Πανεπιστήμιο Θεσσαλίας
Τμήμα Βιοχημείας και Βιοτεχνολογίας

Πρόγραμμα μεταπτυχιακών Σπουδών

ΕΦΑΡΜΟΓΕΣ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ

ΜΟΡΙΑΚΗ ΓΕΝΕΤΙΚΗ

ΔΙΑΓΝΩΣΤΙΚΟΙ ΔΕΙΚΤΕΣ

Κουτρουμπής Κωνσταντίνος του Ηλία

Λάρισα 2022

Μπορεί η ταξινόμηση μοντέλων προτεινόμενων από λογισμικό docking πρωτεΐνης-πρωτεΐνης για ένα πρωτεϊνικό σύμπλοκο να επωφεληθεί από την μέτρηση της διπολικής ροπής και της γυροσκοπικής ακτίνας του καθώς και από τον υπολογισμό του pK_a;

Can the classification of models proposed by protein-protein docking software for a protein complex benefit from the measurement of its dipole moment length and radius of gyration as well as the calculation of pK_a?

Επιβλέπων:

Παπαδόπουλος Γεώργιος, Αναπληρωτής Καθηγητής Βιοφυσικής του τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας

Τριμελής Συμβουλευτική Επιτροπή:

- Παπαδόπουλος Γεώργιος, Αναπληρωτής Καθηγητής Βιοφυσικής του Τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας
- Λεωνίδας Δημήτριος, Καθηγητής Βιοχημείας του Τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας
- Αμούτζιας Γρηγόριος, Αναπληρωτής Καθηγητής Βιοπληροφορικής του Τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω θερμά τον κύριο Παπαδόπουλο Γεώργιο, Αναπληρωτή Καθηγητή Βιοφυσικής του τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας για την ανάθεση του θέματος καθώς και για την πολύτιμη βοήθειά του κατά την εκπόνηση της παρούσας διπλωματικής εργασίας.

Θα ήθελα επίσης να ευχαριστήσω τα άλλα δύο μέλη της τριμελής επιτροπής, τον κύριο Λεωνίδα Δημήτριο και τον κύριο Αμούτζια Γρηγόριο, διότι δέχτηκαν να συμμετέχουν στην αξιολόγηση της παρούσας εργασίας.

Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου για την ηθική και οικονομική τους υποστήριξη, καθ’ όλη την διάρκεια του μεταπτυχιακού προγράμματος στο τμήμα Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας.

Περίληψη

Η χρήση υπολογιστικών προγραμμάτων για την πρόβλεψη του τρόπου πρόσδεσης (Docking) μεταξύ δύο ή περισσοτέρων μορίων είναι πλέον απαραίτητη για την μελέτη αλλά και την παραγωγή νέων φαρμάκων. Παρά την μεγάλη πρόοδο της έρευνας και των τεχνολογιών σ' αυτόν τον τομέα τα τελευταία χρόνια, δεν έχει βρεθεί κατάλληλος τρόπος βαθμολόγησης των μοντέλων που προκύπτουν από τέτοια προγράμματα, έτσι ώστε τα μοντέλα με την ομοιότερη δομή προς αυτήν του πραγματικού συμπλόκου να εμφανίζονται στις πρώτες θέσεις της κατάταξης. Σ' αυτήν την μελέτη χρησιμοποιήθηκαν μοντέλα συμπλόκων που προέκυψαν από τα προγράμματα docking HADDOCK και InterEvDock2, προκειμένου να ελεγχθεί η υπόθεση, εάν πειραματικά προσδιοριζόμενα μεγέθη σ' ένα πρωτεϊνικό σύμπλοκο, όπως το μέτρο της ηλεκτρικής διπολικής ροπής DML, η γυροσκοπική ακτίνα R_g καθώς και το υπολογιστικά προσδιοριζόμενο pK_a μπορούν να αξιοποιηθούν στην βελτίωση της κατάταξης των μοντέλων. Ο έλεγχος της υπόθεσης έγινε σε 27 σύμπλοκα γνωστής κρυσταλλικής δομής τα οποία επιλέχθηκαν από έναν κατάλογο συμπλόκων (Mintseris *et al.*, 2005) που χρησιμοποιείται για την αξιολόγηση της ικανότητας των προγραμμάτων docking να προβλέψουν την σωστή δομή ενός συμπλόκου. Στο πλαίσιο αυτής της εργασίας τα μεγέθη DML, R_g και pK_a προσδιορίστηκαν υπολογιστικά από τις γνωστές κρυσταλλικές δομές των συμπλόκων και ακολούθως υπολογίσθηκε για κάθε μοντέλο του docking η σχετική απόκλιση αυτών των μεγεθών από τις τιμές του κρυσταλλικού συμπλόκου. Για την αξιολόγηση των μοντέλων χρησιμοποιήθηκε τόσο το RMSD ως μέτρο της διαφοράς στην δομή μεταξύ ενός μοντέλου και του αντίστοιχου γνωστού συμπλόκου, όσο και το DockQ ως μέτρο ομοιότητας. Τέλος με την χρήση γραμμικής παλινδρόμησης δημιουργήθηκαν γραμμικά μοντέλα με στόχο την πρόβλεψη είτε του RMSD είτε του DockQ από τις ποσοστιαίες αποκλίσεις των μεταβλητών DML, R_g και pK_a. Τα γραμμικά μοντέλα που προέκυψαν από την στατιστική ανάλυση με και χωρίς κανονικοποιημένα δεδομένα, είχαν ακρίβεια μεταξύ 3,7% και 31,5%. Η σχετική σημαντικότητα των τριών μεταβλητών στα μοντέλα αυτά ήταν $R_g > pK_a >$ DML.

Abstract

The use of computer programs to predict the way of docking between two or more molecules is now necessary for the study and production of new drugs. Despite the rise of research and technology in this field these recent years, no suitable way has been found to grade the models resulting from such programs, so that models with a structure similar to that of the real complex appear in the first ranks. In this study, models of complexes derived from the docking programs HADDOCK and InterEvDock2 were used to test the hypothesis of experimentally determined quantities in a protein complex, such as the measurement of the dipole moment length (DML), the radius of gyration (R_g) and the computationally determined pK_a can be utilized to improve the classification of models. The hypothesis was tested on 27 complexes of known crystal structures selected from a list of complexes (Mintseris et al., 2005) used to evaluate the ability of docking programs to predict the correct structure of a complex. In this work, the quantities DML, R_g and pK_a were computed from the known crystal structures of the complexes and then the relative deviation of these quantities from the values of the crystal complex was calculated for each docking model. Both, RMSD, a measure of the difference in structure between a model and the corresponding known complex as well as DockQ, a measure of similarity were used to evaluate the models. Finally, using linear regression, linear models were created with the aim of predicting either RMSD or DockQ from the percentage deviations of the variables DML, R_g and pK_a. The linear models obtained from the statistical analysis with and without normalized data, had an accuracy between 3.7% and 31.5%. The relative importance of the three variables in these models was $R_g > pK_a >$ DML.

Περιεχόμενα

1. Εισαγωγή

1.1. Πρωτεΐνη

Οι πρωτεΐνες αποτελούν ένα από τα βασικά στοιχεία των οργανισμών, καθώς είναι απαραίτητες για τις βασικές τους λειτουργίες. Είναι πολυμερή τα οποία αποτελούνται από μονομερή αμινοξέα που συνδέονται μεταξύ τους. Ο συνδυασμός των 20 διαφορετικών αμινοξέων με διαφορετική σειρά αποτελεί την πρωτοταγή δομή μιας πρωτεΐνης και ταυτόχρονα τον βασικότερο λόγο της διαφοράς μεταξύ τους. Οι αλληλεπιδράσεις μεταξύ των αμινοξέων μιας πρωτεΐνης προσδίδουν στην πρωτεΐνη αυτή μια ιδιαίτερη διαμόρφωση στον χώρο, από την οποία εξαρτάται και η λειτουργία της (Berg *et al.*, 2015).

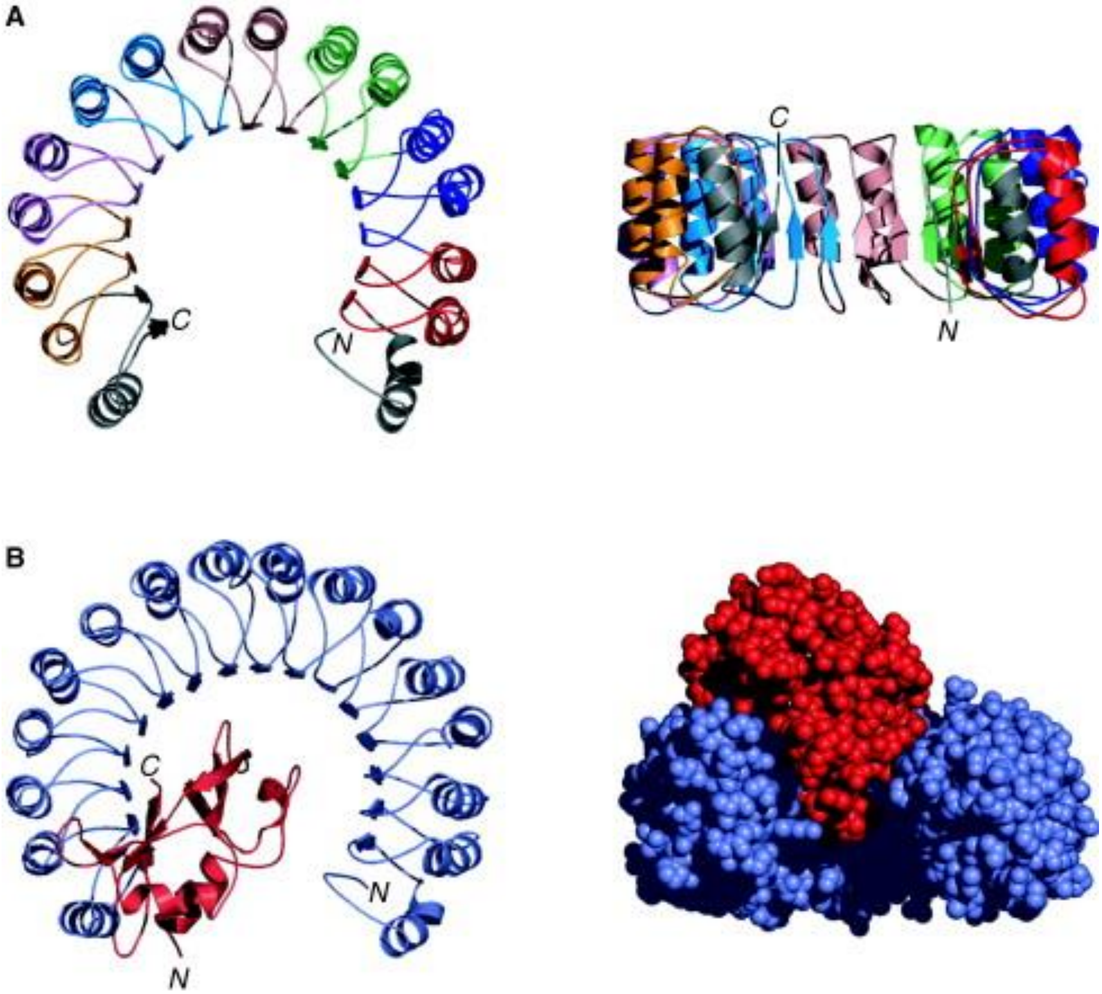
Συγκεκριμένα οι δεσμοί που σχηματίζονται μεταξύ των αμινοξέων για την δημιουργία της πρωτοταγούς δομής μιας πρωτεΐνης ονομάζονται πεπτιδικοί δεσμοί και δημιουργούνται μεταξύ της αμινικής ομάδας του ενός αμινοξέως με την καρβοξυτελική ομάδα του επόμενου. Εκτός από τις δύο αυτές ομάδες των αμινοξέων υπάρχει και μια ακόμα ομάδα, η χαρακτηριστική ομάδα R που διαφοροποιεί μεταξύ τους τα αμινοξέα. Αυτή η ομάδα ονομάζεται και πλευρική αλυσίδα του αμινοξέως (Berg *et al.*, 2015).

Με την δημιουργία δεσμών υδρογόνου μεταξύ των ομάδων N-H και C=O «γειτονικών» αμινοξέων στην γραμμική αλληλουχία μιας πρωτεΐνης σχηματίζονται μορφές στον χώρο, α-έλικες, β-πτυχωτά φύλλα και στροφές, οι οποίες αποτελούν την δευτεροταγή μορφή της πρωτεΐνης. Η τριτοταγής δομή της πρωτεΐνης προκύπτει από την αναδίπλωση των προαναφερθέντων μορφών κυρίως σύμφωνα με την υδροφοβικότητα ή υδροφιλικότητά τους καθώς και τους δισουλφιδικούς δεσμούς που σχηματίζονται μεταξύ αμινοξέων. Τέλος, η τεταρτοταγής δομή μιας πρωτεΐνης, εάν υπάρχει, αφορά στην διάταξη στον χώρο των υπομονάδων μιας πρωτεΐνης και στις αλληλεπιδράσεις αυτών. (Berg *et al.*, 2015).

1.2. Αλληλεπιδράσεις πρωτεϊνών

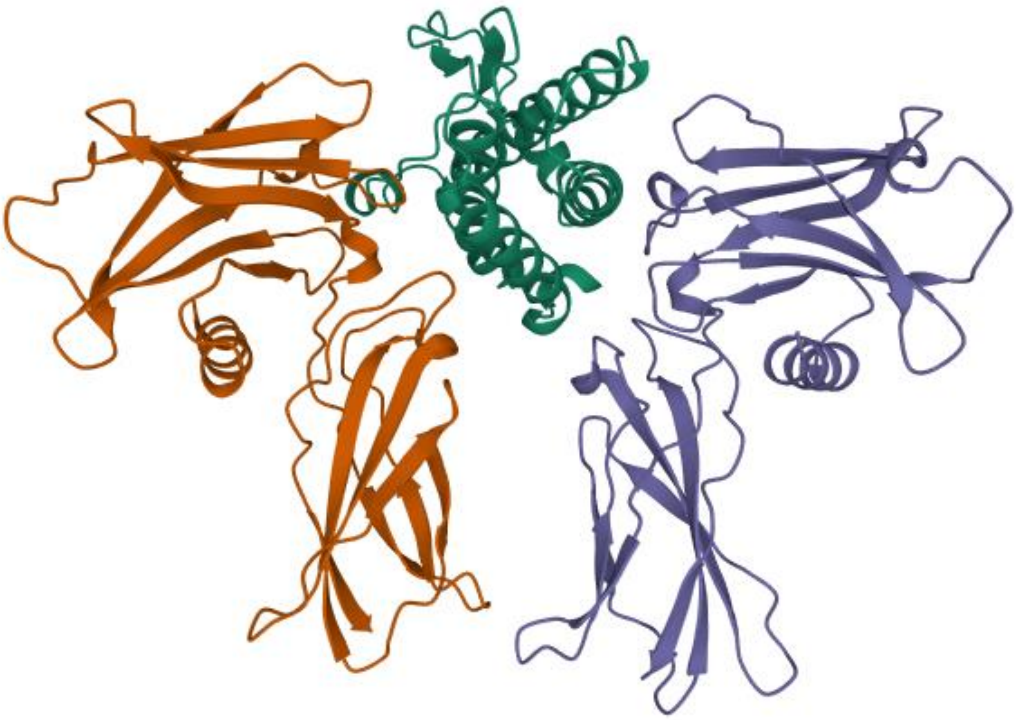
Κάθε μία από τις φυσιολογικές λειτουργίες των οργανισμών περιλαμβάνει τη χρήση πρωτεϊνών, είτε αυτές αλληλεπιδρούν μεταξύ τους, είτε με άλλου είδους μόρια. Το σύνολο των αλληλεπιδράσεων των πρωτεϊνών αποτελεί και την κυτταρική λειτουργία (Ruyck *et al.*, 2016). Η σύνδεση δύο ή και περισσοτέρων πρωτεϊνών μεταξύ τους για την δημιουργία συμπλόκου είναι μια εξαιρετικά εξειδικευμένη διαδικασία, καθώς η συμπληρωματικότητά τους είναι αντίστοιχη με αυτή του κλειδιού με την κλειδαριά (Fischer E., 1894). Αυτή η συμπληρωματικότητα βασίζεται στις στερεοχημικές και φυσικοχημικές ιδιότητες των επιμέρους πρωτεϊνών.

Ένα τέτοιο παράδειγμα αλληλεπίδρασης πρωτεϊνών αποτελεί και αυτή μεταξύ της ριβονουκλεάσης A με τον αναστολέα της. Η ριβονουκλεάση είναι γνωστό πως καταλύει την αποδόμηση του RNA σε μικρότερα τμήματα. Ο αναστολέας της ριβονουκλεάσης αναγνωρίζει τις ριβονουκλεάσες και προσδένεται με αυτές ισχυρά, όπως φαίνεται στην εικόνα 1, καθιστώντας τες ανενεργές (Dickson *et al.*, 2005, Blackburn and Moore, 1982). Ο αναστολέας της ριβονουκλεάσης (RI) περιέχει επαναλαμβανόμενες δομικές μονάδες, οι οποίες περιέχουν πολλά κατάλοιπα λευκίνης και ονομάζονται επαναλήψεις πλούσιες σε λευκίνη (leucine-rich repeats LLR). Αυτά τα μοτίβα έχουν αναγνωριστεί σε διάφορες πρωτεΐνες και φαίνεται να παίζουν καθοριστικό ρόλο στην δημιουργία αλληλεπιδράσεων πρωτεϊνών (Kobe and Kajava, 2001).



Εικόνα 1. Τριδιάστατες δομές του αναστολέα της ριβονουκλεάσης (RI) και του συμπλόκου του με την ριβονουκλεάση. Α) Χοίρεια RI με τα χρώματα να αντιστοιχούν στα διαφορετικά εξώνια που κωδικοποιούν τα αντίστοιχα δομικά στοιχεία. Β) Σύμπλοκο χοίρειας RI με RNase A. (Dickson *et al.*, 2005)

Η ερυθροποιητίνη είναι μια ορμόνη που παίζει καθοριστικό ρόλο στην ρύθμιση της παραγωγής ερυθροκυττάρων. Συγκεκριμένα ενεργοποιεί μέσω του υποδοχέα της τον πολλαπλασιασμό και την διαφοροποίηση των πρόδρομων κυττάρων σε ερυθρά. Ο υποδοχέας της ερυθροποιητίνης αποτελείται από ένα διμερές συνδεδεμένο με κινάση. Με την πρόσδεση της ερυθροποιητίνης στο εξωκυτταρικό τμήμα του υποδοχέα έρχονται τα δύο μονομερή κοντά, με αποτέλεσμα την φωσφορυλίωση των ενδοκυττάριων τμημάτων μεταξύ τους και επομένως την ενεργοποίησή του και συνέχιση του μονοπατιού (Bunn Franklin, 2013, Li and Kang, 2017, Syed *et al*., 1998).



Εικόνα 2. Τριδιάστατη δομή του συμπλόκου της ερυθροποιητίνης με τον υποδοχέα της. Στην εικόνα φαίνεται με πράσινο η ερυθροποιητίνη και το διμερές του υποδοχέα με μπλε και πορτοκαλί. (https://www.rcsb.org/structure/1EER)

Για την κατανόηση αυτών των αλληλεπιδράσεων είναι απαραίτητη η γνώση της δομής των συμπλόκων των εν λόγω πρωτεϊνών. Με βάση τις δομές των επιμέρους πρωτεϊνών, είναι δυνατόν να γίνουν με υπολογιστικές μεθόδους προβλέψεις για την δομή του συμπλόκου των πρωτεϊνών αυτών. Η διαδικασία αυτή ονομάζεται “Protein-Protein Docking” (Vakser I. E., 2014).

1.3. Protein-Protein Docking (Πρόσδεσης πρωτεΐνης-πρωτεΐνης)

Όπως προαναφέρθηκε, για να είναι δυνατή η πρόβλεψη των δομών συμπλόκων πρωτεϊνών, πρέπει να είναι γνωστές οι δομές των επιμέρους πρωτεϊνών. Οι τριδιάστατες δομές των πρωτεϊνών είναι αντικείμενο έρευνας από το 1970 και μετά με την χρήση τεχνολογιών κρυσταλλογραφίας. Οι δομές αυτές είναι αποθηκευμένες με την μορφή αρχείων κειμένου και είναι προσιτές μέσω της βάσης δεδομένων PDB (RCSB PDB, rcsb.org). Με την χρήση αυτών των αρχείων, τα οποία περιέχουν πληροφορίες για την ακολουθία των αμινοξέων και την θέση στον χώρο κάθε ατόμου αυτών, είναι δυνατή η παραγωγή πιθανών μοντέλων πρόσδεσης των εν λόγω πρωτεϊνών. Χρησιμοποιώντας λοιπόν αυτά τα αρχεία έχουν δημιουργηθεί ανά τα χρόνια δεκάδες υπολογιστικές μέθοδοι πρόσδεσης πρωτεϊνών, με σκοπό την διευκόλυνση της μελέτης συμπλόκων (Vajda *et al.*, 2013, Vakser I. E., 2014). Στην παρούσα εργασία γίνεται χρήση δύο προγραμμάτων docking, του HADDOCK και του InterEvDock.

1.4. Ανάγκη υπολογιστικών και πειραματικών μεθόδων

Έπειτα από την αλληλούχιση του γονιδιώματος διαφόρων οργανισμών έχει γίνει γνωστή σε μεγάλο βαθμό η ύπαρξη σχεδόν όλων των μακρομορίων αυτών των οργανισμών. Στα μακρομόρια αυτά ανήκουν φυσικά και οι πρωτεΐνες. Με την συνεχή ανάπτυξη των μεθόδων προσδιορισμού της δομής με κρυσταλλογραφία ακτίνων X και NMR γίνεται όλο και ταχύτερα η ανακάλυψη των τριδιάστατων δομών των πρωτεϊνών (Russell et al., 2004). Παρά το γεγονός πως η δομή των συμπλόκων μεταξύ μακρομορίων φαίνεται να παίζει σημαντικό ρόλο στην κατανόηση των μηχανισμών των κυττάρων, ο αριθμός των δομών των συμπλόκων που έχουν ανακαλυφθεί με τις προαναφερθείσες μεθόδους είναι πολύ μικρότερος από την ανακάλυψη των δομών πρωτεϊνών ξεχωριστά. Αυτό οφείλεται κατά κύριο λόγο στην δυσκολία δημιουργίας και απομόνωσης κρυσταλλωμένων συμπλόκων για την μετέπειτα μελέτη τους (Russell et al., 2004). Για την αντιμετώπιση αυτού του προβλήματος γίνεται χρήση υπολογιστικών μεθόδων, πράγμα που αναδεικνύει την ανάγκη χρήσης τους για την διευκόλυνση της έρευνας και την κατανόηση της κυτταρικής λειτουργίας.

Η χρήση των υπολογιστικών μεθόδων έχει μετατραπεί σε ένα από τα βασικότερα εργαλεία στην διάθεσή μας για την ανακάλυψη νέων φαρμάκων. Η ανακάλυψη νέων ουσιών με πιθανές φαρμακευτικές ιδιότητες με την χρήση μεθόδων docking έχοντας ως βάση ήδη γνωστές δομές πρωτεϊνών, ονομάζεται “structure-based drug design” (SBDD) (Verkhivker *et al.*, 2000, Ferreira *et al.*, 2015). Όπως αναφέρουν και οι Ruyck *et al.*, η χρήση των υπολογιστικών μεθόδων για τον σχεδιασμό φαρμάκων έχει εντατικοποιηθεί και έχει σε αρκετές περιπτώσεις οδηγήσει σε θετικά αποτελέσματα (Finn John, 2012, Lin *et al.*, 2020, Joon *et al.*, 2022).

