ΠΡΑΚΤΙΚΗ ΑΣΚΗΣΗ

Πρόβλεψη ενδοκυτταρικής θέσης πρωτεϊνών με χρήση αλγορίθμων μηχανικής μάθησης

Πολίτη Θεοφύλακτη | ΑΕΜ 2735

Υπεύθυνος: Ψωμόπουλος Φώτης

Δημοκρίτειο Πανεπιστήμιο Θράκης | Τμήμα Μοριακής Βιολογίας και Γενετικής (ΔΠΘ | ΜΒΓ)

Ινστιτούτο Εφαρμοσμένων Βιοεπιστημών | Εθνικό Κέντρο Έρευνας και Τεχνολογικής Ανάπτυξης (ΙΝΕΒ | ΕΚΕΤΑ)

Institute of Applied Biosciences | Centre for Research and Technology Hellas (INAB|CERTH)

Περιεχόμενα

Εισαγωγή:

Ο όγκος των βιολογικών δεδομένων τα τελευταία χρόνια έχει αυξηθεί εκθετικά, κυρίως λόγω της προόδου στις τεχνικές αλληλούχισης νέας γενιάς αλλά και τη συνεχή εξέλιξη της βιοπληροφορικής. Αυτή η αύξηση έχει ως επακόλουθο την ανάγκη για ανάλυση, συνδυασμό και ερμηνεία των συνόλων βιολογικών δεδομένων με σκοπό την εξαγωγή γνώσης και χρήσιμων πληροφοριών, κάτι που μπορεί να επιτευχθεί με υπολογιστικά εργαλεία όπως η μηχανική μάθηση.

Η **Μηχανική Μάθηση** (Machine Learning-ML) είναι κλάδος της Τεχνητής Νοημοσύνης (Artificial Intelligence-AI) που χρησιμοποιεί αλγορίθμους μάθησης οι οποίοι 'μαθαίνουν' μοτίβα, χρησιμοποιώντας γνωστά δεδομένα, και έπειτα έχουν την ικανότητα να προβλέψουν χαρακτηριστικά με βάση αυτά τα μοτίβα σε νέα, άγνωστα δεδομένα. Δηλαδή, προσομοιάζουν τον τρόπο που μαθαίνουν οι άνθρωποι και αυξάνουν σταδιακά την απόδοση τους. Οι προβλέψεις αυτές δεν προκύπτουν από κάποιο ρητά προσαρμοσμένο αλγόριθμο που έχει δημιουργηθεί από τον προγραμματιστή, αλλά από πολύπλοκα μαθηματικά μοντέλα - όπως κρυμμένα νευρωνικά δίκτυα - τα οποία εντοπίζουν και γενικεύουν μοτίβα στα δεδομένα. Επομένως η απόδοση του αλγορίθμου μάθησης βασίζεται στα δεδομένα που θα χρησιμοποιηθούν. Τα χαρακτηριστικά των αλγορίθμων μηχανικής μάθησης τα καθιστούν πολύ χρήσιμα εργαλεία στην ερμηνεία βιολογικών δεδομένων, καθώς επιτρέπουν την αυτοματοποίηση και την μαζική ανάλυση ενώ ταυτόχρονα μπορούν να αντικαθιστούν ακριβά και απαιτητικά πρωτόκολλα εργαστηρίου.

Υπάρχουν πολλοί τύποι μηχανικής μάθησης, αλλά σε αυτή την εργασία αναπτύχθηκε αλγόριθμος ταξινόμησης με **Εποπτευόμενη Μάθηση** (Supervised Learning). Σε αυτόν τον τύπο μάθησης, ο αλγόριθμος εκπαιδεύεται σε δεδομένα τα οποία διαθέτουν ζεύγη εισόδων-εξόδων (x,y) όπου το x αντιπροσωπεύει τα **χαρακτηριστικά** (features) και το y την αντίστοιχη **ετικέτα** (labels), η οποία αποτελεί σημείο αναφοράς, ο αλγόριθμος δημιουργεί συσχέτιση ανάμεσα στα x,y. Έπειτα χρησιμοποιείται σε νέα άγνωστα δεδομένα x' και προβλέπει την ετικέτα τους y'.

