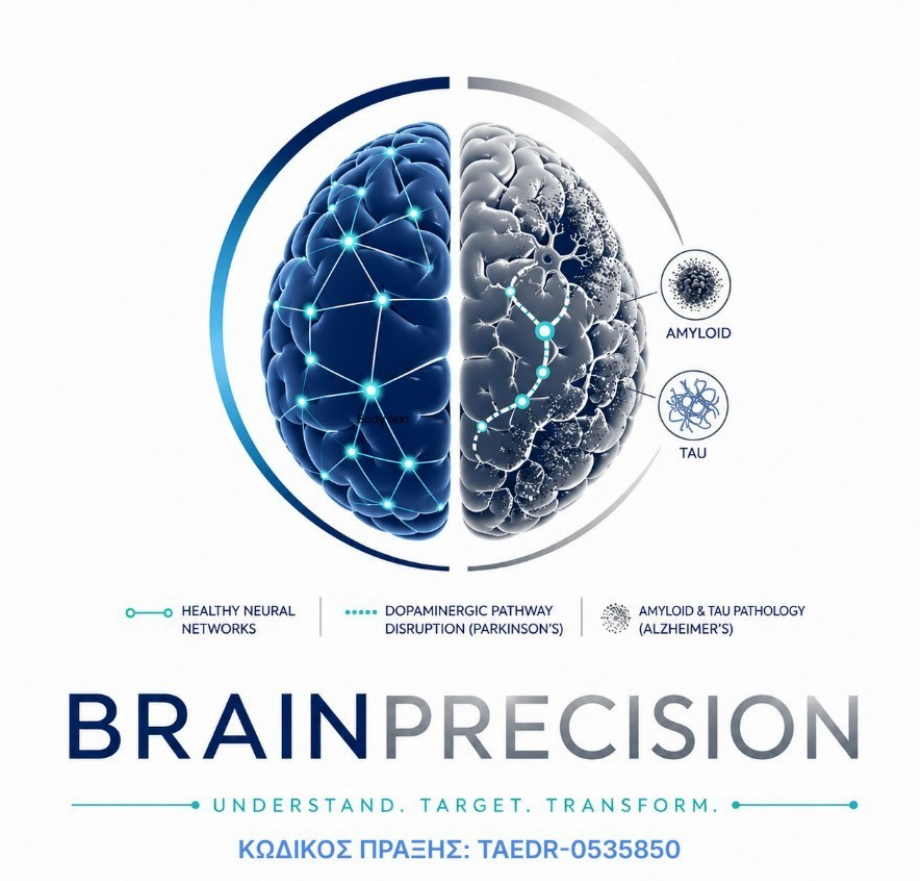


Δράση «Εμβληματικές δράσεις σε διαθεματικές επιστημονικές περιοχές με ειδικό ενδιαφέρον για την σύνδεση με τον παραγωγικό ιστό» ID 16618

Εθνικό δίκτυο έρευνας για την ανάδειξη της γενετικής βάσης των νευροεκφυλιστικών νόσων Alzheimer και Parkinson, την ανίχνευση αξιόπιστων βιοδεικτών και την ανάπτυξη καινοτόμων υπολογιστικών τεχνολογιών και θεραπευτικών στρατηγικών στη βάση της ιατρικής ακριβείας (BRAIN PRECISION, TAEDR-0535850)

ΤΙΤΛΟΣ ΠΑΡΑΔΟΤΕΟΥ: Σύνδεση των δεδομένων της εμβληματικής δράσης BrainPrecision με υπολογιστικές δομές για εκτέλεση ροών εργασιών σε ασφαλές και σύμφωνο με το GDPR περιβάλλον από χρήστες ακόμη και εκτός Ελλάδος

ΕΝΟΤΗΤΑ ΕΡΓΑΣΙΑΣ 2: Καταγραφή πασχόντων και δημιουργία τράπεζας βιολογικού υλικού των νευροεκφυλιστικών νόσων Alzheimer και Parkinson και των πρόδρομων μορφών τους

