Contents

Αναφορά — Εργασία 3: Αναζήτηση “Απομακρυσμένων” Ομολόγων με ESM-2 και ANN

Το παρόν κείμενο τεκμηριώνει την πειραματική μελέτη και τη βιολογική αξιολόγηση της αναζήτησης απομακρυσμένων ομολόγων πρωτεϊνών με χρήση εμβυθίσεων (embeddings) από ESM-2 και προσεγγιστικών μεθόδων πλησιέστερων γειτόνων (ANN). Η δομή και οι απαιτήσεις ακολουθούν την εκφώνηση και το συνοδευτικό υλικό της εργασίας. Η υλοποίηση αναπτύχθηκε σε αποθετήριο Git, με σταδιακές καταγραφές αλλαγών.

1. Εισαγωγή και θεωρητικό πλαίσιο

Η ανίχνευση ομολόγων πρωτεϊνών μέσω στοίχισης ακολουθιών λειτουργεί ικανοποιητικά όταν η ομοιότητα αλληλουχίας είναι σχετικά υψηλή. Στη “Twilight Zone” (περίπου κάτω από 30% ταυτότητα) η απόδοση υποβαθμίζεται, επειδή διατηρείται συχνά η δομή και/ή η λειτουργία ενώ η αλληλουχία έχει αποκλίνει σημαντικά. Στόχος είναι η αξιοποίηση ενός διανυσματικού χώρου, όπου η γειτνίαση αντανακλά πλουσιότερα χαρακτηριστικά από την άμεση ομοιότητα χαρακτήρων.

2. Δεδομένα και πειραματικό setup

Η βάση δεδομένων αποτελείται από πρωτεΐνες SwissProt (50.000 ακολουθίες), ενώ τα ερωτήματα προέρχονται από ένα σύνολο πρωτεϊνών-στόχων. Η αναφορά περιγράφει επίσης το πώς ορίζεται “ground truth” μέσω BLAST στο ίδιο corpus.

3. Pipeline και μέθοδοι

Η διαδικασία χωρίζεται σε τρία στάδια: παραγωγή embeddings με ESM-2, κατασκευή/εκτέλεση ANN ευρετηρίων και αξιολόγηση έναντι BLAST.

3.1 Παραγωγή embeddings (ESM-2)

Χρησιμοποιείται το προεκπαιδευμένο μοντέλο `facebook/esm2_t6_8M_UR50D`. Για κάθε πρωτεΐνη λαμβάνεται η αναπαράσταση του τελευταίου επιπέδου και παράγεται ένα διάνυσμα σταθερού μήκους μέσω mean pooling στις αναπαραστάσεις των residues. Για λόγους υπολογιστικού κόστους και συμβατότητας με τα ειδικά tokens, οι πολύ μεγάλες ακολουθίες αποκόπτονται στο μέγιστο επιτρεπτό μήκος.

3.2 Μέθοδοι ANN και υπερπαράμετροι

Εξετάζονται οι εξής προσεγγιστικές μέθοδοι στο χώρο των embeddings (L2):

Euclidean LSH με υπερπαραμέτρους (k, L, w), Hypercube Projection με (kproj, w, M, probes), IVF-Flat με (kclusters, nprobe), IVF-PQ με (kclusters, nprobe, M, nbits), και Neural LSH με (m, T) και υπερπαραμέτρους MLP (epochs, layers, hidden units, learning rate). Για την επιλογή υπερπαραμέτρων έχει υλοποιηθεί ξεχωριστό script grid search, ώστε να αναλυθεί εμπειρικά το trade-off Recall@N έναντι QPS· στην τελική εκτέλεση που παρουσιάζεται παρακάτω χρησιμοποιήθηκαν οι ρυθμίσεις που επιλέχθηκαν από αυτή τη διερεύνηση.

3.3 BLAST ως baseline και ground truth

Για κάθε query εκτελείται BLAST έναντι της ίδιας βάσης. Ως αναφορά (“ground truth”) θεωρούνται τα Top-N αποτελέσματα του BLAST ταξινομημένα κατά bit-score. Η ταυτότητα (BLAST identity) χρησιμοποιείται στη βιολογική αξιολόγηση, ειδικά όταν βρίσκεται κάτω από 30%.