Λόγω της πολυπλοκότητας, της υπολογιστικής ισχύς και του μεγέθους των δεδομένων που χρειάζεται για την εκπαίδευση μοντέλων, έχουν δημιουργηθεί **προεκπαιδευμένα μοντέλα** (pre-trained Models) τα οποία έχουν εκπαιδευτεί σε τεράστια σετ δεδομένων -για παράδειγμα σε ολόκληρο το γονιδίωμα του ανθρώπου - και έχουν αποθηκευτεί σε βιβλιοθήκες για χρήση σε άλλες εργασίες. Στις επιμέρους εργασίες χρησιμοποιούνται αυτά τα μοντέλα και πραγματοποιείται **βελτιστοποίηση** (fine-tuning) για να 'ταιριάζουν' στα εκάστοτε δεδομένα. Συγκεκριμένα, στην

παρούσα εργασία χρησιμοποιήθηκε το προεκπαιδευμένο μοντέλο "**RaphaelMourad/Mistral-Peptide-v1-15M"** από τη βιβλιοθήκη Transformers, το οποίο έχει εκπαιδευτεί πάνω σε χιλιάδες αλληλουχίες πεπτιδίων[6].

Η ενδοκυτταρική θέση των πρωτεϊνών είναι πολύ σημαντική για τη σωστή λειτουργία τους και τις αλληλεπιδράσεις με άλλες πρωτεΐνες, ενώ σε πολλές περιπτώσεις η λανθασμένη κατανομή τους στο κύτταρο έχει παθογόνο χαρακτήρα. Για παράδειγμα, φαίνεται ότι σε κάποιες περιπτώσεις το ογκοκατασταλτικό γονίδιο BRCA1 που φυσιολογικά βρίσκεται στον πυρήνα, μπορεί να μην περιέχει το σήμα πυρηνικού εντοπισμού λόγω μετάλλαξης, με αποτέλεσμα να παραμένει στο κυτταρόπλασμα και να χάνει την ογκοκατασταλτική του δράση προκαλώντας κακοήθειες (Wang, X., & Li, S., 2014). Συνεπώς, έχοντας αντιληφθεί την σημασία του εντοπισμού των πρωτεϊνών, και την χρησιμότητα τον μοντέλων μηχανικής μάθησης στην βιοπληροφορική, στην παρούσα εργασία έγινε εκπαίδευση μοντέλων εποπτευόμενης μηχανικής μάθησης, σε πρωτεΐνες του ανθρώπου που έχουν χαρακτηριστεί ότι ανήκουν σε συγκεκριμένες θέσεις του κυττάρου οι οποίες χρησιμοποιούνται ως ετικέτες. Έπειτα δημιουργήθηκε ένα πρόγραμμα το οποίο δέχεται πρωτεϊνικές αλληλουχίες (χαρακτηριστικό) σαν είσοδο και επιστρέφει την θέση του κυττάρου (ετικέτα) που αντιστοιχεί περισσότερο σε αυτή την αλληλουχία.

Μεθοδολογία:

> Αναλυτικές οδηγίες για την εκτέλεση του κώδικα: README.md

1. Συλλογή δεδομένων

Αρχικά συλλέχθηκαν 9.999 ανθρώπινες αλληλουχίες πρωτεϊνών από τη βάση δεδομένων πρωτεϊνών της Entrez, χρησιμοποιώντας το πακέτο BioPython. Από αυτό το σύνολο δεδομένων υπολογίστηκε το μήκος των αλληλουχιών, η διακύμανση, καθώς και το χαμηλότερο (5%) και ανώτερο (95%) όριο της κατανομής τους. Οι αλληλουχίες έξω από αυτά τα όρια δεν χρησιμοποιούνται διότι συχνά πολύ μεγάλες ή μικρές πρωτεΐνες αποτελούν χίμαιρες ή θραύσματα και δεν αντιπροσωπεύουν το σύνολο των πρωτεϊνών. Από υπολογιστική άποψη αυτές οι αλληλουχίες αποσταθεροποιούν την εκπαίδευση και στοιχίζουν στην χρήση μνήμης και την τελική απόδοση των μοντέλων.

