέσει σημαντική βάση για μια σειρά περαιτέρω αναλύσεων και ερμηνειών (47). Το γενετικό υλικό των ιών εμφανίζει τόσο μεγάλη εξέλιξη σε μικρό χρονικό διάστημα, όσο οι πολυπλοκότεροι οργανισμοί θα είχαν σε εκατομμύρια χρόνια (49). Για παράδειγμα, η εξέλιξη του HIV-1 έχει εκτιμηθεί σε ποσοστά μεταξύ 1×10^{-3} και 17×10^{-3} αντικαταστάσεις ανά έτος (50-51). Το μοριακό ρολόι χρησιμοποιείται συνήθως στη μοριακή εξέλιξη για να εκτιμήσει τους χρόνους συσσώρευσης ή ακτινοβολίας (48). Το μοριακό ρολόι παρουσιάζει ένα σταθερό ρυθμό γενετικής αλλαγής μεταξύ των αλληλουχιών, έτσι ώστε οι εκτιμήσεις των ποσοστών να μπορούν να εξηγηθούν στο Δέντρο της Ζωής για να συνταχθεί το χρονικό σημείο των εξελικτικών γεγονότων απόκλισης. Διάφορες μέθοδοι έχουν προταθεί για την εκτίμηση του μοριακού ρολογιού (ρυθμού εξέλιξης). Αρχικά, το μοριακό ρολόι υπολογίστηκε ως μια σταθερή συσσώρευση υποκαταστάσεων επί του χρόνου (52). Αλλά αυτή η απλοποίηση μπορεί να μην είναι πάντα κατάλληλη. Ορισμένες άλλες μέθοδοι επιτρέπουν επίσης τη λήψη δειγμάτων με διαφορετικές ημερομηνίες συλλογής. Επίσης έχουν αναπτυχθεί και μοντέλα βασιζόμενα στην Μπεϋζιανή συμπερασματολογία που ονομάζονται μοντέλα χαλαρού ρολογιού (relaxed clock models) τα οποία είναι πιο ρεαλιστικά και δίνουν πιο ακριβείς εκτιμήσεις, αλλά είναι πιο περίπλοκα και δύσκολα να υπολογιστούν.

3.4.6 Διαδικασία Birth-Death (BD)

Η διαδικασία Birth-Death (γεννήσεως-θανάτου) είναι μια ειδική περίπτωση της διαδικασίας Markov συνεχούς χρόνου όπου οι μεταβάσεις στην επόμενη κατάσταση είναι μόνο δύο τύπων: «γεννήσεις», οι οποίες αυξάνουν την κατάσταση κατά μία μονάδα και «θανάτους», που μειώνουν την κατάσταση κατά μία μονάδα. Οι διαδικασίες Birth-Death έχουν πολλές εφαρμογές στη δημογραφία, στη θεωρία των ουρών, στη μηχανική των επιδόσεων, στην επιδημιολογία και στη βιολογία. Όταν

συμβαίνει μια γέννηση, η διαδικασία πηγαίνει από την κατάσταση n στην n + 1. Όταν συμβαίνει θάνατος, η διαδικασία μεταβαίνει από την κατάσταση n στην κατάσταση n - 1. Η διαδικασία καθορίζεται από τα ποσοστά γεννήσεων $\{\lambda_\iota\}_{\iota=0\ldots\infty}$ και τα ποσοστά θανάτων $\{\mu_\iota\}_{\iota=1\ldots\infty}$ (53-54)



Εικόνα 13: Διάγραμμα των καταστάσεων της διαδικασίας Γενέσεων – Θανάτων

Μια διαδικασία καθαρών γεννήσεων (Poisson διαδικασία) είναι μια διαδικασία Birth-Death όπου $\mu_\iota = 0$ για κάθε ι ≥ 0.
Μια διαδικασία καθαρών θανάτων είναι μια διαδικασία Birth-Death όπου $\lambda_\iota = 0$ για κάθε ι ≥ 0.

3.4.6.1 Οι πιθανότητες ισορροπίας μιας διαδικασίας BD

Χρησιμοποιούμε τη μέθοδο μιας συνθήκης τομή = σφαιρική ισορροπία (οι εξισώσεις της σφαιρικής ισορροπίας είναι ένα σύνολο εξισώσεων που χαρακτηρίζουν την κατανομή ισορροπίας μιας αλυσίδας Markov, όταν υπάρχει τέτοια κατανομή) που εφαρμόζεται στο σύνολο των καταστάσεων 0, 1,. . ., κ.

Σε ισορροπία η ροή πιθανότητας σε όλες της τομές είναι ισορροπημένη (καθαρή ροή = 0)

$\lambda_\kappa \pi_\kappa = \mu_{\kappa+1} \pi_{\kappa+1}$ κ = 0,1,2, …

Παίρνουμε την επανάληψη,

$$\pi_{\kappa+1} = \lambda_\kappa \mu_{\kappa+1} \pi_\kappa$$

Μέσω των επαναλήψεων, όλες οι πιθανότητες των καταστάσεων μπορούν να εκφραστούν με όρους των καταστάσεων 0 (καθαρή ροή) και π_0

$$\pi_\kappa = \frac{\lambda_{\kappa-1}\lambda_{\kappa-2}\ldots\lambda_0}{\mu_\kappa \mu_{\kappa-1}\ldots\mu_1}\pi_0 = \prod_{i=0}^{\kappa-1}\frac{\lambda_i}{\mu_{i+1}}\pi_0$$

Η πιθανότητα π_0 καθορίζεται από την συνθήκη κανονικοποίησης π_0

$$\pi_0 = \frac{1}{1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0\lambda_1}{\mu_1\mu_2} + \cdots} = \frac{1}{1 + \sum_{\kappa=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}$$

3.4.6.2 Διαδικασίες BD εξαρτημένου χρόνου

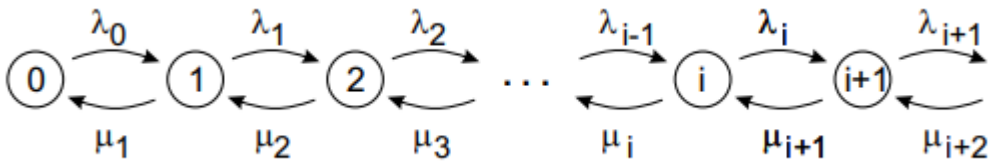
Κάποιες φορές οι πιθανότητες καταστάσεων π_0 είναι γνωστές στον χρόνο 0.
Για μια δεδομένη κατάσταση κ όπου $\pi_\kappa(0) = 1$ και $\pi_j(0) = 0$ όταν j ≠ k , αν θέλουμε να καθορίσουμε πως εξελίσσεται η πιθανότητα καταστάσεων ως συνάρτηση του χρόνου π(t) τότε το όριο που παίρνουμε για το π(t) είναι:

$$\lim_{t\to\infty} \pi(t) = \pi$$

Αυτό καθορίζεται από την εξίσωση:

$\frac{d}{dt}\pi(t)$=π(t)· Q, όπου

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & \cdots \\ 0 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 \\ \vdots & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 \\ \vdots & 0 & \mu_3 & -(\lambda_3 + \mu_3) \end{pmatrix}$$



Οι εξισώσεις των συνιστωσών είναι:

