



Πολυτεχνική Σχολή
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

ΕΙΣΑΓΩΓΗ ΣΤΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ

Δεύτερο Σύνολο Ασκήσεων 2023-24:

Αυγέρης Νικόλαος
Α.Μ. 1067508

Τζερμιά Ασπασία
Α.Μ. 1067455

Πάτρα, 2024

Εισαγωγή-Γενικές έννοιες:

Για την μελέτη των δοσμένων papers χρησίμευσαν κάποιες γενικές έννοιες βιολογίας που είτε χρειάστηκαν άμεσα για την κατανόηση των κειμένων είτε μπορούν να παρομοιαστούν με άλλες έννοιες που εμφανίζονται σε αυτά:

- Το DNA και το RNA, είναι μακρομόρια, αποτελούμενα από νουκλεοτίδια. Κάθε νουκλεοτίδιο του DNA αποτελείται από μια πεντόζη, τη δεοξυριβόζη, ενωμένη με μια φωσφορική ομάδα και μια αζωτούχο βάση. Στα νουκλεοτίδια DNA η αζωτούχος βάση μπορεί να είναι αδενίνη (A), γουανίνη (G), κυτοσίνη (C) ή θυμίνη (T). Σε κάθε νουκλεοτίδιο η αζωτούχος βάση συνδέεται με τον 1ο άνθρακα δεοξυριβόζης και η φωσφορική ομάδα με τον 5ο.
- Μια πολυνουκλεοτιδική αλυσίδα σχηματίζεται από την ένωση πολλών νουκλεοτιδίων με ομοιοπολικό δεσμό, που ονομάζεται φωσφοδιεστερικός. Η πολυνουκλεοτιδική αλυσίδα που δημιουργείται έχει έναν σκελετό, που αποτελείται από επανάληψη των μορίων φωσφορική ομάδα-πεντόζη-φωσφορική ομάδα-πεντόζη.
- Ανεξάρτητα από τον αριθμό των νουκλεοτιδίων από τα οποία αποτελείται η πολυνουκλεοτιδική αλυσίδα, το πρώτο της νουκλεοτίδιο έχει πάντα μια ελεύθερη φωσφορική ομάδα συνδεδεμένη στον 5ο άνθρακα της πεντόζης του και το τελευταίο νουκλεοτίδιο της, έχει ελεύθερο το υδροξύλιο του 3ου άνθρακα της πεντόζης του. Έτσι καθορίζεται ο προσανατολισμός της πολυνουκλεοτιδικής αλυσίδας 5->3.
- Το DNA αποτελείται από 2 πολυνουκλεοτιδικές αλυσίδες που σχηματίζουν στο χώρο μια δεξιόστροφη διπλή έλικα. Η διπλή έλικα έχει έναν σταθερό σκελετό, από επαναλαμβανόμενα μόρια φωσφορικής ομάδας-δεοξυριβόζης ενωμένων με φωσφοδιεστερικό δεσμό. Ο σκελετός αυτός είναι υδρόφιλος και βρίσκεται στο εξωτερικό του μορίου. Προς το εσωτερικό του σκελετού βρίσκονται οι αζωτούχες βάσεις που είναι υδρόφοβες.
- Οι αζωτούχες βάσεις της μιας αλυσίδας συνδέονται με δεσμούς υδρογόνου με τις αζωτούχες βάσεις της απέναντι αλυσίδας με βάση τον κανόνα της συμπληρωματικότητας. Η αδενίνη συνδέεται μόνο με θυμίνη και αντίστροφα, ενώ η κυτοσίνη μόνο με γουανίνη και αντίστροφα. Οι δεσμοί υδρογόνου που αναπτύσσονται μεταξύ των βάσεων σταθεροποιούν την δευτεροταγή δομή του μορίου.
- Οι 2 αλυσίδες ενός μορίου DNA είναι συμπληρωματικές, και αυτό υποδηλώνει ότι η αλληλουχία της μια καθορίζει την αλληλουχία της άλλης. Για την περιγραφή του μήκους ή της αλληλουχίας ενός νουκλεϊκού οξέος χρησιμοποιείται ο όρος αλληλουχία βάσεων. Στην πραγματικότητα εννοούμε την ακολουθία των νουκλεοτιδίων του νουκλεϊκού οξέος.
- Η συμπληρωματικότητα έχει τεράστια σημασία για τον αυτοδιπλασιασμό του DNA, μια ιδιότητα που το καθιστά το καταλληλότερο μόριο για τη διατήρηση και μεταβίβαση της γενετικής πληροφορίας. Το 1953 υπήρξε η πρώτη υπόθεση ότι το DNA αυτοδιπλασιάζεται. Κάθε αλυσίδα DNA μπορεί να χρησιμεύσει ως καλούπι για τη σύνθεση μιας συμπληρωματικής αλυσίδας, ώστε τελικά να σχηματίζονται 2 δίκλωνα μόρια DNA πανομιότυπα με το αρχικό μόριο.
- Μια λειτουργία του DNA είναι η έκφραση των γενετικών πληροφοριών, που επιτυγχάνεται με τον έλεγχο της σύνθεσης των πρωτεϊνών.
- Τα κύτταρα διαθέτουν εξειδικευμένα ένζυμα και άλλες πρωτεΐνες που δρουν

ταυτόχρονα και καταλύουν τις χημικές αντιδράσεις της αντιγραφής DNA με ταχύτητα και ακρίβεια.

- Η αντιγραφή ξεκινά από τις θέσεις έναρξης αντιγραφής. Στα ευκαρυωτικά κύτταρα το DNA κάθε χρωμοσώματος είναι ένα μακρύ γραμμικό μόριο, με πολυάριθμες θέσεις έναρξης αντιγραφής. Έτσι η αντιγραφή γίνεται ταυτόχρονα σε εκατοντάδες σημεία σε όλο το μήκος και τα τμήματα που δημιουργούνται ενώνονται. Για να αρχίσει η αντιγραφή, είναι απαραίτητο να ξετυλιχθούν στις θέσεις εκκίνησης οι δύο αλυσίδες. Αυτό επιτυγχάνεται με τη βοήθεια ειδικών ενζύμων που σπάζουν τους δεσμούς υδρογόνου μεταξύ των αλυσίδων. Τα ένζυμα αυτά ονομάζονται DNA ελικάσες. Τα κύρια ένζυμα που συμμετέχουν στην αντιγραφή ονομάζονται DNA πολυμεράσες. Επειδή αυτά δεν μπορούν μόνοι τους να αρχίσουν την αντιγραφή, το κύτταρο έχει ένα ειδικό σύμπλοκο, αποτελούμενο από πολλά ένζυμα, το πριμόσωμα, το οποίο συνθέτει στις θέσεις έναρξης αντιγραφής μικρά τμήματα RNA, συμπληρωματικά προς τις μητρικές αλυσίδες. Οι πολυμεράσες επιμηκώνουν τα αρχικά τμήματα, τοποθετώντας συμπληρωματικά δεοξυριβονουκλεοτίδια απέναντι από τις μητρικές αλυσίδες. Επίσης οι πολυμεράσες επιδιορθώνουν λάθη κατά το στάδιο αντιγραφής.

Η μεταγραφή είναι η διαδικασία δημιουργίας του RNA με δοσμένη μια αλυσίδα DNA, που χρησιμοποιείται ως πρότυπο και με βάση αυτήν παράγεται η συμπληρωματική αλυσίδα, μόνο που η αδερίνη συνδέεται με την βάση ουρακίλη(U) αντί για θυμίνη.

http://ebooks.edu.gr/ebooks/v/html/8547/2726/Biologia-T2_G-Lykeiou-ThSp-SpYg_html-empl/index2_1.html

- Τα ένζυμα είναι πρωτεΐνες και οι πρωτεΐνες είναι μόρια αποτελούμενα από αλληλουχίες αμινοξέων. Οι αλληλουχίες νουκλεϊκών οξέων καθορίζουν τη σειρά των αμινοξέων στις πρωτεΐνες μέσω της μεταγραφής και της μετάφρασης. Οι πρωτεΐνες είναι υπεύθυνες για ποικίλες λειτουργίες στον οργανισμό, όπως η δομή και η λειτουργία των κυττάρων, η ανοσολογική αντίδραση, η μεταφορά οξυγόνου και η κατασκευή των ιστών.
- Η δευτεροταγής δομή των πρωτεϊνών αναφέρεται στην τρισδιάστατη μορφή των τοπικών τμημάτων της πρωτεΐνης. Η δευτεροταγής δομή ορίζεται επίσημα από το σχήμα των δεσμών υδρογόνου μεταξύ των ατόμων υδρογόνου της αμίνης και του καρβοξυλικού οξυγόνου στο πεπτίδιο σκελετού. Άλλοι τύποι βιοπολυμερών, όπως τα νουκλεϊκά οξέα, διαθέτουν επίσης χαρακτηριστικές δευτεροταγείς δομές. Συνολικά, η δευτεροταγής δομή παρέχει σημαντικές πληροφορίες για την τρισδιάστατη διάταξη των πρωτεϊνών και την εκτέλεση των βιολογικών τους λειτουργιών.
- Τα πεπτίδια είναι δομικά στοιχεία των πρωτεϊνών, ενζύμων, κυττάρων και ιστών του σώματος. Συνολικά, τα πεπτίδια αποτελούνται από δύο ή περισσότερα αμινοξέα που συνδέονται με πεπτιδικό δεσμό. Ο αριθμός των αμινοξέων σε ένα πεπτίδιο υποδηλώνεται με αριθμητικό πρόθεμα, όπως διπεπτίδιο (2 αμινοξέα), οκταπεπτίδιο (με 8 αμινοξέα) ή ακόμα ολιγοπεπτίδιο ή πολυπεπτίδιο (περισσότερα από 50). Η διάκριση μεταξύ πολυπεπτιδίου και πρωτεΐνης είναι μάλλον ασαφής, χωρίς καμία πρακτική σημασία.

- Στην εξατομικευμένη ιατρική οι ασθενείς χωρίζονται σε κατηγορίες με βάση ένα εξατομικευμένο προφίλ που δημιουργείται με βάση το γονιδίωμά τους. Συχνά για τη δημιουργία του προφίλ χρησιμοποιούνται μέθοδοι ανάλυσης του γονιδιώματος σε επίπεδο μορίων, πρωτεϊνών, ενζύμων ή αμινοξέων. Οι μέθοδοι πρόληψης και θεραπείας στην εξατομικευμένη ιατρική δεν είναι γενικευμένες όπως στην παραδοσιακή ιατρική, αλλά σχεδιάζονται με υπόβαθρο το γονιδίωμα κάθε ασθενή και την αναμενόμενη απόκρισή του σε αυτές.
(πηγές: <https://www.genome.gov/genetics-glossary/Personalized-Medicine>,
https://en.wikipedia.org/wiki/Personalized_medicine)
- Η συσταδοποίηση είναι η διαδικασία ανακάλυψης ομάδων και δομών στα δεδομένα που είναι “παρόμοια” κατά κάποιο τρόπο, χωρίς να χρησιμοποιούνται γνωστές δομές στα δεδομένα. Στην ουσία, η συσταδοποίηση επιτρέπει την ομαδοποίηση αντικειμένων σε λογικές ομάδες, με τα αντικείμενα που ανήκουν στην ίδια ομάδα να είναι ομοιόμορφα ή σχετικά μεταξύ τους. Η διαδικασία της συσταδοποίησης περιλαμβάνει τρία βασικά βήματα:
 1. Επιλογή χαρακτηριστικών γνωρισμάτων.
 2. Ομαδοποίηση με χρήση αλγορίθμων συσταδοποίησης.
 3. Επικύρωση αποτελεσμάτων.
 (πηγή: Βικιπαίδεια)
- Το Variable (V) domain είναι ένα είδος δομικής μονάδας που χαρακτηρίζει μια αλυσίδα πρωτεϊνών που ανήκει στην υπεροικογένεια των ανοσοσφαιρίνων (IgSF). Το V domain περιλαμβάνει το V-DOMAIN των ανοσοσφαιρίνων (IG) ή αντισωμάτων.
Η δομή του V domain είναι πολύ παρόμοια ανεξάρτητα από τον τρόπο που παράγεται.
(πηγή: https://link.springer.com/referenceworkentry/10.1007/978-1-4419-9863-7_266)
- Στην ανοσολογία, αντιγόνο είναι μια μοριακή δομή που μπορεί να εισέλθει στο σώμα μέσω ενός παθογόνου μικροοργανισμού και προκαλεί υπό φυσιολογικές συνθήκες την αντίδραση του σώματος που είναι γνωστή ως ανοσολογική απόκριση. Τα αντισώματα είναι μοριακές δομές που παράγονται από το ανοσοποιητικό σύστημα. Κάθε αντίσωμα έχει την ικανότητα να αναγνωρίζει και να προσδένεται με συγκεκριμένο αντιγόνο, δεσμεύοντάς το και αναστέλλοντας έτσι τη δράση του.
(πηγή: <https://el.wikipedia.org/wiki/%CE%91%CE%BD%CF%84%CE%B9%CE%B3%CF%8C%CE%BD%CE%BF>)
- Το heavy chain και το light chain είναι περιοχές της δομής ενός αντισώματος που σχηματίζουν ζεύγη. Το είδος των αλυσίδων και οι συνδέσεις μεταξύ τους καθορίζουν τη λειτουργία του αντισώματος και την κλάση όπου ανήκει. Η λειτουργικότητα του μορίου καθορίζεται κατά κύριο λόγο από το heavy chain.
(πηγή: <https://www.ncbi.nlm.nih.gov/books/NBK27144/>)
- Η πρωτοταγής δομή μιας πρωτεΐνης είναι η αλληλουχία των αμινοξέων της. Η δευτεροταγής δομή μιας πρωτεΐνης είναι η τρισδιάστατη μορφή τοπικών τμημάτων της και σχετίζεται με