Στη συνέχεια συλλέχθηκαν 11.615 ανθρώπινες αλληλουχίες πρωτεϊνών από τη βάση δεδομένων πρωτεϊνών της Entrez οι οποίες περιείχαν στο τίτλο τους μια από τις παρακάτω λέξεις που αντιστοιχούν σε ενδοκυτταρική τοποθεσία:

- Extracellular
- Golgi
- Lysosome
- Membrane
- Mitochondria
- Nuclear
- Peroxisome
- Reticulum
- Ribosome

Έπειτα γίνετε φιλτράρισμα για απαλλαγή των διπλασιασμένων αλληλουχιών καθώς και των αλληλουχιών που βρίσκονται εκτός των ορίων που έχουν οριστεί. Τέλος δημιουργήθηκαν 8 σετ δεδομένων το καθένα περιέχει όλες τις αλληλουχίες από μια διαφορετική τοποθεσία σημασμένες με την ετικέτα ‘1’ και ίδιο αριθμό από αλληλουχίες διάφορων άλλων διαμερισμάτων του κυττάρου με την ετικέτα ‘0’. Δηλαδή δημιουργήθηκαν 8 σετ δεδομένων για δυαδική ταξινόμηση, με

κατανομή των ετικετών 50:50 διότι η ισορροπία ανάμεσα στους δύο τύπους δεδομένων είναι πολύ σημαντική.

Τα δύο πρώτα προγράμματα που αναπτύχθηκαν υπολογίζουν τα όρια των πρωτεϊνών: το πρώτο για το πλήρες σύνολο πρωτεϊνικών αλληλουχιών, ενώ το δεύτερο για το υποσύνολο των πρωτεϊνών που εντοπίζονται σε συγκεκριμένες κυτταρικές θέσεις, αυτές οι τιμές διαφέρουν μεταξύ τους. Αυτό συμβαίνει διότι οι πρωτεΐνες που βρίσκονται εξωκυτταρικά τείνουν να έχουν μεγαλύτερο μέγεθος, επομένως αυξάνουν το ανώτατο όριο. Ο καθορισμός των ορίων του μεγέθους των αλληλουχιών προκύπτει από συλλογή πρωτεϊνικών αλληλουχιών για το πλήρες σύνολο, για να αποφευχθεί αυτή η μεροληψία.

2. Βελτιστοποίηση Μοντέλων

Τα 8 σετ δεδομένων που δημιουργήθηκαν χρησιμοποιούνται για την βελτιστοποίηση 8 μοντέλων για δυαδική ταξινόμηση πρωτεϊνικών αλληλουχιών. Οι παράμετροι που χρησιμοποιήθηκαν για την βελτιστοποίηση είναι προσαρμοσμένες στα δεδομένα μετά από υπολογισμό της καλύτερης απόδοσης ανάμεσα σε διαφορετικές εκτελέσεις του κώδικα με αλλαγές στις παραμέτρους.

Για όλα τα μοντέλα χρησιμοποιήθηκε το προεκπαιδευμένο μοντέλο Mistral-Peptide-v1-15M, το οποίο είναι σχεδιασμένο για ανάλυση πεπτιδικών αλληλουχιών. Πραγματοποιήθηκε fine-tuning με χρήση της μεθοδολογίας **LoRA (Low-Rank Adaptation)**. Το LoRA είναι μια τεχνική που επιτρέπει αποδοτική προσαρμογή πολύ μεγάλων μοντέλων σε νέες εργασίες χωρίς να χρειάζεται να εκπαιδευτούν όλοι οι παράμετροι. Αντί να αλλάξει το αρχικό μοντέλο, διατηρεί τα βάρη του και προσθέτει μικρούς πίνακες χαμηλής διάστασης (low-rank matrices) που μαθαίνουν τις νέες πληροφορίες. Έτσι μειώνεται δραστικά το υπολογιστικό κόστος και η μνήμη που απαιτείται, διατηρώντας παράλληλα υψηλή απόδοση. Κάθε μοντέλο ακολουθεί τα παρακάτω βήματα:

1. Διαχωρισμός δεδομένων: 80% εκπαίδευση(training), 10% επικύρωση(validation), 10% δοκιμή(testing).
2. Tokenization: οι πρωτεϊνικές αλληλουχίες κωδικοποιούνται σε tokens, τα οποία είναι αριθμητικές αναπαραστάσεις μικρών τμημάτων αλληλουχίας, που έπειτα συνδέονται και δημιουργούν αριθμητικές ‘αλληλουχίες’ μέχρι 1.000 αμινοξέα.
3. Ρυθμίσεις εκπαίδευσης:

- Batch size: 16
- Learning Rate: 2×10^{-5}
- Αριθμός επαναλήψεων(epochs): 20
- Early stopping: διακοπή μετά από τρείς επαναλήψεις χωρίς βελτίωση

4. Αξιολόγηση: το τελικό μοντέλο αξιολογείται στο test set και καταγράφονται μετρικές όπως accuracy, precision, recall και F1-score.