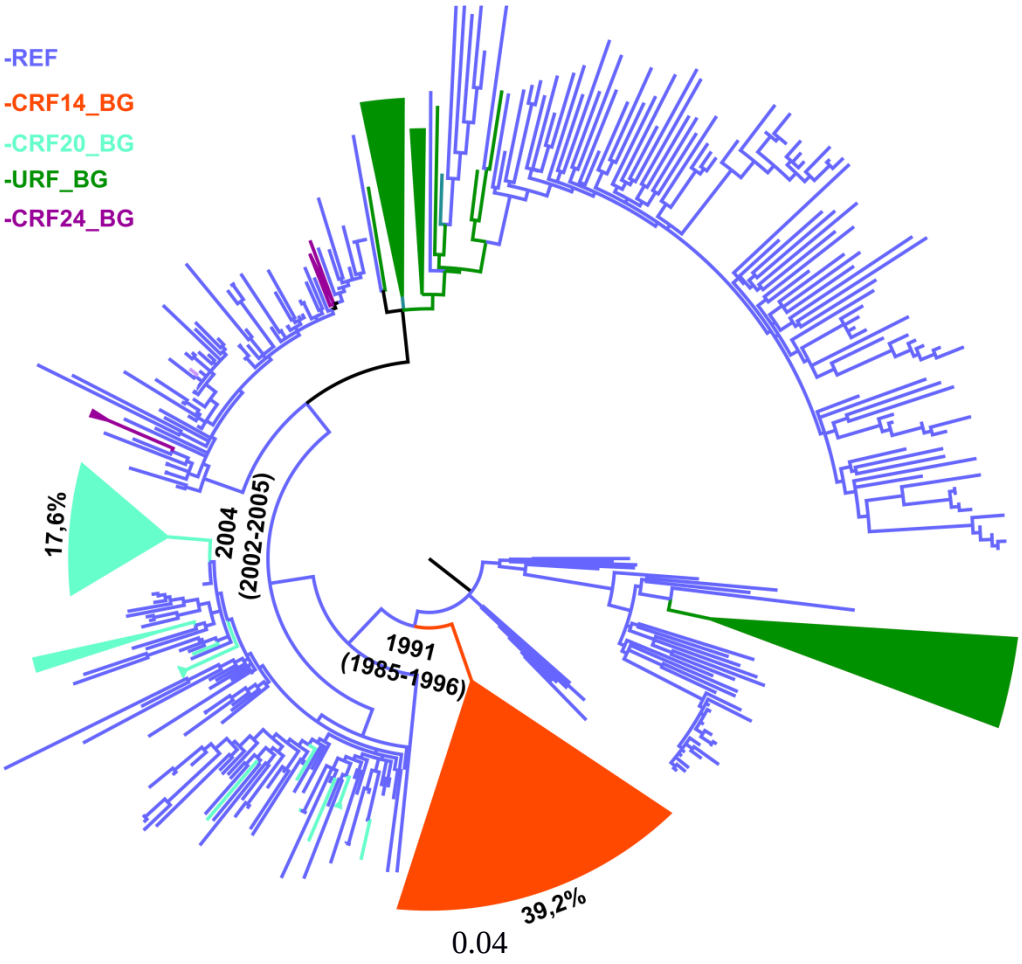
$$\begin{cases} \frac{d\pi_i(t)}{dt} = -(\lambda_i + \mu_i)\pi_i(t) + \lambda_{i-1}\pi_{i-1}(t) + \mu_{i+1}\pi_{i+1}(t) \\ \frac{d\pi_0(t)}{dt} = -\mu_0\pi_0(t) + \mu_1\pi_1(t) \end{cases}$$

1. Αποτελέσματα

4.1 Φυλογενετικά δέντρα

Από την φυλογενετική ανάλυση με το πρόγραμμα RAxML εκτιμήθηκαν δυο φυλογενετικά δέντρα (best tree και bipartition), για τον κάθε υπότυπο. Τα δέντρα αυτά απεικονίζουν την εξέλιξη της επιδημίας και τις μονοφυλετικές ομάδες (φυλογενετικές ομάδες με πάνω από 2 αλληλουχίες με το 70% εξ αυτών να έχει συλλεγεί στην Ισπανία) που προέκυψαν καθώς και το έτος που έγινε η μετάδοση. Για την διεξαγωγή των δέντρων χρησιμοποιηθήκαν τα GTR εξελικτικά μοντέλα για την εκτίμηση των αποστάσεων και η Γ-κατανομή για την εκτίμηση των διαφορετικών ρυθμών νουκλεοτιδικής αντικατάστασης.
Οι B/G και B/F βρέθηκαν σε υψηλότερα ποσοστά τη Ναβάρρα (B/G:7,4%; B/F:14,8%) και στη Χώρα των Βάσκων (B/G:4,9%; B/F:4,9).

4.1.1 B/G

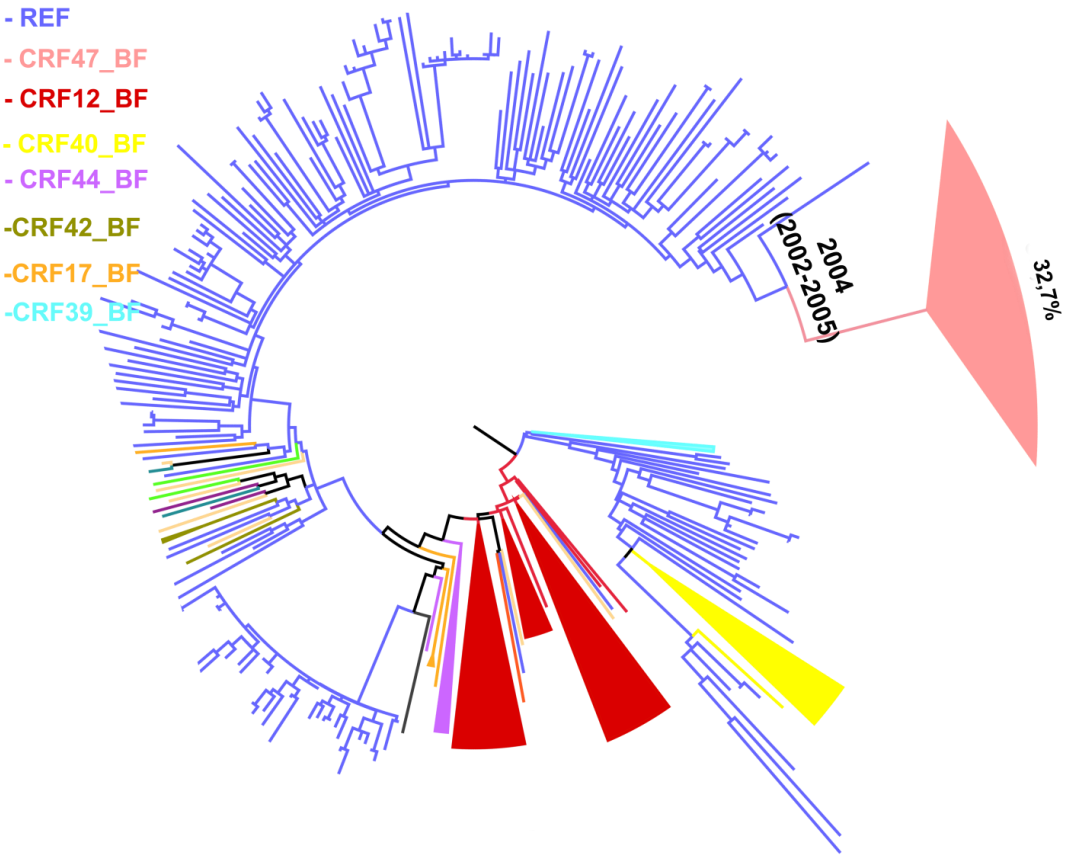


Εικονα.14: Φυλογενετικό δέντρο ανασυνδυασμένων στελεχών B/G στην Ισπανία βασιζόμενο στο εξελικτικό μοντέλο GTR για την εκτίμηση της απόστασης και στην Γ- κατανομής για τους

διαφορετικούς ρυθμούς νουκλεοτιδικής αντικατάστασης. Με μπλε χρώμα απεικονίζονται οι αλληλουχίες αναφοράς. Το κάθε ανασυνδυασμένο στέλεχος έλαβε διαφορετικό χρώμα στο δέντρο και οι μεγάλες τοπικές επιδημίες αναπαρίστανται με τρίγωνα Το έτος που αναγράφεται πάνω από ορισμένες τοπικές επιδημίες (τρίγωνα) αναφέρεται στον εκτιμώμενο χρόνο έναρξης τους.

Η φυλογενετική ανάλυση έδειξε ότι το 86,3% (N=88) των αλληλουχιών B/G από την Ισπανία βρέθηκε εντός 9 LTNs (τοπικά δίκτυα μετάδοσης). Το κάθε δίκτυο είναι μια μονοφυλετική ομάδα, όπου ως μονοφυλετική ορίστηκε κάθε φυλογενετική ομάδα που απαρτίζονταν από τουλάχιστον δύο αλληλουχίες ανασυνδιασμένων στελεχών (CRFs) με το 70% εξ αυτών να έχει συλλεγεί στην Ισπανία. Τα μεγαλύτερα δίκτυα αποτελούνταν από 40 (39,2%; CRF14_BG), 18 (17,6%; CRF20_BG) αλληλουχίες (**Εικονα.14**) και από 5 (4,9%; CRF24_BG). Το 94,4% (N=17) των αλληλουχιών του CRF20_BG LTN βρέθηκε ότι είχε απομονωθεί στη Μαδρίτη από άνδρες που έκαναν σεξ με άνδρες (ΑΣΑ), ενώ η γεωγραφική του προέλευση ήταν πιθανότατα από την Κούβα.