τους δεσμούς υδρογόνου στο σκελετό της. Γενικά η τρισδιάστατη δομή των πρωτεϊνών είναι εξαιρετικά μεταβλητή, σε αντίθεση με τη διπλή έλικα του DNA, που είναι συμμετρική και έχει σταθερή δομή. Υπάρχουν οκτώ τύποι δευτεροταγούς δομής στο DSSP (είδος λεξικού για αντιστοίχιση δευτεροταγών δομών με κωδικούς). Με τον όρο τριτοταγή δομή περιγράφεται το τρισδιάστατο σχήμα μιας πρωτεΐνης στο χώρο. Μια τριτοταγής δομή περιλαμβάνει περισσότερες από μία δευτεροταγείς και καθορίζεται από τις αλληλεπιδράσεις μεταξύ των αλυσίδων των αμινοξέων που την αποτελούν.

(πηγές:

https://el.wikipedia.org/wiki/%CE%A0%CF%81%CF%89%CF%84%CE%BF%CF%84%CE%B1%CE%B3%CE%AE%CF%82_%CE%B4%CE%BF%CE%BC%CE%AE_%CF%80%CF%81%CF%89%CF%84%CE%B5%CE%90%CE%BD%CE%B7%CF%82_,

https://el.wikipedia.org/wiki/%CE%94%CE%B5%CF%85%CF%84%CE%B5%CF%81%CE%BF%CF%84%CE%B1%CE%B3%CE%AE%CF%82_%CE%B4%CE%BF%CE%BC%CE%AE_%CF%80%CF%81%CF%89%CF%84%CE%B5%CE%90%CE%BD%CE%B7%CF%82_,

https://el.wikipedia.org/wiki/%CE%A4%CF%81%CE%B9%CF%84%CE%BF%CF%84%CE%B1%CE%B3%CE%AE%CF%82_%CE%B4%CE%BF%CE%BC%CE%AE_%CF%80%CF%81%CF%89%CF%84%CE%B5%CE%90%CE%BD%CE%B7%CF%82_,

Εισαγωγή στους Αλγορίθμους Βιοπληροφορικής, Neil C. Jones, Pavel Pevzner, Εκδόσεις Κλειδάριθμος, 2008)

- Φάσμα ενέργειας ενός σήματος είναι η ενέργειά σε κάθε συχνότητά του. Το φάσμα ενέργειας μιας μοριακής επιφάνειας σχετίζεται έμμεσα με τις θέσεις των ατόμων σε αυτή, όπως τις μελετά κάποιος στις τρεις διαστάσεις του χώρου.

(πηγές: https://simple.wikipedia.org/wiki/Power_spectrum_,
https://en.wikipedia.org/wiki/Potential_energy_surface)

- Φασματομετρία μάζας είναι μια τεχνική που χρησιμοποιείται για την μελέτη ιδιοτήτων των βιολογικών μακρομορίων με βάση τις αλληλεπιδράσεις των φορτίων ανάμεσα στα μόρια που τα αποτελούν (πχ: ιόντα).

(πηγή: https://en.wikipedia.org/wiki/Mass_spectrometry)

- Η μορφοποίηση DSSP (Define Secondary Structure of Proteins) είναι μέθοδος με την οποία προβλέπεται η δευτεροταγής δομή των πρωτεϊνών από γεωμετρικά και μοριακά χαρακτηριστικά της δομής τους και κάθε πρωτεΐνη κατηγοριοποιείται σε μία από 8 κατηγορίες.

(πηγή: <https://www.sciencedirect.com/topics/immunology-and-microbiology/protein-secondary-structure>)

- Folding πρωτεΐνης είναι η διαδικασία κατά την οποία ένα πολυπεπτίδιο «τυλίγεται» γύρω από τον εαυτό του για να σχηματιστεί η τρισδιάστατη δομή του στο χώρο. Το πολυπεπτίδιο από μόνο του δεν έχει τρισδιάστατη δομή(είναι απλώς μια αλληλουχία αμινοξέων). Η αλληλεπίδραση των αμινοξέων της μεταξύ τους με βάση την υδροφοβία και με ορισμένες χημικές αντιδράσεις είναι που της δίνει τελικά σχήμα.

(πηγή: https://comis.med.uvm.edu/vic/coursefiles/MD540/MD540-Protein_Organization_10400_574581210/Protein-org/Protein_Organization8.html)

- Οι αυτοοργανούμενοι χάρτες είναι μέθοδος αυτοματοποιημένης ανάλυσης κατά την οποία στοιχεία δεδομένων ταξινομούνται σε ομάδες-κλάσεις με βάση ένα πλέγμα και έναν αλγόριθμο που βασίζεται σε διανύσματα και ευρετικές μεθόδους. Οι κόμβοι του γράφου, στην περίπτωση που αναπαριστούν βιολογικές συμβολοσειρές, είναι γειτονικοί όταν οι αντίστοιχες συμβολοσειρές παρουσιάζουν μεγαλύτερη ομοιότητα μεταξύ τους, δηλαδή μικρότερο edit distance.

(πηγή:

<https://hasler.ece.gatech.edu/Courses/MachineLearning/FoundationalPapers/KohonenSOM2013.pdf>)

- Ένας βιολογικός δείκτης προκύπτει από μετρήσεις είτε στις φυσιολογικές λειτουργίες ενός οργανισμού ή από ανάλυση δείγματος(πχ: αίματος). Δείχνει την κατάσταση της υγείας σε μια δεδομένη στιγμή. Κάποιοι βιολογικοί δείκτες αποκαλύπτουν στοιχεία για τα γονιδιακά χαρακτηριστικά του οργανισμού.

(πηγή: <https://www.niehs.nih.gov/health/topics/science/biomarkers>)

- Genomic driver είναι μια μετάλλαξη του γονιδιώματος ενός ασθενή που οδηγεί στο σχηματισμό και την εξάπλωση καρκινικών ιστών.

(πηγή: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/driver-mutation>)

- Υπερμεταβλητή περιοχή ενός αντισώματος είναι περιοχή του heavy ή του light chain με την οποία ένα αντιγόνο έρχεται απευθείας σε επαφή. Λέγεται υπερμεταβλητή γιατί μεταβάλλεται συχνά και ώστε να επιτρέπει την αναγνώριση πολλών και διαφορετικών αντιγόνων από το αντίσωμα.

Complementarity-Determining Region είναι είδος υπερμεταβλητής περιοχής που καθορίζει πώς γίνεται η πρόσδεση αντιγόνου σε αντίσωμα και σχετίζεται με τα πεπτιδία(ενώσεις που αποτελούνται από αμινοξέα).

(πηγές:

https://link.springer.com/referenceworkentry/10.1007/3-540-29662-X_1303,

https://en.wikipedia.org/wiki/Hypervariable_region,

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2396520/>)

- Μια διατηρούμενη αλληλουχία είναι μια αλληλουχία νουκλεοτιδίων που έχει αλλάξει ελάχιστα ή καθόλου σε μια συγκεκριμένη εξελικτική χρονική περίοδο ή που συναντάται κοινή σε διαφορετικά είδη οργανισμών. Με βάση τέτοιες αλληλουχίες προκύπτουν συμπεράσματα για την εξελικτική πορεία των οργανισμών και τις διαφοροποιήσεις μεταξύ τους.

(πηγή:

<https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095633355>,

<https://profiles.umassmed.edu/display/115933>,
https://en.wikipedia.org/wiki/Conserved_sequence)

- Το molecular potential περιγράφεται από μια συνάρτηση που περιγράφει τους όρους με τους οποίους τα σωματίδια που αποτελούν ένα βιολογικό μόριο κινούνται και αλληλεπιδρούν σε μια προσομοίωση με χρήση υπολογιστή. Τέτοιου είδους προσομοίωση με molecular dynamics είναι συνηθισμένη στη μελέτη πρωτεϊνών γιατί επιτρέπει τη μελέτη της κίνησης των σωματιδίων με προσεγγιστικό τρόπο, κάτι που δεν είναι εύκολο να γίνει στο εργαστήριο.
(πηγή: διαφάνειες μαθήματος)

1ο Paper: Advanced Protein Alignments Based on Sequence, Structure and Hydropathy Profiles; The Paradigm of the Viral Polymerase Enzyme

Σύμφωνα με το paper αυτό, ένα από τα μειονεκτήματα που έχουν οι αλγόριθμοι αναζήτησης ομοιοτήτων ανάμεσα σε διαφορετικές πρωτεΐνες με βάση την ακολουθία των αμινοξέων σε αυτές, είναι ότι δεν λαμβάνουν υπ' όψη χαρακτηριστικά της πραγματικής φυσικής και χημικής δομής τους.

Για το λόγο αυτό έγινε ανάπτυξη ενός νέου εργαλείου, του PSSP (Protein Signature Structure Profile) tool. Σκοπός του είναι η ταχύτερη αναζήτηση ομοιοτήτων ανάμεσα σε πρωτεΐνες με βάση πρωτοταγή και δευτεροταγή δομικά χαρακτηριστικά τους.

Πολλές πρωτεΐνες, παρόλο που δομικά είναι ομόλογες, παρουσιάζουν χαμηλό ποσοστό ομοιότητας μεταξύ των αλληλουχιών των αμινοξέων από τα οποία αποτελούνται (μικρότερο από 10%). Ομόλογες (homologous) είναι δύο πρωτεΐνες που ανήκουν στην ίδια "οικογένεια", έχουν την ίδια λειτουργικότητα και δομικά είναι όμοιες κατά περισσότερο από 30%. Τέτοια παραδείγματα είναι τα ένζυμα των οικογενειών helicases (ελικάσες), proteases (πρωτεάσες) και polymerases (πολυμεράσες).