ΥΠΕΥΘΥΝΗ ΕΡΕΥΝΗΤΙΚΗ ΟΜΑΔΑ (ΦΟΡΕΑΣ): MARTIN RECZKO (ΕΚΕΒΕ "Αλέξανδρος Φλέμιγκ")

Σύνδεση των δεδομένων της εμβληματικής δράσης BrainPrecision με υπολογιστικές δομές για εκτέλεση ροών εργασιών σε ασφαλές και σύμφωνο με το GDPR περιβάλλον από χρήστες ακόμη και εκτός Ελλάδος

ΠΕΡΙΓΡΑΦΗ

A. **Ασφαλές Περιβάλλον Εκτέλεσης**: Δημιουργήθηκαν Trusted Data Processing Environments για την ασφαλή εκτέλεση βιοπληροφορικών ροών εργασίας (workflows).
B. **Διαδικτυακή Υπηρεσία Πρόβλεψης RNA**: Παράδειγμα ροής εργασίας που θα εκτελεστεί σε ασφαλές περιβάλλον αποτελεί η ενσωμάτωση διαφόρων εργαλείων που σχετίζονται με την τρισδιάστατη πρόβλεψη δομής του RNA με βάση την αλληλουχία του. Η υπηρεσία ενσωματώνει προηγμένους αλγορίθμους (DeepFoldRNA, trRosettaRNA, RhoFold) και παρέχει τη δυνατότητα αξιολόγησης τρισδιάστατων δομών μέσω των μεθόδων REA, ARES και SVR.

ΠΑΡΑΔΟΤΕΑ

Υλοποίηση Έμπιστου Περιβάλλοντος Έρευνας

Ένα Trusted Research Environment (TRE) είναι μια ασφαλής, ελεγχόμενη και απομονωμένη ψηφιακή τοποθεσία που επιτρέπει σε ερευνητές να έχουν πρόσβαση και να αναλύουν ευαίσθητα δεδομένα (όπως γονιδιωματικά ή κλινικά) χωρίς τα δεδομένα αυτά να "φεύγουν" ποτέ από την υποδομή του παρόχου. Σε αντίθεση με την παραδοσιακή μέθοδο όπου ο ερευνητής κατεβάζει τα δεδομένα στον υπολογιστή του, στο TRE ισχύει η αρχή: «φέρνουμε τον ερευνητή στα δεδομένα».
Ένα τέτοιο περιβάλλον παρέχει όλα τα απαραίτητα εργαλεία ανάλυσης (R, Python, Jupyter, VS Code) σε μια εικονική επιφάνεια εργασίας, η οποία είναι αποκομμένη από το διαδίκτυο για την πρόληψη διαρροών.

Η υιοθέτηση των TREs κρίνεται επιβεβλημένη καθώς παρέχει πολλά οφέλη όπως:

1. Προστασία Προσωπικών Δεδομένων και συμμόρφωση με το GDPR

   Η διαχείριση ευαίσθητων δεδομένων απαιτεί αυστηρή συμμόρφωση με τον Γενικό Κανονισμό Προστασίας Δεδομένων (GDPR). Το TRE διασφαλίζει ότι τα ακατέργαστα δεδομένα (raw data) παραμένουν προστατευμένα, επιτρέποντας την εξαγωγή μόνο των τελικών στατιστικών αποτελεσμάτων, τα οποία έχουν ελεγχθεί ότι δεν αποκαλύπτουν την ταυτότητα των συμμετεχόντων

2. Ασφάλεια και Έλεγχο δεδομένων (Data Sovereignty)