4.1.2 B/F



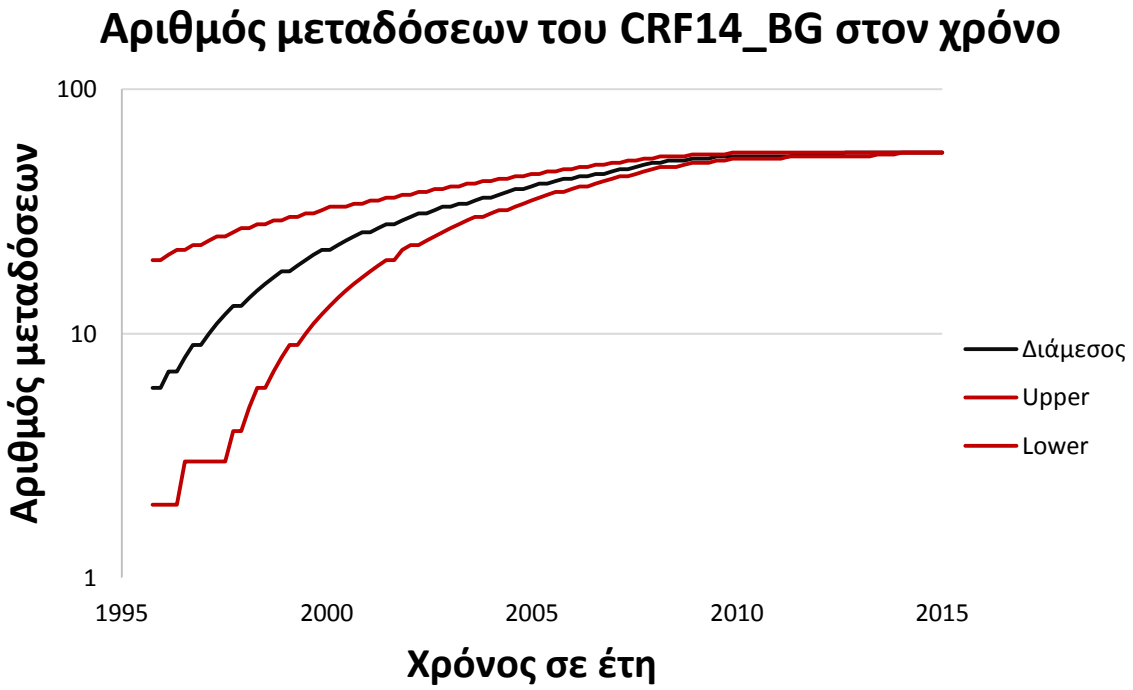
Εικονα.15: Φυλογενετικό δέντρο ανασυνδυασμένων στελεχών B/F στην Ισπανία βασισμένο στο εξελικτικό μοντέλο GTR για την εκτίμηση της απόστασης και στην Γ- κατανομή για τους διαφορετικούς ρυθμούς νουκλεοτιδικής αντικατάστασης. Με μπλε χρώμα απεικονίζονται οι αλληλουχίες αναφοράς. Το κάθε ανασυνδυασμένο στέλεχος έλαβε διαφορετικό χρώμα στο δέντρο και οι μεγάλες τοπικές επιδημίες αναπαρίστανται με τρίγωνα. Το έτος που αναγράφεται πάνω από ορισμένες τοπικές επιδημίες (τρίγωνα) αναφέρεται στον εκτιμώμενο χρόνο έναρξης τους.

Από την παραπάνω εικόνα βλέπουμε ότι το 74,5% (N=73) των αλληλουχιών B/F από Ισπανία βρέθηκε εντός 9 LTNs (τοπικά δίκτυα μετάδοσης). Το μεγαλύτερο εξ' αυτών ήταν το δίκτυο του CRF47_BF το οποίο αποτελούταν από 32 (32,7%) αλληλουχίες, με το 78,1% (N=25) των αλληλουχιών να έχει απομονωθεί από ετεροφυλόφιλους. Το στέλεχος CRF12_BF αποτελούσε το 9,42% (N=29) των ανασυνδιασμένων στελεχών BF, το CRF40_BF αποτελούσε το 2,27% (N=7) και τα CRF44,42,39,17 αποτελούσαν το 1,30 % (N=4)

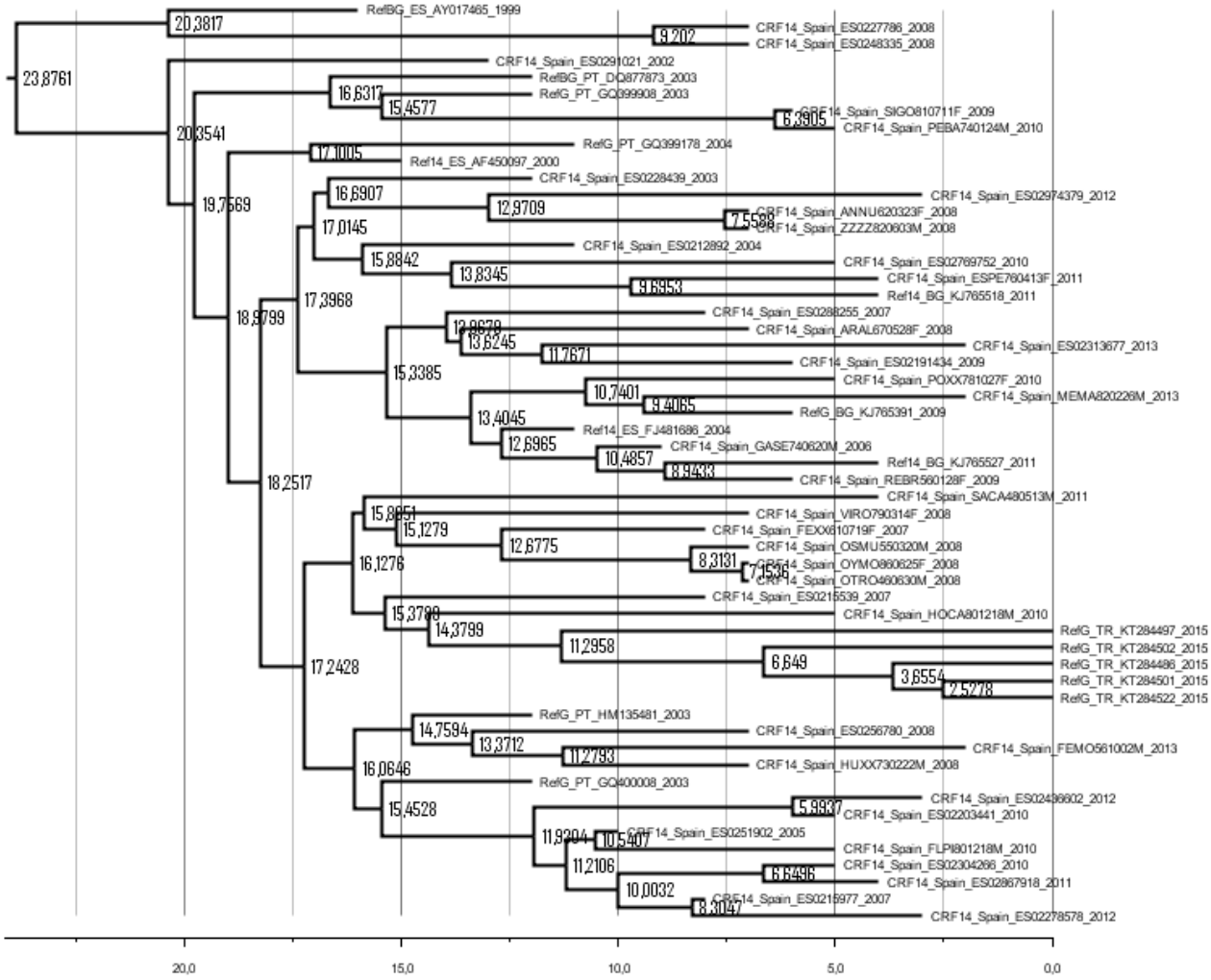
4.2 Φυλοδυναμική ανάλυση

Στην συνέχεια πραγματοποιήθηκε φυλοδυναμική ανάλυση για την εκτίμηση του αριθμού των μεταδόσεων για τις τοπικές επιδημίες των CRF47_BF, CRF20_BG και CRF14_BG με την χρήση του προγράμματος BEAST 1.8 (Εικονα.15-17). Για την παραμετροποίηση του προγράμματος χρησιμοποιήσαμε το εξελικτικό μοντέλο GTR σε συνδυασμό με την Γ-κατανομή και τα μοντέλα Birth-Death.

