Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Χημικών Μηχανικών

Διπλωματική Εργασία

Μεγιστοποίηση της απόδοσης της συναινετικής βαθμολόγησης κατά την σύνδεση πρωτεΐνης-προσδέτη

Φώτης Α. Γιάγκας

Επιβλέπων:
Αν.Καθηγητής ΕΜΠ Σαρίμβεης Χαράλαμπος

Ιούλιος 2015

(η σελίδα αυτή αφέθηκε σκόπιμα λευκή)

Περίληψη

Οι υπολογισμοί μοριακής σύνδεσης πρωτεΐνης-προσδέτη χρησιμοποιούνται ευρέως στην ιατρική χημεία για την ανακάλυψη ή την βελτιστοποίηση νέων υποψήφιων φαρμάκων. Αυτοί οι υπολογισμοί χρησιμοποιούν διάφορους τύπους συναρτήσεων βαθμολόγησης (ΣΒ), όπως είναι οι βασισμένες στη μοριακή μηχανική ,οι εμπειρικές , οι βασισμένες στη γνώση ή υβρίδια τους. Εντούτοις, οι επιδόσεις τους ποικίλλουν ανάλογα με το στόχο που μελετάται. Έχοντας ένα σύνολο προσδετών με πειραματικές σταθερές πρόσδεσης μπορεί κανείς να προσαρμόσει ή να επιλέξει διαφορετικές ΣΒ οι οποίες έχουν καλές επιδόσεις στο στόχο του ενδιαφέροντος. Η μέθοδος που χρησιμοποιήθηκε βασίζεται σε έναν γενετικό αλγόριθμο ο οποίος προτείνει αποδοτικούς συνδυασμούς ΣΒ (Συναινετική Βαθμολόγηση) έχοντας να επιλέξει μέσα από συνολικά 46 ΣΒ. Η αποτελεσματικότητα της προτεινόμενης μεθόδου παρουσιάζεται μέσα από την εφαρμογή της σε 8 διαφορετικούς στόχους. Οι στόχοι αυτοί είναι η AKT κινάση (AKT1), η β'-λακταμάση (AMPC), το κυτοχρώμα P450 3A4 (CP3A4),ο υποδοχέας χημειοκινών CXC τύπου 4 (CXCR4), ο υποδοχέας γλυκοκορτικοειδών (GCR) , η HIV-1 αντίστροφη μεταγραφάση (HIVRT), η κινεσίνη 1(KIF11) και η πρωτεάση του HIV, HIV-1 (HIVPR). Επίσης αξιολογείται η απόδοση των άκαμπτων και εύκαμπτων προσεγγίσεων σύνδεσης, στο πλαίσιο συναινετικής βαθμολόγησης, χρησιμοποιώντας δύο δημοφιλή προγράμματα υπολογισμού μοριακής πρόσδεσης, τα Vina και Glide, καθώς και αναβαθμολογώντας τα αποτελέσματα των προγραμμάτων αυτών με 45 επιπλέον ΣΒ. Ο αλγόριθμος συναινετικής βαθμολόγησης που αναπτύχθηκε είναι συμβατός με οποιοδήποτε λογισμικό υπολογισμού μοριακής σύνδεσης (FlexX, eHiTS, Surflex, GOLD,DOCK, κ.α.).

Λέξεις-Κλειδιά

Υπολογισμός Μοριακής Πρόσδεσης ,Συναινετική Βαθμολόγηση , Σχεδιασμός Φαρμάκων , Αναβαθμολόγηση ,Υποδοχέας ,Προσδέτης , Γενετικός Αλγόριθμος

Abstract

Protein-ligand docking calculations are widely used in medicinal chemistry for discovering or optimizing new drug candidates. These calculations employ various types of scoring functions (SFs), such as molecular mechanics-based, empirical, knowledge-based, or hybrids of them. However, their performance varies according to the target receptor. Having a set of ligands with experimentally derived binding constants one can customize a SF or select a few different SFs that perform well on the target of interest. The method used is based on a genetic algorithm that proposes efficient combinations SF(Consensus Scoring) having to choose from a total of 46 SF. The efficiency of the proposed method is illustrated through its application to 8 diverse targets: AKT kinase (AKT1), Beta-lactamase (AMPC), Cytochrome P450 3A4 (CP3A4), C-X-C chemokine receptor type 4 (CXCR4), Glucocorticoid receptor (GCR), HIV-1 reverse transcriptase (HIVRT) Kinesin-like protein 1 (KIF11), HIV-1 protease (HIVPR). We also assess the performance of rigid and flexible docking approaches in the consensus scoring framework using two popular docking programs, Vina and Glide, and rescoring the results of these programs with 45 additional SF. Our Consensus Scoring Algorithm is compatible with any docking software (FlexX, eHiTS, Surflex, GOLD, DOCK, and others).

Keywords

Docking, Consensus Scoring,Drug Design , Rescoring, Receptor, Ligand, Genetic Algorithm

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Πρόλογος και Ευχαριστίες

Η διπλωματική εργασία που κρατάτε στα χέρια σας είναι μια προσέγγιση βελτίωσης του μέχρι τώρα τρόπου υπολογισμού μοριακής πρόσδεσης στον τομέα σχεδιασμού φαρμάκων. Η προσέγγιση αυτή είναι ένα προϊόν συνεργασίας της μονάδας Αυτόματης Ρύθμισης και Πληροφορικής της Σχολής Χημικών Μηχανικών ΕΜΠ με τον τομέα Φαρμακευτικής Χημείας του Τμήματος Φαρμακευτικής ΕΚΠΑ .

Θα ήθελα να ευχαριστήσω τον επιβλέποντά μου , αναπληρωτή καθηγητή Χ. Σαρίμβεη για την δυνατότητα που μου έδωσε να συμμετέχω στη μελέτη αυτή και να βοηθήσω στην εκπόνησή της. Μαζί με αυτόν ευχαριστώ και τον υποψήφιο διδάκτορα Θ.Ευαγγελίδη για την καθοδήγησή του καθ' όλη τη διάρκεια της μελέτης.

Κλείνοντας τον κύκλο των προπτυχιακών μου σπουδών θα ήθελα να θυμηθώ και να ευχαριστήσω τους ανθρώπους με διαμόρφωσαν όλα αυτά τα χρόνια ως φοιτητή εντός και εκτός σχολής φίλους, συμφοιτητές, καθηγητές.

Τέλος, γνωρίζοντας ότι δεν μας δίνεται συχνά η ευκαιρία να ευχαριστήσουμε επίσημα τους ανθρώπους που μας βοήθησαν στην εκπλήρωση των στόχων μας, δεν θα μπορούσα να παραλείψω την οικογένειά μου: τους γονείς μου Ανδρέα Γιάγκα και Μαρία Βαρθάλη και τον αδερφό μου Γιώργο Γιάγκα. Χάρη στις δικές τους θυσίες, την υπομονή και την κατανόηση τους μπόρεσα να φτάσω έως εδώ, αντιμετωπίζοντας τα εμπόδια που συχνά συναντούσα.

Στην οικογένειά μου

Κεφάλαιο 1

1 Σύνδεση πρωτεΐνης-προσδέτη

1.1 Διαδικασία Ανακάλυψης φαρμάκων

Η δημιουργία ενός φαρμάκου αρχίζει με τον προσδιορισμό μιας πρωτεΐνης που σχετίζεται με μια νόσο. Οι πρωτεΐνες αυτές είναι γνωστές ως «στόχοι». Όταν επιβεβαιωθεί ότι ένας στόχος παίζει ρόλο στην εκδήλωση μιας νόσου, με τη χρήση τεχνολογιών ταχείας ανάλυσης (high-throughput screening) ανευρίσκεται ένα μόριο ή αντίσωμα το οποίο προσδένεται ή «προσβάλλει» τον στόχο (υποδοχέα) έτσι ώστε να ανασταλεί η δραστηριότητά του, δηλαδή να τροποποιηθεί η νόσος. Με υπολογιστικές μεθόδους τελειοποιείται η ενεργός περιοχή πρόσδεσης μορίου και υποδοχέα ώστε να βελτιωθεί η ασφάλεια και η δραστικότητα του υποψήφιου φαρμάκου. Αυτή η διαδικασία διαρκεί από 2 έως 4 έτη.

Ένα αρχικό προφίλ ασφάλειας και αποτελεσματικότητας του υποψήφιου φαρμάκου πρέπει να προσδιορισθεί πριν αυτό μελετηθεί σε ανθρώπους. Σε αυτή τη φάση, οι ερευνητές χρησιμοποιούν υπολογιστικά μοντέλα και εργαστηριακές δοκιμασίες για να αξιολογήσουν την ασφάλεια του εν λόγω φαρμάκου. Συγκεκριμένα, αυτές οι δοκιμασίες αξιολογούν την απορρόφηση, την κατανομή στους ιστούς, την βιομετατροπή και την απέκκριση από τον οργανισμό.

Στις μελέτες Proof-of-Concept (PoC), το υποψήφιο φάρμακο χορηγείται σε μικρό αριθμό ασθενών (5-15) για να κατανοηθεί ο μηχανισμός δράσης του και να σχηματισθεί μια πρώτη αντίληψη του τρόπου με τον οποίο το φάρμακο τροποποιεί τη νόσο. Εάν η μελέτη PoC ευοδωθεί, το υποψήφιο φάρμακο μπορεί να εισέλθει σε μελέτες Φάσης I (20-80 ασθενείς ή υγιείς εθελοντές) ώστε να αξιολογηθεί η ασφάλειά του, να καθορισθεί η ασφαλής δοσολογία και να εξακριβωθούν ανεπιθύμητες ενέργειες. Συχνά, τα υποψήφια φάρμακα εισέρχονται απευθείας από την PoC στις μελέτες Φάσης II. Αυτή η διαδικασία διαρκεί από 4 έως 7 έτη.

Στις μελέτες Φάσης II, το φάρμακο χορηγείται σε μια μεγαλύτερη ομάδα ασθενών (100-300) για να δοκιμασθεί η αποτελεσματικότητά του, να καθορισθεί η κατάλληλη δοσολογία και να αξιολογηθεί περαιτέρω η ασφάλειά του. Στις μελέτες Φάσης II, το φάρμακο χορηγείται σε μεγάλες ομάδες ασθενών (1.000-3.000) ώστε να επιβεβαιωθεί η αποτελεσματικότητά του, να καταγραφούν οι ανεπιθύμητες ενέργειες, να συγκριθεί με άλλα συνήθως χορηγούμενα φάρμακα και να συγκεντρωθούν πληροφορίες που θα επιτρέψουν την ασφαλή χορήγησή του. Αυτή η διαδικασία διαρκεί από 1 έως 2 έτη.

Για να εγκριθεί ένα φάρμακο, τα αποτελέσματα όλων των προκλινικών και κλινικών μελετών μαζί με την περιγραφή της διαδικασίας παραγωγής υποβάλλονται στις Ρυθμιστικές Αρχές. Εφ’ όσον τα δεδομένα αποδεικνύουν την ασφάλεια, την αποτελεσματικότητα και την ποιότητα, τότε χορηγείται άδεια κυκλοφορίας και το φάρμακο μπορεί να διατεθεί στην αγορά. Οι ανεπιθύμητες ενέργειες επιβάλλεται να παρακολουθούνται και να αναφέρονται στις Ρυθμιστικές Αρχές. Επιπρόσθετα, διεξάγονται

συχνά μελέτες Φάσης IV για νέες ενδείξεις ή βελτίωση των υφισταμένων φαρμακοτεχνικών μορφών.

1.2 Σχεδιασμός φαρμάκων

Ο Σχεδιασμός φαρμάκων, είναι η διαδικασία για την εξεύρεση νέων φαρμάκων με βάση τη γνώση ενός βιολογικού στόχου. Το φάρμακο είναι συνήθως ένα μικρό οργανικό μόριο που ενεργοποιεί ή αναστέλλει την λειτουργία ενός βιομορίου, όπως μία πρωτεΐνη, και το οποίο οδηγεί σε ένα θεραπευτικό όφελος για τον ασθενή. Στην πιο βασική έννοια, ο σχεδιασμός φαρμάκων περιλαμβάνει το σχεδιασμό μικρών μορίων που είναι συμπληρωματικά σε σχήμα και φορτίο στο βιομοριακό στόχο με τα οποία αλληλεπιδρούν και συνεπώς θα συνδέονται με αυτό. Ο σχεδιασμός φαρμάκων στηρίζεται συχνά, αλλά όχι κατ 'ανάγκη σε τεχνικές μοντελοποίησης με υπολογιστή.

Υπάρχουν δύο κύριοι τύποι του σχεδιασμού φαρμάκων. Ο πρώτος αναφέρεται ως βασισμένος στον προσδέτη σχεδιασμός φαρμάκων, ενώ ο δεύτερος ως βασισμένος στην δομή σχεδιασμός φαρμάκων **.**

Ο σχεδιασμός φαρμάκων βάσει προσδέτη (έμμεσος σχεδιασμός φαρμάκων) βασίζεται στη γνώση άλλων μορίων που δεσμεύονται με το βιολογικό στόχο του ενδιαφέροντος. Αυτά τα μόρια μπορούν να χρησιμοποιηθούν για να αναπτυχθεί ένα μοντέλο φαρμακοφόρου που καθορίζει τα ελάχιστα απαραίτητα δομικά χαρακτηριστικά τα οποία πρέπει να έχει ένα μόριο , ώστε να συνδέεται στον στόχο. Με άλλα λόγια, μπορεί να κατασκευαστεί ένα μοντέλο του βιολογικού στόχου γνωρίζοντας τι συνδέεται με αυτό και αυτό το μοντέλο με τη σειρά του μπορεί να χρησιμοποιηθεί για το σχεδιασμό νέων μοριακών οντοτήτων που αλληλεπιδρούν με το στόχο. Εναλλακτικά, μπορούν να χρησιμοποιηθούν μοντέλα ποσοτικής σχέσης δομής-δραστικότητας (QSAR), που συσχετίζουν τις υπολογισμένες ιδιότητες των μορίων με την πειραματικά προσδιορισμένη βιολογική τους δράση.

Ο σχεδιασμός φαρμάκων βάσει δομής (άμεσος σχεδιασμός φαρμάκων) στηρίζεται στη γνώση της τρισδιάστατης δομής του βιολογικού στόχου που λαμβάνεται μέσω μεθόδων όπως κρυσταλλογραφία ακτινών Χ ή φασματοσκοπία NMR. Εάν η πειραματική δομή ενός στόχου δεν είναι διαθέσιμη, είναι δυνατόν να δημιουργηθεί ένα μοντέλο του στόχου με βάση την πειραματική δομή μιας σχετικής πρωτεΐνης. Χρησιμοποιώντας τη δομή του βιολογικού στόχου, μπορούν να σχεδιαστούν υποψήφια φάρμακα που προβλέπεται να προσδένονται με υψηλή συγγένεια και εκλεκτικότητα στον στόχο.

Οι τρέχουσες μέθοδοι για το σχεδιασμό φαρμάκων που βασίζονται στη δομή μπορούν να χωριστούν χονδρικά σε δύο κατηγορίες. Η πρώτη κατηγορία είναι η "εύρεση" προσδετών για ένα συγκεκριμένο υποδοχέα, η οποία συνήθως αναφέρεται ως αναζήτηση σε βάση δεδομένων. Σε αυτή την περίπτωση, επιλέγεται ένας μεγάλος αριθμός πιθανών μορίων ώστε να βρεθούν εκείνα που «ταιριάζουν» στον θύλακα σύνδεσης του υποδοχέα. Μια άλλη κατηγορία σχεδιασμού φαρμάκων που βασίζεται στη δομή είναι η "κατασκευή" προσδετών, η οποία συνήθως αναφέρεται ως σχεδιασμός φαρμάκου με βάση τον υποδοχέα. Σε αυτή την περίπτωση, τα μόρια του προσδέτη δημιουργούνται εντός των περιορισμών του θύλακα σύνδεσης με τη συναρμολόγηση μικρών κομματιών με σταδιακό τρόπο. Αυτά τα κομμάτια μπορούν να είναι μεμονωμένα άτομα ή μοριακά θραύσματα. Το βασικό πλεονέκτημα μιας τέτοιας μεθόδου είναι ότι μπορούν να προταθούν νέες δομές, που δεν περιέχονται σε καμία βάση δεδομένων.

Αν και η δομή μιας πρωτεΐνης αποτελεί το κλειδί για τη βιολογική της λειτουργία, για πολλές πρωτεΐνες η επίλυση της δομής τους δεν είναι αρκετή για να καθοριστεί η λειτουργία τους. Πολλά ένζυμα εντείνουν την καταλυτική τους λειτουργία με βάση μια μικρή περιοχή στην πρωτεϊνική επιφάνεια που ονομάζεται ενεργός περιοχή (active site) ή ενεργό κέντρο του ενζύμου. Αυτή η περιοχή χαρακτηρίζεται από γεωμετρικά και φυσικοχημικά χαρακτηριστικά που είναι σχεδόν συμπληρωματικά ενός άλλου μορίου, του υποστρώματος. Έτσι το ενεργό κέντρο μιας πρωτεΐνης ενεργεί σαν υποδοχέας. Αυτή η διαδικασία πρόσδεσης υποδοχέα και υποστρώματος(docking).

Η προσπάθεια εντοπισμού του ενεργού κέντρου και της κατανόησης με ακρίβεια της διαδικασίας προσάραξης αποτελεί ένα πολύ σημαντικό βήμα στην προσπάθεια αποκρυπτογράφησης των περισσότερων μεταβολικών αντιδράσεων. Με την κατανόηση της πρωτεϊνικής λειτουργίας ο σχεδιασμός φαρμάκων μπορεί να αναπτυχθεί σημαντικά.

Προκειμένου η πρωτεΐνη να βρεθεί σε μια ενεργειακή ισορροπία, ιδανική για την προσάραξή της, περνά από ένα σύνολο στεροδιαμορφώσεων. Υπάρχουν εκατομμύρια διαμορφώσεις οι οποίες μπορούν να διαφέρουν σημαντικά. Ακριβώς η πρόβλεψη του τρόπου με τον οποίο δυο πρωτεΐνες έρχονται σε επαφή μεταξύ τους αποτελεί το πρόβλημα της πρωτεϊνικής σύνδεσης (protein docking), και βοηθά στην κατανόηση της αλληλεπίδρασης μεταξύ δυο πρωτεϊνών.

Ο ηλεκτρονικός υπολογιστής ελέγξει τον μεγάλο αριθμό πιθανών στεροδιαμορφώσεων και μειώνει την υπολογιστική πολυπλοκότητα των πειραμάτων που πρέπει να πραγματοποιηθούν.

1.3 Υπολογισμός μοριακής πρόσδεσης

Ο υπολογισμός μοριακής πρόσδεσης μπορεί να παρομοιαστεί ως ένα πρόβλημα «κλειδιού-κλειδαριάς» στο οποίο κάποιος θέλει να βρει τον σωστό σχετικό προσανατολισμό του «κλειδιού» που θα ανοίξει την "κλειδαριά. Εδώ, η πρωτεΐνη μπορεί να θεωρηθεί ως η «κλειδαριά» και ο προσδέτης μπορεί να θεωρηθεί ως ένα «κλειδί». Μπορεί να οριστεί ως ένα πρόβλημα βελτιστοποίησης, το οποίο θα περιγράφει τον "καλύτερο" προσανατολισμό ενός προσδέτη που δεσμεύεται σε μια συγκεκριμένη πρωτεΐνη. Κατά τη διάρκεια της διαδικασίας του υπολογισμού μοριακής πρόσδεσης , ο προσδέτης και η πρωτεΐνη προσαρμόζουν την διάπλαση τους ώστε να επιτύχουν το καλύτερο «ταίριασμα» σε αυτό το είδος της διαμορφωτικής προσαρμογής με αποτέλεσμα την συνολική δέσμευση .

Ο υπολογισμός μοριακής πρόσδεσης επικεντρώνεται στην υπολογιστική προσομοίωση της διαδικασίας μοριακής αναγνώρισης. Έχει ως στόχο να επιτύχει μια βελτιστοποιημένη διαμόρφωση τόσο για την πρωτεΐνη και τον προσδέτη όσο και για τον σχετικό προσανατολισμό μεταξύ της πρωτεΐνης και του προσδέτη , έτσι ώστε η ελεύθερη ενέργεια του συνολικού συστήματος να ελαχιστοποιείται .

Δύο προσεγγίσεις είναι ιδιαίτερα δημοφιλείς στην κοινότητα του υπολογισμού μοριακής πρόσδεσης .Η πρώτη προσέγγιση χρησιμοποιεί μια τεχνική ταιριάσματος που περιγράφει την πρωτεΐνη και τον προσδέτη ως συμπληρωματικές επιφάνειες (γεωμετρικές μέθοδοι). Η δεύτερη προσέγγιση προσομοιώνει την **πραγματική διαδικασία σύνδεσης** με την οποία υπολογίζονται οι κατά ζεύγη ενέργειες αλληλεπίδρασης προσδέτη-πρωτεΐνης . Τόσο η πρώτη όσο και η δεύτερη προσέγγιση έχουν σημαντικά πλεονεκτήματα, καθώς και ορισμένους περιορισμούς.

Οι **γεωμετρικές μέθοδοι συμπληρωματικότητας** περιγράφουν την πρωτεΐνη και τον προσδέτη σαν ένα σύνολο από χαρακτηριστικά που τα καθιστούν δυνατά για μοριακή πρόσδεση. Η συμπληρωματικότητα μεταξύ των δύο επιφανειών ισοδυναμεί με το ταίριασμα της μορφής που μπορεί να βοηθήσει στην εύρεση της συμπληρωματικής πόζας μοριακής πρόσδεσης του στόχου με τα μόρια του προσδέτη. Μια άλλη προσέγγιση είναι η περιγραφή των υδρόφοβων χαρακτηριστικών της πρωτεΐνης χρησιμοποιώντας τις στροφές στα άτομα κύριας αλυσίδας. Μια ακόμα προσέγγιση είναι να χρησιμοποιηθεί μια τεχνική Fourier. Ενώ οι προσεγγίσεις συμπληρωματικότητας κατά σχήμα είναι συνήθως γρήγορες και ισχυρές, δεν μπορούν να μοντελοποιήσουν τις κινήσεις ή τις δυναμικές μεταβολές στις διαμορφώσεις προσδέτη-πρωτεΐνης με ακρίβεια, παρόλο που οι πρόσφατες εξελίξεις επιτρέπουν σε αυτές τις μεθόδους να διερευνήσουν ακόμα και την ευελιξία του προσδέτη . Χρησιμοποιώντας αυτές τις μεθόδους έχουμε την δυνατότητα να σαρώσουμε γρήγορα αρκετές χιλιάδες προσδέτες σε λίγα δευτερόλεπτα, να διαπιστώσουμε αν μπορούν να συνδεθούν στο ενεργό κέντρο της πρωτεΐνης και να επεκταθούμε ακόμη και σε αλληλεπιδράσεις μεταξύ πρωτεϊνών.

Η **προσομοίωση της διαδικασίας σύνδεσης** είναι αρκετά πιο σύνθετη. Η πρωτεΐνη και ο συμπλοκοποιητής διαχωρίζονται από κάποια φυσική απόσταση, και ο προσδέτης βρίσκει τη θέση του στο ενεργό κέντρο της πρωτεΐνης μετά από ένα ορισμένο αριθμό «κινήσεων» στο χώρο διαμόρφωσης . Οι κινήσεις αφορούν μεταμορφώσεις του σταθερού σώματος όπως μεταφράσεις και περιστροφές, καθώς και τις εσωτερικές αλλαγές στη δομή του συνδέτη, συμπεριλαμβανομένων των περιστροφών των γωνιών στρέψης. Κάθε μία από αυτές τις κινήσεις στο χώρο διαμόρφωσης του συνδέτη προκαλεί ένα συνολικό ενεργειακό κόστος του συστήματος. Ως εκ τούτου, η συνολική ενέργεια του συστήματος υπολογίζεται μετά από κάθε κίνηση.

Το προφανές πλεονέκτημα της προσομοίωσης σύνδεσης είναι ότι η ευελιξία του συνδέτη ενσωματώνεται εύκολα, ενώ οι τεχνικές συμπληρωματικότητας σχήματος πρέπει να χρησιμοποιούν έξυπνες μεθόδους για να συμπεριλάβουν την ευελιξία των συνδετών. Επίσης, είναι πιο κοντά στην πραγματικότητα, ενώ οι τεχνικές συμπληρωματικότητας σχήματος είναι περισσότερο ανακριβείς .

Σαφώς, η προσομοίωση είναι υπολογιστικά δαπανηρή καθώς χρειάζεται να εξερευνήσει ένα μεγάλο ενεργειακό τοπίο. Οι τεχνικές βάσει πλέγματος, οι μέθοδοι βελτιστοποίησης και η αύξηση της ταχύτητας των υπολογιστών έχουν κάνει την προσομοίωση της μοριακής πρόσδεσης πιο ρεαλιστική.

Μηχανική της σύνδεσης

Η επιτυχία μιας μεθοδολογίας σύνδεσης εξαρτάται από δύο συνιστώσες. Αυτές είναι ο αλγόριθμος αναζήτησης και η συνάρτηση βαθμολόγησης. Ο χώρος αναζήτησης θεωρητικά αποτελείται από όλες τις πιθανές κατευθύνσεις και διαμορφώσεις της πρωτεΐνης σε συνδυασμό με τον προσδέτη. Ωστόσο, στην πράξη με τους υπάρχοντες υπολογιστικούς πόρους, είναι αδύνατο να διερευνηθεί διεξοδικά ο χώρος καθώς αυτή η αναζήτηση περιλαμβάνει την απαρίθμηση όλων των πιθανών στρεβλώσεων του κάθε μορίου (τα μόρια είναι δυναμικά και υπάρχουν σε ένα σύνολο καταστάσεων διαμόρφωσης) και όλων των πιθανών περιστροφών και των προσανατολισμών μετατόπισης του προσδέτη σε σχέση με την πρωτεΐνη σε ένα δεδομένο επίπεδο. Τα περισσότερα προγράμματα σύνδεσης που εφαρμόζονται για ένα ευέλικτο προσδέτη, και αρκετά άλλα προσπαθούν να διαμορφώσουν ένα ευέλικτο υποδοχέα της πρωτεΐνης. Κάθε «στιγμιότυπο» του ζεύγους αναφέρεται ως πόζα.

Ευελιξία Προσδέτη

Οι διαμορφώσεις του προσδέτη μπορεί να δημιουργηθούν απουσία του υποδοχέα και στη συνέχεια να πραγματοποιηθεί μοριακή πρόσδεση(docking) ή μπορούν να παραχθούν παρουσία της κοιλότητας δέσμευσης του υποδοχέα (binding cavity) . Επίσης με πλήρη ευελιξία περιστροφής κάθε δίεδρης γωνίας χρησιμοποιώντας σύνδεση βάσει κλάσματος (fragment based docking). Η αξιολόγηση ενεργειακού πεδίου (Force field energy evaluation) χρησιμοποιείται συχνά για την επιλογή ενεργειακά λογικών διαμορφώσεων ,αλλά έχουν επίσης χρησιμοποιηθεί μέθοδοι που βασίζονται στη γνώση(knowledge-based methods).

Ευελιξία υποδοχέα

Η υπολογιστική ικανότητα έχει αυξηθεί δραματικά την τελευταία δεκαετία, καθιστώντας δυνατή τη χρήση των πιο εξελιγμένων και υπολογιστικά δύσκολων μεθόδων στον σχεδιασμό φαρμάκων μέσω υπολογιστή . Ωστόσο, η ευελιξία του υποδοχέα στις μεθοδολογίες σύνδεσης εξακολουθεί να είναι ένα ακανθώδες ζήτημα. Ο κύριος λόγος πίσω από αυτή την δυσκολία είναι ο μεγάλος αριθμός των βαθμών ελευθερίας που πρέπει να ληφθούν υπόψη σε αυτό το είδος των υπολογισμών. Η αγνόηση του μεγάλου αριθμού βαθμών ελευθερίας οδηγεί σε άσχημα αποτελέσματα σύνδεσης όσον αφορά τη πρόβλεψη της πόζας δέσμευσης (docking pose prediction). Έχουν προσδιοριστεί πειραματικά πολλαπλές στατικές δομές για την ίδια πρωτεΐνη σε διαφορετικές διαμορφώσεις και χρησιμοποιούνται συχνά για να μιμηθούν την ευελιξία του υποδοχέα.

1.4 Συνάρτηση Βαθμολόγησης (Scoring Function)

Η συνάρτηση βαθμολόγησης παίρνει μια πόζα ως είσοδο και επιστρέφει έναν αριθμό που υποδηλώνει την πιθανότητα ότι η πόζα αποτελεί θετική αλληλεπίδραση δέσμευσης (favorable binding interaction).

Οι περισσότερες συναρτήσεις βαθμολόγησης είναι βασισμένες στη φυσική ,στη μοριακή μηχανική και στο δυναμικό πεδίο και υπολογίζουν την ενέργεια της πόζας. Μια χαμηλή τιμή (αρνητική) ενέργειας δείχνει ένα σταθερό σύστημα και έτσι μια πιθανή αλληλεπίδραση δέσμευσης. Μια εναλλακτική προσέγγιση είναι η άντληση μιας στατιστικής πιθανότητας αλληλεπιδράσεων από μια μεγάλη βάση δεδομένων πρωτεΐνης-προσδέτη , όπως η Protein Data Bank καθώς και η αξιολόγηση της προσαρμογής της πόζας, σύμφωνα με το παρόν δυναμικό.

Ο υπολογισμός μοριακής πρόσδεσης σε συνδυασμό με μια συνάρτηση βαθμολόγησης μπορεί να χρησιμοποιηθεί για την γρήγορη διαλογή, από μεγάλες βάσεις δεδομένων, των πιθανών φαρμάκων in silico για τον εντοπισμό μορίων τα οποία είναι πιθανόν να συνδεθούν με την πρωτεΐνη που μας ενδιαφέρει (virtual screening hit identification).Επίσης μπορεί να χρησιμοποιηθεί για να προβλέψει πού και με ποιο σχετικό προσανατολισμό ένας προσδέτης συνδέεται με μία πρωτεΐνη (αναφέρεται επίσης ως λειτουργία πρόσδεσης ή πόζα). Αυτή η πληροφορία μπορεί στη συνέχεια να χρησιμοποιηθεί για το σχεδιασμό πιο ισχυρών και επιλεκτικών αναλόγων (lead optimization). Ο υπολογισμός μοριακής πρόσδεσης πρωτεΐνης-προσδέτη μπορεί επίσης να χρησιμοποιηθεί για να προβλέψει τους ρύπους που μπορεί να αποικοδομούνται από τα ένζυμα (Bioremediation).