Παρά την εντατική χρήση των υπολογιστικών μεθόδων στην έρευνα, ο συνδυασμός υπολογιστικών αλλά και πειραματικών μεθόδων είναι ο αποτελεσματικότερος τρόπος αναγνώρισης και ανάπτυξης νέων ουσιών (Ferreira *et al.*, 2015). Προτού όμως γίνει χρήση οποιασδήποτε μεθόδου, είτε αυτή είναι υπολογιστική είτε πειραματική, είναι ανάγκη να περάσει από μια διαδικασία αξιολόγησης, έτσι ώστε να θεωρείται αξιόπιστη (Viceconti *et al.*, 2021, Sabe *et al.*, 2021). Η πειραματική επομένως αξιολόγηση των μοντέλων είναι ζωτικής σημασίας, καθώς δίνει την δυνατότητα προσαρμογής της μεθόδου στις παραμέτρους της πραγματικότητας (Ruyck *et al.*, 2016, Sabe *et al.*, 2021). Στην παρούσα μελέτη πραγματοποιήθηκαν υπολογιστικές μέθοδοι έναντι πειραματικών μεθόδων για την αξιολόγηση, καθώς έγινε χρήση των γνωστών συμπλόκων των πρωτεϊνών από την λίστα αναφοράς της βάσης δεδομένων PDB (Mintseris *et al.*, 2005). Από τα 70 σύμπλοκα πρωτεϊνών που ανήκουν σε αυτή την λίστα αναφοράς χρησιμοποιήθηκαν τα 20 για αυτή την εργασία. Η αξιολόγηση αυτή γίνεται με βάση ορισμένα μεγέθη, τα οποία είναι δυνατόν να μετρηθούν πειραματικά ή να υπολογιστούν "in silico" με βάση γνωστά σύμπλοκα. Οι μεταβλητές αυτές είναι: το μέτρο της διπολικής ροπής (DML), η γυροσκοπική ακτίνα (R_g), το συνολικό pK_a του συμπλόκου. Ως μεγέθη αξιολόγησης της ποιότητας των μοντέλων μπορεί να χρησιμοποιηθούν η τετραγωνική ρίζας της μέσης τετραγωνικής απόκλισης των θέσεων των ατόμων (RMSD), καθώς και το DockQ.

1.5. HADDOCK

Η συγκεκριμένη υπολογιστική μέθοδος, πρόσδεση πρωτεϊνών με βάση την υψηλή ασάφεια (High ambiguity driven protein−protein docking), αποτελεί μέθοδο πρόσδεσης άκαμπτων δομών (rigid body docking) με εύκαμπτες πλευρικές αλυσίδες και αναφέρεται αναλυτικά στη δημοσίευση των Dominguez *et al*. το 2003. Εν συντομία, το πρόγραμμα αυτό χρησιμοποιεί αρχικά τους ασαφείς περιορισμούς αλληλεπίδρασης (Ambiguous Interaction Restraints-AIR) που προέρχονται από οποιεσδήποτε πειραματικές πληροφορίες, δίνοντας τελικά τα πιθανά «ενεργά» ή «παθητικά» κατάλοιπα μιας πρωτεΐνης. Τα «ενεργά» κατάλοιπα αντιστοιχούν σε αυτά που θεωρείται ότι συμμετέχουν σε αλληλεπίδραση με άλλες πρωτεΐνες. Ο υπολογισμός αυτός είναι ενσωματωμένος στον CPORT web server (Vries *et al.*, 2011). Το HADDOCK χρησιμοποιεί επίσης και το CNS, ένα πρόγραμμα κρυσταλλογραφίας και NMR για τον υπολογισμό των δομών, ενώ παράλληλα χρησιμοποιεί και τον υπολογισμό ενεργειών εντός αλλά και μεταξύ μορίων για περαιτέρω βαθμολόγηση κατά τον υπολογισμό των δομών (Dominguez *et al.*, 2003, Zundert *et al.*, 2016). Οι ιστοσελίδες των web server του CPORT και HADDOCK είναι η http://alcazar.science.uu.nl/services/CPORT/ και η https://wenmr.science.uu.nl/haddock2.4/ αντίστοιχα.

1.6. InterEvDock

Ο web server InterEvDock2 που χρησιμοποιήθηκε στην παρούσα μελέτη, εφαρμόζει ένα πρόγραμμα πρόσδεσης (FRODOCK2) των πρωτεϊνών και τρεις αλγορίθμους βαθμολόγησης των αποτελεσμάτων αυτού του προγράμματος. Το FRODOCK2 ασκεί άκαμπτο docking (rigid-body docking) που συμπληρώνεται με ένα "knowledge-based" δυναμικό. Είναι μια μέθοδος άκαμπτου docking που συνδυάζει τον αλγόριθμο γρήγορου μετασχηματισμού Fourier (FFT) (Garzon *et al.*, 2009) με έναν αλγόριθμο αναζήτησης βασισμένον στις σφαιρικές αρμονικές για την επιτάχυνση αναζήτησης μοντέλων κατά την περιστροφή των επιμέρους πρωτεϊνών, μια συνάρτηση βαθμολόγησης βασιζόμενη στην ενέργεια αλληλεπίδρασης με όρους van der Waals, ηλεκτροστατικούς καθώς και αποδιαλυτοποίησης, και συμπληρωματικά σε ένα δυναμικό "knowledge-based" (Tobi D., 2010, Ramírez-Aportela *et al.*, 2016). Πέρα από την πρόσδεση και την βαθμολόγηση με το FRODOCK, το InterEvDock2 βασίζεται και στην επαναξιολόγηση των αποτελεσμάτων με το ατομικό στατιστικό δυναμικό SOAP-PP (Dong *et al.*, 2013) και με το InterEvScore (Andreani *et al.*, 2013, Yu *et al.*, 2016, Quignot *et al.*, 2018). Η ιστοσελίδα του του InterEvDock2 είναι https://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::InterEvDock2.

1.7. Υπολογισμός μεταβλητών ελέγχου

Όπως προαναφέρθηκε, στην παρούσα εργασία ερευνήθηκε η χρήση των παραπάνω μεταβλητών για την περαιτέρω βαθμολόγηση των μοντέλων που προέρχονταν από την πρόσδεση πρωτεϊνών με υπολογιστικές μεθόδους. Σκοπός είναι η δημιουργία ενός νέου μοντέλου βαθμολόγησης, το οποίο με την χρήση επιπλέον μεταβλητών που θα μπορούν να υπολογιστούν πειραματικά, θα έχει την δυνατότητα να απομονώνει τα καλύτερα μοντέλα από τα ήδη υπάρχοντα.

Η πρώτη από τις τρεις μεταβλητές της μεθόδου είναι το μέτρο της διπολικής ροπής DML (Dipole Moment Length). Η διπολική ροπή μιας κατανομής ηλεκτρικών φορτίων (εδώ των μερικών φορτίων) ορίζεται στην εξίσωση 1 και μετράται σε Debye (Housecroft and Sharpe, 2008, Tro, 2008).

$$\boldsymbol{\mu} = \sum_{1}^{n} q_i \cdot \boldsymbol{r}_{\boldsymbol{i}} \qquad \text{Εξίσωση 1}$$

Όπου $\boldsymbol{\mu}$ το διάνυσμα της διπολικής ροπής, q_i το φορτίο του ατόμου i και $\boldsymbol{r}_i$ το διάνυσμα θέσης του ατόμου i ως προς το γεωμετρικό κέντρο της πρωτεΐνης, ενώ τέλος το n είναι ο συνολικός αριθμός των ατόμων.

Πειραματικά είναι δυνατή η μέτρηση μόνο του μέτρου της διπολικής ροπής μέσω της διηλεκτρικής φασματοσκοπίας. Έχοντας λοιπόν την τιμή των DML των μοντέλων και αυτού του απομονωμένου συμπλόκου που μετρήθηκε πειραματικά, θα μπορούσαμε να αποκλείσουμε τις τιμές των μοντέλων που αποκλίνουν πολύ από εκείνη του πραγματικού συμπλόκου. Παρόλα αυτά, η τιμή του DML δεν είναι από μόνη της αρκετά αξιόπιστη, καθώς δύο πρωτεΐνες μπορεί να δίνουν δύο εντελώς διαφορετικά σύμπλοκα, όσον αφορά την δομή, που να έχουν ακριβώς την ίδια τιμή DML, λόγω της συμμετρίας της διπολικής ροπής γύρω από τον άξονά της. Επίσης οι αλγόριθμοι που χρησιμοποιούνται για την προσθήκη των μερικών φορτίων μπορεί να παράγουν λίγο διαφορετικές κατανομές ηλεκτρικών φορτίων και επομένως και DML. Για τον λόγο αυτό έγινε χρήση και της απόκλισης της γυροσκοπικής ακτίνας ως δεύτερης δομικής μεταβλητής για την βαθμολόγηση.

Η γυροσκοπική ακτίνα R_g ορίζεται με την βοήθεια της εξίσωσης 2. Εδώ, είναι καθοριστική η κατανομή των ατομικών μαζών και όχι του ηλεκτρικού φορτίου. Πειραματικά προσδιορίζεται με την μέθοδο SAXS (σκέδαση ακτινών Χ μικρής γωνίας) (Miura K., 2018, Gräwert and Svergun, 2020). Είναι εύλογο να υποθέσουμε ότι τα διάφορα μοντέλα θα διαφέρουν ως προς την κατανομή των ατομικών μαζών και ότι η R_g θα μπορούσε να βοηθήσει στην αναγνώριση του "σωστού μοντέλου".

$$R_g^2 = \frac{\sum_{i=1}^{n} m_i (r_i - r_{cm})^2}{\sum_{i=1}^{n} m_i} \qquad \text{Εξίσωση 2}$$

Όπου R_g η γυροσκοπική ακτίνα, m_i η μάζα του ατόμου i και r_i η θέση του ατόμου i ως προς το γεωμετρικό κέντρο της πρωτεΐνης, ενώ τέλος το $\boldsymbol{n}$ είναι ο συνολικός αριθμός των ατόμων.

Όπως και για το DML, αναμένουμε πως η απόκλιση των R_g των μοντέλων που προκύπτουν από τις υπολογιστικές μεθόδους σε σχέση με το πραγματικό σύμπλοκο που θα προκύπτει από την πειραματική μέθοδο θα μας βοηθήσει να αξιολογήσουμε τα μοντέλα. Συγκεκριμένα όσο μικρότερη είναι η απόκλιση του R_g του μοντέλου σε σχέση με το πραγματικό σύμπλοκο, τόσο πιο κοντά θα βρίσκεται το μοντέλο στην πραγματική δομή. Δυστυχώς και με αυτή την μεταβλητή αντιμετωπίζονται παρόμοια προβλήματα όσον αφορά την αξιοπιστία της. Συγκεκριμένα, όπως και με το DML, η περιστροφή ενός ή και των δύο πρωτεϊνών του συμπλόκου γύρω από τον άξονα συμμετρίας τους μπορεί να δίνει παρόμοιες ή ακόμα και ίδιες τιμές R_g, έχοντας διαφορετικές τριδιάστατες δομές και επομένως και με το σύμπλοκο αναφοράς. Έτσι, μια τρίτη μεταβλητή, το pKa, κλήθηκε να συμπληρώσει το κενό.

Το pK_a είναι ο αρνητικός δεκαδικός λογάριθμος της σταθεράς διάστασης ενός οξέος (εξίσωση 3) και είναι μέτρο του βαθμού απόδοσης ή πρόσληψης πρωτονίων. Το pK_a σχετίζεται με το pH σύμφωνα με την εξίσωση Henderson-Hasselbach (εξίσωση 4), όπου οι δύο αυτές τιμές ισούνται σε περίπτωση που οι συγκεντρώσεις ενός οξέος και της συζυγούς βάσης του είναι ίσες σε ένα διάλυμα.

$$pK_a = -\log_{10} K_a = \log_{10} \frac{[HA]}{[A^-][H^+]} \qquad \text{Εξίσωση 3}$$

$$pH = pK_a + \log_{10}\left(\frac{[A^-]}{[HA]}\right) \qquad \text{Εξίσωση 4}$$

Όπου [HA] είναι η συγκέντρωση οξέος, $[A^-]$ η συγκέντρωση της συζυγούς βάσης και $[H^+]$ η συγκέντρωση των πρωτονίων.

Οι τιμές του pK_a των αμινοξέων των πρωτεϊνών σε ουδέτερο pH και υδάτινο περιβάλλον,διαφέρουν μεταξύ τους ανάλογα με την πλευρική τους αλυσίδα. Καθώς το pK_a επηρεάζεται σε μεγάλο βαθμό από την πολικότητα του περιβάλλοντος, αναμένεται το pK_a ενός είδους αμινοξέος να διαφέρει αρκετά εντός μιας πρωτεΐνης σύμφωνα με το σημείο στο οποίο βρίσκεται στη δομή της πρωτεΐνης. Για παράδειγμα, ένα αμινοξύ στο εσωτερικό μιας πρωτεΐνης, όσον αφορά τη δομή της στον χώρο, σε σχέση με ένα αμινοξύ στην επιφάνειά της, που θα έρχεται σε επαφή με μόρια νερού, θα έχει διαφορετικό pK_a. Αυτή η διαφορά μπορεί να παρατηρηθεί και κατά την σύνδεση των δύο πρωτεϊνών μεταξύ τους, ιδιαίτερα στο σημείο αλληλεπίδρασης μεταξύ τους, καθώς αμινοξέα μπορεί να έρχονται πλέον σε επαφή με αμινοξέα της άλλης πρωτεΐνης έναντι των μορίων νερού όπου αλληλοεπιδρούσαν πριν.

Ορισμένες μελέτες πάνω στην αλλαγή της κατάστασης πρωτονίωσης των αμινοξέων κατά την πρόσδεση πρωτεϊνών, και επομένως του pK_a, έχουν αποδειχθεί αντιφατικές. Παρόλο που εκ πρώτης όψεως το pK_a φαίνεται να αλλάζει μεμονωμένα στα αμινοξέα των πρωτεϊνών κατά την πρόσδεση, συνολικά το σύνολο των pK_a των αμινοξέων του συμπλόκου είναι πιθανότερο να μην αλλάξει, ή να αλλάξει πολύ λίγο (Onufriev and Alexov, 2013). Επομένως σε σταθερό pH, η διαφορά του συνολικού pK_a των δύο πρωτεϊνών του συμπλόκου πριν από την πρόσδεση με το συνολικό pK_a των μοντέλων που προκύπτουν από το docking, μπορεί να αποτελέσει μια μεταβλητή αξιολόγησής τους. Με βάση την παραπάνω υπόθεση, όσο μικρότερη είναι αυτή η διαφορά, τόσο πιο κοντά βρίσκεται το εν λόγω μοντέλο στο πραγματικό σύμπλοκο. Το pK_a πριν και μετά από την πρόσδεση μπορεί να υπολογισθεί από την δομή των υπομονάδων με την βοήθεια του προγράμματος PROPKA (Li *et al.*, 2005). Με την χρήση και των τριών αυτών μεταβλητών είναι πιθανόν να δημιουργήσουμε ένα σχετικά αξιόπιστο σύστημα κατάταξης των μοντέλων.

1.8. Αξιολόγηση μεταβλητών ελέγχου

Για την αξιολόγηση του παραπάνω συστήματος έγινε χρήση δύο ευρέως χρησιμοποιούμενων μεγεθών αξιολόγησης μοντέλων, του RMSD και του DockQ. Το RMSD (Root means square deviation) υπολογίζεται σύμφωνα με την εξίσωση 5. Στην προκειμένη περίπτωση αφορά στις αποστάσεις των ατόμων του πρωτεϊνικού σκελετού μεταξύ του γνωστού συμπλόκου και των μοντέλων που προκύπτουν από τις υπολογιστικές μεθόδους. Αυτό υποδηλώνει πως όσο μικρότερος είναι αυτός ο αριθμός, τόσο πιο κοντά βρίσκονται χωροταξικά τα δύο σύμπλοκα και επομένως τόσο πιο σωστό είναι το μοντέλο που διερευνάται. Το RMSD μετριέται σε Ångström (Å) που ισούται με 10^{-10}m, ενώ ένα κοινώς αποδεκτό κατώφλι για docking πρωτεϊνών είναι τα 2.0 Å (Scarpino *et al.*, 2018). Η εξίσωση υπολογισμού του είναι (εξίσωση 5):

$$RMSD = \sqrt{\frac{1}{n}[(x_{ib} - x_{id})^2 + (y_{ib} - y_{id})^2 + (z_{ib} - z_{id})^2]} \qquad \text{Εξίσωση 5}$$

Όπου n είναι το σύνολο των ατόμων, και οι δείκτες b και d στα x_{ib}, x_{id}, y_{ib}, y_{id}, z_{ib}, z_{id}, αντιστοιχούν στο γνωστό σύμπλοκο (bound) και στο εν λόγω μοντέλο (docking model). Τέλος τα x, y και z αντιστοιχούν στις θέσεις των ατόμων στον χώρο.

Το DockQ είναι ο δεύτερος τρόπος αξιολόγησης που χρησιμοποιήθηκε και ακολουθεί την κατηγοριοποίηση των μοντέλων σύμφωνα με το σύστημα της κοινότητας Κριτικής Αξιολόγησης Προβλεπόμενων Αλληλεπιδράσεων (CAPRI). Το CAPRI βασίζεται σε τρείς συσχετιζόμενες μεταβλητές, το F_{nat}, LRMS και iRMS (Lensink *et al.*, 2007).

- F_{nat}: Το σύνολο των ενδογενών επαφών στο προβλεφθέν σύμπλοκο (μοντέλο) προς το σύνολο των επαφών στο πραγματικό σύμπλοκο
- LRMS: Ligand Root mean square. H τετραγωνική ρίζα του μέσου όρου των αποστάσεων των ατόμων του “backbone”, δηλαδή των Cα, N, C, O, μεταξύ του προσδέτη στο μοντέλο και στο γνωστό σύμπλοκο
- iRMS: Interface Root mean square. H τετραγωνική ρίζα του μέσου όρου των αποστάσεων των ατόμων της διεπαφής των πρωτεϊνών των συμπλόκων μεταξύ αυτών του μοντέλου και του γνωστού συμπλόκου. Για να θεωρούνται άτομα στην διεπαφή των πρωτεϊνών, θα πρέπει να έχουν απόσταση από την άλλη πρωτεΐνη το πολύ 10 Å.

Το CAPRI και κατ' επέκταση και το DockQ ομαδοποιούν τα μοντέλα σε τέσσερις κατηγορίες, σύμφωνα με τις παραπάνω μεταβλητές, τις εξής: λανθασμένο, αποδεκτό, μεσαίας και υψηλής ποιότητας (incorrect, acceptable, medium or high quality αντίστοιχα). Το DockQ συγκεκριμένα ενσωματώνει και τις τρεις τιμές (F_{nat}, LRMS και iRMS) σε μία με την εξίσωση 6, με εύρος τιμών [0,1]. Για την συγκεκριμένη τιμή, όσο πιο κοντά στο 1 βρίσκεται, τόσο καλύτερο είναι το μοντέλο υπό διερεύνηση. Ο διαχωρισμός γίνεται με τις παρακάτω τιμές: 0 ≤ DockQ < 0.23 είναι “Incorrect”, 0.23 ≤ DockQ < 0.49 είναι “Acceptable”, 0.49 ≤ DockQ < 0.8 είναι “Medium quality” και 0.8 ≤ DockQ είναι “High quality” (Basu and Wallner, 2016a and b).

$$DockQ = \frac{F_{nat} + LRMS + iRMS}{3} \qquad \text{Εξίσωση 6}$$

2. Μέθοδοι και ανάλυση

2.1. Προγράμματα Docking

Όπως προαναφέρθηκε, και τα δύο προγράμματα που χρησιμοποιήθηκαν στην έρευνα αυτή, HADDOCK και InterEvDock2, χρησιμοποιήθηκαν μέσω των ιστοσελίδων των web server τους https://wenmr.science.uu.nl/haddock2.4/ και https://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::InterEvDock2 αντίστοιχα με τις προκαθορισμένες τους ρυθμίσεις.

2.2. Προετοιμασία αρχείων των συμπλόκων

Για την ανάλυση χρησιμοποιήθηκαν αρχεία από 20 ζεύγη πρωτεϊνών από την βάση δεδομένων PDB, και συγκεκριμένα από την λίστα αναφοράς, όπου είχαν ταξινομηθεί με βάση την δυσκολία τους για την εύρεση μοντέλων από την πρόσδεση τους. Τα επίπεδα δυσκολίας είναι: εύκολο με άκαμπτα σύμπλοκα άκαμπτου docking (RB), μεσαίας δυσκολίας (MD), υψηλής δυσκολίας (HD). Τα 20 ζεύγη που χρησιμοποιήθηκαν ανήκουν στα παρακάτω επίπεδα δυσκολίας, 10 στην κατηγορία RB, και από 5 στις κατηγορίες MD και HD.

Τα σύμπλοκα αυτά είναι:

- 1BVN: (RB) Σύμπλοκο παγκρεατικής αμυλάσης χοίρου με τον πρωτεϊνικό αναστολέα tendamistat (μικροβιακός αναστολέας α-αμυλάσης)
- 1CGI: (RB) Σύμπλοκο βόειου χημοθρυψινογόνου με δύο παραλλαγές ανασυνδυασμένου αναστολέα της ανθρώπινης παγκρεατικής τρυψίνης
- 1D6R: (RB) Σύμπλοκο βόειας τρυψίνης με τον αναστολέα Bowman-Birk
- 1DFJ: (RB) Σύμπλοκο ριβονουκλεάσης Α με αναστολέα ριβονουκλεάσης
- 1EAW: (RB) Σύμπλοκο ματριπτάσης με BPTI
- 1EWY: (RB) Σύμπλοκο φερεδοξίνης με αναγωγάση της φερεδοξίνης
- 1EZU: (RB) Σύμπλοκο τρυψίνης D102N με ecotin Y69F
- 1F34: (RB) Σύμπλοκο χοίρειας πεψίνης με αναστολέα πεψίνης
- 1HIA: (RB) Σύμπλοκο καλλικρεΐνη με hirustasin (αναστολέα της καλλικρεΐνης)
- 1MAH: (RB) Σύμπλοκο ακετυλοχολινεστεράσης με φασκικουλίνη (νευροτοξίνη)
- 1ACB: (MD) Σύμπλοκο χοίρειας χημοτρυψίνης με eglin C (αναστολέας ελεστάσης)
- 1KKL: (MD) Σύμπλοκο πρωτεϊνικής κινάσης HprK με την πρωτεΐνη HPr
- 1GP2: (MD) Συνδεδεμένο ετεροτριμερές G πρωτεΐνης των υπομονάδων Alpha και Beta-Gamma
- 1GRN: (MD) Σύμπλοκο μικρής G πρωτεΐνης CDC42 με τον ενεργοποιητή της CDC42 Gap
- 1HE8: (MD) Σύμπλοκο μικρής G πρωτεΐνης Ras με κινάση PIP3
- 1ATN: (HD) Σύμπλοκο ακτίνης με DNase I
- 1EER: (HD) Σύμπλοκο ανθρώπινης ερυθροποιητίνης με τον υποδοχέα της
- 1FAK: (HD) Σύμπλοκο παράγοντα ανθρώπινου ιστού με παράγοντα πήξης αίματος VIIA με παρεμποδιστή μεταλλαγμένη BPTI
- 1FQ1: (HD) Σύμπλοκο κινάσης CDK2 με CDK αναστολέα 3
- 1H1V: (HD) Σύμπλοκο γκελσολίνης με ακτίνη

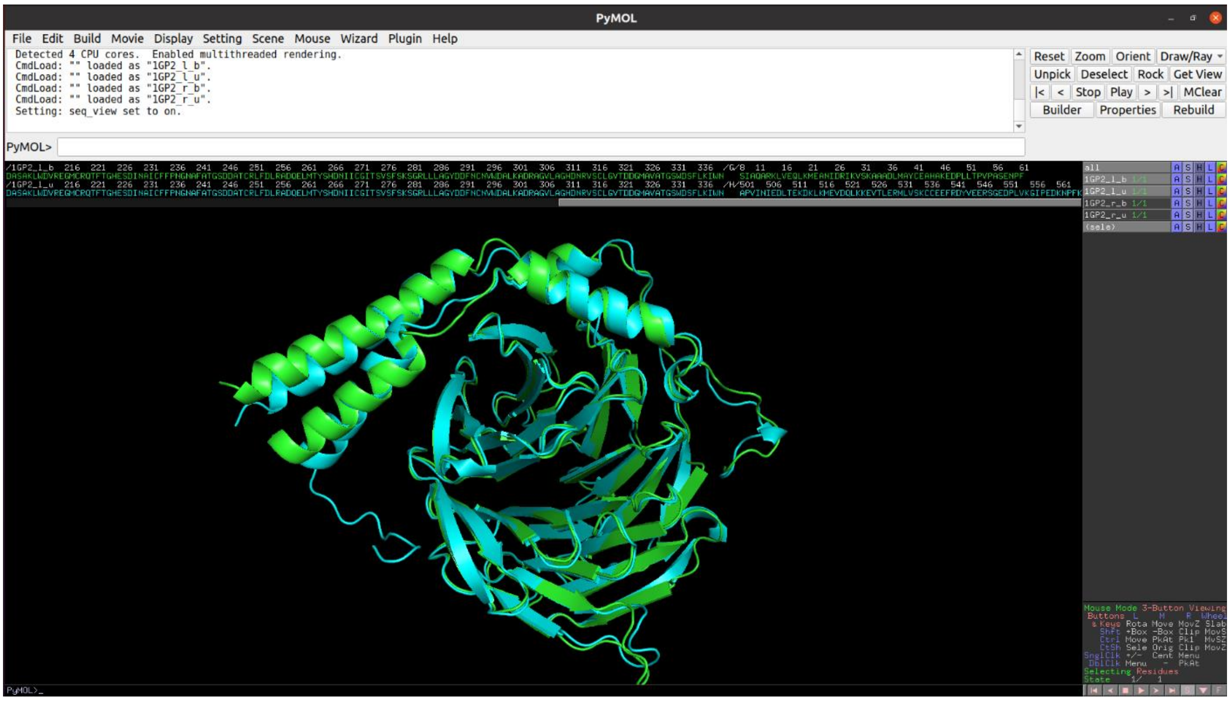
Για κάθε ένα από τα παραπάνω σύμπλοκα είχαμε στην διάθεσή μας τέσσερα αρχεία pdb. Δύο από αυτά ανήκουν σε υπομονάδες του συμπλόκου μετά από την πρόσδεση (**b**ound), ενώ τα άλλα δύο προ του σχηματισμού συμπλόκου (**u**nbound). Τα αρχεία pdb των δύο μη προσδεδεμένων υπομονάδων χρησιμοποιήθηκαν από τα προγράμματα πρόσδεσης για την δημιουργία μοντέλων συμπλόκων, ενώ τα αρχεία από τις προσδεδεμένες υπομονάδες χρησιμοποιήθηκαν για την αξιολόγηση των παραπάνω μοντέλων. Σημειωτέον, ότι μια unbound υπομονάδα μπορεί να αποτελείται από περισσότερες της μίας αλυσίδες.

