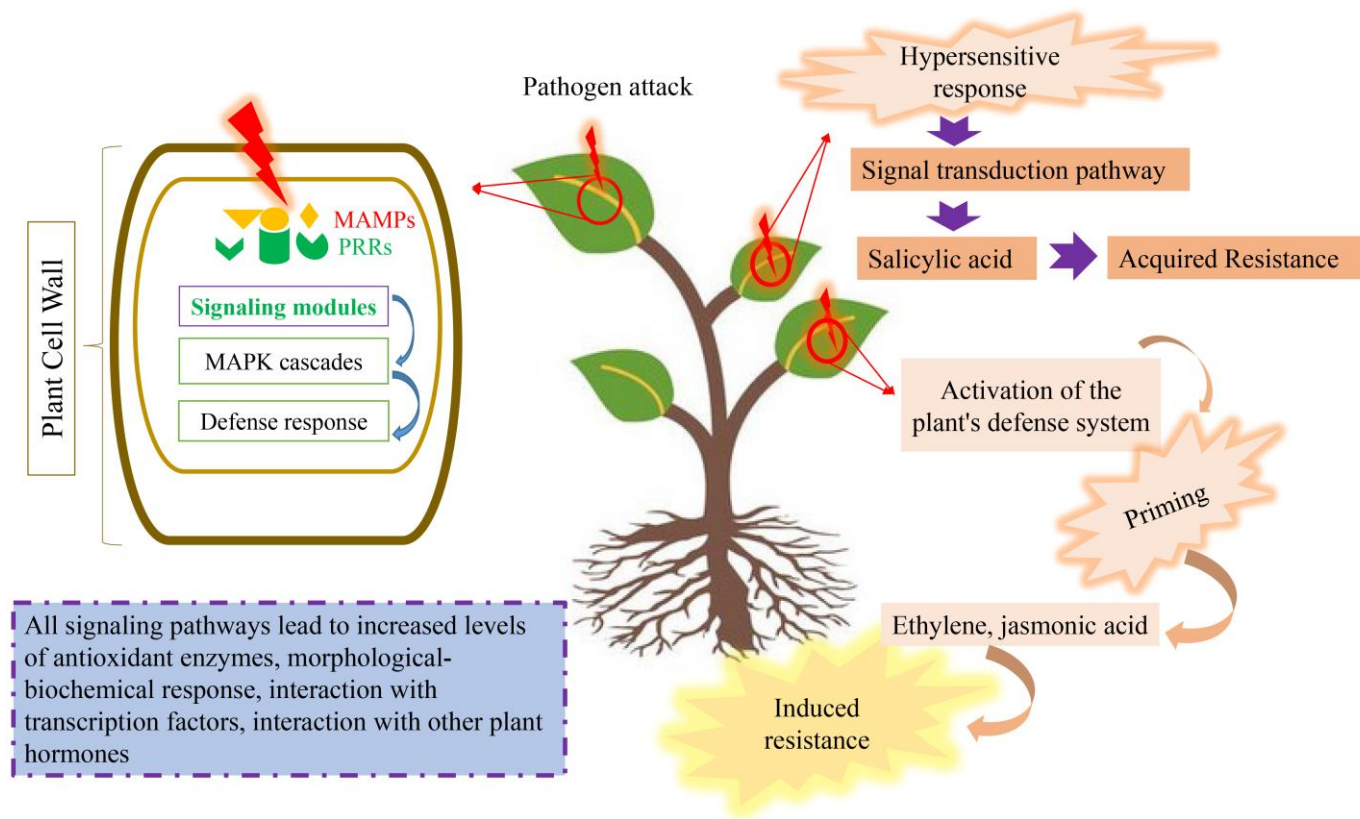


Növényi rezisztencia gének azonosítása NGS bioinformatikai adatokból AI modell tanításához

Hidvégi Norbert Tibor
HUN-REN ATK Mezőgazdasági Intézet

- Növények folyamatos kölcsönhatásoknak vannak kitéve
- Gombák, baktériumok, vírusok, fonálférgek, rovarok
- 30-40%-os terméskiesés
- Klímaváltozás tovább növeli a
- járványok előfordulását



Riseh, Roohallah Saberi, and Mozghan Gholizadeh Vazvani. "Unveiling methods to stimulate plant resistance against pathogens." *Frontiers in Bioscience-Landmark* 29.5 (2024): 188.

