Protein signatures using electrostatic molecular surfaces in harmonic space.

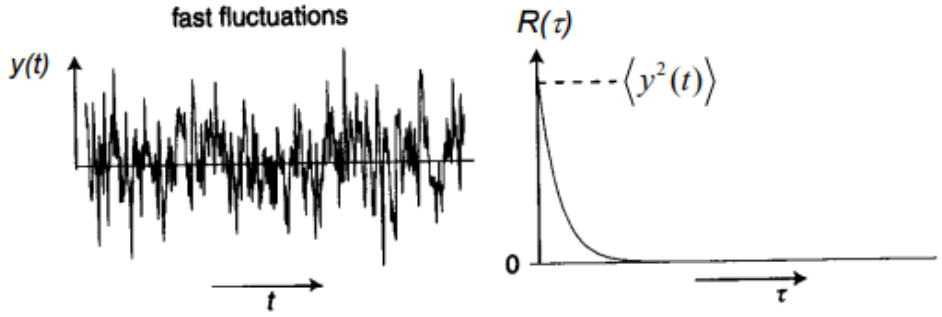
Το κείμενο καταλήγει σε έναν αλγόριθμο που καταφέρνει να αποθηκεύει και να αναλύει τρισδιάστατα μοντέλα πρωτεϊνών, χρησιμοποιώντας ελάχιστη υπολογιστική ισχύ (για τι συγκρίσεις) και λιγότερη μνήμη. Βασίζεται πάνω στα εργαλεία-θεωρήματα της διακύμανσης, της συνάρτησης αυτοσυσχέτισης, του μετασχηματισμού Fourier, της density functional theory (DTF), της στατιστικής ανάλυσης, του Molecular electrostatic Potential (MESP), και του Power spectrum density (PSD). Αφού τελειώσει με την επεξήγηση και την κατασκευή του μοντέλου, η εργασία μας μιλάει για τα θετικά αποτελέσματά του πάνω σε πρωτεϊνες και έντζυμα των ιών Υπατίτηδα C , Δάγκειος πυρετός, Κίτρινου πυρετού , ιογενής διάρροια βοοειδών, Πυρετού του δυτικού Νείλου. Πιο συγκεκριμένα βρέθηκε ότι το PSD ορίζει τα διαφορετικά μόρια με μοναδικό τρόπο και έτσι μπορεί να γίνει η αναγνώρισή του πολύ πιο αποδοτική. Η μόνη δυσκολία έγκειται στον σωστό υπολογισμό της μεθόδου σε τόσο μικρή κλίμακα.

Αρχικά ας αναλύσουμε κάποια μαθηματικά εργαλεία που θα χρησιμεύσουν. 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. [1]

$$Var(X) = E\left[(X-\mu)^2\right]$$ με Xi=δείγματα , μ=στατιστική μέση τιμή

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. [2]

a. Αν η διακύμανση αλλάζει γρήγορα, η R(τ) συγκλίνει άμεσα.



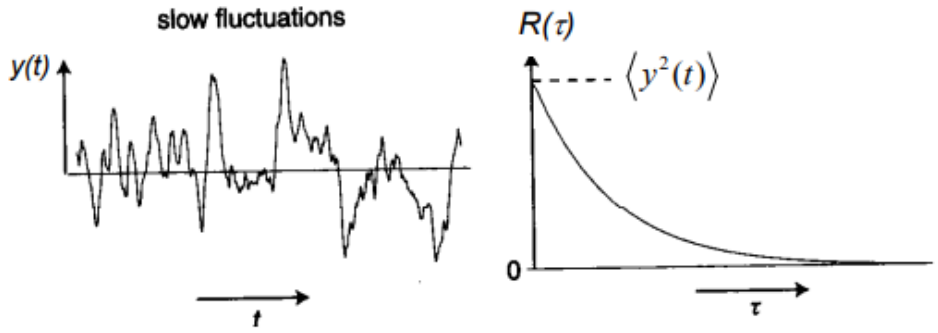
1

https://eclass.upatras.gr/modules/document/file.php/CEID1081/lecture5.pdf

2

https://physics.mcmaster.ca/phys4d06/LectureSlides-Ch9_Autocorr-Power-Noise.pdf

b. Αν η διακύμανση αλλάζει αργά, η R(τ) συγκλίνει αργά.



Η εξίσωση της αυτοσυσχέτισης είναι:

$$R(\tau) = \langle y(t)y(t+\tau) \rangle = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} y(t)y(t+\tau)dt \neq 0$$

.

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.