   Σε ένα TRE, ο πάροχος των δεδομένων έχει τον πλήρη έλεγχο και εμποδίζεται οποιαδήποτε προσπάθεια αποστολής δεδομένων σε εξωτερικούς διακομιστές. Κάθε ενέργεια του χρήστη καταγράφεται (Auditing), διασφαλίζοντας τη διαφάνεια και τη λογοδοσία

3. Διευκόλυνση της Συνεργασίας

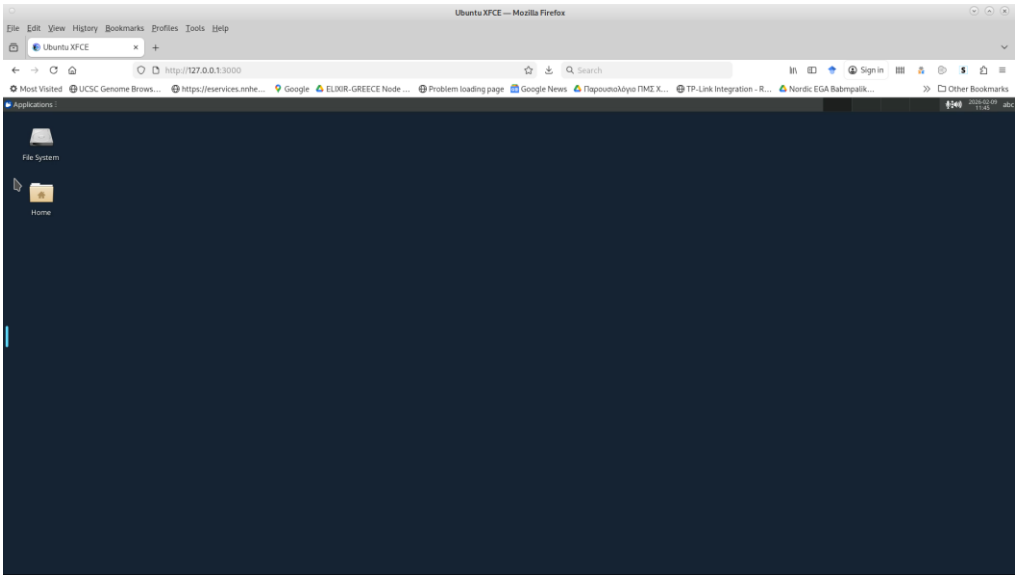
   Τα TREs επιτρέπουν σε ομάδες από διαφορετικές χώρες (π.χ. μέσω του δικτύου ELIXIR) να συνεργάζονται πάνω στο ίδιο σύνολο δεδομένων μέσα σε ένα κοινό, ασφαλές περιβάλλον. Αυτό καταργεί την ανάγκη για πολλαπλά αντίγραφα δεδομένων σε διαφορετικά ιδρύματα, μειώνοντας δραστικά το ρίσκο ασφάλειας.

4. Υπολογιστική Ισχύς και Εργαλεία

Τα TREs προσφέρουν πρόσβαση σε υπολογιστική ισχύ υψηλών επιδόσεων (HPC), έχοντας προ-εγκατεστημένα και βελτιστοποιημένα όλα τα επιστημονικά πακέτα λογισμικού.

Για τους παραπάνω λόγους, στο πλαίσιο του έργου *Brain Precision* αναπτύχθηκε ένα τέτοιο TRE περιβάλλον που στηρίζεται στο docker image webtop που εκτελεί την διανομή XFCE Ubuntu. Το συγκεκριμένο TRE κάνει mount τμήματα από το filesystem που περιέχουν τα δεδομένα που είναι προς διαμοιρασμό με δικαιώματα εγγραφής και ανάγνωσης ενώ μπορεί να έχει πρόσβαση σε δεδομένα με μόνο δικαίωμα ανάγνωση.

