Πανεπιστήμιο Δυτικής Αττικής

Σχολή Μηχανικών

Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών

Θεωρία Γραφημάτων και Εφαρμογές

Απαλλακτική εργασία

Εφαρμογές της Θεωρίας Γραφημάτων στο Πρόβλημα Αναδίπλωσης Πρωτεϊνών

Ονοματεπώνυμο: Άγγελος Τζώρτζης
Αριθμός Μητρώου: ice18390094
Εξάμηνο: 12ο
Πρόγραμμα Σπουδών: Πα.Δ.Α.

Περιεχόμενα

Σύντομη Περιγραφή Θέματος

Η αναδίπλωση των πρωτεΐνών είναι η φυσική διαδικασία με την οποία μία πρωτεΐνη, μετά την σύνθεση της από ένα ριβόσωμα ως μία γραμμική αλυσίδα αμινοξέων, μετατρέπεται από μία ασταθή τυχαία σπείρα σε μία πιο οργανωμένη τρισδιάστατη δομή. Αυτή η δομή επιτρέπει την πρωτεΐνη να γίνει βιολογικά λειτουργική.

Το πρόβλημα της αναδίπλωσης πρωτεϊνών θέτει το ερώτημα του πώς η αλληλουχία των αμινοξέων μιας πρωτεΐνης καθορίζει την τρισδιάστατη ατομική δομή της. Η επίλυση αυτού του προβλήματος είναι σημαντικός στόχος της υπολογιστικής βιολογίας καθώς θα αντικαταστήσει τα αργά και ακριβά πειράματα βιολογίας με ταχύτερες και φθηνότερες προσομοιώσεις στους υπολογιστές.

Σύντομη Επισκόπηση Διαφορετικών Προσεγγίσεων

Διάφορες προσεγγίσεις για την εφαρμογή της θεωρίας γραφημάτων για την αναδίπλωση των πρωτεϊνών:

- **Αναπαράσταση με χρήση γραφήματος:** Η τρισδιάστατη δομή μιας πρωτεΐνης μπορεί να μοντελοποιηθεί ως χάρτης όπου οι κόμβοι αντιπροσωπεύουν τα αμινοξέα και οι ακμές τις συνδέσεις μεταξύ τους. Απλοποιώντας τον τρισδιάστατος χώρο σε γράφημα, μπορούμε να εφαρμόσουμε αλγόριθμους (π.χ. graph clustering, community detection) που μπορούν να προβλέψουν δομικά χαρακτηριστικά, και στην συνέχεια τις τελικές δομές των πρωτεϊνών.
- **Ελαχιστοποίηση ενέργειας μέσω αναζήτησης γραφημάτων:** Η αναδίπλωσης των πρωτεϊνων καθοδηγείται από τον στόχο της ελαχιστοποίησης της ενέργειας. Οι κόμβοι αντιπροσωπεύουν πιθανές διαμορφώσεις πρωτεϊνων και οι ακμές αναπαριστούν μεταβάσεις μεταξύ αυτών με τις σχετικές μεταβολές στην ενέργεια. Αλγόριθμοι αναζήτησης γράφου μπορούν να εξερευνήσουν το τοπίο ενέργειας και να βρουν την διαμόρφωση με την χαμηλότερη ενέργεια. Διασχίζοντας αποτελεσματικά

αυτό το γράφημα, μπορεί να προσδιοριστεί η βέλτιστη αναδιπλωμένη δομή.

- **Μοντέλα κατάστασης Markov:** Αναπαράσταση της διαδικασίας αναδίπλωσης ώς αλυσίδα Markov όπου οι κόμβοι είναι διαφορετικές διαμορφωτικές καταστάσεις και οι ακμές αντιπροσωπεύουν πιθανοτικές μεταβάσεις μεταξύ καταστάσεων. Χρησιμοποιώντας αλγορίθμους όπως ο PageRank, οι ερευνητές μπορούν να προσδιορίσουν τις πιο πιθανές διαδρομές αναδίπλωσης και τις σταθερές ενδιάμεσες δομές. Αυτό το στοχαστικό μοντέλο βοηθά στην προσομοίωση του τρόπου με τον οποίο μια πρωτεΐνη αναδιπλώνεται με την πάροδο του χρόνου, προσδιορίζοντας την πιο πιθανή τελική δομή.

Αυτές οι προσεγγίσεις που βασίζονται σε γράφους επιτρέπουν την αποτελεσματική μοντελοποίηση, πρόβλεψη και ανάλυση των διαδικασιών αναδίπλωσης των πρωτεϊνών.

Περιγραφή Του Προβλήματος

Δυσκολία για τον προσδιορισμό των πρωτεϊνικών δομών

Εδώ και αρκετές δεκαετίες είναι δυνατός ο προσδιορισμός των δομών των πρωτεϊνών σε ατομικά επίπεδα λεπτομέρειας, με τη χρήση πειραματικών μεθόδων. Ωστόσο, παρά την πρόοδο που έχει σημειωθεί στις πειραματικές τεχνικές, ο πειραματικός προσδιορισμός της πρωτεϊνικής δομής είναι εγγενώς χρονοβόρος και επίπονος και περιλαμβάνει πολλά στάδια αντιμετώπισης προβλημάτων. Ως αποτέλεσμα, ο αριθμός των γνωστών πρωτεϊνικών αλληλουχιών ήταν πάντα μεγαλύτερος από τον αριθμό των διαθέσιμων πρωτεϊνικών δομών. Τον Ιανουάριο του 2024, ο αριθμός των πειραματικών δομών που ήταν διαθέσιμες στην PDB ήταν λίγο πάνω από 215.000, σε σύγκριση με τις εντυπωσιακές ~250 εκατομμύρια πρωτεϊνικές αλληλουχίες που είναι διαθέσιμες μέσω του UniProt. Επιπλέον, το χάσμα έχει διευρυνθεί σημαντικά τα τελευταία χρόνια, χάρη στις προόδους στην αλληλούχιση του DNA. Ο πειραματικός προσδιορισμός της δομής δεν μπορεί να συμβαδίσει με