Η προετοιμασία των παραπάνω αρχείων αποτελεί το σημαντικότερο βήμα στην διαδικασία αξιολόγησης, καθώς τα αρχεία οφείλουν να είναι σε κατάλληλη μορφή για την περαιτέρω αξιοποίησής τους από τα προγράμματα docking. Αρχικά έγινε μετονομασία όλων των αλυσίδων των αρχείων σε "A" για τον υποδοχέα και "B" για τον προσδέτη χάριν ευκολίας καθώς και οπτικός έλεγχος μεταξύ των αρχείων συζευγμένης και μη μορφής ως προς την αρίθμηση και την ακολουθία τους. Ο οπτικός έλεγχος πραγματοποιήθηκε με την βοήθεια δύο προγραμμάτων οπτικοποίησης και επεξεργασίας αρχείων PDB, του UCSF Chimera και του PyMOL τα οποία προμηθευτήκαμε από τις ιστοσελίδες τους https://www.cgl.ucsf.edu/chimera/ και https://pymol.org/2/ αντίστοιχα. Σε περιπτώσεις που τα αρχεία εμφάνιζαν σχετικά μικρά χάσματα στις ακολουθίες, δηλαδή της τάξης των 10-12 αμινοξέων, η επιδιόρθωση πραγματοποιήθηκε με την χρήση του RCD+ (Chys *et* al., 2013) στην ιστοσελίδα του http://rcd.chaconlab.org, το οποίο απαιτεί την δομή με το χάσμα καθώς και την αλληλουχία των αμινοξέων του χάσματος.

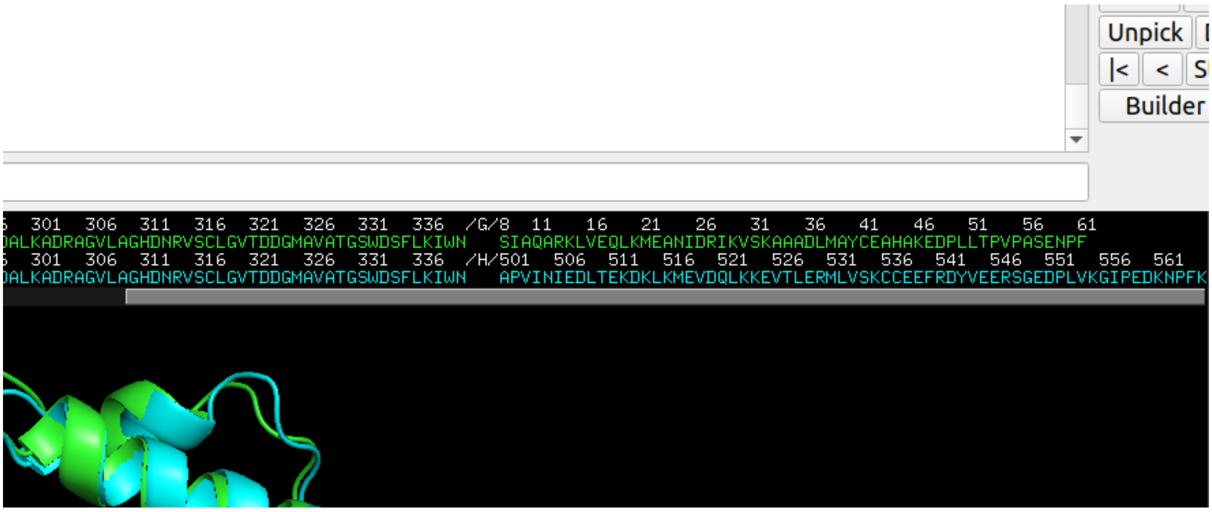
Κατά την σύγκριση των αρχείων των συζευγμένων και μη πρωτεϊνών, bound και unbound αντίστοιχα, διαπιστώθηκε πως σε ορισμένα σύμπλοκα, μεγάλα τμήματα καρβοξυτελικών ή αμινοτελικών άκρων έλειπαν σε ένα από τα δύο αρχεία, bound και unbound, ενώ υπήρχαν στο άλλο. Για την επιδιόρθωση τέτοιων προβλημάτων έγινε χρήση του Swiss-Pdbviewer έκδοσης 4.1 (Guex and Peitsch, 1996), το οποίο προμηθευτήκαμε από την ιστοσελίδα https://spdbv.unil.ch/. Η επιδιόρθωση αυτή ήταν απαραίτητη καθώς η διαφορά αυτή θα δημιουργούσε μεγάλες αποκλίσεις στις μεταβλητές αξιολόγησης. Το πρόγραμμα αυτό δίνει την δυνατότητα επεξεργασίας της δομής στα αρχεία pdb, επομένως επιλέχθηκε για την προσθήκη των ελλείψεων στα αρχεία από τα αντίστοιχα αρχεία στα οποία δεν υπήρχε η εν λόγω έλλειψη. Για την πραγματοποίηση αυτή γίνεται αντιγραφή των άκρων των αλυσίδων που υπάρχουν στο ένα αρχείο και προσθήκη τους στο αρχείο στο οποίο υπάρχει η έλλειψη. Έπειτα γίνεται χρήση της επιλογής "Ligate Backbone" του προγράμματος για να γίνει σύνδεση του τμήματος στην υπόλοιπη αλυσίδα. Σε ορισμένες περιπτώσεις έγινε χρήση αυτής της επιλογής του προγράμματος για την προσθήκη ελλείψεων και εντός της αλληλουχίας μιας αλυσίδας, όταν αυτή φυσικά υπήρχε μόνο στο ένα από τα δύο αρχεία, bound ή unbound. Το πρόγραμμα Swiss-Pdbviewer είναι το μοναδικό από αυτά που χρησιμοποιήθηκαν για την προετοιμασία των αρχείων, το οποίο χρησιμοποιήθηκε σε περιβάλλον Windows, σε αντίθεση με όλα τα υπόλοιπα που χρησιμοποιήθηκαν σε περιβάλλον Linux.

Η φάση της προετοιμασίας ορισμένων αρχείων των συμπλόκων 1GP2, 1HE8 και 1H1V, μας απέτρεψε από το να προχωρήσουμε στο επόμενο βήμα με τις προσομοιώσεις αλληλεπιδράσεων πρωτεϊνών μέσω των προγραμμάτων docking. Συγκεκριμένα:

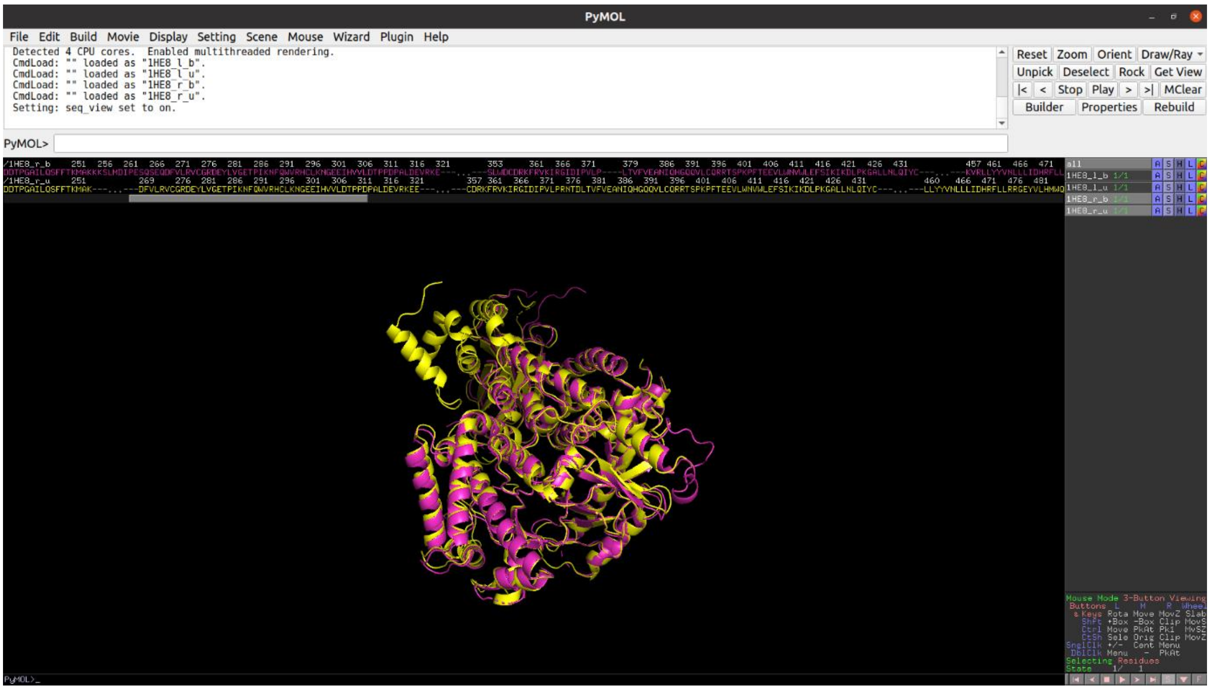
- Στο σύμπλοκο 1GP2, το οποίο ανήκει στην κατηγορία μεσαίας δυσκολίας και αφορά σε μια ετεροτριμερή πρωτεΐνη, η αλυσίδα Gamma είχε την δυνατότητα να εμφανίζεται με δύο διαφορετικές αλληλουχίες και επομένως διαμόρφωση στον χώρο. Αυτό το φαινόμενο εμφανιζόταν μεταξύ των αρχείων bound και unbound, το οποίο καθιστούσε αδύνατο τον υπολογισμό και την σύγκριση των μεταβλητών που σχετίζονταν με τα φορτία των ατόμων. (Εικόνα 3 και 4)
- Στο σύμπλοκο 1HE8, το οποίο ανήκει στην κατηγορία μεσαίας δυσκολίας, αντιμετωπίσαμε διαφορετικά προβλήματα σχετικά με την δομή του. Συγκεκριμένα τα αρχεία του συμπλόκου, bound και unbound, περιείχαν πολλά αλληλουχικά χάσματα που δεν είχαμε την δυνατότητα να τα αντιμετωπίσουμε με κανέναν από τους δύο τρόπους που είχαμε στην διάθεσή μας με αξιοπιστία. Αυτό ίσχυε καθώς τα κενά αυτά εμφανίζονταν και στα δύο αρχεία pdb, ενώ ταυτόχρονα άγγιζαν τα 30 αμινοξέα σε μήκος (Εικόνα 5 και 6).
- Στο σύμπλοκο 1H1V, το οποίο ανήκει στην κατηγορία υψηλής δυσκολίας, βρέθηκε πως οι απομονώσεις των πρωτεϊνών από τις οποίες προέκυψαν τα αρχεία bound και unbound είχαν πραγματοποιηθεί σε διαφορετικά είδη οργανισμών. Για τον λόγο αυτό τα αρχεία του ligand είχαν μια διαφορά μεταξύ τους 400 περίπου αμινοξέων, ενώ μετά από την στοίχιση των αλληλουχιών των αρχείων βρέθηκαν και σημειακές μεταλλάξεις. (Εικόνα 7)



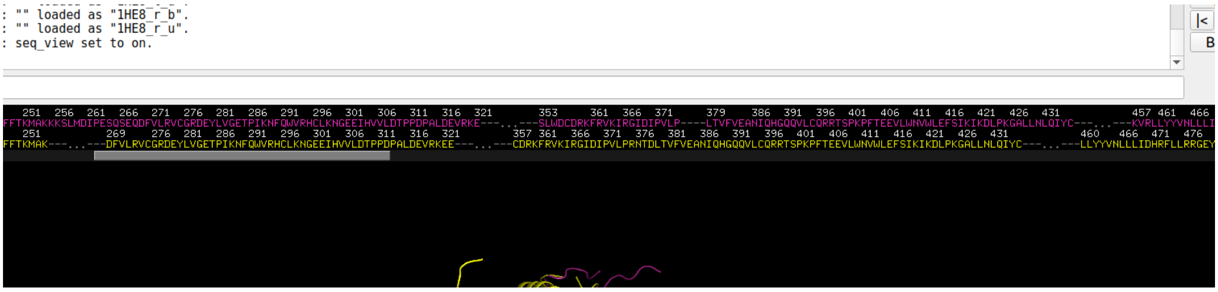
Εικόνα 3. Ο προσδέτης του συμπλόκου 1GP2 στις δύο καταστάσεις, bound και unbound. Η οπτικοποίηση γίνεται με το πρόγραμμα pyMOL. Στην εικόνα με πράσινο απεικονίζεται ο προσδέτης στην bound κατάσταση, ενώ με γαλάζιο ο προσδέτης στην unbound κατάσταση.



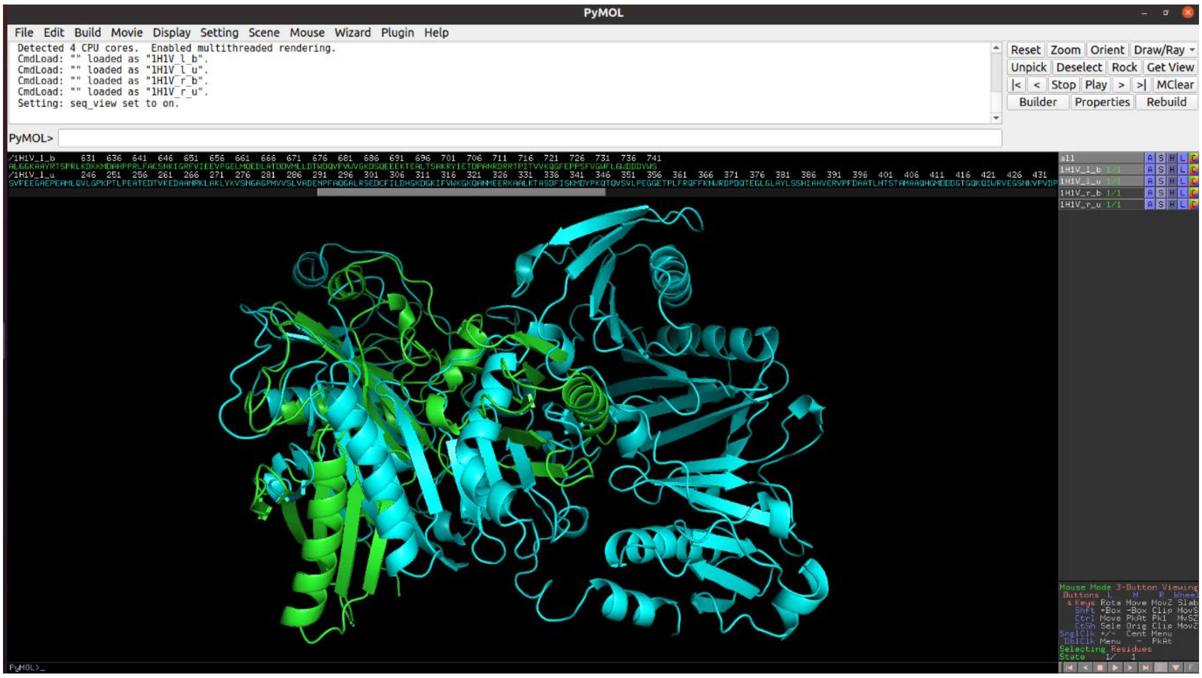
Εικόνα 4. Μεγέθυνση τμήματος εικόνας 1. Συγκεκριμένα απεικονίζονται οι αλληλουχίες του προσδέτη του συμπλόκου 1GP2 στις καταστάσεις bound και unbound. Στην εικόνα με πράσινο απεικονίζονται τα αμινοξέα του προσδέτη στην bound κατάσταση, ενώ με γαλάζιο τα αμινοξέα του προσδέτη στην unbound κατάσταση.



Εικόνα 5. Ο υποδοχέας του συμπλόκου 1HE8 στις δύο καταστάσεις, bound και unbound. Η οπτικοποίηση γίνεται με το πρόγραμμα pyMOL. Στην εικόνα με ματζέντα απεικονίζεται ο υποδοχέας στην bound κατάσταση, ενώ με κίτρινο ο υποδοχέας στην unbound κατάσταση.



Εικόνα 6. Μεγέθυνση τμήματος εικόνας 3. Συγκεκριμένα απεικονίζονται οι αλληλουχίες του υποδοχέα του συμπλόκου 1HE8 στις καταστάσεις bound και unbound. Στην εικόνα με ματζέντα απεικονίζονται τα αμινοξέα του υποδοχέα στην bound κατάσταση, ενώ με κίτρινο τα αμινοξέα του υποδοχέα στην unbound κατάσταση.



Εικόνα 7. Ο προσδέτης του συμπλόκου 1H1V στις δύο καταστάσεις, bound και unbound. Η οπτικοποίηση γίνεται με το πρόγραμμα pyMOL. Στην εικόνα με πράσινο απεικονίζεται ο προσδέτης στην bound κατάσταση, ενώ με γαλάζιο ο προσδέτης στην unbound κατάσταση.

2.2.1. Πραγματοποίηση docking με HADDOCK

Η πρόσδεση πρωτεΐνης - πρωτεΐνης με τον web server του HADDOCK προϋποθέτει την εύρεση των "ενεργών" καταλοίπων κάθε πρωτεΐνης στα αρχεία πριν την πρόσδεση (unbound). Τα «ενεργά» κατάλοιπα αντιστοιχούν σε αυτά που θεωρείται ότι συμμετέχουν σε αλληλεπίδραση με άλλες πρωτεΐνες. Για την εύρεση τους έγινε χρήση του εργαλείου CPORT (Vries *et al.*, 2011) του προγράμματος στην ιστοσελίδα http://alcazar.science.uu.nl/services/CPORT/ , δίνοντας το υπό διερεύνηση αρχείο pdb και το όνομα της αλυσίδα της πρωτεΐνης. Η πληροφορία αυτή μαζί με τα αρχεία "unbound" των πρωτεϊνών δίνονται στο HADDOCK στην σελίδα τους https://wenmr.science.uu.nl/haddock2.4/ για την πραγματοποίηση της πρόσδεσης με τις προκαθορισμένες ρυθμίσεις. Με το τέλος της πρόσδεσης γίνεται κατέβασμα του συνόλου των αποτελεσμάτων. Στον φάκελο υπάρχουν 1000 πιθανά μοντέλα συμπλόκων εκ των οποίων τα 200 καλύτερα με βάση την αξιολόγηση του HADDOCK απομονώθηκαν και χρησιμοποιήθηκαν για τις μετέπειτα επεξεργασίες και υπολογισμούς.

2.2.2. Πραγματοποίηση docking με InterEvDock2

Για την χρήση του web server του InterEvDock2 είναι απαραίτητη να του διατεθούν τα αρχεία “unbound” των πρωτεϊνών. Η προσθήκη πραγματοποιείται στην ιστοσελίδα https://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::InterEvDock2. Και σε αυτόν τον web server έγινε χρήση μόνο των προκαθορισμένων ρυθμίσεων. Τα αποτελέσματα κατέβηκαν από την ιστοσελίδα και από αυτά απομονώθηκαν τα καλύτερα με βάση την τριών ειδών αξιολόγηση του προγράμματος. Τα αποτελέσματα αυτά ήταν 50 με βάση την αξιολόγηση FRODOCK, 50 με βάση την αξιολόγηση του InterEvScore και 50 με βάση SOAP_PP. Τα αποτελέσματα είναι συνολικά λιγότερα από 150, καθώς ορισμένα από αυτά συνέπιπταν μεταξύ των καλύτερων από τις τρεις αξιολογήσεις. Αυτά μετέπειτα χρησιμοποιήθηκαν για επεξεργασία και υπολογισμούς.

2.2.3. Προσθήκη μερικών φορτίων στα αρχεία

Καθώς τα αρχεία PDB που βρίσκονται στην διάθεσή μας δεν εμπεριέχουν τα ηλεκτρικά φορτία των ατόμων καθώς και τα άτομα υδρογόνου, δεν θα ήταν δυνατός ο υπολογισμός του DML και του R_g. Για τον λόγο αυτό είναι αναγκαίο να γίνει πρώτα η προσθήκη των υδρογόνων έτσι ώστε να πραγματοποιηθεί μετέπειτα ο υπολογισμός του πραγματικού R_g καθώς και η προσθήκη των μερικών φορτίων στα άτομα για τον υπολογισμό του DML. Για την πραγματοποίηση αυτής της προσθήκης έγινε χρήση του προγράμματος PDB2PQR (Dolinsky *et al.*, 2004).

Το PDB2PQR αποτελεί πρόγραμμα που λειτουργεί σε περιβάλλον Linux και απαιτεί υπολογιστική δύναμη. Προκειμένου να γίνει η χρήση του, το πρόγραμμα εγκαταστάθηκε σε τοπικό διακομιστή (server) περιβάλλοντος Linux. Η διαδικασία προσθήκης των υδρογόνων και τον φορτίων έγινε έπειτα αυτοματοποιημένα σε όλα τα μοντέλα κάθε συμπλόκου, με την χρήση ενός “bash script” που είναι διαθέσιμο στο *Παράρτημα Β1*.

2.3. Υπολογισμός των μεταβλητών

Ο υπολογισμός των μεταβλητών πραγματοποιήθηκε για τα μοντέλα που προέκυψαν από το HADDOCK και το InterEvDock2 για κάθε σύμπλοκο, καθώς και του πραγματικού συμπλόκου με την πρόσδεση των δύο bound αρχείων για κάθε σύμπλοκο. Για τον υπολογισμό του DML και του R_g έγινε χρήση του προγράμματος οπτικοποίησης VMD, το οποίο μπορεί να διαβάζει αρχεία όπως PDB και PQR, ενώ δίνει και την δυνατότητα υπολογισμού πολλών μεταβλητών συμπεριλαμβανομένων και των παραπάνω με χρήση scripts σε γλώσσα Tcl. Το πρόγραμμα VMD (Humphrey *et al.*, 1996) εγκαταστάθηκε μέσω της σελίδας https://www.ks.uiuc.edu/Research/vmd/. Το ίδιο πρόγραμμα χρησιμοποιήθηκε και για τον υπολογισμό του RMSD μεταξύ των δομών.

Για τον υπολογισμό του συνολικού pK_a έγινε χρήση του προγράμματος PROPKA, το οποίο εγκαταστάθηκε μαζί με το PDB2PQR, καθώς εμπεριέχεται στο πακέτο αυτού. Τέλος το DockQ υπολογίστηκε από το ομώνυμο πρόγραμμα (Vajda *et al.*, 2013, Ferreira *et al.*, 2015, Kirchmair *et al.*, 2008, Chen Y.C., 2014), που διατίθεται στην ιστοσελίδα http://github.com/bjornwallner/DockQ/.