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 εως άπειρα απλούστερα σήματα διάφορων συχνοτήτων. Λόγω της φύσης της R(τ), στη μέτρηση PSD περιέχεται στατιστικά στατιστικά η τιμή του σήματος σε όλες τις κυματομορφές (το άθροισμά τους), χωρίς να αποθηεύεται πουθενά η φάση. Αυτό είναι όμως ότι 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 χαρτογράφιση υψομέτρων h(x,y) σε οριζόντια θέση x,y με μηδενική μέση τιμή, οι 3 διαστάσεις μπορούν να υπολογιστούν με την μέθοδο της real-space topology. Η μέση τετραγωνική ρίζα δίνεται ως:

$$h_{\mathrm{rms}}^2 = \langle h^2 \rangle$$

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 $h_{\mathrm{rms}}'^2 = \langle |\nabla h|^2 \rangle$ όπου το ∇ υποδεικνύει τη μερική παράγωγο του h. Η καμπυλότητα δίνεται από το $h_{\mathrm{rms}}''^2 = \frac{1}{2}\langle (\nabla^2 h)^2 \rangle$. 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 h_{rms}, h'_{rms}, and h''_{rms} είναι μοναδικές για κάθε επιφάνεια αν μετρηθούν με αρκετή ακρίβεια, και αυτό είναι μια πολύ χρήσιμη ιδιότητα για τον προσδιορισμό μιας πρωτεΐνης. 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. Όμως επειδή οι συγγραφείς δεν αναφέρθηκαν πολύ στις τεχνικές ακρίβειας μέτρησης, θα αναφέρουμε κάποια πράγματα απλά ονομαστικά.

Πρόκληση Α: Ποικιλίες στον ορισμό του PSD μπορεί να περιλαμβάνουν διαφορετικές μονάδες μέτρησης όπως m^2 to m^3 to m^4 to "arbitrary units".

Στρατηγική Α: Χρησιμοποιούμε μια προτεινόμενη μέθοδο ανάλογα αν το σχήμα μας είναι ισοτροπικό ή μη. (Ισοτροπικό : Οι ιδιότητες διαφέρουν ανάλογα την κατεύθυνση των μετρήσεων).

Πρόκληση Β: Όταν χωρίζουμε σε κυματομορφές την 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 συνδυάζουμε πολλές μετρήσεις από διαφορετικές διαστάσεις και από διαφορετικές τεχνικές μέτρησης. Με μια λέξη, επαλήθευση.

Πρόκληση Γ: Μέτρηση της τοπογραφίας σε πολύ μικρές κλίμακες.

Στρατηγική Γ: Πρέπει να καθορίσουμε τα όρια της ακτίνας και να χρησιμοποιήσουμε απόλυτη τιμή, σε συνδυασμό με καθορισμό της μέγιστης συχνότητας αναπαράστασης.

Η εργασία που πλέον έχουμε τις κατάλληλες μαθηματικές γνώσεις να κατανοήσουμε καλύτερα διαισθητικά βασίζεται και σε μια ακόμη ανακάλυψη. Στο 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 αναλύσουμε. Όσον αφορά το MESP, πήραμε πληροφορίες από το abstract διότι το κείμενο ήταν επί πληρωμή. [3]

Στο κυρίως θέμα

Στο βασικό κείμενο που αναλύουμε, συζητάμε αρχικά την μεγάλη ανάγκη για αποδοτική διαχείριση γενετικών πληροφοριών. Δεν μένουμε εκεί όμως. Το κείμενο προτείνει οτι οι κλασσικοί χάρτες πρωτεϊνών και αμινοξέων μπορούν να προσδιορίσουν την οικογένεια των στοιχείων αυτών. Έχουν όμως ένα μειονέκτημα, δεν μπορούν να προσδιορίσουν τη λειτουργία τους. Οι λειτουργίες των μορίων σχετίζονται πολύ με την δομή τους (MESP) και το σχήμα τους. 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 που μελετούν επ ακριβείας τη δομή των πρωτεϊνων, αλλά χρειάζονται πολύ χρόνο και πόρους και δεν είναι αποδοτικές. Με το PSD θα αναλύσουμε λιγότερο ακριβείς μετρήσεις. Θα χρησιμοποιήσουμε το μετασχηματισμό Fourier ο οποίος ανεξαρτήτως κλίμακας αναπαριστά την ίδια πληροφορία. Τέλος, 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

Χρησιμοποιήθηκαν 4 διαφορετικές οικογένειες πρωτεϊνων 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. Αυτές οι πρωτεϊνες αποτελούν πολύ συχνά συστατικά πολλών ασθενειών όπως οι Υπατίτηδα C , Δάγκειος πυρετός, Κίτρινος πυρετός , ιογενής διάρροια βοοειδών, Πυρετός του δυτικού

[3] https://pubs.rsc.org/en/content/articlelanding/2022/cp/d2cp03244a

Νείλου. Επίσης χρησιμοποιήθηκε πρωτεΐνη από κινάση ποντικιού λόγω της πολύ διαφορετικής δομής της με τις υπόλοιπες.

Οι πρωτεΐνες της 1ης οικογένειας, ξεδιπλώνουν την έλικα DNA ή RNA κατά τον πολλαπλάσιασμό.