Λόγω διαφορών στον αριθμό των αλληλουχιών που προκύπτουν ανάμεσα στα 8 διαφορετικά διαμερίσματα του κυττάρου, η απόδοση των μοντέλων δεν είναι ομοιόμορφη. Οι κατηγορίες με μεγαλύτερο αριθμό αλληλουχιών (όπως *Membrane*, *Mitochondrial* και *Nuclear*) αναμένεται να δώσουν καλύτερα αποτελέσματα σε σχέση με τις μικρότερες (π.χ. *Extracellular*, *Ribosome*). Σημειώνεται ότι για την κατηγορία *Lysosome* δεν εκπαιδεύτηκε μοντέλο, καθώς δεν υπάρχει επαρκής αριθμός αλληλουχιών.

3. Πρόβλεψη ενδοκυτταρικής θέσης

Δημιουργήθηκε ένα πρόγραμμα στο οποίο ο χρήστης συμπληρώνει μια πρωτεϊνική αλληλουχία και χρησιμοποιώντας τα αποθηκευμένα πλέον μοντέλα, υπολογίζεται η πιθανότητα η πρωτεΐνη να ανήκει στην καθεμία από τις 8 ενδοκυτταρικές θέσεις. Ως έξοδο εμφανίζεται η θέση που υπολογίστηκε η μεγαλύτερη πιθανότητα καθώς και τις πιθανότητες που υπολογίστηκαν στις υπόλοιπες θέσεις. Μια ικανοποιητική πρόβλεψη θα πρέπει να έχει υψηλό σκορ πιθανότητας στην θέση που εκτυπώθηκε σαν πρόβλεψη και χαμηλό σκορ πιθανότητας σε όλες τις υπόλοιπες θέσεις.

Η έξοδος του προγράμματος φαίνεται παρακάτω:

```
## Example 2.2 execution
# Models were trained on very few data, for output example purposes.
# Peroxisome and Reticulum models were not trained or used in the localization function.

(SLP) ~\Subcellular-Localization-Prediction>python models/Localization.py
Please provide protein sequence:
MAALRRLLWPPPRVSPPLCAHQPLLGPWGRPAVTTLGLPGRPFSSREDEERAVAEAAWRRRRRWGELSVAAAAGGGLVGLVCYQLYGDPRAGSPATGRPSKSAATEPEDPPRGRGMLPIPVAAAKETVAIGRTDIE
DLDLYATSRERRFRLFASIECEGQLFMTPYDFILAVTTDEPKVAKTWKSLSKQELNQMLAETPPVWKGSSKLFRNLKEKEPHAGFRIAFNMFDTDGNEMVDKKEFLVLQEIFRKKNEKREIKGDEEKRAMLRLQLY
GYHSPTNSVLKTDAEELVSRSYWDTLRRNTSQALFSDLAERADDITSLVTDTTLLVHFFGKKGKAELNFEDFYRFMDNLQTEVLEIEFLSYSNGMNTISEEDFAHILLRYTNVENTSVFLENVRYSIPEEKGITFD
EFRSFFQFLNNLEDFAIALNMYNFASRSIGQDEFKRAVYVATGLKFSPHLVNTVFKIFDVDKDDQLSYKEFIGIMKDRLHRGFRGYKTVQKYPTFKSCLKKELHSR

Predicted type: mitochondria
All model scores: {'extracellular': 0.4101673960685573, 'golgi': 0.38762950897216797, 'membrane': 0.3010360598564148, 'mitochondria':
0.6791390776634216, 'nuclear': 0.2867254316806793, 'ribosome': 0.44316619634628296}
```

Εικόνα 1: Εκτέλεση κώδικα πρόβλεψης κυτταρικής θέσης πρωτεινικής αλληλουχίας

Μελλοντικές Βελτιώσεις:

1. Επέκταση δεδομένων

Ο αριθμός αλληλουχιών είναι διαφορετικός για κάθε μοντέλο, το οποίο έχει ως αποτέλεσμα σημαντικά διαφορετικές αποδόσεις και προβλέψεις των μοντέλων, ενώ στην περίπτωση του λυσοσώματος ο αριθμός των πρωτεϊνών δεν ήταν επαρκής για την εκπαίδευση μοντέλου. Για αυτό τον λόγο υπάρχουν τεχνικές επέκτασης των δεδομένων (data augmentation)[ιι]. Σε μια από αυτές τις τεχνικές, οι αλληλουχίες χωρίζονται σε υπό αλληλουχίες μικρότερου μήκους, συνδέονται σε περιοχές επικαλύψεις ανάμεσά τους και δημιουργούν παραλλαγές τις αρχικής αλληλουχίας. Με αυτό τον τρόπο πολλαπλασιάζεται ο αριθμός των αλληλουχιών οι οποίες θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου χωρίς να αλλοιώνεται η θεμελιώδης πληροφορία των αρχικών αλληλουχιών.