2.3.1. Υπολογισμός των διπολικών ροπών (DML)

Για τον υπολογισμό του DML, όπως προαναφέρθηκε, έγινε με βάση την εξίσωση 1 μέσω του προγράμματος VMD. Για την αυτοματοποίηση της διαδικασίας υπολογισμού για όλα τα μοντέλα από τα προγράμματα docking έγινε χρήση ενός Tcl script με την ονομασία “`DML_calculation_pqr.tcl`”. Η διαδικασία αυτή πραγματοποιήθηκε για όλα τα μοντέλα των συμπλόκων, καθώς και για το σύμπλοκο bound έτσι ώστε να μπορεί να γίνει η σύγκρισή τους. Η διαφορά μεταξύ των script για το HADDOCK και για το InterEvDock2 βρισκόταν μόνο στην ονομασία των αρχείων και τον αριθμό τους. Το Tcl script για την αυτοματοποίηση υπολογισμού του DML από το VMD είναι διαθέσιμο ολόκληρο στο *Παράρτημα B2*.

Μετά από τον υπολογισμό των DML όλων των συμπλόκων, πραγματοποιήθηκε η επεξεργασία τους μέσω του Excel με την χρήση της εξίσωσης 7, για τον υπολογισμό της ποσοστιαίας απόλυτης απόκλισης μεταξύ των μοντέλων και των bound συμπλόκων.

$$dDML\% = \frac{|DipoleLength_{bound} - DipoleLengnth_{model}|}{DipoleLength_{bound}} * 100\% \qquad \text{Εξίσωση 7}$$

2.3.2. Υπολογισμός των γυροσκοπικών ακτίνων (R_g)

Για την αυτοματοποίηση του υπολογισμού της R_g με βάση την εξίσωση 2 για όλα τα μοντέλα του docking του κάθε συμπλόκου έγινε χρήση του VMD και του Tcl script “`Rg_calc.tcl`”. Τα αποτελέσματα αποθηκεύονταν σε αρχεία TXT για την περαιτέρω επεξεργασία τους. Το Tcl script για την αυτοματοποίηση του υπολογισμού της R_g από είναι διαθέσιμο στο *Παράρτημα B3*.

Από τις τιμές της R_g των μοντέλων $R_{g.model}$ και της γνωστής δομής $R_{g.bound}$ υπολογίσθηκε με την βοήθεια του Excel για κάθε μοντέλο η ποσοστιαία απόλυτη απόκλιση της R_g από εκείνη του συμπλόκου bound κάνοντας χρήση της εξίσωσης 8.

$$dR_g\% = \frac{\left|R_{g,bound} - R_{g,model}\right|}{R_{g,bound}} * 100\% \qquad \text{Εξίσωση 8}$$

2.3.3. Υπολογισμός των pK_a

Για τον υπολογισμό του συνολικού pK_a των μοντέλων έγινε χρήση του PROPKA, ενός εργαλείου του PDB2PQR μέσω του "script αυτοματοποίηση του PROPKA" (B4 Παράρτημα). Τα αποτελέσματα της εκτέλεσης αυτού του script αποθηκεύονταν για κάθε ένα μοντέλο σε ένα αρχείο "`.propka`". Τα αρχεία αυτά εμπεριέχουν πληροφορίες, όπως αλληλεπιδράσεις με άλλα κατάλοιπα, την ενέργεια πρόσδεσης καθώς και τις τιμές pK_a για κάθε κατάλοιπο. Ένα παράδειγμα τέτοιου αρχείου υπάρχει στο *Παράρτημα Α3*.

Από τα μεμονωμένα αρχεία "`.propka`" που προέκυψαν στο προηγούμενο βήμα έγινε υπολογισμός του συνολικού pK_a κάθε μοντέλου με το bash script "`get_total-pKa_loop_compl.sh`" ως άθροισμα των επιμέρους pK_a κάθε καταλοίπου (Παράρτημα B5). Το συγκεκριμένο script απομονώνει από τα αρχεία PROPKA μόνο τις πληροφορίες με τα pK_a που βρίσκονται κάτω από τον πίνακα "SUMMARY OF THE PREDICTION" και έπειτα απομονώνει και προσθέτει μόνο την τέταρτη στήλη του, που περιέχει τα pK_a. Αυτά αποθηκεύονται σε ένα αρχείο TXT για την περαιτέρω ανάλυση. Η παραπάνω διαδικασία πραγματοποιείται σε όλα τα μοντέλα των συμπλόκων, καθώς και στα unbound αρχεία pdb του υποδοχέα και του προσδέτη ξεχωριστά, τα αρχεία δηλαδή που χρησιμοποιήθηκαν από το HADDOCK και το InterEvDock2 για την παραγωγή των μοντέλων.

Ακολούθως, υπολογίσθηκε η απόλυτη τιμή της απόκλισης των pK_a με την βοήθεια του Excel και της εξίσωσης 9:

$$\Delta pK_a = \left|\left(Total\ pK_{a,bound\ receptor} + Total\ \ pK_{a,bound\ ligand}\right) - Total\ \ pK_{a,model}\right| \qquad \text{Εξίσωση 9}$$

2.3.4. Υπολογισμός RMSD

Όπως προαναφέρθηκε, το RMSD υπολογίστηκε με την βοήθεια του VMD με σκοπό την αξιολόγηση των παραπάνω μεταβλητών για την εύρεση των καλύτερων μοντέλων. Για τον υπολογισμό του χρησιμοποιήθηκε το Tcl script "`RMSD_calculation.tcl`" (Παράρτημα B6). Το script κάνει πρώτα δομική υπέρθεση κάθε ενός μοντέλου επάνω στην bound δομή με βάση τα άτομα του πρωτεϊνικού σκελετού ($C_α$, N, C, O). Για την πραγματοποίηση της δομικής υπέρθεσης απαιτείται τα δύο σύμπλοκα να έχουν ακριβώς τα ίδια κατάλοιπα και κατ' επέκταση άτομα, πράγμα που πρέπει να αντιμετωπισθεί στην προετοιμασία των αρχείων. Τα αποτελέσματα αυτά αποθηκεύονται σε αρχείο TXT.

2.3.5. Υπολογισμός DockQ

Για τον υπολογισμό του DockQ χρησιμοποιήθηκε το ομώνυμο πρόγραμμα σε περιβάλλον Linux. Καθώς το πρόγραμμα αυτό χρησιμοποιεί για τον υπολογισμό ορισμένες εντολές που δεν υπάρχουν στην γλώσσα προγραμματισμού Python, ήταν αναγκαίο να γίνει εγκατάσταση και των βιβλιοθηκών των Biopython και Numpy έτσι ώστε να λειτουργήσει. Το πρόγραμμα DockQ.py εκτελέσθηκε επαναληπτικά για όλα τα μοντέλα με την βοήθεια του script “`run_DockQ.sh`” που είναι διαθέσιμο στο *Παράρτημα B7*. Τα αποτελέσματα αποθηκεύτηκαν σε αρχείο TXT. Καθώς τα αρχεία αυτά δεν περιείχαν μόνο τις τιμές DockQ, αλλά και πληροφορίες σχετικά με τον διαχωρισμό σε ομάδες που κάνει το πρόγραμμα, χρησιμοποιήθηκε ένα ακόμα bash script για την απομόνωσή τους. Το script με όνομα “`collect_dockq.sh`” έχει την δυνατότητα απομόνωσης συγκεκριμένης γραμμής από αρχεία και είναι διαθέσιμο στο *Παράρτημα B8*.

2.4. Συλλογή υπολογισμών

Οι παραπάνω υπολογισμοί συλλέχθηκαν σε πίνακες Excel για κάθε ένα σύμπλοκο και πρόγραμμα docking ξεχωριστά, με σκοπό την μετέπειτα διαλογή των καλύτερων μοντέλων με βάση τις τρεις μεταβλητές που εξετάσαμε. Στους πίνακες εμπεριέχονται οι αριθμοί κατάταξης των συμπλόκων με βάση το αντίστοιχο πρόγραμμα docking, η ποσοστιαία απόκλιση του DML, η ποσοστιαία απόκλιση του R_g, η διαφορά του pKa, καθώς και οι δύο μεταβλητές αξιολόγησης, RMSD και DockQ.

Η στατιστική ανάλυση πραγματοποιήθηκε με το πρόγραμμα SPSS που προμηθευτήκαμε από την ιστοσελίδα https://www.ibm.com/analytics/spss-statistics-software , και συγκεκριμένα έγινε χρήση της γραμμικής παλινδρόμησης (linear regression) ως μέθοδος ανάλυσης. Για την πραγματοποίησή της έγινε προσθήκη των δεδομένων στο πρόγραμμα ξεχωριστά για το HADDOCK και για το InterEvDock2. Προέκυψαν δύο γραμμικές σχέσεις, μία για την πρόβλεψη του RMSD και μία για του DockQ (εξισώσεις 10 και 11):

$$RMSD = a \cdot dDML + b \cdot dR_g + c \cdot \Delta pK_a + d \qquad \text{Εξίσωση 10}$$

και

$$DockQ = a \cdot dDML + b \cdot dR_g + c \cdot \Delta pK_a + d \qquad \text{Εξίσωση 11}$$

3. Αποτελέσματα

Τα δεδομένα που χρησιμοποιήθηκαν για τις παρακάτω αναλύσεις προέρχονται από 27 σύμπλοκα τις λίστας αναφοράς της βάσης δεδομένων PDB. Τα 10 από αυτά προέρχονται από τους υπολογισμούς στο πλαίσιο της πτυχιακής εργασίας της Γεωργίας-Μαρίας Κεφαλά (Kefala Georgia-Maria, 2021) , ενώ τα υπόλοιπα 17 προέρχονται από τους υπολογισμούς στα 20 σύμπλοκα που έχουν αναφερθεί στην ενότητα 2.2. Τα τρία από τα 20 σύμπλοκα που δεν εμπεριέχονται στα αποτελέσματα αφορούν σε σύμπλοκα με πολλά προβλήματα στα αρχεία PDB που μας οδήγησαν στον αποκλεισμό τους. Από τα 27 σύμπλοκα, μόνα τα 22 χρησιμοποιήθηκαν για την παραγωγή της γραμμικής εξίσωσης, ενώ τα υπόλοιπα 5 χρησιμοποιήθηκαν για την αξιολόγηση της απόδοσής της να προβλέπει το καλύτερο μοντέλο. Τα σύμπλοκα αυτά είναι: 1MAH (RB), 1KKL (MD), 1HIA (RB), 1GRN (MD) και 1FQ1 (HD), τα οποία ανήκουν σε όλες τις ομάδες δυσκολίας της λίστας.

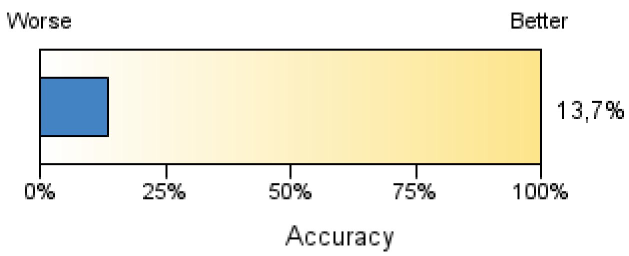
3.1. Γραμμική εξίσωση με τις 3 μεταβλητές

Για την πραγματοποίηση της γραμμικής παλινδρόμησης στο πρόγραμμα SPSS χρησιμοποιήθηκαν ως επεξηγηματικές ή ελεγχόμενες μεταβλητές τα dDML, dRg και ΔpKa, ενώ οι μεταβλητές αξιολόγησης προστέθηκαν ως στόχοι (εξαρτημένες μεταβλητές ή απόκρισης) το RMSD ή το DockQ αναλόγως.

3.1.1. Αξιολόγηση με RMSD

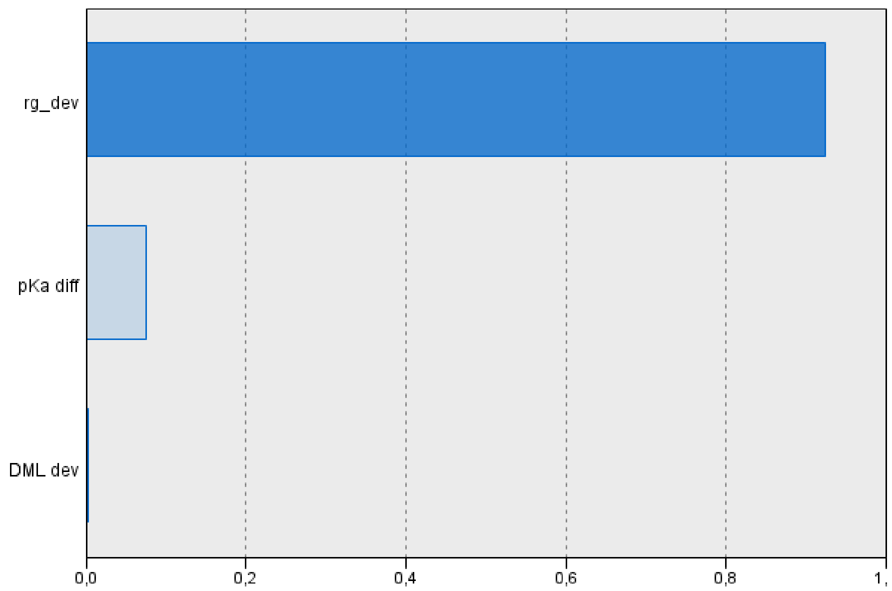
3.1.1.1. Σύμπλοκα HADDOCK

Όπως φαίνεται και στο πρώτο διάγραμμα που προέκυψε από τις αναλύσεις των τιμών, η γραμμική εξίσωση που προέκυψε δεν αποτελούσε ικανοποιητικό μοντέλο πρόβλεψης του RMSD και σωστής αξιολόγησής των μοντέλων, καθώς η ακρίβειά του δεν ξεπερνούσε το 14%. Παρόλο που η τιμή αυτή απορρίπτει το μοντέλο αξιολόγησης, είναι δυνατόν να προκύψουν χρήσιμες πληροφορίες για τις μεταβλητές που χρησιμοποιήσαμε.



Διάγραμμα 1. Ακρίβεια μοντέλου αξιολόγησης με στόχο το RMSD για τα μοντέλα συμπλόκων από το HADDOCK. Στο διάγραμμα απεικονίζεται το ποσοστό των μοντέλων των οποίων οι μεταβλητές προσαρμόζονται στην γραμμική εξίσωση με μικρές αποκλίσεις από το σύνολο των μοντέλων.

Στο διάγραμμα 2 παρατηρείται η σχετική σημαντικότητα των μεταβλητών μεταξύ τους. Με βάση αυτό το διάγραμμα είναι εύκολο να διακρίνουμε τη σειρά με την οποία το SPSS θεωρεί σημαντικές τις μεταβλητές. Προς έκπληξή μας δεν φαίνεται να παίζουν καθοριστικό ρόλο και οι δύο δομικές μεταβλητές, αλλά μόνο η R_g. Ταυτόχρονα σε σχέση με την R_g, η διαφορά του pK_a μεταξύ των συζευγμένων και μη πρωτεϊνών δεν αποτέλεσε καθοριστική μεταβλητή στην πρόβλεψη.



Διάγραμμα 2. Σχετική σημαντικότητα των μεταβλητών στο μοντέλο αξιολόγησης με στόχο το RMSD για τα μοντέλα συμπλόκων από το HADDOCK. Στον άξονα y αντιστοιχούν οι μεταβλητές dR_g, ΔpKa και dDML, ενώ στον άξονα x η σημαντικότητά τους με μέγιστο 1. Συγκεκριμένα η σημαντικότητα των μεταβλητών είναι 0,924 για την R_g, 0,074 για το pKa και μόλις 0,001 για το DML.

Η γραμμική εξίσωση για τα μοντέλα των συμπλόκων που προέρχονται από το HADDOCK με στόχο το RMSD, έχει συντελεστές για την εξίσωση 10: a=-0,003, b=0,743, c=0,005 και d=11,692.

Παρακάτω παρατίθενται τα αποτελέσματα της εφαρμογής των γραμμικών μοντέλων στο σύμπλοκο 1FQ1, το οποίο αποτελεί σύμπλοκο υψηλής δυσκολίας. Όπως είναι εμφανές από τον πίνακα 1, οι τιμές αξιολόγησής τους, δηλαδή DockQ και RMSD, χαρακτηρίζουν τα μοντέλα ως λανθασμένα. Συγκεκριμένα οι τιμές των DockQ είναι μικρότερες από 0,23 και ταυτόχρονα οι τιμές του RMSD είναι μεγαλύτερες από 2 Å. Παρόλο που τα μοντέλα έχουν ταξινομηθεί με βάση τα 10 καλύτερα μοντέλα σύμφωνα με το μοντέλο γραμμικής παλινδρόμησης, δεν φαίνεται οπτικά να ακολουθείται κάποιο μοτίβο πέραν της σχεδόν πιστής ακολουθίας της κατάταξης της R_g. Αυτό είναι εμφανές και από το διάγραμμα 2 που υποδηλώνει την υψηλή σημαντικότητα της R_g. Σε αντίθεση με την R_g, τα DML και pK_a δεν ακολουθούν την αύξουσα σειρά του μοντέλου, ενώ φαίνεται πως το καλύτερο μοντέλο με βάση το μοντέλο αξιολόγησης μας αντιστοιχεί σε αρκετά χαμηλότερη κατάταξη σε αυτές τις δύο μεταβλητές. Όσον αφορά την κατάταξη των μοντέλων από το πρόγραμμα που πραγματοποίησε το docking, φαίνεται πως ορισμένα έχουν υψηλή ενώ άλλα χαμηλή σειρά κατάταξης, χωρίς να ακολουθούν κάποιο μοτίβο. Η σειρά κατάταξης από το ίδιο το πρόγραμμα είναι διαθέσιμη στην ονομασία κάθε μοντέλου.

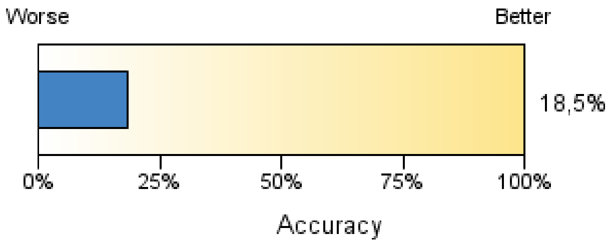
Πίνακας 1. Υπολογισμοί μεταβλητών του συμπλόκου 1FQ1 που προέκυψαν από μοντέλα του HADDOCK. Στον πίνακα φαίνονται οι μεταβλητές και η σχετική τους θέση στο σύνολο των μοντέλων του συμπλόκου από την καλύτερη στην χειρότερη.

	DML dev	DML rank	Rg dev	Rg rank	pKa diff	pKa rank	DockQ	DockQ rank	RMSD	RMSD rank	LiReg. RMSD	Model RMSD rank
1FQ1_HD_8	4,648443	53	0,065715	1	3,24	34	0,073	7	8,079714	7	11,7430811	1
1FQ1_HD_195	3,025694	34	0,101885	2	2,48	27	0,011	192	20,22361	193	11,77102378	2
1FQ1_HD_66	32,56945	195	0,345766	3	6	62	0,017	169	18,03267	184	11,88119599	3
1FQ1_HD_125	4,931086	58	0,353669	4	6,59	70	0,009	197	21,87285	200	11,97293276	4
1FQ1_HD_188	17,67587	119	0,545352	7	1,5	16	0,023	116	16,08559	148	12,05166925	5
1FQ1_HD_106	2,583018	27	0,454766	5	14,3	172	0,09	2	6,566229	1	12,09364236	6
1FQ1_HD_72	32,71863	196	0,618794	9	9,08	111	0,02	147	13,55693	74	12,09900835	7
1FQ1_HD_99	0,253789	2	0,528117	6	6,63	72	0,038	57	8,195863	9	12,11677921	8
1FQ1_HD_133	19,57089	127	0,569106	8	12,32	156	0,024	113	14,23892	85	12,11773333	9
1FQ1_HD_175	24,38697	154	0,654643	10	7,33	79	0,021	133	14,9897	114	12,14188903	10

Το ίδιο πρόβλημα εμφανίζεται και στα μοντέλα HADDOCK των υπόλοιπων συμπλόκων όσον αφορά στις ταξινομήσεις με βάση το γραμμικό μοντέλο. Καθώς τα μοντέλα δεν πληρούν τις προϋποθέσεις που έχουν τεθεί από τα DockQ και RMSD, δηλαδή τις ουδούς των 0,23 και 2 Å αντίστοιχα, που προαναφέρθηκαν, δεν αντιστοιχούν σε «αποδεκτά» μοντέλα για κανένα εκ των τεσσάρων συμπλόκων. Μοναδική εξαίρεση αποτελεί το σύμπλοκο 1MAH, το οποίο ανήκει στην κατηγορία “χαμηλής δυσκολίας”, όπου τρία εκ των διακοσίων μοντέλων χαρακτηρίζονται ως «αποδεκτά».

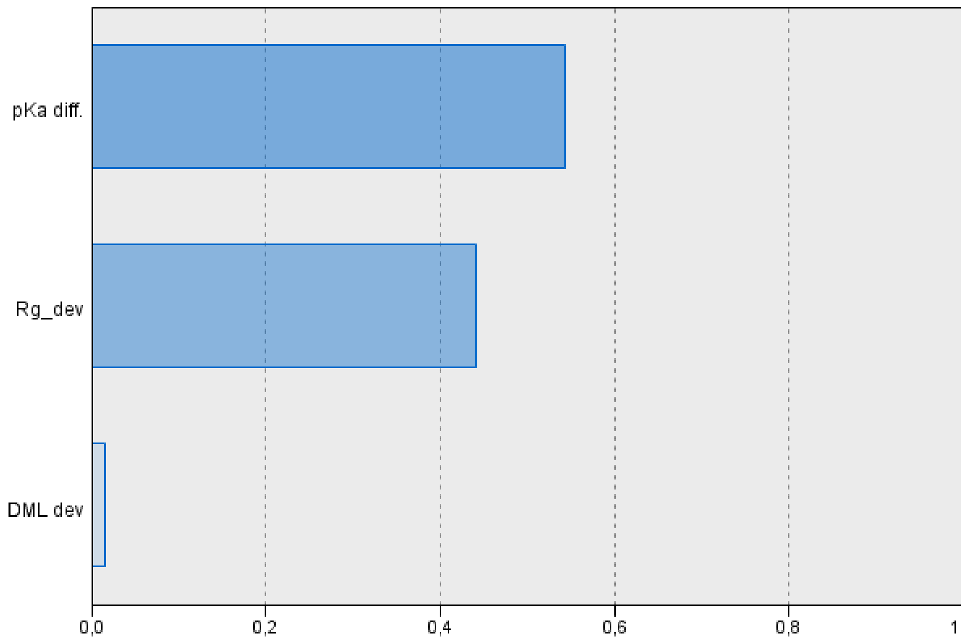
3.1.1.2. Σύμπλοκα InterEvDock2

Στο τρίτο διάγραμμα που προέκυψε από τις αναλύσεις των τιμών, η γραμμική εξίσωση για τα μοντέλα του InterEvDock δεν αποτελούσε ικανοποιητικό μοντέλο πρόβλεψης νέων μοντέλων και σωστής αξιολόγησής τους, καθώς η ακρίβειά του δεν ξεπερνούσε το 20%. Όπως και για τα μοντέλα του HADDOCK η τιμή αυτή δεν είναι επαρκής για να θεωρηθεί το γραμμικό μοντέλο αξιολόγησης χρήσιμο. Παρόλα αυτά είναι δυνατόν να προκύψουν χρήσιμες πληροφορίες για τις μεταβλητές που χρησιμοποιήσαμε.



Διάγραμμα 3. Ακρίβεια του γραμμικού μοντέλου αξιολόγησης με στόχο το RMSD για τα μοντέλα συμπλόκων από το InterEvDock2. Στο διάγραμμα απεικονίζεται το ποσοστό των μοντέλων των οποίων οι μεταβλητές προσαρμόζονται στην γραμμική εξίσωση με μικρές αποκλίσεις από το σύνολο των μοντέλων.

Η γραμμική παλινδρόμηση για την πρόβλεψη του RMSD με βάση τα μοντέλα του InterEvDock2 έδωσε διαφορετική εικόνα για την αξιολόγηση της σχετικής σημαντικότητας των dDML, dR_g και ΔpKa. Εδώ φαίνεται το ΔpK_a να παίζει σημαντικότερο ρόλο ακολουθούμενο από την dR_g.