4. Μετρικές

Η ακρίβεια ποσοτικοποιείται με Recall@N έναντι των Top-N χτυπημάτων του BLAST. Για κάθε query q, ορίζεται η ανάκτηση ως το πλήθος κοινών στοιχείων ανάμεσα στο σύνολο των N γειτόνων που επιστρέφει η ANN μέθοδος και στο σύνολο των N κορυφαίων χτυπημάτων του BLAST, διαιρεμένο με N. Η συνολική επίδοση υπολογίζεται ως μέσος όρος πάνω στα queries.

Η ταχύτητα αναφέρεται ως QPS (Queries Per Second), δηλαδή το πλήθος queries που εξυπηρετούνται ανά δευτερόλεπτο, με μέτρηση χρόνου αναζήτησης (χωρίς να συνυπολογίζεται το κόστος παραγωγής query embeddings όταν το ζητούμενο είναι η σύγκριση των ευρετηρίων στον ίδιο χώρο).

5. Πειραματικά αποτελέσματα (ποσοτικά)

Η τελική αξιολόγηση πραγματοποιήθηκε σε βάση 50.000 πρωτεϊνών (SwissProt), με (N=50) για τον υπολογισμό της Recall@N (σε σχέση με BLAST Top-50) και με εκτύπωση Top-10 γειτόνων ανά μέθοδο για αναγνωσιμότητα. Το σύνολο αξιολόγησης περιλαμβάνει 12 queries (από `targets.fasta`).

Η παραγωγή query embeddings έγινε με το ίδιο ESM-2 μοντέλο και mean pooling. Οι χρόνοι που αναφέρονται για τις ANN μεθόδους αφορούν αποκλειστικά το στάδιο αναζήτησης (`index.query(...)`) στον χώρο των embeddings, ενώ η παραγωγή των query embeddings μετράται χωριστά και δεν συμπεριλαμβάνεται στους χρόνους του πίνακα. Για λόγους πληρότητας, η BLAST αναφορά μετράται ως συνολικός χρόνος στοίχισης για το ίδιο σύνολο queries.

5.1 Επιλογή υπερπαραμέτρων (grid search)

Η επιλογή υπερπαραμέτρων πραγματοποιήθηκε με το `protein_grid_search.py`, χρησιμοποιώντας ως “ground truth” τα BLAST Top-50 στο ίδιο corpus. Για κάθε μέθοδο δοκιμάστηκε ένα μικρό αλλά αντιπροσωπευτικό πλέγμα τιμών και επιλέχθηκε ρύθμιση που ισορροπεί Recall@50 και QPS. Τα αναλυτικά αποτελέσματα των πλεγμάτων αποθηκεύονται στα αρχεία `grid_*.csv` (LSH/Hypercube/IVF-Flat/IVF-PQ/Neural) για αναπαραγωγιμότητα.

Για το Euclidean LSH εξετάστηκαν (k∈{2,4,6}), (L∈{5,10}), (w∈{2,4,6}) και επιλέχθηκε (k=6, L=10, w=4.0), που αύξησε την Recall@50 σε ~0.203 με QPS ~118 στο grid.

Για το Hypercube εξετάστηκαν (k_{proj}∈{12,14,16}), (M∈{500,1000}), (probes∈{5,10,20}) (με (w=4.0)) και επιλέχθηκε (k_{proj}=14, M=1000, probes=20), που βελτίωσε την Recall@50 σε ~0.097 με διατήρηση υψηλού QPS.

Για το IVF-Flat εξετάστηκαν (k_{clusters}∈{25,50,100}), (n_{probe}∈{1,5,10}) και επιλέχθηκε (k_{clusters}=100, n_{probe}=5), που διατήρησε Recall@50 ~0.252 και έδωσε σημαντικά υψηλότερο QPS από μικρότερα (k_{clusters}).

Για το IVF-PQ εξετάστηκαν (n_{probe}∈{1,5,10}) και (nbits∈{6,8}) (με (M=16)), και επιλέχθηκε (nbits=8). Επιπλέον ελέγχθηκε η επίδραση του (k_{clusters}) στα (50) και (100)· επιλέχθηκε (k_{clusters}=100, n_{probe}=5) ως συμβιβασμός που διατηρεί Recall@50 ~0.185 και αυξάνει αισθητά τον QPS σε σχέση με (k_{clusters}=50).