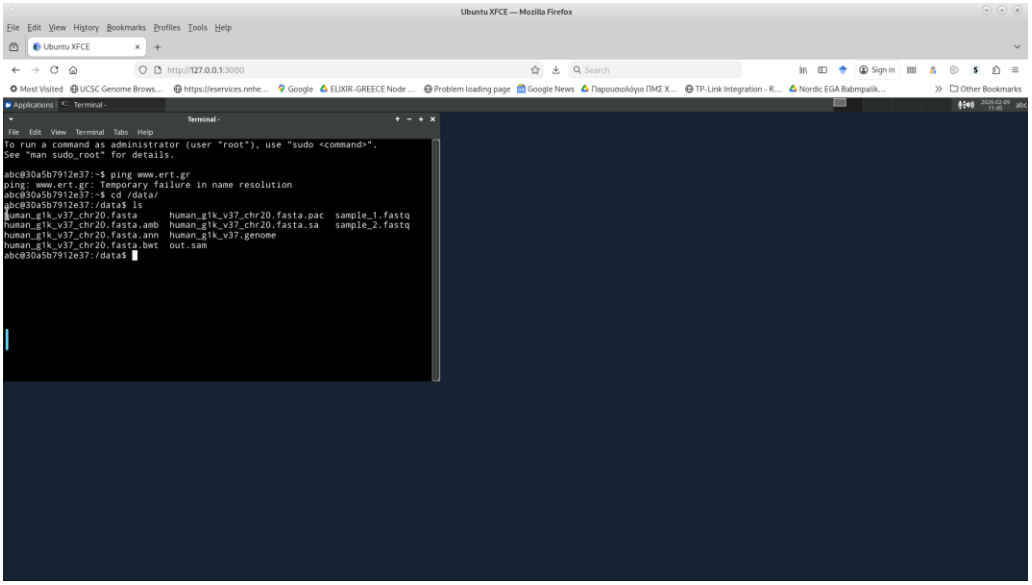
Για προσπέλαση στο συγκεκριμένο TRE χρειάζεται μόνο ένα web browser, όπως φαίνεται και στο ακόλουθο σχήμα.



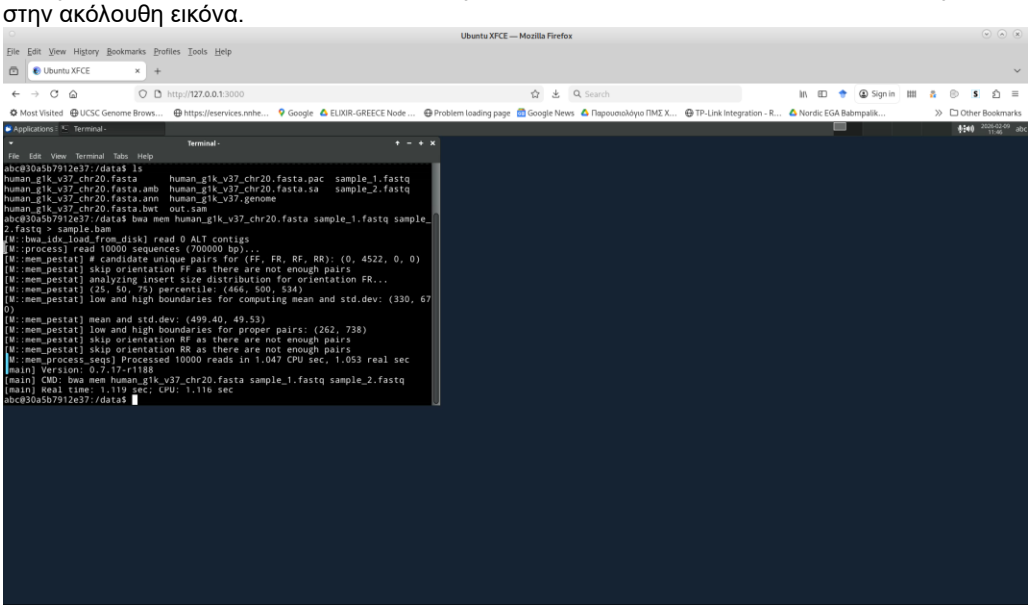
Πρόσβαση στο TRE μέσω της διεύθυνσης http://127.0.0.1:3000

