Η δημοσίευση καταλήγει σε έναν αλγόριθμο-μέθοδο που καταφέρνει να αποθηκεύει και να αναλύει
τρισδιάστατα μοντέλα πρωτεϊνών, χρησιμοποιώντας ελάχιστη υπολογιστική ισχύ για τις συγκρίσεις. Βασίζεται πάνω σε κάποια βασικά μαθηματικά εργαλεία-θεωρήματα, όπως της διακύμανσης, της
συνάρτησης αυτο-συσχέτισης, του μετασχηματισμού Fourier, της Density Functional Theory (DTF), της στατιστικής ανάλυσης, του Molecular Electrostatic Potential (MESP), και του Power Spectrum Density (PSD).

~~Αφού τελειώσει με~~ Μετά την επεξήγηση και την κατασκευή ~~του μοντέλου~~ της μεθόδου, η εργασία ~~μας μιλάει για~~ αναφέρει τα θετικά αποτελέσματά της εφαρμογής της πάνω σε πρωτεϊνες και έντζυμα ιών που σχετίζονται με διαδεδομένες ασθένειες όπως Ηπατίτηδα C , Δάγκειος πυρετός, Κίτρινος πυρετός , ιογενής διάρροια βοοειδών και Πυρετός του δυτικού Νείλου.

Πιο συγκεκριμένα βρέθηκε ότι οι μετρήσεις μέσω PSD ορίζουν τα διαφορετικά μόρια με μοναδικό τρόπο, δηλαδή δίνουν σε κάθε διαφορετικό μόριο μοναδική «υπογραφή» και έτσι μπορεί να γίνει η αναγνώρισή τους αποδοτικότερη. Η μόνη δυσκολία έγκειται στη σωστή ~~υπολογισμό~~ εφαρμογή της μεθόδου σε τόσο μικρή κλίμακα.

Αρχικά ας αναλύσουμε κάποια μαθηματικά εργαλεία που ~~θα χρησιμεύσουν~~ χρησιμοποιήθηκαν. Ένα από αυτά είναι η διακύμανση. Η διακύμανση δείχνει το πόσο πολύ απλώνεται ένα σύνολο αριθμών (ή τιμών όταν μιλάμε για μια συνάρτηση ή σήμα), από την στατιστική μέση τιμή του. Με Xi=δείγματα, μ=στατιστική μέση τιμή (πηγή:
https://eclass.upatras.gr/modules/document/file.php/CEID1081/lecture5.pdf ).
Η συνάρτηση αυτοσυσχέτισης ενός σήματος, εκφράζει την διακύμανση του σήματος την χρονική στιγμή t συγκριτικά με μια επόμενη χρονική στιγμή t+τ. Είναι σημαντική διότι μας δείχνει την ταχύτητα μεταβολής της διακύμανσης. Αν η διακύμανση αλλάζει γρήγορα, η R(τ) συγκλίνει άμεσα, ενώ αν η διακύμανση αλλάζει αργά, η R(τ) συγκλίνει αργά. Η εξίσωση της αυτο-συσχέτισης είναι:
.

Όπως βλέπουμε εδώ μετράμε την μεταβολή της διακύμανσης χωρίς κάποια αφαίρεση (όπως στον παραπάνω τύπο του var). Αυτό διότι, αντί για αφαίρεση θέσαμε τα άκρα του ολοκληρώματος σε T και -T. Ο όρος της μέσης τιμής επιτυγχάνεται διαιρώντας με 2T, που αν ήμασταν στο διακριτό χώρο θα ήταν ο αριθμός των δειγμάτων της δειγματοληψίας (πηγή:
https://physics.mcmaster.ca/phys4d06/LectureSlides-Ch9_Autocorr-Power-Noise.pdf ).