τον ρυθμό της αλληλούχισης, οπότε έχουν υπάρξει εκατοντάδες εκατομμύρια γνωστές πρωτεΐνες με άγνωστες δομές. Επομένως, είναι εξαιρετικά σημαντικό να προσδιοριστούν οι δομές αυτών των πρωτεϊνών με ακρίβεια και ταχύτητα, προκειμένου να μειωθεί το χάσμα μεταξύ των αλληλουχιών και των διαθέσιμων δομών.

Πρόβλημα της πρωτεϊνικής αναδίπλωσης

Το πρόβλημα της πρωτεϊνικής αναδίπλωσης περιλαμβάνει δύο αλληλένδετες προκλήσεις, την κατανόηση της διαδικασίας αναδίπλωσης της πρωτεϊνικής αλυσίδας και την ακριβή πρόβλεψη της τελικής αναδιπλωμένης δομής μιας πρωτεΐνης. Το 1972 ο Christian Anfinsen πρότεινε ότι,, η δομή μιας πρωτεΐνης καθορίζεται από την αλληλουχία των αμινοξέων που την αποτελούν. Αυτό έμεινε γνωστό ως το δόγμα του Άνφινσεν. Η υπόθεση αυτή ήταν σημαντική, διότι πρότεινε ότι θα έπρεπε να μπορούμε να προβλέψουμε τη δομή μιας πρωτεΐνης από την αλληλουχία των αμινοξέων της. Δεκαετίες έρευνας στη δομική βιολογία έχουν δείξει έκτοτε ότι ο Άνφινσεν ήταν σε μεγάλο βαθμό σωστός.

Η υπολογιστική πρόκληση

Ωστόσο, αποδεικνύεται ότι η πρόβλεψη της πρωτεϊνικής δομής δεν είναι τόσο απλή. Αυτό οφείλεται σε μια δεύτερη έννοια που ονομάζεται παράδοξο του Levinthal. Στη δεκαετία του 1960, ο Cyrus Levinthal έδειξε ότι υπάρχει ένας πολύ μεγάλος αριθμός πιθανών διαμορφώσεων που θα μπορούσε θεωρητικά να υιοθετήσει μια πρωτεϊνική αλυσίδα. Αν μια πρωτεΐνη έπρεπε να τις εξερευνήσει όλες, θα χρειαζόταν ένα ασύλληπτο χρονικό διάστημα, συγκρίσιμο με τη διάρκεια ζωής του Σύμπαντος. Παρ' όλα αυτά, τα ευρήματα του Anfinsen ενέπνευσαν την αναζήτηση ενός αποτελεσματικού συστήματος που θα μπορούσε να προσδιορίσει αξιόπιστα την πιθανότερη εγγενή δομή μιας πρωτεΐνης, με βάση αποκλειστικά την αλληλουχία των αμινοξέων της. Αν και ήταν δύσκολο, αυτό ήταν τουλάχιστον θεωρητικά εφικτό.

Ο ρόλος της τεχνητής νοημοσύνης

Σε αυτό το σημείο έρχεται η τεχνητή νοημοσύνη. Οι σύγχρονες μέθοδοι μηχανικής μάθησης μπορούν να βοηθήσουν στον εντοπισμό πολύπλοκων

σχέσεων σε μεγάλα σύνολα δεδομένων, επιτρέποντας την πρόβλεψη πρωτεϊνικών δομών. Το κρίσιμο είναι ότι το δόγμα του Anfinsen υπονοεί ότι η πρόβλεψη της αναδιπλωμένης κατάστασης μιας πρωτεΐνης δεν απαιτεί απαραίτητα την κατανόηση της διαδικασίας αναδίπλωσης. Δηλαδή, θα πρέπει να είναι δυνατόν να προβλεφθεί το τελικό τρισδιάστατο σχήμα μιας πρωτεΐνης χωρίς να προβλεφθεί η ακολουθία των κινήσεων που οδηγεί σε αυτό το σχήμα - παρακάμπτοντας το παράδοξο του Levinthal.

Περιγραφή αλγορίθμου

Το μοντέλο AlphaFold2

Ο αλγόριθμος που θα δούμε είναι δεν είναι στην ουσία ένα αλγόριθμος αλλά ένας συνδυασμός αλγορίθμων. Το AlphaFold2 είναι ένα πρωτοποριακό μοντέλο βαθιάς μάθησης που αναπτύχθηκε από την DeepMind για την πρόβλεψη της τρισδιάστατης δομής των πρωτεϊνών με βάση αποκλειστικά τις αλληλουχίες των αμινοξέων τους. Βελτίωσε σημαντικά την ακρίβεια της πρόβλεψης της δομής των πρωτεϊνών, η οποία είναι κρίσιμη για την κατανόηση των βιολογικών διεργασιών, την ανακάλυψη φαρμάκων και τους μηχανισμούς ασθενειών.