4.2.1 CRF14_BG

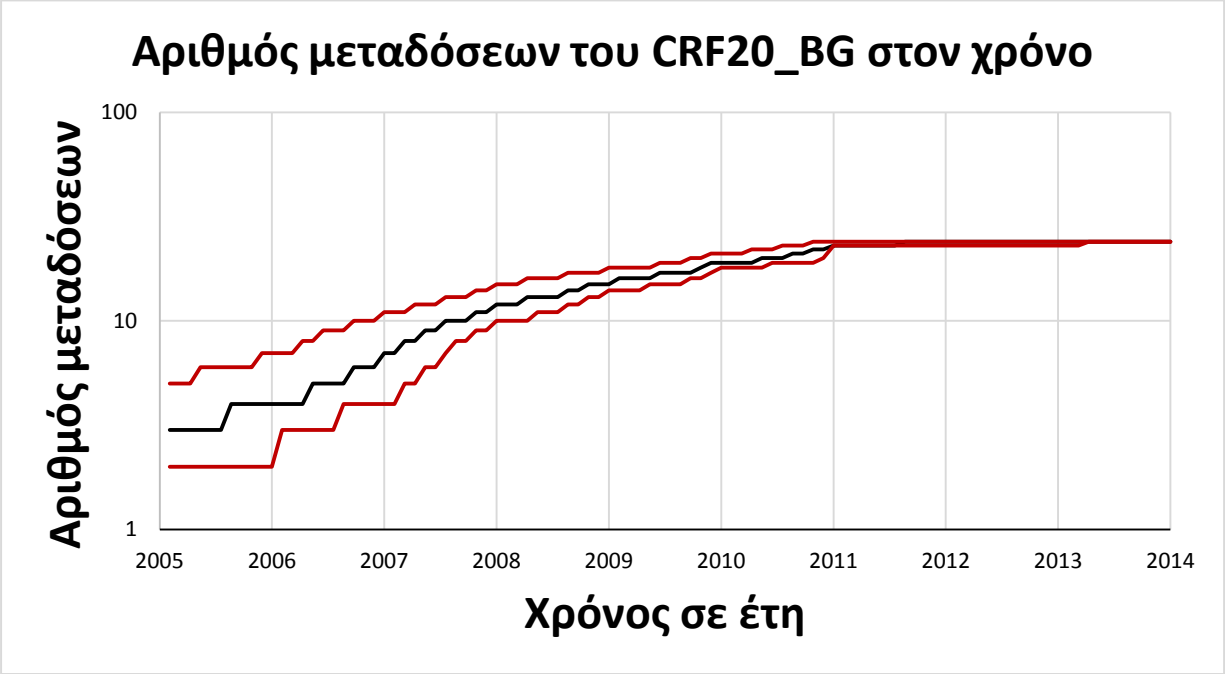


Εικονα.16: Αποτελέσματα της φυλοδυναμικής ανάλυσης (BEAST) των ανασυνδυασμένων στελεχών CRF14_BG στην Ισπανία. Στον κάθετο άξονα αναπαρίσταται ο εκτιμώμενος αριθμός των μεταδόσεων σε λογαριθμική κλίμακα και στον οριζόντιο τα έτη στα οποία έγιναν οι μεταδόσεις. Με μαύρη γραμμή αναπαρίσταται η διάμεσος του εκτιμώμενου αριθμού των μεταδόσεων ανά έτος και με διακεκομμένες κόκκινες γραμμές τα 95% διαστήματα αξιοπιστίας τους.

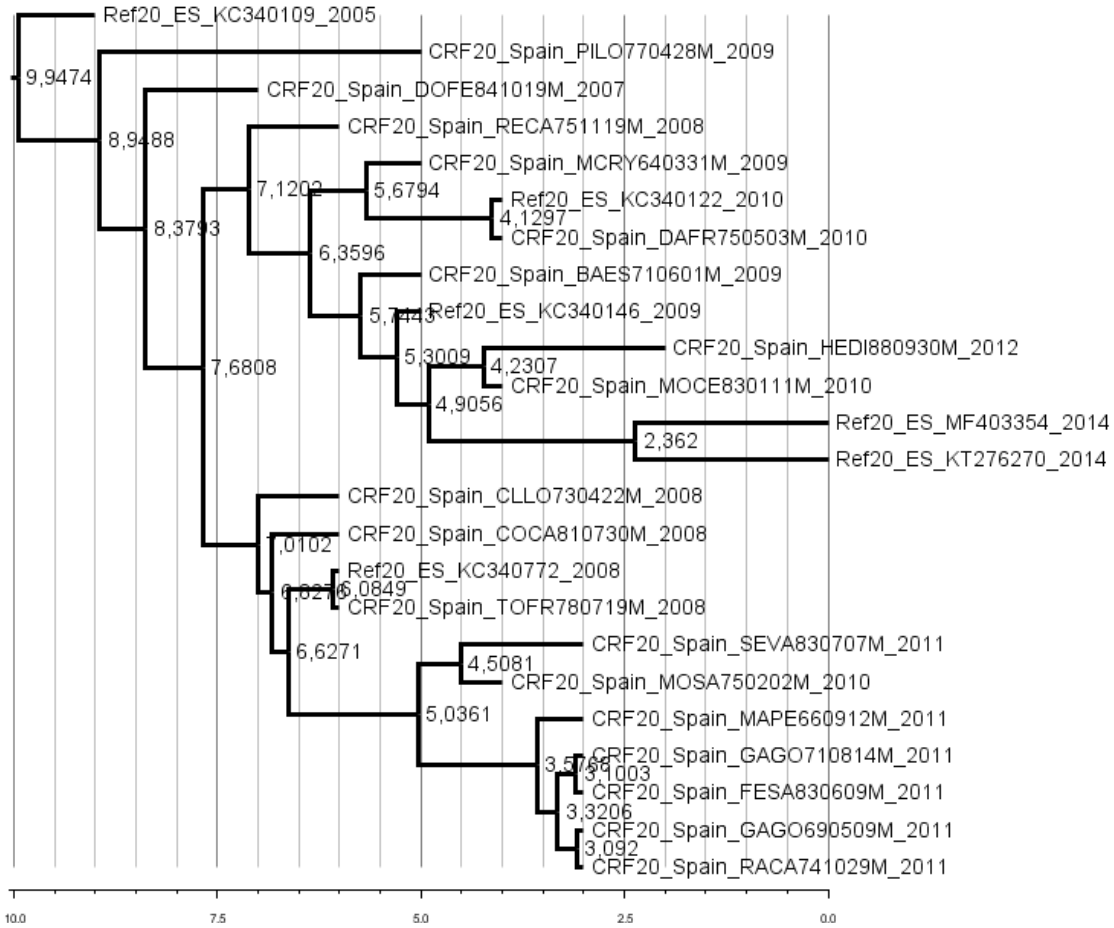


Εικόνα.17: Εκτίμηση της χρονολογημένης φυλογένειας για το ανασυνδιασμένο στέλεχος CRF14_BG. Ο οριζόντιος άξονας αναπαριστά την κλίμακα του χρόνου όπου το σημείο μηδέν είναι η πιο πρόσφατη ημερομηνία δειγματοληψίας στο δείγμα μας. Στην άκρη κάθε κλαδιού αναγράφεται το όνομα του ανασυνδιασμένου στελέχους.