Μόλις συνδεθεί ο χρήστης στην διεύθυνση http://127.0.0.1:3000 αποκτά πρόσβαση σε μια εικονική μηχανή που περιέχει προεγκατεστημένα απαραίτητα εργαλεία για βιοπληροφορικές αναλύσεις.
Στο παράδειγμα που ακολουθεί, γίνεται aligment αρχείων fastq πάνω στο ανθρώπινο γονιδίωμα με χρήση του εργαλείου bwa.

Όπως φαίνεται στην παρακάτω εικόνα το VM δεν έχει πρόβαση στο διαδίκτυο αλλά έχει πρόσβαση σε συγκεκριμένα δεδομένα μέσω του mountpoint /data :



Τέλος, το VM έχει και δικαιώματα εγγραφής, και αποθηκεύει εκεί τα δεδομένα που μόλις έγιναν align, όπως φαίνεται στην ακόλουθη εικόνα.



Επομένως, η χρήση TRE γεφυρώνει το χάσμα μεταξύ της ανάγκης για επιστημονική πρόοδο και της υποχρέωσης για απόλυτη προστασία της ιδιωτικότητας, δημιουργώντας ένα περιβάλλον εμπιστοσύνης μεταξύ ερευνητών, ασθενών και θεσμικών οργάνων.

Β. Διαδικτυακή Υπηρεσία Πρόβλεψης RNA

Παράδειγμα ροής εργασίας που θα εκτελεστεί σε ασφαλές περιβάλλον αποτελεί η διαδικτυακή υπηρεσία για την πρόβλεψη τρισδιάστατων δομών RNA με βάση αποκλειστικά την αλληλουχία τους, αξιοποιώντας σύγχρονες μεθόδους βαθιάς μάθησης. Η υπηρεσία παράγει εννέα (9) εναλλακτικά τρισδιάστατα μοντέλα για κάθε αλυσίδα RNA, χρησιμοποιώντας τις μεθόδους DeepFoldRNA [1], trRosettaRNA [2,3] και RhoFold [4].

Η αξιολόγηση και η κατάταξη των παραγόμενων μοντέλων πραγματοποιείται με αλγορίθμους μηχανικής μάθησης REA, ARES και SVR και συνδυασμούς αυτών, με στόχο την ανάδειξη του ακριβέστερου μοντέλου, βάσει της προβλεπόμενης απόκλισης RMSD (Root Mean Square Deviation) από τη φυσική τρισδιάστατη διαμόρφωση, σύμφωνα με τη σχετική πρόσφατη βιβλιογραφία [5, 6]. Η υπηρεσία επιτρέπει τη μετατροπή λειτουργικών αλληλουχιών RNA από πειράματα αλληλούχισης σε τρισδιάστατα δομικά μοντέλα υψηλής ακρίβειας, τα οποία μπορούν να αξιοποιηθούν για περαιτέρω δομικό και λειτουργικό σχεδιασμό, υποστηρίζοντας τους στόχους της ιατρικής ακρίβειας και της μελέτης νευροεκφυλιστικών νόσων.

Περαιτέρω, στη διαδικτυακή υπηρεσία ενσωματώθηκε η δυνατότητα επιλογής από τον χρήστη της χρήσης πολλαπλής στοίχισης αλληλουχιών (Multiple Sequence Alignment, MSA) ως επιπλέον δεδομένου εισόδου στους αλγορίθμους πρόβλεψης. Η αξιοποίηση της πληροφορίας αυτής δύναται να βελτιώσει σημαντικά την ποιότητα των προβλέψεων, ιδίως σε περιπτώσεις όπου η υπό μελέτη αλληλουχία εμφανίζει ομοιότητα με ήδη γνωστές, καθώς τα μοντέλα βαθιάς μάθησης μπορούν να εξάγουν εξελικτικό σήμα από τους πίνακες MSA, συμβάλλοντας στον ακριβέστερο προσδιορισμό της τρισδιάστατης δομής.

[1] C. Feng, et al., “Accurate de novo prediction of RNA 3D structure with transformer network,” bioRxiv, p. 2022.10.24.513506, Oct. 25, 2022. doi: 10.1101/2022.10.24.513506
[2] W. Wang et al., “trRosettaRNA: automated prediction of RNA 3D structure with transformer network,” Nature Communications, vol. 14, no. 1, Nov. 2023, doi: 10.1038/s41467-023-42528-4.
[3] C. J. Williams et al., “MolProbity: More and better reference data for improved all-atom structure validation,” Protein Science, vol. 27, no. 1, pp. 293–315, Nov. 2017, doi: 10.1002/pro.3330.
[4] T. Shen et al., “E2Efold-3D: End-to-End Deep Learning Method for accurate de novo RNA 3D Structure Prediction,” arXiv, Jul. 04, 2022. Available: http://arxiv.org/abs/2207.01586

[5] A. Kalampaliki, A. C. Dimopoulos, and M. Reczko, “A Root Mean Square Deviation Estimation Algorithm (REA) and its use for improved RNA Structure Prediction,” bioRxiv (Cold Spring Harbor Laboratory), Mar. 2024, doi: 10.1101/2024.02.28.582508.
[6] R. Pearce, G. S. Omenn, and Y. Zhang, “De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning,” bioRxiv, p. 2022.05.15.491755, May 15, 2022. doi: 10.1101/2022.05.15.491755

ΑΝΑΛΥΣΗ ΤΗΣ ΠΟΡΕΙΑΣ ΥΛΟΠΟΙΗΣΗΣ ΤΗΣ ΕΕ

Η ΕΕ3 περιέλαβε τη δημιουργία Έμπιστων Περιβαλλόντων Έρευνας (Trusted Research Environments - TREs) που επιτρέπουν την ανάλυση ευαίσθητων δεδομένων χωρίς αυτά να μετακινούνται από την υποδομή του παρόχου. Εντός αυτού του περιβάλλοντος αναπτύχθηκε και ενσωματώθηκε η διαδικτυακή υπηρεσία REA για την πρόβλεψη τρισδιάστατης δομής RNA. Η υπηρεσία αξιοποιεί προηγμένους αλγορίθμους βαθιάς μάθησης (DeepFoldRNA, RhoFold κ.α.) και μεθόδους αξιολόγησης (REA, ARES).

Αναμενόμενα Αποτελέσματα – Οφέλη Κάθε Φορέα

Για το Φλέμιγκ: Εδραιώθηκε ως πάροχος ασφαλών υπολογιστικών περιβαλλόντων «Data-to-Researcher», ενισχύοντας την κυριαρχία των δεδομένων (data sovereignty). Η ανάπτυξη της υπηρεσίας REA προσφέρει ένα ισχυρό εργαλείο για τη δομική και λειτουργική μελέτη μορίων RNA που σχετίζονται με νευροεκφυλιστικές νόσους.
Για τους Λοιπούς Φορείς: Οι ερευνητές εκτός Φλέμιγκ μπορούν να εκτελούν βιοπληροφορικές ροές εργασίας πάνω σε ευαίσθητα δεδομένα μέσα σε ένα πλήρως εξοπλισμένο (R, Python, κ.λπ.) και ασφαλές περιβάλλον, εξασφαλίζοντας απόλυτη συμμόρφωση με τον GDPR. Η πρόσβαση στα εργαλεία πρόβλεψης δομής RNA υψηλής ακρίβειας υποστηρίζει άμεσα τους επιστημονικούς στόχους όλων των φορέων που συμμετέχουν στην εμβληματική δράση.