Για το Neural LSH εξετάστηκαν (m∈{100,200}), (T∈{3,5,10}) και επιλέχθηκε (m=200, T=5) (με epochs=10, layers=3, hidden_units=256), που έδωσε τη μεγαλύτερη Recall@50 (~0.260) στο πλέγμα, με υψηλό QPS.

Στην τελική εκτέλεση χρησιμοποιήθηκαν οι παραπάνω ρυθμίσεις: Euclidean LSH (k=6, L=10, w=4.0), Hypercube (kproj=14, w=4.0, M=1000, probes=20), IVF-Flat (kclusters=100, nprobe=5), IVF-PQ (kclusters=100, nprobe=5, M=16, nbits=8), Neural LSH (m=200, T=5, epochs=10, layers=3, hidden_units=256, lr=1e-3, batch_size=256).

Ο Πίνακας 1 συνοψίζει τους μέσους όρους (πάνω στα 12 queries) του χρόνου ανά query, του QPS και της Recall@50.

Table 1: Μέσοι όροι χρόνου ανά query, QPS και Recall@50 (έναντι BLAST Top-50) για τα 12 queries.

Method	Time/query (s)	QPS	Mean Recall@50 (vs BLAST Top-50)
Euclidean LSH	0.0187	127.80	0.203
Hypercube	0.0028	754.39	0.0967
Neural LSH	0.0041	478.66	0.260
IVF-Flat	0.0039	359.93	0.252
IVF-PQ	0.0014	901.83	0.185
BLAST (Ref)	7.715	0.130	1.000

Στην παραπάνω ρύθμιση, το Hypercube διατηρεί πολύ υψηλό QPS αλλά χαμηλή ανάκτηση έναντι BLAST Top-50, ενώ το IVF-Flat και το Neural LSH επιτυγχάνουν υψηλότερη Recall. Το IVF-PQ παρουσιάζει υψηλό QPS με χαμηλότερη Recall από IVF-Flat, χαρακτηριστικό της επιπλέον ποσοτικοποίησης. Το BLAST παραμένει κατά τάξεις μεγέθους βραδύτερο, όπως αποτυπώνεται από το QPS.

Το Σχήμα 1 παρουσιάζει το trade-off μεταξύ Recall@50 και QPS για όλες τις μεθόδους, βασισμένο στα αποτελέσματα του grid search. Κάθε σημείο αντιπροσωπεύει μία διαφορετική ρύθμιση υπερπαραμέτρων. Το διάγραμμα επιτρέπει την οπτική σύγκριση των μεθόδων και την επιλογή ρυθμίσεων που ισορροπούν ταχύτητα και ακρίβεια ανάλογα με τις ανάγκες κάθε εφαρμογής.