1. Érzékelik a kórokozó specifikus molekuláit (effektorok)
 2. Immunválaszt indítanak
- Leggyakoribb R-gén típusok (NLR, RLP, RLK)
 - Az NLR típusú gének a rezisztenciagének 80%-át adják
 - Kulcsfontosságúak a fenntartható mezőgazdaságban, hogy vegyszeres védekezés helyett genetikai ellenállást használjunk

- Hagyományos nemesítés
- Marker-alapú szelekció
- Transzgénikus megoldások
- Genomszerkesztés
- Pan-genom alapú rezisztenciafeltérképezés

Cél a stabil, tartós, széles spektrumú rezisztencia kialakítása

Az újgenerációs szekvenálás (NGS) lehetővé teszi:

- Teljes genomok olcsó és gyors szekvenálását
- Transzkriptomok feltérképezését (RNA-seq)
- Génexpressziós mintázatok megismerését
- Nagy populációk genotipizálását

R-gének főbb kihívásai:

- Nagy variabilitásúak
- Gyakran ismétlődő szekvenciákat tartalmaznak
- Klaszterekben rendeződnek

- Képes felismerni mintázatokat szekvenciákban
- Sokdimenziós adatokból tanul
- Képes a nem annotált genomi szekvenciákban jelölteket találni

Input adatok:

Szekvencia motívumok
Doménarchitektúrák
Evolúciós konzerváció
Génexpressziós profilok

Alkalmazások:

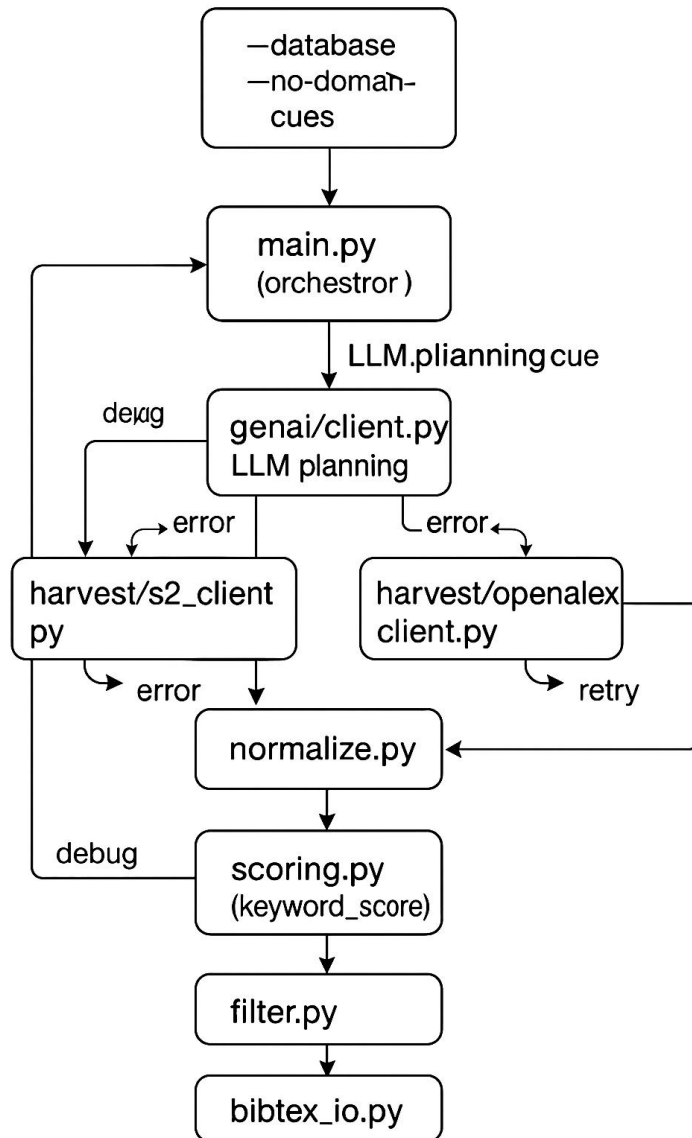
R-gén predikció
Egyedi doménkombinációk felismerése
Új jelölt géncsoportok felfedezése
Genetikai variánsok funkcionális értékelése