Διάγραμμα 4. Σχετική σημαντικότητα των μεταβλητών στο μοντέλο αξιολόγησης με στόχο το RMSD για τα μοντέλα συμπλόκων από το InterEvDock2. Στον άξονα y αντιστοιχούν οι μεταβλητές dR_g, $ΔpK_a$ και dDML, ενώ στο άξονα x η σημαντικότητά τους με μέγιστο 1. Συγκεκριμένα η σημαντικότητα των μεταβλητών είναι 0,442 για την dR_g, 0,544 για το $ΔpK_a$ και μόλις 0,015 για το dDML.

Η γραμμική εξίσωση για τα μοντέλα των συμπλόκων που προέρχονται από το InterEvDock2 με στόχο το RMSD, έχει συντελεστές για την εξίσωση 10: a=0,012, b=0,359, c=0,018 και d=12,204.

Για την σύγκριση των αποτελεσμάτων μεταξύ του HADDOCK και InterEvDock2 βλέπουμε στον πίνακα 2 τα 10 καλύτερα μοντέλα του συμπλόκου 1FQ1 με βάση το γραμμικό μοντέλο. Για τα μοντέλα αυτά ισχύει πως δεν είναι αποδεκτά καθώς δεν βρίσκονται εντός των ορίων για τις μεταβλητές DockQ και RMSD. Σε αντίθεση με τα μοντέλα του HADDOCK, για το InterEvDock2 δεν φαίνεται να υπάρχει κάποιο μοτίβο με καμία από τις τρεις μεταβλητές. Καθώς οι κατατάξεις που δίνονται από το InterEvDock2 είναι με βάση τρεις διαφορετικές αξιολογήσεις, στις ονομασίες των μοντέλων αναγράφονται και η κατάταξη που αντιστοιχεί στο κάθε ένα αλλά και ο τρόπος αξιολόγησης. Τα μοντέλα με βάση την κατάταξη που δόθηκε από το πρόγραμμα, φαίνεται να έχουν κυρίως χαμηλή κατάταξη.

Για το σύμπλοκο 1GRN που ανήκει στην κατηγορία “μεσαίας δυσκολίας”, τα αποτελέσματα και η κατάταξή τους φαίνονται στον πίνακα 3. Για το σύμπλοκο αυτό πολλά από τα μοντέλα φαίνεται να είναι “αποδεκτά” σύμφωνα με τις μεταβλητές DockQ και RMSD. Συγκεκριμένα φαίνεται πως και οι κατατάξεις των μεταβλητών dR_g και dDML δίνουν επίσης υψηλές θέσεις στα 10 καλύτερα μοντέλα με βάση την γραμμική εξίσωση, ενώ το ίδιο παρατηρήθηκε και για τις μεταβλητές αξιολόγησης. Και για το σύμπλοκο αυτό η κατάταξη των μοντέλων με βάση το πρόγραμμα είναι χαμηλή.

Πίνακας 2. Υπολογισμοί μεταβλητών του συμπλόκου 1FQ1 που προέκυψαν από μοντέλα του InterEvDock2. Στον πίνακα φαίνονται οι μεταβλητές και η σχετική τους θέση στο σύνολο των μοντέλων του συμπλόκου από την καλύτερη στην χειρότερη.

	DML dev	**DML rank**	**Rg dev**	**Rg rank**	**pKa diff**	**pKa rank**	**DockQ**	**DockQ rank**	**RMSD**	**RMSD rank**	**LiReg. RMSD**	**Model RMSD rank**
1FQ1_FRODOCK_45	5,00959855	39	0,12269	2	6,2	78	0,01	125	19,825	123	12,41975957	1
1FQ1_IES_41	1,10728151	7	0,55535	10	2,16	29	0,025	55	15,4173	62	12,45553848	2
1FQ1_SOAP_PP_39	6,34502589	52	0,63208	12	0,46	6	0,026	50	15,4727	63	12,51533804	3
1FQ1_SOAP_PP_38	5,03441892	41	0,23456	4	9,97	101	0,027	49	14,9755	59	12,52808111	4
1FQ1_FRODOCK_4	2,98051521	23	0,58767	11	5	66	0,014	100	19,4362	120	12,54073817	5
1FQ1_FRODOCK_31	4,5994799	35	0,47051	7	6,71	80	0,023	59	12,9186	35	12,54888703	6
1FQ1_SOAP_PP_19	7,28769673	62	0,71129	13	2,46	39	0,031	37	14,4599	51	12,5910838	7
1FQ1_IES_28	3,30284528	28	0,96585	19	1,95	25	0,027	47	14,21	47	12,62547557	8
1FQ1_FRODOCK_12	5,15865258	42	0,30592	5	15,13	130	0,026	52	11,3648	13	12,64807055	9
1FQ1_IES_17	2,33775575	19	1,08919	23	2,43	37	0,025	56	14,7957	57	12,66681233	10

Πίνακας 3. Υπολογισμοί μεταβλητών του συμπλόκου 1GRN που προέκυψαν από μοντέλα του InterEvDock2. Στον πίνακα φαίνονται οι μεταβλητές και η σχετική τους θέση στο σύνολο των μοντέλων του συμπλόκου από την καλύτερη στην χειρότερη.

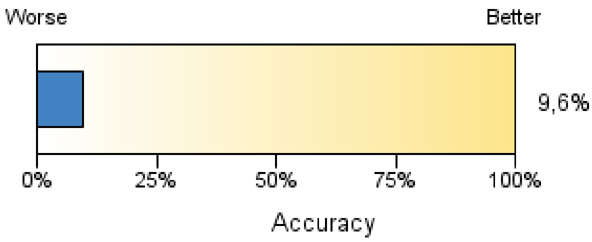
	DML dev	**DML rank**	**Rg dev**	**Rg rank**	**pKa diff**	**pKa rank**	**DockQ**	**DockQ rank**	**RMSD**	**RMSD rank**	**LiReg. RMSD**	**Model RMSD rank**
1GRN_FRODOCK_37	0,909609	19	0,057084	2	5,495427	62	0,455	5	2,920428	5	12,33432603	1
1GRN_FRODOCK_46	0,585265	13	0,281861	4	1,806971	11	0,752	1	1,0612	1	12,34473673	2
1GRN_IES_22	0,130111	3	0,404689	6	0,735828	8	0,043	48	20,31543	128	12,36408954	3
1GRN_FRODOCK_30	2,739985	41	0,378232	5	0,372571	3	0,524	4	2,104367	3	12,37937149	4
1GRN_FRODOCK_21	1,817184	27	0,245241	3	5,19737	53	0,538	2	2,051551	2	12,40740052	5
1GRN_FRODOCK_25	6,636497	69	0,419143	7	2,561428	22	0,089	24	10,06141	34	12,4802161	6
1GRN_FRODOCK_7	0,163849	4	0,609779	10	3,818856	35	0,078	26	5,724243	10	12,49361629	7
1GRN_FRODOCK_38	5,766991	62	0,463154	8	5,411598	59	0,041	52	14,23134	55	12,53688498	8
1GRN_FRODOCK_24	2,641407	39	0,883601	13	0,577486	6	0,052	39	19,3832	109	12,56330444	9
1GRN_FRODOCK_17	1,752298	26	0,892927	14	3,381085	31	0,537	3	2,112104	4	12,60644775	10

Για τα μοντέλα των υπόλοιπων συμπλόκων εκ των πέντε που χρησιμοποιήθηκαν για την αξιολόγηση, δεν εντοπίζονταν μοτίβα που να συσχέτιζαν με κάποιο τρόπο τις κατατάξεις μεταξύ του γραμμικού μοντέλου με τις μεταβλητές.

3.1.2. Αξιολόγηση με DockQ

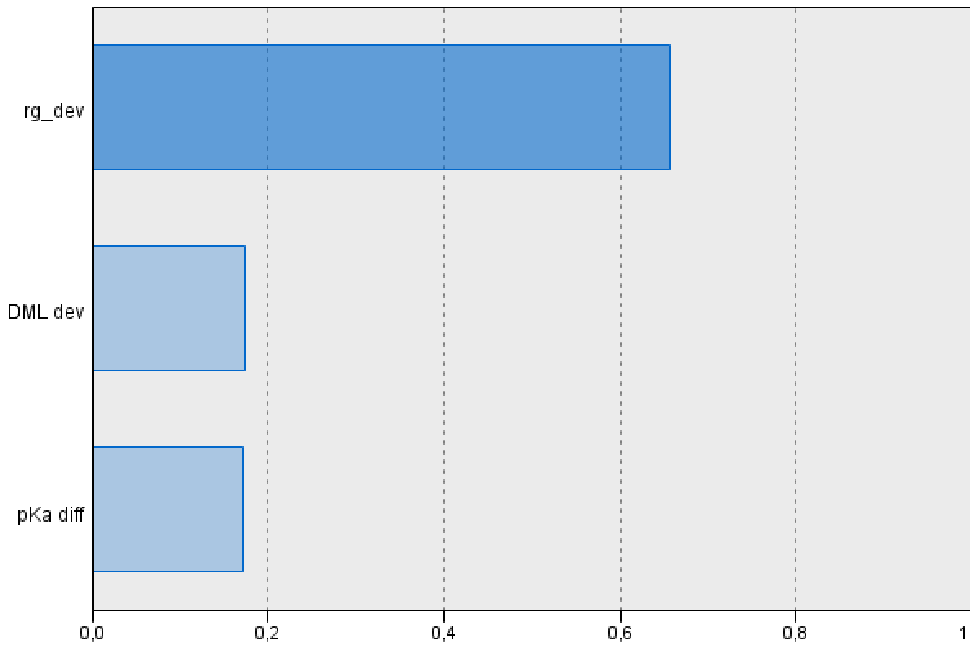
3.1.2.1. Σύμπλοκα HADDOCK

Στο πέμπτο διάγραμμα που προέκυψε από τις αναλύσεις των τιμών, η γραμμική εξίσωση για τα μοντέλα του HADDOCK επίσης δεν αποτελούσε ικανοποιητικό μοντέλο πρόβλεψης νέων μοντέλων και σωστής αξιολόγησής τους, καθώς η ακρίβειά του δεν ξεπερνούσε το 10%. Παρόλα αυτά είναι δυνατόν να προκύψουν χρήσιμες πληροφορίες για τις μεταβλητές που χρησιμοποιήσαμε.



Διάγραμμα 5. Ακρίβεια γραμμικού μοντέλου αξιολόγησης με στόχο το DockQ για τα μοντέλα συμπλόκων από το HADDOCK. Στο διάγραμμα απεικονίζεται το ποσοστό των μοντέλων των οποίων οι μεταβλητές προσαρμόζονται στην γραμμική εξίσωση με μικρές αποκλίσεις από το σύνολο των μοντέλων.

Όπως είδαμε και προηγουμένως στο μοντέλο αξιολόγησης των μοντέλων του HADDOCK με στόχο το RMSD, έτσι και σε αυτό με στόχο το DockQ, η σχετική σημαντικότητα της dR_g είναι είναι αρκετά μεγαλύτερη από τις άλλες δύο μεταβλητές. Η παρατήρηση αυτή είναι εμφανής στο διάγραμμα 6. Σε αυτό το διάγραμμα φαίνεται πως και το dDML έχει μια σημαντικότητα, συγκριτικά με το μοντέλο με στόχο το RMSD, στο οποίο ήταν σχεδόν μηδαμινή.



Διάγραμμα 6. Σχετική σημαντικότητα των μεταβλητών στο μοντέλο αξιολόγησης με στόχο το DockQ για τα μοντέλα συμπλόκων από το HADDOCK. Στον άξονα y αντιστοιχούν οι μεταβλητές dR_g, ΔpKa και dDML, ενώ στο άξονα x η σημαντικότητά τους με μέγιστο 1. Συγκεκριμένα η σημαντικότητα των μεταβλητών είναι 0,656 για την dRg, 0,171 για το ΔpKa και 0,173 για το dDML.

Η γραμμική εξίσωση για τα μοντέλα των συμπλόκων που προέρχονται από το HADDOCK με στόχο το DockQ, έχει συντελεστές για την εξίσωση 11: a=-0,001, b=-0,009, c=0 και d=0,17.

Στο πίνακα 4 φαίνονται οι κατατάξεις για το σύμπλοκο 1FQ1 με βάση το γραμμικό μοντέλο με στόχο το DockQ όπως προέκυψε από τα μοντέλα του HADDOCK. Τα μοντέλα του συμπλόκου δεν ανήκουν στην κατηγορία "αποδεκτά" με βάση τις μεταβλητές DockQ και RMSD. Είναι εμφανές πως η κατάταξη της dR_g ακολουθεί αυτή του γραμμικού μοντέλου, ενώ δεν παρατηρείται κάποιο μοτίβο για τις υπόλοιπες μεταβλητές. Όσον αφορά την κατάταξη των μοντέλων με βάση το πρόγραμμα, τα μοντέλα ανήκουν σε χαμηλή κατάταξη, ενώ δεν φαίνεται να ακολουθείται κάποιο μοτίβο.

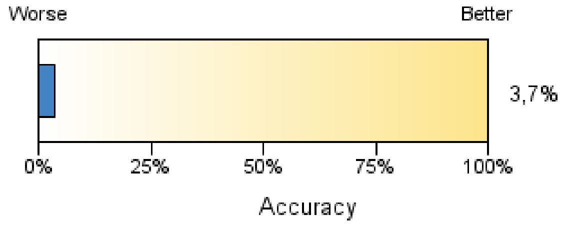
Πίνακας 4. Υπολογισμοί μεταβλητών του συμπλόκου 1FQ1 που προέκυψαν από μοντέλα του HADDOCK. Στον πίνακα φαίνονται οι μεταβλητές και η σχετική τους θέση στο σύνολο των μοντέλων του συμπλόκου από την καλύτερη στην χειρότερη.

name	DML dev	DML rank	Rg dev	Rg rank	pKa diff	pKa rank	DockQ	DockQ rank	RMSD	RMSD rank	LiReg. DockQ	Model DockQ rank
1FQ1_HD_195	3,025694	34	0,101885	2	2,48	27	0,011	192	20,22361	193	0,166057337	1
1FQ1_HD_99	0,253789	2	0,528117	6	6,63	72	0,038	57	8,195863	9	0,164993163	2
1FQ1_HD_8	4,648443	53	0,065715	1	3,24	34	0,073	7	8,079714	7	0,16476012	3
1FQ1_HD_106	2,583018	27	0,454766	5	14,3	172	0,09	2	6,566229	1	0,163324085	4
1FQ1_HD_125	4,931086	58	0,353669	4	6,59	70	0,009	197	21,87285	200	0,161885894	5
1FQ1_HD_194	1,71803	20	0,788139	16	3,59	38	0,008	199	21,53813	199	0,161188716	6
1FQ1_HD_59	2,892368	30	0,801106	17	0,74	7	0,008	198	21,21905	196	0,159897674	7
1FQ1_HD_153	4,055964	45	0,675704	11	9,99	125	0,045	39	7,964553	6	0,159862697	8
1FQ1_HD_180	3,349692	38	1,095991	20	2,17	23	0,026	105	9,595669	17	0,156786387	9
1FQ1_HD_2	0,707118	10	1,418241	24	10,5	132	0,039	56	8,089218	8	0,156528716	10

Τα ίδια προβλήματα που αναφέρθηκαν στην ενότητα 3.1.1.1. ισχύουν και στην παρούσα ενότητα καθώς οι αναλύσεις έγιναν στα ίδια μοντέλα του HADDOCK. Συγκεκριμένα και στα μοντέλα των υπόλοιπων συμπλόκων δεν εμφανίζεται κάποιο μοτίβο στις κατατάξεις των μεταβλητών σε σχέση με την γραμμική εξίσωση. Επίσης, όσον αφορά τις τιμές των DockQ και RMSD δεν αντιστοιχούν σε «αποδεκτά» μοντέλα κανένα από αυτά των τεσσάρων συμπλόκων. Μοναδική εξαίρεση αποτελεί το σύμπλοκο 1MAH, το οποίο ανήκει στην κατηγορία "χαμηλής δυσκολίας", όπου τα τρία εκ των διακοσίων μοντέλων χαρακτηρίζονται ως «αποδεκτά».

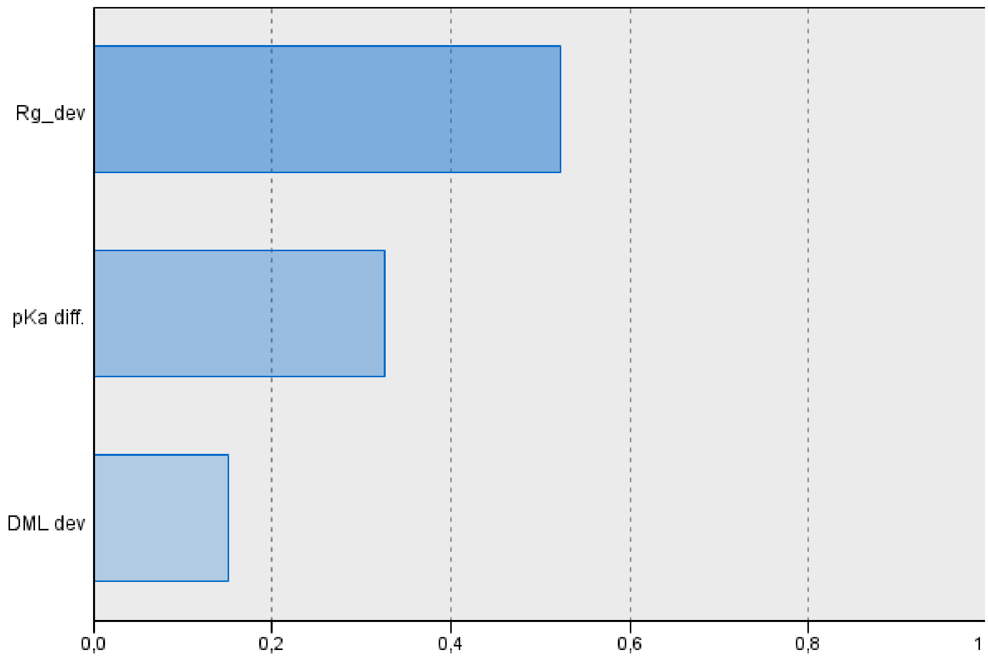
3.1.2.2. Σύμπλοκα InterEvDock2

Στο τρίτο διάγραμμα που προέκυψε από τις αναλύσεις των τιμών, η γραμμική εξίσωση για τα μοντέλα του InterEvDock με στόχο το DockQ δεν αποτέλεσε ικανοποιητικό μοντέλο της σωστής αξιολόγησής τους, καθώς η ακρίβειά του δεν ξεπερνούσε το 5%. Το γραμμικό μοντέλο αυτό ήταν το χειρότερο εκ των τριών προηγούμενων με βάση την ακρίβειά τους.



Διάγραμμα 7. Ακρίβεια μοντέλου αξιολόγησης με στόχο το DockQ για τα μοντέλα συμπλόκων από το InterEvDock2. Στο διάγραμμα απεικονίζεται το ποσοστό των μοντέλων των οποίων οι μεταβλητές προσαρμόζονται στην γραμμική εξίσωση με μικρές αποκλίσεις από το σύνολο των μοντέλων.

Η σχετική σημαντικότητα των μεταβλητών για το συγκεκριμένο γραμμικό μοντέλο φαίνεται στο διάγραμμα 8. Όπως φαίνεται παρακάτω, και σε αυτό το μοντέλο η μεγαλύτερη σχετική σημαντικότητα αναφέρεται στην δομική μεταβλητή dR_g. Αυτή ακολουθείται από την μεταβλητή για το pKa και έπειτα για το DML.



Διάγραμμα 8. Σχετική σημαντικότητα των μεταβλητών στο μοντέλο αξιολόγησης με στόχο το DockQ για τα μοντέλα συμπλόκων από το InterEvDock2. Στον άξονα y αντιστοιχούν οι μεταβλητές dR_g, ΔpKa και dDML, ενώ στο άξονα x η σημαντικότητά τους με μέγιστο 1. Συγκεκριμένα η σημαντικότητα των μεταβλητών είναι 0,523 για την dRg, 0,326 για το ΔpKa και 0,150 για το dDML.

Η γραμμική εξίσωση για τα μοντέλα των συμπλόκων που προέρχονται από το InterEvDock2 με στόχο το DockQ, έχει συντελεστές για την εξίσωση 11: a=0, b=-0,005, c=0 και d=0,155.

Τα μοντέλα του InterEvDock2 για το σύμπλοκο 1FQ1 ανήκουν όπως προαναφέρθηκε στην κατηγορία "μη αποδεκτά" με βάση τις μεταβλητές αξιολόγησης DockQ και RMSD. Στον πίνακα 5 απεικονίζονται οι υπολογισμοί και οι κατατάξεις των μεταβλητών των μοντέλων με βάση την κατάταξή τους από το γραμμικό μοντέλο. Φαίνεται πως και σε αυτό το μοντέλο, η Rg ακολουθεί την κατάταξη του μοντέλου, ενώ οι υπόλοιπες μεταβλητές δεν ακολουθούν κάποιο μοτίβο. Το ίδιο ισχύει και για την κατάταξη του InterEvDock2, δηλαδή δεν ακολουθείται κάποιο μοτίβο σε σχέση με την γραμμική εξίσωση.

Τα ίδια προβλήματα φαίνονται και στα υπόλοιπα τέσσερα σύμπλοκα εκ των πέντε που χρησιμοποιήθηκαν για τον έλεγχο των γραμμικών μοντέλων.

Πίνακας 5. Υπολογισμοί μεταβλητών του συμπλόκου 1FQ1 που προέκυψαν από μοντέλα του InterEvDock2. Στον πίνακα φαίνονται οι μεταβλητές και η σχετική τους θέση στο σύνολο των μοντέλων του συμπλόκου από την καλύτερη στην χειρότερη.

name	**DML dev**	**DML rank**	**Rg dev**	**Rg rank**	**pKa diff**	**pKa rank**	**DockQ**	**DockQ rank**	**RMSD**	**RMSD rank**	**LiReg. DockQ**	**Model DockQ rank**
1FQ1_FRODOCK_35	48,9701365	144	0,01978	1	9,49	95	0,013	108	18,6153	106	0,154901083	1
1FQ1_FRODOCK_45	5,00959855	39	0,12269	2	6,2	78	0,01	125	19,825	123	0,154386568	2
1FQ1_FRODOCK_20	44,6148175	141	0,16098	3	5,39	70	0,013	104	18,6636	107	0,154195122	3
1FQ1_SOAP_PP_38	5,03441892	41	0,23456	4	9,97	101	0,027	49	14,9755	59	0,153827185	4
1FQ1_FRODOCK_12	5,15865258	42	0,30592	5	15,13	130	0,026	52	11,3648	13	0,15347038	5
1FQ1_IES_1	25,4575751	123	0,40919	6	0,98	18	0,015	91	19,0039	115	0,152954068	6
1FQ1_FRODOCK_31	4,5994799	35	0,47051	7	6,71	80	0,023	59	12,9186	35	0,152647448	7
1FQ1_SOAP_PP_17	41,7106239	138	0,47381	8	2,29	32	0,012	113	18,7345	108	0,152630961	8
1FQ1_FRODOCK_23	47,9956391	143	0,47605	9	0,02	1	0,011	118	19,9347	124	0,152619754	9
1FQ1_IES_41	1,10728151	7	0,55535	10	2,16	29	0,025	55	15,4173	62	0,152223244	10

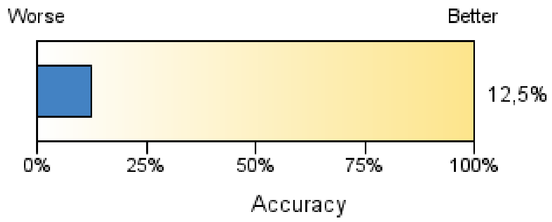
3.2. Γραμμική εξίσωση με κανονικοποιημένες τιμές των μεταβλητών

Καθώς οι γραμμικές παλινδρομήσεις που αναλύθηκαν παραπάνω δεν φαίνεται να είχαν αρκετά υψηλή ακρίβεια, έγινε χρήση της αυτόματης κανονικοποίησης των τιμών από το SPSS. Με τον τρόπο αυτό είναι δυνατόν να πάρουμε διαφορετική συμπεριφορά των μεταβλητών ως προς τον τρόπο που επηρεάζουν την αξιολόγηση των μοντέλων του docking. Για την κανονικοποίηση έγινε χρήση της επιλογής του SPSS στην αυτόματη γραμμική παλινδρόμηση της επεξεργασίας δεδομένων.