2. Φιλικό προς τον χρήστη

Τα προγράμματα που αναπτύχθηκαν έχουν πολλές σταθερές μεταβλητές όπως τον οργανισμό από των οποίο προέρχονται οι αλληλουχίες, τα οργανίδια του κυττάρου, τα όρια στα μήκη των αλληλουχιών, τις παραμέτρους βελτιστοποίησης των μοντέλων. Για παράδειγμα, στα προγράμματα βελτιστοποίησης, υπάρχει η επιλογή ενεργοποίησης bf16/gpu, διότι δεν υποστηρίζεται από όλα τα υπολογιστικά συστήματα. Σε μελλοντική εκδοχή θα μπορούσε να αναπτυχθεί κάποιο γραφικό περιβάλλον χρήστη, στο οποίο θα υπάρχει η δυνατότητα επιλογής παραπάνω από μια μεταβλητή.

3. Multilabel Classification

Με αλλαγές στο πρόγραμμα, το μοντέλο μπορεί να χρησιμοποιηθεί για ταξινόμηση παραπάνω από 2 ετικέτες, δηλαδή ένα ενιαίο μοντέλο να αντικαταστήσει τα οκτώ. Ένα τέτοιο μοντέλο προϋποθέτει ισορροπημένα σετ δεδομένων και χρησιμοποιεί μικρότερη υπολογιστική ισχύ σε σχέση με την εκπαίδευση 8 διαφορετικών μοντέλων, ενώ ταυτόχρονα η προβλέψεις της θέσης των πρωτεϊνών είναι πιο ακριβείς γιατί το μοντέλο έχει μια ενιαία απόδοση για όλα διαμερίσματα του κυττάρου.

Κώδικας:

Το προεκπαιδευμένο μοντέλο που χρησιμοποιήθηκε από την βιβλιοθήκη Transformers:

Mourad, R. (2024). *Mistral-Peptide-v1-15M* [Machine learning model]. Hugging Face. https://huggingface.co/RaphaelMourad/Mistral-Peptide-v1-15M

Ο κώδικας δημιουργήθηκε σε Python και υπάρχει στο link: (GitHub Repository).

Βιβλιογραφία:

1. **Abbasi-Vineh, M. A., Rouzbahani, S., Kavousi, K., & Emadpour, M. (2025).** Innovative data augmentation strategy for deep learning on biological datasets with limited gene representations focused on chloroplast genomes. *Scientific Reports, 15*(1), 27079.
2. **Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., & Winther, O. (2017).** DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics, 33*(21), 3387–3395.
3. **Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., ... & Robles, V. (2006).** Machine learning in bioinformatics. *Briefings in Bioinformatics, 7*(1), 86–112.
4. **Min, S., Lee, B., & Yoon, S. (2017).** Deep learning in bioinformatics. *Briefings in Bioinformatics, 18*(5), 851–869.
5. **Müller, A. C., & Guido, S. (2016).** *Introduction to machine learning with Python: A guide for data scientists*. O'Reilly Media, Inc. Chapter 2 – Supervised learning.
6. **Raphael Mourad & Bérénice Batut.** Fine-tuning a LLM for DNA Sequence Classification (Galaxy Training Materials). https://training.galaxyproject.org/training-material/topics/statistics/tutorials/genomic-llm-finetuning/tutorial.html Online; accessed Thu Aug 21 2025
7. **Τζεδάκης, Χ. Ε. (2014).** Ανασκόπηση της εφαρμογής των μεθόδων μηχανικής μάθησης στη βιοπληροφορική (Bachelor's thesis).
8. **Wang, X., & Li, S. (2014).** Protein mislocalization: mechanisms, functions and clinical applications in cancer. *Biochimica et Biophysica Acta (BBA) – Reviews on Cancer, 1846*(1), 13–25.