Ο όρος ~~του~~ Power Spectrum Density δηλώνεται στην εργασία ως power spectrum of surface. Το PSD ενός σήματος είναι ο μετασχηματισμός Fourier της συνάρτησης αυτοσυσχέτισης του σήματος. Μέσω του Fourier, παίρνουμε την αποσύνθεσή των σημάτων(των επιφανειών στη συγκεκριμένη εργασία) σε πολλά έως άπειρα απλούστερα σήματα διάφορων συχνοτήτων. Λόγω της φύσης της R(τ), στη μέτρηση PSD περιέχεται στατιστικά η τιμή του σήματος σε όλες τις κυματομορφές (το άθροισμά τους), χωρίς να αποθηεύεται πουθενά η φάση. Αυτό είναι όμως ό,τι χρειαζόμαστε αργότερα που θα ορίσουμε το MESP.

Ένα μεγάλο πλεονέκτημα του PSD ως μαθηματικού εργαλείου είναι ότι, οι στατιστικές πληροφορίες που περιέχει για μια επιφάνεια δεν επηρεάζονται από το μέγεθος του σχήματος ή

την αναπαράσταση σε pixels. Στην περίπτωση που γνωρίζουμε τέλεια την εξεταζόμενη επιφάνεια, ως συνεχή χαρτογράφηση υψομέτρων h(x,y) σε οριζόντια θέση x,y με μηδενική μέση τιμή, οι 3 διαστάσεις μπορούν να υπολογιστούν με την μέθοδο real-space topology. Η μέση τετραγωνική ρίζα δίνεται ως:
Όπου οι δύο αγκύλες υποδηλώνουν την μέση τιμή στον x-y άξονα, που υποδηλώνεται μέσω της h. Η κλίση ισούται με την παράγωγο, δηλαδή με όπου το υποδεικνύει τη μερική παράγωγο του h. Η
καμπυλότητα δίνεται από το . Από αυτές τις διαστάσεις συνήθως μόνο μία είναι σημαντική και κυρίαρχη να γνωρίζουμε. Έτσι δίνεται ένα παράδειγμα για όταν θέλουμε να μετρήσουμε την τραχύτητα μεταξύ δύο επιφανειών αρκεί η τετραγωνική διαφορά υψών hrms.
Οι τιμές είναι μοναδικές για κάθε επιφάνεια αν μετρηθούν με αρκετή ακρίβεια, και αυτό είναι μια πολύ χρήσιμη ιδιότητα για τον προσδιορισμό μιας πρωτεϊνης. Το πρόβλημα στα πραγματικά συστήματα έγκειται στο να υπολογιστεί με άπειρη ακρίβεια η τριάδα των υψών, άρα θα πρέπει να κάνουμε συμβιβασμούς, να χρησιμοποιούμε τεχνικές προσαρμοσμένες στις ιδιαιτερότητες κάθε συστήματος και να επαληθεύουμε τα αποτελέσματα κάθε μοντέλου.
Πιο συγκεκριμένα υπάρχει αλγόριθμος για μείωση σφαλμάτων μέτρησης αλλά και αναπαράστασης στον τρισδιάστατο χώρο. Υπάρχουν επίσης και αλγόριθμοι για ακρίβεια όταν πρόκειται για απειροελάχιστα μικρές επιφάνειες. Αυτοί οι αλγόριθμοι είναι πιθανό να αξιοποιήθηκαν στην παρούσα εργασία.

Οι συγγραφείς του paper δεν αναφέρθηκαν πολύ στις τεχνικές ακρίβειας μέτρησης, διότι αυτές μπορούν να αλλάξουν σε κάθε περίσταση. ~~Οι τελευταίοι,~~ Αφοσιώθηκαν περισσότερο στον αλγόριθμο υλοποίησης και ανέφεραν ονομαστικά-αριθμητικά τις-κατανομές και τα ανεκτά μεγέθη των σφαλμάτων της πειραματικής επαλήθευσής τους. Έτσι θα αναφέρουμε κάποια πράγματα απλά ονομαστικά.