Το λογισμικό που προτείνεται είναι ανοικτού κώδικα και το εργαλείο PSSP βασίζεται σε αυτό. Αξιοποιεί την βάση δεδομένων RCSB Protein Data Base για να ανακτά με blast αναζήτηση αποθηκευμένες δευτεροταγείς πληροφορίες για τη δομή των πρωτεϊνών όταν υπάρχουν αποθηκευμένες στη βάση PDB. Η πρόβλεψη της τρισδιάστατης δομής πρωτεϊνών που έχουν άγνωστη δομή γίνεται με χρήση της δομής άλλων ομόλογων με αυτές πρωτεϊνών ως πρότυπα (templates) και ευθυγράμμιση των ακολουθιών τους. Όταν δεν υπάρχουν πληροφορίες στην PDB για κάποιο συγκεκριμένο μοντέλο πρωτεϊνών, το λογισμικό τις προβλέπει με μερική στοίχιση και αντιγραφή της γεωμετρίας της δομής των προτύπων και μετά τις επαληθεύει χρησιμοποιώντας και άλλα εργαλεία. Οι συγκρίσεις γίνονται έτσι με μεγαλύτερη ακρίβεια, ενώ γίνεται καλύτερη πρόβλεψη της δομής άγνωστων μορίων πρωτεϊνών, επειδή εξετάζονται τόσο η σειρά της αλληλουχίας αμινοξέων όσο και η τρισδιάστατη δομή των πρωτεϊνών που έχουν ήδη καταγραφεί.

Οι πληροφορίες που χρησιμοποιούνται για την μοντελοποίηση της δομής των πρωτεϊνών είναι δευτεροταγείς, αλλά εξαιρούνται οι πλευρικές αλυσίδες. Ένα μητρώο ομοιότητας (similarity matrix) που κατασκευάζεται με βάση τα σκορ ομοιότητας (similarity scores) και περιέχει τις δευτεροταγείς πληροφορίες των μορίων χρησιμοποιείται ώστε να μεταφράσει μια σειρά αμινοξέων σε σχηματική μορφή σύμφωνα με τις πρωτεΐνες-πρότυπα όταν είναι απαραίτητο. Αυτό το μητρώο κατασκευάζεται με βάση τις ιδιότητες υδροπάθειας (hydropathicity) των μορίων, που αφορά το folding πρωτεϊνών και συνδέεται άμεσα με τις φυσικοχημικές ιδιότητές τους και τη δομή τους. Η σύγκριση βασισμένη στην υδροπάθεια επιτρέπει να εντοπίζεται η ομοιότητα και σε περιπτώσεις μακρινής συγγένειας μεταξύ πρωτεϊνών όταν δεν παρατηρείται απλά με σύγκριση των αλληλουχιών των αμινοξέων τους.

Μια εικόνα για τη δευτεροταγή δομή των πρωτεϊνών μπορεί να δοθεί όταν αυτές βρίσκονται σε επαφή με το νερό. Η υδροπάθεια μελετάει τις ιδιότητες του μορίου που συμβαίνουν λόγω της υδροφοβίας ή υδροφιλίας των αμινοξέων του. Υδροφοβία(hydrophoby) είναι η τάση μη-πολικών ουσιών να απωθούνται όταν έρχονται σε επαφή με νερό. Βασίζεται στις χημικές ιδιότητες των πλευρικών αλυσίδων του μορίου(αφού αυτές έχουν την ελευθερία κίνησης) και βοηθάει να καθορίσουμε τον προσανατολισμό των πλευρικών αλυσίδων στις 3 διαστάσεις. Αν οι πλευρικές αλυσίδες είναι υδρόφοβες, μέσα στο νερό διπλώνουν προς τον πυρήνα του μορίου. Τα 20 γνωστά αμινοξέα έχουν κατηγοριοποιηθεί με βάση έναν δείκτη υδροφοβίας(amino acid hydrophathy index) σε τρεις κλάσεις: τα υδρόφοβα (I, V, F, C, M, A, W), τα ουδέτερα (G, T, S, Y, P, H) και τα υδρόφιλα (D, N, E, Q, K, R). Ο προσδιορισμός των κλίσεων των υδρόφοβων ή υδρόφυλων πλευρικών αλυσίδων γίνεται με πολλούς τρόπους. Ένας είναι να μετρηθούν μέσα στο νερό και μετά σε ένα μη αντιδραστικό περιβάλλον ως ισοτροπικές, αφαιρώντας την διαφορά ενέργειας σε κάθε κατάσταση. Ένας άλλος τρόπος βασίζεται στον υπολογισμό του μέσου όρου της θέσης των ατομικών συντεταγμένων 12 πρωτεϊνών. Ένας τρίτος τρόπος είναι ο συνδυασμός των παραπάνω.

Τα αμινοξέα με μεγάλες μη πολικές πλευρικές αλυσίδες τείνουν να είναι υδροφοβικά, ενώ τα πολύ πολικά τείνουν να είναι υδροφιλικά. Πολλά αμινοξέα έχουν και υδροφιλικά και υδροφοβικά τμήματα. Για τη μέτρησή τους χωρίζουμε τις περιόδους μέτρησης. Οι δομές των πρωτεϊνών σχετίζονται με πολλά πράγματα όπως τη δυναμική ενέργεια των μορίων, την εντροπία και τις ηλεκτροστατικές δυνάμεις. Όμως η υδροπάθεια είναι υπεύθυνη για το folding πρωτεϊνών και αποτελεί τον βασικότερο παράγοντα που καθορίζει ποιά θα είναι η τρισδιάστατη δομή μιας πρωτεΐνης.

Η περιοδικότητα πολικών και μη πολικών αμινοξέων, είναι καθοριστικός παράγοντας για τη δευτερογενή δομή των αυτο-συναρμολογούμενων ολιγομετρικών πεπτιδίων. Παρόλο που τα αμινοξέα μοιάζουν να διαφέρουν κάποιες φορές στη δομή τους, μέσα στο νερό, ο σχηματισμός υδρόφυλων δομών δείχνει πατέντες που μπορεί να είναι κοινές και αυτό έχει τεράστια σημασία. Αυτές οι πατέντες, που είναι κοινές ανάμεσα σε οικογένειες, μπορούν εύκολα να διακριθούν όταν βάλουμε ένα μόριο μέσα στο νερό και θα ήταν πολύ δύσκολο να παρατηρηθούν διαφορετικά. Αυτό έκανε δυνατή την ταυτοποίηση μέχρι και πιο μακρινών συγγενικών πρωτεϊνών. Ομόλογες πρωτεΐνες μιας οικογένειας με παρόμοιες δομές, έχει βρεθεί να εμφανίζουν πολλές ομοιότητες όσον αφορά την υδροπάθειά τους, παρόλο που είχαν μηδαμινές ομοιότητες στην αλληλουχία αμινοξέων.

Συνοπτικά, ο αλγόριθμος που χρησιμοποιεί το προτεινόμενο λογισμικό-εργαλείο για να ψάχνει ομόλογα μόρια πρωτεϊνών ακολουθεί τα εξής βήματα:

1. Αφού δεχθεί ως είσοδο από το χρήστη είτε μια ακολουθία αμινοξέων ή μια πρωτεϊνική ακολουθία που έχει προηγουμένως υποστεί DSSP-formatting, ψάχνει στη βιολογική βάση δεδομένων για παρόμοια μοντέλα μορίων.

- 1.1. Χρησιμοποιεί γρήγορες μεθόδους εξαντλητικής αναζήτησης. Ψάχνει στη βάση, μέσω μιας ευρετικής συνάρτησης, τμήματα ίδιων δομών πρωτεϊνών, με την είσοδο(άγνωστη δομή). Αν η ομοιότητα είναι μεγάλη, η άγνωστη δομή θα αντιγράψει στοιχεία από το σκελετό και τις πλευρικές αλυσίδες. Αν η ομοιότητα είναι μικρή, θα αντιγράψει μόνο το σκελετό. Σε μηδέν ομοιότητα δεν αντιγράφει τίποτα.

- 1.2. Υπολογίζει ένα σκόρ ομοιότητας ανάμεσα στην είσοδο και τα μοντέλα της βάσης δεδομένων. Κατασκευάζει το μητρώο ομοιότητας και διατάσσει τα πιθανά μοντέλα.

- 1.3. Χρησιμοποιεί μια μέθοδο για μια ακόμη καλύτερη ταξινόμηση.

- 1.4. Παράγει μια πολλαπλή αντιστοίχιση με τα καλύτερα μοντέλα.

2. Τέλος για να ελεγχθεί η ποιότητα της αντιγραφής, ελέγχεται αν τα τελικά στοιχεία έχουν ίδιες ιδιότητες υδροπάθειας.

Λίγα πράγματα για τη λειτουργία της διεπαφής λογισμικού:

Αρχικά, το λογισμικό δίνει επιλογές εισαγωγής τύπου-ερωτήματος. Ο χρήστης μπορεί να επιλέξει μια σειρά αμινοξέων ή DSSP-δευτεροταγών στοιχείων σειρών πρωτεϊνών.

Η αναζήτηση για όμοια μοντέλα γίνεται είτε με τις κλασσικές σειρές αμινοξέων είτε

με δευτερογενή στοιχεία με το STRAP μοντέλο. Το μοντέλο STRAP προβλέπει μια πρώτη εικόνα για την δομή. Έπειτα κάνει αναζήτηση για δευτερογενή ταιριάσματα με άλλα μόρια. Μετά γίνεται αναζήτηση στη βάση δεδομένων πρωτεϊνών. Η αναζήτηση ομοιότητας ακολουθεί δύο βήματα: Πρώτα γίνεται αναζήτηση blast, με την οποία παράγεται ένα προσωρινό αρχείο με όλες τις ακολουθίες που έχουν πάνω από 30% ομοιότητα με τη δοσμένη πρωτεϊνική ακολουθία. Στη συνέχεια με τη βοήθεια ενός μητρώου (hydropathicity matrix) οι είσοδοι κατατάσσονται με βάση το προφίλ υδροπάθειάς τους. Υπολογίζονται οι στοιχίσεις μεταξύ τους και τα ποσοστά ομοιότητας, τα οποία στη συνέχεια αναπαριστώνται γραφικά και προβάλλονται στο χρήστη σύμφωνα με το IMGT coloring scheme for hydropathy. Στα αποτελέσματα χρησιμοποιούνται χρώματα για να περιγράψουν κατά πόσο το μοντέλο είναι ακριβές. Τα πανομοιότυπα στοιχεία χρωματίζονται πράσινα, τα υδρόφοβα μπλε, τα ουδέτερα κόκκινα και τα υδρόφιλα κίτρινα. Οι δευτεροταγείς πληροφορίες κωδικοποιούνται με αλφάβητο 8 συμβόλων για τα αμινοξέα αντί για 20. Αυτό επιτυγχάνει μεγαλύτερη ταχύτητα μοντελοποίησης.

Για τον έλεγχο της αποτελεσματικότητας του εργαλείου επιλέχθηκε να γίνει μοντελοποίηση του ιού της πολυμεράσης με το πακέτο MOE για αναπαράσταση πρωτεϊνών 3-διαστάσεων. Οι υπολογισμοί μοριακών δυναμικών έγιναν με το GROMACS, λαμβάνοντας υπόψη την τοπολογία και σύστημα. Το εργαλείο αυτό περιέχει και βάση με τις τοπολογίες νουκλεοτιδίων και αμινοξέων. Κάποιες δυσκολίες εντοπίζονται όμως λόγω της πολυπλοκότητάς του. Ένας έλεγχος που έγινε με συναρτήσεις του MOE έδειξε ότι τα λάθη στην μοντελοποίηση ήταν ασήμαντα και ότι το μοντέλο που παράχθηκε ήταν ικανοποιητικά ακριβές. Η μοντελοποίηση της HCV πολυμεράσης δεν θα ήταν δυνατή με χρήση προγενέστερων τεχνικών μοντελοποίησης. Το παράδειγμα αυτό δείχνει τη χρησιμότητα του εργαλείου-πλατφόρμας PSSP.

Δυστυχώς όταν μπαίνουμε στη σελίδα του λογισμικού για να πειραματιστούμε με αυτό αφού πατάμε enter οδηγούμαστε στον σύνδεσμο: <http://ww25.bioinfoteam.com/?subid1=20240406-2046-3316-8900-afcc8ae4214a> που οδηγεί σε σελίδα που δεν φορτώνει σωστά.