3.2.1. Αξιολόγηση με RMSD

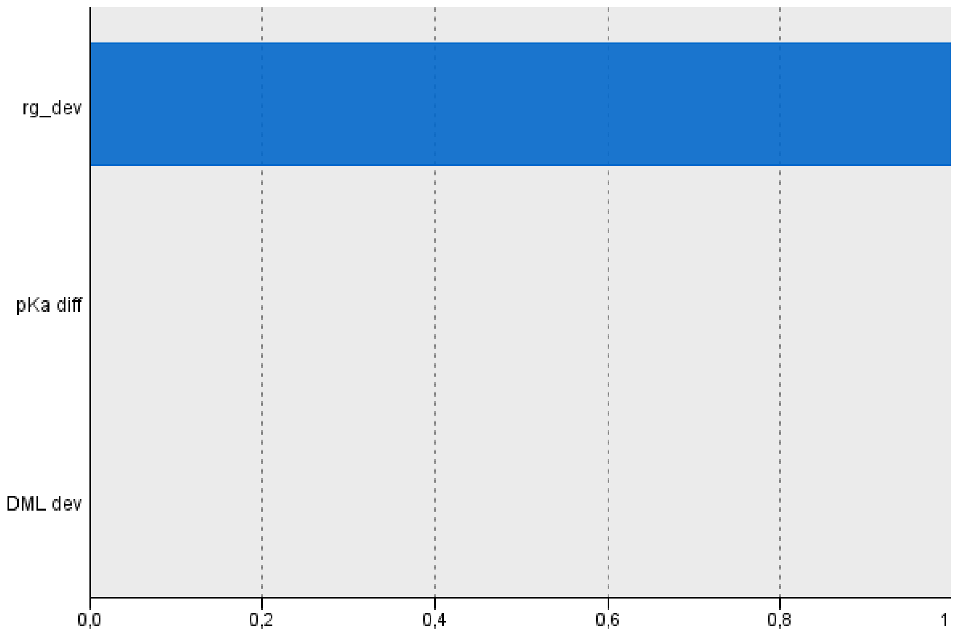
3.2.1.1. Σύμπλοκα HADDOCK

Η ακρίβεια του μοντέλου για τα αποτελέσματα των μοντέλων του HADDOCK με στόχο το RMSD φαίνεται να έχει μειωθεί έπειτα από την κανονικοποίηση των δεδομένων κατά περίπου κατά 1%.



Διάγραμμα 9. Ακρίβεια γραμμικού μοντέλου αξιολόγησης με στόχο το RMSD για τα μοντέλα συμπλόκων από το HADDOCK έπειτα από κανονικοποίηση των δεδομένων. Στο διάγραμμα απεικονίζεται το ποσοστό των μοντέλων των οποίων οι μεταβλητές προσαρμόζονται στην γραμμική εξίσωση με μικρές αποκλίσεις από το σύνολο των μοντέλων.

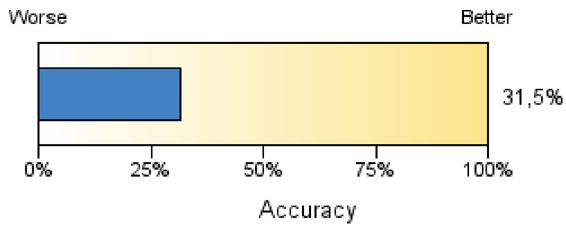
Όσον αφορά την σχετική σημαντικότητα των κανονικοποιημένων μεταβλητών για το συγκεκριμένο μοντέλο απεικονίζονται στο διάγραμμα 10. Είναι εμφανές πως δεν υπάρχει καμία σχετική σημαντικότητα στις μεταβλητές dDML και ΔpKa, ενώ για το dRg η τιμή αυτή είναι 1.



Διάγραμμα 10. Σχετική σημαντικότητα των κανονικοποιημένων μεταβλητών στο μοντέλο αξιολόγησης με στόχο το RMSD για τα μοντέλα συμπλόκων από το HADDOCK. Στον άξονα y αντιστοιχούν οι μεταβλητές dRg, ΔpKa και dDML, ενώ στο άξονα x η σημαντικότητά τους με μέγιστο 1. Συγκεκριμένα η σημαντικότητα των μεταβλητών είναι 1 για το dRg, 0 για το ΔpKa και 0 για το dDML.

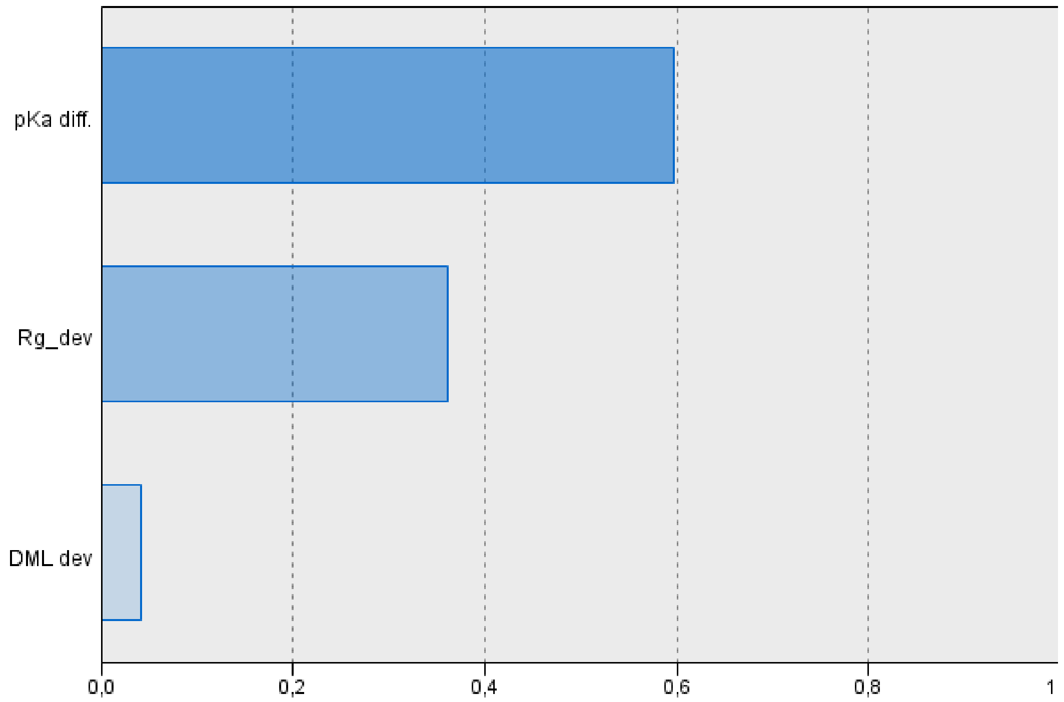
3.2.1.2. Σύμπλοκα InterEvDock2

Αντιθέτως με την ακρίβεια του μοντέλου για το HADDOCK, στο παρόν μοντέλο, που πραγματοποιήθηκε με την χρήση των κανονικοποιημένων τιμών από τα μοντέλα του InterEvDock2, φαίνεται να αυξήθηκε κατά πολύ η ακρίβειά του. Συγκεκριμένα παρατηρήθηκε αύξηση 12% αύξηση.



Διάγραμμα 11. Ακρίβεια μοντέλου αξιολόγησης με στόχο το RMSD για τα μοντέλα συμπλόκων από το InterEvDock2 έπειτα από κανονικοποίηση των δεδομένων. Στο διάγραμμα απεικονίζεται το ποσοστό των μοντέλων των οποίων οι μεταβλητές προσαρμόζονται στην γραμμική εξίσωση με μικρές αποκλίσεις από το σύνολο των μοντέλων.

Η σχετική σημαντικότητα των μεταβλητών του μοντέλου αυτού σε σχέση με αυτή των μη κανονικοποιημένων μεταβλητών δεν φαίνεται να άλλαξε δραματικά. Τα δεδομένα αυτά απεικονίζονται στο παρακάτω διάγραμμα.

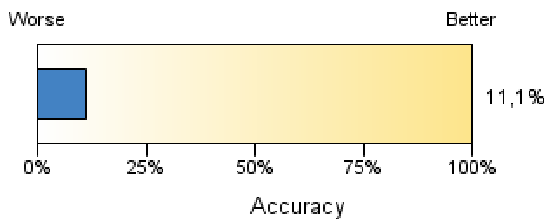


Διάγραμμα 12. Σχετική σημαντικότητα των κανονικοποιημένων μεταβλητών στο μοντέλο αξιολόγησης με στόχο το RMSD για τα μοντέλα συμπλόκων από το InterEvDock2. Στον άξονα y αντιστοιχούν οι μεταβλητές dRg, ΔpKa και dDML, ενώ στο άξονα x η σημαντικότητά τους με μέγιστο 1. Συγκεκριμένα η σημαντικότητα των μεταβλητών είναι 0,362 για το dRg, 0,597 για το ΔpKa και μόλις 0,041 για το dDML.

3.2.2. Αξιολόγηση με DockQ

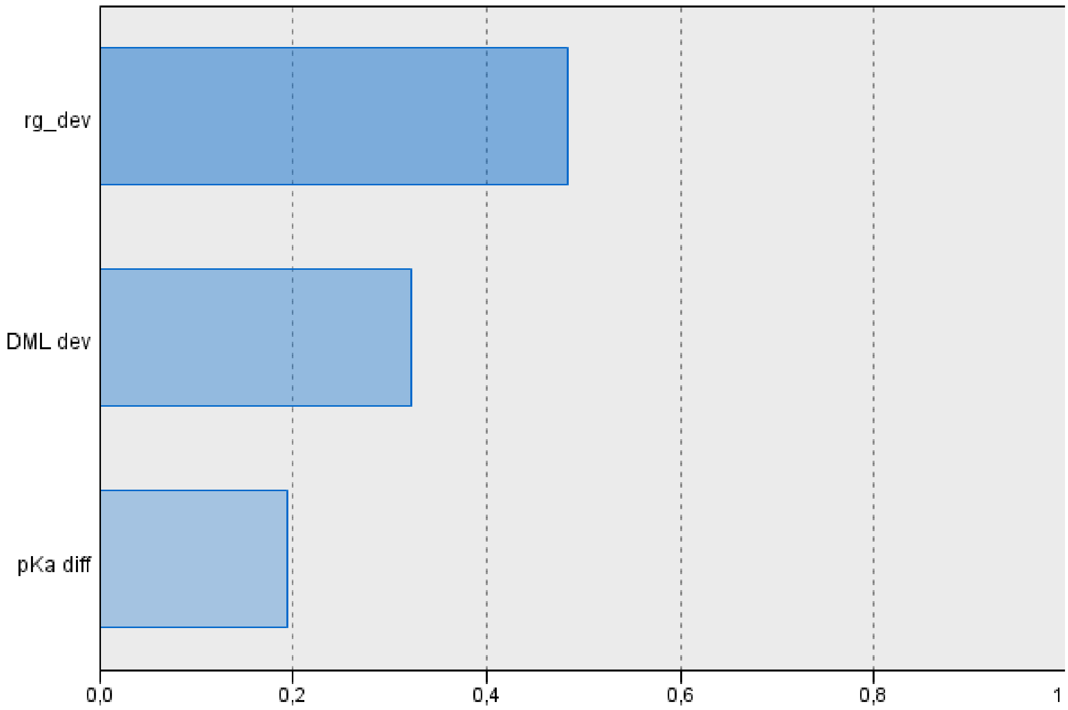
3.2.2.1. Σύμπλοκα HADDOCK

Στο παρόν γραμμικό μοντέλο φαίνεται πως έπειτα από την κανονικοποίηση των μεταβλητών υπήρξε μια αύξηση της ακρίβειας αν και αυτή ήταν μικρή. Η αύξηση αυτή είναι μόλις 1,5%.



Διάγραμμα 13. Ακρίβεια μοντέλου αξιολόγησης με στόχο το DockQ για τα μοντέλα συμπλόκων από το HADDOCK έπειτα από κανονικοποίηση των δεδομένων. Στο διάγραμμα απεικονίζεται το ποσοστό των μοντέλων των οποίων οι μεταβλητές προσαρμόζονται στην γραμμική εξίσωση με μικρές αποκλίσεις από το σύνολο των μοντέλων.

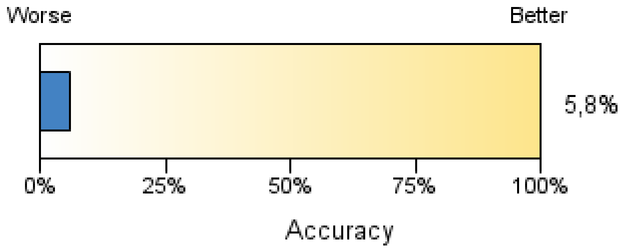
Η σημαντικότητα των μεταβλητών απεικονίζεται στο παρακάτω διάγραμμα. Η σημαντικότητα ακολουθεί το μοτίβο των μη κανονικοποιημένων μεταβλητών, δηλαδή η μεγαλύτερη αυτών είναι της dRg και ακολουθείται από το dDML και ΔpKa.



Διάγραμμα 14. Σχετική σημαντικότητα των κανονικοποιημένων μεταβλητών στο μοντέλο αξιολόγησης με στόχο το DockQ για τα μοντέλα συμπλόκων από το HADDOCK. Στον άξονα y αντιστοιχούν οι μεταβλητές dRg, ΔpKa και dDML, ενώ στο άξονα x η σημαντικότητά τους με μέγιστο 1. Συγκεκριμένα η σημαντικότητα των μεταβλητών είναι 0,484 για την dRg, 0,193 για το ΔpKa και 0,323 για το dDML.

3.2.2.2. Σύμπλοκα InterEvDock2

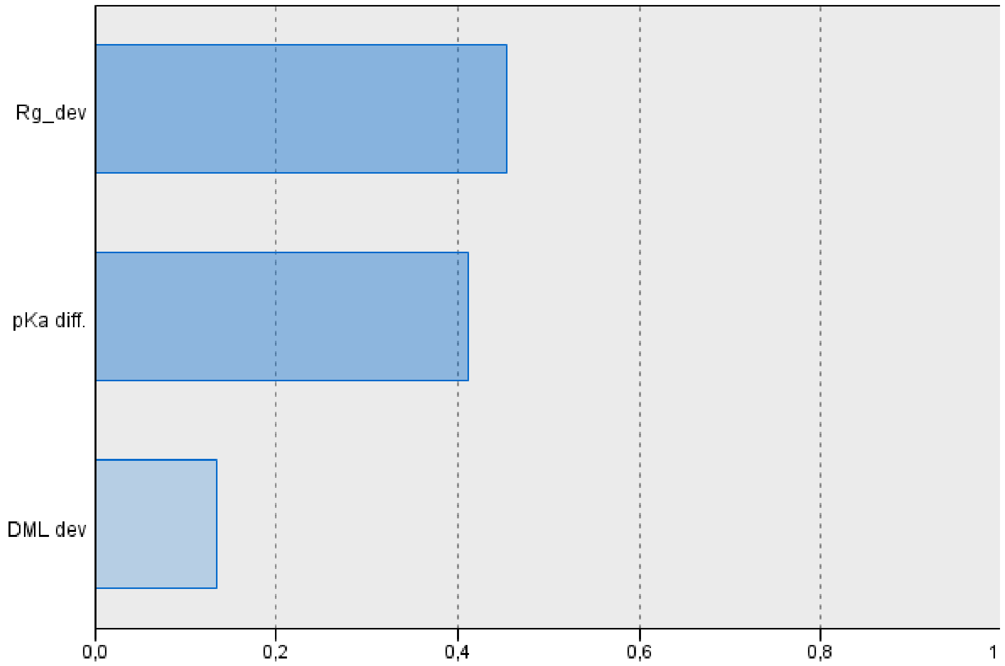
Όπως παρατηρήθηκε και στο μοντέλο με τα μη κανονικοποιημένα δεδομένα, η ακρίβειά του αντιστοιχεί στην χαμηλότερη εκ των γραμμικών μοντέλων. Συγκεκριμένα ανέρχεται στο 5,8%, μόλις 1,1% πάνω από το αρχικό.



Διάγραμμα 15. Ακρίβεια μοντέλου αξιολόγησης με στόχο το DockQ για τα μοντέλα συμπλόκων από το InterEvDock2 έπειτα από κανονικοποίηση των δεδομένων. Στο διάγραμμα απεικονίζεται το

ποσοστό των μοντέλων των οποίων οι μεταβλητές προσαρμόζονται στην γραμμική εξίσωση με μικρές αποκλίσεις από το σύνολο των μοντέλων.

Η σχετική σημαντικότητα των μοντέλων είναι όμοια με αυτήν στο γραμμικό μοντέλο μη κανονικοποιημένων μεταβλητών και απεικονίζεται στο διάγραμμα 16.



Διάγραμμα 16. Σχετική σημαντικότητα των κανονικοποιημένων μεταβλητών στο μοντέλο αξιολόγησης με στόχο το DockQ για τα μοντέλα συμπλόκων από το InterEvDock2. Στον άξονα y αντιστοιχούν οι μεταβλητές dRg, ΔpKa και dDML, ενώ στο άξονα x η σημαντικότητά τους με μέγιστο 1. Συγκεκριμένα η σημαντικότητα των μεταβλητών είναι 0,454 για την dRg, 0,412 για το ΔpKa και 0,134 για το dDML.

4. Συμπεράσματα-Συζήτηση

Η αξιολόγηση των δομικών μοντέλων αλληλεπίδρασης των πρωτεϊνών αποτελεί το σημαντικότερο τμήμα της πρόσδεσης δύο ή περισσοτέρων πρωτεϊνών μέσω υπολογιστικών μεθόδων. Η δυνατότητα της αξιολόγησης αυτής με υπολογιστικές αλλά και πειραματικές μεθόδους αποτελεί κομβικό σημείο για την υποβοηθούμενη με υπολογιστικές μεθόδους έρευνα των πρωτεϊνικών αλληλεπιδράσεων. Στην παρούσα μελέτη έγινε χρήση τριών μεταβλητών για την αξιολόγηση τέτοιων μοντέλων, του μέτρου της διπολικής ροπής (DML), της γυροσκοπικής ακτίνας (R_g) και του συνολικού pK_a του συμπλόκου.

Για την διερεύνηση αυτή έγινε χρήση μοντέλου γραμμικής παλινδρόμησης για τα μοντέλα των δύο προγραμμάτων docking, HADDOCK και InterEvDock2, με χρήση των μεταβλητών αυτών και στόχους είτε το RMSD, είτε το DockQ. Όπως φαίνεται από τα διαγράμματα ακρίβειας των γραμμικών μοντέλων, δεν αποτελούν αποδεκτά μοντέλα, καθώς το ποσοστό ακρίβειας δεν ξεπερνούσε το 31,5%. Αυτό το ποσοστό ακρίβειας ανήκει στο μοντέλο κανονικοποιημένων δεδομένων των μοντέλων του InterEvDock2 με στόχο το RMSD.

Η σχετική σημαντικότητα των τριών μεταβλητών στα γραμμικά μοντέλα υποδηλώνει πως τον μεγαλύτερο ρόλο στην αξιολόγηση έχει η γυροσκοπική ακτίνα. Αυτό το αποτέλεσμα διαφοροποιείται μόνο στα διαγράμματα 12 και 4, που αφορούν στα γραμμικά μοντέλα για το InterEvDock2 με στόχο το RMSD με κανονικοποιημένα και μη δεδομένα αντίστοιχα. Αυτά τα μοντέλα αποτελούν τα δύο εκ των συνολικά οκτώ γραμμικών μοντέλων με την υψηλότερη ακρίβεια. Την γυροσκοπική ακτίνα ακολουθεί κατά πλειοψηφία το συνολικό pK_a και τέλος το μέτρο της διπολικής ροπής.

Είναι δυνατόν από αυτά τα αποτελέσματα να συμπεράνουμε τα εξής:

- Η γραμμική παλινδρόμηση δεν αποδίδει αρκετά ακριβή μοντέλα πρόβλεψης του RMSD και του DockQ με ελεγχόμενες μεταβλητές τα dRg, dDML και ΔpKa. Αυτό πιθανόν να οφείλεται στα μοντέλα των προγραμμάτων docking, τα οποία παρόλο που αξιολογούνται από τα εν λόγω προγράμματα, δεν αφορούν στην πλειοψηφία τους «αποδεκτά» μοντέλα με βάση τις μεταβλητές αξιολόγησης RMSD και DockQ. Αυτό είναι ιδιαίτερα εμφανές στα σύμπλοκα που με βάση την λίστα αναφοράς κατηγοριοποιούνται σε μεσαίας και υψηλής δυσκολίας. Τα σύμπλοκα αυτά αφορούν συνήθως πρωτεΐνες που προσδένονται με εύκαμπτο τρόπο, αλλάζοντας δηλαδή την δομή τους κατά την πρόσδεση. Τόσο το HADDOCK όσο και το InterEvDock2 δεν κάνουν εύκαμπτο docking.
- Η χαμηλή ακρίβεια μπορεί επίσης να οφείλεται στο γεγονός ότι εκτός από το σωστό σύμπλοκο υπάρχει και ένας αριθμός μοντέλων με το ίδιο μέτρο διπολικής ροπής αλλά διαφορετική δομή, καθώς η περιστροφή της μίας εκ των δύο πρωτεϊνών ή και των δύο γύρω από τον εκάστοτε άξονα της διπολικής τους ροπής δεν αλλάζει την διπολική ροπή. Κάτι ανάλογο ισχύει και για την R_g.
- Η υψηλότερη σχετική σημαντικότητα του dR_g έναντι του dDML ήταν αρκετά έντονη στα παραπάνω διαγράμματα. Όπως είναι φανερό, το dDML επηρεάζεται από την κατανομή των ηλεκτρικών φορτίων, των οποίων η απόδοση εξαρτάται από το πρόγραμμα που χρησιμοποιήθηκε και από το force field. Αυτή η αβεβαιότητα δεν υπάρχει στην R_g καθώς οι μάζες των ατόμων έχουν συγκεκριμένες τιμές.
- Η μη ικανοποιητική επεξηγηματική ικανότητα του ΔpK_a οφείλεται πιθανότατα στο γεγονός ότι τόσο το HADDOCK όσο και το InterEvDock δεν προβλέπουν χαλάρωση της δομής κατά τον σχηματισμό των μοντέλων. Αυτό συχνά έχει ως αποτέλεσμα αφύσικες σχετικές θέσεις των ατόμων στην διεπιφάνεια αλληλεπίδρασης και κατά

συνέπεια λανθασμένη πρόβλεψη του pK_a. Ένας άλλος λόγος είναι ότι δεν λαμβάνεται κατά την πρόσδεση πιθανή αλλαγή στην κατάσταση πρωτονίωσης.

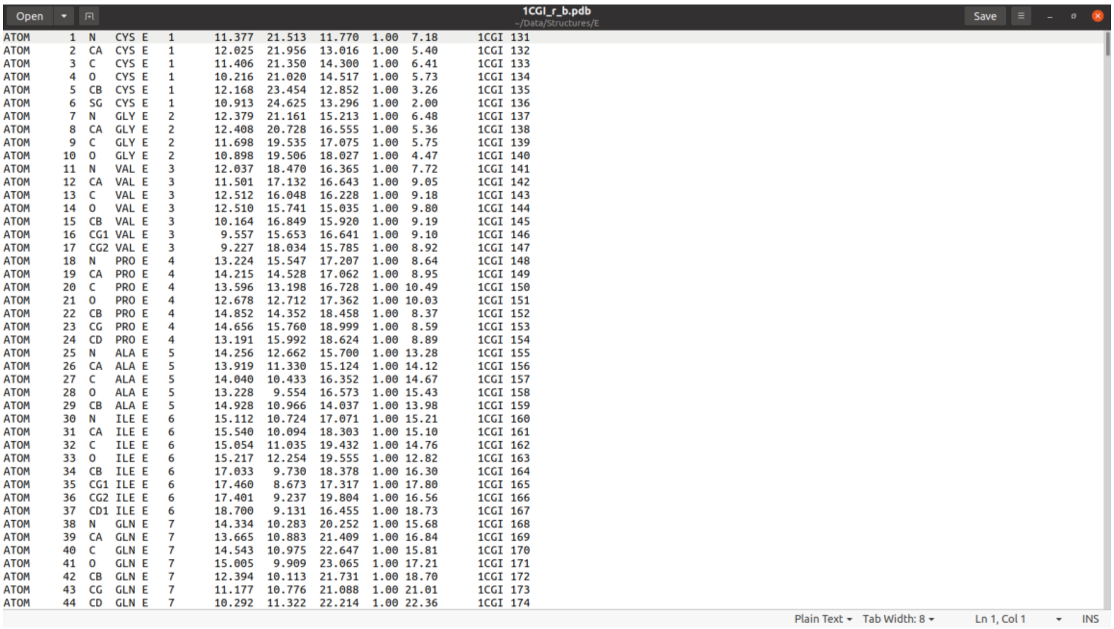
- Η ακρίβεια των γραμμικών μοντέλων με στόχο το RMSD είναι μεγαλύτερη από αυτή των μοντέλων με στόχο το DockQ, πιθανόν λόγω του ότι το DockQ δίνεται με τρία μόνο σημαντικά ψηφία.