Πρόκληση Α: Ποικιλίες στον ορισμό του PSD μπορεί να περιλαμβάνουν διαφορετικές μονάδες μέτρησης όπως στην εργασία κινούμαστε στις 3 διαστάσεις και πιο συγκεκριμένα στην σφαίρα του Fourier.
Στρατηγική Α: Χρησιμοποιούμε μια προτεινόμενη μέθοδο ανάλογα αν το σχήμα μας είναι ισοτροπικό ή μη. Ισοτροπικό είναι ένα σχήμα όταν οι ιδιότητές του διαφέρουν ανάλογα την κατεύθυνση των μετρήσεων. Στην εργασία που μελετούμε έγινε υπόθεση για τους υπολογισμούς ότι τα μόρια είναι ισοτροπικά. Έτσι απορρίπτεται κάθε ενασχόληση με την κατεύθυνση της μέτρησης και ~~πλέον~~ η μόνη μέτρηση που καθορίζει την τιμή είναι η απόσταση μεταξύ σημείων.

Πρόκληση Β: Όταν ~~χωρίζουμε~~ χωρίζεται σε κυματομορφές ~~την~~ η αρχική μέτρηση της επιφάνειας, αυτές δεν μπορούν να είναι άπειρες όπως στο θεωρητικό μοντέλο. Άρα χρειάζεται προσοχή στην ανακατασκευή του σήματος. Επίσης το εύρος συχνοτήτων είναι περιορισμένο.
Στρατηγική Β: Να ~~παίρνουμε και να~~ συνδυάζονται πολλές μετρήσεις από διαφορετικές διαστάσεις και από διαφορετικές τεχνικές μέτρησης, ~~Με μια λέξη,~~ με άλλα λόγια να γίνεται επαλήθευση. Στην εργασία, έγιναν μετρήσεις των μορίων σε κατάσταση υψηλής και χαμηλής ενέργειας ψάχνοντας τυχόν σημαντικές διαφορές.

Πρόκληση Γ: Μέτρηση της τοπογραφίας σε πολύ μικρές κλίμακες.
Στρατηγική Γ: Πρέπει να ~~καθορίσουμε~~ καθοριστούν τα όρια της ακτίνας και να ~~χρησιμοποιήσουμε~~ χρησιμοποιείται απόλυτη τιμή, σε συνδυασμό με καθορισμό της μέγιστης συχνότητας αναπαράστασης. Τα όρια της ακτίνας καθορίστηκαν στην εργασία πουμελετάμε.

Καθορίστηκε επίσης η διακριτή συχνότητα να είναι διπλάσια από το ρυθμό που αλλάζουν οι συχνότητες
στις κυματομορφές, με σεβασμό στο θεώρημα Nyqist.
Η εργασία, που ~~πλέον έχουμε τις κατάλληλες μαθηματικές γνώσεις να κατανοήσουμε καλύτερα διαισθητικά~~ στηρίζεται στις παραπάνω μαθηματικές γνώσεις, τις οποίες μπορούμε πλέον να κατανοήσουμε διαισθητικά, βασίζεται και σε μια ακόμη καινοτομία, το Molecular electrostatic potential (MESP). ~~Η πηγή που χρησιμοποιήσαμε μας λέει~~
~~ότι~~ Το MESP μπορεί να αποτελέσει ένδειξη για τις χημικές ιδιότητες των μορίων. ~~Το MESP που~~ Κατά μία έννοια είναι η ενέργεια που βρίσκεται αποθηκευμένη σε κάθε μόριο, μετριέται με το Density Functional Theory (DFT) και με στατιστικά στοιχεία. Το DFT με τη σειρά του χρησιμοποιεί το PSD και πλέον γίνεται αντιληπτό ότι μετρώντας την επιφάνεια του μορίου με το PSD μπορούμε να συμπεράνουμε τις ιδιότητές του. Αυτό είναι και το βασικό θεώρημα της εργασίας που ~~θα~~ αναλύουμε. Όσον αφορά το MESP, πήραμε πληροφορίες ~~μόνο~~ από το abstract της παρακάτω
εργασίας ~~διότι το κείμενο ήταν επί πληρωμή~~ :
https://pubs.rsc.org/en/content/articlelanding/2022/cp/d2cp03244a .
~~Στο κυρίως θέμα~~