Κατά την μελέτη του συγκεκριμένου paper σταθήκαμε περισσότερο στην περιγραφή και ανάλυση όρων. Μια ιδέα συνέχισης της εργασίας σε συνδυασμό με το προηγούμενο paper είναι η περαιτέρω ανάπτυξη του εργαλείου ώστε εκτός από τις συγκρίσεις υδροπαθειών να συγκρίνεται ταυτόχρονα και η ηλεκτροστατική δυναμική για κάθε ομοιότητα.

2^ο Paper: Protein signatures using electrostatic molecular surfaces in harmonic space

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 (DFT), της στατιστικής ανάλυσης, του Molecular Electrostatic Potential (MESP), και του Power Spectrum Density (PSD).

Μετά την κατασκευή και επεξήγηση της μεθόδου, η εργασία αναφέρει τα θετικά αποτελέσματά της εφαρμογής της πάνω σε πρωτεΐνες και ένζυμα ιών που σχετίζονται με διαδεδομένες ασθένειες όπως Ηπατίτιδα C, Δάγκειος πυρετός, Κίτρινος πυρετός, Ιογενής διάρροια των 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 εφαρμογή της μεθόδου σε τόσο μικρή κλίμακα.

Ένα από τα μαθηματικά εργαλεία που χρησιμοποιήθηκαν 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.

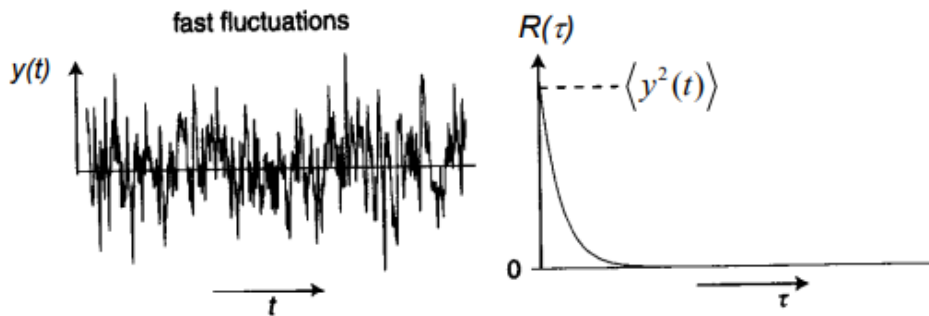
$$Var(X) = E[(X - \mu)^2]$$

, με X_i =δείγματα, μ =στατιστική μέση τιμή.

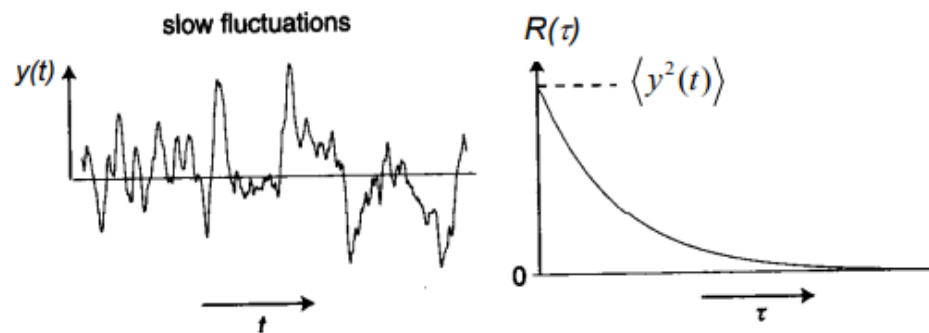
(πηγή: <https://eclass.upatras.gr/modules/document/file.php/CEID1081/lecture5.pdf>).

Η συνάρτηση αυτοσυσχέτισης ενός σήματος, εκφράζει την διακύμανση του σήματος την χρονική στιγμή t συγκριτικά με μια επόμενη χρονική στιγμή $t+\tau$. Είναι σημαντική διότι μας δείχνει την ταχύτητα μεταβολής της διακύμανσης.

Αν η διακύμανση αλλάζει γρήγορα, η $R(\tau)$ συγκλίνει άμεσα.



Αν η διακύμανση αλλάζει αργά, η $R(\tau)$ συγκλίνει αργά.



Η εξίσωση της αυτο-συσχέτισης είναι:

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

Όπως φαίνεται μετρήθηκε η μεταβολή της διακύμανσης χωρίς κάποια αφαίρεση (όπως στον παραπάνω τύπο του var). Αυτό διότι, αντί για αφαίρεση τέθηκαν τα άκρα του ολοκληρώματος σε T και $-T$. Ο όρος της μέσης τιμής επιτυγχάνεται διαιρώντας με $2T$, που στον διακριτό χώρο θα ήταν ο αριθμός των δειγμάτων της δειγματοληψίας(πηγή:

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

Ο όρος Power Spectrum Density δηλώνεται στην εργασία και ως power spectrum of surface. Το PSD ενός σήματος είναι ο μετασχηματισμός Fourier της συνάρτησης αυτοσυσχέτισης του σήματος. Μέσω του μετασχηματισμού Fourier, προκύπτει η αποσύνθεσή των σημάτων(των επιφανειών στη συγκεκριμένη εργασία) σε πολλά έως άπειρα απλούστερα σήματα διάφορων

συχνοτήτων. Λόγω της φύσης της $R(\tau)$, στη μέτρηση PSD περιέχεται στατιστικά η τιμή του σήματος σε όλες τις κυματομορφές (το άθροισμά τους), χωρίς να αποθηκεύεται πουθενά η φάση. Αυτό όμως είναι που χρειάζεται αργότερα στον ορισμό του 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 της επιφάνειας σε pixels. Στην περίπτωση που είναι τέλεια γνωστή η εξεταζόμενη επιφάνεια ως συνεχής 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:

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

$$h'_{rms} = \langle |\nabla h|^2 \rangle$$
, όπου το ∇ υποδεικνύει τη μερική παράγωγο του h . Η καμπυλότητα δίνεται από το

$$h''_{rms} = \frac{1}{2} \langle (\nabla^2 h)^2 \rangle$$
.

Από αυτές τις διαστάσεις συνήθως μόνο μία είναι κυρίαρχη. Έτσι για παράδειγμα για όταν θέλουμε να μετρήσουμε την τραχύτητα μεταξύ δύο επιφανειών αρκεί η τετραγωνική διαφορά υψών h_{rms} . 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 γίνουν συμβιβασμοί, να χρησιμοποιηθούν τεχνικές προσαρμοσμένες στις ιδιαιτερότητες κάθε συστήματος και να επαληθεύονται τα αποτελέσματα κάθε μοντέλου.

Πιο συγκεκριμένα υπάρχει αλγόριθμος για μείωση των 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 αξιοποιηθούν στην παρούσα εργασία.

Οι συγγραφείς δεν αναφέρθηκαν πολύ στις τεχνικές ακρίβειας μέτρησης, διότι αυτές μπορούν να αλλάξουν σε κάθε περίπτωση. Επικεντρώθηκαν περισσότερο στον αλγόριθμο υλοποίησης και ανέφεραν ονομαστικά-αριθμητικά τις-κατανομές και τα ανεκτά μεγέθη των σφαλμάτων της πειραματικής επαλήθευσής τους. Έτσι αναφέρονται κάποια πράγματα απλά ονομαστικά.

1. 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”. Στην εργασία γίνονται υπολογισμοί στις 3 διαστάσεις και πιο συγκεκριμένα στην σφαίρα του Fourier.
Στρατηγική Α: Χρήση μιας προτεινόμενης μεθόδου ανάλογα αν το σχήμα 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 η μόνη μέτρηση που καθορίζει την τιμή είναι η απόσταση μεταξύ σημείων.
2. Πρόκληση Β: Όταν χωρίζεται σε κυματομορφές 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 B: Να συνδυάζονται πολλές μετρήσεις από διαφορετικές διαστάσεις και από διαφορετικές τεχνικές μέτρησης, με άλλα λόγια να γίνεται επαλήθευση. Στην εργασία, έγιναν μετρήσεις των μορίων σε κατάσταση υψηλής και χαμηλής ενέργειας ψάχνοντας τυχόν σημαντικές διαφορές.

3. Πρόκληση Γ: Μέτρηση της τοπογραφίας σε πολύ μικρές κλίμακες.

Στρατηγική Γ: Πρέπει να καθοριστούν τα όρια της ακτίνας και να χρησιμοποιείται απόλυτη τιμή, σε συνδυασμό με καθορισμό της μέγιστης συχνότητας αναπαράστασης. Τα όρια της ακτίνας καθορίστηκαν στην εργασία που μελετάμε. Καθορίστηκε επίσης η διακριτή συχνότητα να είναι διπλάσια από το ρυθμό που αλλάζουν οι συχνότητες στις κυματομορφές, με σεβασμό στο θεώρημα Nyquist.

Η εργασία, που στηρίζεται στις παραπάνω μαθηματικές γνώσεις τις οποίες μπορούμε πλέον να κατανοήσουμε διαισθητικά, βασίζεται και σε μια ακόμη καινοτομία, το Molecular electrostatic potential (MESP). Το MESP μπορεί να αποτελέσει ένδειξη για τις χημικές ιδιότητες των μορίων. 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 μετρώντας την επιφάνεια του μορίου με το PSD μπορούν να βγουν συμπεράσματα για τις ιδιότητές του. Αυτό είναι και το βασικό θεώρημα της εργασίας που θα αναλύουμε. Όσον αφορά το MESP, πληροφορίες είναι διαθέσιμες στον σύνδεσμο που υπάρχει στο abstract της εργασίας: <https://pubs.rsc.org/en/content/articlelanding/2022/cp/d2cp03244a>.

Το βασικό κείμενο που αναλύουμε, 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».

Η αναζήτηση ομοιοτήτων ανάμεσα σε πρωτεΐνες βασίζεται κατά κύριο λόγο στις ομοιότητες των ακολουθιών των στοιχείων τους. Μια προσέγγιση για να λυθεί το πρόβλημα της αναζήτησης ομοιοτήτων χρησιμοποιούνται αυτοοργανούμενοι χάρτες (self-organizing maps). Οι 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 γιατί διαφορετικές μοριακές επιφάνειες πρωτεϊνών κωδικοποιούν διαφορετικές λειτουργικές πληροφορίες.

Έτσι η εργασία αυτή προτείνει έναν νέο τρόπο αναγνώρισης των μορίων. Όπως αναλύθηκε παραπάνω, οι μετρήσεις σχημάτων με PSD έχουν ένα βαθμό μοναδικότητας και αυτό είναι το στοιχείο που αξιοποιεί η εργασία. Υπάρχουν μέθοδοι που μελετούν με ακριβεία τη δομή των πρωτεϊνών, αλλά χρειάζονται πολύ χρόνο και πόρους, απαιτούν ευθυγράμμιση των επιφανειών και δεν είναι αποδοτικές σε χώρο/χρόνο. Με το PSD αναλύονται λιγότερο ακριβείς μετρήσεις και χρησιμοποιείται ο μετασχηματισμός Fourier, ο οποίος ανεξαρτήτως κλίμακας αναπαριστά την ίδια πληροφορία. Το μοντέλο παρέχει πολύ γρήγορες συγκρίσεις, κάτι πολύ βασικό, αφού εξετάζει απλά την διαφορά των μετρήσεων PSD και δεν κολλάει σε συγκρίσεις ολόκληρου του σχήματος του μορίου.

Για την εκπαίδευση του μοντέλου χρησιμοποιήθηκαν δείγματα από 4 διαφορετικές οικογένειες πρωτεϊνών: 12 ελικάσες, 4 πολυμεράσες, 6 μεθυλτρανσφεράσες ή μεθυλάσες και 4 γλυκοπρωτεΐνες. Επίσης χρησιμοποιήθηκε η πρωτεΐνη κινάση ποντικού (Mouse kinase) λόγω της πολύ διαφορετικής δομής της με τις υπόλοιπες. Οι πρωτεΐνες της πρώτης οικογένειας είναι

υπεύθυνες για το ξεδίπλωμα της έλικας DNA ή RNA κατά 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 ακτίνων 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 ιογενούς διάρροιας βοοειδών.