Τα μοντέλα γραμμικής παλινδρόμησης μπορεί να βελτιωθούν εάν γίνει η ανάλυση για όλα τα σύμπλοκα της λίστας αναφοράς. Επίσης, είναι δυνατόν να χρησιμοποιηθούν και άλλες επεξηγηματικές μεταβλητές σε μελλοντικές μελέτες, για καλύτερα και ποιοτικότερα αποτελέσματα. Τέλος, ίσως να προκύψουν καλύτερα αποτελέσματα κάνοντας χρήση των μεθόδων τεχνητής νοημοσύνης (AI).

Παράρτημα

A1. Αρχείο PDB

Τα αρχεία PDB περιλαμβάνουν πληροφορίες για τις συντεταγμένες των ατόμων ενός μορίου στον χώρο και αξιοποιούνται από κατάλληλα προγράμματα οπτικοποίησης των μορίων. Το αρχείο όπως φαίνεται και στην εικόνα 8, αποτελείται από ορισμένες στήλες, κάθε μια από της οποίες δίνει διαφορετικές πληροφορίες για τα άτομα. Στην πρώτη στήλη είναι δυνατόν να βρεθούν τρεις διαφορετικές λέξεις κλειδιά, REMARK, ATOM, HETATM. Το REMARK σηματοδοτεί εγγραφές που αποτελούν σχόλια. Το ATOM αντιστοιχεί σε γραμμή συντεταγμένων και άλλων πληροφοριών που αφορούν σε άτομα της πρωτεΐνης, ενώ το HETATM αντιστοιχεί σε άτομα που δεν ανήκουν στην πρωτεΐνη, όπως νερό, ATP κ.α. Η αμέσως επόμενη στήλη αντιστοιχεί στον αριθμό του κάθε ατόμου. Η τρίτη στήλη είναι το όνομα του κάθε ατόμου, όπως N, CA, C. Η τέταρτη στήλη δίνει πληροφορία για το αμινοξύ στο οποίο ανήκει κάθε άτομο και η πέμπτη στήλη δίνει με ένα γράμμα την αλυσίδα στην οποία ανήκουν. Στην παρούσα εικόνα η αλυσίδα ονομάζεται E. Η έκτη στήλη δίνει τον αριθμό του αμινοξέως στο αρχείο της πρωτεΐνης. Οι επόμενες 5 στήλες, δηλαδή από την έβδομη έως την ενδέκατη, αφορούν με την σειρά τις τρεις συντεταγμένες των ατόμων στις τρεις διαστάσεις, η επόμενη στήλη δίνει μια εκτίμηση του ποσοστού κατάληψης της συγκεκριμένης θέσης από το άτομο και τέλος τον παράγοντα θερμοκρασίας. Στην δωδέκατη στήλη αναγράφετε η πληροφορία για το τμήμα στο οποίο ανήκει το εν λόγω άτομο στην πεπτιδική αλυσίδα. Τέλος η δέκατη τρίτη δείχνει το χημικό στοιχείο του ατόμου.

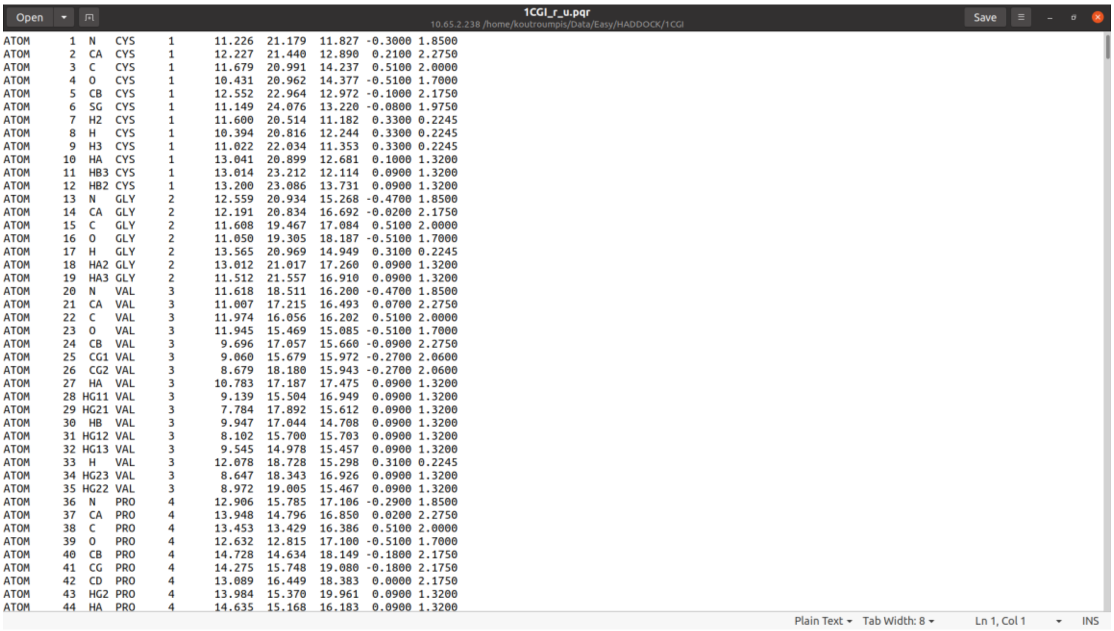


```
ATOM      1  N   CYS E   1      11.377  21.513  11.770  1.00  7.18      1CGI 131
ATOM      2  CA  CYS E   1      12.025  21.956  13.016  1.00  5.40      1CGI 132
ATOM      3  C   CYS E   1      11.406  21.350  14.300  1.00  6.41      1CGI 133
ATOM      4  O   CYS E   1      10.216  21.020  14.517  1.00  5.73      1CGI 134
ATOM      5  CB  CYS E   1      12.168  23.454  12.852  1.00  3.26      1CGI 135
ATOM      6  SG  CYS E   1      10.913  24.625  13.296  1.00  2.00      1CGI 136
ATOM      7  N   GLY E   2      12.379  21.161  15.213  1.00  6.48      1CGI 137
ATOM      8  CA  GLY E   2      12.408  20.728  16.555  1.00  5.36      1CGI 138
ATOM      9  C   GLY E   2      11.698  19.535  17.075  1.00  5.75      1CGI 139
ATOM     10  O   GLY E   2      10.898  19.506  18.027  1.00  4.47      1CGI 140
ATOM     11  N   VAL E   3      12.037  18.470  16.365  1.00  7.72      1CGI 141
ATOM     12  CA  VAL E   3      11.501  17.132  16.643  1.00  9.05      1CGI 142
ATOM     13  C   VAL E   3      12.512  16.048  16.228  1.00  9.18      1CGI 143
ATOM     14  O   VAL E   3      12.510  15.741  15.035  1.00  9.80      1CGI 144
ATOM     15  CB  VAL E   3      10.164  16.849  15.920  1.00  9.19      1CGI 145
ATOM     16  CG1 VAL E   3       9.557  15.653  16.641  1.00  9.10      1CGI 146
ATOM     17  CG2 VAL E   3       9.227  18.034  15.785  1.00  8.92      1CGI 147
ATOM     18  N   PRO E   4      13.224  15.547  17.207  1.00  8.64      1CGI 148
ATOM     19  CA  PRO E   4      14.215  14.528  17.062  1.00  8.95      1CGI 149
ATOM     20  C   PRO E   4      13.596  13.198  16.728  1.00 10.49      1CGI 150
ATOM     21  O   PRO E   4      12.678  12.712  17.362  1.00 10.03      1CGI 151
ATOM     22  CB  PRO E   4      14.852  14.352  18.458  1.00  8.37      1CGI 152
ATOM     23  CG  PRO E   4      14.656  15.760  18.999  1.00  8.59      1CGI 153
ATOM     24  CD  PRO E   4      13.191  15.992  18.624  1.00  8.89      1CGI 154
ATOM     25  N   ALA E   5      14.256  12.662  15.700  1.00 13.28      1CGI 155
ATOM     26  CA  ALA E   5      13.919  11.330  15.124  1.00 14.12      1CGI 156
ATOM     27  C   ALA E   5      14.040  10.433  16.352  1.00 14.67      1CGI 157
ATOM     28  O   ALA E   5      13.228   9.554  16.573  1.00 15.43      1CGI 158
ATOM     29  CB  ALA E   5      14.928  10.966  14.037  1.00 13.98      1CGI 159
ATOM     30  N   ILE E   6      15.112  10.724  17.071  1.00 15.21      1CGI 160
ATOM     31  CA  ILE E   6      15.540  10.094  18.303  1.00 15.10      1CGI 161
ATOM     32  C   ILE E   6      15.054  11.035  19.432  1.00 14.76      1CGI 162
ATOM     33  O   ILE E   6      15.217  12.254  19.555  1.00 12.82      1CGI 163
ATOM     34  CB  ILE E   6      17.033   9.730  18.378  1.00 16.30      1CGI 164
ATOM     35  CG1 ILE E   6      17.460   8.673  17.317  1.00 17.80      1CGI 165
ATOM     36  CG2 ILE E   6      17.401   9.237  19.804  1.00 16.56      1CGI 166
ATOM     37  CD1 ILE E   6      18.700   9.131  16.455  1.00 18.73      1CGI 167
ATOM     38  N   GLN E   7      14.334  10.283  20.252  1.00 15.68      1CGI 168
ATOM     39  CA  GLN E   7      13.665  10.883  21.409  1.00 16.84      1CGI 169
ATOM     40  C   GLN E   7      14.543  10.975  22.647  1.00 15.81      1CGI 170
ATOM     41  O   GLN E   7      15.005   9.909  23.065  1.00 17.21      1CGI 171
ATOM     42  CB  GLN E   7      12.394  10.113  21.731  1.00 18.70      1CGI 172
ATOM     43  CG  GLN E   7      11.177  10.776  21.088  1.00 21.01      1CGI 173
ATOM     44  CD  GLN E   7      10.292  11.322  22.214  1.00 22.36      1CGI 174
```

Εικόνα 8. Αρχείο PDB του υποδοχέα του συμπλόκου 1CGI του bound αρχείου.

A2. Αρχείο PQR

Η διαφορά μεταξύ των αρχείων PDB και PQR αφορά τις στήλες 10 και 11. Συγκεκριμένα στην δέκατη στήλη αναγράφεται το μερικό φορτίο των ατόμων, ενώ στην ενδέκατη στήλη αναγράφετε η ακτίνα του ατόμου.



```
ATOM      1  N   CYS     1      11.226  21.179  11.827 -0.3000 1.8500
ATOM      2  CA  CYS     1      12.227  21.440  12.890  0.2100 2.2750
ATOM      3  C   CYS     1      11.679  20.991  14.237  0.5100 2.0000
ATOM      4  O   CYS     1      10.431  20.962  14.377 -0.5100 1.7000
ATOM      5  CB  CYS     1      12.552  22.964  12.972 -0.1000 2.1750
ATOM      6  SG  CYS     1      11.149  24.076  13.220 -0.0800 1.9750
ATOM      7  H2  CYS     1      11.600  20.514  11.182  0.3300 0.2245
ATOM      8  H   CYS     1      10.394  20.816  12.244  0.3300 0.2245
ATOM      9  H3  CYS     1      11.022  22.034  11.353  0.3300 0.2245
ATOM     10  HA  CYS     1      13.041  20.899  12.681  0.1000 1.3200
ATOM     11  HB3 CYS     1      13.014  23.212  12.114  0.0900 1.3200
ATOM     12  HB2 CYS     1      13.200  23.086  13.731  0.0900 1.3200
ATOM     13  N   GLY     2      12.559  20.934  15.268 -0.4700 1.8500
ATOM     14  CA  GLY     2      12.191  20.834  16.692 -0.0200 2.1750
ATOM     15  C   GLY     2      11.608  19.467  17.084  0.5100 2.0000
ATOM     16  O   GLY     2      11.050  19.305  18.187 -0.5100 1.7000
ATOM     17  H   GLY     2      13.565  20.969  14.949  0.3100 0.2245
ATOM     18  HA2 GLY     2      13.012  21.017  17.260  0.0900 1.3200
ATOM     19  HA3 GLY     2      11.512  21.557  16.910  0.0900 1.3200
ATOM     20  N   VAL     3      11.618  18.511  16.200 -0.4700 1.8500
ATOM     21  CA  VAL     3      11.007  17.215  16.493  0.0700 2.2750
ATOM     22  C   VAL     3      11.974  16.056  16.202  0.5100 2.0000
ATOM     23  O   VAL     3      11.945  15.469  15.085 -0.5100 1.7000
ATOM     24  CB  VAL     3       9.696  17.057  15.660 -0.0900 2.2750
ATOM     25  CG1 VAL     3       9.060  15.679  15.972 -0.2700 2.0600
ATOM     26  CG2 VAL     3       8.679  18.180  15.943 -0.2700 2.0600
ATOM     27  HA  VAL     3      10.783  17.187  17.475  0.0900 1.3200
ATOM     28 HG11 VAL     3       9.139  15.504  16.949  0.0900 1.3200
ATOM     29 HG21 VAL     3       7.784  17.892  15.612  0.0900 1.3200
ATOM     30  HB  VAL     3       9.947  17.044  14.708  0.0900 1.3200
ATOM     31 HG12 VAL     3       8.102  15.700  15.703  0.0900 1.3200
ATOM     32 HG13 VAL     3       9.545  14.978  15.457  0.0900 1.3200
ATOM     33  H   VAL     3      12.078  18.728  15.298  0.3100 0.2245
ATOM     34 HG23 VAL     3       8.647  18.343  16.926  0.0900 1.3200
ATOM     35 HG22 VAL     3       8.972  19.005  15.467  0.0900 1.3200
ATOM     36  N   PRO     4      12.906  15.785  17.106 -0.2900 1.8500
ATOM     37  CA  PRO     4      13.948  14.796  16.850  0.0200 2.2750
ATOM     38  C   PRO     4      13.453  13.429  16.386  0.5100 2.0000
ATOM     39  O   PRO     4      12.632  12.815  17.100 -0.5100 1.7000
ATOM     40  CB  PRO     4      14.728  14.634  18.149 -0.1800 2.1750
ATOM     41  CG  PRO     4      14.275  15.748  19.080 -0.1800 2.1750
ATOM     42  CD  PRO     4      13.089  16.449  18.383  0.0000 2.1750
ATOM     43  HG2 PRO     4      13.984  15.370  19.961  0.0900 1.3200
ATOM     44  HA  PRO     4      14.635  15.168  16.183  0.0900 1.3200
```

Εικόνα 9. Αρχείο PQR του υποδοχέα του συμπλόκου 1CGI του unbound αρχείου.

A3. Αρχείο PROPKA

Το αρχείο PROPKA αποτελείται από ένα σύνολο τεσσάρων πινάκων. Ο πρώτος πίνακας (Εικόνα 10) αφορά στις βασικές πληροφορίες των αμινοξέων ξεχωριστά. Στην πρώτη στήλη αναγράφεται το όνομα του αμινοξέως, έπειτα ο αριθμός του, και έπειτα η αλυσίδα στην οποία ανήκει. Η τέταρτη στήλη δείχνει το pK_a του αμινοξέως, ενώ η πέμπτη το το ποσοστό κάλυψης του αμινοξέως στο εσωτερικό της πρωτεΐνη. Στον πίνακα αναγράφονται επίσης με την σειρά, μια εκτίμηση του βαθμού διαλυτοποίησης, καθώς και οι αλληλεπιδράσεις με άλλα αμινοξέα. Ο δεύτερος πίνακας (Εικόνα 11) περιέχει πληροφορίες για το pKa των αμινοξέων στην πρωτεΐνη και το ελεύθερο pKa, αν το αμινοξύ δεν επηρεαζόταν από το περιβάλλον. Στον τρίτο πίνακα (Εικόνα 12) αναγράφεται εκτίμηση της ελεύθερης ενέργειας αναδίπλωσης της πρωτεΐνης ως συνάρτηση του pH και στον τέταρτο πίνακα (Εικόνα 12) εκτίμηση του συνολικού φορτίου της στα διάφορα pH.

Open | 1CGI_r_u.propka | 10.65.2.238 /home/koutroumpis/Data/Easy/HADDOCK/1CGI | Save

```
---------  -----   ------   ---------------------    --------------    --------------    --------------
                            DESOLVATION  EFFECTS       SIDECHAIN          BACKBONE         COULOMBIC
 RESIDUE    pKa    BURIED     REGULAR      RE        HYDROGEN BOND     HYDROGEN BOND      INTERACTION
---------  -----   ------   ---------   ---------    --------------    --------------    --------------

ASP  35 B   3.03     0 %    0.67  229   0.00    0   -0.20 THR  37 B   -0.07 ASP  35 B   -0.03 LYS  36 B
ASP  35 B                                            0.00 XXX   0 X   -0.14 THR  37 B    0.00 XXX   0 X
ASP  35 B                                            0.00 XXX   0 X   -0.28 GLY  38 B    0.00 XXX   0 X
ASP  35 B                                            0.00 XXX   0 X   -0.71 PHE  39 B    0.00 XXX   0 X

ASP  64 B   4.59    25 %    1.14  350   0.30    0    0.00 XXX   0 X   -0.75 THR  61 B   -0.02 LYS  36 B
ASP  64 B                                            0.00 XXX   0 X    0.00 XXX   0 X    0.12 ASP  35 B

ASP  72 B   3.32     0 %    0.54  273   0.00    0   -0.12 ARG 154 B   -0.49 GLY  74 B   -0.22 ARG 154 B
ASP  72 B                                            0.00 XXX   0 X   -0.19 ASP 153 B    0.00 XXX   0 X

ASP 102 B   3.69    77 %    3.19  498   0.43    0   -0.84 SER 214 B   -0.30 ALA  56 B   -0.95 HIS  57 B
ASP 102 B                                           -0.85 HIS  57 B   -0.79 HIS  57 B    0.00 XXX   0 X

ASP 128 B   3.10     0 %    0.50  259   0.00    0   -0.15 SER 127 B   -0.83 SER 125 B   -0.23 LYS 203 B

ASP 129 B   3.91     0 %    0.25  240   0.00    0    0.00 XXX   0 X   -0.00 PHE 130 B   -0.12 LYS 203 B
ASP 129 B                                            0.00 XXX   0 X    0.00 XXX   0 X   -0.04 ARG 230 B
ASP 129 B                                            0.00 XXX   0 X    0.00 XXX   0 X    0.02 ASP 128 B

ASP 153 B   6.20    50 %    1.75  420   0.59    0    0.00 XXX   0 X    0.00 XXX   0 X   -0.30 HIS  40 B
ASP 153 B                                            0.00 XXX   0 X    0.00 XXX   0 X    0.10 ASP  72 B
ASP 153 B                                            0.00 XXX   0 X    0.00 XXX   0 X    0.27 ASP 194 B

ASP 178 B   3.73     0 %    0.16  237   0.00    0    0.00 XXX   0 X    0.00 XXX   0 X   -0.13 LYS 169 B
ASP 178 B                                            0.00 XXX   0 X    0.00 XXX   0 X   -0.01 LYS 177 B
ASP 178 B                                            0.00 XXX   0 X    0.00 XXX   0 X   -0.10 ARG 230 B

ASP 194 B   4.94   100 %    3.54  587   0.64    0   -0.85 HIS  40 B   -0.65 TRP 141 B   -1.44 HIS  40 B
ASP 194 B                                            0.00 XXX   0 X   -0.10 ASP 194 B    0.00 XXX   0 X

GLU  20 B   3.11     8 %    0.82  304   0.11    0   -0.83 SER  11 B   -0.80 SER  11 B    0.00 XXX   0 X
GLU  20 B                                           -0.70 GLN 157 B    0.00 XXX   0 X    0.00 XXX   0 X

GLU  21 B   4.33     8 %    0.52  303   0.03    0   -0.42 ARG 154 B    0.00 XXX   0 X    0.02 GLU  20 B
GLU  21 B                                            0.00 XXX   0 X    0.00 XXX   0 X    0.05 ASP  72 B
GLU  21 B                                            0.00 XXX   0 X    0.00 XXX   0 X   -0.37 ARG 154 B
```

Plain Text ▾ Tab Width: 8 ▾ Ln 1, Col 1 ▾ INS

Εικόνα 10. Πρώτος πίνακας του αρχείου PROPKA του υποδοχέα συμπλόκου 1CGI του unbound αρχείου.

```
1CGI_r_u.propka
10.65.2.238 /home/koutroumpis/Data/Easy/HADDOCK/1CGI
------------------------------------------------------------------------------------------------------------
SUMMARY OF THIS PREDICTION
     RESIDUE    pKa   pKmodel   ligand atom-type
   ASP  35 B    3.03      3.80
   ASP  64 B    4.59      3.80
   ASP  72 B    3.32      3.80
   ASP 102 B    3.69      3.80
   ASP 128 B    3.10      3.80
   ASP 129 B    3.91      3.80
   ASP 153 B    6.20      3.80
   ASP 178 B    3.73      3.80
   ASP 194 B    4.94      3.80
   GLU  20 B    3.11      4.50
   GLU  21 B    4.33      4.50
   GLU  49 B    4.16      4.50
   GLU  70 B    6.33      4.50
   GLU  78 B    4.39      4.50
   C-  245 B    2.34      3.20
   HIS  40 B    7.19      6.50
   HIS  57 B    6.57      6.50
   CYS   1 B   99.99     99.99
   CYS  42 B   99.99     99.99
   CYS  58 B   99.99     99.99
   CYS 122 B   99.99     99.99
   CYS 136 B   99.99     99.99
   CYS 168 B   99.99     99.99
   CYS 182 B   99.99     99.99
   CYS 191 B   99.99     99.99
   CYS 201 B   99.99     99.99
   CYS 220 B   99.99     99.99
   TYR  94 B   11.51     10.00
   TYR 146 B   11.00     10.00
   TYR 171 B   10.21     10.00
   TYR 228 B   12.58     10.00
   LYS  36 B   10.48     10.50
   LYS  79 B   10.44     10.50
   LYS  82 B   10.21     10.50
   LYS  84 B   10.41     10.50
   LYS  87 B   10.31     10.50
   LYS  90 B   10.26     10.50
   LYS  93 B   10.30     10.50
   LYS 107 B   11.27     10.50
   LYS 169 B   10.41     10.50
   LYS 170 B   10.35     10.50
```

Εικόνα 11. Δεύτερος πίνακας του αρχείου PROPKA του υποδοχέα συμπλόκου 1CGI του unbound αρχείου.

```
1CGI_r_u.propka
10.65.2.238 /home/koutroumpis/Data/Easy/HADDOCK/1CGI
   ARG 154 B   13.36     12.50
   ARG 230 B   11.24     12.50
   N+    1 B    7.67      8.00
------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------
Free energy of   folding (kcal/mol) as a function of pH (using neutral reference)
  0.00      9.39
  1.00      9.36
  2.00      9.09
  3.00      7.85
  4.00      6.47
  5.00      7.75
  6.00     10.03
  7.00     11.48
  8.00     11.81
  9.00     11.79
 10.00     12.11
 11.00     13.33
 12.00     14.62
 13.00     15.22
 14.00     15.64

The pH of optimum stability is  4.0 for which the free energy is   6.5 kcal/mol at 298K
Could not determine pH values where the free energy is within 80 % of maximum
Could not determine where the free energy is positive

Protein charge of folded and unfolded state as a function of pH
    pH  unfolded  folded
  0.00     21.00   20.99
  1.00     20.98   20.92
  2.00     20.78   20.34
  3.00     19.23   17.87
  4.00     13.41   13.30
  5.00      7.69    9.34
  6.00      5.72    7.20
  7.00      4.40    5.03
  8.00      3.48    3.45
  9.00      2.30    2.37
 10.00     -2.37   -1.88
 11.00    -11.39  -10.16
 12.00    -15.49  -14.95
 13.00    -17.99  -17.56
 14.00    -18.87  -18.72
The pI is  9.71 (folded) and  9.66 (unfolded)
```

Εικόνα 12. Τρίτος και τέταρτος πίνακας του αρχείου PROPKA του υποδοχέα συμπλόκου 1CGI του unbound αρχείου.