Πως λειτουργεί το AlphaFold2

1. **Είσοδος: Ακολουθία πρωτεΐνης:** Η είσοδος στο AlphaFold2 είναι η γραμμική αλληλουχία αμινοξέων μιας πρωτεΐνης. Η ακολουθία είναι απλώς μια σειρά γραμμάτων όπου κάθε γράμμα αντιπροσωπεύει ένα από τα 20 τυποποιημένα αμινοξέα.
2. **Πολλαπλή ευθυγράμμιση ακολουθιών (MSA):**
   Το AlphaFold2 χρησιμοποιεί πολλαπλή στοίχιση ακολουθιών (MSA) για τη συλλογή εξελικτικών πληροφοριών. Συγκρίνοντας την πρωτεϊνική αλληλουχία εισόδου με σχετικές αλληλουχίες από άλλους οργανισμούς, το μοντέλο εξάγει μοτίβα, τα οποία παρέχουν ενδείξεις σχετικά με το ποια κατάλοιπα είναι πιθανό να αλληλεπιδρούν στον τρισδιάστατο χώρο. Αυτό το βήμα είναι καθοριστικής σημασίας, επειδή οι πρωτεΐνες που σχετίζονται εξελικτικά τείνουν να υιοθετούν παρόμοιες δομές.

3. **Δίκτυο Evoformer (μετασχηματιστής):** Ο πυρήνας του AlphaFold2 είναι ο Evoformer, μια αρχιτεκτονική βαθιού νευρωνικού δικτύου που βασίζεται στο μοντέλο Transformer. Επεξεργάζεται δύο πληροφορίες:
    - Πληροφορίες MSA: Τα εξελικτικά δεδομένα από την διαδικασία του MSA, που βοηθάνε το μοντέλο να κατανοήσει τα μοτίβα στην εξέλιξη της πρωτεϊνικής αλληλουχίας και να ανιχνεύσει ποια τμήματα της αλληλουχίας μπορούν να αναδιπλωθούν μαζί.
    - Πληροφορίες pairwise residue: Το AlphaFold2 μοντελοποιεί τις σχέσεις μεταξύ ζευγών αμινοξέων (καταλοίπων) στην πρωτεϊνική ακολουθία, που αποτυπώνει τις σχέσεις μεταξύ ζευγών καταλοίπων (πόσο πιθανό είναι να αλληλεπιδράσουν στην αναδιπλωμένη πρωτεΐνη). Αυτές οι σχέσεις είναι κρίσιμες για τον προσδιορισμό της τρισδιάστατης δομής της πρωτεΐνης
4. **Μονάδα Δομής:** Αφού προβλεφθούν οι αλληλεπιδράσεις ανά ζεύγη, το AlphaFold2 χρησιμοποιεί μια Ενότητα Δομής για να προβλέψει τις τελικές τρισδιάστατες συντεταγμένες κάθε ατόμου στην πρωτεΐνη. Υπολογίζεται ένας χάρτης που προβλέπει τις αποστάσεις μεταξύ ζευγών καταλοίπων. Αυτός ο χάρτης βοηθά το μοντέλο να καθοδηγήσει τον προσδιορισμό της χωρικής διάταξης των καταλοίπων. Επίσης το μοντέλο προβλέπει τις ακριβείς θέσεις των ατόμων στον τρισδιάστατο χώρο βελτιώνοντας επαναληπτικά αυτές τις θέσεις για να διασφαλίσει ότι είναι φυσικά ρεαλιστικές. Αυτό περιλαμβάνει την ικανοποίηση γεωμετρικών περιορισμών (π.χ. μήκη και γωνίες δεσμών) για την παραγωγή μιας φυσικά αληθοφανους δομής.
5. **Μηχανισμός ανακύκλωσης:** Μια σημαντική καινοτομία στο AlphaFold2 είναι ο μηχανισμός ανακύκλωσης. Μετά την αρχική πρόβλεψη της πρωτεϊνικής δομής, το μοντέλο μπορεί να τροφοδοτήσει την ίδια την έξοδό του πίσω στον εαυτό του για να βελτιώσει επαναληπτικά την πρόβλεψη. Αυτός ο βρόχος ανατροφοδότησης βοηθά στη βελτίωση της τελικής δομής επιτρέποντας στο μοντέλο να μαθαίνει από τις δικές του ενδιάμεσες προβλέψεις.
6. **Εκπαίδευση άκρη σε άκρη:** Το AlphaFold2 εκπαιδεύεται από άκρο σε άκρο, πράγμα που σημαίνει ότι ολόκληρη η διαδικασία - από την είσοδο της ακολουθίας μέχρι την έξοδο της δομής - μαθαίνεται από το μοντέλο

σε έναν ενιαίο αγωγό. Κατά τη διάρκεια της εκπαίδευσης, το AlphaFold2 βελτιστοποιεί τις προβλέψεις του συγκρίνοντας με πρωτεϊνικές δομές από βάσεις δεδομένων όπως η Τράπεζα Δεδομένων Πρωτεϊνών (PDB). Το μοντέλο χρησιμοποιεί μια συνάρτηση απωλειών που λαμβάνει υπόψη τόσο την ακρίβεια της προβλεπόμενης τρισδιάστατης δομής όσο και τις ενδιάμεσες προβλέψεις, όπως οι χάρτες απόστασης ανά ζεύγη. Αυτό διασφαλίζει ότι το μοντέλο μαθαίνει να παράγει ακριβείς δομές σε όλα τα στάδια.