Adattípusok:

- Genom (WGS)
- Transzkriptom (RNA-seq)
- Pan-genom adatok

Főbb nyilvános adatbázisok:

- NCBI
- Ensembl Plants
- PLAZA, Phytozome
- PRGdb (Plant Resistance Genes Database)
- Sematic Scholar, OpenAlex



Bemenet és kimenet – bibtex_io.py

- Beolvassa a felhasználó által megadott BibTex formátumú fájlt

Lekérdezés-generálás – client.py és prompts.py

- GenAI LLM (GPT-OSS:120B) elemzi a cikkeket
- Kulcsszavak és boolean keresések tervezése

Keresés – s2_client.py és openalex_client.py

- Semantic Scholar
- OpenAlex

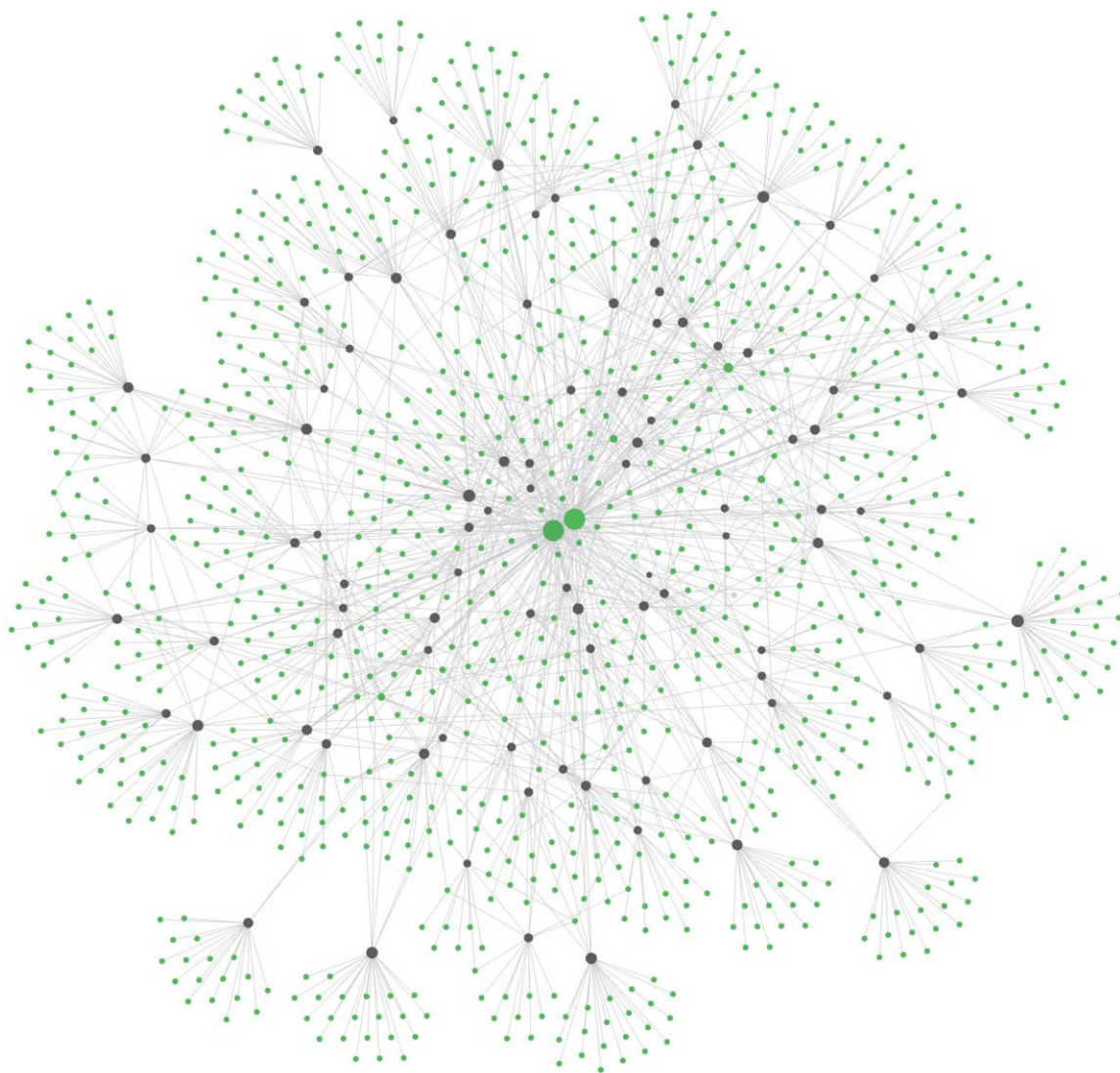
Adattisztítás – normalize.py és dedup.py