A4. Αρχείο DockQ

Το αρχείο DockQ αναγράφει σε έναν πίνακα πάνω από κάθε υπολογισμό την κατηγοριοποίηση των συμπλόκων με βάση το πρόγραμμα DockQ. Έπειτα αναγράφονται το μοντέλο που αξιολογείται και το μοντέλο αναφοράς. Στο αρχείο αναγράφονται επίσης τα αποτελέσματα των τριών μεταβλητών (iRMS, LRMS, Fnat) και τέλος ο αριθμός DockQ.

```
****************************************************************
*                       DockQ                                  *
*   Scoring function for protein-protein docking models        *
*   Statistics on CAPRI data:                                  *
*    0.00 <= DockQ <  0.23 - Incorrect                         *
*    0.23 <= DockQ <  0.49 - Acceptable quality                *
*    0.49 <= DockQ <  0.80 - Medium quality                    *
*            DockQ >= 0.80 - High quality                      *
*   Reference: Sankar Basu and Bjorn Wallner, DockQ: A quality *
*   measure for protein-protein docking models, submitted      *
*                                                              *
*   For the record:                                            *
*   Definition of contact <5A (Fnat)                           *
*   Definition of interface <10A all heavy atoms (iRMS)        *
*   For comments, please email: bjorn.wallner@.liu.se          *
*                                                              *
****************************************************************
Model  : complex_1.pdb
Native : 1CGI_r-l_b.pdb
Number of equivalent residues in chain A 245 (receptor)
Number of equivalent residues in chain B 56 (ligand)
Fnat 0.388 33 correct of 85 native contacts
Fnonnat 0.421 24 non-native of 57 model contacts
iRMS 2.984
LRMS 4.735
DockQ 0.451
```

Εικόνα 13. Τμήμα αρχείου TXT με πληροφορίες σύγκρισης μέσω DockQ μοντέλων του συμπλόκου 1CGI με το γνωστό σύμπλοκο.

B1. Script αυτοματοποίησης του PDB2PQR

Για το αρχείο αυτό το “n” αναφέρεται στον αριθμό των συμπλόκων, δηλαδή στην παρούσα εργασία το 200 για το HADDOCK και το 50 για το κάθε τρόπο αξιολόγησης του InterEvDock2. Όπου αναγράφεται το “`file_name$i`” αφορά στο όνομα του αρχείου του συμπλόκου. Για παράδειγμα “`complex_$i`” για το HADDOCK, ενώ “Complex_FRODOCK$i”, “Complex_IES$i” και “`Complex_SOAP_PP$i`” για το InterEvDock2. Για να αποφευχθεί η ανάγκη προσθήκης και τον μονοπατιών για την εύρεση των αρχείων των μοντέλων, το script εκτελείται από τον ίδιο φάκελο με τα μοντέλα του κάθε συμπλόκου.

Bash script:

```
#n=number of complexes
for i in {1..n}
do
pdb2pqr --ff=CHARMM --with-ph=7.0 file_name$i.pdb
file_name$i.pqr
done
```

B2. Tcl script αυτοματοποίησης του υπολογισμού του DML μέσω του VMD

Εδώ το “number” αφορά στον αριθμό των αρχείων των μοντέλων και “file_name$i” στο όνομά τους όπως αναφέρθηκε στο Β1. Σε περίπτωση που το script δεν βρίσκεται στον ίδιο φάκελο με τα αρχεία των μοντέλων, είναι απαραίτητο να προστεθεί με την εντολή “cd” το μονοπάτι για την εύρεσή τους. Το όνομα του αρχείου που προκύπτει είναι “DML_file.txt” και εμπεριέχει όλα τα αποτελέσματα των μοντέλων ενός συμπλόκου.

Tcl script:

```
#directory path where the PQR files are or run the file in the
same directory as the file
#cd pqr_directory_path
#number of files + 1
set n number
#creat text file
set outfile [open "DML_file.txt" w];
#dipole moment length calculation loop
for { set i 1 } { $i < $n } { incr i } {
#upload the PQR file in VMD
mol new file_name$i.pqr
#select the last protein that was uploaded
set sel [atomselect top "protein"]
#calculate dipole vector from the geometrical center of the
protein
set DM [measure dipole $sel -debye -geocenter]
#calculate the vector’s length
set DM_length [veclength $DM]
#write the value in the text file
puts $DM_length
puts $outfile "$i $DM_length"
mol delete top
}
close $outfile
```

B3. Tcl script αυτοματοποίησης του υπολογισμού του R_g μέσω του VMD

Αντίστοιχα για τον υπολογισμό του R_g στο αρχείο το “number” αφορά στον αριθμό των αρχείων των μοντέλων και “file_name$i”στο όνομά τους όπως αναφέρθηκε στο Β1. Σε περίπτωση που το script δεν βρίσκεται στον ίδιο φάκελο με τα αρχεία των μοντέλων, είναι απαραίτητο να προστεθεί με την εντολή “cd” το μονοπάτι για την εύρεσή τους. Το όνομα του αρχείου που προκύπτει είναι “Rg.txt” και εμπεριέχει τα αποτελέσματα για όλα τα μοντέλα του συμπλόκου.

Tcl script:

```
#directory path where the PQR files are or run the file in the
same directory as the file
#cd pqr_directory_path
#number of files + 1
```

```
set npdb number
set outfile [open "Rg.txt" w]
#rg calculation loop
for { set i 1 } { $i < $npdb } { incr i } {
mol load pdb file_name$i.pqr
#select the last protein that was uploaded
set pdb [atomselect top "protein"]
#calculate the Rg
set Rg [measure rgyr $pdb]
puts $outfile "$i $Rg"
#delete the last complex that was loaded
mol delete top
}
close $outfile
```

B4. Script αυτοματοποίησης του PROPKA

Για το script αυτό το "`n`" αναφέρεται στον αριθμό των συμπλόκων, δηλαδή στην παρούσα εργασία το 200 για το HADDOCK και το 50 για το κάθε τρόπο αξιολόγησης του InterEvDock2. Όπου αναγράφεται το "`file_name$i`" αφορά στο όνομα του αρχείου του συμπλόκου. Για παράδειγμα "`complex_$i`" για το HADDOCK, ενώ "`Complex_FRODOCK$i`", "`Complex_IES$i`" και "`Complex_SOAP_PP$i`" για το InterEvDock2. Για να αποφευχθεί η ανάγκη προσθήκης και τον μονοπατιών για την εύρεση των αρχείων των μοντέλων, το script εκτελείται από τον ίδιο φάκελο με τα μοντέλα του κάθε συμπλόκου.

Bash script:

```
#n=number of complexes
for i in {1..n}
do
pdb2pqr --ff=CHARMM --with-ph=7.0 --ph-calc-method=propka
file_name$i.pdb file_name$i.pqr
done
```

B5. Συλλογή πληροφοριών από τα αρχεία PROPKA

Το συγκεκριμένο script έχει την δυνατότητα να αντιγράφει την πληροφορία από τον δεύτερο πίνακα των αρχείων PROPKA που αφορά στο pKa του κάθε αμινοξέως και έπειτα να κάνει την πρόσθεσή τους ώστε να υπολογιστεί το σύνολο του pK_a για το μοντέλο. Και εδώ το "`n`" αναφέρεται στον αριθμό των αρχείων των μοντέλων και "`file_name$i`" στο όνομα τους όπως αναφέρθηκε στο B1. Το τελικό αρχείο που προκύπτει με τα συνολικά pK_a των μοντέλων ενός συμπλόκου αναγράφεται ως "`output_file`".

Bash script:

```
# Start copying lines from line next to "SUMMARY OF THIS
PREDICTION"
# Stop copying lines at line before "-----------------"
# Write into a temp.txt file
#n=number of complexes
for i in {1..n}
do
  sed -n '/SUMMARY OF THIS PREDICTION/,/-----------------/ w
temp.txt' file_name_$i.propka
# Remove first and last line from temp.txt and save to
pKa_table.txt
  sed -i '1d' temp.txt
#  sed -e '$d' temp.txt > pKa_table_model_$i.txt
  sed -i '1d' temp.txt
#  Extract pKa (column 4) from temp.txt and save it into
temp1.txt
  awk '{print $4}' temp.txt > pKa.txt
# Sum pKa from pKa.txt
awk -v var="$i" '{Total=Total+$1} END{print var " " Total}'
pKa.txt >> output_file.txt
  rm temp.txt
  rm pKa.txt
done
```

B6. Tcl script αυτοματοποίησης του υπολογισμού του RMSD μέσω του VMD

Στο script αυτό το "`number`" αναφέρεται στον αριθμό των αρχείων των μοντέλων και "`file_name$i`" στο όνομά τους όπως αναφέρθηκε στο B1. Στο συγκεκριμένο Tcl script είναι αναγκαίο να προστεθεί και το αρχείο του γνωστού συμπλόκου που αναγράφεται παρακάτω ως "`bound_file.pdb`". Σε περίπτωση που το script δεν βρίσκεται στον ίδιο φάκελο με τα αρχεία των μοντέλων, είναι απαραίτητο να προστεθεί με την εντολή "`cd`" το μονοπάτι για την εύρεσή τους. Το όνομα του αρχείου που προκύπτει είναι "`rmsd_file.txt`" και εμπεριέχει όλα τα αποτελέσματα των μοντέλων ενός συμπλόκου.

Tcl script:

```
#directory path where the PDB files are or run the file in the
same directory as the file
#cd pqr_directory_path
#load the known bound state of the protein complex
mol load pdb bound_file.pdb
#specify the number of complexes + 1
set npdb number
#creates the text file where the values will be written
set outfile [open rmsd_file.txt w];
#selects the backbone of the whole protein without the
hydrogens
set pdb_0 [atomselect top "protein and backbone and noh"]
```

```
# rmsd calculation loop
for {set i 1 } {$i < $npdb } { incr i } {
#loads each complex (specify their location)
mol load pdb file_name$i.pdb
set pdb [atomselect top "protein and backbone and noh"]
#fits the selected areas of the two complexes (known bound –
complex)
$pdb move [measure fit $pdb $pdb_0]
#calculates rmsd
set rmsd [measure rmsd $pdb $pdb_0]
puts $outfile "$i $rmsd"
mol delete top
}
close $outfile
```

B7. Το bash script αυτοματοποίησης του προγράμματος DockQ

Το script αυτό αυτοματοποιεί την διαδικασία υπολογισμού του DockQ για όλα τα μοντέλα ενός συμπλόκου. Στο αρχείο το “n” αναφέρεται στον αριθμό των μοντέλων του συμπλόκου με όνομα “file_name$i”, ενώ το “bound_file” αναφέρεται στο σύμπλοκο αναφοράς. Για την λειτουργία της εντολής “DockQ.py” πρέπει να προστεθεί και το μονοπάτι εγκατάστασης των αρχείων του DockQ. Το αρχείο TXT που προκύπτει ονομάζεται “dockq_calc”.

Bash script:

```
#This file runs the DockQ program for a number of inputs
#n=number of inputs
for i in {1..n}
do
 /DockQ_file_path/DockQ.py file_name$i.pdb bound_file.pdb >>
dockq_calc.txt
 echo $i
done
```

B8. Συλλογή πληροφοριών από το αρχείο DockQ

Από το αρχείο που προκύπτει από το B7, γίνεται συλλογή των μεταβλητών DockQ. Το τελικό αρχείο TXT με τις πληροφορίες αποθηκεύεται με την ονομασία “DockQ_final”.

Bash script:

```
#Collecting the information we need from two separate lines
from one file
#Info from first line
line1=$(sed -n '/^Model/p' dockq_calc.txt)
echo "$line1" > one.txt
#Separating the unwanted word from the file into another
```

```
sed 's/Model  : //g' one.txt > three.txt
#Info from second line
line2=$(sed -n '/^DockQ/p' dockq_calc.txt)
echo "$line2" > two.txt
#Separating the unwanted word from the file into another
sed 's/DockQ //g' two.txt > four.txt
pr -m -t three.txt four.txt > DockQ_final.txt
#Deleting the junk files
rm one.txt two.txt three.txt four.txt
```

Βιβλιογραφία

- Andreani, J., Faure, G., & Guerois, R. (2013). InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics (Oxford, England)*, *29*(14), 1742–1749. https://doi.org/10.1093/bioinformatics/btt260
- Basu, S., & Wallner, B. (2016a). DockQ: A Quality Measure for Protein-Protein Docking Models. *PloS one*, *11*(8), e0161879. https://doi.org/10.1371/journal.pone.0161879
- Basu, S., & Wallner, B. (2016b). Finding correct protein-protein docking models using ProQDock. *Bioinformatics (Oxford, England)*, *32*(12), i262–i270. https://doi.org/10.1093/bioinformatics/btw257
- Berg, J. M., Tymoczko, J. L., & Stryer, L. (2015) *Βιοχημεία*, 2η έκδοση, ΠΑΝΕΠΙΣΤΗΜΙΑΚΕΣ ΕΚΔΟΣΕΙΣ ΚΡΗΤΗΣ, ΗΡΑΚΛΕΙΟ, Κεφάλαιο 2, 27-66.
- Blackburn, P., & Moore, S. (1982). 12 Pancreatic Ribonuclease. In *The enzymes* (Vol. 15, pp. 317-433). Academic Press. https://doi.org/10.1016/S1874-6047(08)60284-X
- Bunn H. F. (2013). Erythropoietin. *Cold Spring Harbor perspectives in medicine*, *3*(3), a011619. https://doi.org/10.1101/cshperspect.a011619
- Chen Y. C. (2015). Beware of docking!. *Trends in pharmacological sciences*, *36*(2), 78–95. https://doi.org/10.1016/j.tips.2014.12.001
- Chys, P., & Chacón, P. (2013). Random Coordinate Descent with Spinor-matrices and Geometric Filters for Efficient Loop Closure. *Journal of chemical theory and computation*, *9*(3), 1821–1829. https://doi.org/10.1021/ct300977f
- de Ruyck, J., Brysbaert, G., Blossey, R., & Lensink, M. F. (2016). Molecular docking as a popular tool in drug design, an in silico travel. *Advances and applications in bioinformatics and chemistry : AABC*, *9*, 1–11. https://doi.org/10.2147/AABC.S105289
- de Vries SJ, Bonvin AMJJ (2011) CPORT: A Consensus Interface Predictor and Its Performance in Prediction-Driven Docking with HADDOCK. PLoS ONE 6(3): e17695. https://doi.org/10.1371/journal.pone.0017695
- Dickson, K. A., Haigis, M. C., & Raines, R. T. (2005). Ribonuclease inhibitor: structure and function. *Progress in nucleic acid research and molecular biology*, *80*, 349–374. https://doi.org/10.1016/S0079-6603(05)80009-1
- Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. (2004). PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics

calculations. *Nucleic acids research*, *32*(Web Server issue), W665–W667. https://doi.org/10.1093/nar/gkh381

- Dominguez, C., Boelens, R., & Bonvin, A. M. (2003). HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, *125*(7), 1731-1737.
- Dong, G. Q., Fan, H., Schneidman-Duhovny, D., Webb, B., & Sali, A. (2013). Optimized atomic statistical potentials: assessment of protein interfaces and loops. *Bioinformatics (Oxford, England)*, *29*(24), 3158–3166. https://doi.org/10.1093/bioinformatics/btt560
- Ferreira, L. G., Dos Santos, R. N., Oliva, G., & Andricopulo, A. D. (2015). Molecular docking and structure-based drug design strategies. *Molecules (Basel, Switzerland)*, *20*(7), 13384–13421. https://doi.org/10.3390/molecules200713384
- Finn J. (2012). Application of SBDD to the discovery of new antibacterial drugs. *Methods in molecular biology (Clifton, N.J.)*, *841*, 291–319. https://doi.org/10.1007/978-1-61779-520-6_13
- Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft*, *27*(3), 2985-2993.
- Garzon, J. I., Lopéz-Blanco, J. R., Pons, C., Kovacs, J., Abagyan, R., Fernandez-Recio, J., & Chacon, P. (2009). FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics (Oxford, England)*, *25*(19), 2544–2551. https://doi.org/10.1093/bioinformatics/btp447
- Gräwert, T. W., & Svergun, D. I. (2020). Structural Modeling Using Solution Small-Angle X-ray Scattering (SAXS). *Journal of molecular biology*, *432*(9), 3078–3092. https://doi.org/10.1016/j.jmb.2020.01.030
- Grosberg A. Y., Khokhlov A. R., Pande V. S. (1994). *Statistical Physics of Macromolecules*. AIP Press. ISBN 1-56396-071-0
- Guex, N and Peitsch, M.C. (1996). Swiss-PdbViewer: A Fast and Easy-to-use PDB Viewer for Macintosh and PC. *Protein Data Bank Quaterly Newsletter*, 77, 7.
- Guex, N., & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, *18*(15), 2714–2723. https://doi.org/10.1002/elps.1150181505
- H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank *Nucleic Acids Research, 28: 235-242.*
- Housecroft, C. E., & Sharpe, A. G. (2008). *Inorganic chemistry* (Vol. 1). Pearson Education. Chapter 2, 44-46.

- Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), 33–28. https://doi.org/10.1016/0263-7855(96)00018-5
- IBM Corp. (2020). IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp. (Released 2020) https://www.ibm.com/analytics/spss-statistics-software
- Joon, S., Singla, R. K., & Shen, B. (2022). In Silico Drug Discovery for Treatment of Virus Diseases. *Advances in experimental medicine and biology*, *1368*, 73–93. https://doi.org/10.1007/978-981-16-8969-7_4
- Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L. E., Brookes, D. H., Wilson, L., Chen, J., Liles, K., Chun, M., Li, P., Gohara, D. W., Dolinsky, T., Konecny, R., Koes, D. R., Nielsen, J. E., Head-Gordon, T., Geng, W., Krasny, R., ... Baker, N. A. (2018). Improvements to the APBS biomolecular solvation software suite. *Protein science : a publication of the Protein Society*, *27*(1), 112–128. https://doi.org/10.1002/pro.3280
- Kefala Georgia-Maria, (2021). "Improving ranking of models derived from protein-protein docking using Dipole Moment, Rg and pKa", *Aristotle University of Thessaloniki Faculty of Science, Physics Department*
- Kirchmair, J., Markt, P., Distinto, S., Wolber, G., & Langer, T. (2008). Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection--what can we learn from earlier mistakes?. *Journal of computer-aided molecular design*, *22*(3-4), 213–228. https://doi.org/10.1007/s10822-007-9163-6
- Kobe, B., & Kajava, A. V. (2001). The leucine-rich repeat as a protein recognition motif. *Current opinion in structural biology*, *11*(6), 725–732. https://doi.org/10.1016/s0959-440x(01)00266-4
- Lensink, M. F., Méndez, R., & Wodak, S. J. (2007). Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins*, *69*(4), 704–718. https://doi.org/10.1002/prot.21804
- Li, H., Robertson, A. D., & Jensen, J. H. (2005). Very fast empirical prediction and rationalization of protein pKa values. *Proteins*, *61*(4), 704–721. https://doi.org/10.1002/prot.20660
- Li, Q., & Kang, C. (2017). Erythropoietin Receptor Structural Domains. *Vitamins and hormones*, *105*, 1–17. https://doi.org/10.1016/bs.vh.2017.02.005
- Lin, X., Li, X., & Lin, X. (2020). A Review on Applications of Computational Methods in Drug Screening and Design. *Molecules (Basel, Switzerland)*, *25*(6), 1375. https://doi.org/10.3390/molecules25061375

- López-Blanco, J. R., Canosa-Valls, A. J., Li, Y., & Chacón, P. (2016). RCD+: Fast loop modeling server. *Nucleic acids research*, *44*(W1), W395–W400. https://doi.org/10.1093/nar/gkw395
- Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J., & Weng, Z. (2005). Protein-Protein Docking Benchmark 2.0: an update. *Proteins*, *60*(2), 214–216. https://doi.org/10.1002/prot.20560
- Miura K. (2018). An Overview of Current Methods to Confirm Protein-Protein Interactions. *Protein and peptide letters*, *25*(8), 728–733. https://doi.org/10.2174/0929866525666180821122240
- Onufriev, A. V., & Alexov, E. (2013). Protonation and pK changes in protein-ligand binding. *Quarterly reviews of biophysics*, *46*(2), 181–209. https://doi.org/10.1017/S0033583513000024
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084
- Quignot, C., Rey, J., Yu, J., Tufféry, P., Guerois, R., & Andreani, J. (2018). InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. *Nucleic acids research*, *46*(W1), W408–W416. https://doi.org/10.1093/nar/gky377
- Ramírez-Aportela, E., López-Blanco, J. R., & Chacón, P. (2016). FRODOCK 2.0: fast protein-protein docking server. *Bioinformatics (Oxford, England)*, *32*(15), 2386–2388. https://doi.org/10.1093/bioinformatics/btw141
- RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences (2021) *Nucleic Acids Research* 49: D437–D451 doi: 10.1093/nar/gkaa1038
- Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., Topf, M., & Sali, A. (2004). A structural perspective on protein-protein interactions. *Current opinion in structural biology*, *14*(3), 313–324. https://doi.org/10.1016/j.sbi.2004.04.006
- Sabe, V. T., Ntombela, T., Jhamba, L. A., Maguire, G., Govender, T., Naicker, T., & Kruger, H. G. (2021). Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *European journal of medicinal chemistry*, *224*, 113705. https://doi.org/10.1016/j.ejmech.2021.113705

- Scarpino, A., Ferenczy, G. G., & Keserű, G. M. (2018). Comparative Evaluation of Covalent Docking Tools. *Journal of chemical information and modeling*, *58*(7), 1441–1458. https://doi.org/10.1021/acs.jcim.8b00228
- Syed, R. S., Reid, S. W., Li, C., Cheetham, J. C., Aoki, K. H., Liu, B., Zhan, H., Osslund, T. D., Chirino, A. J., Zhang, J., Finer-Moore, J., Elliott, S., Sitney, K., Katz, B. A., Matthews, D. J., Wendoloski, J. J., Egrie, J., & Stroud, R. M. (1998). Efficiency of signalling through cytokine receptors depends critically on receptor orientation. *Nature*, *395*(6701), 511–516. https://doi.org/10.1038/26773
- The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC. https://pymol.org/2/
- Tobi, D. (2010). Designing coarse grained-and atom based-potentials for protein-protein docking. *BMC Structural Biology*, *10*(1), 1-11.
- Tro, Nivaldo J. (2008). *Chemistry: A Molecular Approach*. Upper Saddle River: Pearson Education. 379-386.
- Vajda, S., Hall, D. R., & Kozakov, D. (2013). Sampling and scoring: a marriage made in heaven. *Proteins*, *81*(11), 1874–1884. https://doi.org/10.1002/prot.24343
- Vakser, I. A. (2014). Protein-protein docking: From interaction to interactome. *Biophysical journal*, *107*(8), 1785-1793.
- Van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., Melquiond, A. S. J., Van Dijk, M., de Vries S.J. & Bonvin, A. M. J. J. (2016). The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *Journal of molecular biology*, *428*(4), 720-725.
- Verkhivker, G. M., Bouzida, D., Gehlhaar, D. K., Rejto, P. A., Arthurs, S., Colson, A. B., Freer, S. T., Larson, V., Luty, B. A., Marrone, T., & Rose, P. W. (2000). Deciphering common failures in molecular docking of ligand-protein complexes. *Journal of computer-aided molecular design*, *14*(8), 731–751. https://doi.org/10.1023/a:1008158231558
- Viceconti, M., Pappalardo, F., Rodriguez, B., Horner, M., Bischoff, J., & Musuamba Tshinanu, F. (2021). In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products. *Methods (San Diego, Calif.)*, *185*, 120–127. https://doi.org/10.1016/j.ymeth.2020.01.011
- Yu, J., Vavrusa, M., Andreani, J., Rey, J., Tufféry, P., & Guerois, R. (2016). InterEvDock: a docking server to predict the structure of protein-protein interactions using evolutionary information. *Nucleic acids research*, *44*(W1), W542–W549. https://doi.org/10.1093/nar/gkw340