Οι επιφάνειες του ηλεκτροστατικού δυναμικού 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. Η θερμοκρασία τέθηκε στους 300 K και η πίεση στο 1 atm. 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 εξίσωση για το $R(r)$ που αναλύθηκε παραπάνω:

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

Η μέση τιμή του κανονικοποιημένου όγκου ήταν $1/L^*L^*L$.

Έπειτα έγινε διαχωρισμός σε κυματοσυναρτήσεις μέσω Fourier:

$$F(\mathbf{x}) = \sum_k F_k \exp[-i\mathbf{k} \cdot \mathbf{x}]$$

,με επιλογή ιδανικού $k=2\pi/v$ που δεν μπορεί να επιτευχθεί.

Μετά από απλοποιήσεις η τελική εξίσωση ήταν:

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

,που υπολογίζει διαφορές πλάτους μεταξύ ενός σήματος και δεν ενδιαφέρεται για τη φάση του.

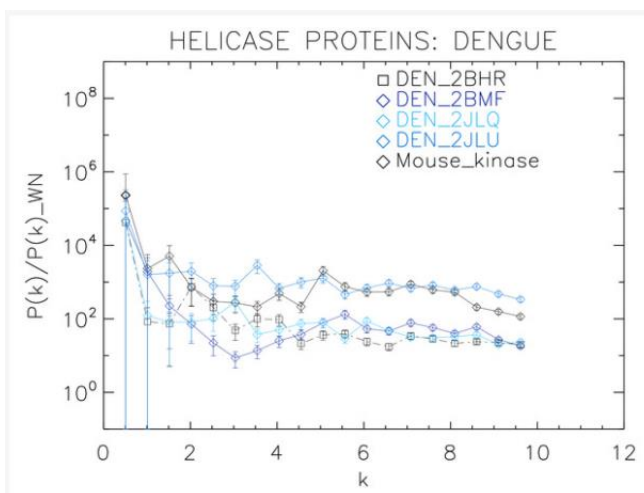
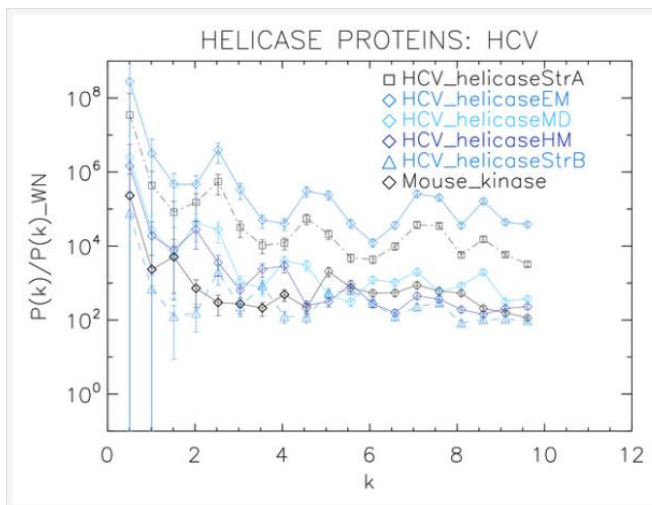
Μετά επιλέχθηκε για την αναπαράσταση ένα μέγεθος που να μπορεί να είναι διπλάσιο από την μικρότερη συχνότητα των κυματομορφών, σύμφωνα με το Nyquist. Για τις αποκλίσεις και τα λάθη χρησιμοποιήθηκε Gaussian ομοιόμορφη κατανομή. Παίρνοντας το μέσο όρο των σημείων 2-κορυφών, εξασφαλίστηκε αναπαράσταση σε 1 διάσταση, μέσω και των κυματομορφών.

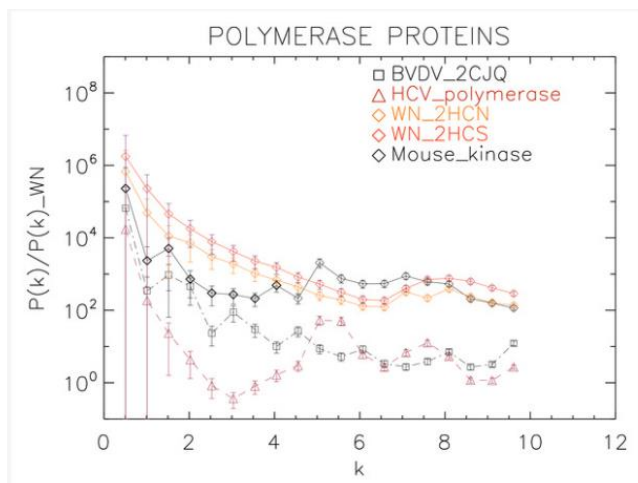
Παρατηρήθηκε ότι τα power spectra (φάσματα ενέργειας) όλων των μοριακών επιφανειών έχουν συγκρίσιμα εύρη. Για ευκολότερη σύγκριση των αποτελεσμάτων, διαιρέθηκαν τα φάσματα αυτά με τον μέσο όρο των φασμάτων λευκού θορύβου και εκτιμήθηκε συνολικά ο θόρυβος με βάση το k . Σε όλα τα αντίστοιχα γραφήματα που προέκυψαν, αναπαραστάθηκε γραφικά και το φάσμα της Mouse-κινάσης (ένα σχεδόν επίπεδο φάσμα με ανομοιόμορφες κορυφώσεις).

Αποτελέσματα:

Οι πρωτεΐνες από την πρώτη οικογένεια έχουν όλες παρόμοια δυναμική ενέργεια, κάποιες περισσότερο και κάποιες λιγότερο. Για παράδειγμα οι πρωτεΐνες του Κίτρινου πυρετού και του Δάγκειου πυρετού είχαν πολλές ομοιότητες. Παρόλα αυτά μπορούσαν να διαχωριστούν μεταξύ τους. Σε σύγκριση με τις HCV ελικάσες, το φάσμα ενέργειας 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.

Σύμφωνα και με τα γραφήματα παρακάτω, είναι διακριτό με γυμνό μάτι 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 πολυμερείς ή όχι.





Συμπεράσματα:

Αν μελετηθούν σε βάθος τα παραπάνω μοντέλα (πχ: μέσω τεχνητής νοημοσύνης) και βρεθούν τα λεπτά όρια ώστε να μπορούμε να ξεχωρίσουμε, μέσω του PSD, την οικογένεια και την λειτουργία των μορίων, σε συνδυασμό με την αναπαράσταση σε μονοδιάστατα διανύσματα που επιτρέπει ο μετασχηματισμός Fourier, η αναζήτηση και σύγκριση μορίων στη βιοπληροφορική μπορούν αλλάξουν ριζικά (πηγή: <https://peerj.com/articles/185/#>).

3^ο Paper: Antibody Clustering Using a Machine Learning Pipeline that Fuses Genetic, Structural, and Physicochemical Properties

Στο paper αυτό παρουσιάζεται μια υβριδική προσέγγιση για την αναγνώριση και συσταδοποίηση ομάδων αντισωμάτων. Η συσταδοποίηση (clustering) του V domain των αντισωμάτων είναι σημαντική για τη βελτίωση τους με στόχο να χρησιμοποιηθούν σε ανοσολογικές θεραπείες. Χάρη στην πολυπλοκότητα των αντισωμάτων, το πρόβλημα της κατηγοριοποίησής τους σε συστάδες φαίνεται να διαφέρει διαρκώς ανάλογα με το μέγεθος των διαθέσιμων δεδομένων. Δεδομένου του τεράστιου αριθμού αντισωμάτων από διαφορετικά είδη οργανισμών και τις νέες διαθέσιμες (μέσω προσομοίωσης) τρισδιάστατες κατασκευές των μεταβλητών περιοχών heavy-chain(VH) και light-chain(VL) των αντισωμάτων, η προσοχή έχει στραφεί στην υπερμεταβλητή περιοχή (hypervariable region(HVR)) του αντισώματος και συγκεκριμένα στις περιοχές συμπληρωματικότητας (complementarity-determining regions(CDRs)) που συμμετέχουν κυρίως στην αλληλεπίδραση αντισώματος-αντιγόνου.

Η προτεινόμενη προσέγγιση σύμφωνα με τους συγγραφείς διατηρεί τα πλεονεκτήματα της συσταδοποίησης αλλά αξιοποιεί παράλληλα την πληροφορία που προσφέρει η τρισδιάστατη αναπαράσταση. Οι συγγραφείς υποστηρίζουν ότι η τεχνική αυτή είναι χρήσιμη για την συσταδοποίηση αντισωμάτων που δεν μπορούν να κατηγοριοποιηθούν αποτελεσματικά με φυλογενετικές τεχνικές ή αποκλειστικά με αλγορίθμους συσταδοποίησης, και ότι μπορεί να εφαρμοστεί και για άλλες κατηγορίες πρωτεϊνών πέρα από τα αντισώματα.

Οι περισσότερες υπερμεταβλητές περιοχές στα αντισώματα έχουν έναν μικρό αριθμό διακριτών διατάξεων που ονομάζονται κανονικές δομές και έναν μικρό αριθμό υπολειμμάτων-κλειδιών που θα μπορούσαν να χρησιμοποιηθούν για την πρόβλεψη της συστάδας στην οποία θα πρέπει να κατηγοριοποιηθεί κάποιο συγκεκριμένο αντίσωμα. Επιπλέον, έχουν ανακαλυφθεί με βάση τις CDR νέες συστάδες κανονικών δομών σε συγκεκριμένες θέσεις συντήρησης που είναι υπεύθυνες για τις διαφοροποιήσεις ανάμεσα σε συστάδες αντισωμάτων. Σε μια παλαιότερη

προσέγγιση για την κατηγοριοποίηση των παρατηρούμενων CDRs, έγινε μελέτη των συστάδων στον εσωτερικό χώρο συντεταγμένων (internal coordinate space), ακολουθούμενη από συγχώνευση ομάδων δομών στον Καρτεσιανό χώρο χρησιμοποιώντας την απόκλιση της μέσης τετραγωνικής τιμής.

Παρατηρήθηκε όμως ότι παρότι δύο συστάδες μπορεί να ήταν πιο κοντά σε κάποια κανονική κλάση όσο αφορούσε την αλληλουχία τους, οι δομές τους μπορεί να ήταν πιο κοντά σε κάποια άλλη. Παρατηρήθηκε επίσης ότι όταν το heavy chain χωρίστηκε σε δύο περιοχές, κεφαλή (head) και σώμα (torso) και μελετήθηκαν τα CDRs σε αυτό, το σώμα συνήθως έπαιρνε μία από δύο μορφές, ενώ η δομή της κεφαλής εξαρτόταν από τη δομή του σώματος. Μπορεί επομένως να γίνει ανάλυση και κατηγοριοποίηση με βάση το μήκος των CDRs ή δομικές πληροφορίες στην επιφάνεια των σημείων συνένωσης του V domain.

Για τους σκοπούς της έρευνας συλλέχθηκαν δείγματα δομών αντισωμάτων V domain από τη βάση δεδομένων Protein Data Bank (PDB). Χρησιμοποιήθηκαν μέθοδοι εξόρυξης δεδομένων, χαρακτηριστικά της αλληλουχίας και της δομής των αντισωμάτων, ένα υβριδικό μητρώο απόστασης (hybrid distance matrix) και αλγόριθμοι φυλογενετικής ανάλυσης. Για έλεγχο της ποιότητας τους, τα αποτελέσματα συγκρίθηκαν με τα 3D δεδομένα που υπάρχουν στην βάση δεδομένων PDB. Το τελικό σύνολο δεδομένων περιείχε μοναδικά V domains, το καθένα με τουλάχιστον ένα ολοκληρωμένο variable domain ενός heavy ή light chain. Κατασκευάστηκε μια βάση δεδομένων με τα δομικά τους στοιχεία και ανακαλύφθηκαν πάνω από 50 υπογραφές αντισωμάτων. Στη συνέχεια τα δείγματα αριθμήθηκαν και καθορίστηκαν οι CDR περιοχές μέσω κανονικών εκφράσεων για τα μοτίβα συντηρημένων ακολουθιών (conserved sequence).