- Az eltérő formátumú válaszokat egységes szerkezetű JSON rekordokká alakítja
- Azonos DOI-jú cikkeket összevonja

Pontozás – scoring.py és filter.py

- Tartalmi és kulcsszavas hasonlóság
- LLM dönt

1. PDF és JSON/BibTex fájlok beolvasása
2. Metaadatok kinyerése (CrossRef, Semantic Scholar, OpenAlex)
3. Cikk összefoglalása LLM-el
4. Egységes markdown fájl
5. Cikkben lévő adatbázisok kinyerése
6. Kapcsolódó cikkek listázása
7. LLM-el csoportosított kulcsszavak



Summary

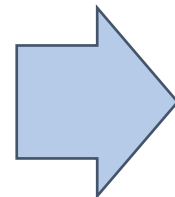
- Produced a **chromosome-scale, annotated reference genome** for rye (*Secale cereale*) cultivar **Lo7** ($\approx 6.2\text{--}7.9\text{ Gb}$, 96 % + BUSCO completeness).
- Integrated **long-read, Hi-C, 10× linked-reads, Bionano optical maps** and genetic maps to achieve high contiguity ($N50 \approx 29\text{ Mb}$, >92 % of sequence placed in pseudomolecules).
- Validated the assembly with BAC libraries, FISH, population resequencing, and extensive comparative analyses with barley and wheat.
- Demonstrated immediate utility for **structural-variation (SV) discovery, population genomics, introgression detection, and breeding applications** (fertility restoration, disease resistance, cold-tolerance, yield gains).

Key Findings

- **Genome architecture**
 - Rye genome $\sim 7\text{--}8\text{ Gb}$, $\sim 50\%$ larger than wheat/barley subgenomes.
 - Pericentromeric low-collinearity regions and high repeat density ($\approx 15\text{ k}$ full-length LTR-RTs, four dominant families).
 - Gene density increases >10-fold toward telomeres; 34 441 high-confidence genes annotated.
- **Structural variation**
 - Four megabase-scale inversions between Lo7 and Lo225, including a 50-Mb inversion on 5R that suppresses recombination.
 - Large SVs concentrate in pericentromeric regions, reducing collinearity and contributing to genetic isolation.
- **Population genomics** ($\approx 955\text{ GBS samples} + 347\text{ wild } Secale$)
 - Seven genetic clusters (domesticated-like, *S. strictum*-like, *S. sylvestre*-like).
 - Elevated Pn/Ps ratios in low-recombination regions indicate weakened purifying selection.
 - D-statistics reveal recent gene flow between rye groups and introgression into wheat (especially chromosome 1R).
- **Introgression detection**
 - Developed a **support-vector-machine (SVM) classifier** using normalized read-depth (rpm) across 1-Mb bins to discriminate non-introgressed, 1A·1R, 1B·1R, and ambiguous karyotypes ($\geq 97\%$ accuracy).
 - Applied to >5 800 wheat lines (WHEALBI, CIMMYT, KSU, USDA-RPN) and >1 200 predicted 1R carriers, quantifying yield effects (0.86–4.55 % increase).

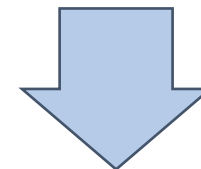
Preprocessing

- QC: FastQC
- Trimming: Trimmomatic
- Read mapping: BWA / HISAT2
- Assembly (ha kell): Trinity, SPAdes



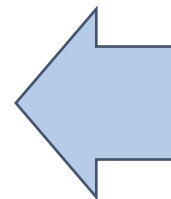
Annotáció

- BLAST
- InterProScan
- Pfam/SMART domain keresés



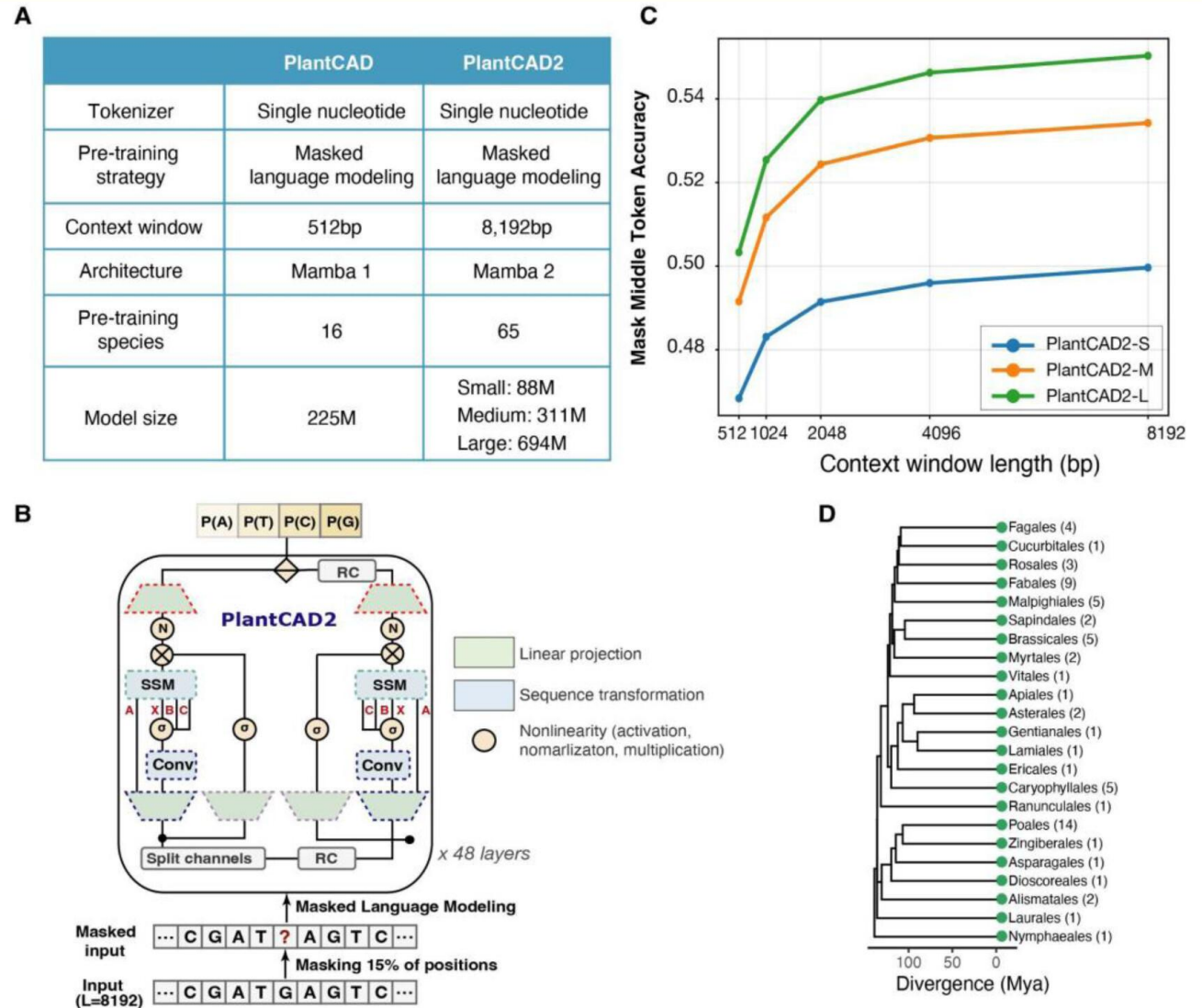
R-gén azonosítás

- NLR-Annotator
- DRAGO2
- RGAugury



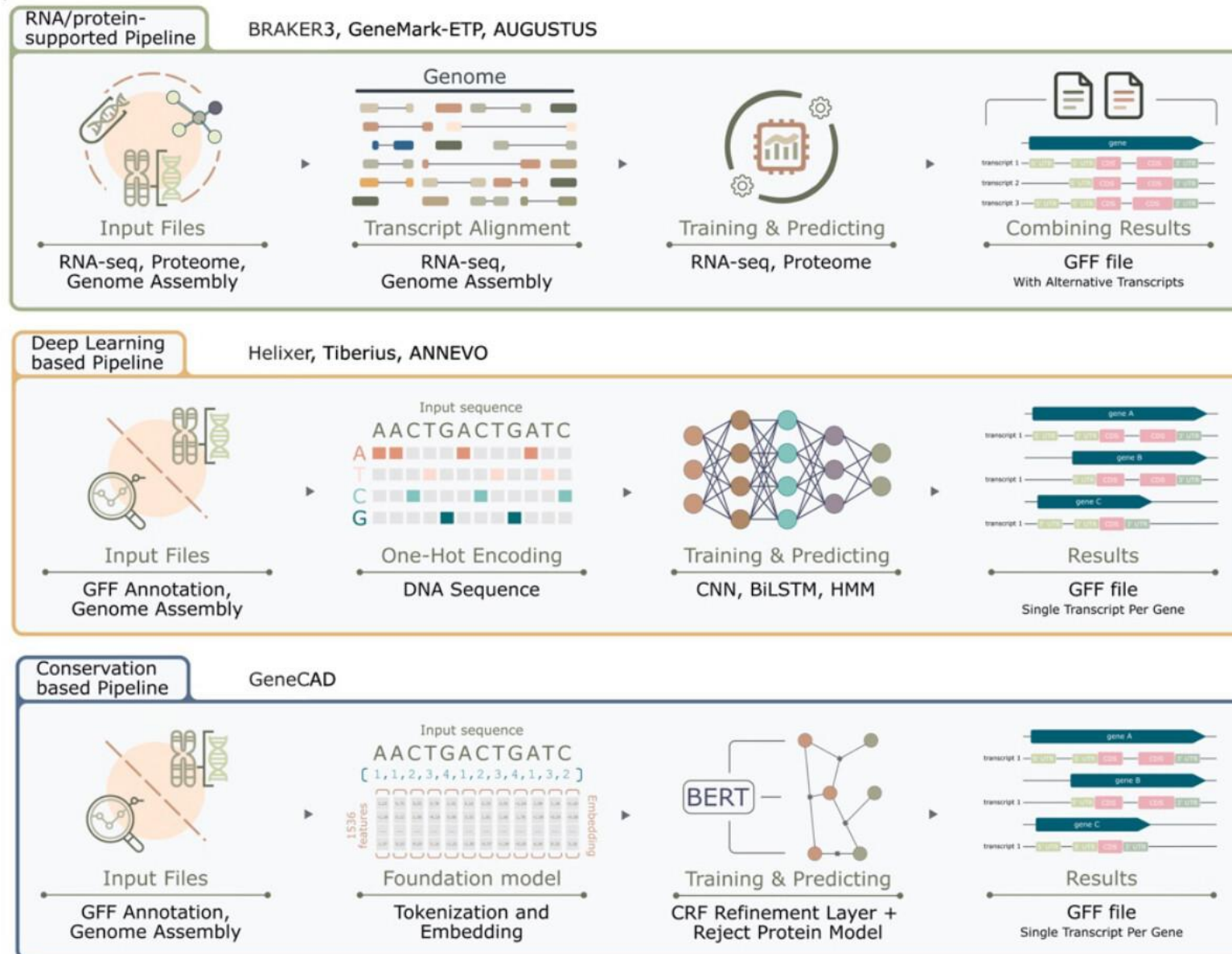
Szekvenciák
Domain struktúra
Motívumok
Kifejeződési profil

- PlantCAD v1 – 16 angiosperm genom
- PlantCAD v2 – 65 angiosperm genom

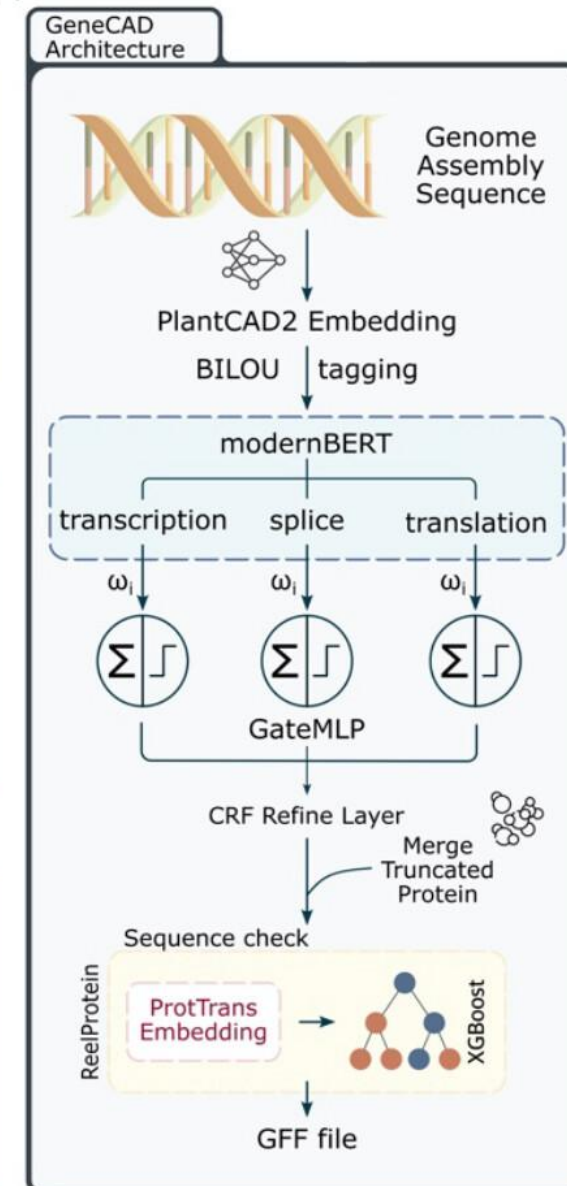


Zhai, Jingjing, et al. "PlantCAD2: A Long-Context DNA Language Model for Cross-Species Functional Annotation in Angiosperms." *bioRxiv* (2025): 2025-08.

(a)



(b)



1. HUN-REN Cloud (Validálás)
 - 32 vCPU, 64 GB RAM, 1 TB SSD, 4 TB HDD
2. HUN-REN ATK MGI saját szerver (Adatbányászat, NGS workflow)
 - 128 CPU, 1 TB RAM, 69 TB SSD
3. DKF HPC Kompetencia Központ – Komondor HPC (PlantCAD2 és GeneCAD pre-training)
 - BigData partíció: 288 CPU, 12 TB RAM
 - MI partíció: 512 CPU, 2 TB RAM, 32 db Nvidia A100 GPU

Tudományos cikkek és adatbázisok alapján a PlantCAD2modell pre-training

Saját idegenfajú keresztezéseinkből származó növénypopulációkon
tesztelés/validálás

Köszönöm a figyelmet!