Οι πρωτεΐνες της 2ης οικογένειας, είναι ένζυμα που αντιγράφουν γεννετικό υλικό.

Οι πρωτεΐνες της 3ης οικογένειας, είναι ένζυμα που μεταφέρουν methyl από δωρητές σε αποδοχείς.

Οι πρωτεΐνες της 4ης οικογένειας, χρησιμοποιούνται από τους ιούς για μοριακή αναγνώριση.

Το μεγαλύτερο μέρος των σχημάτων των πρωτεϊνών ανακτήθηκε από την βάση RCSB όπου σχηματίστηκε με κρυσταλλογραφία με ακτίνες Χ.

Από την 1η οικογένεια επιλέχθηκαν οι 1A1V και 80HM, πρωτεΐνες του ιού της Υπατίτιδας C . Οι 1YMF, 1YKS και 2V80, πρωτεΐνες του ιού του Κίτρινου πυρετού . Οι 2JLU, 2BHR, 2BMF και 2JLQ, πρωτεΐνες του ιού του Δάγκειου πυρετού . Από την 2η οικογένεια επιλέχθηκαν οι 2CJQ, 2HCS και SHCN, πρωτεΐνες του ιού του Δυτικού Νείλου. Από την 3η οικογένεια επιλέχθηκαν οι 3EVA, 3EVB, 3EVC, 3EVE και 3EVF, πρωτεΐνες του ιού του κίτρινου πυρετού. Από την 4η οικογένεια επιλέχθηκαν οι 1NB7, 4DVN, 4DW4 και 4DW3, πρωτεΐνες του ιού της ιογενείς διάρροιας βοοειδών.

Μοριακές επιφάνειες των επιλεγέντων πρωτεϊνών

Οι επιφάνειες των μορίων ακολουθούν τη μη γραμμική εξίσωση των Poisson-Boltzmann αρκετά ικανοποιητικά. Για τον υπολογισμό της δυναμικής ενέργειας χρησιμοποιήθηκε το 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 τις τιμές της ηλεκτροστατικής δύναμης (ως πλάτος του σήματος) σε διάφορες κορυφές αλλά και τις διαφορές τους μεταξύ των σημείων(για ακόμη πιο αποδοτική αναπαράσταση). Συγκεκριμένα μεταξύ δύο σημείων μετρήθηκε η διασπορά, με ελάχιστα διαφορετική την εξίσωση του R(τ) που είδαμε:

$$\xi(\boldsymbol{r}) \equiv \langle F^*(\boldsymbol{x})F(\boldsymbol{x}+\boldsymbol{r})\rangle = \frac{1}{L^3}\int d^3x\, F^*(\boldsymbol{x})F(\boldsymbol{x}+\boldsymbol{r}).$$

Η μέση τιμή του κανονικοποιημένου όγκου εδώ είναι 1/L*L*L.

Έπειτα χωρίζουμε σε κυματοσυναρτήσεις μέσω Fourier

$$F(\boldsymbol{x}) = \sum_{\boldsymbol{k}} F_{\boldsymbol{k}} \exp[-i\boldsymbol{k}\cdot\boldsymbol{x}]$$

Με επιλογή διακριτού κ=2π/ν .

Τέλος καταλήγουμε μετά από απλοποιήσεις στην

$$\xi(\boldsymbol{r}) = \left(\frac{L}{2\pi}\right)^3 \int d^3k\, |F_{\boldsymbol{k}}|^2 \exp[-i\boldsymbol{k}\cdot\boldsymbol{r}].$$

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 διπλάσσιο από την μικρότερη συχνότητα των κυματομορφών, όπως προτείνει και ο Niquist. Για τις αποκλείσεις και τα λάθη χρησιμοποιήθηκε Gaussian κατανομή.

Αποτελέσματα

Οι πρωτεϊνες από την 1η οικογένεια έχουν όλες παρόμοια δυναμική ενέργεια. Κάποιες περισσότερο και κάποιες λιγότερο. Οι πρωτεϊνες πχ της πρώτης οικογένειας του Κίτρινου πυρετού και του Δάγκειου είχαν πολλές ομοιότητες. Παρόλα αυτά μπορούσαν να διαχωριστούν μεταξύ τους. Παρόμοια αποτελέσματα παρατηρήθηκαν για όλες τις οικογένειες με μικρές αποκλίσεις σε αριθμούς νανοκλίμακας.

Συμπέρασμα

Αν μελετηθούν σε βάθος τα παραπάνω μοντέλα (πχ μέσω τεχνητής νοημοσύνης) και βρεθούν τα λεπά όρια ώστε να ξεχωρίζουμε μέσω του PSD, την οικογένεια και την λειτουργία των μορίων, σε συνδυασμό με την αναπαράσταση σε 1διάστατα διανύσματα που επιτρέπει ο Fourier, η αναζήτηση και σύγκριση μορίων στη βιοπληροφορική θα αλλάξουν προς το καλύτερο επαναστατικά.