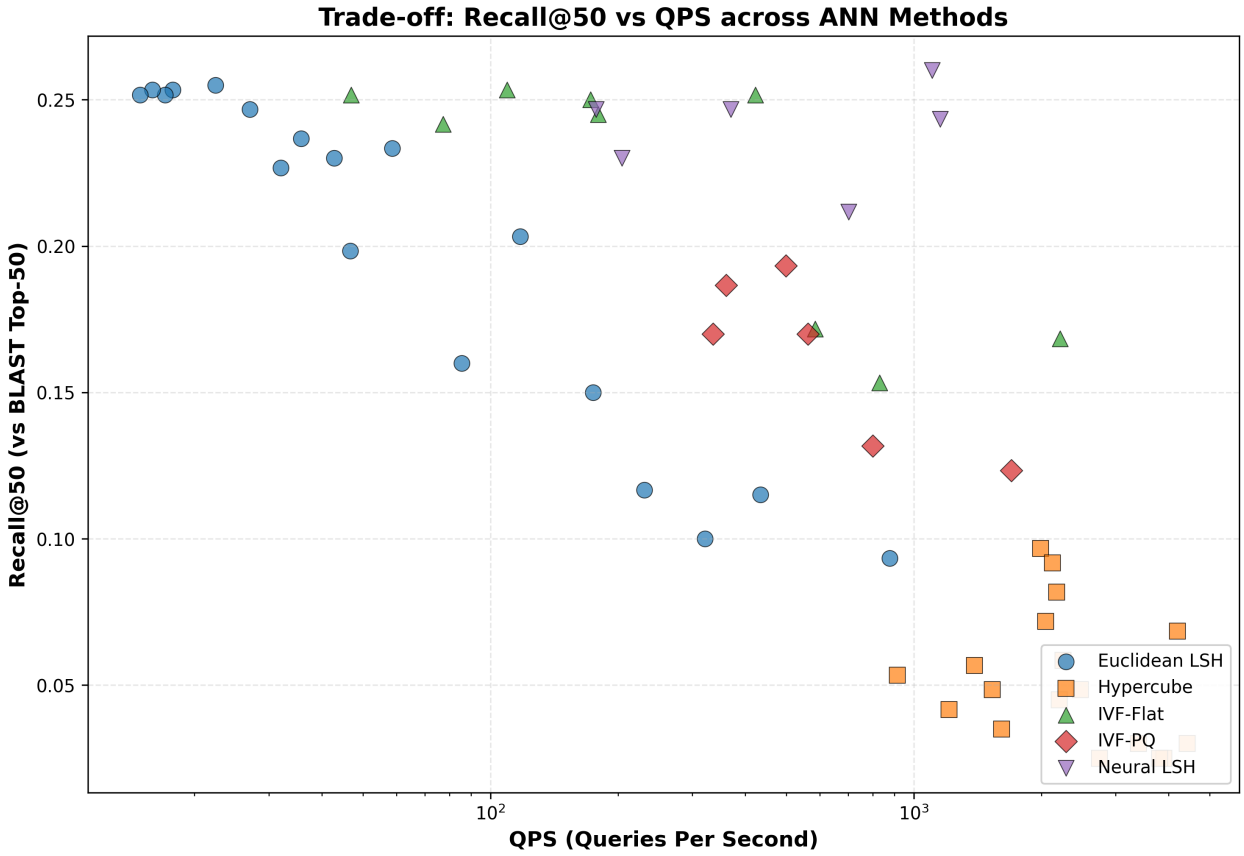

Figure 1: Trade-off: Recall@50 vs QPS

Σχήμα 1: Trade-off μεταξύ Recall@50 και QPS για τις πέντε ANN μεθόδους. Ο άξονας QPS είναι σε λογαριθμική κλίμακα. Κάθε σημείο αντιπροσωπεύει μία ρύθμιση υπερπαραμέτρων από το grid search.

6. Βιολογική αξιολόγηση (ποιοτικά)

6.0 Ορισμός "απομακρυσμένου ομόλογου" στην πράξη

Στην πράξη, ως υποψήφιος απομακρυσμένος ομόλογος ορίζεται ένας γείτονας που πληροί τα ακόλουθα κριτήρια:

1. **Κριτήρια από τα αποτελέσματα ANN**: Ο γείτονας επιστρέφεται ψηλά (Top-10) από μία ή περισσότερες embedding-based μεθόδους, με μικρή L2 απόσταση στον χώρο των embeddings.

2. **Κριτήρια από BLAST**: Η BLAST identity είναι χαμηλή (κάτω από 30%, δηλαδή εντός της "Twilight Zone"), και ο γείτονας δεν ανήκει στο BLAST Top-50 του ίδιου query (δηλαδή δεν ανακαλύπτεται εύκολα από την παραδοσιακή στοίχιση ακολουθιών).

3. **Βιολογική τεκμηρίωση**: Ο γείτονας συνοδεύεται από συμβατές επισημειώσεις σε UniProt/SwissProt που υποστηρίζουν ομολογία, όπως:

- Κοινά Pfam domains που υποδεικνύουν διατηρημένη δομική/λειτουργική μονάδα
- Συμβατοί GO όροι (Gene Ontology) που υποδεικνύουν παρόμοια βιολογική λειτουργία
- Συμβατή λειτουργική περιγραφή που υποδεικνύει συγγένεια σε επίπεδο οικογένειας ή ομάδας πρωτεϊνών

Αυτός ο ορισμός επιτρέπει την ταυτοποίηση πρωτεϊνών που διατηρούν λειτουργική ή δομική συγγένεια παρά την χαμηλή ομοιότητα αλληλουχίας, και που δεν ανακαλύπτονται εύκολα από το BLAST λόγω της εξελικτικής απόκλισης.

Τα παρακάτω παραδείγματα είναι χαρακτηριστικά και προέρχονται απευθείας από το `results.txt`.