Υπάρχουν τρεις γενικές τάξεις των λειτουργιών βαθμολόγησης:

Συναρτήσεις Βαθμολόγησης δυναμικού πεδίου

Οι συγγένειες υπολογίζονται αθροίζοντας την αντοχή των διαμοριακών δυνάμεων van der Waals με τις ηλεκτροστατικές αλληλεπιδράσεις μεταξύ όλων των ατόμων των δύο μορίων στο συγκρότημα. Οι ενδομοριακές ενέργειες των δύο μορίων μπορούν επίσης να ληφθούν υπόψη. Τέλος, επειδή η σύνδεση γίνεται συνήθως με την παρουσία νερού, λαμβάνονται υπόψη η ενέργεια αποδιαλύτωσης του συνδέτη και της πρωτεΐνης με τη χρήση μεθόδων άμεσης διαλύτωσης όπως GBSA ή PBSA .

Εμπειρικές Συναρτήσεις Βαθμολόγησης

Οι συναρτήσεις αυτές βασίζονται στην καταμέτρηση του αριθμού των διαφόρων τύπων αλληλεπιδράσεων μεταξύ των δύο εταίρων δέσμευσης. Η μέτρηση μπορεί να βασίζεται στον αριθμό των ατόμων του προσδέτη και του υποδοχέα που έρχονται σε επαφή με το άλλο ή με τον υπολογισμό της μεταβολής της προσβάσιμης επιφάνειας διαλύτη στο σύμπλοκο σε σύγκριση με το μη συμπλοκοποιημένο προσδέτη-πρωτεΐνη. Οι συντελεστές της συνάρτησης βαθμολόγησης συνήθως συμπληρώνονται με τη χρήση πολλαπλών μεθόδων γραμμικής παλινδρόμησης. Αυτές οι αλληλεπιδράσεις στους όρους της λειτουργίας μπορεί να περιλαμβάνουν επαφές υδρόφοβου–υδρόφοβου(ευνοϊκή), υδρόφοβου–υδρόφιλου(δυσμενής).

Συναρτήσεις Βαθμολόγησης βασισμένες στη γνώση

Οι βασισμένες στη γνώση (στατιστικά δυναμικά) βασίζονται σε στατιστικές παρατηρήσεις των διαμοριακών επαφών από μεγάλες τρισδιάστατες βάσεις δεδομένων , οι οποίες χρησιμοποιούνται για την παραγωγή «δυναμικών μέσης δύναμης» . Η μέθοδος αυτή βασίζεται στην υπόθεση ότι οι μοριακές αλληλεπιδράσεις μεταξύ ορισμένων τύπων ατόμων ή λειτουργικών ομάδων που συμβαίνουν πιο συχνά από ό, τι θα περίμενε κανείς από μια τυχαία κατανομή είναι πιθανό να είναι ενεργειακά ευνοϊκές και , επομένως, συμβάλλουν θετικά στην συγγένεια πρόσδεσης .

Συναινετική Βαθμολόγηση

Παρά τον υψηλό αριθμό των συναρτήσεων βαθμολόγησης που έχουν αναπτυχθεί, κανένα από αυτά δεν είναι τέλειο από την άποψη της ακρίβειας και γενικής εφαρμογής. Κάθε συνάρτηση βαθμολόγησης έχει πλεονεκτήματα και περιορισμούς. Για να αξιοποιηθούν τα πλεονεκτήματα και να ισορροπηθούν οι ελλείψεις των διαφορετικών συναρτήσεων βαθμολόγησης καθώς και να αυξηθεί η πιθανότητα εύρεσης σωστών λύσεων έχει εισαχθεί η τεχνική συναινετικής βαθμολόγησης η οποία συνδυάζει τα αποτελέσματα από πολλαπλές συναρτήσεις βαθμολόγησης. Το κρίσιμο βήμα στη συναινετική βαθμολόγηση είναι ο σχεδιασμός μιας κατάλληλης στρατηγικής των επιμέρους βαθμολογιών, έτσι ώστε οι πραγματικοί προσδέτες να μπορούν να διακριθούν από τους άλλους.

1.4.1 Κριτήρια επίδοσης μιας συνάρτησης βαθμονόμησης

Ένα από τα βασικά κριτήρια επίδοσης μιας συνάρτησης βαθμονόμησης είναι η ικανότητά της **να διακρίνει τους ενεργούς προσδέτες(actives) από τους μη ενεργούς(decoys).** Συγκεκριμένα, σε ένα σύνολο από μη ενεργά μόρια για ένα σύμπλοκο πρωτεΐνης-προσδέτη, μια αξιόπιστη συνάρτηση βαθμολόγησης πρέπει να είναι σε θέση να κατατάσσει στη κορυφή τη φυσική δομή υπολογίζοντας βαθμολογίες δέσμευσης. Σε εφαρμογές υπολογισμού μοριακής σύνδεσης, η επιτυχής πρόβλεψη ενός φυσικού τρόπου πρόσδεσης

ορίζεται συνήθως από την τιμή της μέσης τετραγωνικής απόκλισης RMSD μεταξύ των κορυφαίων διαμορφώσεων συνδέτη και την πειραματικά παρατηρούμενη (μητρική) δομή.

Η ρίζα της RMSD είναι το μέτρο της μέσης απόστασης μεταξύ των ατόμων επάλληλων πρωτεϊνών. Η τιμή RMSD υπολογίζεται από την εξίσωση :

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

όπου δ είναι η απόσταση μεταξύ N ζευγών ισοδύναμων ατόμων (συνήθως Cα και μερικές φορές C, N, O, Cβ).

Λόγω της απλότητάς της και της ευκολίας εφαρμογής, το κριτήριο RMSD έχει χρησιμοποιηθεί ευρέως στο πεδίο για τη πρόβλεψη σύνδεσης. Ωστόσο, το κριτήριο αυτό θα μπορούσε να παρουσιάσει προβλήματα σε ορισμένες περιπτώσεις. Για παράδειγμα, μικροί ή σχεδόν συμμετρικοί συνδέτες είναι πιθανό να αποκτήσουν καλές τιμές RMSD ακόμη και όταν τοποθετούνται τυχαία σε ένα μικρό ενεργό κέντρο. Αντίθετα, για ένα μεγάλο εύκαμπτο συνδέτη, η μεγάλη τιμή RMSD λόγω διαλύτη που εκτίθενται, μπορεί να κρύψει την ορθότητα στην πρόβλεψη της συνολικής λειτουργίας δέσμευσης.

Ένα δεύτερο σημαντικό κριτήριο για την συνάρτηση βαθμολόγησης είναι **η ικανότητα να προβλέψει την συγγένεια δέσμευσης ενός συμπλόκου**, δηλαδή πόσο σφικτά ο προσδέτης δεσμεύει την πρωτεΐνη. Κριτήρια για την ποσοτικοποίηση της συγγένειας είναι η συσχέτιση Pearson και Spearman μεταξύ των υπολογιζόμενων βαθμολογιών και των πειραματικών δεδομένων.

Το τρίτο κριτήριο για την αξιολόγηση μιας συνάρτησης βαθμολογίας είναι **η δυνατότητα επιλογής πιθανών συνδετών** από μια μεγάλη βάση δεδομένων ενώσεων για ένα συγκεκριμένο πρωτεϊνικό στόχο. Η πρακτική εφαρμογή της είναι η εικονική διαλογή σε υπολογιστή κατά το σχεδιασμό φαρμάκων. Η εικονική διαλογή της βάσης δεδομένων ελέγχει αν μια συνάρτηση βαθμολόγησης είναι σε θέση να κατατάξει τους γνωστούς προσδέτες σε υψηλότερη θέση από τις αδρανείς ενώσεις. Θεωρητικά, μια ακριβής συνάρτηση βαθμολόγησης θα πρέπει να είναι σε θέση να εκτελεί εξίσου καλά και στα τρία κριτήρια σε κάθε σύνολο δοκιμών. Ωστόσο, λόγω των εγγενών περιορισμών, οι περισσότερες από τις υπάρχουσες συναρτήσεις βαθμολόγησης συνήθως αποδίδουν καλά σε μόνο ένα ή δύο από τα κριτήρια και αποτυγχάνουν στα υπόλοιπα. Οι δείκτες που εφαρμόστηκαν στη συγκεκριμένη διπλωματική εργασία περιγράφονται πιο αναλυτικά στη υποκεφάλαιο 1.5.

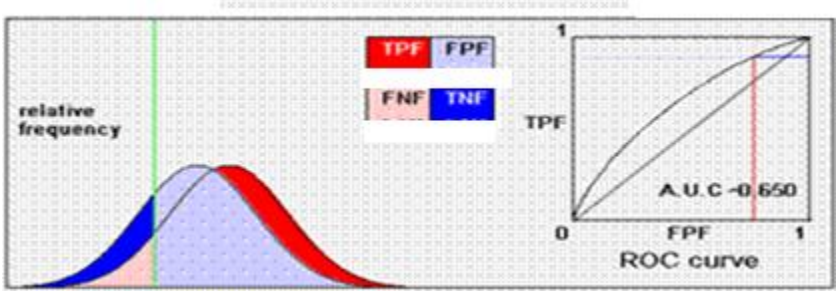
1.5 Μετρητές Ταξινόμησης

1.5.1 Καμπύλες ROC-CROC

Η καμπύλη ROC χρησιμοποιείται για να αξιολογήσει την απόδοση ενός ταξινομητή και παριστάνει γραφικά το ποσοστό των αληθώς θετικών(True Positive Rate) ως προς το ποσοστό ψευδώς θετικών (False Positive Rate, FPR) προβλέψεων όταν μεταβάλλεται η

παράμετρος βάσει της οποίας διαχωρίζονται τα θετικά από τα αρνητικά αποτελέσματα. Η καμπύλη CROC συνδέεται με την ROC από ένα γραμμικό μετασχηματισμό του άξονα χ'χ. Η περιοχή κάτω από την καμπύλη ROC(AUC [ROC)] μπορεί να χρησιμοποιηθεί για την ποσοτικοποίηση της απόδοσης.

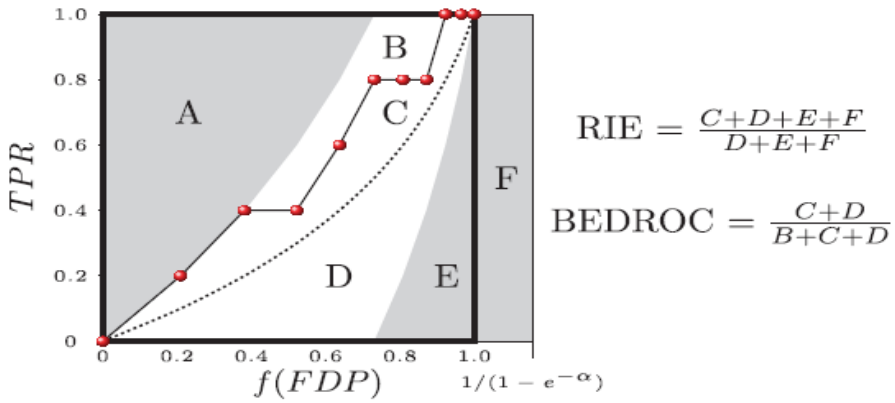
Εντός της ROC, ένας ταξινομητής μπορεί να αξιολογηθεί συγκρίνοντας τις επιδόσεις του με την απόδοση ενός τυχαίου ταξινομητή. Για την ROC , κατά μέσο όρο οι καμπύλες απόδοσης μιας τυχαίας ταξινόμησης είναι ευθείες γραμμές που συνδέουν τα σημεία (0,0) με (1,1).[5]



Εικόνα 1 Παράδειγμα Καμπύλης ROC

1.5.2 Καμπύλες BEDROC

Η BEDROC (Boltzmann-Enhanced Discrimination of ROC) είναι μια γενικευμένη ROC στην οποία έχει ενσωματωθεί μια συνάρτηση στάθμισης. Η τροποποίηση αυτή καθιστά την BEDROC χρήσιμη σε προβλήματα «έγκαιρης αναγνώρισης». Η μόνη διαφορά της BEDROC με την ROC είναι ότι η τελευταία έχει ομοιόμορφη κατανομή ενώ η BEDROC εκθετική κατανομή.[11]



Εικόνα 2 :Η γεωμετρική σχέση μεταξύ μιας καμπύλης CAC(exp) και δύο από τα μετρητικά «έγκαιρης αναγνώρισης». Κάθε γράμμα , από το A έως F, αντιστοιχεί σε μία περιοχή οριοθετείται από τις καμπύλες και τα όρια του διαγράμματος

1.5.3 Συντελεστής συσχέτισης Kendall's τ

Στη στατιστική, ο συντελεστής συσχέτισης τάξης Kendall, που συνήθως αναφέρεται ως συντελεστής Kendall's tau (από ελληνικό γράμμα τ) (Teles, 2012) είναι ένας στατιστικός δείκτης που χρησιμοποιείται για τη μέτρηση της συσχέτισης μεταξύ δύο ποσοτήτων.

Έστω (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) ένα σύνολο παρατηρήσεων των κοινών τυχαίων μεταβλητών Χ και Υ αντίστοιχα, έτσι ώστε όλες οι αξίες του (x_i) και (y_i) είναι μοναδικά . Κάθε ζεύγος των παρατηρήσεων (xi, yi) και (x_j, y_j) λέγεται ότι είναι σύμφωνα εάν οι τάξεις και για τα δύο στοιχεία συμφωνούν, δηλαδή, αν τα δύο $x_i > x_j$ και $y_i > yj$ ή αν τα δύο $x_i < x_j$ και $y_i < Y_J$.Είναι ασύμφωνα, αν $x_i > x_j$ και $y_i < y_j$ ή αν $x_i < x_j$ και $y_i > y_j$. Αν $x_i = x_j$ ή $y_i = y_j$, το ζευγάρι δεν είναι ούτε σύμφωνο ούτε ασύμφωνο.

Ο συντελεστής Kendall's τ ορίζεται ως εξής:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}.$$

Ο παρονομαστής είναι ο συνολικός αριθμός των συνδυασμών των ζευγών, έτσι ώστε ο συντελεστής να έχει εύρος $-1 \le \tau \le 1$.Εάν η συμφωνία μεταξύ των δύο βαθμολογιών είναι τέλεια (δηλαδή, οι δύο ταξινομήσεις είναι οι ίδιες), ο συντελεστής έχει τιμή 1.Εάν η διαφωνία μεταξύ των δύο βαθμολογιών είναι τέλεια (δηλαδή, μία κατάταξη είναι η αντίστροφη της άλλης) ο συντελεστής έχει τιμή -1. Αν Χ και Υ είναι ανεξάρτητες, τότε θα περιμέναμε ο συντελεστής να είναι περίπου μηδέν.

Ο συντελεστής κατάταξης Kendall's τ χρησιμοποιείται συχνά για να καθοριστεί αν δύο μεταβλητές μπορεί να θεωρηθούν στατιστικά εξαρτώμενες. Αυτή η δοκιμή είναι μη-παραμετρική, καθώς δεν βασίζεται σε οποιεσδήποτε υποθέσεις σχετικά με τις κατανομές των Χ ή Υ ή τη κατανομή του (Χ, Υ).[7]

1.5.4 Συντελεστής αντιστοιχίας Top-Down Concordance

Ο συντελεστής αντιστοιχίας Top-down Concordance[7] επικεντρώνεται στη συμφωνία κυρίως στο χαμηλότερο ή στο υψηλότερο επίπεδο μιας κατάταξης. Ο συντελεστής αντιστοιχίας Top-Down Concordance υπολογίζεται από τον τύπο :

$$\mathbf{C_\tau} = \frac{\sum_{t=1}^{n}(Ri^2 - M^2 * n)}{M^2(n - S1)}$$

όπου $\mathrm{R_i} = \sum_{j=1}^{m} S_{ij}$, και Sij είναι η βαθμολογία Savage για το αντικείμενο i τη στιγμή j για i = 1,….,n j = 1,…..,m Smax = max[i: 1≤i≤n]Sij είναι η υψηλότερη βαθμολογία Savage. Για ένα δείγμα n-διαστάσεων , εάν ο στόχος είναι να επισημανθεί η συμφωνία στο κατώτερο επίπεδο της κατάταξης , η βαθμολογία Savage θα είναι Sk= $\sum_{r=k}^{n}(\frac{1}{r})$, και ο συντελεστής αντιστοιχίας Top-down Concordance είναι $C_T^{(l)}$. Αν ο στόχος είναι να επισημανθεί η συμφωνία στο ανώτερο επίπεδο της κατάταξης, η βαθμολογία Savage θα είναι Sk= $\sum_{r=n-k+1}^{n}(\frac{1}{r})$, και ο συντελεστής αντιστοιχίας Top-down Concordance είναι $C_T^{(h)}$.Η Top-Down Concordance για ένα ζεύγος μεταβλητών είναι η τιμή της Pearson's R των βαθμολογιών Savage των μεταβλητών.

1.6 Διασταυρωμένη επαλήθευση(Cross-Validation)

Η διασταυρωμένη επαλήθευση, είναι ένα μοντέλο τεχνικής επαλήθευσης. Σε ένα πρόβλημα πρόβλεψης, δίνεται συνήθως ένα σύνολο γνωστών δεδομένων με το οποίο η λειτουργεί η εκπαίδευση (σύνολο δεδομένων εκπαίδευσης), καθώς και ένα σύνολο αγνώστων δεδομένων βάσει τα οποία ελέγχεται το μοντέλο. Ο στόχος της διασταυρωμένης επαλήθευσης είναι να καθορίσει ένα σύνολο δεδομένων για να «δοκιμάσει» το μοντέλο στη φάση της εκπαίδευσης (σύνολο δεδομένων επικύρωσης), προκειμένου να δώσει μια εικόνα για το πώς το μοντέλο θα γενικευθεί σε ένα ανεξάρτητο σύνολο δεδομένων (άγνωστο σύνολο δεδομένων). Ένας γύρος διασταυρωμένης επικύρωσης περιλαμβάνει τον διαχωρισμό ενός δείγματος δεδομένων σε συμπληρωματικά υποσύνολα, την εκτέλεση της διαδικασίας μοντελοποίησης σε ένα υποσύνολο (που ονομάζεται σύνολο εκπαίδευσης) και την επικύρωση της ανάλυσης από το άλλο υποσύνολο (σύνολο επικύρωσης). Για να μειωθεί η μεταβλητότητα, εκτελούνται πολλαπλοί γύροι διασταυρούμενης επαλήθευσης χρησιμοποιώντας διαφορετικά χωρίσματα και τα αποτελέσματα της επαλήθευσης συνυπολογίζονται.

Στην διασταυρωμένη επικύρωση k-υποσυνόλων, το αρχικό δείγμα χωρίζεται τυχαία σε k επιμέρους σύνολα ίσου μεγέθους. Από τα επιμέρους δείγματα k, ένα δείγμα διατηρείται ως δείγμα επικύρωσης για τη δοκιμή του μοντέλου, και τα υπόλοιπα k - 1 δείγματα χρησιμοποιούνται ως δεδομένα εκπαίδευσης. Η διαδικασία διασταυρούμενης επικύρωσης κατόπιν επαναλαμβάνεται k φορές, με κάθε ένα από τα επιμέρους δείγματα k χρησιμοποιείται ακριβώς μια φορά ως δεδομένο επικύρωσης. Τα αποτελέσματα από τα k-υποσύνολα μπορεί στη συνέχεια να χρησιμοποιηθούν συνολικά για την παραγωγή μιας ενιαίας εκτίμησης. Το πλεονέκτημα αυτής της μεθόδου είναι ότι όλες οι παρατηρήσεις χρησιμοποιούνται τόσο ως σύνολα εκπαίδευσης όσο και ως σύνολα επικύρωσης, και κάθε παρατήρηση χρησιμοποιείται σε σύνολο επικύρωσης ακριβώς μια φορά. Στη διασταυρωμένη επικύρωση k-υποσυνόλων, τα υποσύνολα επιλέγονται έτσι ώστε η μέση τιμή απόκρισης είναι περίπου ίση σε όλες τα υποσύνολα. Η μέθοδος αυτή χρησιμοποιήθηκε εκτενέστατα κατά την διεξαγωγή αυτής τις μελέτης καθώς απαιτεί προϋπόθεση για την δημιουργία των αρχείων εισόδου τόσο για την αναβαθμολόγηση όσο και για την μέθοδο συναινετικής βαθμολόγησης που χρησιμοποιήθηκε.

1.7 Ομοιότητα αποτυπώματος δύο διαστάσεων

Στην καρδιά του κάθε συστήματος εικονικής διαλογής που βασίζεται στην ομοιότητα είναι το μέτρο που χρησιμοποιείται για την ποσοτικοποίηση του βαθμού ομοιότητας μεταξύ των δομών αναφοράς και κάθε μιας από τις δομές της βάσης δεδομένων (πραγματικές ή εικονικές) που υποβάλλονται σε εικονική διαλογή. Ένα μέτρο ομοιότητας περιλαμβάνει τρεις συνιστώσες: Τον τρόπο αναπαράστασης των μορίων που συγκρίνονται, το σύστημα συντελεστών στάθμισης που χρησιμοποιείται για να αποδοθούν βαθμοί σπουδαιότητας στα συστατικά αυτών, και ο συντελεστής ο οποίος χρησιμοποιείται για να προσδιοριστεί ο βαθμός συγγένειας μεταξύ δύο δομικών αναπαραστάσεων. Μια ιδιαίτερη αναπαράσταση της δομής, είναι το αποτύπωμα δύο διαστάσεων.

Παρά το γεγονός ότι πολλά από αυτά έχουν σχεδιαστεί με συνεχείς, πραγματικές τιμές των δεδομένων μπορούν συχνά να εκφράζονται σε μια μορφή που τις καθιστά κατάλληλες για τον προσδιορισμό των ομοιοτήτων μεταξύ του ζεύγους αρχείων, όπως τα δισδιάστατα(2D) δακτυλικά αποτυπώματα [8] .

Η αναζήτηση ομοιότητας χρησιμοποιώντας δακτυλικά αποτυπώματα δύο διαστάσεων είναι ένα από τα απλούστερα εργαλεία εικονικής διαλογής και έτσι χρησιμοποιείται ευρέως στα πρώιμα στάδια των προγραμμάτων ανακάλυψης φαρμάκων, όταν υπάρχουν διαθέσιμα δομικά στοιχεία. Κύρια λειτουργία του είναι να εντοπίσει μερικά ενεργά συστατικά που μπορούν, στη συνέχεια, να αποτελέσουν τη βάση για πιο λεπτομερείς μελέτες εικονικής διαλογής που απασχολούν περισσότερο εξελιγμένες τεχνικές όπως είναι χαρτογράφηση, και η σύνδεση φαρμακοφόρου.

1.8 Δομική αλληλεπίδραση δακτυλικών αποτυπωμάτων

Η αναπαράσταση και η κατανόηση των τρισδιάστατων (3D) δομικών πληροφοριών ενός συμπλόκου πρωτεΐνης-προσδέτη είναι ένα κρίσιμο βήμα προς την ορθολογική διαδικασία ανακάλυψης φαρμάκων. Οι παραδοσιακές αναλυτικές μέθοδοι αποδεικνύονται ανεπαρκείς και αναποτελεσματικές στην αντιμετώπιση του μεγάλου ποσού των δομικών πληροφοριών που εξάγονται από κρυσταλλογραφία ακτινών X, NMR, και in silico προσεγγίσεις όπως τα πειράματα πρόσδεσης βασίζονται στη δομή . Μια μέθοδος για την αναπαράσταση και την ανάλυση τρισδιάστατων αλληλεπιδράσεων δέσμευσης πρωτεΐνης-προσδέτη είναι η δομική αλληλεπίδραση δακτυλικών αποτυπωμάτων(Structural Interaction Fingerprint,SIFt)[9]. Κλειδί για αυτή την προσέγγιση είναι η παραγωγή ενός αποτυπώματος αλληλεπίδρασης που μεταφράζει τις τρισδιάστατες δομικές πληροφορίες δέσμευσης από ένα σύμπλοκο πρωτεΐνης-προσδέτη σε μία μονοδιάστατη δυαδική αλυσίδα. Κάθε δακτυλικό αποτύπωμα αντιπροσωπεύει το «προφίλ» δομικής αλληλεπίδρασης του συγκροτήματος που μπορεί να χρησιμοποιηθεί για την οργάνωση, την ανάλυση, καθώς και να απεικονίσει την πλούσια ποσότητα των πληροφοριών που κωδικοποιούνται στα σύμπλοκα υποδοχέα-προσδέτη και επίσης για να βοηθήσει την εξόρυξη δεδομένων. Η μέθοδος SIFt είναι σε θέση να οργανώσει τις δομές και να αποκαλύψει σημαντικές ομοιότητες και ποικιλομορφίες μεταξύ των αλληλεπιδράσεων μικρών μορίων πρόσδεσης. Τέλος, η μέθοδος SiFt μπορεί να χρησιμοποιείται ως ένα αποτελεσματικό μοριακό φίλτρο κατά τη διάρκεια της διαδικασίας εικονικού ελέγχου χημικής βιβλιοθήκης για επιλογή μορίων με επιθυμητές δεσμευτικές λειτουργίες ή και επιθυμητά πρότυπα αλληλεπίδρασης με την πρωτεϊνικό στόχο.

1.9 Γενετικοί Αλγόριθμοι

Η μέθοδος που ακολουθήθηκε βασίζεται στην ανάπτυξη ενός γενετικού αλγορίθμου. Οι Γενετικοί αλγόριθμοι ανήκουν στο κλάδο της επιστήμης υπολογιστών και αποτελούν μια μέθοδο αναζήτησης βέλτιστων λύσεων σε συστήματα που μπορούν να περιγραφούν ως μαθηματικό πρόβλημα. Είναι μια τεχνική προγραμματισμού που εισήγαγε στα τέλη της δεκαετίας του 1960 ο Τζον Χόλαντ, ερευνητής του Ινστιτούτου της Σάντα Φε (ΗΠΑ). Είναι χρήσιμοι σε προβλήματα που περιέχουν πολλές παραμέτρους/διαστάσεις και δεν υπάρχει αναλυτική μέθοδος που να μπορεί να βρει το βέλτιστο συνδυασμό τιμών για τις μεταβλητές ώστε το υπό εξέταση σύστημα να αντιδρά με όσο το δυνατόν με το επιθυμητό τρόπο.

Ο τρόπος λειτουργίας των Γενετικών Αλγορίθμων είναι εμπνευσμένος από τη βιολογία. Χρησιμοποιεί την ιδέα της εξέλιξης μέσω γενετικής μετάλλαξης, φυσικής επιλογής και διασταύρωσης. Οι τιμές για τις παραμέτρους του συστήματος πρέπει να κωδικοποιούνται με τρόπο ώστε να αναπαρασταθούν από μια μεταβλητή που περιέχει σειρά χαρακτήρων ή δυαδικών ψηφίων (0/1). Αυτή η μεταβλητή μιμείται το γενετικό κώδικα που υπάρχει στους ζωντανούς οργανισμούς. Αρχικά, ο Γενετικός Αλγόριθμος παράγει πολλαπλά αντίγραφα της μεταβλητής/γεννητικού κώδικα, συνήθως με τυχαίες τιμές, δημιουργώντας ένα πληθυσμό λύσεων. Κάθε λύση (τιμές για τις παραμέτρους του συστήματος) δοκιμάζεται για το πόσο κοντά φέρνει την αντίδραση του συστήματος στην επιθυμητή, μέσω μιας συνάρτησης που δίνει το μέτρο ικανότητας της λύσης και η οποία ονομάζεται συνάρτηση ικανότητας (Σ.Ι).

Οι πιο ικανές λύσεις για ένα συγκεκριμένο πρόβλημα συνεχίζουν να εξελίσσονται και ανασυνδυάζονται τυχαία, μέχρις ότου "επιβιώσουν" οι καλύτερες. Συνήθως, όσο περισσότερες γενιές περνούν τόσο καλύτερες λύσεις βρίσκονται, μπορεί όμως ο αλγόριθμος να βρεθεί σε σημείο του πεδίου των λύσεων από όπου και δεν μπορεί να προχωρήσει λόγο του ότι βρίσκεται σε τοπικό μέγιστο. Για το λόγο αυτό έχουν υπάρχουν διαφορετικές εκδοχές του αλγόριθμου ανάλογα με τη μορφή του προβλήματος.

Περιλαμβάνουν τη διασταύρωση (ζευγάρωμα) γονιδίων/λύσεων ώστε ο αλγόριθμος να φτάσει στο αποτέλεσμα πιο γρήγορα.

Κεφάλαιο 2

2 Μέθοδος Βελτιστοποίησης

Σε αυτή τη εργασία παρουσιάζουμε μια μεθοδολογία συναινετικής βαθμολόγησης που βασίζεται στο σχεδιασμό ενός γενετικού αλγορίθμου ο οποίος βρίσκει βέλτιστους γραμμικούς συνδυασμούς ΣΒ . Αρχικά χρησιμοποιήθηκαν δύο ΣΒ, Glide και Vina, οι οποίες βαθμολόγησαν τις πόζες σύνδεσης σε 8 διαφορετικούς υποδοχείς. Οι πόζες σύνδεσης των δύο ΣΒ, αναβαθμολογήθηκαν με 45 επιπλέων ΣΒ ώστε να βρεθούν οι καλύτερες ΣΒ για κάθε υποδοχέα. Η διαδικασία αυτή έγινε τόσο για εύκαμπτους όσο και για άκαμπτους υποδοχείς. Σε κάθε περίπτωση, χρησιμοποιήθηκαν 46 ΣΒ από τον προτεινόμενο γενετικό αλγόριθμο ώστε να βρεθεί ο βέλτιστος γραμμικός συνδυασμός ΣΒ για κάθε υποδοχέα. Τα αποτελέσματα τις μεθόδου επεξεργάστηκαν ώστε να αξιολογηθεί η αποτελεσματικότητά της και παρουσιάζονται στο Κεφάλαιο 3.

Ως σύνολο εκπαίδευσης για τη βελτιστοποίηση της συναινετικής βαθμολόγησης έχουμε χρησιμοποιήσει το σύνολο μη-δραστικών ενώσεων DUD-E *[4]* από οκτώ διαφορετικούς υποδοχείς [Mysinger2012]. Αυτό το σύνολο δεδομένων αποτελείται από ενώσεις και δεσμευτικά στοιχεία από το ChEMBL[10] .Για να χρησιμοποιηθεί ένα «τυφλό» σύνολο δοκιμών, δημιουργήθηκε ένα σύνολο δεδομένων χρησιμοποιώντας αρχεία ChEMBL , που δεν περιλαμβάνονται στο DUD-E, και επεξεργάστηκαν τα αντίστοιχα μόρια με τον ίδιο τρόπο όπως και με εκείνα του συνόλου DUD-E.

Η βάση DUD, είναι ένας κατάλογος με χρήσιμες μη-δραστικές ενώσεις για τη συγκριτική αξιολόγηση της εικονικής διαλογής .Η βάση DUD είναι σχεδιασμένη ώστε να βοηθάει στην δοκιμή αλγορίθμων με την παροχή των μη δραστικών ενώσεων(decoys).Ο κατάλογος περιέχει ένα σύνολο 2.950 δραστικών ενώσεων έναντι συνολικά 40 στόχων. Για κάθε δραστική ένωση υπάρχουν 36 μη δραστικές με παρόμοιες φυσικές ιδιότητες (μοριακό βάρος ,LogP) αλλά ανόμοια τοπολογία. Μια πιο πλούσια έκδοση της βάσης DUD, η DUD-E έχει σχεδιαστεί για να συμβάλει στη συγκριτική αξιολόγηση των προγραμμάτων μοριακής σύνδεσης. Ο κατάλογος περιέχει ένα σύνολο 22886 δραστικών ενώσεων καθώς και τις συγγένειες πρόσδεσης τους έναντι 102 στόχων και κατά μέσο όρο 224 προσδέτες ανά στόχο 50 μη-δραστικών ενώσεις για κάθε δραστική οι οποίες έχουν παρόμοιες φυσικοχημικές ιδιότητες, αλλά ανόμοια διστιάστατη(2D) τοπολογία.

Η ChEMBL είναι μία βάση δεδομένων που περιέχει δεδομένα δέσμευσης, λειτουργικές και ADMET πληροφορίες για ένα μεγάλο αριθμό βιοδραστικών ενώσεων. Το ADMET είναι μια συντομογραφία που χρησιμοποιείτε στη φαρμακοκινητική και στη φαρμακολογία για την απορρόφηση, κατανομή, τον μεταβολισμό και την απέκκριση, και περιγράφει τη διάθεση μιας φαρμακευτικής ένωσης εντός ενός οργανισμού . Τα τέσσερα κριτήρια επηρεάζουν τα επίπεδα του φαρμάκου και την κινητική της έκθεσης του φάρμακο στους ιστούς επηρεάζοντας την απόδοση και την φαρμακολογική δραστικότητα της ένωσης ως φάρμακο.

Επί του παρόντος, η βάση δεδομένων ChEMBL περιέχει 5,4 εκατομμύρια μετρήσεις βιοδραστικότητας για περισσότερες από 1 εκατομμύριο ενώσεις και 5200 πρωτεϊνικών στόχων.

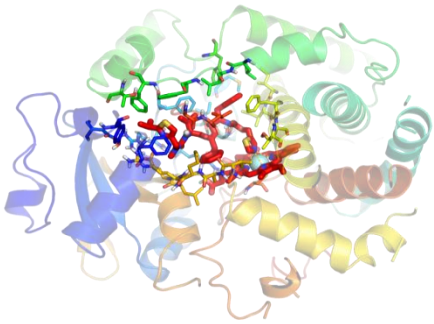
2.1 Υποδοχείς DUD-E

Οι οκτώ υποδοχείς που χρησιμοποιήθηκαν από την βάση DUD-E είναι οι εξής :

Κυτόχρωμα P450 3A4(CYP3A4 ή CP3A4)

Το κυτόχρωμα P450 3A4 (CYP3A4) , είναι ένα σημαντικό ένζυμο στο σώμα οποίο βρίσκεται κυρίως στο ήπαρ και στο έντερο. Σκοπός του είναι να οξειδώνει μικρά ξένα οργανικά μόρια (ξενοβιοτικά), όπως τοξίνες ή φάρμακα, ούτως ώστε να μπορούν να αφαιρεθούν από το σώμα.

Ενώ πολλά φάρμακα απενεργοποιούνται από το CYP3A4, υπάρχουν ορισμένα φάρμακα τα οποία ενεργοποιούνται από το ένζυμο. Ορισμένες ουσίες, όπως ο χυμός γκρέιπφρουτ και ορισμένα φάρμακα, παρεμποδίζουν τη δράση του CYP3A4. Αυτές οι ουσίες συνεπώς ενισχύουν ή εξασθενούν τη δράση αυτών των φαρμάκων που έχουν τροποποιηθεί από το CYP3A4.

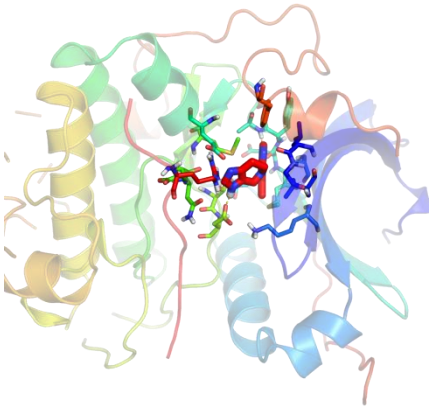


Εικόνα 3: Cytochrome P450 3A4 (CP3A4)

Το CYP3A4 είναι ένα μέλος της οικογένειας P450 κυτοχρώματος οξειδωτικών ενζύμων. Αρκετά άλλα μέλη αυτής της οικογένειας εμπλέκονται επίσης στο μεταβολισμό των φαρμάκων, αλλά το CYP3A4 είναι το πιο κοινό και το πιο ευέλικτο . Όπως όλα τα μέλη αυτής της οικογένειας ,το CYP3A4 είναι μια αιμοπρωτεΐνη, δηλαδή μια πρωτεΐνη που περιέχει μια ομάδα αίμης με ένα άτομο σιδήρου. Στον άνθρωπο, η πρωτεΐνη CYP3A4 κωδικοποιείται από το γονίδιο CYP3A4. Αυτό το γονίδιο είναι μέρος ενός συμπλέγματος των γονιδίων του κυτοχρώματος P450 στο χρωμόσωμα 7q21.1.

Κινάση σερίνης-θρεονίνης(AKT1)

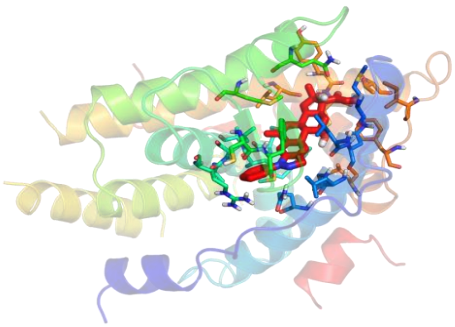
Η RAC-άλφα κινάση σερίνης-θρεονίνης είναι ένα ένζυμο που κωδικοποιείται στον άνθρωπο από το γονίδιο AKT1. Αυτό το ένζυμο ανήκει στην υποοικογένεια της AKT κινασών σερίνης-θρεονίνης .Η AKT1 είναι μία από τις τρείς κινάσες σερίνης / θρεονίνης πρωτεΐνης (AKT1, AKT2 και AKT3) οι οποίες ρυθμίζουν πολλές διαδικασίες όπως ο μεταβολισμός, ο πολλαπλασιασμός, η κυτταρική επιβίωση, η ανάπτυξη και η αγγειογένεση. Αυτό επιτυγχάνεται με τη φωσφορυλίωση της σερίνης ή και της θρεονίνης ενός φάσματος υποστρωμάτων. Πάνω από 100 υποψήφια υποστρώματα έχουν αναφερθεί μέχρι τώρα, αλλά για τα περισσότερα από αυτά, δεν έχει αναφερθεί κανένα ισόμορφο χαρακτηριστικό.



Εικόνα 4: Ser/Thr-protein kinase AKT (AKT1)

Υποδοχέας γλυκοκορτικοειδών(GCR)

Ο υποδοχέας γλυκοκορτικοειδών (GR, ή GCR), επίσης γνωστή ως NR3C1 (υποοικογένεια πυρηνικού υποδοχέα 3 , ομάδα Γ , μέλος 1) είναι ο υποδοχέας στον οποίο προσδένονται η κορτιζόλη και άλλα γλυκοκορτικοειδή. Ο GR εκφράζεται σχεδόν σε κάθε κύτταρο του σώματος και ρυθμίζει γονίδια που ελέγχουν την ανάπτυξη , το μεταβολισμό, και την ανοσοαπόκριση. Επειδή το γονίδιο υποδοχέας εκφράζεται σε διάφορες μορφές , έχει πολλά διαφορετικά (πλειοτροπικές) αποτελέσματα σε διάφορα μέρη του σώματος.



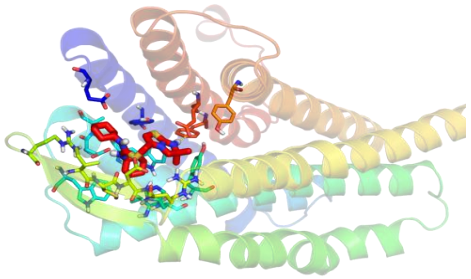
Εικόνα 5: Glucocorticoid receptor (GCR)

Όταν το GR συνδέεται με γλυκοκορτικοειδή, πρωταρχικός μηχανισμός δράσης της είναι η ρύθμιση της γονιδιακής μεταγραφής. Ο μη δεσμευμένος υποδοχέας βρίσκεται στο κυτοσόλιο του κυττάρου. Αφού ο υποδοχέας συνδέεται με γλυκοκορτικοειδές, το σύμπλοκο υποδοχέα γλυκοκορτικοειδές μπορεί να λάβει οποιαδήποτε από τις δύο διαδρομές . Το ενεργοποιημένο σύμπλοκο GR ρυθμίζει την έκφραση των αντιφλεγμονωδών πρωτεϊνών

στον πυρήνα ή καταστέλλει την έκφραση των προ-φλεγμονωδών πρωτεϊνών στο κυτοσόλιο (εμποδίζοντας την μετατόπιση των άλλων παραγόντων μεταγραφής από το κυτοσόλιο στον πυρήνα).Στον άνθρωπο , η πρωτεΐνη GR κωδικοποιείται από NR3C1 γονίδιο που βρίσκεται στο χρωμόσωμα 5 (5q31) .

Υποδοχέας χημειοκινών(CXCR4)

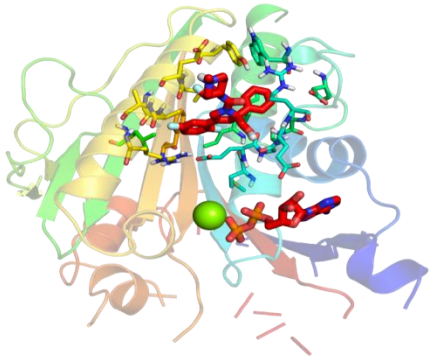
Οι υποδοχείς χημειοκινών CXC είναι ενσωματωμένες μεμβρανικές πρωτεΐνες που δεσμεύονται ειδικά και ανταποκρίνονται στις κυτοκίνες της οικογένειας χημειοκίνης CXC. Αντιπροσωπεύουν μία υποοικογένεια των υποδοχέων χημειοκινών, μια μεγάλη οικογένεια των G πρωτεΐνων που συνδέεται υποδοχείς που είναι γνωστές ως επτά διαμεμβρανικές (7-TM) πρωτεΐνες, δεδομένου ότι καλύπτουν την κυτταρική μεμβράνη επτά φορές. Υπάρχουν επί του παρόντος επτά γνωστοί CXC υποδοχείς χημειοκινών στα θηλαστικά, που ονομάζονται CXCR1 έως CXCR7.
Ο CXCR4 (επίσης γνωστή ως fusin) είναι ο υποδοχέας για ένα χημειοκινικό γνωστό ως CXCL12 (ή SDF-1) και, όπως και με το CCR5, χρησιμοποιείται από τον ιό HIV-1 για να εισβάλει στα κύτταρα-στόχους. Αυτός ο υποδοχέας έχει μία ευρεία κυτταρική κατανομή, με την έκφραση στους αιμοποιητικούς τύπους κυττάρων (π.χ. ουδετερόφιλα, μονοκύτταρα, Τ και Β κύτταρα, δενδριτικά κύτταρα, κύτταρα Langerhans και μακροφάγα). Επιπλέον, CXCR4 μπορεί επίσης να βρεθεί σε αγγειακά ενδοθηλιακά κύτταρα και νευρωνικά κύτταρα.



Εικόνα 6: C-X-C chemokine receptor type 4 (CXCR4)

Κινεσίνη-5(KIF11)

Η κινεσίνη-5 είναι μία πρωτεΐνη μοριακού κινητήρα που είναι απαραίτητη στη μίτωση. Οι κινεσίνες-5 είναι μέλη της υπεροικογένειας κινεσίνης, νανοκινητήρων που κινούνται κατά μήκος των διαδρομών των μικροσωληνίσκων στο κύτταρο. Ονομάστηκε από μελέτες κατά τις πρώτες ημέρες της ανακάλυψης, είναι επίσης γνωστό ως μέλος της οικογένειας κινεσίνης 11, που κωδικοποιείται από το γονίδιο KIF11. Ο όρος κινεσίνη-5 βασίζεται σε τυποποιημένη ονοματολογία που έχει εγκριθεί από την επιστημονική κοινότητα.

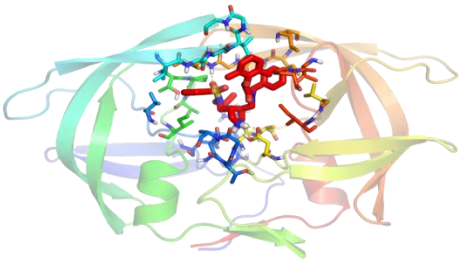


Εικόνα 7: Kinesin-like protein 1 (KIF11)

Επί του παρόντος, υπάρχουν πάνω από 70 διαφορετικές ευκαρυωτικές κινεσίνες-5 που προσδιορίζονται από την ομοιότητα αλληλουχίας. Μέλη αυτής της οικογένειας πρωτεϊνών είναι γνωστό ότι εμπλέκονται σε διάφορα είδη της δυναμικής ατράκτου και είναι ουσιώδη για μίτωση. Η λειτουργία αυτού του γονιδίου περιλαμβάνει τοποθέτηση χρωμοσωμάτων, κεντροσωματίων και το διαχωρισμό και τη σύσταση διπολικής ατράκτου κατά τη διάρκεια της κυτταρικής μίτωσης. Η ανθρώπινη πρωτεΐνη κινεσίνης 5 έχει μελετηθεί ενεργά για το ρόλο της στη μίτωση και το δυναμικό της ως θεραπευτικός στόχος για τη θεραπεία του καρκίνου.

Πρωτεάση HIV-1(HIVPR)

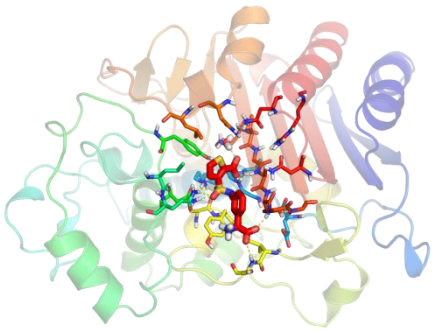
Η πρωτεάση HIV-1 είναι ένας ρετροϊικός που είναι απαραίτητος για τον κύκλο ζωής του HIV, του ρετροϊού που προκαλεί το AIDS. Η πρωτεάση του HIV διασπά νεοσυντιθέμενες πολυπρωτεΐνες στα κατάλληλα μέρη για να δημιουργήσει στην πρωτεΐνη συνιστώσες ενός μολυσματικού βιριόντος HIV. Έτσι, η μετάλλαξη της δραστική θέσης της πρωτεάσης του HIV ή η αναστολή της δραστηριότητας της διαταράσσει την ικανότητα του HIV να αναπαραχθεί και να μολύνει επιπλέον κύτταρα, , καθιστώντας την αναστολή της πρωτεάσης του HIV αντικείμενο σημαντικής φαρμακευτικής έρευνας. Η πρωτεϊνική δομή της πρωτεάσης του HIV έχει διερευνηθεί χρησιμοποιώντας κρυσταλλογραφία ακτινών X. Υπάρχει ως ένα ομοδιμερές, με κάθε υπομονάδα αποτελείται από 99 αμινοξέα. Η ενεργή θέση βρίσκεται μεταξύ των ταυτόσημες υπομονάδες και έχει το χαρακτηριστική Asp-Thr-Gly (Asp25, Thr26 και Gly27) αλληλουχία κοινή σε ασπαρτικές πρωτεάσες. Τα δύο υπολείμματα Asp25 (ένα από κάθε αλυσίδα) ενεργούν ως καταλυτικά υπολείμματα. Με πρωταγωνιστικό ρόλο στην αντιγραφή του HIV, η πρωτεάση του HIV ήταν ένας πρωταρχικός στόχος στην φαρμακευτική θεραπεία. Οι αναστολείς της πρωτεάσης του HIV λειτουργούν με ειδική δέσμευση στο ενεργό κέντρο που μιμούνται το τετραεδρικό ενδιάμεσο του υποστρώματος του και ουσιαστικά "κολλάνε" απενεργοποιώντας το ένζυμο. Μετά τη σύνδεση , ικά σωματίδια τα οποία στερούνται ενεργής πρωτεάσης και δεν μπορούν να ωριμάσουν σε μολυσματικά ιοσωμάτια. Αρκετοί αναστολείς πρωτεάσης έχουν εγκριθεί για τη θεραπεία του HIV.



Εικόνα 8: HIV type 1 protease (HIVPR)

Β'-λακταμάση(AMPC)

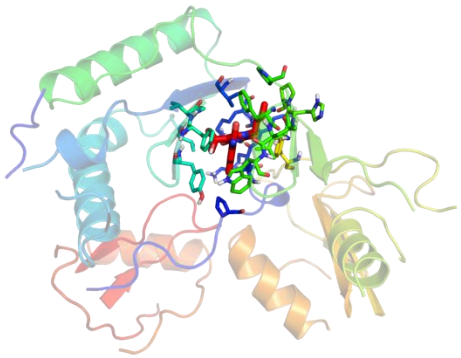
Οι Βήτα-λακταμάσες είναι ένζυμα που παράγονται από ορισμένα βακτήρια που προβάλουν αντίσταση σε αντιβιοτικά β-λακτάμης, όπως πενικιλλίνες, κεφαμυκίνες και καρβαπενέμες, αν και καρβαπενέμες είναι σχετικά ανθεκτικά στην βήτα-λακταμάση.Η Βήτα-λακταμάση παρέχει ανθεκτικότητα στα αντιβιοτικά καθώς επεμβαίνει στη δομή των αντιβιοτικών . Αυτά τα αντιβιοτικά έχουν ένα κοινό στοιχείο στη μοριακή τους δομή, ένα δαχτυλίδι τεσσάρων άτομο γνωστό ως β-λακτάμη. Μέσω υδρόλυσης, το ένζυμο βήτα-λακταμάσης διασπά τον ανοικτό δακτύλιο της β-λακτάμης, απενεργοποιώντας τις αντιβακτηριακές ιδιότητες του μορίου. Τα αντιβιοτικά β-λακτάμης χρησιμοποιούνται συνήθως για τη θεραπεία ένος ευρέως φάσματος βακτηρίων.



Εικόνα 9: Beta-lactamase (AMPC)

Αντίστροφη μεταγραφάση HIV(HIVRT)

Η αντίστροφη μεταγραφάση είναι ένα ιογενής κωδικοποιημένο ένζυμο που μετατρέπει το ιικό γονιδίωμα μονόκλωνου RNA σε δίκλωνο DNA προϊού που είναι ενσωματωμένος στο χρωμόσωμα του ξενιστή στον πυρήνα του κυττάρου ξενιστή. Η διαδικασία της μετατροπής ιικών ssRNA σε dsDNA που μπορούν να ενσωματωθούν στο χρωμόσωμα του ξενιστή ονομάζεται ρετρομεταγραφή, και είναι χαρακτηριστικό όλων των ρετροϊών. HIV-1 ανάστροφη μεταγραφάση κωδικοποιείται από τον ιό ανθρώπινης ανοσοανεπάρκειας, γνωστή ως αιτιολογικός παράγοντας του συνδρόμου επίκτητης ανοσολογικής ανεπάρκειας (AIDS).Η HIV-1 είναι χρόνια και απαιτεί ισόβια θεραπεία με έναν συνδυασμό από τουλάχιστον τρία διαφορετικά αντι-ιικά φάρμακα.



Εικόνα 10: HIV type 1 reverse transcriptase (HIVRT)

2.2 Λογισμικό Συναρτήσεων Βαθμολόγησης

Οι 8 πρωτεϊνικοί υποδοχείς διαφορετικών υποσυνόλων του DUD-E*[4]* χρησιμοποιήθηκαν μέσω της ιστοσελίδας του DUD-E(http://dude.docking.org/subsets/diverse). Ο υπολογισμός μοριακής πρόσδεσης μέσω της ΣΒ Vina, AutoDockTools 1.5.6 χρησιμοποιήθηκε για τη συγχώνευση μη-πολικών ατόμων υδρογόνου με τα μητρικά τους άτομα, και να αναθέσει τους τύπους των ατόμων και φορτία Gasteiger. Οι καταστάσεις πρωτονίωσης των υποδοχέων δεν μεταβλήθηκαν καθώς είχαν ήδη βελτιστοποιηθεί. Επίσης διατηρούνται τα κρυσταλλικά νερά και οι συν-παράγοντες. Ο αιμικός σίδηρος του υποδοχέα CP3A4 ορίστηκε στην κατάσταση τρισθενούς σιδήρου (Fe3 +), όπως στην κρυσταλλική δομή του υποδοχέα . Οι ενώσεις από τα σύνολα DUD-E χρησιμοποιήθηκαν χωρίς να αλλοιώνεται η δομή τους. Πρώτα μετατράπηκαν σε αρχεία της μορφής .mol2 χρησιμοποιώντας fconv [12] και, στη συνέχεια, επεξεργάστηκαν με τα AutoDockTools. Η απόδοση των άκαμπτων και εύκαμπτων προσεγγίσεων σύνδεσης στο πλαίσιο συναινετικής βαθμολόγησης χρησιμοποιώντας δύο δημοφιλή προγράμματα υπολογισμού μοριακής σύνδεσης, τα Vina και Glide.

Οι σταθεροί ή εύκαμπτοι υποδοχείς επεξεργάστηκαν χρησιμοποιώντας το AutoDock Vina[13] μέσω του PyMOL. Στην τελευταία περίπτωση, η επιλεκτικές πλευρικές αλυσίδες που ήταν μέσα σε 5 Å από τον συνδέτη παρέμειναν ευέλικτες. Οι ίδιες δομές υποδοχέα και προσδέτη σε μορφή .sdf χρησιμοποιήθηκαν για την προσομοίωση μέσω Glide [1] κατά την ανάθεση φορτίων ανάλογα σύμφωνα με το OPLS2005 (Banks et al., 2005). Οι διαστάσεις του πλαισίου για την προσάρτηση τόσο του Vina όσο και του Glide ρυθμίστηκαν χειροκίνητα, έτσι ώστε να περιέχουν όλους τους συν-παράγοντες, καθώς και αρκετό χώρο για τις ενώσεις που πρόκειται να συνδεθούν. Επιπλέον, οι καλύτερες 10 πόζες Vina και Glide αναβαθμολογηθήκαν χρησιμοποιώντας NNScore 1.0 και 2.0[3], και DSX [2]. Τα δύο πρώτα περιλαμβάνουν 24 και 20 ατομικές ΣΒ βασισμένες σε νευρωνικά δίκτυα αντίστοιχα, ενώ η τρίτη αποτελείται από μία ΣΒ στατιστικού δυναμικού .Για την ανάλυση χρησιμοποιήθηκαν μόνο οι ενώσεις που συνδέθηκαν (τόσο από Vina όσο και από το Glide) και αναβαθμολογήθηκαν επιτυχώς από όλες τις ΣΒ. Η προαναφερόμενη διαδικασία διεξάγεται σε πολλαπλές CPUs με τη χρήση του εργαλείου RescoringTK.py. Το εργαλείο RescoringTK.py μπορεί να διαβάζει και να επεξεργάζεται τα αποτελέσματα σύνδεσης των Vina και Glide.

2.2.1 Vina

Η γενική λειτουργική μορφή της εξαρτώμενης από τη διαμόρφωση μέρος της λειτουργίας βαθμολόγησης Autodock Vina (αναφέρεται ως Vina)[13] έχει σχεδιαστεί για να λειτουργεί με βάση το άθροισμα

$$c=\sum_{i<j} f_{t_i t_j}(r_{ij})$$

όπου το άθροισμα είναι από όλα τα ζεύγη των ατόμων που μπορούν να κινούνται σε σχέση με το άλλο, με εξαίρεση 1-4 αλληλεπιδράσεις, δηλαδή άτομα που χωρίζονται από 3 διαδοχικούς ομοιοπολικούς δεσμούς. Εδώ, κάθε άτομο i αντιστοιχεί σε ένα τύπο t_i, και θα πρέπει να οριστεί ένα συμμετρικό σύνολο συναρτήσεων αλληλεπίδρασης $f_{t_i t_j}$ της ενδοατομικής απόστασης r_{ij} .Η τιμή αυτή μπορεί να θεωρηθεί ως ένα άθροισμα και ενδομοριακών και διαμοριακών επιδράσεων:

$$c=c_{inter}+c_{intra}$$

Ο αλγόριθμος βελτιστοποίησης του Vina, επιχειρεί να βρει το ολικό ελάχιστο του c και άλλα χαμηλής βαθμολόγησης διαμορφώσεις, τις οποίες και κατατάσσει.

Η προβλεπόμενη ελεύθερη ενέργεια της πρόσδεσης υπολογίζεται από το ενδομοριακό τμήμα της χαμηλότερα βαθμολογημένης διάπλασης οποία ορίζεται ως 1:

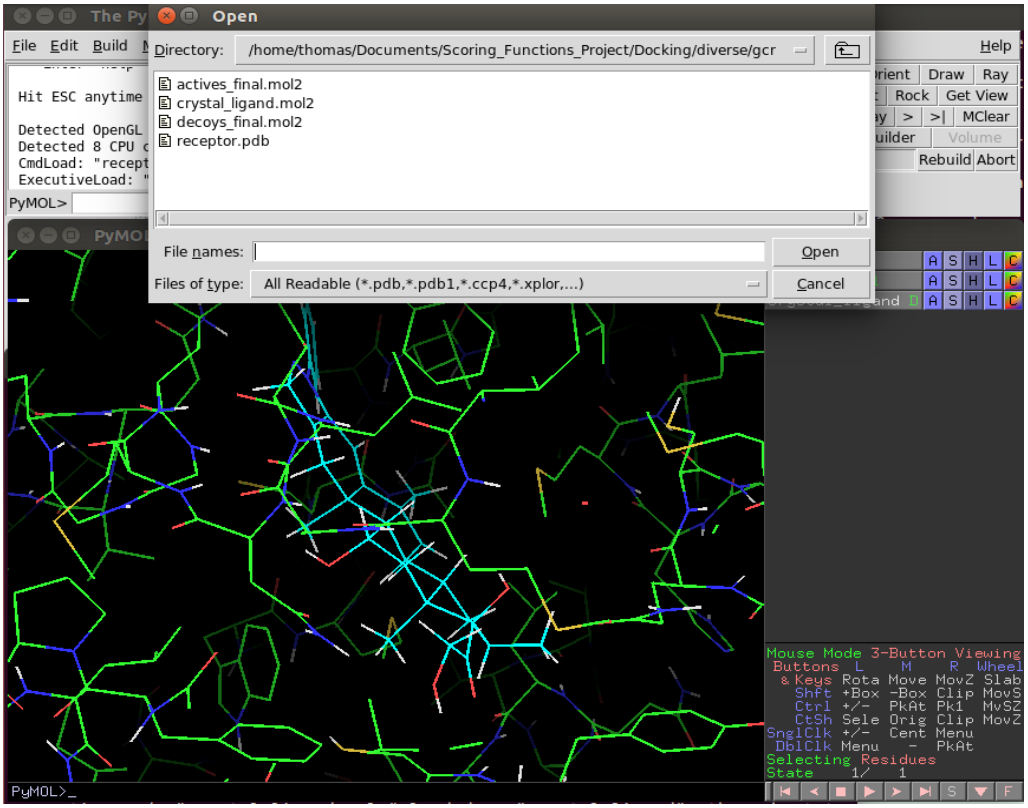
$$s_1=g(c_1-c_{intra1})=g(c_{inter1}),$$

όπου η συνάρτηση g μπορεί να είναι αυθαίρετη γνησίως αύξουσα ομαλή πιθανώς μη γραμμική.

Στην έξοδο, άλλες διαμορφώσεις χαμηλής βαθμολόγησης δίνονται επίσης επίσημα s αξίες, αλλά, για να διατηρηθεί η κατάταξη, χρησιμοποιώντας Cintra του καλύτερου τρόπου δέσμευσης:

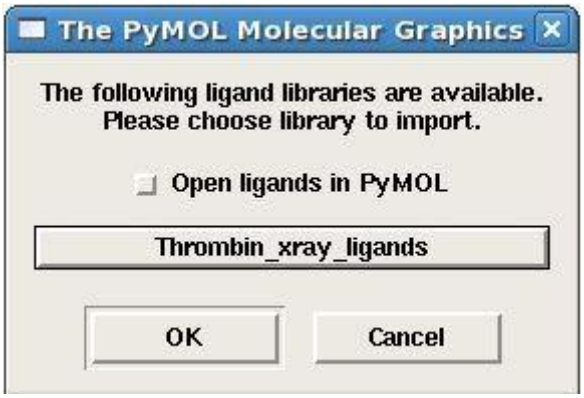
$$s_i=g(c_i-c_{intra1}).$$

Η συνάρτηση βαθμολόγησης Vina συνδυάζει ορισμένα πλεονεκτήματα των δυνατοτήτων των βασισμένων στη γνώση και των εμπειρικές συναρτήσεων βαθμολόγησης. Εξάγει εμπειρικές πληροφορίες τόσο από τις διαμορφωτικές προτιμήσεις των συμπλεγμάτων υποδοχέα-προσδέτη και τις πειραματικές μετρήσεις συγγένειας. Η συνάρτηση βαθμολόγησης Vina έχει σχεδιαστεί για να είναι συμβατή με τη μορφή αρχείου : PDBQT, η οποία μπορεί να θεωρηθεί ως επέκταση της μορφής αρχείου PDB. Αυτό το καθιστά εύκολο στη χρήση Vina με το υπάρχων βοηθητικό λογισμικό που αναπτύχθηκε , για την προετοιμασία των αρχείων, η επιλογή του χώρου αναζήτησης και την προβολή των αποτελεσμάτων. Επιπλέον, η χειροκίνητη επιλογή των τύπων ατόμων για τους χάρτες ο δικτύου, ο υπολογισμός των αρχείων πλέγματος με την βοήθεια του AutoGrid καθώς και η επιλογή «παραμέτρων αναζήτησης» και ομαδοποίησης των αποτελεσμάτων μετά τη σύνδεση δεν είναι απαραίτητη, καθώς το Vina υπολογίζει τους δικούς της χάρτες δικτύου γρήγορα και αυτόματα. Επίσης, κατηγοριοποιεί και κατατάσσει τα αποτελέσματα χωρίς την έκθεση του χρήστη σε ενδιάμεσες λεπτομέρειες. Η ΣΒ Vina χρησιμοποιήθηκε μέσω της πλατφόρμας Pymol . Το AutoDock Vina είναι ένα πρόσθετο πρόγραμμα και περιέχει την ΣΒ Vina που χρησιμοποιήθηκε. Το AutoDock διαβάζει τα αρχεία δομής υποδοχέα και προσδέτη και τα παρουσιάζει στην πλατφόρμα.



Εικόνα 11 : Pymol plugin AutoDock Vina

Το AutoDock απαιτεί ένα αρχείο **AUTODOCK_DICTIONARY** στο οποίο υπάρχει ο αριθμός των προσδετών καθώς και το όνομα της βιβλιοθήκης προσδετών που πρόκειται να χρησιμοποιηθεί .



Εικόνα 12 : AutoDock Vina Library

Στην συνέχεια ρυθμίζεται το μέγεθος και το κέντρο του πλαισίου χειροκίνητα για να καλύψει τον προσδέτη καθώς και λίγο ακόμα χώρο. Διευκρινίζεται επίσης αν ο υποδοχέας θα είναι σταθερός κατά τον υπολογισμό μοριακής πρόσδεσης η όχι . Τέλος καθορίζεται ο μέγιστος αριθμός των προβλεπόμενων δεσμευτικών ποζών(10 στην περίπτωσή μας).

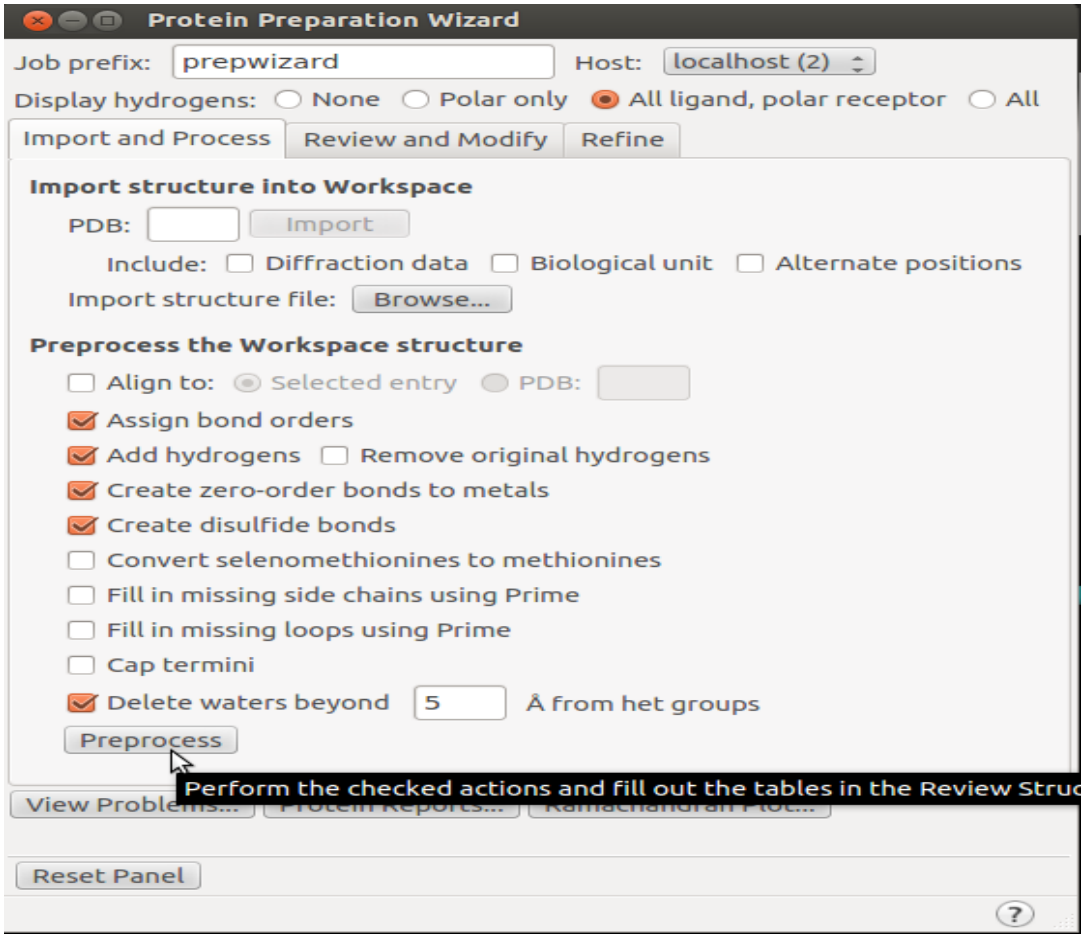
2.2.2 Glide

Η ΣΒ του Glide[1] χρησιμοποιεί μια σειρά από ιεραρχικά φίλτρα για να αναζητήσει πιθανές θέσεις του συνδέτη στην δραστική περιοχή του υποδοχέα . Το σχήμα και οι ιδιότητες του υποδοχέα παρουσιάζονται σαν ένα πλέγμα από διαφορετικά σύνολα των πεδίων που παρέχουν προοδευτικά πιο ακριβή αποτελέσματα της θέσης του προσδέτη (πόζα) . (Με τον όρο "πόζα" εννοούμε ένα ολοκληρωμένο προσδιορισμό του προσδέτη: θέση και προσανατολισμός σε σχέση με τον υποδοχέα, διάπλαση πυρήνα , διαμορφώσεις τις ροτομερούς ομάδας). Τα πεδία αυτά δημιουργούνται ως προκαταρκτικά βήματα στον υπολογισμό και χρειάζονται να υπολογιστούν μόνο μία φορά για κάθε υποδοχέα. Το επόμενο βήμα παράγει ένα σύνολο αρχικών διαμορφώσεων του προσδέτη. Αυτές οι διαμορφώσεις που επιλέγονται από μια εξαντλητική απαρίθμηση των ελαχίστων στον χώρο στρέψης του προσδέτη και παρουσιάζονται σε μια συμπαγή συνδυαστική μορφή. Λαμβάνοντας υπόψη τις διαμορφώσεις του προσδέτη, η διαδικασία πραγματοποιείται σε όλο το χώρο που έχει στην διάθεση του ο προσδέτης ώστε να εντοπιστούν καλύτερες πόζες για αυτόν .Αυτή η προκαταρτική διαλογή μειώνει δραστικά την περιοχή στον χώρο κίνησης του προσδέτη για την οποία θα λάμβαναν χώρα υπολογιστικά δαπανηρές ενεργειακές αξιολογήσεις ενώ την ίδια στιγμή αποφεύγεται η χρήση των στοχαστικών μεθόδων .

Ξεκινώντας από τις πόζες που επιλέγονται από την αρχική διαλογή, ο προσδέτης ελαχιστοποιείται στο πεδίο του υποδοχέα χρησιμοποιώντας ένα πρότυπο λειτουργίας της ενεργειακής μοριακής μηχανικής (σε αυτή την περίπτωση, αυτή του OPLS-AA) σε συνδυασμό με μια απόσταση που εξαρτάται από διηλεκτρικό μοντέλο. Τέλος, τρείς με έξι χαμηλότερες σε κατανάλωση ενέργειας πόζες που λαμβάνονται με αυτόν τον τρόπο υποβάλλονται σε μία διαδικασία Monte Carlo που εξετάζει κοντινά στρεπτικά ελάχιστα. Η διαδικασία αυτή είναι απαραίτητη σε ορισμένες περιπτώσεις ώστε να προσανατολίζει σωστά περιφερειακές ομάδες και περιστασιακά μεταβάλλει την εσωτερική γωνία στρέψης .

Η διαδικασία δημιουργίας των αρχικών ποζών προσδετών για την ΣΒ Glide έγινε μέσω του προγράμματος Maestro(λογισμικό της Schrodinger).

Αρχικά είναι απαραίτητα τα αρχεία στα οποία υπάρχουν οι δομές του υποδοχέα και του προσδέτη . Αυτά τα αρχεία είναι συνήθως σε receptor.mae ή μορφή ligand.mae. Αφού εισαχθούν τα αρχεία στην πλατφόρμα γίνεται προετοιμασία της πρωτεΐνης.

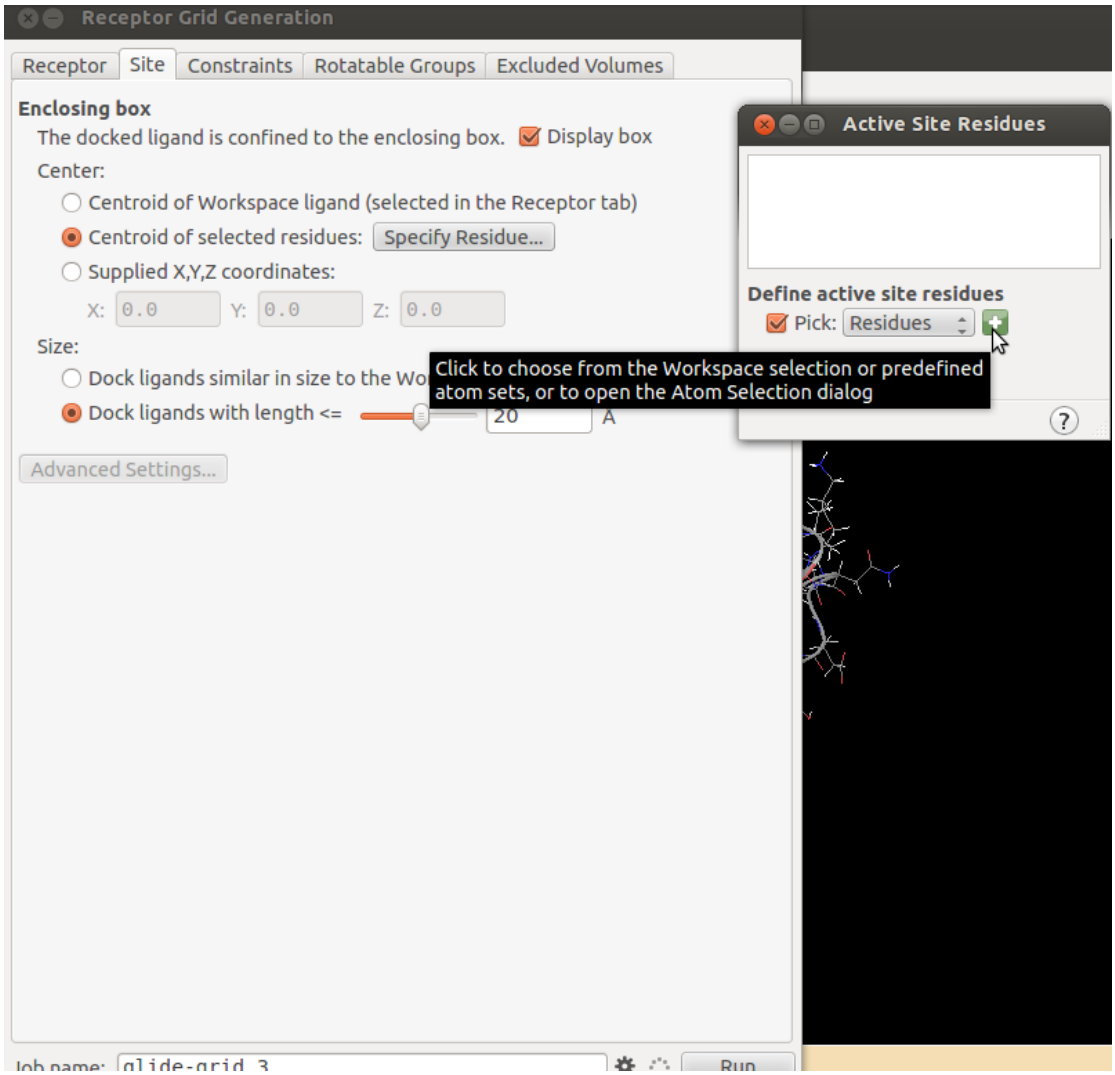


Εικόνα 13: Maestro Glide Προετοιμασία της Πρωτεΐνης

Όπως φαίνεται στο παραπάνω σχήμα η πρωτεΐνη προετοιμάζεται ώστε να μπορεί να δεχτεί τον προσδέτη . Οι επιλογές που έγιναν είναι η προσθήκη υδρογόνων και η δημιουργία δισουλφιδικών δεσμών .

Δημιουργία Πλέγματος

Η πλατφόρμα MAESTRO δημιουργεί πλέγμα του υποδοχέα



Εικόνα 14: Maestro Glide Επιλογή Ορίων πλαισίου

Η επιλογή του πλαισίου από το ενεργό κέντρο του υποδοχέα στο οποίο μπορεί ο υποδοχέας να κινηθεί προσδιορίζεται όπως φαίνεται στην παραπάνω εικόνα μέσω της πλατφόρμας . Για κάθε έναν από τους οκτώ υποδοχείς τα όρια επιλέχτηκαν χειροκίνητα όμοια με την ΣΒ Vina .

Η επιλογή δημιουργία πλέγματος δημιουργεί αρχεία που είναι απαραίτητα για την ταχεία αξιολόγηση στη σύνδεση . Για το λόγο αυτό και επειδή ο αριθμός τον προσδετών που μελετάμε σε κάθε υποδοχέα είναι μεγάλος χρησιμοποιήθηκε το παρακάτω script .

```
#!/bin/tcsh
foreach f ( kif11_cluster_45.maegz )
echo GRIDLIG NO >> $f:r_grid.in
echo USECOMPMAE YES >> $f:r_grid.in
echo INNERBOX 10, 10, 10 >> $f:r_grid.in
echo ACTXRANGE 28.000000 >> $f:r_grid.in
echo ACTYRANGE 28.000000 >> $f:r_grid.in
echo GRID_CENTER 20.615557, 15.259138, 109.482277 >> $f:r_grid.in
echo OUTERBOX 28.000000, 28.000000, 28.000000 >> $f:r_grid.in
echo ENTRYTITLE kif11_1_$f >> $f:r_grid.in
echo GRIDFILE ./glide-grid_kif11_$f:r.zip >> $f:r_grid.in
echo RECEP_FILE $f >> $f:r_grid.in
end
```

στο οποίο προσδιορίζονται τα όρια καθώς όμως και το αρχείο με τις δομές των προσδετών που μας ενδιαφέρουν. Το script αυτό δημιούργησε τα αρχεία πλέγματος για κάθε προσδέτη και στην συνέχεια με τις εντολές

```
#!/bin/sh
for input in $(ls kif11_cluster_45*_grid.in)
do /opt/suite_2012/glide -WAIT $input
done
```

Η πλατφόρμα χρησιμοποιήθηκε και για άλλες δυσκολίες που προέκυψαν π.χ. για κατηγοριοποίηση του τεράστιου αριθμού των προσδετών ώστε ο αριθμός αυτός να μειωθεί.

2.3 Συναρτήσεις Αναβαθμολόγησης

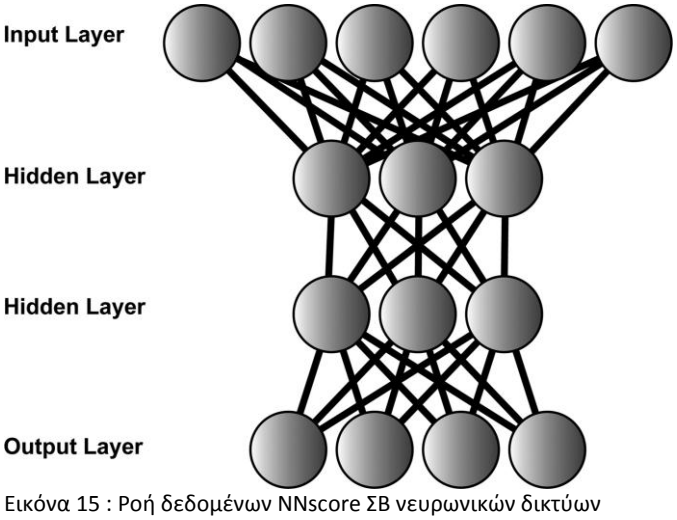
Έχει αναπτυχθεί ένα αυτόματο πλαίσιο για την επιλογή μεταξύ 46 ΣΒ μέσω αναβαθμολόγησης των ποζών σύνδεσης και αξιολόγησης τους με διάφορες μετρήσεις κατάταξης, ταξινόμησης και συσχέτισης. Για κάθε υποδοχέα, ευέλικτο η μη, τα αποτελέσματα του Vina καθώς και του Glide αναβαθμολογήθηκαν με βάση άλλες 45 ΣΒ. Αυτές είναι οι NNscore(1 και 2 αποτελούμενες από 24 και 20 ΣΒ αντίστοιχα) και η ΣΒ DSX :

2.3.1 NNScore

Οι συναρτήσεις βαθμολόγησης NNScore[3] βασίζονται σε νευρωνικά δίκτυα, δηλαδή υπολογιστικά μοντέλα που προσομοιώνουν τη μικροσκοπική οργάνωση του εγκεφάλου. Ο

υποβοηθούμενος από υπολογιστή σχεδιασμός φαρμάκων εξαρτάται από το πόσο γρήγορη και ακριβής είναι συνάρτηση βαθμολόγησης ώστε να βοηθήσει στην ταυτοποίηση των προσδετών μικρού μορίου.

Τα νευρωνικά δίκτυα είναι μοντέλα υπολογιστή σχεδιασμένα να μιμούνται, αν και ανεπαρκώς, τη μικροσκοπική αρχιτεκτονική και η οργάνωση του εγκεφάλου. Εν συντομία, διάφοροι νευρώνες ανάλογη με τη βιολογική έννοια νευρώνες, ενώνονται με «συνδέσεις» ανάλογες με τις νευρωνικές . Η συμπεριφορά του δικτύου δεν καθορίζεται μόνο από την οργάνωση και τον αριθμό των νευρώνων , αλλά και από την δύναμη των συνδέσεων. Όλα τα νευρωνικά δίκτυα έχουν τουλάχιστον δύο στρώματα. Το πρώτο, που ονομάζεται στρώμα εισόδου, λαμβάνει πληροφορίες σχετικά με το σύστημα του δικτύου που πρόκειται να αναλυθεί. Το δεύτερο, που ονομάζεται στρώμα εξόδου, κωδικοποιεί τα αποτελέσματα της εν λόγω ανάλυσης. Επιπλέον, κρυμμένα στρώματα μεταφέρουν πληροφορίες από το στρώμα εισόδου στο στρώμα εξόδου, επιτρέποντας ακόμα πιο πολύπλοκη συμπεριφορά.



Εικόνα 15 : Ροή δεδομένων NNscore ΣΒ νευρωνικών δικτύων

Κατά το σχεδιασμό ενός νευρωνικού δικτύου προς ανάλυση ενός σύνθετου συνόλου δεδομένων, οι ειδικοί τύποι που περιγράφουν τις σχέσεις μεταξύ των χαρακτηριστικών των δεδομένων δεν χρειάζεται να οριοθετηθούν πλήρως. Ο σχεδιαστής οφείλει μόνο να παρέχει στο δίκτυο με μια επαρκή περιγραφή του συστήματος, έτσι ώστε το δίκτυο να μπορέσει να συμπεράνει αυτές τις σχέσεις. Στο σημερινό πλαίσιο, η δημιουργία των νευρωνικών δικτύων για το χαρακτηρισμό της συγγένειας δέσμευσης του συμπλόκου πρωτεΐνης-προσδέτη δεν απαιτούν την εφαρμογή ή ακόμη και να κατανόηση των συγκεκριμένων σχέσεων που περιγράφουν αλληλεπιδράσεις van der Waals, δεσμού υδρογόνου, αν και οι ενέργειες που υπολογίζονται από αυτές τις αλληλεπιδράσεις μπορούν θεωρητικά να συμπεριληφθούν στη στοιβάδα εισόδου.

2.3.2 DSX

Η συνάρτηση βαθμολόγησης DSX [2] συνδυάζει τη δυναμικότητα των συναρτήσεων βασισμένων στη γνώση για αποστάσεις ατόμων με τα δυναμικά γωνιών στρέψης και μια νέα μέτρηση των αλλαγών στην προσιτή επιφάνεια διαλύτη.

Η συνολική βαθμολογία για ένα δεδομένο σύμπλεγμα των ατόμων πρωτεΐνης a_p και προσδέτη άτομα a_l υπολογίζεται από τις εξισώσεις

$$\text{total score}_{\text{pair}} = \sum_{a_p} \sum_{a_l} \text{score}(p(a_p), l(a_l), r(a_p, a_l))$$

$$\text{score}_{\text{pair}}(p, l, r) = -\ln\left(\frac{\rho(p, l, r)}{\rho_{\text{ref}}}\right)$$

όπου $p(a_p)$ και $L(a_l)$ είναι οι τύποι άτομο και $r(a_p, a_l)$ είναι η απόσταση a_p και a_l.Η εξίσωση δεν περιορίζεται απαραίτητα από τις εξαρτώμενες από την απόσταση βαθμολογίες των ατόμων, αλλά μπορεί επίσης να εφαρμοστεί σε πολλά άλλα δομικά χαρακτηριστικά, όπως οι δεσμοί ή η δίεδρη γωνία. Με τη χρήση της θεωρίας των πιθανοτήτων Bayesian, μπορούν να προκύψουν παρόμοιες εξισώσεις, αλλά το ουσιαστικό πρόβλημα που απορρέει από τις συναρτήσεις πιθανότητας και η προϋπόθεση της ανεξαρτησίας κατά ζεύγη δεν πληρείται.

2.4 Μετρήσεις Στατιστικών Δεικτών

Για την αξιολόγηση της απόδοσης της κάθε συνάρτησης βαθμολόγησης, υλοποιούνται στο λογισμικό μας αρκετές στατιστικές μετρήσεις, οι οποίες ομαδοποιούνται σε 3 κατηγορίες:

1. Κατάταξης
2. Τάξης
3. Συσχέτισης

Οι στατιστικές ταξινόμησης και τάξης είναι μη-παραμετρικές στατιστικές, με την έννοια ότι δεν εξαρτώνται από την κατανομή των πιθανοτήτων του πληθυσμού. Σε αντίθεση, η συσχέτιση είναι μια τυπική παραμετρική στατιστική και εξαρτάται από την υποκείμενη κατανομή πιθανότητας των δεδομένων του δείγματος.

Για να μειωθεί η ασυμμετρία των πειραματικών σταθερών σύνδεσης και να είναι σε θέση να συνδέονται με τα αποτελέσματα σύνδεσης, υπολογίστηκαν εικονικές ελεύθερες ενέργειες δέσμευσης από την προηγούμενες χρησιμοποιώντας τον ακόλουθο εμπειρικό τύπο:

$$E = -R\,T\,\ln\left(\frac{10^{-6}}{K}\right)$$

όπου το K μπορεί να είναι Ki, Kd, IC50 ή EC50.

Για κάθε κατάταξη υπολογίζονται οι δείκτες που αναφέρθηκαν προηγούμενα, δηλαδή, η Top-Down Concordance (Iman και Conover, 1987 Teles, 2012), η Kendall's tau, η καμπύλη λειτουργικού χαρακτηριστικού δέκτη (ROC), αλλά και δύο από τις παραλλαγές του, η συμπυκνωμένη καμπύλη (CROC)[5] και η ενισχυμένη Boltzmann (BEDROC).

2.5 Αναβαθμολόγηση μέσω `RescoringTK.py`

Το script της python `RescoringTK.py` εκτελείται για κάθε υποδοχέα χρησιμοποιώντας τα αρχεία με τις βαθμολογημένες πόζες που δημιουργήθηκαν από το Glide και το Vina. Οι παράμετροι του αλγόριθμου είναι οι εξής:

`--dir`
Οπού διευκρινίζεται η τοποθεσία του καταλόγου σύνδεσης.
`--ki`
Όπου διευκρινίζεται το αρχείο με τις σταθερές πρόσδεσης των προσδετών.
`--deltag`
Ίδια με --ki αλλά αντί οι τιμές ki αυτό περιέχει τις εκτιμώμενες τιμές πειραματική ΔG.
`--activity`
Αντίστοιχη τις παραμέτρου --ki αλλά αντί Ki τιμές περιέχει "1", εάν ο προσδέτης είναι μία δραστική ένωση και "0", εάν αυτό είναι ανενεργή.

Π.χ.

< ονομασία προσδέτη > 1
< ονομασία προσδέτη > 0
κ.λπ.

Αυτή η επιλογή χρησιμοποιήθηκε μόνο για να υπολογιστούν οι μετρήσεις ταξινόμησης ROC, CROC, BEDROC . Δηλαδή για τις περιπτώσεις μεγάλου συνόλου μορίων(DUD-E) , όπου μόνο την πληροφορία αν ο κάθε προσδέτης είναι ενεργός ή όχι.

`--nn1`
Χρήση των NNscore1 για αναβαθμολόγηση

`--nn2`
Χρήση NNscore2 για αναβαθμολόγηση

`--dsx`
Χρήση DSX για αναβαθμολόγηση

`--noprep`
Να μην προετοιμάζονται τα αρχεία υποδοχέα και προσδέτη για αναβαθμολόγηση, με την προϋπόθεση ότι αυτό έχει ήδη γίνει.

`--cluster`
Αξιολογήσει μόνο τα αποτελέσματα για το καθορισμένο αναγνωριστικό συμπλέγματος.

`--nposes`
Να χρησιμοποιηθούν μόνο οι πρώτες πόζες N (σύμφωνα με την αρχική κατάταξη Vina ή Glide)

`--write`
Δημιουργεί αρχεία διαγραμμάτων για τα δεδομένα συσχέτισης (Top-Down Concordance,Kendall's τ,Pearson's R) και καμπύλες (ROC, CROC, BEDROC).

`--eval`
Αξιολόγηση των αποτελεσμάτων από τις ΣΒ Vina, NNscore1, NNscore2, ή DSX.

`--report`
Εξαγωγή αρχείου, κάθε γραμμή του οποίου περιέχει την ταυτότητα του συμπλέγματος υποδοχέα, το όνομα, το ισομερές με την υψηλότερη συγγένεια δέσμευσης (εάν υπάρχει), τον αριθμό της καλύτερης πόζας και την βαθμολογία.

Ένα παράδειγμα χρήσης `RescoringTK.py` για τον έναν ευέλικτο υποδοχέα που χρησιμοποιήθηκε με ένα σύνολο προσδετών χρησιμοποιώντας αρχείο με πειραματικές μετρήσεις σταθερών πρόσδεσης(all_curated_ligands.dat)

```
RescoringTK.py --dir `pwd` --eval vina --ki reports.all/all_curated_ligands.dat --flex --report >& reports/Vina_evaluation.log

RescoringTK.py --dir `pwd` --nn1 --ki reports.all/ all_curated_ligands.dat --flex

RescoringTK.py --dir `pwd` --eval nn1 –ki  reports.all/ all_curated_ligands.dat --flex –report >& reports/NNScore1_evaluation.log

RescoringTK.py --dir `pwd` --nn2 --ki all_curated_ligands.dat --flex --noprep

RescoringTK.py --dir `pwd` --eval nn2 --ki reports.all/ all_curated_ligands.dat --flex --report >& reports/NNScore2_evaluation.log

RescoringTK.py --dir `pwd` --dsx --ki all_curated_ligands.dat --flex --noprep

RescoringTK.py --dir `pwd` --eval dsx --ki reports.all/ all_curated_ligands.dat --flex --report >& reports/DSX_evaluation.log
```

Σύμφωνα με τις παραπάνω εντολές(**`--nn1`**,**`--nn2`**,**`--dsx`**) γίνεται αναβαθμολόγηση των αποτελεσμάτων Vina με χρήση των ΣΒ NNscore1(24 ΣΒ) ,NNscore2(20 ΣΒ) και DSX καθώς και αξιολόγηση αυτών(**`--eval`**). Επίσης γίνεται εξαγωγή των αρχείων με τα αποτελέσματα (**`--report`**).

Τα αρχεία που προκύπτουν από την διαδικασία αυτή είναι τα εξής :

```
NNScore1_evaluation.log
NNScore2_evaluation.log
Rescoring_Report_Consensus_Score.dat
Rescoring_Report_DSX_all_isomers.dat
Rescoring_Report_DSX_averageIsomerE.dat
Rescoring_Report_DSX_lowestEisomer.dat
Rescoring_Report_NNscore1_all_isomers.dat
Rescoring_Report_NNscore1_averageIsomerE.dat
Rescoring_Report_NNscore1_lowestEisomer.dat
Rescoring_Report_NNscore2_all_isomers.dat
Rescoring_Report_NNscore2_averageIsomerE.dat
Rescoring_Report_NNscore2_lowestEisomer.dat
Vina_evaluation.log
Rescoring_Report_Vina_all_isomers.dat
Rescoring_Report_Vina_averageIsomerE.dat
Rescoring_Report_Vina_lowestEisomer.dat
```

Τα αρχεία αξιολόγησης (evaluation.log) δίνουν τις τιμές των ROC,CROC,BEDROC,Pearson's R,Kendall's tau,Top-Down Concordance) π.χ.

```
Area under ROC curve of lowest Energy/top Scored isomers: 0.356060606061 under CROC curve 0.0716570315913 under ROC (croc package) 0.356060606061 BEDROC 0.448760020837
```

Ενώ τα αρχεία που προκύπτουν από την εντολή **--report** περιέχουν την ταυτότητα του υποδοχέα , το όνομα της ΣΒ που χρησιμοποιήθηκε , το ισομερές με υψηλότερη συγγένεια δέσμευσης, τον αριθμό της καλύτερης πόζας και την βαθμολογία.

CLUSTER	***SCORING_FUNCTION***	***LIGAND***
000001	*Vina*	*chembl1096003*
BEST_ISOMER	***BEST_POSE***	***ENERGY/SCORE***
chembl1096003_iso2	*1*	*-9.6*
Z-SCORE	***EXPERIMENTAL_ENERGY***	***Z-SCORE***
1.49022160366	*-7.81782198725*	*-1.6042012568*

Τα αποτελέσματα της αναβαθμολόγησης συγκρίθηκαν ώστε να εξεταστεί εάν βελτιώνουν η όχι την απόδοση τις αρχικής εκτίμησης ΣΒ(Vina και Glide) στις περιπτώσεις εύκαμπτου και άκαμπτου υποδοχέα.

2.6 Συναινετική Βαθμολόγηση

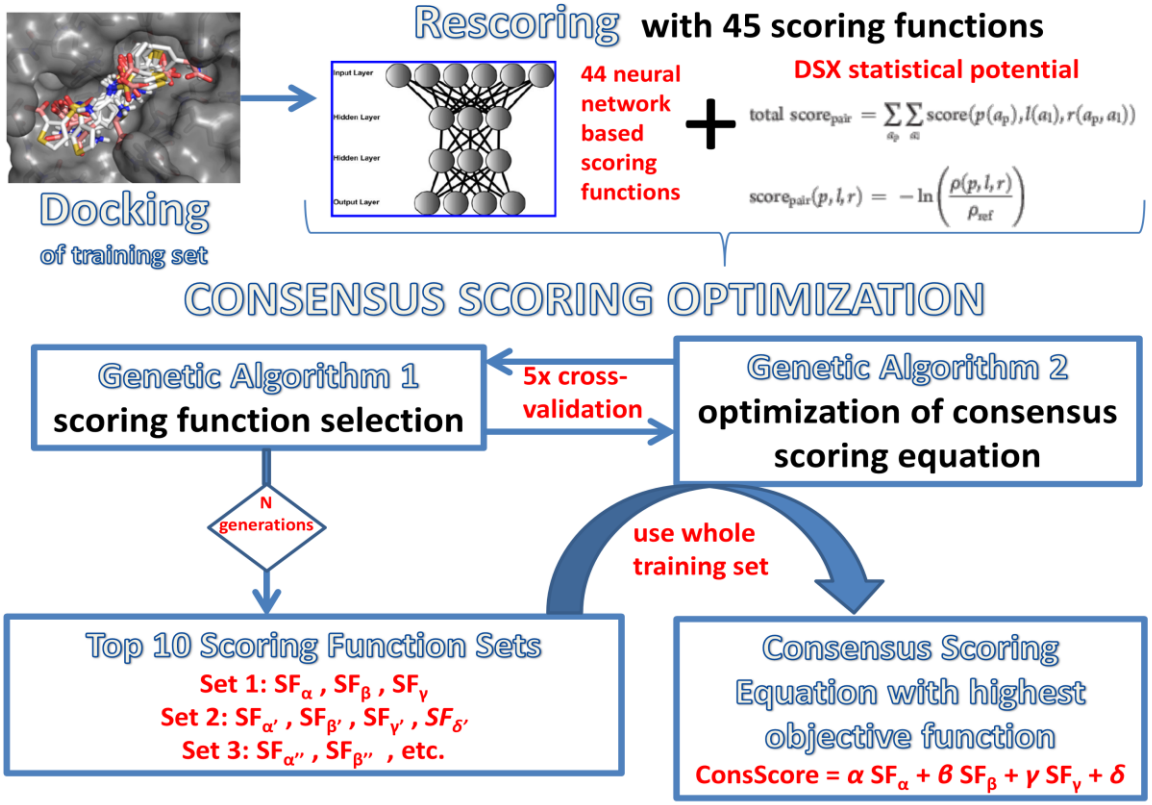
Όπως προαναφέρθηκε, η συναινετική βαθμολόγηση χρησιμοποιείται εδώ και πολλά χρόνια, αλλά είτε περιορίζεται σε μικρό αριθμό συναρτήσεων βαθμολόγησης (ΣΒ), ή συνδυάζει πολλές ΣΒ αλλά με μέτριο αποτέλεσμα. Συνήθως περιλαμβάνουν μοντέλα παλινδρόμησης που προσπαθούν να μεγιστοποιήσουν τη συσχέτιση μεταξύ της συναινετικής βαθμολόγησης και των πειραματικών ενεργειακών βαθμολογιών (π.χ.LogIC50).

Στην παρούσα εργασία, αναπτύχθηκε και εφαρμόστηκε ένας γενετικός αλγόριθμος για να επιτευχθεί η βελτιστοποίηση συναινετικής βαθμολόγησης συνάρτησης συγκεκριμένου στόχου. Οι κατάλληλες μεταβλητές επιλέγονται σε δύο στάδια χρησιμοποιώντας βελτιστοποίηση πολλαπλών στόχων. Στο πρώτο στάδιο, ένας ειδικά σχεδιασμένος γενετικός αλγόριθμος επιλέγει τα σύνολα των ΣΒ που έχουν καλές επιδόσεις, ενώ στο δεύτερο στάδιο ένας δεύτερος γενετικός αλγόριθμος βρίσκει το βέλτιστο γραμμικό συνδυασμό τους. Η αντικειμενική συνάρτηση που επιδιώκεται να μεγιστοποιηθεί μπορεί να είναι παραμετρική, μη-παραμετρική, βασισμένη στην ταξινόμηση ή ένας συνδυασμός αυτών.

2.6.1 Λειτουργία γενετικού αλγορίθμου

Το σύστημα συναινετικής βαθμολόγησης που ακολουθήθηκε περιγράφεται στην Εικόνα 16, και απαιτεί τον κατάλογο των ποζών σύνδεσης καθώς και τις αντίστοιχες βαθμολογίες τους ως είσοδο . Όσο περισσότερες είναι οι ατομικές ΣΒ στην είσοδο , τόσο μεγαλύτερη είναι η απόδοση της συναίνεσης βαθμολόγησης. Σε αυτή τη μελέτη, η είσοδος ήταν το σύνολο DUD-E*[4]* μετά από υπολογισμό από Vina[13] ή Glide[1] (ανάλογα με το λογισμικό που είχε τις καλύτερες στατιστικές μετά την αναβαθμολόγηση) και αναβαθμολογήθηκε από 45

ατομικές ΣΒ. Ως εκ τούτου, χρησιμοποιήθηκαν 46 ΣΒ για την βελτιστοποίηση της συναινετικής βαθμολόγησης.



Εικόνα 16: Διάγραμμα ροής της βελτιστοποίησης συναινετικής βαθμολόγησης

Η μέθοδος που χρησιμοποιήθηκε απασχολεί δύο γενετικούς αλγορίθμους (ΓΑ1, ΓΑ2) οι οποίοι αλληλοσυνδέονται. Ο γενετικός ΓΑ1 παράγει τυχαίους συνδυασμούς των συναρτήσεων βαθμολόγησης , τους οποίους «στέλνει» στον δεύτερο ΓΑ2 ώστε ο τελευταίος να υπολογίσει το βέλτιστο γραμμικό συνδυασμό τους, και, τέλος, ο ΓΑ2 επεξεργάζεται κάθε γραμμικό συνδυασμό που παράγεται από τον ΓΑ1 ώστε να κρατήσει τους καλύτερους συνδυασμούς των συναρτήσεων βαθμολόγησης για την επόμενη γενιά. Η ακριβής ροή εργασίας είναι περιγράφεται παρακάτω:

Σύμφωνα με την μέθοδο διασταυρωμένης επαλήθευσης διαχωρίζεται το σύνολο για κάθε υποδοχέα DUD-E σε 5 μέρη. Τα 4 υποσύνολα διαμορφώνουν το σύνολο εκπαίδευσης και το άλλο υποσύνολο αποτελεί το σύνολο επαλήθευσης. Η διαδικασία επαναλαμβάνεται 5 φορές, κάθε φορά με διαφορετικό υποσύνολο ως σύνολο επαλήθευσης. Στη συνέχεια, για κάθε υποδοχέα DUD-E:

1. Με την εκκίνηση του ΓΑ1 δημιουργείται ένας πληθυσμός τυχαίων ατόμων (σύνολα συναρτήσεων βαθμολόγησης). Για κάθε άτομο επικαλείται ο ΓΑ2 5 φορές, μία για κάθε υποσύνολο επικύρωσης. Ο ΓΑ2 υπολογίζει τον βέλτιστο γραμμικό συνδυασμό των παρεχόμενων συναρτήσεων βαθμολόγησης και επιστρέφει την τιμή της συνάρτησης στόχου , καθώς και οι συναινετικές βαθμολογίες των ενώσεων του συνόλου επικύρωσης. Όταν οι 5 επικλήσεις του GA2 ολοκληρωθούν, ο ΓΑ1 θα συγκεντρώσει τις συναινετικές βαθμολογίες όλων των ενώσεων του συνόλου δεδομένων και θα υπολογίσει της τιμή της συνάρτησης στόχου. Αυτή η τιμή είναι η φυσική κατάσταση του ατόμου και θα καθορίζει εάν θα περάσει στην επόμενη γενιά.

2. Η γενιά του ΓΑ1 συνεχίζει με τη διατήρηση των καλύτερων ατόμων και με την εφαρμογή των μεταλλάξεων και των διασταυρώσεων μεταξύ τους. Οι πράξεις αυτές παράγουν ένα νέο πληθυσμό ατόμων.

3. Τα βήματα 2 και 3 επαναλαμβάνονται μέχρι να επιτευχθεί ο καθορισμένος αριθμός των γενεών ΓΑ1 .

4. Τέλος, ο ΓΑ2 επιστρέφει τα 10 μοναδικά, πιο ισχυρά σύνολα των συναρτήσεων βαθμολόγησης . Για το καθένα από αυτά, τρέχει ο ΓΑ2 ανεξάρτητα χρησιμοποιώντας όλες τις ενώσεις του τρέχοντος DUD-E υποδοχέα ως σύνολο εκπαίδευσης. Μετά την ολοκλήρωση, ο ΓΑ2 θα επιστρέψει το βέλτιστο γραμμικό συνδυασμό του κάθε σύνολο συναρτήσεων βαθμολόγησης , μαζί με την αντίστοιχη τιμή του.

5. Μεταξύ των κορυφαίων 10 γραμμικοί συνδυασμοί των ΔΤ επιλέγουμε το ένα με τη μέγιστη τιμή ως τη βέλτιστη εξίσωση συναινετικής βαθμολόγησης για τον τρέχων DUD-E υποδοχέα.

Έχουμε ρυθμίσει προσεκτικά τις παραμέτρους του ΓΑ (μετάλλαξη, ποσοστό διασταυρώσεων, μέγεθος του πληθυσμού, αριθμός γενιάς) για να εξασφαλιστεί η σύγκλιση των αποτελεσμάτων. Επιπρόσθετα , έχουμε προσαρμόσει το μέγεθος του πληθυσμού του ΓΑ1 ανάλογα με το μέγιστο αριθμό των ΣΒ στην εξίσωση συναινετικής βαθμολόγησης (όσο περισσότερες είναι οι ΣΒ τόσο υψηλότερη είναι η τιμή του πληθυσμού). Έχουμε επίσης εισαγάγει μια λειτουργία που ονομάζεται «συνωστισμός» ο οποίος ελέγχει πόσο τα μεταλλαγμένα άτομα θα μοιάζουν με τους γονείς τους. Παρατηρήθηκε ότι με την αύξηση του συνωστισμού σε κάθε γενιά μπορούμε να πάρουμε καλύτερη σύγκλιση προς το μέγιστο της συνάρτησης στόχου. Από την εφαρμογή όλων των παραπάνω, μπορούμε να επιτύχουμε πολύ καλή σύγκλιση τόσο σε επίπεδο επιλογής των συναρτήσεων βαθμολόγησης όσο και στη βελτιστοποίηση των γραμμικών συνδυασμών των ΣΒ.

Ως εκ τούτου, η επανάληψη της διαδικασία αρκετές φορές είναι περιττή.

Η συνάρτηση στόχου που επιδιώκεται να μεγιστοποιηθεί μπορεί να αποτελείται από έναν μόνο στατιστικό δείκτη (βελτιστοποίηση μοναδικού στόχου) ή πολλαπλούς στατιστικούς δείκτες (βελτιστοποίηση πολλαπλών – στόχων). Στην τελευταία περίπτωση, η τιμή της συνάρτησης είναι ένας σταθμισμένος συνδυασμός των τιμών των επιμέρους στατιστικών δεικτών, με την βαρύτητα που καθορίζονται από το χρήστη . Η επιλογή των καλύτερων ατόμων σε κάθε γενιά γίνεται με τη χρήση του Pareto . Στην περίπτωση όπου πολλά άτομα βρίσκονται στην «πρώτη γραμμή» του Pareto, η επιλογή γίνεται σύμφωνα με τους σταθμισμένη απόσταση που υπολογίζεται από την εξίσωση :

$$Distance = \sqrt{\Sigma_m^{all\ metrics} w_m m^2}$$

όπου w_m είναι το βάρος της στατιστικής μέτρησης m . Για παράδειγμα , στην περίπτωση των εύκαμπτων υποδοχέων σύνδεσης μεγιστοποιήθηκαν τρείς στατιστικές μετρήσεις , το Top - Down Concordance[7] , η Kendall's τ και η Pearson 's R , χρησιμοποιώντας τις τιμές βαρών 1.0 , 0.5 και 0.5 , αντίστοιχα. Η σταθμισμένη απόσταση του κάθε ατόμου μετρήθηκε χρησιμοποιώντας την εξίσωση:

$$Distance = \sqrt{C^2 + 0.5\tau^2 + 0.5R^2}$$

Στην βελτιστοποίηση μοναδικού στόχου χρησιμοποιήθηκε ο δείκτης CROC [5]. Αυτό έγινε για το λόγο ότι οι ο δείκτης CROC βρέθηκε πλήρως συσχετισμένος με της ROC και BEDROC οπότε η βελτιστοποίηση και των τριών δεικτών θα ήταν άσκοπα χρονοβόρα. Η συσχέτιση υπολογίστηκε μέσω της Pearson's R. Η Pearson's R μεταξύ AU-ROC και AU-CROC ήταν 0,8960, μεταξύ AU-ROC και AU-BEDROC ήταν 0,7467, και μεταξύ AU-CROC και AU-BEDROC 0,94750715.

Η βελτιστοποίηση της συναινετικής βαθμολόγησης διεξήχθη χρησιμοποιώντας τα python scripts scoringfunction_selection_script.py και consscore_optimization_script.py. Και τα δύο αυτά scripts απασχολούν τροποποιημένων λειτουργιών της ενότητας DEAP της Python[15] για τους γενετικούς αλγόριθμους , και λειτουργίες από τη μονάδα SCOOP της Python[14] .

Ως είσοδο Γ.Α. είναι αναγκαία μια μετατροπή των αρχείων διασταυρωμένης επικύρωσης για την οποία χρησιμοποιείται ένα python script (**`create_ConsScorTK_file.py`**).
Το script αυτό χρησιμοποιεί τα αρχεία που έχουν προκύψει από το script που χρησιμοποιήθηκε για την αναβαθμολόγηση(**`RescoringTK.py`**) καθώς και τα αρχεία που έχουν προκύψει από την διαδικασία της διασταυρωμένης επικύρωσης(Cross-Validation).

Για παράδειγμα για τον άκαμπτο υποδοχέα CXCR4 χρησιμοποιήσαμε τις εντολές :

```
python2.7 /usr/local/bin/create_ConsScorTK_file.py --scorefile
Rescoring_Report_Glide_lowest_isomers.dat --scorefile
Rescoring_Report_NNscore1_lowest_isomers.dat \

--scorefile Rescoring_Report_NNscore2_lowest_isomers.dat --scorefile
Rescoring_Report_DSX_lowest_isomers.dat --trainfile rigid_cxcr4_cross_validation.txt \

--cluster 000001
```

Έτσι δημιουργήθηκε ένα νέο αρχείο το οποίο χρησιμοποιείται ως είσοδος στον ΓΑ1 και περιέχει τις πληροφορίες όλων τον παραπάνω αρχείων .

Ο ΓΑ1 καλείται μέσω του αρχείου **`run_scoring_function_selection.sh`** απλά με την εντολή

```
./run_scoring_function_selection.sh rigid_cxcr4_cross_validation.txt.ConsScorTK.dat
```

Όπου το **`rigid_cxcr4_cross_validation.txt.ConsScorTK.dat`** είναι το αρχείο που δημιουργήθηκε από το προηγούμενο script .

Οι τιμές των παραμέτρων του γενετικού αλγορίθμου φαίνονται στο αρχείο αυτό όπως αναφέρθηκε επιλέχτηκαν προσεκτικά και μετά από πολλές δοκιμές για να εξασφαλιστεί η σύγκλιση των αποτελεσμάτων .

Το αρχείο αυτό δημιουργεί τους συνδυασμούς των ΣΒ και τους «στέλνει» σε ένα δεύτερο (**`run_consscore_optimization.sh`**) ο οποίος θα επιστρέψει τους δέκα κορυφαίους γραμμικούς συνδυασμό του κάθε συνόλου συναρτήσεων βαθμολόγησης , μαζί με την αντίστοιχη τιμή τους.

Από τα τους δέκα κορυφαίους γραμμικούς συνδυασμούς επιλέγουμε τον καλύτερο και στη συνέχεια χρησιμοποιώντας πάλι το python script `RescoringTK.py` και την εντολή

--consscoreq

Εισάγουμε την εξίσωση της συναινετικής βαθμολόγησης που επιλέξαμε. π.χ.

```
"0.742475884462 * NNscore1_net4 + 0.764218249134 * NNscore1_net16 + 0.412941310405 * NNscore1_net14 + 0.051710300194"
```

Στο ίδιο script με την εντολή **–write** δημιουργήθηκαν τα δεδομένα για τα διαγράμματα των καμπυλών ROC , CROC , BEDROC τα οποία και μελετήθηκαν .

Η διαδικασία αναβαθμολόγησης και συναινετικής βαθμολόγησης έγινε και για τους οκτώ υποδοχείς που μελετήθηκαν .Για τους άκαμπτους υποδοχείς χρησιμοποιήθηκαν μεγάλα σύνολα δεδομένων προσδετών(DUD-E*[4]*) καθώς και μικρότερα(CHEMBL[10]), ενώ για τους ευέλικτους υποδοχείς μικρότερα επιλεγμένα σύνολα CHEMBL[10] καθώς ο υπολογισμός σύνδεσης ευέλικτου υποδοχέα χρησιμοποιώντας μεγάλα σύνολα θα ήταν εξαιρετικά χρονοβόρος .

Για την ενίσχυση του δυναμικού πρόβλεψης της αναβαθμολόγησης και συναινετικής βαθμολόγησης, ενσωματώσαμε ένα βήμα ομοιότητας δακτυλικών αποτυπωμάτων το οποίο φιλτράρει το σύνολο ενώσεων για τις οποίες είναι απίθανο να γίνει σωστή πρόβλεψη. Σε αυτό το βήμα, όπως περιγράφηκε προηγουμένως από [Yun et al., 2013], ορίζεται ένα επίπεδο εμπιστοσύνης ("μεγάλη", "μέτρια" ή "χαμηλή") σε κάθε μόριο της βιβλιοθήκης εικονικής διαλογής βρίσκοντας τις M όμοιες ενώσεις από το σύνολο εκπαίδευσης .

Υποθέτοντας ότι μια ακριβής συνάρτηση βαθμολόγησης ΣBj πρέπει να δίνει χαμηλότερες βαθμολογίες στα δραστικά μόρια και υψηλότερες σε μη-δραστικά ,η ΣBj είναι πιο ακριβής όταν σε προσδέτη I η πρόβλεψη για το I έχει χαμηλότερη τιμή από τις προβλέψεις της για το D(decoys). Ομοίως, η ΣBj είναι πιο ακριβής για μη-δραστικό μόριο d όταν η τιμή πρόβλεψης της για το d είναι υψηλότερη από τις προβλέψεις της για τον L. Η απόδοση για κάθε συνάρτηση βαθμολόγησης υπολογίζεται από :

$$p(i,j) = \begin{cases} \frac{|\{c_k \in D | s(c_k,j) > s(c_i,j)\}|}{|\{c_k \in L\}|}, & c_i \in L \\ \frac{|\{c_k \in L | s(c_k,j) < s(c_i,j)\}|}{|\{c_k \in L\}|}, & c_i \in L \end{cases}$$

όπου D είναι το σύνολο των μη δραστικών μορίων, L το σύνολο το δραστικών και s ($c_κ$, j) είναι η βαθμολογία της ένωσης c_k από την ΣBj. Η τιμή που παράγεται από την εξίσωση βρίσκεται στο διάστημα [0, 1]. Όσο υψηλότερη η τιμή αυτή, τόσο καλύτερη είναι η πρόβλεψη της ΣB για το εν λόγω μόριο. Η απόδοση ΣB σε ένα σύνολο ενώσεων εκπαίδευσης M τα οποία είναι όμοια με ένα μόριο j από την βιβλιοθήκη εικονικής διαλογής μπορεί να υπολογιστεί ως εξής :

$$perf(M) = \frac{\Sigma_i^M sim(i,j) p(i)}{\Sigma_i^M sim(i,j)}$$

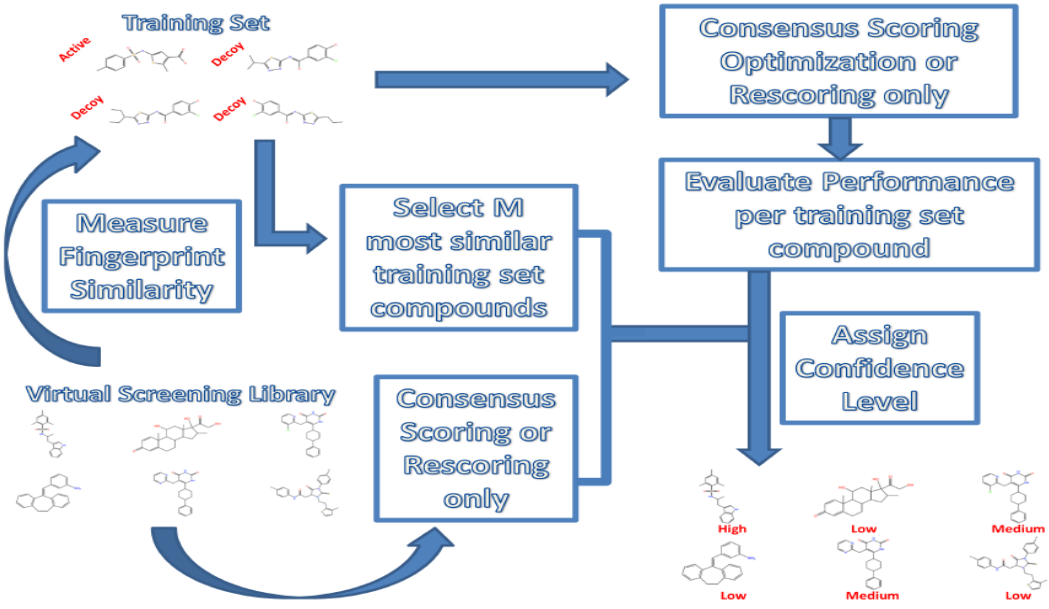
όπου το sim(i,j) είναι η ομοιότητα μεταξύ των μορίων I και j . Η συνολική ομοιότητας αποτυπώματος ενός μορίου j με ένα σύνολο μορίων εκπαίδευσης υπολογίζεται από :

$$sim(M,j) = \frac{\Sigma_i^M sim(i,j)}{M}$$

Το επίπεδο εμπιστοσύνης μίας ΣΒ σε ένα μόριο j ορίζεται ως :

$$conf(j) = \begin{cases} \text{“}HIGH\text{“}, & sim(M,j) > R_{sim} \; and \; perf(M) > R_{perf} \\ \text{“}MEDIUM\text{“}, & sim(M,j) < R_{sim} \; and \; perf(M) > R_{perf} \\ \text{“}MEDIUM\text{“}, & sim(M,j) > R_{sim} \; and \; perf(M) < R_{perf} \\ \text{“}LOW\text{“}, & sim(M,j) < R_{sim} \; and \; perf(M) < R_{perf} \end{cases}$$

Όπου R_{sim} και R_{perf} είναι αυθαίρετες τιμές αποκοπής που καθορίζονται από το χρήστη. Στην παρούσα μελέτη αξιολογείται η απόδοση των 5 διαφορετικών δισδιάστατων δακτυλικών αποτυπωμάτων ομοιότητας (linear, maccs, radial, dendritic, molprint2D) και των δακτυλικών αποτυπωμάτων αλληλεπίδρασης δομής (SIFT). Η τελευταία χρησιμοποιεί τις τρισδιάστατες δομικές πληροφορίες δέσμευσης πρωτεΐνης-προσδέτη κωδικοποιώντας την φύση των αλληλεπιδράσεων μεταξύ του ενεργού κέντρου και του προσδέτη . Με αυτόν τον τρόπο, το SiFt μπορεί δυνητικά προσδιορίσει καλύτερα νέους προσδέτες από ότι τα δισδιάστατα δακτυλικά αποτυπώματα ομοιότητας. Η ομοιότητα δισδιάστατων δακτυλικών αποτυπωμάτων μεταξύ δύο μικρών μορίων είναι σχεδόν πάντα ανεξάρτητη από το ταυτομερές που επιλέχτηκε.



Εικόνα 17: Διάγραμμα ροής για το συνδυασμό των δακτυλικών αποτυπωμάτων ομοιότητας με τον υπολογισμό μοριακής πρόσδεσης .

Ωστόσο, αυτό δεν συμβαίνει με τα δακτυλικά αποτυπώματα αλληλεπίδραση δομής (SIFT). Στη περίπτωση αυτή που υπολογίστηκε το SiFT του καλύτερα βαθμολογημένου ισομερούς ανά ένωση σε περιπτώσεις μεμονωμένων ΣΒ. Στην περίπτωση συναινετικής βαθμολόγησης,

κάθε ΣΒ της εξίσωσής της έχει διαφορετική συμβολή στη συναινετική βαθμολογία και κατατάσσει μια διαφορετική πόζα σύνδεσης στην κορυφή. Η διαδικασία αυτή καθιστά τον υπολογισμό της ομοιότητας Sift υπερβολικά περίπλοκο. Για το λόγο αυτό η ομοιότητα SiFt δεν συνδυάστηκε με τη μελέτη της συναινετικής βαθμολόγησης. Ο υπολογισμός των επιπέδων εμπιστοσύνης για το «τυφλό» δοκιμαστικό σύνολο έγιναν μέσω ενός module (thecacl_confidence.py) του ConsScorTK. Για την ανάλυση επιλέχτηκαν μόνο οι ενώσεις οι οποίες χαρακτηρίστηκαν ως «υψηλής» εμπιστοσύνης .

Η αναβαθμολόγηση των αποτελεσμάτων των Vina ή Glide με τις ΣΒ NNScore1, NNScore2 και DSX, καθώς και η βελτιστοποιημένη συναινετική βαθμολόγηση που χρησιμοποιεί οποιαδήποτε επιπλέον ΣΒ, είναι διαθέσιμα στην εφαρμογή στη διεύθυνση: http://consscortk.molsim.pharm.uoa.gr

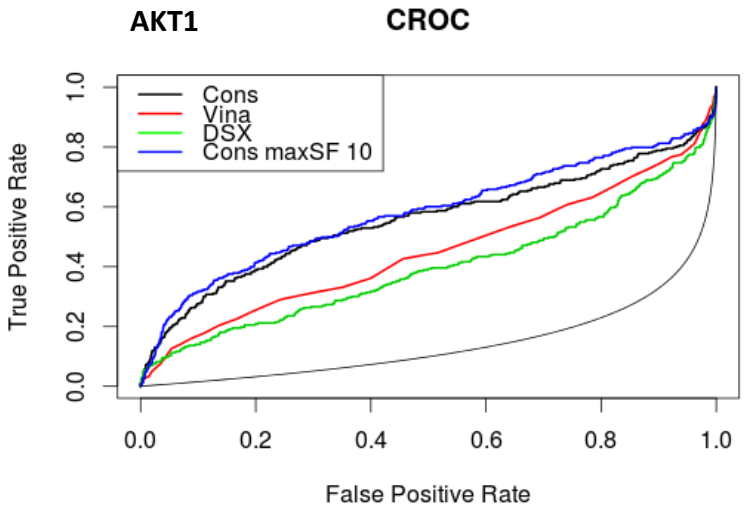
Κεφάλαιο 3

3 Αποτελέσματα

3.1 Άκαμπτος υποδοχέας σύνδεσης μεγάλου συνόλου δεδομένων: Μεγιστοποίηση της CROC

Η καμπύλη συμπυκνωμένου λειτουργικού χαρακτηριστικού δέκτη (CROC) μετρά πόσο καλά μια λειτουργία βαθμολόγησης μπορεί να διακρίνει τους πιο ισχυρούς συνδέτες από τους μη-συνδέτες (decoys). Όσο μεγαλύτερη είναι η περιοχή κάτω από την καμπύλη CROC (AU-CROC) τόσο καλύτερη είναι η συνάρτηση βαθμολόγησης. Στο παρακάτω διάγραμμα παρουσιάζονται τα αποτελέσματα της αναβαθμολόγησης καθώς και της συναινετικής βαθμολόγησης για τον άκαμπτο υποδοχέα AKT1 οπού χρησιμοποιήθηκαν 293 δραστικές ουσίες και 16338 μη δραστικές.

AKT1(293 actives, 16338 decoys)

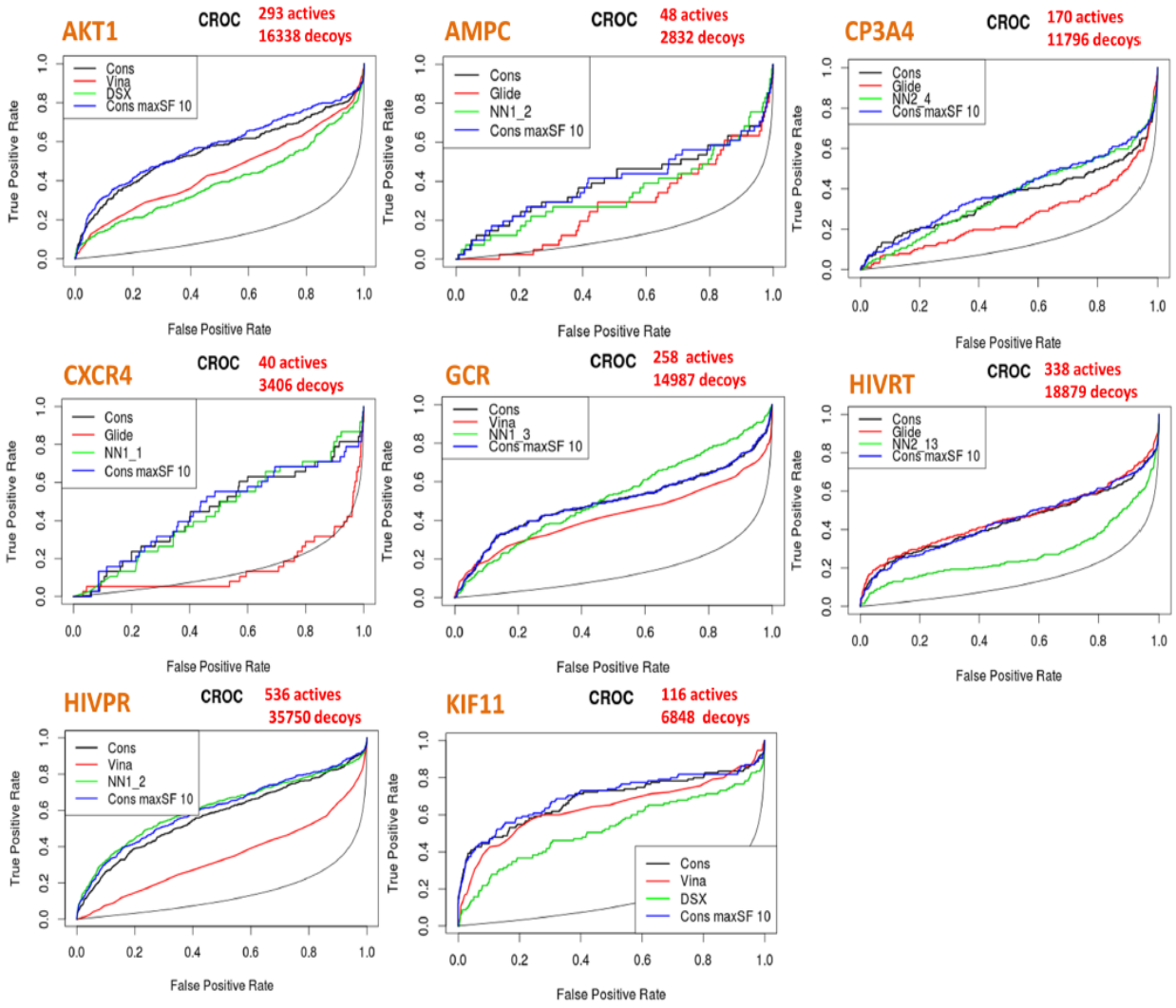


Σύγκριση του λογισμικού σύνδεσης (κόκκινη καμπύλη), της ΣΒ με την υψηλότερη AU-CROC (πράσινη καμπύλη), της συναινετικής βαθμολογίας με 2 έως 5 ΣΒ (μαύρη καμπύλη) και συναινετικής βαθμολογίας με 6 έως 10 ΣΒ (μπλε καμπύλη).

- Στη συναινετική βαθμολόγηση (6-10 ΣΒ) βρέθηκε 46% των δραστικών ουσιών στο 5% της κορυφής των μορίων.
- Στη συναινετική βαθμολόγηση (2-5 ΣΒ) βρέθηκε το 44,7% των δραστικών ουσιών στο 5% της κορυφής των μορίων.
- Στη ΣΒ Vina βρέθηκε 31% των δραστικών ουσιών στο 5% της κορυφής των μορίων.
- Στη ΣΒ DSX βρέθηκε 25,6% των δραστικών στο 5% της κορυφής των μορίων.

Η συναινετική βαθμολόγηση με 6 ως 10 ΣΒ παίρνει την υψηλότερη τιμή CROC(**0.5784)** με μικρή διαφορά από την τιμή της CROC για συναινετική βαθμολόγηση 2 έως 5 ΣΒ(**0.5547)**.Η ΣΒ λογισμικού(Vina) παίρνει τιμή CROC(**0.4478**) ενώ η ΣΒ που αναβαθμολόγησε καλυτέρα την ΣΒ λογισμικού , στην αυτή την περίπτωση η ΣΒ DSX είχε την χαμηλότερη τιμή από τις προηγούμενες τρείς προσεγγίσεις(**0.3980**).

Οι καμπύλες CROC για τους υπόλοιπους υποδοχείς παρουσιάζονται παρακάτω .



Εικόνα 18: Απόδοση των ΣΒ μέσω CROC στο σύνολο εκπαίδευσης με τη χρήση διασταυρωμένης επικύρωσης

Οι καμπύλες CROC λήφθηκαν όπως περιγράφεται στις μεθόδους. Η σύσταση του καθενός συνόλου στοιχείων σε δραστικές ενώσεις και μη φαίνονται στην παραπάνω εικόνα, με κόκκινο χρώμα. Η ΣΒ "NNscore" έχει συντομογραφία «NN». Δηλαδή NN1_2 είναι μια συντομογραφία για τη δεύτερη ΣΒ της NNscore 1.0.

Παρατηρώντας τα διαγράμματα διαπιστώνουμε ότι η συνάρτηση βαθμολόγησης του λογισμικού σύνδεσης σπάνια είναι η καλύτερη. Η συναινετική βαθμολόγηση βελτιώνει την απόδοση μέτρια έως οριακά, ανάλογα με τον υποδοχέα.

Για τα μεγάλα σύνολα μη-δραστικών ενώσεων DUD-E ,έγινε υπολογισμός μοριακής πρόσδεσης και αναβαθμολόγηση με 45 επιπλέον ΣΒ πάνω σε 8 διαφορετικούς υποδοχείς. Από αυτά τα 8 σύνολα εκπαίδευσης που αποτελούνται από 46 ΔΤ (συμπεριλαμβανομένης της ΣΒ του λογισμικού σύνδεσης), παρήχθησαν εξισώσεις συναινετικής βαθμολόγησης αποτελούμενες από 5 ή 10 ατομικές ΣΒ κατά ανώτατο όριο, που μεγιστοποίησαν την AUC-CROC μέσω διασταυρωμένης επικύρωσης 5-υποσυνόλων. Η απόδοση του λογισμικού σύνδεσης ΣΒ, η καλύτερη ατομική ΣΒ και τη συναινετικές βαθμολογήσεις αντιπαρατίθενται στην Εικόνα 16. Όπως αναμενόταν, οι εξισώσεις συναίνεσης ξεπέρασαν όλες οι επιμέρους ΣΒ σε 7 από τους 8 υποδοχείς, με την HIVPR να είναι η μόνη εξαίρεση, αφού η ΣΒ NNscore1_2 βρέθηκε οριακά καλύτερη από τις εξισώσεις συναινετικής βαθμολόγησης. Ωστόσο, η πιο σημαντική πληροφορία σε αυτή την Εικόνα είναι ότι για 5 από τους 8 υποδοχείς υπάρχουν ΣΒ που ξεπερνούν την ΣΒ του λογισμικού σύνδεσης. Είναι αξιοσημείωτο ότι σε υποδοχείς όπως ο CXCR4 ή ο AMPC (ειδικά στο πρώτο μέρος της καμπύλης) η ΣΒ του λογισμικού σύνδεσης έχει όσο χαμηλή απόδοση έχει ένας τυχαίος ταξινομητής. Μόνο στην HIVRT και KIF11 το λογισμικό σύνδεσης ΣΒ μπορεί να διακρίνει καλύτερα τους συνδέτες από τους μη-συνδέτες, ενώ στο AKT1 η απόδοση της είναι οριακά καλύτερη.

Με βάση τους πίνακες που παρατίθενται παρακάτω γίνεται σύγκριση των επιδόσεων μεταξύ Glide και Vina στους υποδοχείς για σύνολο ενώσεων DUD-E *[4]*. Επίσης αναφέρονται τα στατιστικά στοιχεία των συναρτήσεων βαθμολόγησης που έδωσαν τις μεγαλύτερες τιμές AUC-ROC, η AUC-CROC, AUC-BEDROC μετά από αναβαθμολόγηση με Glide ή Vina . Κάτω από το όνομα του λογισμικού σύνδεσης βρίσκονται τα ονόματα των ΣΒ που παρήγαγαν τις νέες πόζες σύνδεσης μετά από αναβαθμολόγηση στο εκάστοτε λογισμικό σύνδεσης.

AKT1	**ROC**	**CROC**	**BEDROC**
Glide	0.7008064366	0.400342957377	0.294864015067
NN1_19	0.7827212933	0.421855526663	0.218923849935
DSX	0.7492539414	0.396865432015	0.284350016253
Vina	0.7800145361	0.447849054148	0.307142602287
NN1_19	0.7492877353	0.309366896585	0.110496426531
DSX	0.7398633191	0.397947994547	0.267292298291

Πίνακας 1 : AKT1 αποτελέσματα αναβαθμολόγησης άκαμπτου υποδοχέα

AMPC	**ROC**	**CROC**	**BEDROC**
Glide	0.6919584375	0.301280043015	0.120249322978
NN1_2	0.750439144933	0.374673946234	0.227292808317
Vina	0.6131992143	0.2211722299	0.111342370348
NN1_3	0.685750148182	0.294630735062	0.141192298776
NN2_13	0.6477052171	0.293557365716	0.199152619837

Πίνακας 2 : AMPC αποτελέσματα αναβαθμολόγησης άκαμπτου υποδοχέα

CP3A4	ROC	CROC	BEDROC
Glide	0.6351840344	0.247844648732	0.136208342795
NN1_24	0.72076006131	0.340262044245	0.190882003124
NN2_4	0.6814961665	0.351474900871	0.210153499019
NN2_8	0.6341369589	0.309004089454	0.214314034267

Πίνακας 3 : CP3A4 αποτελέσματα αναβαθμολόγησης άκαμπτου υποδοχέα

CXCR4	ROC	CROC	BEDROC
Glide	0.5323025535	0.137345461471	0.060397108041
NN1_1	0.77543509639	0.430001130865	0.226974776031
NN1_6	0.6916880152	0.373135572487	0.236895357349
Vina	0.5857677627	0.126495290152	0.029842032441
NN1_1	0.727491926013	0.343907118287	0.134314572502
NN1_6	0.6352209336	0.34341068956	0.19192054851

Πίνακας 4 : CXCR4 αποτελέσματα αναβαθμολόγησης άκαμπτου υποδοχέα

GCR	ROC	CROC	BEDROC
Vina	0.6430180057	0.388506464321	0.297240001811
NN1_3	0.798769786528	0.499367626333	0.335491930063
Vina	0.6430180057	0.388506464321	0.297240001811
NN1_3	0.798769786528	0.499367626333	0.335491930063

Πίνακας 5 : GCR αποτελέσματα αναβαθμολόγησης άκαμπτου υποδοχέα

HIVPR	ROC	CROC	BEDROC
Glide	0.7286926186	0.431459900208	0.296573713613
NN1_2	0.8247022984	0.536734249877	0.379036416699
Vina	0.7116527054	0.343057468167	0.198046651804
NN1_2	0.851124991427	0.626731980948	0.496481266892

Πίνακας 6 : HIVPR αποτελέσματα αναβαθμολόγησης άκαμπτου υποδοχέα

HIVRT	ROC	CROC	BEDROC
Glide	0.7294842236	0.425980679491	0.327488593358
NN2_11	0.650795894336	0.26006960778	0.116609071005
NN2_13	0.6424707269	0.274501674609	0.188977757716
Vina	0.6589198460	0.285817194642	0.157290956867
NN2_13	0.6809676931	0.273346880248	0.1492524354
NN2_14	0.6297002918	0.269798224199	0.160425326726

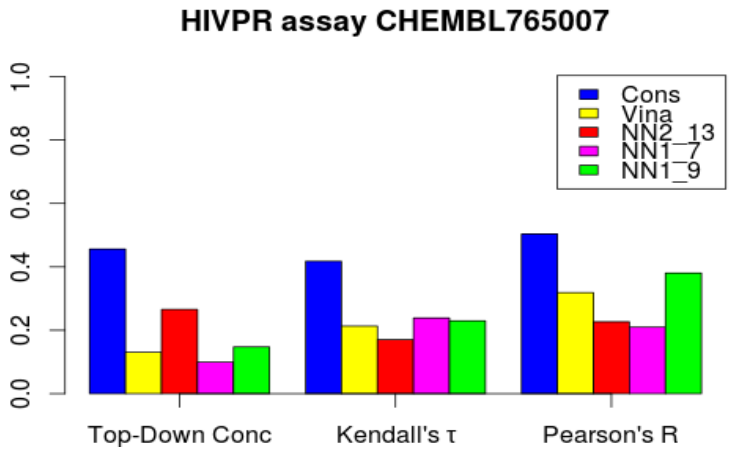
Πίνακας 7 : HIVPR αποτελέσματα αναβαθμολόγησης άκαμπτου υποδοχέα

KIF11	ROC	CROC	BEDROC
Glide	0.7666533446	0.502642971874	0.349803116616
NN1_3	0.802918026444	0.531686630929	0.409159647495
DSX	0.7930951750	0.574275093249	0.464319276035
Vina	0.8623824222	0.635157530399	0.53316610004
DSX	0.7807703734	0.521139573413	0.391568934518
NN1_3	0.7686456403	0.519405871766	0.440223197982

Πίνακας 8 : HIVPR αποτελέσματα αναβαθμολόγησης άκαμπτου υποδοχέα

3.2 Ευέλικτος υποδοχέας σύνδεσης: Ταυτόχρονη Μεγιστοποίηση του Top-Down Concordance, Kendall's τ της Pearson's R

Τα αποτελέσματα αναβαθμολόγησης καθώς και της συναινετικής βαθμολόγησης του ευέλικτο υποδοχέα HIVPR χρησιμοποιώντας το σύνολο δεδομένων CHEBML765007 παρουσιάζονται στο παρακάτω διάγραμμα . Το συγκεκριμένο σύνολο περιέχει σταθερές πρόσδεσης για 107 προσδέτες.



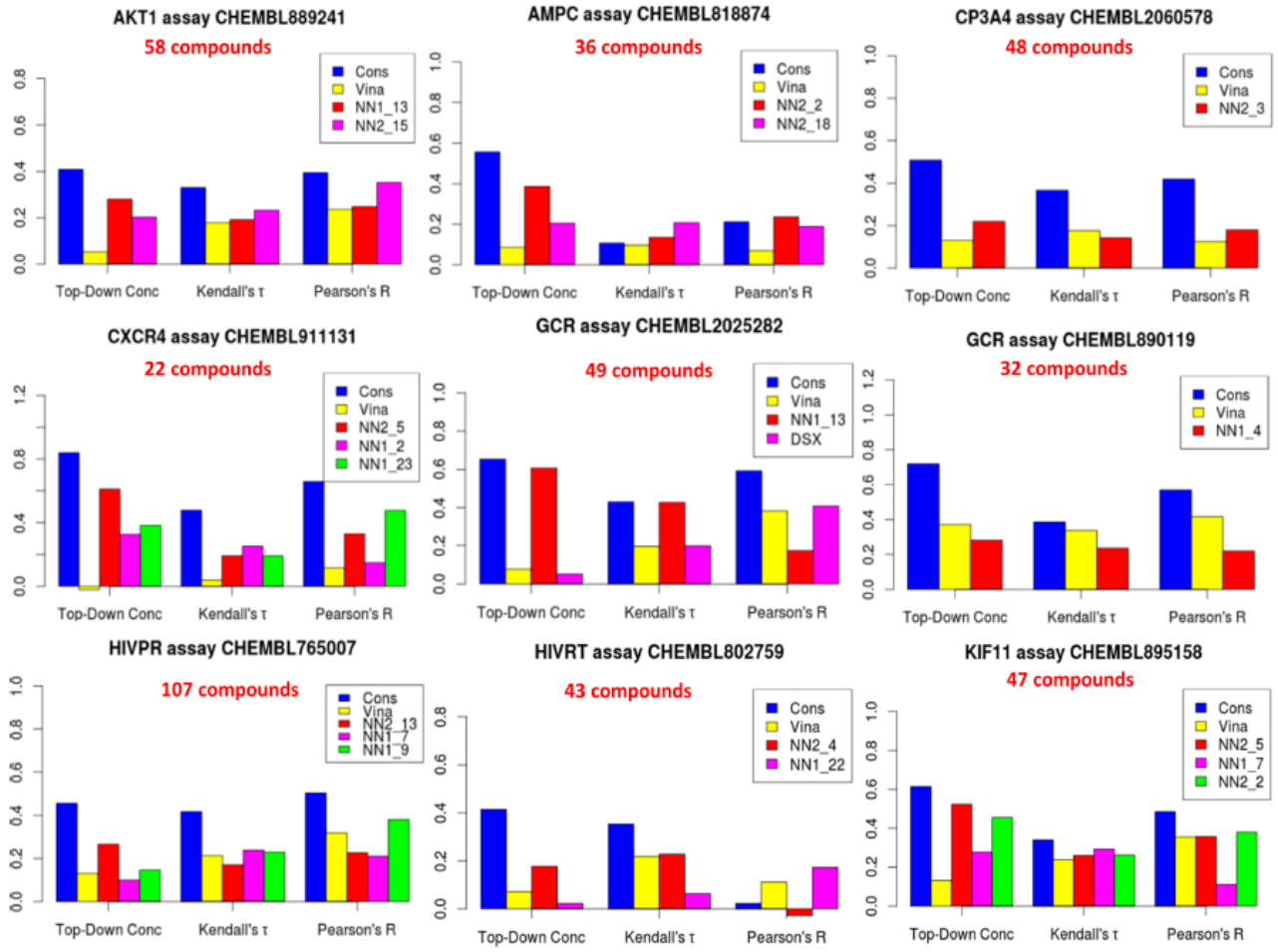
CHEMBL765007(107 ligands)

Οι μπλε μπάρες αναφέρονται στις τιμές της συναινετικής βαθμολόγησης , κίτρινες στην ΣΒ του λογισμικού (σε αυτήν την περίπτωση Vina). Οι υπόλοιπες μπάρες αναφέρονται στις ΣΒ αναβαθμολόγησης οι οποίες έδωσαν τις καλύτερες βαθμολογίες για κάθε μία από τις μετρήσεις κατάταξης (Top-Down Concordance,Kendall's τ, Pearson's R) . Στον παρακάτω πίνακα φαίνονται η ακριβής τιμές που χρησιμοποιήθηκαν για το παραπάνω διάγραμμα.

ΣΒ	Top-Down	Kendall's τ	Pearson's R
Consensus	0.456	0.418	0.504
Vina	0.13	0.213	0.319
NN2_13	0.266	0.171	0.227
NN1_7	0.1	0.239	0.21
NN1_9	0.148	0.23	0.38

Πίνακας 9: HIVPR assay CHEMBL765007

Παρακάτω παρουσιάζεται τα πιο αντιπροσωπευτικά σύνολα CHEMBL με διάγραμμα για κάθε ευέλικτο υποδοχέα.

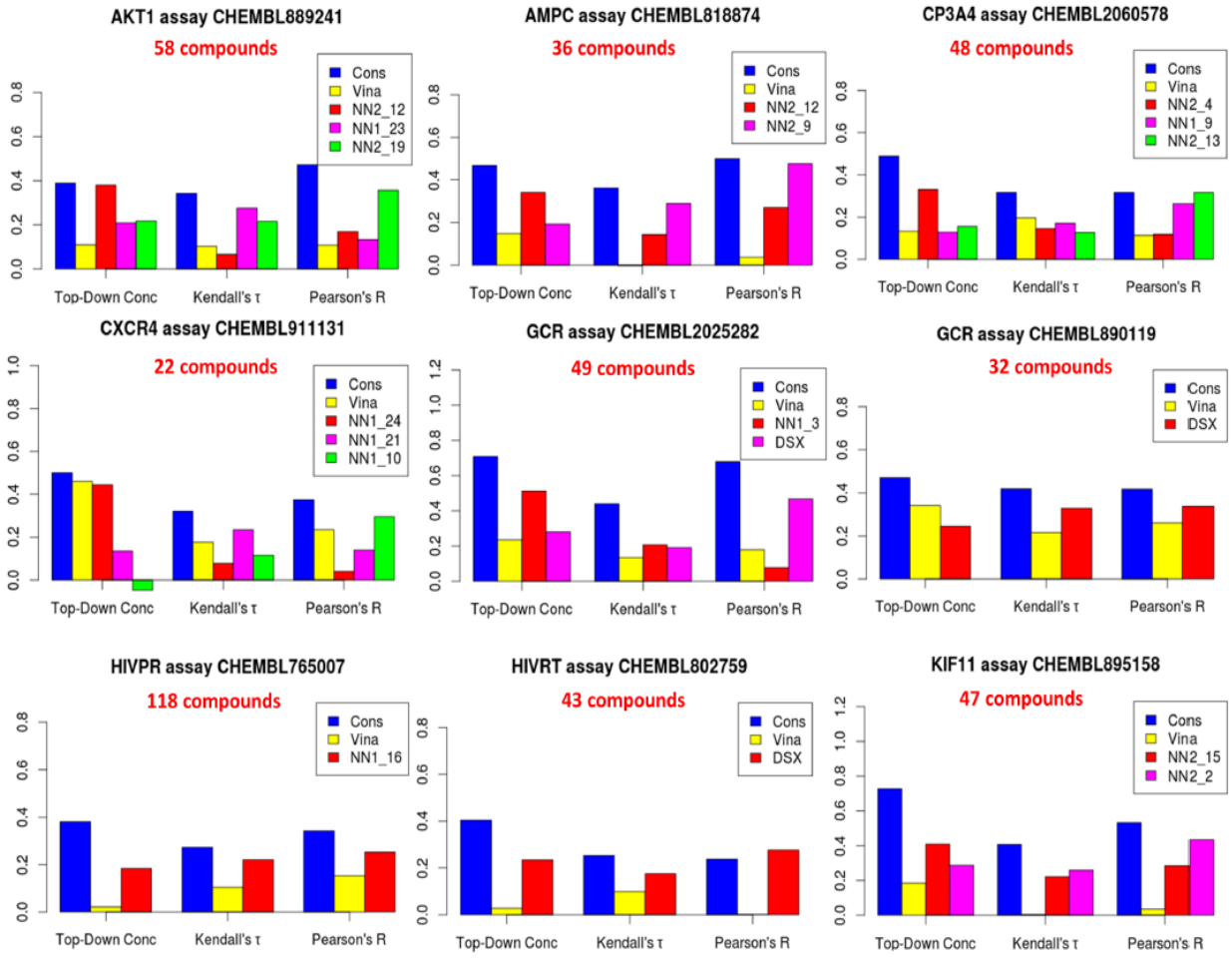


Εικόνα 19 : Αντιπροσωπευτικά σύνολα CHEMBL με διάγραμμα για κάθε ευέλικτο υποδοχέα.

Η συναινετική βαθμολόγηση (μπλε μπάρες) κατατάσσεται πρώτη και στις 3 μετρήσεις ταξινόμησης έχοντας 2 εξαιρέσεις(AMPC και KIF11).Μεταξύ των επιμέρους συναρτήσεων βαθμολόγησης η Vina (κίτρινες μπάρες) κατατάσσεται πρώτη σε μία μόνο περίπτωση(CP3A4).

3.3 Άκαμπτος υποδοχέας σύνδεσης: Ταυτόχρονη Μεγιστοποίηση του Top-Down Concordance, Kendall's τ της Pearson's R

Τα αποτελέσματα αναβαθμολόγησης καθώς και της συναινετικής βαθμολόγησης των άκαμπτων υποδοχέων χρησιμοποιώντας τα ίδια σύνολα δεδομένων CHEBM που εφαρμόστηκαν και για τους ευέλικτους υποδοχείς. Τα ίδια σύνολα εφαρμόστηκαν για να γίνει μια σύγκριση αποτελεσμάτων μεταξύ άκαμπτων και ευέλικτων υποδοχέων.



Εικόνα 20 : αντιπροσωπευτικά σύνολα CHEMBL με διάγραμμα για κάθε ευέλικτο υποδοχέα.

Σύγκριση της συναινετικής βαθμολόγησης που χρησιμοποιεί βελτιστοποίηση συνάρτησης πολλαπλών στόχων με την καλύτερη ΣΒ και την ΣΒ της Vina η οποία χρησιμοποιήθηκε για την άκαμπτη βάση σύνδεσης υποδοχέα. Ένας εκπρόσωπος για κάθε περίπτωση υποδοχέα-στόχο παρουσιάζεται αναφέροντας την ταυτότητα της ανάλυσης CHEMBL.Οι μπλε μπάρες αντιστοιχούν σε συναινετική βαθμολόγηση, οι κίτρινες στην ΣΒ Vina ,οι κόκκινες στην ΣΒ με την υψηλότερο Top-Down Concordance,οι μωβ στην ΣΒ με το υψηλότερο Kendall's τ, οι πράσινες στην ΣΒ με την υψηλότερη Pearson 's R. Αν η ίδια ΣΒ είναι η καλύτερη για πολλαπλές μετρήσεις, εμφανίζονται λιγότερες μπάρες (π.χ. CP3A4).

Στον παρακάτω πίνακα παρουσιάζονται οι τιμές απόδοσης της ΣΒ Vina στη διασταυρωμένη επικύρωση(5-fold cross-validation) χρησιμοποιώντας δεδομένα από 18 διαφορετικές

δοκιμασίες πρόσδεσης, στους 8 διαφορετικούς υποδοχείς. Συγκρίνονται οι τιμές των τριών στατιστικών μεθόδων που αποτελούσαν την συνάρτηση πολλαπλών στόχων στην περίπτωση ευέλικτου υποδοχέα σύνδεσης και άκαμπτου υποδοχέα σύνδεσης .

Receptor	Assay	C(Flex)	C(Rigid)	τ(Flex)	τ(Rigid)	R(Flex)	R(Rigid)
akt1	CHEMBL1049250	0.1782	0.2975	0.3290	0.3223	0.4338	0.5100
akt1	CHEMBL889241	0.0527	0.1097	0.1780	0.1038	0.2364	0.1086
ampc	CHEMBL818874	0.0860	0.1473	0.0976	-0.0035	0.0701	0.0365
cp3a4	CHEMBL2060578	0.1303	0.1330	0.1759	0.1979	0.1249	0.1143
cxcr4	CHEMBL911131	-0.0225	0.4611	0.0397	0.1758	0.1156	0.2344
gcr	CHEMBL2025282	0.0769	0.2340	0.1959	0.1343	0.3819	0.1781
gcr	CHEMBL890119	0.3722	0.3406	0.3380	0.2156	0.4162	0.2605
gcr	CHEMBL941996	0.2761	-0.5080	0.1692	-0.5628	0.2709	-0.7616
gcr	CHEMBL950162	-0.3421	-0.3356	-0.3141	-0.2937	-0.3938	-0.4452
hivpr	CHEMBL763303	0.4788	0.5986	0.5258	0.5375	0.7782	0.6997
hivpr	CHEMBL765007	0.1308	0.0208	0.2131	0.1034	0.3186	0.1529
hivrt	CHEMBL802759	0.0720	0.0284	0.2181	0.1000	0.1121	0.0017
hivrt	CHEMBL803376	0.0016	-0.0143	-0.0262	-0.0195	-0.0236	-0.0526
kif11	CHEMBL1101849	-0.4031	-0.2928	-0.3284	-0.2197	-0.3925	-0.3409
kif11	CHEMBL1833735	0.1110	0.1010	0.0072	0.0804	0.2129	0.1177
kif11	CHEMBL2330556	0.4327	0.3466	0.2045	0.0807	0.2634	0.1250
kif11	CHEMBL895158	0.1309	0.1860	0.2387	0.0047	0.3551	0.0356
kif11	CHEMBL934296	-0.2269	-0.0849	-0.1180	-0.0437	-0.2315	-0.0465

Πίνακας 10: Αποτελέσματα ΣΒ Vina για ευέλικτο και μη υποδοχέα

Επίσης παρουσιάζεται η απόδοση της συναινετικής βαθμολόγησης στην περίπτωση ευέλικτου υποδοχέα σύνδεσης και άκαμπτου υποδοχέα σύνδεσης .

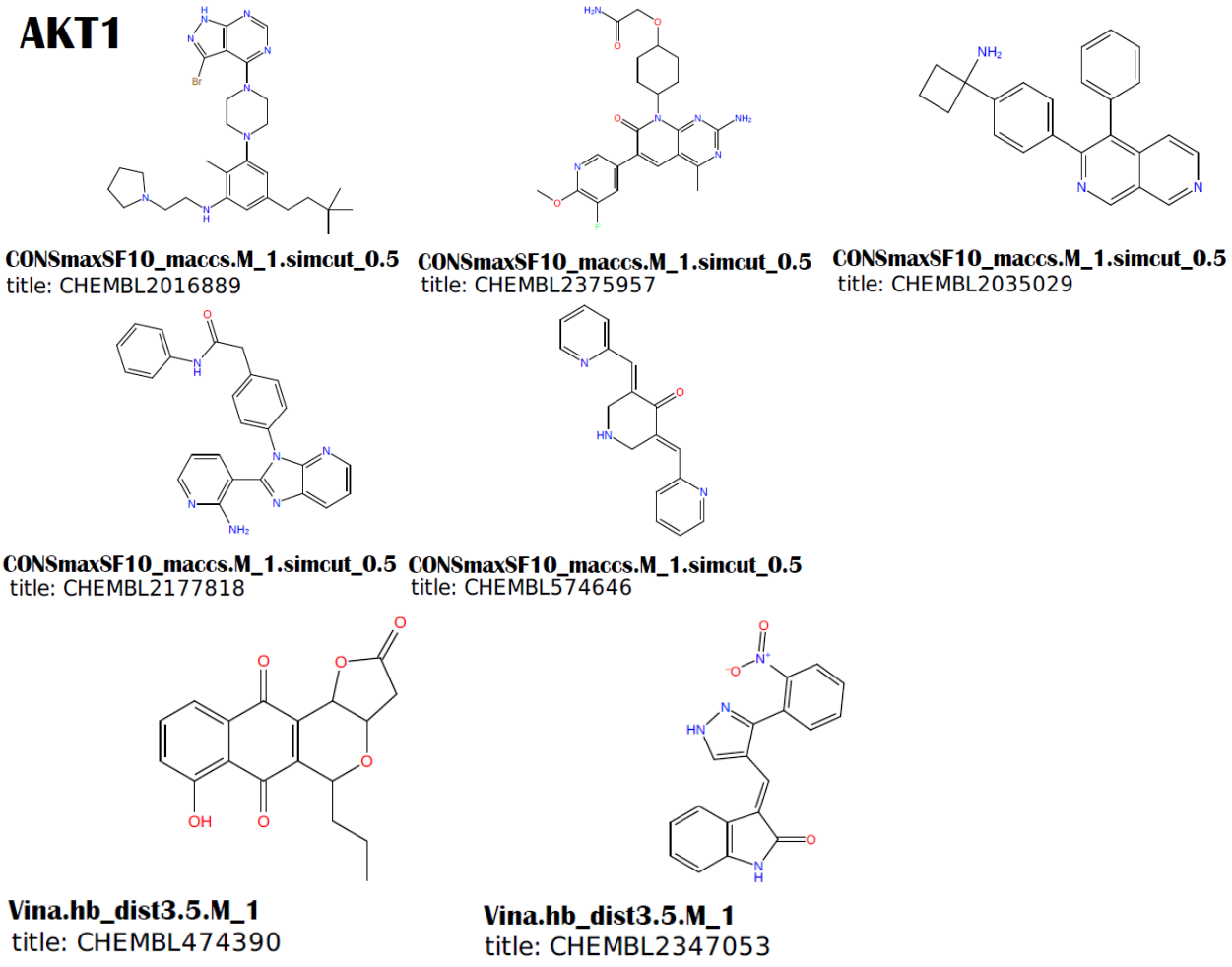
Receptor	Assay	C(Flex)	C(Rigid)	τ(Flex)	τ(Rigid)	R(Flex)	R(Rigid)
akt1	CHEMBL1049250	0.3122	0.4314	0.3272	0.3938	0.4813	0.5856
akt1	CHEMBL889241	0.4081	0.3893	0.3303	0.3424	0.3951	0.4729
ampc	CHEMBL818874	0.5569	0.4683	0.1072	0.3616	0.2119	0.4994
cp3a4	CHEMBL2060578	0.5077	0.4897	0.3658	0.3175	0.4184	0.3174
cxcr4	CHEMBL911131	0.8399	0.5011	0.4772	0.3210	0.6590	0.3749
gcr	CHEMBL2025282	0.6532	0.7086	0.4303	0.4391	0.5919	0.6788
gcr	CHEMBL890119	0.7196	0.4701	0.3863	0.4186	0.5697	0.4174
gcr	CHEMBL941996	0.5170	0.2057	0.3940	0.0586	0.4961	0.0204
gcr	CHEMBL950162	0.3892	0.4149	0.2867	0.1539	0.3531	0.0588
hivpr	CHEMBL763303	0.5242	0.6678	0.5463	0.5638	0.7712	0.7172
hivpr	CHEMBL765007	0.4560	0.3807	0.4176	0.2729	0.5036	0.3416
hivrt	CHEMBL802759	0.4140	0.4044	0.3534	0.2540	0.0218	0.2374
hivrt	CHEMBL803376	0.1935	0.3180	0.2171	0.2680	0.2860	0.2960
kif11	CHEMBL1101849	0.7365	0.7038	0.4972	0.4374	0.6553	0.6780
kif11	CHEMBL1833735	0.5349	0.4255	0.2557	0.3473	0.4225	0.4021
kif11	CHEMBL2330556	0.5941	0.3787	0.3231	0.3417	0.4241	0.4105

kif11	CHEMBL895158	0.6140	0.7287	0.3398	0.4083	0.4860	0.5318
kif11	CHEMBL934296	0.4192	0.5236	0.2792	0.3983	0.3353	0.5085

Πίνακας 11: Αποτελέσματα συναινετική βαθμολόγησης για ευέλικτο και μη υποδοχέα

3.4 Ομοιότητα Αποτυπώματος

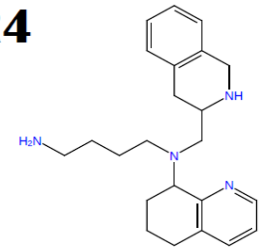
Αν και ο συνδυασμός των δισδιάστατης ομοιότητας δακτυλικών αποτυπώματών[8] και των ΣΒ τείνει να έχει καλύτερες επιδόσεις από το συνδυασμό των SIft[9] με τις ΣΒ, όπως απεικονίζεται στις καμπύλες ROC και CROC[5], οι δύο αυτές μέθοδοι έχουν τη δυνατότητα να ανακαλύψουν νέους συνδέτες με διαφορετικούς τρόπους. Ως εκ τούτου, κάθε μέθοδος συμπληρώνει την άλλη . Στο παρακάτω σχήμα, τα μη κοινά δραστικά μόρια μεταξύ εκείνων που προσδιορίζονται από κάθε μέθοδο κατηγοριοποιήθηκαν και παρουσιάζονται μόνο τα αντιπροσωπευτικά μόρια από το σύνολο τους στα οποία η μέση ανομοιότητα με όλα τα αντικείμενα της κάθε κατηγορίας είναι ελάχιστη (medoids). Έτσι στο σχήμα αυτό παρουσιάζεται η δυναμική τις κάθε μεθόδου στον εντοπισμό δομικά διαφορετικών συνδετών. Παρακάτω παρουσιάζονται τα μη κοινά δραστικά μορίων μεταξύ των δύο μεθόδων. Για κάθε υποδοχέα παρουσιάζεται η ΣΒ ή η συναινετική βαθμολόγηση που απέδωσε καλύτερα από ό, τι όλες οι υπόλοιπες.



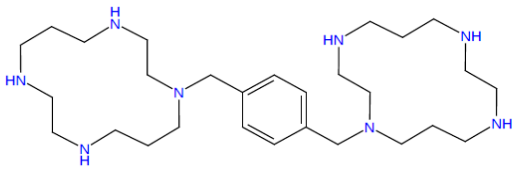
Εικόνα 21: AKT1 2DvsSiFt

Ανακαλύφθηκαν διαφορετικά μόρια από τον συνδυασμό της σύνδεσης (ΣΒ ή συναινετική ΣΒ) με τις δυο μεθόδους ομοιότητας (2D,SIFt) στο σύνολο ελέγχου AKT1. Το σύνολο Cons_10_maccs.M_1_0.5 βρέθηκαν 17 δραστικές ουσίες, ενώ το Vina_SiFt.hb_dist3.5.M_1 βρέθηκαν 5 δραστικές ενώσεις συνολικά.

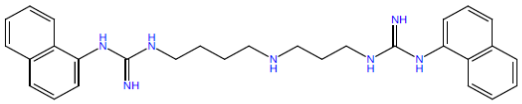
CXCR4



NNscore1_net1_dendritic.M_1.simcut_0.4
title: CHEMBL3091687



NNscore1_net1_dendritic.M_1.simcut_0.4
title: CHEMBL2311028

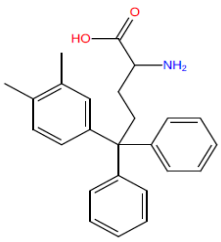


NNscore1_net1.hb_dist2.5M_1
title: CHEMBL2347626

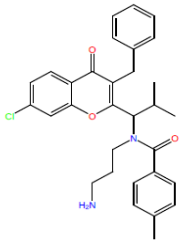
Εικόνα 22: CXCR4 2DvsSIFt

Ανακαλύφθηκαν διαφορετικά μόρια από τον συνδυασμό της σύνδεσης (ΣΒ ή συναινετική ΣΒ) με τις δυο μεθόδους ομοιότητας (2D,SIFt) στο σύνολο ελέγχου CXCR4.Στο σύνολο NN1_1_SiFt.hb_dist2.5.M_1 βρέθηκαν 4 δραστικές ουσίες από την πισίνα, ενώ NN1_1_dendritic.M_1_0.4 βρέθηκαν 3 δραστικές ουσίες συνολικά .

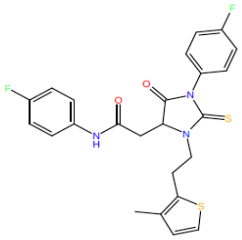
KIF11



Vina_maccs.M_3.simcut_0.5
title: CHEMBL2325418



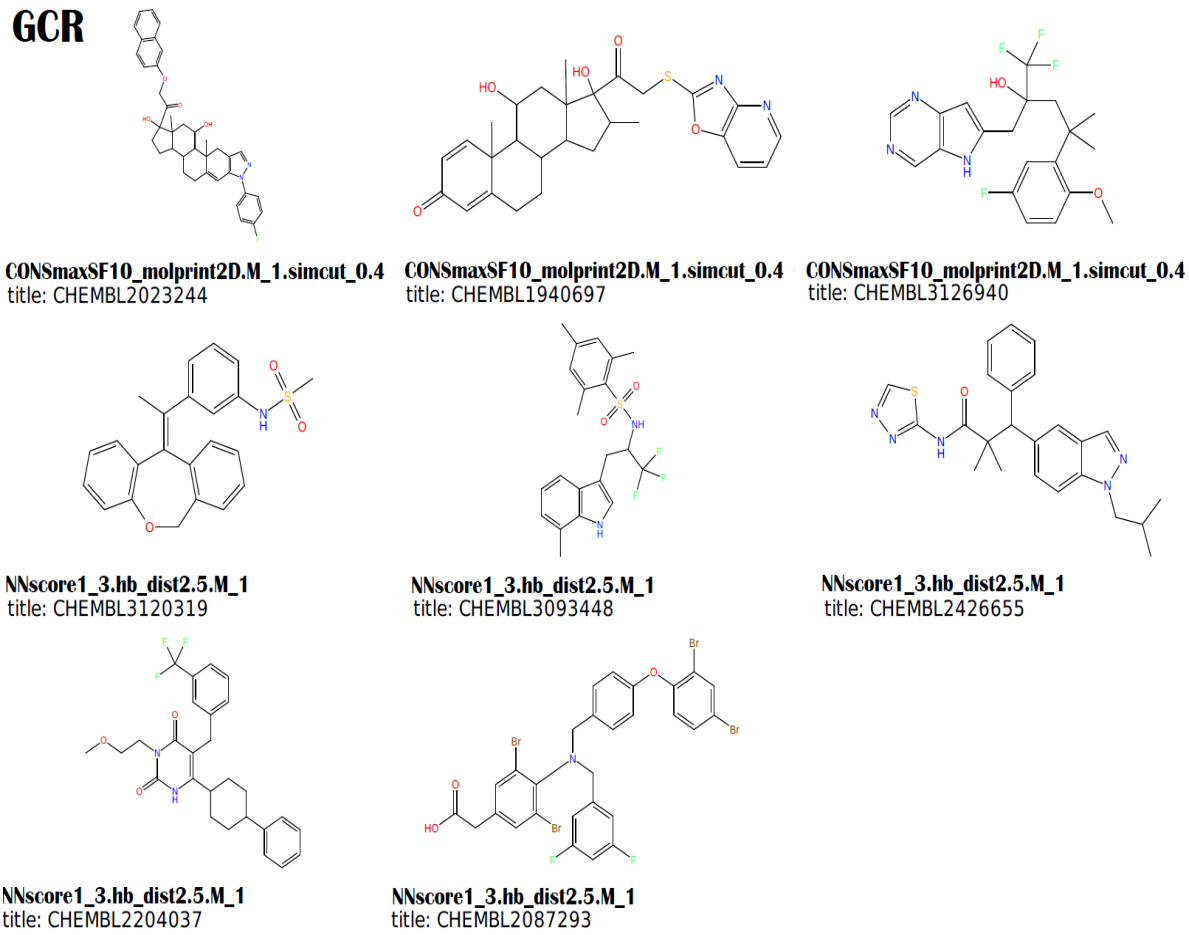
Vina.hb_dist2.5.M_1
title: CHEMBL2325429



Vina.hb_dist2.5.M_1
title: CHEMBL2031570

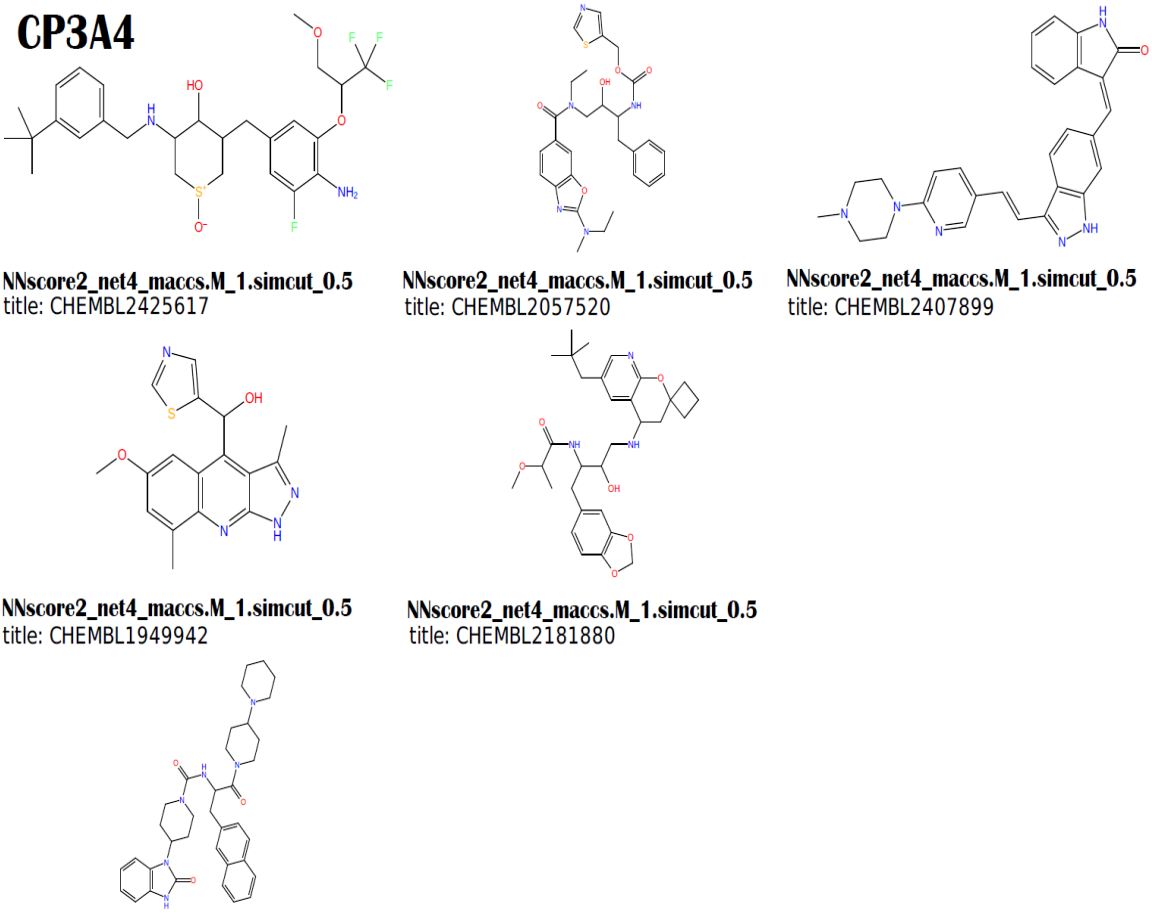
Εικόνα 23: KIF11 2DvsSift

Ανακαλύφθηκαν διαφορετικά μόρια από τον συνδυασμό της σύνδεσης (ΣΒ ή συναινετική ΣΒ) με τις δυο μεθόδους ομοιότητας (2D,SIFt) στο σύνολο ελέγχου KIF11. Στο σύνολο Vina_maccs.M_3.simcut_0.5 βρέθηκε μία ενεργή ουσία, ενώ στο Vina_SiFt.hb_dist2.5.M_1 βρέθηκαν 3 ενεργές ουσίες.



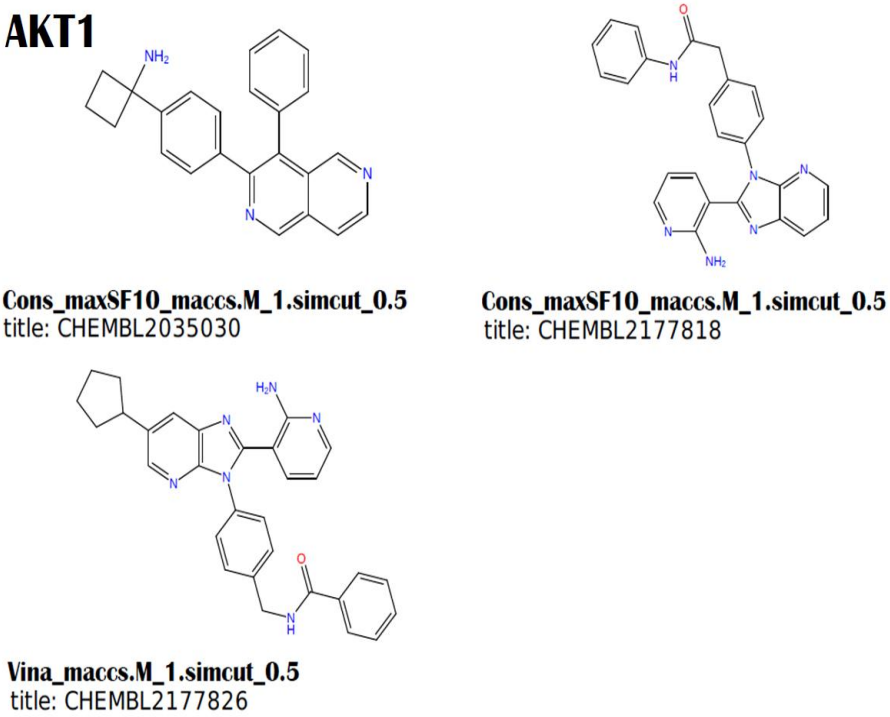
Εικόνα 24: GCR 2DvsSIFt

Ανακαλύφθηκαν διαφορετικά μόρια από τον συνδυασμό της σύνδεσης (ΣΒ ή συναινετική ΣΒ) με τις δυο μεθόδους ομοιότητας (2D,SIFt) στο σύνολο ελέγχου CP3A4 .Στο σύνολο NN2_4_maccs.M_1 0.5 βρέθηκαν 30 δραστικές ουσίες, ενώ στο NN2_4_SiFt.hb_dist2.5.M_1 βρέθηκαν 5 δραστικά συνολικά.



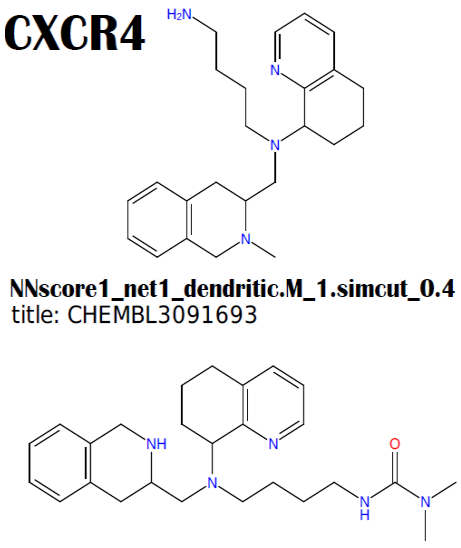
Εικόνα 25: CP3A4 2DvsSIFt

Ανακαλύφθηκαν διαφορετικά μόρια από τον συνδυασμό της σύνδεσης (ΣΒ ή συναινετική ΣΒ) με τις δυο μεθόδους ομοιότητας (2D,SIFt) στο σύνολο ελέγχου GCR. Στο σύνολο Cons_10_molprint2D.M_1_0.4 βρέθηκαν 17 δραστικές ουσίες, ενώ στο NN1_3_SiFt.hb_dist2.5.M_1 16 δραστικές ουσίες συνολικά.



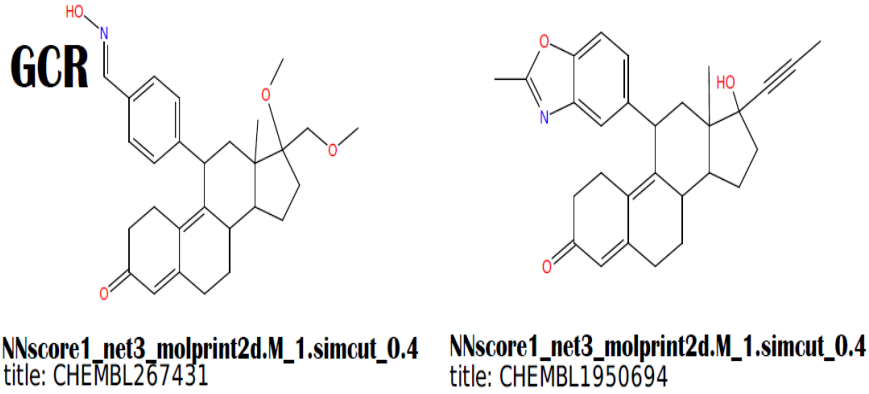
Εικόνα 26: AKT1 SFvsCons

Ανακαλύφθηκαν διαφορετικά μόρια από τον συνδυασμό μιας μόνο ΣΒ με την δισδιάστατη μέθοδο ομοιότητας δακτυλικών αποτυπωμάτων και της συναινετική βαθμολόγηση με την δισδιάστατη μέθοδο ομοιότητας δακτυλικών αποτυπωμάτων στο σύνολο ελέγχου AKT1. Στο σύνολο Cons_10_maccs.M_1_0.5 βρέθηκαν 6 δραστικές ουσίες ενώ στο Vina_maccs.M_1_ 0.5 βρέθηκε μία δραστική ουσία.



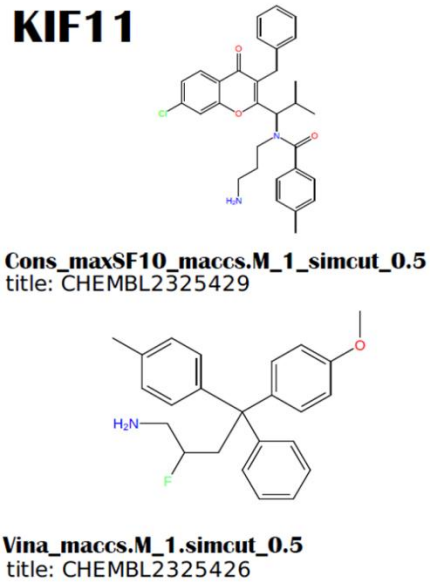
Εικόνα 27: CXCR4 SFvsCons

Ανακαλύφθηκαν διαφορετικά μόρια από τον συνδυασμό μιας μόνο ΣΒ με την δισδιάστατη μέθοδο ομοιότητας δακτυλικών αποτυπωμάτων και της συναινετική βαθμολόγηση με την δισδιάστατη μέθοδο ομοιότητας δακτυλικών αποτυπωμάτων στο σύνολο ελέγχου CXCR4 .Στο σύνολο NN1_1_dendritic.M_1_0.4 βρέθηκαν 3 δραστικές ουσίες, ενώ στο Cons_10_maccs.M_1_0.5 βρέθηκαν 2.



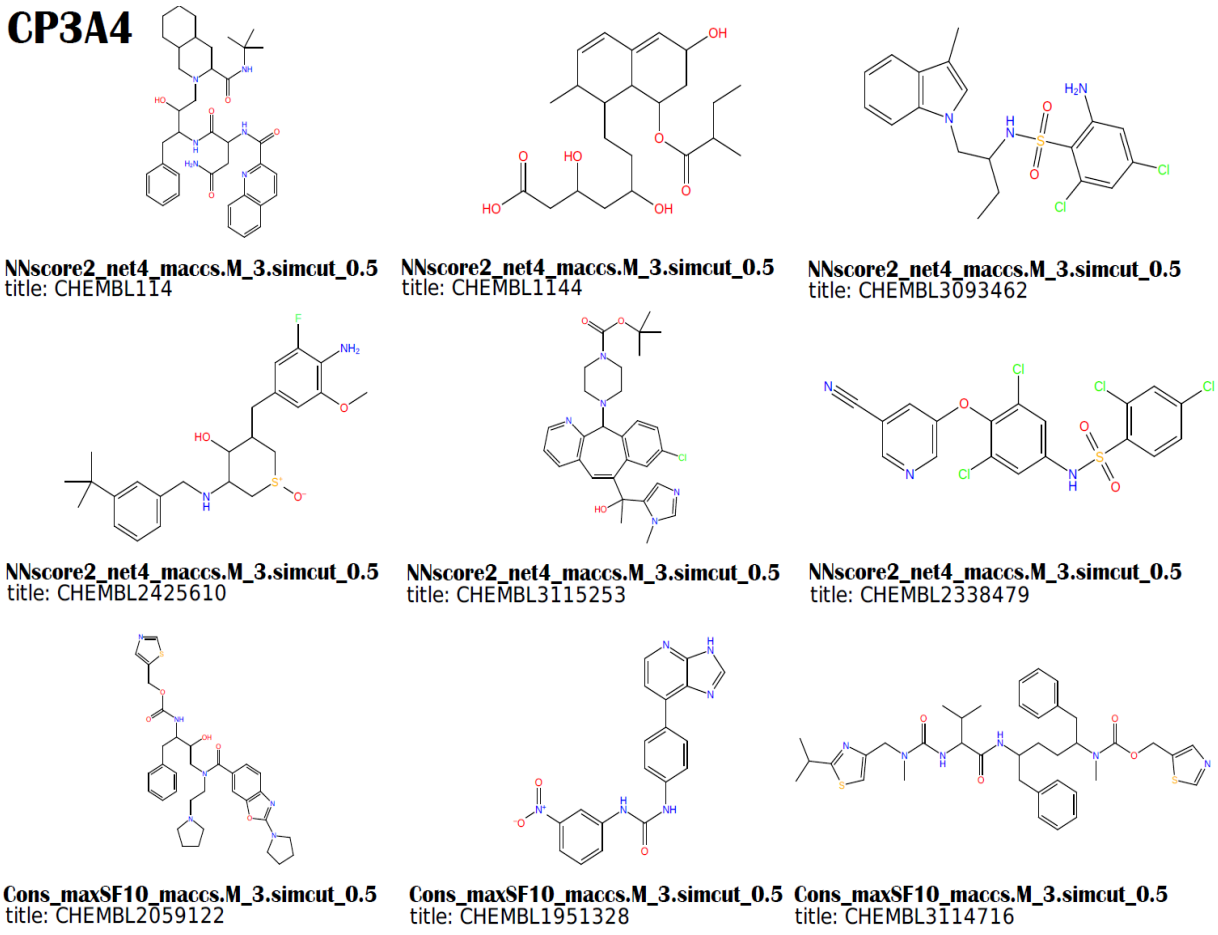
Εικόνα 28: GCR SFvsCons

Ανακαλύφθηκαν διαφορετικά μόρια από τον συνδυασμό μιας μόνο ΣΒ με την δισδιάστατη μέθοδο ομοιότητας δακτυλικών αποτυπωμάτων και της συναινετική βαθμολόγηση με την δισδιάστατη μέθοδο ομοιότητας δακτυλικών αποτυπωμάτων στο σύνολο ελέγχου GCR.Στο σύνολο NN1_3_molprint2D.M_1_0.4 βρέθηκαν 3 δραστικές ουσίες, ενώ στο Cons_10_molprint2D.M_1_0.4 δεν βρέθηκε καμία.



Εικόνα 29: KIF11 SFvsCons

Ανακαλύφθηκαν διαφορετικά μόρια από τον συνδυασμό μιας μόνο ΣΒ με την δισδιάστατη μέθοδο ομοιότητας δακτυλικών αποτυπωμάτων και της συναινετική βαθμολόγηση με την δισδιάστατη μέθοδο ομοιότητας δακτυλικών αποτυπωμάτων στο σύνολο ελέγχου KIF11.Στο σύνολο Cons_10_maccs.M_1_0.5 βρέθηκε μία ενεργή ουσία, ενώ στο Vina_maccs.M_1_0.5 βρέθηκαν 2 .



Εικόνα 30: CP3A4 SFvsCons

Ανακαλύφθηκαν διαφορετικά μόρια από τον συνδυασμό μιας μόνο ΣΒ με την διδιάστατη μέθοδο ομοιότητας δακτυλικών αποτυπωμάτων και της συναινετική βαθμολόγηση με την δισδιάστατη μέθοδο ομοιότητας δακτυλικών αποτυπωμάτων στο σύνολο ελέγχου AKT1. Στο σύνολο Cons_10_maccs.M_1_0.5 βρέθηκαν 6 δραστικές ουσίες ενώ στο Vina_maccs.M_1_ 0.5 βρέθηκε μία δραστική ουσία.

Στους παρακάτω πίνακες παρουσιάζονται οι τιμές των ROC, CROC , BEDROC για τα δοκιμαστικά σύνολα που χρησιμοποιήθηκαν στη διαδικασία ομοιότητας δακτυλικών αποτυπωμάτων καθώς αναφέρονται και ο αριθμός των δραστικών και μη δραστικών ενώσεων.

AKT1	ROC	CROC	BEDROC	ACTIVES	DECOYS
Cons_5_maccs.M_3_0.5	0.8476	0.5616	0.3796	10	1152
Vina_maccs.M_3_0.5	0.9061	0.5177	0.1539	2	1219
Cons_10_maccs.M_3_0.5	0.8196	0.2822	0.0273	2	988
Vina	0.6068	0.1808	0.0741	88	6477
Cons_max10	0.6407	0.2453	0.1079	88	6477
Cons_max5	0.6253	0.1983	0.0769	88	6477
Vina_SiFt.hb_dist2.5.M_1	0.6378	0.1863	0.044	4	468
Vina_SiFt.hb_dist3.5.M_1	0.6642	0.1816	0.0512	6	616
Vina_SiFt.hb_dist2.5.M_3	0.5789	0.0698	0.0015	2	95

Πίνακας 12: Εμβαδόν κάτω από τις καμπύλες της AKT1 για το δοκιμαστικό σύνολο

CXCR4	ROC	CROC	BEDROC	ACTIVES	DECOYS
NN1_1_dendritic.M_1_0.4	0.9636	0.7899	0.6415	10	188
NN1_1_maccs.M_1_0.5	0.9421	0.6849	0.446	13	474
Cons_10_maccs.M_1_0.5	0.7462	0.3272	0.1457	9	225
Cons_5_maccs.M_1_0.5	0.7051	0.2998	0.1011	8	228
NN1_1	0.954	0.7424	0.5166	19	1105
Cons_max10	0.7027	0.3087	0.1409	19	1105
Cons_max5	0.688	0.3124	0.158	19	1105
NN1_1_maccs.M_3_0.6	0.9778	0.8586	0.6756	4	248
Cons_5_maccs.M_3_0.6	0.7827	0.3281	0.1835	8	108
Cons_10_maccs.M_3_0.6	0.7296	0.193	0.0294	7	84
NN1_1_SiFt.hb_dist2.5.M_1	0.9642	0.792	0.5911	11	665
NN1_1_SiFt.hb_dist3.5.M_1	0.9641	0.7914	0.5902	11	661
NN1_1_SiFt.hb_dist2.5.M_3	0.9438	0.6929	0.4349	8	534
NN1_1_SiFt.hb_dist3.5.M_3	0.9439	0.6932	0.4348	8	533

Πίνακας 13: Εμβαδόν κάτω από τις καμπύλες του υποδοχέα CXCR4 για το δοκιμαστικό σύνολο

GCR	ROC	CROC	BEDROC	ACTIVES	DECOYS
Cons_10_molprint2D.M_1_0.4	0.8066	0.9406	0.6677	32	1902
Cons_5_molprint2D.M_1_0.4	0.969	0.8261	0.6703	32	1726
NN1_3_molprint2D.M_1_0.4	0.964	0.8144	0.6384	36	2848
NN1_3	0.8346	0.4729	0.2776	187	9795
Cons_max10	0.7328	0.3866	0.2479	187	9795
Cons_max5	0.731	0.3862	0.2476	187	9795
NN1_3_molprint2D.M_3_0.4	0.9863	0.9106	0.7801	8	1861
Cons_10_molprint2D.M_3_0.4	0.9751	0.8548	0.7071	9	1000
Cons_5_molprint2D.M_3_0.4	0.975	0.8535	0.7039	9	887
NN1_3_SiFt.hb_dist2.5.M_1	0.8559	0.5512	0.38	20	1275
NN1_3_SiFt.hb_dist3.5.M_1	0.8396	0.4825	0.2914	25	1274

Πίνακας 14: Εμβαδόν κάτω από τις καμπύλες του υποδοχέα της GCR για το δοκιμαστικό σύνολο

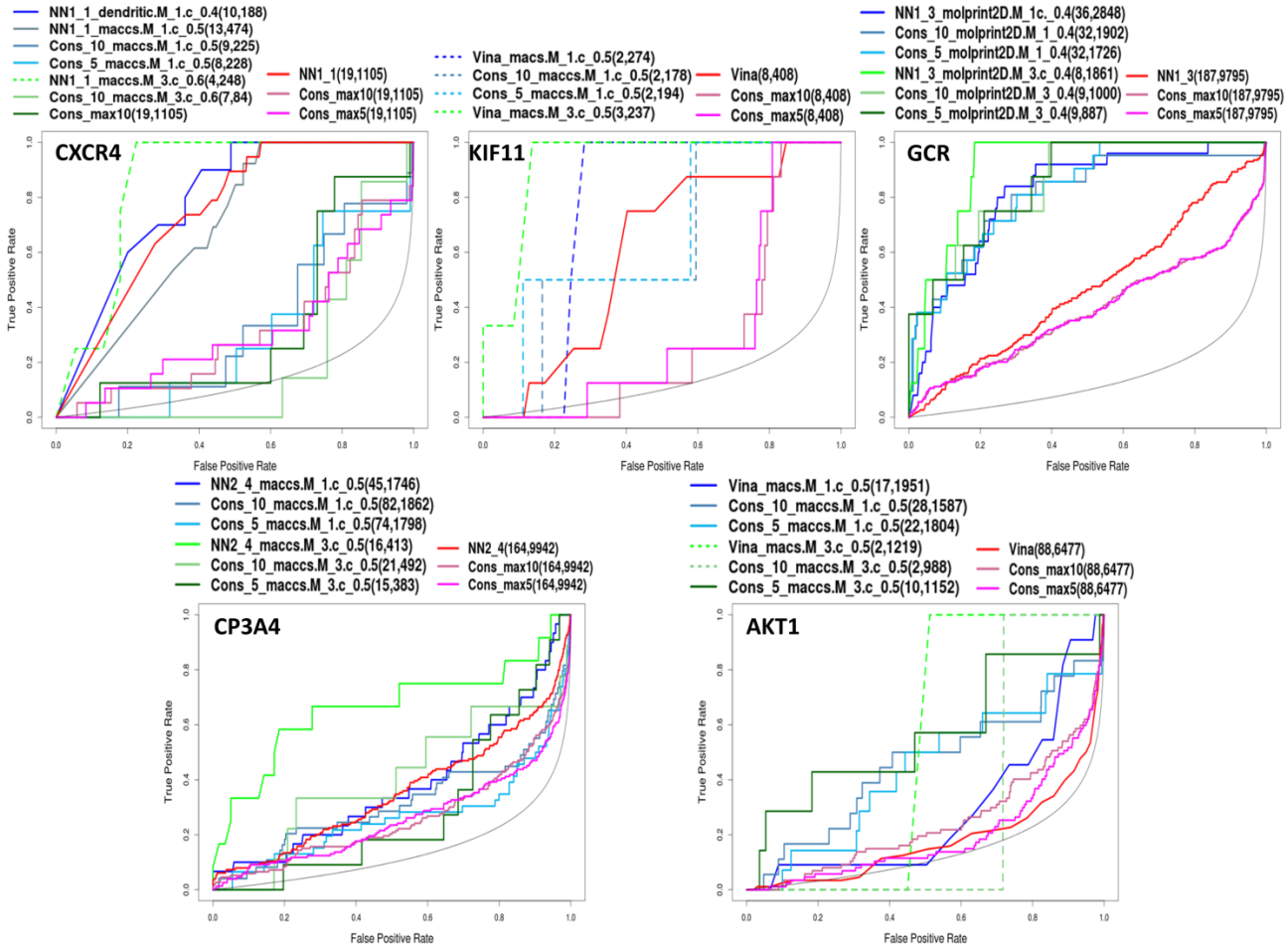
CP3A4	ROC	CROC	BEDROC	ACTIVES	DECOYS
NN2_4_maccs.M_1_0.5	0.7969	0.3756	0.211	45	1746
Cons_10_maccs.M_1_0.5	0.6714	0.3151	0.2108	82	1862
Cons_5_maccs.M_1_0.5	0.653	0.2643	0.1735	74	1798
NN2_4	0.7343	0.3525	0.2069	164	9932
Cons_max10	0.628	0.2612	0.1453	164	9932
Cons_max5	0.6173	0.2566	0.1463	164	9932
NN2_4_maccs.M_3_0.5	0.8973	0.6625	0.5697	16	413
Cons_10_maccs.M_3_0.5	0.7284	0.4005	0.2281	21	492
Cons_5_maccs.M_3_0.5	0.7726	0.2874	0.1067	15	383
NN2_4_SiFt.hb_dist2.5.M_1	0.8205	0.4817	0.3345	5	117
NN2_4_SiFt.hb_dist3.5.M_1	0.7761	0.3944	0.2815	5	117

Πίνακας 15: Εμβαδόν κάτω από τις καμπύλες του CP3A4 για το δοκιμαστικό σύνολο

KIF11	ROC	CROC	BEDROC	ACTIVES	DECOYS
Vina_maccs.M_1.simcut_0.5	0.9589	0.7501	0.4446	2	274
Cons_5_maccs.M_1_0.5	0.9298	0.6545	0.4191	2	194
Cons_10_maccs.M_1_0.5	0.9227	0.62	0.3525	2	178
Vina	0.6068	0.1808	0.0741	88	6477
Cons_max10	0.8082	0.2875	0.0604	8	408
Cons_max5	0.818	0.3134	0.0854	8	408
Vina_maccs.M_3.simcut_0.5	0.9887	0.9257	0.8206	3	237
Vina_SiFt.hb_dist2.5.M_1	0.8533	0.4522	0.2144	5	135
Vina_SiFt.hb_dist3.5.M_1	0.8267	0.4272	0.2104	4	132
Vina_SiFt.hb_dist2.5.M_3	0.8397	0.4018	0.215	2	39
Vina_SiFt.hb_dist3.5.M_3	0.9009	0.5381	0.2943	2	5

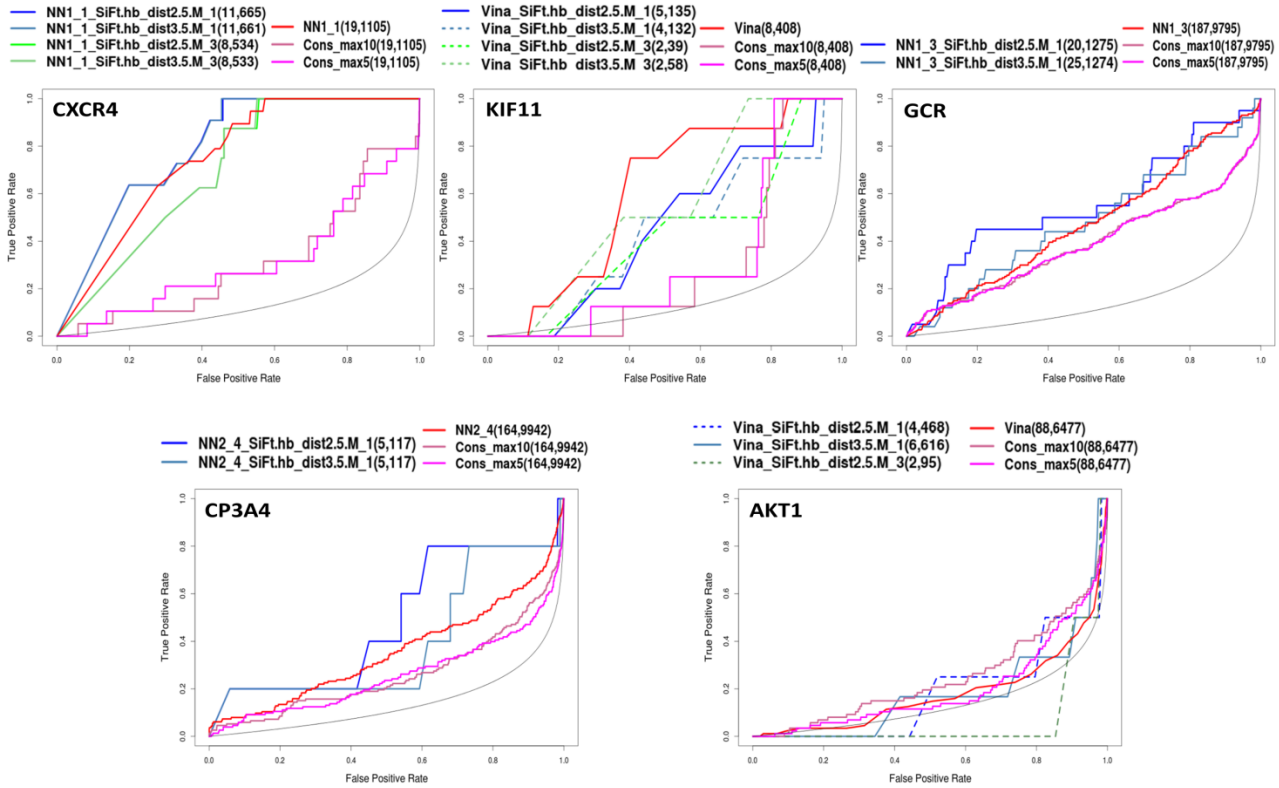
Πίνακας 16: Εμβαδόν κάτω από τις καμπύλες του KIF11 για το δοκιμαστικό σύνολο

Τα συγκεντρωτικά διαγράμματα των καμπύλων ROC ,CROC ,BEDROC για τα δοκιμαστικά σύνολα παρουσιάζονται παρακάτω.



Εικόνα 31: Καμπύλες CROC ατομικών ΣΒ ή συναινετικής βαθμολόγησης, σε συνδυασμό ή όχι με δισδιάστατα δακτυλικά αποτυπώματα ομοιότητας

Σύγκριση των καλύτερων ατομικών ΣΒ ή της συναινετικής βαθμολόγησης, σε συνδυασμό ή όχι με δισδιάστατα δακτυλικά αποτυπώματα ομοιότητας, στο τυφλό σύνολο δοκιμών: Χρησιμοποιήθηκαν μόνο οι ενώσεις έχουν ανατεθεί ως "υψηλής" εμπιστοσύνης από κάθε μέθοδο, και το αντίστοιχο συνόλων δεδομένων αναφέρεται σε παρένθεση (αριθμός συνδετών, αριθμός μη-συνδετών) μετά από την ονομασία της κάθε μεθόδου. Οι καμπύλες των μεθόδων με «υψηλό» επίπεδο εμπιστοσύνης για λιγότερο από 5 συνδετικά μόρια ήταν περιορισμένης στατιστικής σημασίας και ως εκ τούτου έχουν σημανθεί με διακεκομμένες γραμμές. Η ονοματολογία που ακολουθεί τη σύμβαση είναι η εξής: <συνάρτηση βαθμολόγησης> _ <τύπος δακτυλικών αποτυπωμάτων> .M_ <τιμή M> .c_ <ομοιότητα αποκοπής>, όπου M είναι ο αριθμός του παρόμοιου συνόλου εκπαίδευσης .



Εικόνα 32: Καμπύλες CROC των καλύτερων ατομικών ΣΒ ή συναινετικής βαθμολόγησης, σε συνδυασμό ή όχι με δακτυλικά αποτυπώματα αλληλεπίδραση δομής(SIFT)

Ακριβώς ή ίδια με την προηγούμενη διαδικασία χρησιμοποιήθηκε για τη σύγκριση των καλύτερων ατομικών ΣΒ ή της συναινετικής βαθμολόγησης, σε συνδυασμό ή όχι με δακτυλικά αποτυπώματα αλληλεπίδραση δομής(SIFT), στο τυφλό σύνολο δοκιμών .

Για την σειρά δοκιμών τυφλής επικύρωσης αξιολογήσαμε την απόδοση των καλύτερων σε βαθμολογία ΣΒ και της συναινετικής βαθμολόγησης στο σύνολο εκπαίδευσης. Επίσης ενσωματώνονται διάφορες μέθοδοι ομοιότητας δακτυλικών αποτυπωμάτων στην βαθμολογία. Στις εικόνες 31 και 32 χρησιμοποιήθηκε η AUC-CROC για να συγκριθεί η απόδοση των καλύτερων ΣΒ καθώς σε συνδυασμό με δύο τύπους των δακτυλικών αποτυπωμάτων: δισδιάστατης ομοιότητας δακτυλικά αποτυπώματα και αλληλεπίδραση δομής (SIFt),αντίστοιχα. Οι τιμές του AUC-CROC αυτών των καμπυλών, αναφέρονται στους πίνακες 5 έως 9. Συνολικά, έχουμε ελέγξει την απόδοση 5 δισδιάστατων δακτυλικών αποτυπωμάτων ομοιότητας (linear, maccs, radial , dendritic ,molprint2D). Εμφανίζονται μόνο τα αποτελέσματα από τα καλύτερα σε κάθε περίπτωση.

Σύμφωνα με την Εικόνα 31 οι προσεγγίσεις συναινετικής βαθμολόγησης τείνουν να έχουν τη χειρότερη απόδοση, εκτός από την περίπτωση της AKT1 όπου έχει σχετικά καλύτερη βαθμολογία από την καλύτερη ατομική ΣΒ (Vina σε αυτή την περίπτωση). Σε αντίθεση, ο συνδυασμός δισδιάστατων δακτυλικών αποτυπωμάτων ομοιότητας είτε με την καλύτερη ΣΒ ή με συναινετική βαθμολόγηση ενισχύει περαιτέρω την αποτελεσματικότητα της ανακάλυψης νέων συνδετών. Εκτός από την περίπτωση του CXCR4 , όπου η καλύτερη ατομική ΣΒ (NNscore1_net1) είναι σχεδόν τέλεια , σε όλες τις άλλες περιπτώσεις, η εισαγωγή των δακτυλικών αποτυπωμάτων ομοιότητας στην βαθμολόγηση ξεπερνά κάθε ΣΒ ή συναινετική βαθμολόγηση. Τέλος , αν και οι ατομικές ΣΒ έχουν την τάση να είναι πιο αξιόπιστες από ότι η συναινετική ΣΒ, όταν συνδυάζονται με δισδιάστατο δακτυλικό

αποτύπωμα ομοιότητας , αυτές οι δύο μέθοδοι έχουν την τάση να προσδιορίζουν συνδέτες με διαφορετικά ικριώματα από το σύνολο των δραστικών μορίων .

Σε αντίθεση με την δισδιάστατη δομή ομοιότητας, η ομοιότητα SIFt δεν βελτίωσε την απόδοση της κορυφαία ΣΒ (Εικόνα 32.Οι καμπύλες με χρώμα ροζ, πορφυρό και κόκκινο είναι ακριβώς το ίδιες όπως στην Εικόνα 31). Υπάρχουν περιπτώσεις, όπως αυτή του GCR και του CP3A4 όπου η ομοιότητα SIFt έχει οριακά καλύτερη απόδοση από τις πρώτους υπολογισμούς σύνδεσης. Θα πρέπει επίσης να σημειωθεί ότι η συνολική απόδοση του Sift είναι χειρότερη από ότι αυτή της δισδιάστατης ομοιότητα όπως μπορούμε να δούμε στους Πίνακες 5-9. Αυτά τα δύο στοιχεία υπογραμμίζουν για μια ακόμη φορά το πρόβλημα της σωστής πρόβλεψης ποζών σύνδεσης τόσο της ΣΒ Vina όσο και του Glide .

Τέλος, τόσο από διασταυρούμενη επικύρωση όσο και από το τυφλό σύνολο δοκιμών, καταλήγουμε στο συμπέρασμα ότι η ενσωμάτωση άνω των 5 επιμέρους ΣΒ στην εξίσωση συναινετικής βαθμολόγησης δεν βελτιώνει την απόδοση και προκαλεί μια σημαντική επιβράδυνση της ταχύτητας.

Κεφάλαιο 4

4 Συζήτηση Αποτελεσμάτων

Παρόλο που τα σύνολα ΣΒ NNscore 1.0 και 2.0 NNscore[3] αναπτύχθηκαν χρησιμοποιώντας τη ΣΒ Vina[13], η μέση απόδοση της σύνδεσης δεν βελτιώθηκε ούτε επιδεινώθηκε όταν επανεκτιμήθηκαν η πόζες που προέκυψαν από την ΣΒ Glide[1]. Αυτό επαληθεύεται από τα αποτελέσματα της μελέτης αυτής, δεδομένου ότι μόνο κατά το ήμισυ των περιπτώσεων η Vina είχε πόζες σύνδεσης με καλύτερες τιμές στατιστικών δεικτών. Ως εκ τούτου, η επιλογή της κατάλληλης μεθόδου εξαρτάται από το σύστημα .

Σε γενικές γραμμές υπάρχει μια τάση για την τιμή της παράστασης $\frac{\mathrm{TP}}{(\mathrm{TP+FP})}$ (εξειδίκευση,TP: αληθώς θετικά, FP: ψευδώς θετικά) να μειώνεται όσο μειώνεται η ελάχιστη βαθμολογία σύνδεσης (υποθέτοντας ότι όσο χαμηλότερη είναι η βαθμολογία της σύνδεσης τόσο καλύτερη είναι η σύνδεση). Κατά συνέπεια, το ποσοστό ψευδώς θετικών στοιχείων (FPR = 1- $\frac{\mathrm{TP}}{(\mathrm{TP+FP})}$) αυξάνει καθώς η ελάχιστη βαθμολογία σύνδεσης μειώνεται. Ως εκ τούτου, οι χαμηλές τιμές του x-άξονα (FPR) της ROC που μεγεθύνονται στην καμπύλη CROC αντιστοιχούν στην απόδοση της ΣΒ με υψηλή βαθμολογία σύνδεσης. Αυτό το αποτέλεσμα, επίσης γνωστό ως το "πρόβλημα έγκαιρης αναγνώρισης", είναι υψίστης σημασίας στο σχεδιασμό φαρμάκων, όπου οι περιορισμοί κόστους εξαναγκάζουν τους χημικούς να δοκιμάζουν μόνο ένα μικρό αριθμό των ενώσεων υψηλής βαθμολογίας . Είναι λοιπόν ιδιαίτερα χρήσιμη η ανάπτυξη μιας ΣΒ που θα μπορούσε να ταξινομήσει σωστά τις κορυφαίες βαθμολογικά δραστικές ενώσεις. Για το σκοπό αυτό, εφαρμόστηκαν ταξινομητές δύο κατηγοριών που ελαχιστοποιούν αυτό το πρόβλημα, τις καμπύλες CROC και BEDROC. Σύμφωνα με την τυφλή δοκιμαστική επαλήθευση μου εφαρμόστηκε, η συναινετική βαθμολόγηση ήταν πολύ καλύτερη σε επιδόσεις στις περισσότερες περιπτώσεις από ότι μια απλή ΣΒ.

Αυτό συνέβη για τους εξής λόγους:

1) τόσο το σύνολο εκπαίδευσης όσο και στο δοκιμαστικό σύνολο στοιχείων περιείχαν πολλά υποθετικά ανενεργά στοιχεία (decoys), δεδομένου ότι τα τελευταία συνήθως δεν δημοσιεύονται.

2) μια ΣΒ χρησιμοποιεί μόνο το καλύτερα βαθμολογημένο ταυτομερές της κάθε ένωσης, ενώ η συναινετική βαθμολόγηση το ταυτομερές που έχει τη μεγαλύτερη βαθμολογία από τις περισσότερες από τις ΣΒ που χρησιμοποιούνται για την διαδικασία βελτιστοποίησης.

3) Λάθος πόζες σύνδεσης. Δυστυχώς, τόσο Vina όσο και το Glide, σε πολλές περιπτώσεις δεν ήταν σε θέση να βρουν τη σωστή δεσμευτική στάση του κρυσταλλικού προσδέτη. Μπορεί να υποστηριχθεί ότι μια συναινετική βαθμολόγηση πρέπει να αποτελείται από ΣΒ που τείνουν να κατατάξουν την ίδια πόζα πρώτη. Αν και αυτό ακούγεται σαν προϋπόθεση, σε περιπτώσεις σύνδεσης υποστρώματος καθώς και σε περιπτώσεις σύνδεσης μιας μη δραστικής ένωσης δεν υπάρχει πάντα μία μοναδική καλύτερη πόζα. Δεδομένου ότι η πλειοψηφία του συνόλου εκπαίδευσης ήταν μη δραστικές ενώσεις, δεν είχαμε την δυνατότητα ελέγχου της συνοχής μεταξύ των καλύτερων ποζών σύνδεσης καθεμίας από

τις ΣΒ σε μία εξίσωση συναινετικής βαθμολόγησης. Ίσως αυτό να αποτελεί ελάττωμα σχεδιασμού στον αλγόριθμό και θα είναι ένα αντικείμενο μελλοντικών ερευνών.

Δυστυχώς, η τυφλή δοκιμαστική επαλήθευση δεν κατέστη δυνατό να εφαρμοστεί στην πιο ελπιδοφόρα περίπτωση των πολλαπλών στόχων βελτιστοποίησης, καθώς τα δεδομένα που χρησιμοποιούνται θα πρέπει να προέρχονται από το τα ίδια σύνολα με τιμές σύνδεσης, όπως εκείνες που χρησιμοποιήθηκαν για την εξίσωση συναινετικής βαθμολόγησης. Οι μελλοντικές προσπάθειες προσανατολίζονται στην επαλήθευση αυτής της μεθόδου με τη χρήση άλλων υποδοχέων για τους οποίους υπάρχουν περισσότερο "ομοιογενή" δεδομένα.

Πιστεύουμε ότι η συναινετική βαθμολόγηση θα μπορούσε να έχει ευρύτερη εφαρμογή βελτιστοποίηση προβλήματα καθοδήγησης. Η πρόβλεψη της σωστής κατάταξης της συγγένειας δέσμευσης για νέα μόρια είναι πιο σημαντική στον τομέα της ανακάλυψης φαρμάκων από τον απλό διαχωρισμό συνδετών από μη-συνδέτες. Έχοντας μια συνάρτηση βαθμολόγησης που θα μπορούσε να ταξινομήσει γρήγορα και με ακρίβεια μερικές εκατοντάδες μορίων θα μπορούσε δυνητικά να μειώσει το υπολογιστικό κόστος της πιο χρονοβόρας μεθόδου υπολογισμού ελεύθερης ενέργειας.

Λογισμικό

Για την εκπόνηση αυτής της εργασίας χρησιμοποιήθηκαν αρκετά πακέτα λογισμικού τα οποία χρησιμοποιήθηκαν σε περιβάλλον Linux :

Pymol : Το pymol είναι ένα σύστημα μοριακής απεικόνιση. Μέσω αυτού του συστήματος χρησιμοποιήθηκε το **AutoDock Vina Pymol plugin** ένα πρόγραμμα ανοιχτού κώδικα το ο οποίο υπολογίζει την μοριακή πρόσδεση . Σχεδιάστηκε και υλοποιήθηκε από τον Dr. Oleg Trott στη Μοριακό εργαστήριο γραφικών Scripps Research Institute. Τα αρχεία χρήσης διατίθενται στη διεύθυνση http://vina.scripps.edu/. Ενώ το σύστημα μοριακής απεικόνισης pymol διατίθεται στη διεύθυνση https://www.pymol.org/

Maestro : Ένα ισχυρό, περιβάλλον μοριακής μοντελοποίησης για ακαδημαϊκή χρήση το οποίο διατίθεται στη διεύθυνση http://www.schrodinger.com/Maestro/

R–Studio : Το R-Studio είναι ένα ελεύθερο περιβάλλον λογισμικού για στατιστικούς υπολογισμούς και γραφικά . Το περιβάλλον αυτό χρησιμοποιήθηκε για την εξαγωγή των γραφικών παραστάσεων της μελέτης .Μπορεί να βρεθεί στη διεύθυνση http://www.rstudio.com/ .

Χρησιμοποιήθηκαν επίσης χρήσιμα πακέτα της R , όπως το **ggplot2 , CROC-package**.

SPSS : Το SPSS (Superior Performance Software System) είναι ένα διαδεδομένο πρόγραμμα για τη στατιστική ανάλυση δεδομένων της IBM . Το πρόγραμμα διατίθεται στην ηλεκτρονική διεύθυνση http://www-01.ibm.com/software/analytics/spss/

Βιβλιογραφία

- [1] Friesner RA1, B. J. (2004). *Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.* New York.
- [2] Gerd Neudert, G. K. (n.d.). *DSX: A Knowledge-Based Scoring Function for the Assessment of Protein Ligand Complexes.* Department of Pharmaceutical Chemistry, Philipps-Universit€at Marburg, Marbacher Weg 6, D-35032, Germany.
- [3] Jacob D. Durrant, J. Andrew McCammon. (2010). *NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes.* Department of Chemistry & Biochemistry, NSF Center for Theoretical Biological Physics, National.
- *[4] Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking.* (Michael M. Mysinger,Michael Carchia,John. J. Irwin,Brian K. Shoichet). †Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California 94158-2330, United.
- [5] S. Joshua Swamidass, Chloé-Agathe Azencot, Kenny Daily. (2010). *A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval.* USA.
- [6] Sheng-You Huang, S. Z. (2010). *Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions.*
- [7] Teles, J. (2012). *Concordance coefficients to measure the agreement among several sets of ranks.* Unit of Mathematics and CIPER, Faculty of Human Kinetics,Technical University of Lisbon, Lisbon, Portugal.
- [8] Willett, P. (2006). *Similarity-based virtual screening using 2D fingerprints.* Krebs Institute for Biomolecular Research and Department of Information Studies,University of Sheffield, 211 Portobello, Sheffield S1 4DP, UK.
- [9]Zhan Deng, C. C. (2004). *Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions.* Department of Structural Informatics, Biogen, Inc., 12 Cambridge Center, Cambridge, Massachusetts 02142.
- [10] Anna Gaulton, 1. L.-L. (2011). *ChEMBL: a large-scale bioactivity database for drug discovery.*
- [11]Truchon JF1, B. C. (2007). *Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem.*

- [12]Neudert G., K. G. (2011). *fconv: Format conversion, manipulation and feature computation of molecular data.*
- [13]Olson*, O. T. (n.d.). *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading.* 2011.
- [14]Hold, M. P. (2014). *Scalable Concurrent Operations in Python.*
- [15]François-Michel De Rainville, F.-A. F.-A. (n.d.). *DEAP: A Python Framework for Evolutionary Algorithms.*