Μελετήθηκαν τα διμερή αντιγόνου-αντισώματος που ανήκαν στο ίδιο PDB αρχείο και όλες οι αναγνωρίσιμες αλληλεπιδράσεις εντοπίστηκαν, σημειώθηκαν και μετατράπηκαν σε ψηφιακές υπογραφές βασισμένες στο V domain και στη συνάρτηση:

$I(absi) = 0$, if all the elements of $R(absi \ ag \ m) > 4$

$I(absi) = 1$, if one of the elements of $R(abs \ ag \ i \ im) \leq 4$

, όπου R η ευκλείδεια απόσταση μεταξύ δύο σημείων στον τρισδιάστατο χώρο, abs το αντίσωμα, ag το αντιγόνο, n ο αριθμός των ατόμων στο αρχείο PDB για το αντίσωμα και m ο αριθμός των ατόμων στο PDB αρχείο για το αντιγόνο. Τα V domains για τα αντισώματα και οι τρισδιάστατες κατασκευές για τα αντιγόνα θεωρήθηκαν τρισδιάστατα σημεία στο χώρο. Όλες οι εντοπιζόμενες αλληλεπιδράσεις αντικαταστάθηκαν στις πρωτεϊνικές αλληλουχίες με 0 ή 1 και κάθε περιοχή του V domain αντικαταστάθηκε με 0 ή 1 αν μία ή περισσότερες αλληλεπιδράσεις εντοπίστηκαν σε αυτήν. Το αποτέλεσμα ήταν ένα διάγραμμα «υπογραφής» 5 τιμών όπου κάθε τιμή αντιστοιχούσε σε μια περιοχή του V domain.

Το υβριδικό μητρώο αποστάσεων υπολογίστηκε από τη συνάρτηση:

$$H = A * B,$$

όπου A Jukes-Cantor μητρώο και B μητρώο για τις αποστάσεις μεταξύ ατόμων στα αντισώματα. Οι υβριδικές αποστάσεις που υπολογίστηκαν αναλύθηκαν ιεραρχικά και υποβλήθηκαν σε πολλαπλή στοίχιση μέσω του εργαλείου MATLAB. Έγινε φυλογενετική ανάλυση και τα αποτελέσματα οπτικοποιήθηκαν σε μορφή δέντρων με χρήση ακτινοβολίας. Κάθε μοναδικό V domain αντισώματος προστέθηκε στη βάση με τις δομικές πληροφορίες και διαχωρίστηκε σε δύο μέρη, τα “heavy chains” και “light chains” και σημειώθηκαν τα αντισώματα που σχημάτιζαν διμερές σύμπλεγμα με κάποιο αντιγόνο. Τα δεδομένα της βάσης συμπληρώθηκαν με τις απαραίτητες πληροφορίες από τη βάση NCBI. Τέλος, όλα τα δεδομένα της βάσης χωρίστηκαν σε συστοιχίες με βάση την υπογραφή τους.

Τα φυλογενετικά δέντρα που κατασκευάστηκαν από το μοντέλο είχαν παρόμοια τοπολογία και μοτίβα διακλάδωσης. Τα αντισώματα με ασαφή συμπεριφορά ομαδοποιήθηκαν με αντισώματα που είχαν γνωστή συμπεριφορά για έναν αριθμό πιθανών αντιγόνων. Κάθε διαφορετική ομάδα-συστοιχία αντιστοιχίστηκε σε συγκεκριμένο αριθμό μοτίβων αλληλεπίδρασης από όλα τα μοτίβα που βρέθηκαν συνολικά και οι πιο συχνά εμφανιζόμενες αλληλεπιδράσεις φάνηκε πως υπήρχαν μέσα στις συστάδες 1, 2, 3, 5 και 8.

Τα αποτελέσματα της μελέτης έδειξαν ότι δεν υπάρχει αλληλεπίδραση μόνο μεταξύ CDR και αντιγόνων, αλλά ότι διαφορετικά αμινοξέα αλληλεπιδρούν με τα αντιγόνα, και μάλιστα με διαφορετική συχνότητα από περιοχή σε περιοχή. Σε κάποιες περιπτώσεις υπάρχει αλληλεπίδραση μόνο με την περιοχή FR(fragment region) του αντισώματος μέσα στα VL και VH domains. Επιπλέον, πολλά αμινοξέα «κλειδιά» φαίνεται να παρουσιάζουν διαφορές στην κάθε εμφάνισή τους, ενώ κάποια άλλα φαίνεται να χάνονται. Αυτή η παρατήρηση είναι σημαντική για την πειραματική κατασκευή αντισωμάτων, πιθανών και νέων, που πιθανόν θα χρησιμοποιηθούν σε θεραπείες στο μέλλον. Στον μηχανισμό αλληλεπίδρασης του CDR καθοριστικές ήταν οι τυροσίνη (tyrosine) και σερίνη (serine), ενώ στον μηχανισμό αλληλεπίδρασης των περιοχών FR οι γλυκίνη (glycine), τρυπτοφάνη (tryptophan) και φαινυλαλανίνη (phenylalanine).

Ένα από τα θετικά του προτεινόμενου μοντέλου σύμφωνα με τους συγγραφείς είναι ότι με τη χρήση του συνδυασμού των αποστάσεων στο υβριδικό μητρώο μπορεί να γίνει συσταδοποίηση άγνωστων μορίων αντισωμάτων με γνωστά, εύκολα και βέλτιστα, χωρίς να χρειάζονται επιπλέον ευρετικές τεχνικές, οι οποίες θα απαιτούσαν και περισσότερο χρόνο. Το μοντέλο επίσης επιτρέπει να ληφθούν υπόψη και οι τρισδιάστατες δομές των μορίων. Η συνάρτηση που χρησιμοποιήθηκε είναι ευέλικτη γιατί αντιστοιχεί κάθε συσχέτιση σε επίπεδο αλληλουχίας με την ομοιότητα σε επίπεδο τρισδιάστατης δομής.

Συμπεράσματα της συγκεκριμένης μελέτης είναι ότι η κατηγοριοποίηση των αντισωμάτων σε συστάδες αποτελεί πρόκληση για την ανοσολογία. Ακόμα και αν δύο αντισώματα έχουν ακριβώς την ίδια αλληλουχία μπορεί στον τρισδιάστατο χώρο να παρουσιάζουν εντελώς διαφορετική δομή και συμπεριφορά απέναντι στα αντιγόνα. Για να αντιμετωπιστεί το πρόβλημα, πρέπει να ληφθούν υπόψη όλες οι απαιτούμενες βιολογικές παράμετροι σε κάποιας μορφής υβριδικό μοντέλο, όπως το προτεινόμενο. Τα αποτελέσματα της χρήσης τέτοιων μοντέλων θα μπορούσαν να οδηγήσουν σε ριζοσπαστικές ανακαλύψεις που αφορούν την έρευνα γύρω από την ανοσολογία, την βιοτεχνολογία, τη διάγνωση, την ανοσοθεραπεία και άλλους τομείς.

Επεκτάσεις της παραπάνω δημοσίευσης:

1. Μια εφαρμογή της συσταδοποίησης αντισωμάτων με ερευνητικό ενδιαφέρον είναι το antibody engineering, η κατασκευή δηλαδή, αντισωμάτων σε εργαστηριακό περιβάλλον. Προς το παρόν η βάση δεδομένων PDB περιέχει κυρίως θραύσματα (fragments) αντισωμάτων και κάποια ολόκληρα δείγματα. Η μοντελοποίηση των αντισωμάτων έχει μεγάλη σημασία γιατί ο τρόπος που γίνεται η πρόσδεση αντισώματος σε αντιγόνο είναι πολύ συγκεκριμένος και γίνεται σε συγκεκριμένα σημεία της επιφάνειας των μορίων υπό συγκεκριμένη γωνία (elbow bend). Στην πρόσδεση αυτή καθοριστικός είναι ο προσανατολισμός του V domain και οι κανονικές δομές των CDRs. Μάλιστα, δομικοί περιορισμοί στο σημείο πρόσδεσης είναι πιθανό να εμποδίσουν την αναγνώριση του αντιγόνου από το αντίσωμα. Η γνώση των κανονικών δομών των μορίων επιτρέπει την ευκολότερη μοντελοποίησή τους. Αφού όμως η μοντελοποίηση με προσομοίωση σε υπολογιστή είναι προς το παρόν περιορισμένη, η ύπαρξη πληροφορίας σε βάσεις δεδομένων για την συμπεριφορά συστάδων από αντιγόνα μπορεί να διευκολύνει τη σωστή πρόβλεψη της δομής, συνεπώς και την κατασκευή τους ή ακόμα και την βελτίωση της δομής τους στο εργαστήριο.

(πηγή:

Chiu ML, Goulet DR, Teplyakov A, Gilliland GL. Antibody Structure and Function: The Basis for Engineering Therapeutics. Antibodies (Basel). 2019 Dec 3;8(4):55. doi: 10.3390/antib8040055. PMID: 31816964; PMCID: PMC6963682.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6963682/>)

2. Το clustering αντισωμάτων έχει επίσης εφαρμοστεί στην επιλογή ασθενών για συμμετοχή σε δοκιμές θεραπειών με βάση την κατηγοριοποίηση των αντισωμάτων τους σε συστάδες. Έχει σημασία σε τέτοιες δοκιμές τα ανοσολογικά προφίλ των υποψηφίων να εμφανίζουν ποικιλομορφία, για να μπορεί να θεωρηθεί ότι τα συμπεράσματα της έρευνας αντιπροσωπεύουν το σύνολο του πληθυσμού. Το clustering με βάση το CDR έχει αποδειχθεί χρήσιμο, αν και όχι με πλήρη επιτυχία, στον εντοπισμό ομοιοτήτων και διαφορών ανάμεσα σε σύνολα από δείγματα αντισωμάτων, ειδικά σε συνδυασμό με αλγορίθμους και εργαλεία τεχνητής νοημοσύνης.

(πηγή:




Chomicz D, Kończak J, Wróbel S, Szaława T, Dudzic P, Janusz B, Tarkowski M, Deszyński P, Gawłowski T, Kostyn A, Orłowski M, Klaus T, Schulte L, Martin K, Comeau SR, Krawczyk K. Benchmarking antibody clustering methods using sequence, structural, and machine learning similarity measures for antibody discovery applications. Front Mol Biosci. 2024 Mar 28;11:1352508. doi: 10.3389/fmolb.2024.1352508. PMID: 38606289; PMCID: PMC11008471. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11008471/>)

4^ο Paper: Discovery and Evaluation of Protein Biomarkers as a Signature of Wellness in Late-Stage Cancer Patients in Early Phase Clinical Trials

Το paper περιγράφει την βοηθητική μέθοδο TARGET (tumour characterization to guide experimental targeted therapy), που έχει στόχο την επιλογή καταλληλότερης θεραπείας για τους ασθενείς, με βάση τα γονίδια τους. Ιδανικά η επιλογή υποψηφίων ασθενών για κλινικές μελέτες συχνά γίνεται με την ανάλυση εξατομικευμένων βιολογικών δεικτών (biomarkers), όμως οι περισσότεροι ασθενείς δεν έχουν καθαρό γονιδιωματικό οδηγό (genomic driver) για την ασθένειά τους. Αντίθετα, οι πρωτεΐνες δίνουν πολύ περισσότερη πληροφορία για την κατάσταση ενός οργανισμού μέχρι και σε κυτταρικό επίπεδο. Επομένως, η ανάλυση του πρωτεϊνώματος ενός ασθενούς είναι πολύ καλύτερος δείκτης για την κατάσταση της υγείας του, άρα και για την επιλογή του για συμμετοχή σε κλινικές μελέτες.