7. **Έξοδος: Δομή πρωτεΐνης:** Το τελικό αποτέλεσμα είναι οι τρισδιάστατες ατομικές συντεταγμένες της πρωτεΐνης, που αντιπροσωπεύουν την αναδιπλωμένη δομή της. Το AlphaFold2 παρέχει όχι μόνο τη δομή αλλά και μια εκτίμηση του διαστήματος εμπιστοσύνης για κάθε περιοχή της πρωτεΐνης, η οποία βοηθά τους ερευνητές να αξιολογήσουν την αξιοπιστία της πρόβλεψης.

Πώς Εφαρμόζεται η Θεωρία Γραφημάτων στο μοντέλο AlphaFold2

Το AlphaFold 2 χρησιμοποιεί γραφήματα ως μέρος της αρχιτεκτονικής του νευρωνικού δικτύου του για την αναπαράσταση των πρωτεϊνικών δομών και των αλληλοεπιδράσεων τους. Ειδικότερα, οι γράφοι είναι κρίσιμοι για τη μοντελοποίηση των χωρικών σχέσεων μεταξύ των αμινοξέων σε μια πρωτεϊνική ακολουθία, η οποία είναι απαραίτητη για την ακριβή πρόβλεψη της τρισδιάστατης δομής της.

Τομείς που το AlphaFold2 χρησιμοποιεί γράφους:

1. **Αναπαράσταση της πρωτεϊνικής δομής:** Το AlphaFold 2 αναπαριστά τις πρωτεΐνες χρησιμοποιώντας ένα γράφημα καταλοίπων-υπολειμμάτων. Σε αυτό το γράφημα, κάθε κόμβος αντιστοιχεί σε ένα αμινοξύ (κατάλοιπο) της πρωτεΐνης και οι ακμές μεταξύ των κόμβων αντιπροσωπεύουν τις σχέσεις ή τις αποστάσεις μεταξύ των καταλοίπων. Αυτή η αναπαράσταση με βάση το γράφημα βοηθά στην αποτύπωση των γεωμετρικών σχέσεων που είναι απαραίτητες για την αναδίπλωση της πρωτεΐνης στον τρισδιάστατο χώρο.
2. **Νευρωνικά δίκτυα γραφημάτων (GNN):** Το AlphaFold 2 χρησιμοποιεί νευρωνικά δίκτυα γράφων (GNN) για την επεξεργασία του γράφου καταλοίπων-υπολειμμάτων. Αυτά τα δίκτυα βοηθούν στην ενημέρωση

των ενσωματώσεων των καταλοίπων με τη διαβίβαση μηνυμάτων στις ακμές του γράφου, επιτρέποντας στο μοντέλο να μάθει το δομικό και χωρικό πλαίσιο κάθε αμινοξέος με βάση τους γείτονές του.

3. **Evoformer module:** Ο Evoformer βασίζεται σε μεγάλο βαθμό σε αναπαραστάσεις βασισμένες σε γράφους. Ο Evoformer επεξεργάζεται τόσο μια πολλαπλή στοίχιση ακολουθίας (MSA) όσο και την αναπαράσταση ζεύγους καταλοίπων-υπολειμμάτων (σε μορφή γραφήματος) για να βελτιώσει επαναληπτικά τις πληροφορίες σχετικά με την ακολουθία και τη χωρική της διάταξη. Η MSA καταγράφει εξελικτικές πληροφορίες, ενώ ο γράφος καταγράφει γεωμετρικές πληροφορίες.
4. **Μηχανισμοί προσοχής σε γράφους:** AlphaFold 2 εφαρμόζει μηχανισμούς προσοχής τόσο σε αναπαραστάσεις ακολουθίας όσο και σε αναπαραστάσεις ανά ζεύγη (γράφοι). Αυτή η προσοχή λειτουργεί πάνω στους κόμβους (κατάλοιπα) και βοηθά το μοντέλο να εστιάσει σε συγκεκριμένες σχέσεις μεταξύ αμινοξέων που είναι σημαντικές για την αναδίπλωση της πρωτεΐνης.

Σκοπός των γραφημάτων στο AlphaFold 2

- **Αποτύπωση χωρικών σχέσεων:** Οι γραφικές παραστάσεις επιτρέπουν στο AlphaFold 2 να μοντελοποιεί τον τρόπο με τον οποίο κάθε κατάλοιπο τοποθετείται σε σχέση με άλλα στον τρισδιάστατο χώρο, πράγμα που είναι ζωτικής σημασίας για τον καθορισμό της τελικής δομής.
- **Χειρισμός δεδομένων μεγάλης κλίμακας:** Με τη χρήση γραφημάτων, το AlphaFold 2 μπορεί να διαχειριστεί αποτελεσματικά τις μεγάλες ποσότητες δεδομένων που εμπλέκονται στην αναπαράσταση των πρωτεϊνικών ακολουθιών και των πιθανών διαμορφώσεών τους.
- **Αποτελεσματική μεταβίβαση μηνυμάτων:** Το νευρωνικό δίκτυο γραφημάτων (GNN) επιτρέπει την αποτελεσματική διαβίβαση μηνυμάτων μεταξύ καταλοίπων, επιτρέποντας στο μοντέλο να συγκεντρώνει δομικές πληροφορίες από γειτονικούς κόμβους για τη βελτίωση των προβλέψεων.