4.2.2 CRF20_BG

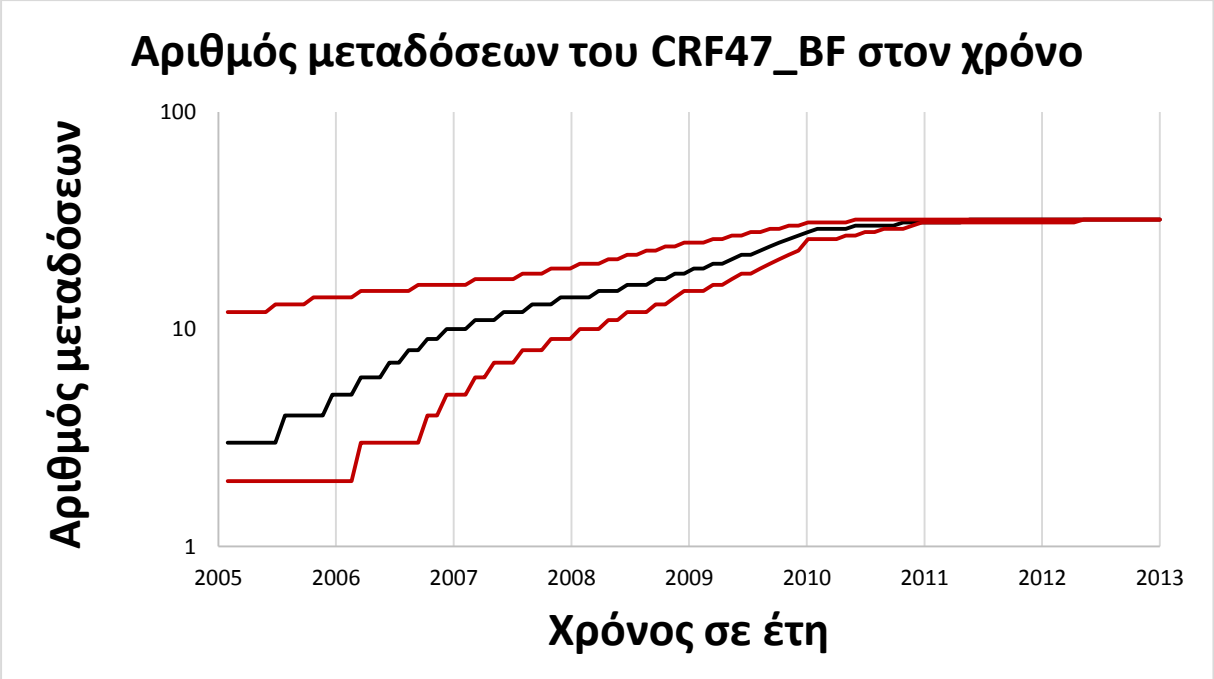


Εικονα.18: Αποτελέσματα της φυλοδυναμικής ανάλυσης (BEAST) των ανασυνδυασμένων στελεχών CRF20_BG στην Ισπανία. Στον κάθετο άξονα αναπαρίσταται ο εκτιμώμενος αριθμός των μεταδόσεων σε λογαριθμική κλίμακα και στον οριζόντιο τα έτη στα οποία έγιναν οι μεταδόσεις. Με μαύρη γραμμή αναπαρίσταται η διάμεσος του εκτιμώμενου αριθμού των μεταδόσεων ανά έτος και με διακεκομμένες κόκκινες γραμμές τα 95% διαστήματα αξιοπιστίας τους.

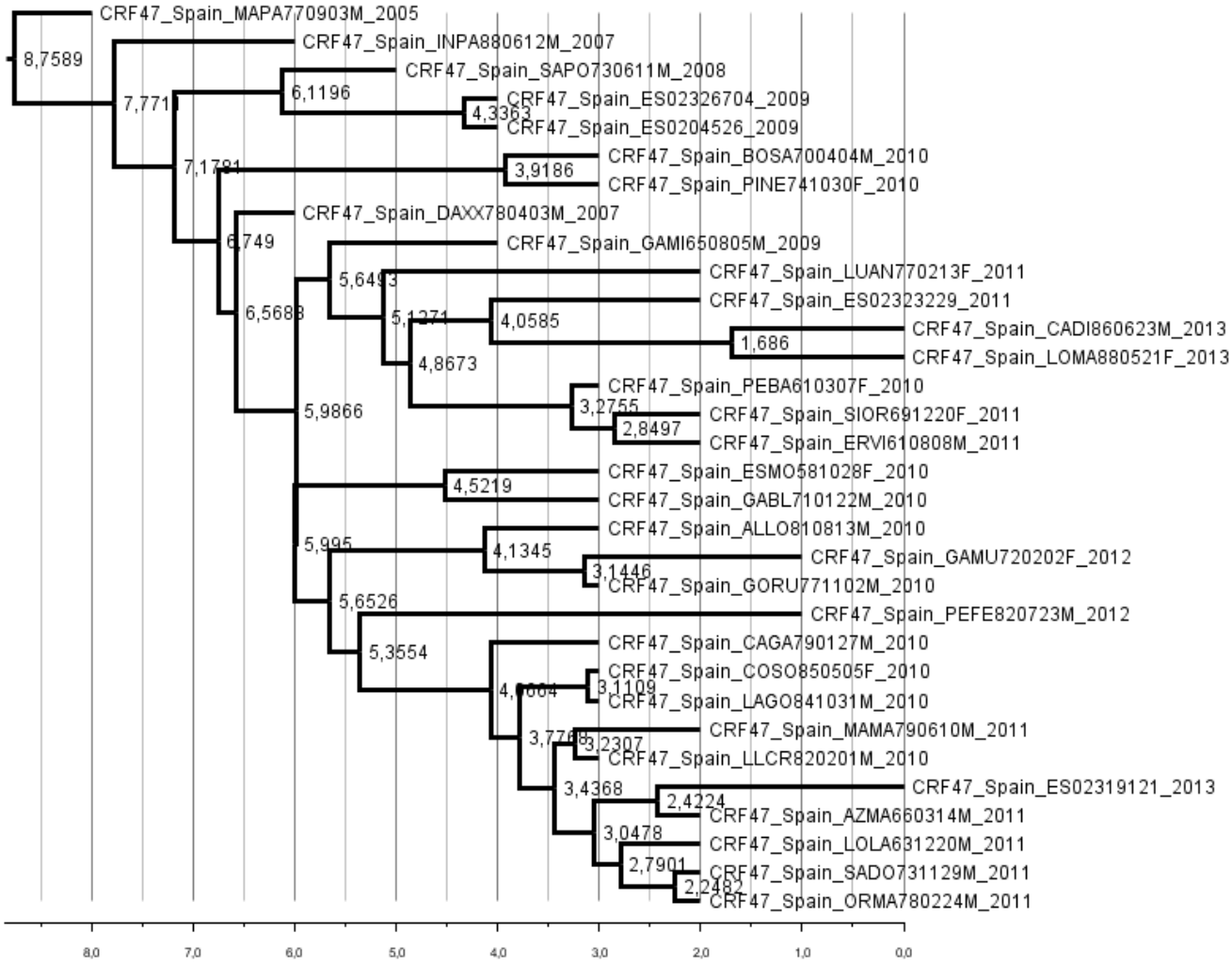


Εικόνα.19: Εκτίμηση της χρονολογημένης φυλογένειας για το ανασυνδιασμένο στέλεχος CRF20_BG. Ο οριζόντιος άξονας αναπαριστά την κλίμακα του χρόνου όπου το σημείο μηδέν είναι η πιο πρόσφατη ημερομηνία δειγματοληψίας στο δείγμα μας. Στην άκρη κάθε κλαδιού αναγράφεται το όνομα του κάθε ανασυνδιασμένου στελέχους.

4.2.3 CRF47_BF



Εικονα.20: Αποτελέσματα της φυλοδυναμικής ανάλυσης (BEAST) των ανασυνδυασμένων στελεχών CRF47_BF στην Ισπανία. Στον κάθετο άξονα αναπαρίσταται ο εκτιμώμενος αριθμός των μεταδόσεων σε λογαριθμική κλίμακα και στον οριζόντιο τα έτη στα οποία έγιναν οι μεταδόσεις. Με μαύρη γραμμή αναπαρίσταται η διάμεσος του εκτιμώμενου αριθμού των μεταδόσεων ανά έτος και με διακεκομμένες κόκκινες γραμμές τα 95% διαστήματα αξιοπιστίας τους.