Το βασικό κείμενο που αναλύουμε, ~~συζητάμε~~ αναφέρεται αρχικά στη μεγάλη ανάγκη για αποδοτική διαχείριση γενετικών πληροφοριών. Τα τελευταία χρόνια το κόστος της ανάλυσης και η ποσότητα της γενετικής πληροφορίας έχουν εκτοξευθεί. Οι σημερινές δυνατότητες των υπολογιστικών συστημάτων δεν επαρκούν για γρήγορη και αποδοτική ανάλυση, και καθώς είναι αδύνατο να αναβαθμίζεται διαρκώς το υλικό των υπολογιστών για να ανταπεξέρχονται στις υπολογιστικές απαιτήσεις, είναι ανάγκη να βρεθεί μια λύση μέσω του λογισμικού, που να αξιοποιεί με τον βέλτιστο τρόπο το υπάρχον υλικό. Κάτι τέτοιο προβλέπεται ότι θα είναι ιδιαίτερα σημαντικό στο μέλλον, όπου είναι πιθανό η ιατρική να είναι «εξατομικευμένη». ~~Δεν μένουμε εκεί όμως. Το κείμενο προτείνει ότι~~ Στην post-genomic εποχή η αναζήτηση ομοιοτήτων ανάμεσα σε πρωτεϊνες βασίζεται κατά κύριο λόγο στις ομοιότητες των ακολουθιών των στοιχείων τους. Μια τέτοια προσέγγιση είναι οι αυτοοργανούμενοι χάρτες (self-organizing maps). Οι ~~κλασσικοί~~
χάρτες αυτοί πρωτεϊνών και
αμινοξέων μπορούν να προσδιορίσουν την οικογένεια των στοιχείων αυτών. Έχουν όμως ένα μειονέκτημα, ότι δεν μπορούν να προσδιορίσουν τη λειτουργία τους. Οι λειτουργίες των μορίων σχετίζονται πολύ με την δομή τους (MESP) και το σχήμα τους. Διαφορετικές μοριακές επιφάνειες πρωτεϊνών κωδικοποιούν διαφορετικές λειτουργικές πληροφορίες. Έτσι η εργασία αυτή προτείνει έναν νέο τρόπο αναγώρισής τους. Όπως είδαμε και στην παραπάνω ανάλυση, οι μετρήσεις σχημάτων με PSD έχουν ένα βαθμό μοναδικότητας και αυτό είναι το στοιχείο που αξιοποιεί η εργασία. Υπάρχουν ~~ωστόσο εργασίες~~ μέθοδοι που μελετούν με ακριβεία τη δομή των πρωτεϊνων, αλλά χρειάζονται πολύ χρόνο και πόρους, απαιτούν ευθυγράμμιση των επιφανειών και δεν είναι αποδοτικές σε χώρο/χρόνο. Με το PSD ~~θα αναλύσουμε~~ αναλύονται λιγότερο ακριβείς μετρήσεις. ~~Θα χρησιμοποιήσουμε~~ Χρησιμοποιείται ο μετασχηματισμός Fourier ο οποίος ανεξαρτήτως κλίμακας αναπαριστά την ίδια πληροφορία. Το μοντέλο παρέχει πολύ γρήγορες συγκρίσεις, κάτι πολύ βασικό στην βιοπληροφορική, αφού εξετάζει απλά την διαφορά των μετρήσεων PSD και δεν κάθεται να συγκρίνει ολόκληρο το σχήμα του μορίου.