Αυτή η προσέγγιση που βασίζεται στους γράφους, σε συνδυασμό με τα εξελικτικά δεδομένα, είναι ένας από τους λόγους για τους οποίους το AlphaFold2 επιτυγχάνει τόσο μεγάλη ακρίβεια στην πρόβλεψη πρωτεϊνικών δομών.

Υποθέσεις και περιορισμοί

Τι μπορεί να κάνει το AlphaFold2

Το AlphaFold2 εκπαιδεύτηκε αρχικά σε μεμονωμένες πρωτεϊνικές αλυσίδες, οπότε είναι εξαιρετικό στην πρόβλεψη των δομών τους. Αργότερα, μια επέκταση του AlphaFold2 εκπαιδεύτηκε ειδικά για την πρόβλεψη συμπλόκων πρωτεϊνών: αυτή η έκδοση είναι πλέον γνωστή ως AlphaFold-Multimer. Μπορεί να προβλέψει τις δομές των πρωτεϊνικών συμπλόκων που αποτελούνται από πολλά αντίγραφα της ίδιας αλυσίδας (ομο-πολυμερή, όπως διμερή και εξαμερή), καθώς και εκείνων που αποτελούνται από πολλές διαφορετικές πρωτεϊνικές αλυσίδες (ετερο-πολυμερή). Είναι σημαντικό ότι το AlphaFold2 δεν αναπαράγει απλώς γνωστές πρωτεϊνικές δομές. Ανεξάρτητοι ερευνητές έχουν δείξει ότι το AlphaFold2 μπορεί να προβλέψει δομές που δεν έχουν εμφανιστεί ποτέ στο PDB, δηλαδή νέες πρωτεϊνικές πτυχές. Φυσικά, το σύστημα δεν μπορεί να προβλέψει δομές αλληλουχιών που δεν υπάρχουν σε μία καθορισμένη διαμόρφωση στη φύση. Τέτοιες περιοχές είναι πραγματικά δυναμικές και δεν έχουν μια σταθερή δομή για να προβλεφθεί. Ωστόσο, η μετρική τοπικής εμπιστοσύνης του AlphaFold2 (pLDDT) παρουσιάζει ισχυρή συσχέτιση με την εγγενή αταξία, καθιστώντας το AlphaFold ένα κορυφαίο εργαλείο για τον εντοπισμό των περιοχών εκτός τάξης.

Που δυσκολεύεται το AlphaFold2

Το AlphaFold δεν είναι ευαίσθητος σε σημειακές μεταλλάξεις που αλλάζουν ένα μόνο κατάλοιπο, λόγω της αλλαγής της αλληλουχίας του DNA. Αυτό οφείλεται στην έλλειψη δεδομένων σχετικά με την επίδραση των παραλλαγών, σε συνδυασμό με την εστίαση του AlphaFold2 σε μοτίβα σε αντίθεση με τον υπολογισμό φυσικών δυνάμεων. Για τους ίδιους λόγους, το AlphaFold2 είναι επίσης λιγότερο ακριβές στην πρόβλεψη των δομών που σχετίζονται με εξαιρετικά μεταβλητές αλληλουχίες, όπως αυτές των μορίων του ανοσοποιητικού

συστήματος, όπως τα αντισώματα. Το AlphaFold2 δυσκολεύεται να προβλέψει τις δομές των «ορφανών» πρωτεϊνών - εκείνων με λίγους στενούς συγγενείς - καθώς λειτουργεί με την εξαγωγή σχέσεων μεταξύ των πρωτεϊνικών αλληλουχιών. Εάν δεν υπάρχουν αρκετές αλληλουχίες για σύγκριση, το AlphaFold2 παράγει συχνά προβλέψεις χαμηλής ποιότητας με χαμηλές βαθμολογίες εμπιστοσύνης. Το πρόβλημα αυτό επιδεινώνεται εάν οι λίγες συγγενικές αλληλουχίες δεν έχουν γνωστές δομές στην PDB. Από την άλλη πλευρά, αυτή η μεθοδολογική επιλογή σημαίνει ότι το AlphaFold2 μπορεί συχνά να προβλέψει τη δομή των πρωτεϊνών, ακόμη και αν δεν υπάρχουν γνωστές συγγενείς δομές στο PDB, υπό την προϋπόθεση ότι μια αλληλουχία έχει χιλιάδες συγγενείς. Οι πρωτεΐνες υφίστανται δομικές αλλαγές όταν εκτελούν τις λειτουργίες τους. Ωστόσο, αυτές οι διαφορετικές διαμορφώσεις περιγράφονται μόνο για μια μειοψηφία των πρωτεϊνών στην PDB. Το AlphaFold2 δεν καταγράφει τέτοιες αλλαγές διαμόρφωσης, καθώς σχεδιάστηκε για να προβλέπει στατικές δομές, δηλαδή δομικά στιγμιότυπα. Ωστόσο, οι ερευνητές έχουν διαπιστώσει ότι, εφαρμόζοντας ορισμένα τεχνάσματα, μπορεί κανείς να αναγκάσει το AlphaFold2 να παράγει μια διαφορετική διαμόρφωση της πρωτεΐνης.