Εικόνα.21: Εκτίμηση της χρονολογημένης φυλογένειας για το ανασυνδιασμένο στέλεχος CRF47_BF. Ο οριζόντιος άξονας αναπαριστά την κλίμακα του χρόνου όπου το σημείο μηδέν είναι η πιο πρόσφατη ημερομηνία δειγματοληψίας στο δείγμα μας. Στην άκρη κάθε κλαδιού αναγράφεται το όνομα του κάθε ανασυνδιασμένου στελέχους.

Από τα παραπάνω σχήματα παρατηρούμε ότι ο αριθμός των μεταδόσεων για το ανασυνδιασμένο στέλεχος CRF14_BG άρχισε να αυξάνεται το 1996 και η αύξηση αυτή διήρκεσε μέχρι το 2005, δηλαδή για περίπου 9 χρόνια. Έπειτα φαίνεται ότι ο αριθμός των μεταδόσεων σταθεροποιείται. Για τα στελέχη CRF20_BG και CRF47_BF παρατηρούμε ότι ο αριθμός των μεταδόσεων άρχισε να αυξάνεται από το 2005 μέχρι το 2011, για περίπου 6 χρόνια δηλαδή.

Ο εκτιμώμενος χρόνος προέλευσης για την τοπική επιδημία του CRF14_BG ήταν το 1991, με διάστημα αξιοπιστίας από το 1985 μέχρι 1996 και για την τοπική επιδημία του CRF20_BG το 2004 με διάστημα αξιοπιστίας από το 2002 μέχρι το 2005. Ο εκτιμώμενος χρόνος προέλευσης για τον CRF47_BF ήταν το 2004 με διάστημα αξιοπιστίας από το 2002 έως το 2005 (Εικονες.17-20).

2. Συμπεράσματα – Συζήτηση

Τα ευρήματα της παρούσας μελέτης υποδεικνύουν ότι υπάρχει σημαντική τοπική διασπορά των ανασυνδιασμένων στελεχών B/G και B/F στην Ισπανία. Συγκεκριμένα, εντοπίστηκαν αρκετά δίκτυα μετάδοσης του ιού που είναι ένδειξη ύπαρξης σημαντικής τοπικής διασποράς. Τα πιο σημαντικά από αυτά ήταν τα τοπικά δίκτυα του CFR14 και του CRF20 για τον B/G και του CRF47 για το B/F. Το επίκεντρο της τοπικής επιδημίας του B/G (CRF20_BG) στην Ισπανία βρέθηκε στη Μαδρίτη, είχε γεωγραφική προέλευση πιθανότατα από την Κούβα, ενώ αφορούσε κυρίως άνδρες που έκαναν σεξ με άντρες (ΑΣΑ). Οι πιο πρόσφατες τοπικές επιδημίες των CRF47_BF και CRF20_BG παρουσίασαν αύξηση στις μεταδόσεις για περίπου 6 έτη, ενώ η αύξηση της επιδημίας του CRF14_BG διήρκησε μεγαλύτερο χρονικό διάστημα.

Μέσω βιβλιογραφικής ανασκόπησης βρέθηκε ότι τα συγκεκριμένα ανασυνδεόμενα στελέχη εμφανίζουν παρόμοια συμπεριφορά και σε άλλες χώρες της Ευρώπης και της Λατινικής Αμερικής ως προς τις ομάδες του πληθυσμού που μολύνουν αλλά και την χρονική στιγμή που εμφανίστηκαν και δημιούργησαν επιμέρους επιδημίες. Συγκεκραμένα, ανασυνδεόμενα στελέχη B/F εντοπίζονται αρκετά συχνά σε χώρες της Λατινικής Αμερικής (Αργεντινή, Ουρουγουάη, Βραζιλία, Χιλή κα.) (56). Το πιο συχνό ανασυνδυασμένο στέλεχος φαίνεται να είναι ο CRF12_BF για το οποίο έχει εντοπιστεί επιμέρους επιδημία και στην Ισπανία στην παρούσα μελέτη. Το στέλεχος CRF47_BF επίσης έχει βρεθεί ότι έχει προκαλέσει τοπική επιδημία στην Ισπανία, η οποία σχετίζονταν κυρίως με ετεροφυλοφιλική επαφή (57). Αναφορικά με τα ανασυνδεόμενα στελέχη B/G έχει βρεθεί ότι έχουν επίκεντρο την Ευρώπη και μολύνουν κυρίως χρήστες ενδοφλεβίων ναρκωτικών ή άντρες που κάνουν σεξ με άντρες (ΑΣΑ). Στελέχη του CRF14_BG εντοπίστηκαν σε τοπικά δίκτυα στην Ισπανία, την Πορτογαλία, τη Γερμάνια, τη Ρουμανία και την Ελλάδα από τις αρχές του 1990 (58)(59)(61). Το επίκεντρο της επιδημίας του CRF14_BG φαίνεται να ήταν στην πόλη Γάλλιζα για την Ισπανία στα τέλη της δεκαετίας του ‘90, με ιδρυτικό στέλεχος με προέλευση από την Πορτογαλία, όπου η επιδημία του CRF14_BG εκτιμάται ότι ξεκίνησε στις αρχές του 1990 (60). Η επιδημία του συγκεκριμένου στελέχους στην Ελλάδα είχε επίκεντρο την Αθήνα, είχε προέλευση τη Ρουμανία και είναι σχετικά πρόσφατη (61). Προηγούμενες μελέτες υποδεικνύουν ότι μετά το 2007 ο επιπολασμός του CRF14_BG μειώθηκε. Η μείωση αυτή οφείλετε κυρίως στο γεγονός ότι το στέλεχος αυτό εμφανίζει πολύ υψηλή παθογένεια με αποτέλεσμα οι ασθενείς που έχουν μολυνθεί να εμφανίζουν γρήγορα AIDS και να οδηγούνται στον θάνατο(62)(63).

Περίληψη

Εισαγωγή: Προηγούμενη ανάλυση μας σε δείγμα 6.632 HIV-1 αλληλουχιών από Ισπανία έδειξε ότι τα ανασυνδυασμένα στελέχη B/G και B/F ήταν μεταξύ των μη-B HIV-1 υπότυπων με υψηλό επιπολασμό στη Ισπανία.

Σκοπός: Η διερεύνηση του τρόπου διασποράς των B/G και B/F στην Ισπανία και η εκτίμηση του χρόνου προέλευσης και της δυναμικής των μεγαλύτερων τοπικών επιδημιών, χρησιμοποιώντας μεθόδους μοριακής επιδημιολογίας.

Υλικό: Μελετήθηκαν 102 αλληλουχίες B/G και 98 B/F με δειγματοληψία το διάστημα 2002-2014 σε 10 περιοχές της Ισπανίας.

Μέθοδος: Η φυλογενετική ανάλυση των υπο μελέτη αλληλουχιών πραγματοποιήθηκε μαζί με τις πιο όμοιες προς αυτές γενετικά αλληλουχίες ως αλληλουχίες αναφοράς. Ως τοπικά δίκτυα μετάδοσης (local transmission networks, LTNs) ορίστηκαν οι φυλογενετικές ομάδες με ποσοστό αλληλουχιών από Ισπανία >70%. Η φυλοδυναμική ανάλυση πραγματοποιήθηκε χρησιμοποιώντας μεθόδους Μπεϋζιανής συμπερασματολογίας.

Αποτελέσματα: Οι B/G και B/F βρέθηκαν σε υψηλότερα ποσοστά στη Ναβάρρα (B/G:7,4%; B/F:14,8%) και στη Χώρα των Βάσκων (B/G:4,9%; B/F:4,9%). Η φυλογενετική ανάλυση έδειξε ότι το 86,3% (N=88) των αλληλουχιών B/G από Ισπανία βρέθηκε εντός 9 LTNs. Τα δυο μεγαλύτερα από αυτά αποτελούνταν από 40 (39,2%; CRF14_BG) και 18 (17,6%; CRF20_BG) αλληλουχίες. Το 94,4% (N=17) των αλληλουχιών του CRF20_BG LTN βρέθηκε ότι είχε απομονωθεί στη Μαδρίτη από άνδρες που έκαναν σεξ με άνδρες (MSM), ενώ η γεωγραφική του προέλευση ήταν πιθανότατα από την Κούβα. Το 74,5% (N=73) των αλληλουχιών B/F από Ισπανία βρέθηκε εντός 9 LTNs. Το μεγαλύτερο εξ' αυτών (CRF47_BF) αποτελούνταν από 32 (32,7%) αλληλουχίες, με το 78,1% (N=25) των αλληλουχιών να έχει απομονωθεί από ετεροφυλόφιλους. Ο εκτιμώμενος χρόνος προέλευσης (tMRCA) ήταν το 1991 (διάμεση τιμή) για τον CRF14_BG, το 2004 για τον CRF20_BG και το 2004 για τον CRF47_BF. Ο αριθμός των μεταδόσεων για τα LTNs των CRF47_BF και CRF20_BG εκτιμήθηκε ότι αυξήθηκε μεταξύ 2005 και 2011. Για τον CRF14_BG η μεγαλύτερη αύξηση στις μεταδόσεις συνέβη το διάστημα 1996-2005.