Είναι σημαντικό να γίνει προσεκτική επιλογή για την έρευνα, τόσο για την ασφάλεια των ίδιων των ασθενών όσο και για να είναι έγκυρα τα αποτελέσματα των δοκιμών. Συνηθισμένο κριτήριο επιλογής ενός ασθενή για συμμετοχή σε τέτοιες δοκιμές ήταν ο δείκτης Performance Status(PS), που υπολόγιζε το προσδόκιμο ζωής του ασθενή με βάση την τρέχουσα κατάσταση της υγείας του. Ο βαθμός κατάστασης του ασθενούς (PS) είναι ένας από τους πιο κεντρικούς παράγοντες στην ογκολογική φροντίδα. Παίζει ρόλο τόσο στον προγνωστικό προσδιορισμό όσο και στην καλύτερη θεραπεία για έναν ασθενή με καρκίνο. Ο βαθμός κατάστασης αξιολογεί την ικανότητα του ασθενούς να εκτελεί συγκεκριμένες δραστηριότητες της καθημερινής ζωής χωρίς τη βοήθεια άλλων. Αυτές οι δραστηριότητες περιλαμβάνουν βασικές δραστηριότητες. Υπάρχουν 2 ευρέως χρησιμοποιούμενες κλίμακες για τον βαθμό κατάστασης. Η πιο συνηθισμένη είναι η κλίμακα Zubrod ή ECOG (Eastern Cooperative Oncology Group). Αυτή η κλίμακα κυμαίνεται από 0 έως 4, με το 0 να αντιστοιχεί σε πλήρη λειτουργικότητα και άνευ συμπτωμάτων, και το 4 να αντιστοιχεί σε ασθενή που βρίσκεται στο κρεβάτι. Η άλλη κλίμακα που χρησιμοποιείται περιστασιακά είναι η κλίμακα

Karnofsky. Αυτή η κλίμακα κυμαίνεται από 10 (σε κατάσταση αγωνίας) έως 100 (χωρίς περιορισμούς). Στην παρακάτω εικόνα φαίνεται ένα παράδειγμα χρήσης τέτοιας κλίμακας:

Zubrod Scale	Karnofsky Scale
0 Normal activity 	100 Normal; no evidence of disease
1 Symptomatic and ambulatory; cares for self	90 Able to perform normal activities with only minor symptoms
2 Ambulatory >50% of time; occasional assistance	80 Normal activity with effort; some symptoms
3 Ambulatory ≤50% of time; nursing care needed 	70 Able to care for self but unable to do normal activities
4 Bedridden 	60 Requires occasional assistance; cares for most needs
	50 Requires considerable assistance
	40 Disabled; requires special assistance
	30 Severely disabled
	20 Very sick; requires active supportive treatment
	10 Moribund

Είναι σημαντικό να σημειωθεί ότι ο βαθμός κατάστασης δεν είναι μόνο ένας αριθμός, αλλά αντιπροσωπεύει τη συνολική κατάσταση του ασθενούς. Οι ιατροί λαμβάνουν υπόψη και άλλους παράγοντες για την κατηγοριοποίηση, όπως τον τύπο του καρκίνου.

(πηγή: <https://jamanetwork.com/journals/jamaoncology/fullarticle/2432463>)

Βλέπουμε ότι το PS από το 1948 εξαρτάται μόνο από τη βασική εικόνα της υγείας του ασθενούς. Σο paper που μελετάμε συνυπολογίζονται περισσότεροι και πιο σύγχρονοι παράγοντες για να εξελίξουν το δείκτη PS. Με τη σημερινή τεχνολογία πλέον αυτό είναι δυνατό με κάποιο αυξημένο κόστος.

Οι συγγραφείς του paper δημιούργησαν ένα νέο σκορ υγείας (wellness score), που κάνει παρόμοιο υπολογισμό του προσδόκιμου ζωής αλλά με βάση πρωτεΐνες στο αίμα και μελετήθηκε σε ασθενείς που αναμένεται να είναι ζωντανοί τουλάχιστον για έξι μήνες μετά τον υπολογισμό του. Αποτελεί ισχυρότερο και πιο αντικειμενικό δείκτη για τον προσδιορισμό του σταδίου του καρκίνου από το Performance Status, γιατί για τον υπολογισμό του εκτός από το PS χρησιμοποιήθηκαν μετρήσεις των επιπέδων LDH, albumin και Hb στο αίμα. Παράλληλα χρησιμοποιήθηκε πρωτεομική ανάλυση (Mass Spectrometry (MS) proteomics), που δείχνει λεπτομέρειες για τον φαινότυπο κάθε πρωτεΐνης, ειδικά όταν η ανάλυση γίνεται και με τη βοήθεια εργαλείων τεχνητής νοημοσύνης. Η πρωτεομική ανάλυση είναι καθοριστική για την έγκαιρη διάγνωση, την πρόβλεψη του τρόπου που μπορεί να εξελιχθεί η ασθένεια και των αιτιών που οδηγούν σε συγκεκριμένη έκβαση, την εύρεση κατάλληλης θεραπείας, και την πρόβλεψη τυχόν ανθεκτικότητας στα φάρμακα/θεραπείες. Ο λόγος είναι ότι λαμβάνει υπόψη την μεγάλη ετερογένεια που υπάρχει στους φαινοτύπους των πρωτεϊνών.

Το εργαλείο SWATH-MS χρησιμοποιήθηκε για την γρήγορη ταυτοποίηση πρωτεϊνών και την παραγωγή των φασμάτων θραυσμάτων (fragment spectra) για όλα τα πεπτίδια κάθε δείγματος από αυτά που μελετήθηκαν με την μέθοδο TARGET. Οι υποψήφιοι ασθενείς αντιστοιχίστηκαν σε θεραπείες με βάση το μοριακό έλεγχο (molecular screening) και το είδος της ασθένειας και αναπτύχθηκε το πρωτεομικό προφίλ τους. Με βάση τα δεδομένα αυτά έγινε προσπάθεια να αναπτυχθεί ένας βελτιωμένος αλγόριθμος.

Η δοκιμή του εργαλείου για μια ομάδα ανακάλυψης (discovery cohort) με 73 ασθενείς και μια ομάδα επιβεβαίωσης (validation cohort) με 79 ασθενείς χωρίστηκε σε δύο μέρη. Ο σκοπός ήταν να αξιολογηθούν οι αλληλουχίες σε καρκινικούς όγκους και η ικανότητα να κατηγοριοποιηθούν με βάση αυτό οι ασθενείς σε ομάδες δοκιμών θεραπείας. Συλλέχθηκαν δείγματα από το πλάσμα των ασθενών και στη συνέχεια αφαιρέθηκαν από αυτό οι 12 πιο συνηθισμένες πρωτεΐνες. Το υλικό που

προέκυψε αναλύθηκε για ποσότητες πρωτεΐνης χρησιμοποιώντας ένα χημικό αντιδραστήριο, εφαρμόστηκε φασματομετρία (spectrometry) και η μέθοδος μεταβλητού παραθύρου (variable window method). Η ανάλυση των δεδομένων έγινε με τη γλώσσα R και υπολογίστηκαν οι συντελεστές της διαφοράς μεταξύ των αντιγράφων. Η σημαντικότητα της παρουσίας ή όχι μεγάλης ποσότητας πρωτεϊνών υπολογίστηκε χρησιμοποιώντας εμπειρική στατιστική Bayes για διαφορικές εξισώσεις. Τα δεδομένα που προέκυψαν χωρίστηκαν σε σύνολα εκπαίδευσης και δοκιμής με τον αλγόριθμο μηχανικής μάθησης RandomForest. Συνολικά δημιουργήθηκαν 1000 μοντέλα και οι σημαντικότητες των πρωτεϊνών αξιολογήθηκαν για όλα τα μοντέλα. Η παράμετρος που χρησιμοποιήθηκε για βελτιστοποίηση ήταν η ακρίβεια.

Ως θεραπεία θεωρήθηκε οποιαδήποτε μέθοδος στοχεύει στην αντιμετώπιση ή την επιβράδυνση της ασθένειας, όχι όμως οι μέθοδοι ανακούφισης από δυσάρεστα συμπτώματα. Ορίστηκαν διαφορετικές "γραμμές" θεραπείας για να καθορίσουν πόσο σοβαρά νοσούσαν οι ασθενείς πριν τη θεραπεία τους. Τα σημεία της νόσου (sites of disease) βρέθηκαν μέσω σαρώσεων υπολογιστικής τομογραφίας. Λήφθηκαν υπ' όψη μόνο αν βρέθηκαν την ώρα του σαρώματος και όχι αν είχαν αφαιρεθεί με χειρουργείο και δεν είχαν επανεμφανιστεί. Χρησιμοποιήθηκαν ένα άνω και ένα κάτω διάφραγμα για τη μέτρηση λεμφαδένων (μέγιστο 2 ανά άτομο). Σε ασθενείς με ένα σημείο νόσου αντιστοιχίστηκε ο αριθμός 0, ενώ σε ασθενείς με περισσότερα από ένα αντιστοιχίστηκε ο αριθμός 2.

Ως χρόνος επιβίωσης (survival) ορίστηκε το διάστημα από την ημερομηνία της συγκατάθεσης του ασθενή για ανάλυση των δεδομένων του με το TARGET μέχρι την ημερομηνία θανάτου του από οποιαδήποτε αιτία. Ο συνολικός χρόνος επιβίωσης καθορίστηκε με καμπύλες Kaplan–Meier.

Τα αποτελέσματα της δοκιμής έδειξαν ότι 77 συνολικά πρωτεΐνες διέφεραν ανάμεσα στα δείγματα ασθενών που πέθαναν μέσα σε 6 μήνες από τη λήψη πλάσματος και αυτούς που πέθαναν μετά τους 6 μήνες. Χρησιμοποιήθηκαν τεχνικές μηχανικής μάθησης για να ταυτοποιηθούν οι πρωτεΐνες και να γίνει στατιστική ανάλυση. Οι πρωτεΐνες φιλτραρίστηκαν περισσότερο με παλινδρόμηση Cox (Cox regression). Βρέθηκε ότι τρεις πρωτεΐνες είχαν σημαντική συσχέτιση με την επιβίωση. Μια από αυτές, η Leucine-rich alpha-2-glycoprotein είχε θετική συσχέτιση, δηλαδή όσο μεγαλύτερη ποσότητά της βρέθηκε στα δείγματα τόσο αυξημένος ήταν ο κίνδυνος θανάτου. Οι άλλες δύο πρωτεΐνες, Apolipoprotein C-III και Plasma serine protease inhibitor είχαν αρνητική συσχέτιση.

Προκειμένου να δημιουργηθεί το wellness score με βάση και τις τρεις αυτές πρωτεΐνες, χρησιμοποιήθηκαν διάφορες προσεγγίσεις μαζί με την ανάπτυξη ενός κλινικού εργαλείου. Ο αριθμός 0 χρησιμοποιήθηκε για να υποδείξει μια μείωση στον κίνδυνο θανάτου και σχεδιάστηκε ένα δέντρο αποφάσεων για το wellness score όλων των ασθενών. Το wellness score για κάθε ασθενή προέκυψε από το άθροισμα όλων των πρωτεϊνικών του σκορ. Αν είχαν συνολικό πρωτεϊνικό σκορ από 0 έως 1 το wellness score τους ήταν 0, ενώ αν είχαν συνολικό πρωτεϊνικό σκορ από 2 έως 3 το wellness score τους ήταν 1. Συνολικά, σημαντική διαφορά στο πλάσμα ασθενών που πέθαναν σε 6 μήνες από τη λήψη δείγματος και στο πλάσμα ασθενών που πέθαναν μετά τους 6 μήνες υπήρχε σε 118 πρωτεΐνες.