~~Σετ εκπαίδευσης~~ Για την εκπαίδευση του μοντέλου χρησιμοποιήθηκαν δείγματα από 4 διαφορετικές οικογένειες πρωτεϊνων: 12 ελικάσες, 4 πολυμεράσες, 6 μεθυλτρανσφεράσεις ή μεθυλάσεις και 4 γλυκοπρωτεϊνες. ~~για τις δοκιμές~~. ~~Θα αποδοθούν με τους αγγλικούς όρους, 12~~

~~από την οικογένεια helicase, 6 από την οικογένεια methylases, 4 από την οικογένεια polymerases και 4 από την οικογένεια glycoproteins.~~ Επίσης χρησιμοποιήθηκε πρωτεϊνη κινάση ποντικιού (Mouse kinase) λόγω της πολύ διαφορετικής δομής της με τις υπόλοιπες. Οι πρωτεΐνες της πρώτης οικογένειας είναι υπέυθυνες για το ξεδίπλωμα της έλικας DNA ή RNA κατά ~~τον πολλαπλάσιασμό~~ την αντιγραφή του γονιδιώματος ιών. Οι πρωτεΐνες της δεύτερης οικογένειας, είναι ένζυμα που αντιγράφουν γενετικό υλικό. Οι πρωτεΐνες της τρίτης οικογένειας, είναι ένζυμα που μεταφέρουν methyl groups από δωρητές σε αποδοχείς. Οι πρωτεΐνες της τέταρτης οικογένειας χρησιμοποιούνται από τους ιούς για μοριακή αναγνώριση.

Το μεγαλύτερο μέρος των σχημάτων των πρωτεϊνών που χρησιμοποιήθηκαν ως treatment set ανακτήθηκε από την βάση RCSB όπου σχηματίστηκαν με κρυσταλλογραφία ακτινών Χ.
Από την πρώτη οικογένεια επιλέχθηκαν οι 1A1V και 80HM πρωτεΐνες του ιού της Ηπατίτιδας C, οι
1YMF, 1YKS και 2V80 πρωτεΐνες του ιού του Κίτρινου πυρετού και οι 2JLU, 2BHR, 2BMF και 2JLQ
πρωτεΐνες του ιού του Δάγκειου πυρετού. Από τη δεύτερη οικογένεια επιλέχθηκαν οι 2CJQ, 2HCS και SHCN πρωτεΐνες του ιού του Δυτικού Νείλου. Από την τρίτη οικογένεια επιλέχθηκαν οι 3EVA, 3EVB, 3EVC, 3EVE και 3EVF πρωτεΐνες του ιού του Κίτρινου πυρετού. Από την τέταρτη οικογένεια επιλέχθηκαν οι 1NB7, 4DVN, 4DW4 και 4DW3 πρωτεΐνες του ιού της ιογενείς διάρροιας βοοειδών.