Συμπεράσματα: Η μελέτη ανέδειξε την ύπαρξη σημαντικής τοπικής διασποράς για τα ανασυνδυασμένα στελέχη B/G και B/F στην Ισπανία. Το επίκεντρο για τον B/G (CRF20_BG) βρέθηκε στη Μαδρίτη, και προήλθε πιθανότατα από την Κούβα. Οι πιο πρόσφατες τοπικές επιδημίες των CRF47_BF και CRF20_BG παρουσίασαν αύξηση στις μεταδόσεις για περίπου 6 έτη, ενώ η αύξηση της επιδημίας του CRF14_BG διήρκησε μεγαλύτερο χρονικό διάστημα.

Abstract

Background: Our previous analysis on 6,632 HIV-1 sequences sampled in Spain revealed that B/G and B/F recombinant forms were among the HIV-1 non-B clades with the higher prevalence in Spain (1.54% and 1.48%, respectively). Our aim was to investigate the patterns of B/G and B/F dispersal across Spain and estimate the spatiotemporal characteristics of their largest regional epidemics, using molecular methods.

Materials and Methods: We studied 102 B/G and 98 B/F sequences, available in the PR/RT regions. Sequences were isolated from HIV-1 diagnosed patients during 2002-2014 from 10 autonomies of Spain. Patients' samples were merged from two datasets: a) CoRIS (2004-2013), and b) Eastern Andalusia Resistance Cohort (2000-2014). We analyzed phylogenetically sequences from our study population along with the most closely related sequences to them (HIV BLAST tool; B/G:N=317; B/F:N=210), using maximum likelihood method with bootstrap evaluation as implemented in RAxML v8.0.20 (GTR+G model). Local transmission networks (LTNs) were phylogenetic clusters including sequences from Spain at proportions >70%. Phylodynamic analysis was performed by using a Bayesian method as implemented in BEAST v1.8.0 (birth-death model).

Results: Navarre (B/G:7.4%; B/F:14.8%) and Basque Country (B/G:4.9%; B/F:4.9%) were the autonomies where B/G and B/F were more frequently found. Phylogenetic analysis revealed that 86.3% (N=88) of the B/G sequences from Spain found within 9 LTNs (CRF14_BG:N=40, 1 LTN; CRF20_BG:N=27, 4 LTNs; URF B/G:N=19, 3 LTNs; CRF24_BG:N=2, 1 LTN). The two largest B/G LTNs included 40 (39.2%; CRF14_BG) and 18 (17.6%; CRF20_BG) sequences. The 94.4% (N=17) of the sequences found within the CRF20_BG LTN were from individuals living in Madrid reported men having sex with men (MSM) as transmission risk group. Analysis revealed Cuba as the most possible source of CRF20_BG subepidemic. Analysis also revealed that 74.5% (N=73) of the B/F sequences from Spain formed 9 LTNs (CRF47_BF:N=32, 1 LTN; CRF12_BF:N=26, 3 LTNs; CRF40_BF:N=6, 1 LTN; CRF44_BF:N=3, 1 LTN; CRF17_BF:N=2; 1 LTN; CRF39_BF:N=2, 1 LTN; CRF42_BF:N=2, 1 LTN). The largest B/F LTN (CRF47_BF) consisted of 32 (32.7%) sequences, the majority of which had been isolated from heterosexuals (N=25, 78.1%) living in Andalusia (N=10, 31.3%), Navarre (N=8, 25%) and Basque Country (N=7, 21.9%). Molecular clock analysis estimated that the time of the most recent common ancestor (tMRCA) of the subepidemics was in 1991 (median estimate; 95%HPD:1985-1996) (CRF14_BG), in 2004 (95%HPD:2002-2005) (CRF20_BG) and in 2004 (95%HPD:2002-2005) (CRF47_BF). The birth-death skylines suggested a large increase in number of infections for the CRF47_BF and CRF20_BG, lasting between 2005 and 2011. For the CRF14_BG the largest increase in number of new infections occurred during 1996-2005.

Conclusions: Our study revealed that the B/G and B/F transmissions are due to regional dispersal at a considerable proportion in Spain. The hot spot for one of the largest B/G regional subepidemics (CRF20_BG) in Spain was in Madrid, associated with MSM risk group and probably originated from Cuba. The tMRCA and transmission dynamics of the three largest outbreaks were diverse. The most recent subepidemics (CRF47_BF, CRF20_BG) showed a rapid increase that lasted for approximately six years, whilst the CRF14_BG epidemic growth occurred over a longer time period.

Βιβλιογραφία

1.http://www.who.int/

2.http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html

3.Keele B.F, Van H.F, Li Y, Bailes E, Takehisa J, et al. (2006*) Chimpanzee reservoirs of pandemic and nonpandemic HIV-1*.Science 313: 523–526

4.Hemelaar J. (2012). *The origin and diversity of the HIV-1 pandemic. Trends Mol Med* 18: 182

5.Capon D.J, Ward R.H. (1991). *The CD4-gp120 interaction and AIDS pathogenesis. Annu Rev Immunol*; :649-678

6.Feng Y, Broder C.C, Kennedy P.E. et al. (*1996). HIV-1 entry cofactor: Functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. Science;*272:872-877.

7.Deng H, Liu R, Ellmeier W, et al. (1996). *Identification of a major co-receptor for primary isolates of HIV-1*. Nature;381:661-666.

8.Gottlinger H.G, Sodroski J.G, Haseltine W.A. (1989). *Role of capsid precursor processing and myristoylation in morphogenesis and infectivity of human immunodeficiency virus type 1.* Proc Natl Acad Sci USA;86:5781-5785.

9.Korber, B, Foley B. T, Kuiken C, Pillai S. K, Sodroski J. G. (1998*), Numbering positions in HIV relative to HXB2CG,* p. III-102–III-111. In B.Korber, C. L. Kuiken, B. Foley, B. Hahn, F. McCutchan, J. W. Mellors, and J. Sodroski (ed.), Human retroviruses and AIDS 1998. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, N.Mex

10.Plantier J. C, et al. (2009). *A new human immunodeficiency virus derived from gorillas.* Nat. Med.15:871–872

11.Taylor B. S, Sobieszczyk M. E, McCutchan F. E, Hammer S. M. (2008). *The challenge of HIV-1 subtype diversity.* N. Engl. J. Med. 358:1590–1602

12.Van der Kuyl A. C, Cornelissen M. (2007). *Identifying HIV-1 dual infections. Retrovirology* 4:67

13.Creighton H. B. and McClintock B. (1931). *«A Correlation of Cytological and Genetical Crossing-Over in Zea Mays»* . Proccedings of the National Academy of Science of the USA 17.8 (2016): 492–497

14.Robertson D.L, Anderson J.P, Bradac J.A, Carr J.K, Foley B, Funkhouser R.K, Gao F, Hahn B.H, Kalish M.L, Kuiken C, et al. (2000). *HIV-1 nomenclature proposal*. Science.;288:55–55

15.http://science.sciencemag.org/content/288/5463/55.4.long

16.Anthony S. Fauci M. D. (1999). N Engl J Med 1999; 341:1046-1050

17.Gallo R. S, Sarin P. S, Gelmann E. P, Robert-Guroff M, Richardson E, Kalyanaraman V. S, Mann D, Sidhu G. D, Stahl R. E, Zolla-Pazner S, Leibowitch J, Popovic M. (1983). «*Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS)».* Science 220 (4599): 865–867.