6.1 Υποψήφιος απομακρυσμένος ομόλογος: UvrB-τύπου repair helicase (NER)

Για το query A0A009HN45 (*Acinetobacter baumannii*), το Hypercube επέστρεψε το Q9ZDW2 (UvrABC system protein B, *Rickettsia prowazekii*) στη θέση 2 με L2=0.74 και BLAST identity 25%, εκτός BLAST Top-50. Το UniProt για το Q9ZDW2 περιλαμβάνει GO:0006289 (nucleotide-excision repair) και GO:0009380 (excinuclease repair complex), ενώ και τα δύο εμφανίζουν helicase-related επισημείωση (GO:0004386, ATP binding). Η σύμπτωση στο επίπεδο μοριακής λειτουργίας/repair συστήματος, σε συνδυασμό με identity εντός Twilight Zone, είναι συμβατή με απομακρυσμένη ομολογία που δεν ανακτάται ψηλά από BLAST.

6.2 Υποψήφιος απομακρυσμένος ομόλογος: Helicase-like λειτουργία (PF00271) με χαμηλή identity

Για το query A0A009HQC9 (RNA polymerase-associated protein RapA), η Euclidean LSH επέστρεψε το Q8SQM5 (ATP-dependent RNA helicase eIF4A, *Encephalitozoon cuniculi*) στη θέση 8 (L2=1.52) με BLAST identity 26% και εκτός BLAST Top-50. Και οι δύο πρωτεΐνες εμφανίζουν "helicase-like" προφίλ: το RapA σχετίζεται με ρύθμιση μεταγραφής και έχει GO:0004386 (helicase activity), ενώ το eIF4A φέρει GO όρους ATP binding/ATP hydrolysis και RNA helicase activity, καθώς και κοινή Pfam "υπογραφή" PF00271. Η συμφωνία στο επίπεδο domain/λειτουργίας, με ταυτότητα στο εύρος Twilight Zone, είναι ενδεικτική περίπτωσης όπου οι embeddings αναδεικνύουν λειτουργικά συγγενείς πρωτεΐνες που δεν ανακτώνται ψηλά από BLAST.

6.3 Υποψήφιος απομακρυσμένος ομόλογος: WD repeats και πλαίσιο ribosome biogenesis (YTM1)

Για το query A0A010Q3W2 (WD domain-containing protein), το Neural LSH επέστρεψε το A1CXL0 (Ribosome biogenesis protein ytm1) στη θέση 5 (L2=1.24) με BLAST identity 29% και εκτός BLAST Top-50. Και τα δύο φέρουν Pfam PF00400 (WD repeats) και σχετίζονται στο UniProt με πυρηνικά/νουκλεολικά συμφραζόμενα και rRNA επεξεργασία (ενδεικτικά: nucleolus/nucleoplasm, ωρίμανση rRNA). Η συμφωνία στο επίπεδο domain και βιολογικού πλαισίου, με identity στο όριο Twilight Zone, είναι ενδεικτική απομακρυσμένης ομολογίας/οικογένειας σε WD-repeat χώρο.

6.4 Υποψήφιος απομακρυσμένος ομόλογος: ABC-type ATPase / transport

Για το query A0A002 (MoeJ5, *Streptomyces viridosporus*), η Euclidean LSH επέστρεψε το P9WQJ4 (OppD, oligopeptide transport ATP-binding protein, *Mycobacterium tuberculosis*) στη θέση 8 (L2=1.89) με BLAST identity 27% και εκτός BLAST Top-50. Τα UniProt annotations είναι συνεπή με ABC-type transporter ATPase (GO:0005524, GO:0016887), ενώ και τα Pfam domains περιλαμβάνουν PF00005. Η περίπτωση αποτελεί τυπικό εύρημα "Twilight Zone" εντός μεγάλης οικογένειας μεταφορέων, όπου η γειτνίαση στο embedding space διατηρεί λειτουργική συνοχή, παρά την απόκλιση αλληλουχίας.

6.5 Παράδειγμα πιθανού false positive

Το query A0A009HPM0 αντιστοιχεί σε biotin carboxylase (*Acinetobacter baumannii*), ένζυμο του acetyl-CoA carboxylase complex. Ωστόσο, στο IVF-PQ εμφανίστηκε ως κοντινός γείτονας το Q5E320 (RlmD, 23S rRNA methyltransferase) στη θέση 6 με BLAST identity 22% και εκτός BLAST Top-50. Η λειτουργική περιγραφή και το βιολογικό πλαίσιο είναι ασύμβατα (καρβοξυλίωση/λιπιδικός μεταβολισμός έναντι τροποποίησης rRNA), ενώ και τα Pfam cross-references δεν συγκλίνουν. Η περίπτωση υποδεικνύει ότι η γειτνίαση στο embedding space δεν αρκεί από μόνη της για βιολογική ερμηνεία, και ότι φίλτρα με βάση Pfam/GO είναι πρακτικά χρήσιμα για τον περιορισμό false positives.