Τι δεν μπορεί να κάνει το AlphaFold2

Το AlphaFold2 δεν γνωρίζει άλλα μόρια που αλληλεπιδρούν με τις πρωτεΐνες, όπως τα νουκλεϊκά οξέα, τους συμπαράγοντες μικρών μορίων, τα ιόντα και άλλα μη πρωτεϊνικά συστατικά. Ομοίως, το AlphaFold2 δεν σχεδιάστηκε για να μοντελοποιεί μετα-μεταφραστικές τροποποιήσεις ή για να μοντελοποιεί δομές ελεύθερων νουκλεϊκών οξέων. Ωστόσο, το AlphaFold2 μπορεί συχνά να προβλέψει μια μορφή μιας πρωτεΐνης συνδεδεμένη με ligand ή ιόν, ακόμη και απουσία του πραγματικού ligand/ιόντος.Το AlphaFold2 δεν γνωρίζει το επίπεδο της μεμβράνης. Κατά συνέπεια, δεν μπορεί να μοντελοποιήσει σωστά τους σχετικούς προσανατολισμούς των διαμεμβρανικών περιοχών και άλλων πρωτεϊνικών περιοχών των μεμβρανικών πρωτεϊνών. Ωστόσο, όπως σημειώνεται, το AlphaFold2 προειδοποιεί τους χρήστες για τις αβεβαιότητες αποδίδοντας χαμηλή βαθμολογία εμπιστοσύνης. Αυτό το συγκεκριμένο ζήτημα αντικατοπτρίζεται συνήθως στη βαθμολογία του προβλεπόμενου σφάλματος ευθυγράμμισης (PAE).

Πίνακας που συνοψίζει τις δυνατότητες του AlphaFold2

AlphaFold2 predicts	AlphaFold2 struggles to predict	AlphaFold2 doesn't predict
• Single protein chains • Protein multimers • Multisubunit protein-protein complexes	• Multiple conformations for the same sequence • Effects of point mutations • Antigen-antibody interactions	• Protein-DNA and protein-RNA complexes • Nucleic acid structure • Ligand and ion binding • Post-translational modifications • Membrane plane for transmembrane domains

Περιορισμοί λόγω εφαρμογής Θεωρία Γραφημάτων

- **Στατική αναπαράσταση:** Δυσκολεύεται με δυναμικές συμπεριφορές πρωτεϊνών, όπως οι διαμορφωτικές αλλαγές.
- **Σύμπλοκα πολλαπλών πρωτεϊνών:** Οι γράφοι του είναι λιγότερο κατάλληλοι για την πρόβλεψη αλληλεπιδράσεων μεταξύ πολλαπλών πρωτεϊνών.
- **Μετα-μεταφραστικές τροποποιήσεις:** Παραβλέπει χημικές τροποποιήσεις όπως η φωσφορυλίωση, οι οποίες μπορούν να μεταβάλουν τη δομή των πρωτεϊνών.
- **Αλληλεπιδράσεις μεγάλης εμβέλειας:** Η καταγραφή των σχέσεων απομακρυσμένων καταλοίπων είναι πρόκληση για μεγάλες πρωτεΐνες.

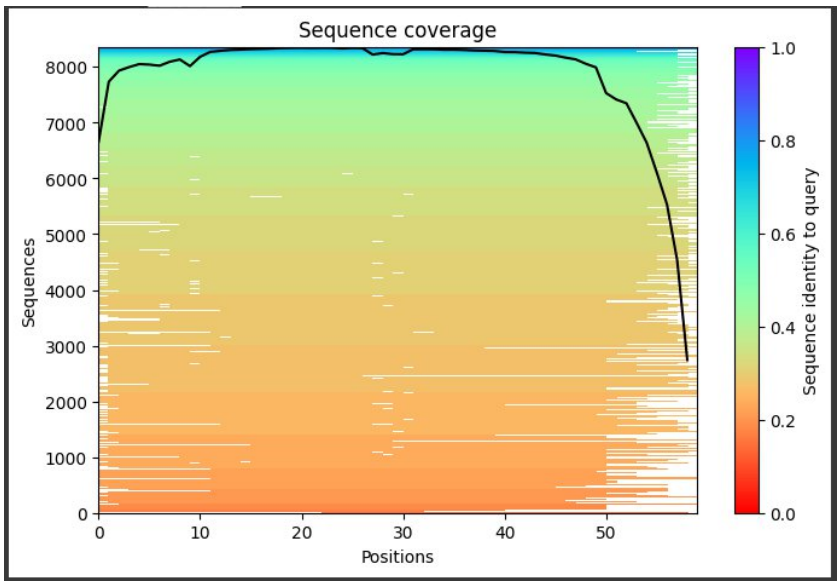
- **Περιοχές με αταξία:** Υποθέτει ότι οι δομές είναι καλά καθορισμένες, καθιστώντας την λιγότερο ακριβή για εγγενώς ατακτοποιήτες περιοχές.

Αυτά τα προβλήματα που σχετίζονται με τα γραφήματα περιορίζουν την απόδοση του AlphaFold 2 σε πολύπλοκα σενάρια πρωτεϊνών.