18.Keele B. F, van Heuverswyn, F, Li Y. Y, Bailes, et al. (2006*). «Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1»*. Science 313 (5786):523

19.Reeves J. D. and Doms R. W. (2002). *«Human Immunodeficiency Virus Type 2». J. Gen. Virol. 83 (Pt 6):* 1253–65

20.Sharp P. M, Bailes E, Chaudhuri R. R, Rodenburg C. M, Santiago M. O, Hahn B. H. (2001). *«The origins of acquired immune deficiency syndrome viruses: where and when?».* Philosophical Transactions of the Royal Society B: Biological Sciences 356 (1410):867-76

21.Wang H, Wolock T. M, Carter A, Nguyen, G, Kyu H. H, Gakidou E, Hay S. I, Mills E. J, Trickey A. (2016). *"Estimates of global, regional, and national incidence, prevalence, and mortality of HIV, 1980–2015: the Global Burden of Disease Study 2015"*. The Lancet HIV Aug;3(8):e361-e387

22.unaids data_2017 σελ.14

23.unaids.org Unaids data_2017 σελ 25, 40

24.http://www.unaids.org/en/regionscountries/countries/spain

25.Caro-Murillo A.M, Castilla J, Pérez-Hoyos S, Miró J.M, Podzamczer D, Rubio R, Riera M, Viciana P, López Aldeguer J, Iribarren J.A, de los Santos-Gil I, Gómez-Sirvent J.L, Berenguer J, Gutiérrez F, Saumoy M, Segura F, Soriano V, Peña A, Pulido F, Oteo J.A, Leal M, Casabona J, del Amo J, Moreno S; Grupo de trabajo de la Cohorte de la Red de Investigación en Sida (CoRIS). *Spanish cohort of naïve HIV-infected patients (CoRIS): rationale, organization and initial results*]. Enferm Infecc Microbiol Clin. 2007 Jan;25(1):23-31

26.Koser C.U, Ellington M.J, Cartwright E.J.P, Gillespie S.H, Brown N.M, et al. (2012). Routine *Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology.* PLoS Pathog journal.ppat.1002824

27.Porta M, Greenland S, Hernán M, dos Santos Silva I, Last J.M. (2014). *A dictionary of epidemiology, 6th. edition. New York: Oxford University Press.*_ISBN 9780199976737

28.Porta M. (2002)."*Incomplete overlapping of biological, clinical, and environmental information in molecular epidemiological studies: a variety of causes and a cascade of consequences"*. J Epidemiol Community Health. 56 (10): 734–738

29.Jukes T.H. and Cantor C.R. (1969). Evolution of Protein Molecules. New York: Academic Press. pp. 21–132

30. Tavaré S. (1986). "*Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences"*. American Mathematical Society. 17: 57–86

31.https://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Phylogenetics/phylo12.html

32.Sokal R and Michener C. (1958). *"A statistical method for evaluating systematic relationships". University of Kansas Science Bulletin.* 38: 1409–1438

33.Saitou N and Nei M. (1987). *"The neighbor-joining method: a new method for reconstructing phylogenetic trees."* Molecular Biology and Evolution, volume 4, issue 4, pp. 406-425

34.https://www.stat.berkeley.edu/~terry/Classes/s246.2006/Week6/2neighbourJoining.pdf

35.Farris, J. S. (1970). *Methods for computing Wagner trees.* Systematic Zoology 19, *83-92*

36.Felsenstein J. (1985). *Phylogenies and the Comparative Method,* The American Naturalist Vol. 125, No. 1 pp. 1-15

37.Fitch, W. M. (1971). *Toward defining the course of evolution: minimum change for a specified tree topology.* Systematic Zoology 20 (4), 406-416

38.Edgar, R. C. (2004), MUSCLE: *multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32(5), 1792-97*

39. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013). *Molecular Biology and Evolution*, Volume 30, Issue 12, Pages 2725–2729, https://doi.org/10.1093/molbev/mst197

40.Stamatakis A. RAxML version 8 (*2014): a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, Volume 30, Issue 9, Pages *1312–1313*, https://doi.org/10.1093/bioinformatics/btu033

41. Grenfell, B. T.; Pybus, O. G.; Gog, J. R.; Wood, J. L.; Daly, J. M.; Mumford, J. A.; Holmes, E. C. (2004). "*Unifying the Epidemiological and Evolutionary Dynamics of Pathogens"*. *Science*. 303 (5656): 327 PMID 14726583

42. http://tree.bio.ed.ac.uk/software/beast/

43. http://beast.community/rates_and_dates

44. http://tree.bio.ed.ac.uk/software/tracer/

45. http://beast.community/treeannotator

46. Metropolis N, Rosenbluth A. W, Rosenbluth M. N, Teller A. H, Teller E. (1953). *Equation of State Calculations by Fast Computing Machines.* J. Chem. Phys. 21, 1087

47.Simon Y. W. Ho, Sebastián Duchêne Mol Ecol. (2014) *Molecular-clock methods for estimating evolutionary rates and timescales* (24):5947

48.(Kasha M & Pullman B) E Zuckerkandl, LB Pauling (1962). *Molecular disease, evolution, and genic heterogeneity. From 'Horizons in Biochemistry*'- Academic Press

49.Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T Science (2000). *Timing the ancestor of the HIV-1 pandemic strains*. ; 288(5472):1789-96

50.Leitner T, Albert J Proc Natl Acad Sci U S A (1999). *The molecular clock of HIV-1 unveiled through analysis of a known transmission history.* ; 96(19):10752-7

51.Salemi M, Strimmer K, Hall WW, Duffy M, Delaporte E, Mboup S, Peeters M, Vandamme AM FASEB J. (2001). *Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution*. (2):276-8

52.Kimura M J Mol Evol (1980). *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences*. Dec; 16(2):111-20

53.Virtamo J,"*Birth-death processesBirth-death processes*", 38.3143 Queueing Theory.

54.https://eclass.upatras.gr/modules/units/?course=CEID1094&id=4251

55.http://www.who.int/

56. Monick L Guimarães,corresponding author, Ketty G Velarde-Dunois, David Segurondo, and Mariza G Morgado (2012) *The HIV-1 epidemic in Bolivia is dominated by subtype B and CRF12_BF "family" strains. Virol J. 2012 Jan 16;9:19*

57. Ferna´ndez-Garcı ´a, Lucı ´a Pe´rez-A´ lvarez, Marı ´a Teresa Cuevas, Elena Delgado,Mercedes et al.(2016) *Identification of a New HIV Type 1 Circulating BF Intersubtype Recombinant Form (CRF47_BF) in Spain*. AIDS Res Hum Retroviruses. 2010 Jul;26(7):827-32

58. Delgado E, Thomson MM, Villahermosa ML, Sierra M, Ocampo A, et al.(2002). *Identification of a newly characterized HIV-1 BG intersubtype circulating recombinant form in Galicia, Spain, which exhibits a pseudotype-like virion structure.* J Acquir Immune Defic Syndr. ;29:536–543

59.Harris B, von Truchsess I, Schatzl HM, Devare SG, Hackett J.(2005) *Genomic characterization of a novel HIV type 1 B/G intersubtype recombinant strain from an injecting drug user in Germany*. AIDS Res Hum Retroviruses.;21:654–660

60.Bártolo I, Abecasis A. B, Borrego P, Barroso H, McCutchan F, Gomes P, Camacho R, Taveira N (2011). *Origin and Epidemiological History of HIV-1 CRF14_BG*. PLoS One. 2011;6(9):e24130

61.Paraskevis D, Paraschiv S, Sypsa V, Nikolopoulos G, Tsiara C, Magiorkinis G, Psichogiou M, Flampouris A, Mardarescu M, Niculescu I, Batan I, Malliori M, Otelea D, Hatzakis A. *Enhanced HIV-1 surveillance using molecular epidemiology to study and monitor HIV-1 outbreaks among intravenous drug users (IDUs) in Athens and Bucharest.* Infect Genet Evol. 2015 Oct;35:109-21

62.Pérez-Alvarez L, Muñoz M, Delgado E, Miralles C, Ocampo A, García V, Thomson M, Contreras G, Nájera R, Spanish Group for Antiretroviral Resistance Studies in Galicia (2006). *Isolation and biological characterization of HIV-1 BG intersubtype recombinants and other genetic forms circulating in Galicia, Spain.* J Med Virol; 78(12):1520-8.

63.Bartolo I, Camacho R, Barroso H, Bezerra V, Taveira N(2009). *Rapid clinical progression to AIDS and death in a persistently seronegative HIV-1 infected heterosexual young man. AIDS*.; 23:2359–2362.