7. Συζήτηση ορίων και βελτιώσεων

Η προσέγγιση που παρουσιάστηκε αντιμετωπίζει αρκετούς περιορισμούς. Η επιλογή της μετρικής L2 στον χώρο των embeddings μπορεί να μην αντανακλά πάντα τη βιολογική συγγένεια, ειδικά όταν οι πρωτεΐνες διαφέρουν σε μήκος ή όταν η λειτουργική ομοιότητα δεν αντιστοιχεί σε γεωμετρική εγγύτητα. Οι υπερπαράμετροι των ANN μεθόδων επηρεάζουν σημαντικά την απόδοση, όπως φαίνεται από τα αποτελέσματα του grid search, αλλά η βέλτιστη επιλογή εξαρτάται από το trade-off μεταξύ ταχύτητας και ακρίβειας που επιδιώκει κάθε εφαρμογή.

Η ελλιπής ή ανακριβής επισημείωση στη βάση δεδομένων μπορεί να οδηγήσει σε λανθασμένες ερμηνείες των αποτελεσμάτων. Στο παράδειγμα του false positive (ενότητα 6.5), η γειτνίαση στο embedding space δεν συνοδεύεται από συμβατές βιολογικές επισημειώσεις, υποδεικνύοντας ότι η χρήση φίλτρων με βάση Pfam domains ή GO όρους είναι απαραίτητη για τον περιορισμό τέτοιων περιπτώσεων.

Πιθανές κατευθύνσεις βελτίωσης περιλαμβάνουν την εξερεύνηση εναλλακτικών μετρικών απόστασης (π.χ. cosine similarity, learned metrics) που μπορεί να είναι πιο κατάλληλες για πρωτεϊνικά embeddings. Η χρήση μεγαλύτερων μοντέλων ESM-2 (π.χ. esm2_t33_650M_UR50D) ή ειδικών μοντέλων που έχουν εκπαιδευτεί για ομολογία μπορεί να βελτιώσει την ποιότητα των embeddings. Επιπλέον, πιο προηγμένα σχήματα ευρετηρίασης, όπως hierarchical navigable small world (HNSW) ή learned indices, μπορεί να προσφέρουν καλύτερο trade-off Recall-QPS για μεγάλες βάσεις δεδομένων.

8. Συμπεράσματα

Η πειραματική μελέτη επιδεικνύει ότι οι προσεγγιστικές μέθοδοι πλησιέστερων γειτόνων στον χώρο των ESM-2 embeddings μπορούν να επιτύχουν σημαντική ταχύτητα έναντι του BLAST, διατηρώντας παράλληλα λογικά επίπεδα ανάκτησης. Το Neural LSH και το IVF-Flat επιτυγχάνουν τις υψηλότερες τιμές Recall@50 (0.260 και 0.252 αντίστοιχα), ενώ το IVF-PQ και το Hypercube προσφέρουν εξαιρετικά υψηλό QPS (901.83 και 754.39 αντίστοιχα) με χαμηλότερη ανάκτηση.

Σημαντικότερα, οι embedding-based μέθοδοι αναδεικνύουν υποψήφιες απομακρυσμένες ομολογίες που δεν ανακτώνται ψηλά από το BLAST, όπως φαίνεται από τα παραδείγματα της ενότητας 6. Αυτές οι περιπτώσεις χαρακτηρίζονται από ταυτότητα αλληλουχίας εντός της Twilight Zone (20-30%) αλλά συμβατές βιολογικές επισημειώσεις σε επίπεδο domain, GO όρων ή λειτουργικής περιγραφής. Η προσέγγιση αποτελεί συμπληρωματικό εργαλείο στο BLAST, ιδιαίτερα χρήσιμο για την ανίχνευση απομακρυσμένων ομολόγων όπου η διατήρηση λειτουργίας ή δομής δεν αντανακλάται πλήρως στην ομοιότητα αλληλουχίας.