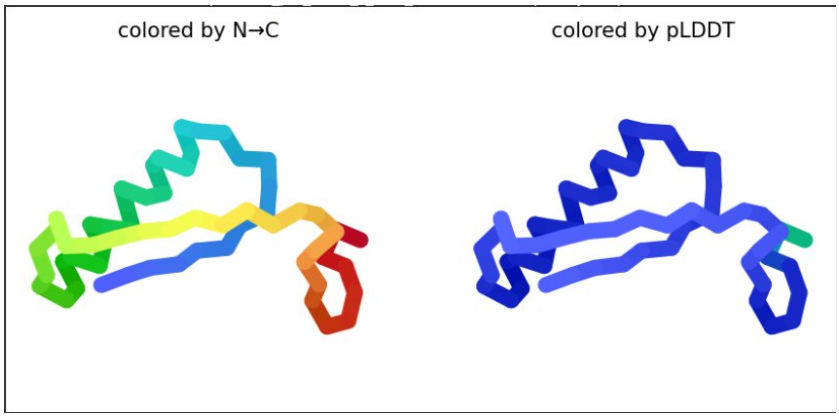
Παραδείγματα και Αποτελέσματα

Αλληλουχία εισόδου:
PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGEL
ASK

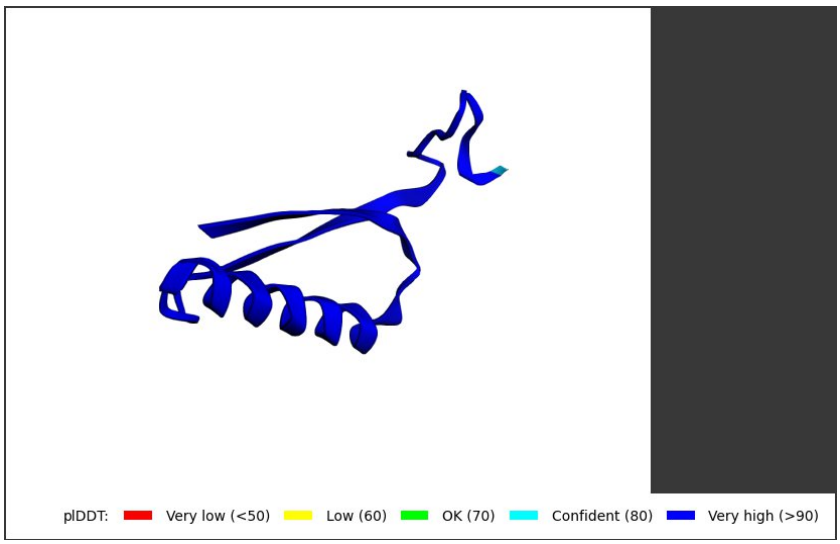
Διάγραμμα Sequence Coverage: Μας δείχνει πόσο καλά υποστηρίζεται το αποτέλεσμα για κάθε τμήμα της πρωτεΐνης από τα εξελιτικά δεδομένα.



N->C και pLDDT:
Colored by N->C αναφέρεται στον χρωματισμό μιας πρωτεΐνης βάση την σειρά τον αμινοξέων της από τό N-terminus (αρχή, κόκκινο) μέχρι τό C-terminus (τέλος, μπλέ).
Colored by pLDDT μας δείχνει τον βαθμό εμπιστοσύνης σε κάθε τμήμα πρωτεΐνης με χρώματα από μπλέ εώς κόκκινο (μεγαλύτερη εμπιστοσύνη-μικρότερη εμπιστοσύνη.