Τα διαφορετικά σκορ αξιολογήθηκαν με χαρακτηριστικές καμπύλες λειτουργίας δέκτη (user receiver operatic characteristics curves), στις οποίες το ψευδοθετικό ποσοστό (false positive rate), αλλιώς specificity, αναπαραστάθηκε γραφικά σε σχέση με το πραγματικό θετικό ποσοστό (true positive rate), αλλιώς sensitivity. Στη συνέχεια μελετήθηκαν σαν μέσα πρόβλεψης τα σκορ σε σημεία του γραφήματος που αντιστοιχούσαν σε διαφορετικές χρονικές στιγμές πριν τον θάνατο. Η περίοδος με την μεγαλύτερη κρισιμότητα ήταν οι 9 μήνες.

Απευθείας σύγκριση δεν ήταν δυνατή με βάση τα δεδομένα των ασθενών που συμμετείχαν στη δοκιμή του TARGET. Γι' αυτό, συγκρίθηκαν μέσω GPS τα δεδομένα από το SWATH-MS και τα επίπεδα αλβουμίνης στα δείγματα πλάσματος των ασθενών μέσω ενός C-reactive protein score (CRP). Δείγματα ασθενών με κανονικοποιημένη ποσότητα CRP μεγαλύτερη από 0 έλαβαν CRP score

1 και δείγματα με κανονικοποιημένη ποσότητα κάτω από 0 έλαβαν CRP score 0. Κανένα δείγμα στην ομάδα ανακάλυψης δεν είχε ταυτόχρονα υψηλό CRP score και χαμηλό σκορ για την αλβουμίνη.

Αποτελέσματα:

Διαπιστώθηκε ότι το wellness score, το PS και ο αριθμός των σημείων νόσου ήταν όλα καθοριστικά για την επιβίωση. Όμως μόνο το wellness score είχε άμεση συσχέτιση με την επιβίωση. Το wellness score κατάφερε να προβλέψει σωστά την έκβαση της ασθένειας σε βάθος 6 μηνών για το 75% των ασθενών στην ομάδα ανακάλυψης και 66% των ασθενών στην ομάδα επιβεβαίωσης. Επιπλέον, παρότι το PS από μόνο του είχε λιγότερη επιτυχία στην πρόβλεψη του αποτελέσματος και στην κατηγοριοποίηση των ασθενών, σε συνδυασμό με το wellness score μελετήθηκε ως ένα ακόμα ενιαίο σκορ, Phase I proteomics (PPM) score. Αυτό έδειξε ότι υπήρχε σημαντική διαφορά ανάμεσα στις δύο κατηγορίες ασθενών σχετικά με την επιβίωση στην ομάδα ανακάλυψης και στην ομάδα επιβεβαίωσης. Συντριπτικός αριθμός ασθενών στο ένα σύνολο έπασχε από καρκίνο του παχέος εντέρου σε σχέση με το άλλο. Η ανάλυση έδειξε επίσης ότι υπήρχε σημαντική διαφορά στην τελική έκβαση της ασθένειας ανάμεσα στους ασθενείς με καλή και κακή πρόγνωση έκβασης της ασθένειάς τους (outcome wellness score) στα δύο σύνολα.

Η μείωση στο αναστολέα πρωτεάσης στο πλάσμα (plasma serine protease inhibitor) έχει συσχετιστεί με την μετάσταση και επιθετικότητα του καρκίνου. Παρότι οι πρωτεΐνες που αφορούν το wellness score έχουν συσχετιστεί με συγκεκριμένους τύπους καρκίνου και ο εντοπισμός τους βοηθά την πρόγνωση, η χρησιμότητά τους στον καθορισμό της συνολικής υγείας των ασθενών για μεγάλο εύρος περιπτώσεων καρκίνων δεν έχει παρατηρηθεί ξανά στο παρελθόν.

Η πρωτεομική ανάλυση για να ανακαλυφθούν νέοι προγνωστικοί βιολογικοί δείκτες με βάση τους οποίους θα μπορούν να ταυτοποιούνται ασθενείς των οποίων η ασθένεια προχωρά τόσο γρήγορα που υπάρχει μόνο ένα μικρό χρονικό παράθυρο όπου η αποτελεσματικότητα μη-δοκιμασμένων φαρμάκων μπορεί να είναι αποτελεσματική, είναι μια πολύ σημαντική παρατήρηση. Επιπλέον η ανακάλυψη νέων βιολογικών δεικτών έδειξε την ικανότητα τεχνικών όπως το SWATH-MS να ανακαλύπτουν προηγουμένως άγνωστους βιολογικούς δείκτες.

Το wellness score ήταν αξιόπιστο στο να προβλέπει την επιβίωση τόσο στην ομάδα ανακάλυψης όσο και στην ομάδα επιβεβαίωσης, ενώ όσο εμπλουτίζεται με πρόσθετες μετρικές, γίνεται ακόμα περισσότερο ακριβές. Παρ' όλα αυτά, περιορισμοί της συγκεκριμένης έρευνας ήταν ο μικρός αριθμός δειγμάτων από ασθενείς ενώ είναι άγνωστο κατά πόσο το wellness score τελικά επηρεάζεται από την ίδια την ασθένεια του καρκίνου. Τέλος, τα συστήματα που βασίζονται σε proteomic scores κοστίζουν περισσότερο από το PS.

Οι περισσότερες πρωϊμες κλινικές δοκιμές απαιτούν προσδόκιμο ζωής 3-6 μήνες. Η υποκειμενική φύση του PS δίνει ανακριβή αποτελέσματα για τους ασθενείς και δυσκολεύει στην επιλογή καταλληλότερης θεραπείας. Μια μέθοδος, όπως τακτικές εξετάσεις αίματος, θα μπορούσε να αποτελέσει ένα πιο αντικειμενικό εναλλακτικό μέτρο καθορισμού της κατάστασης των ασθενών για συμμετοχή σε κλινικές δοκιμές.

Πιθανόν θα μπορούσαν στο μέλλον να αποδειχθούν χρήσιμες τεχνικές βασισμένες στα αντισώματα των τριών πρωτεϊνών που σχετίζονται με τον καρκίνο αν χρησιμοποιηθούν ως προγνωστικοί δείκτες.

Επεκτάσεις για την παραπάνω δημοσίευση:

Παρόμοιοι τύπου wellness scores μπορούν, σε συνδυασμό με άλλες σημαντικές παραμέτρους, να χρησιμοποιηθούν για την έγκαιρη διάγνωση ορισμένων μορφών καρκίνου που δεν εντοπίζονται εύκολα από άλλου είδους εξετάσεις. Χαρακτηριστικό παράδειγμα είναι ο καρκίνος του πνεύμονα, όπου η συσχέτιση μεταξύ πρωτεομικής ανάλυσης και πρόγνωσης σε ασθενείς που υποβάλλονταν σε μερική πνευμονοεκτομή ήταν μέχρι πρόσφατα ασαφής. Τα διαφορετικά είδη καρκίνου του πνεύμονα συχνά εξελίσσονται πολύ διαφορετικά και απαιτούν διαφορετικές μεθόδους θεραπείας, ειδικά σε περιπτώσεις μετάστασης ή επανεμφάνισης στον ίδιο ασθενή. Η πρωτεομική ανάλυση από

μόνη της δεν έχει χρησιμοποιηθεί ευρέως για τον καρκίνο του πνεύμονα. Όμως μπορεί να αξιοποιηθεί σε συνδυασμό με άλλες μετρικές (όπως η μετάσταση λεμφαδένων, το depth of invasion και το μέγεθος του καρκινικού όγκου) για την δημιουργία ενός ειδικού σκορ που θα προβλέπει το προσδόκιμο ζωής ειδικά για ασθενείς με καρκίνο του πνεύμονα.

(πηγή:

Peng J, Zhang J, Zou D, Gong W. Proteomics score: a potential biomarker for the prediction of prognosis in non-small cell lung cancer. *Transl Cancer Res.* 2019 Sep;8(5):1904-1917. doi: 10.21037/tcr.2019.08.39. PMID: 35116940; PMCID: PMC8798976.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8798976/>)

Η ανάπτυξη εξατομικευμένων φαρμάκων βασίζεται στη μελέτη βιολογικών δεικτών, ιδιαίτερα για τη θεραπεία του καρκίνου. Βιολογικοί δείκτες μπορούν να εντοπιστούν με την ανάλυση μοριακών, φυσιολογικών ή ανατομικών χαρακτηριστικών και μπορούν να χρησιμοποιηθούν είτε διάγνωση είτε για πρόβλεψη της ασθένειας. Η στατιστική ανάλυση δειγμάτων όπως του αίματος, μπορούν να προβλέψουν τόσο την εμφάνιση του καρκίνου όσο και την πιθανή εξέλιξή του

(πηγή:

Jung SH. Design and Analysis of Cancer Clinical Trials for Personalized Medicine. *J Pers Med.* 2021 May 4;11(5):376. doi: 10.3390/jpm11050376. PMID: 34064394; PMCID: PMC8147797.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8147797/>)

Η στατιστική ανάλυση μπορεί να συνδυαστεί με πρωτεομική ανάλυση και με σκορ όπως το wellness score και μπορούν να χρησιμοποιηθούν εργαλεία παρόμοια με το TARGET ώστε να γίνεται διάγνωση νωρίτερα απ'ότι είναι σήμερα εφικτό και ακόμα καλύτερη πρόβλεψη.

Συμπεράσματα από το σύνολο των papers:

Κάθε πρωτεΐνη μπορεί να αναγνωριστεί μοναδικά λόγω της αλληλουχίας των αμινοξέων της, της επιφάνειάς της και του τρόπου που αλληλεπιδρά χημικά με άλλες δομές. Τα χαρακτηριστικά αυτά είναι όλα χρήσιμα για την αναγνώριση ασθενειών, τη σύγκριση διαφορετικών γονιδιωμάτων οργανισμών, την προσαρμογή και εξατομίκευση θεραπειών στα γενετικά χαρακτηριστικά των ασθενών και την πρόβλεψη της εξέλιξης μιας ασθένειας. Με την πρόοδο της τεχνολογίας και της βιοπληροφορικής είναι πλέον δυνατό τα στοιχεία αυτά να αναλύονται ή και να προσομοιώνονται με τη βοήθεια υπολογιστή. Αυτό έχει οδηγήσει στην ανάπτυξη νέων μεθόδων και εργαλείων που έχουν διευκολύνει σημαντικά τη μελέτη των πρωτεϊνών και κατά συνέπεια των βιολογικών διεργασιών και των οργανισμών.

Βιβλιογραφία:

1. Carvalho CS, Vlachakis D, Tsiliki G, Megalooikonomou V, Kossida S. 2013. Protein signatures using electrostatic molecular surfaces in harmonic space. *PeerJ* 1:e185
<https://doi.org/10.7717/peerj.185>
2. Papageorgiou L, Maroulis D, Chrousos GP, Eliopoulos E, Vlachakis D. Antibody Clustering Using a Machine Learning Pipeline that Fuses Genetic, Structural, and Physicochemical Properties. *Adv Exp Med Biol.* 2020;1194:41-58. doi: 10.1007/978-3-030-32622-7_4. PMID: 32468522.
<https://pubmed.ncbi.nlm.nih.gov/32468522/>

3. Vlachakis, D., Armaos, A. & Kossida, S. Advanced Protein Alignments Based on Sequence, Structure and Hydropathy Profiles; The Paradigm of the Viral Polymerase Enzyme. *Math.Comput.Sci.* **11**, 197–208 (2017).
<https://doi.org/10.1007/s11786-016-0287-8>
4. Geary B, Peat E, Dransfield S, Cook N, Thistlethwaite F, Graham D, Carter L, Hughes A, Krebs MG, Whetton AD. Discovery and Evaluation of Protein Biomarkers as a Signature of Wellness in Late-Stage Cancer Patients in Early Phase Clinical Trials. *Cancers (Basel)*. 2021 May 18;13(10):2443. doi: 10.3390/cancers13102443. PMID: 34069985; PMCID: PMC8157875.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8157875/>