Οι επιφάνειες του electrostatic potential των μορίων ακολουθούν τη μη-γραμμική εξίσωση των Poisson-Boltzmann αρκετά ικανοποιητικά. Για τον υπολογισμό της δυναμικής ενέργειας χρησιμοποιήθηκε η μέθοδος της πεπερασμένης διαφοράς όπως είναι υλοποιημένη στο λογισμικό APBS. Χρησιμοποιήθηκαν όρια στο μέγιστο μήκος της ακτίνας (όπως προτάθηκε στα θεωρήματα παραπάνω). Η θερμοκρασία τέθηκε στους 300 K και η πίεση στο 1 atm. ~~Μετρώντας~~ Μετρήθηκαν οι τιμές της ηλεκτροστατικής δύναμης (ως πλάτος του σήματος) σε διάφορες κορυφές αλλά και οι διαφορές μεταξύ των σημείων (για ακόμη πιο αποδοτική αναπαράσταση). Συγκεκριμένα, μεταξύ δύο σημείων μετρήθηκε η διασπορά, με ελάχιστα διαφορετική ~~την~~ εξίσωση ~~του~~ για το R(τ) που αναλύθηκε παραπάνω ~~είδαμε~~:
Η μέση τιμή του κανονικοποιημένου όγκου εδώ είναι 1/L*L*L.
Έπειτα χωρίζουμε σε κυματοσυναρτήσεις μέσω Fourier
Με επιλογή ιδανικού κ=2π/ν που δεν μπορεί να επιτευχθεί.
Μετά από απλοποιήσεις καταλήγουμε στην:
που υπολογίζει διαφορές πλάτους μεταξύ ενός σήματος και δεν ενδιαφέρεται για τη φάση του. Μετά διαλέγουμε για την αναπαράσταση ένα μέγεθος να μπορεί να είναι διπλάσιο από την μικρότερη συχνότητα των κυματομορφών, όπως προτείνει και ο Nyquist. Για τις αποκλίσεις και τα λάθη χρησιμοποιήθηκε Gaussian ~~N(0,1)~~ ομοιόμορφη κατανομή. Παίρνοντας το μέσο όρο των σημείων 2-κορυφών, εξασφαλίζουμε αναπαράσταση σε 1 διάσταση, μέσω και των κυματομορφών. Παρατηρήθηκε ότι τα power spectra(φάσματα ενέργειας?) όλων των μοριακών επιφανειών έχουν συγκρίσιμα εύρη. Για ευκολότερη σύγκριση των αποτελεσμάτων, διαιρέθηκαν τα power spectra με τον μέσο όρο των white noise power spectra και εκτιμήθηκε ο θόρυβος με βάση το k. Σε όλα τα αντίστοιχα γραφήματα που προέκυψαν, αναπαραστάθηκε γραφικά και το power spectrum της Mouse-κινάσης(ένα σχεδόν επίπεδο φάσμα με ανομοιόμορφες κορυφώσεις(irregular peaks). Δεν θα σταθούμε περισσότερο στα συγκεκριμένα νούμερα, καθώς η εργασία προτείνει αλγόριθμο κατά το μεγαλύτερο μέρος αλλά θα δείξουμε ένα διάγραμμα που δείχνει κάποια επιτυχία στα αποτελέσματα.

Αποτελέσματα:
Οι πρωτεΐνες από την πρώτη οικογένεια έχουν όλες παρόμοια δυναμική ενέργεια, κάποιες περισσότερο και κάποιες λιγότερο. Για παράδειγμα οι πρωτεΐνες ~~πχ της πρώτης οικογένειας~~ του Κίτρινου πυρετού και του Δάγκειου πυρετού είχαν πολλές ομοιότητες. Παρόλα αυτά μπορούσαν να διαχωριστούν μεταξύ τους. Σε σύγκριση με τις HCV ελικάσες, το power spectrum της HCV_helicaseStrB παρουσιάζει διαφοροποιήσεις ως προς τις θέσεις των κορυφών. Παρόμοια αποτελέσματα παρατηρήθηκαν για όλες τις οικογένειες, με μικρές αποκλίσεις σε αριθμούς νανοκλίμακας. Όλες οι πολυμεράσεις είχαν το ίδιο μοτίβο μεταξύ τους, όμοια και οι μεθυλτρανσφεράσεις και οι γλυκοπρωτεΐνες.

Όπως βλέπουμε και από τα γραφήματα παρακάτω, μπορούμε με γυμνό μάτι να διακρίνουμε ένα μοτίβο στις ελικοειδείς πρωτεΐνες και επίσης μπορούμε να ξεχωρίσουμε ανάλογα την ένταση της δυναμικής τους ενέργειας σε ποια οικογένεια ανήκουν. Μπορούμε επίσης να διακρίνουμε ένα διαφορετικό μοτίβο στις πολυμερείς πρωτεΐνες αλλά να καταλάβουμε αν είναι πολυμερής ή όχι.

Συμπεράσματα:
Αν μελετηθούν σε βάθος τα παραπάνω μοντέλα (πχ μέσω τεχνητής νοημοσύνης) και βρεθούν τα λεπτά όρια ώστε να ξεχωρίζουμε, μέσω του PSD, την οικογένεια και την λειτουργία των μορίων, σε συνδυασμό με την αναπαράσταση σε μονοδιάστατα διανύσματα που επιτρέπει ο Fourier, η αναζήτηση και σύγκριση μορίων στη βιοπληροφορική θα αλλάξουν προς το καλύτερο επαναστατικά. https://peerj.com/articles/185/#