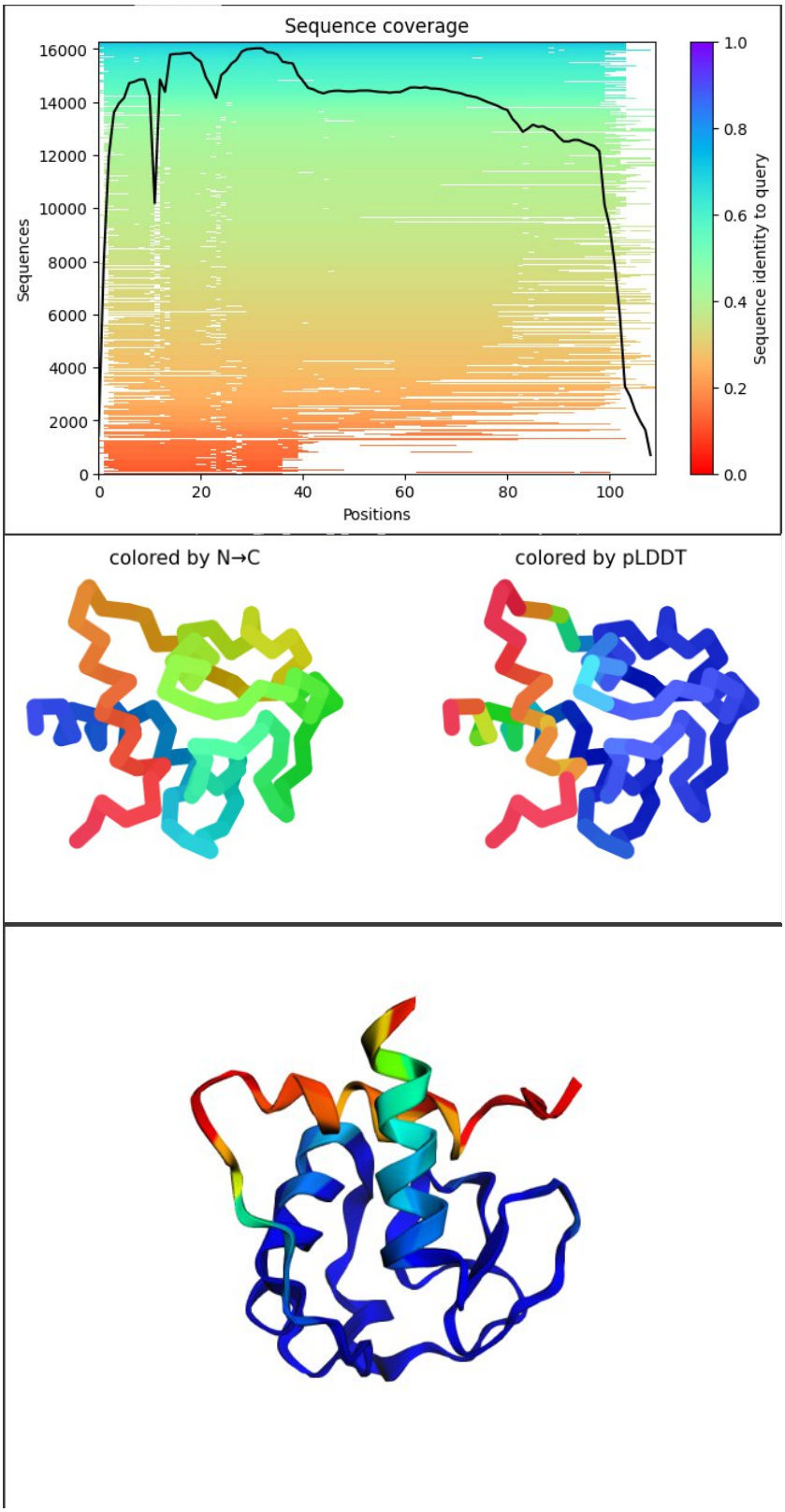


Τρισδιάστατη απεικόνιση της πρωτεΐνης: απεικόνιση της τελικής μορφής της πρωτεΐνης. Στο κάτω μέρος φαίνονται οι βαθμοί εμπιστοσύνης για το τελικό αποτέλεσμα.



Αλληλουχία εισόδου:

MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFSYTDA
NKNKGITWGEETLMEYLENPKKYIPGTKMAFGGLKKEGANNKVGTAFVPFTYT
DANK



Στην πρώτη πρωτεΐνη έχουμε μεγάλο βαθμό εμπιστοσύνης σε όλα τα μέρη της άρα η πρόβλεψη μας έχει μεγάλο βαθμό ακρίβειας, και πιθανό να αντιπροσωπεύει την αναδίπλωση που θα γινόταν στον φυσικό κόσμο.

Βλέπουμε ότι για την δεύτερη πρωτεΐνη ο βαθμός εμπιστοσύνης στα άκρα τής είναι μικρός οπότε πιθανόν αυτά τα σημεία να μήν αναδιπλώνονται έτσι εάν κάναμε πειραματική δοκιμή για την αναδίπλωση αυτής της πρωτεΐνης. Αυτή η μικρή εμπιστοσύνη πιθανόν οφείλεται σε μικρό μέγεθος δεδομένων για συγκεκριμένα αμινοξέα με αποτέλεσμα να μην έχει εκπαιδευτεί κατάλληλα το μοντέλο.

Συμπέρασμα: Το μοντέλο AlphaFold2 είναι αρκετό γρήγορο και ακριβές στις προβλέψεις του. Επίσης καθώς μας εμφανίζει που πιθανόν έχουμε λάθη στις προβλέψεις μας μπορούμε να εμπιστευτούμε τα αποτελέσματα του και να τα αναλύσουμε όπως χρειαζόμαστε για το έργο μας. Η εφαρμογή της θεωρίας γραφημάτων κρίνεται απαραίτητη καθώς βοηθάει στην γρήγορη και ορθή παραγωγή αποτελεσμάτων και οι αδυναμίες που προκύπτουν δεν εμποδίζουν σε μεγάλο βαθμό το αποτέλεσμα.

Αναφορές

[1] Wikipedia (2024), Protein Folding: https://en.wikipedia.org/wiki/Protein_folding

[2] National Library of Medicine (2008), The ProteinFolding Problem: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2443096/

[3] European Bioinformatics Institution (2024), AlphaFold: https://www.ebi.ac.uk/training/online/courses/alphafold/

[4] The Roots of Progress (2020), What is the “protein folding problem”? A brief Explanation: https://blog.rootsofprogress.org/alphafold-protein-folding-explainer

[5] National Library of Medicine (2011), Applications of graph theory in protein structure identification:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3289078/

[6] National Library of Medicine (2011), AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8592092/

[7] PI IP LAW (2022) , AlphaFold 2: Attention Mechanism for Predicting 3D Protein Structures: https://piip.co.kr/en/blog/AlphaFold2_Architecture_Improvements

[8] European Bioinformatics Institution (2024), AlphaFold Protein Structure Database: https://alphafold.ebi.ac.uk/

[9] Github (2024): Open Source Code for AlphaFold: https://github.com/google-deepmind/alphafold

[10] Google Colab (2024), AlphaFold2.ipynb: https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb