

Nagy nyelvi modellek alkalmazása DNS szekvenálási metaadatok standardizálására

Ari Eszter és Kada Norbert

HUN-REN Szegedi Biológiai Kutatóközpont, Biokémiai Intézet

II. Országos HUN-REN Cloud Találkozó

2025. 11. 13.

Dr

Ari Eszter



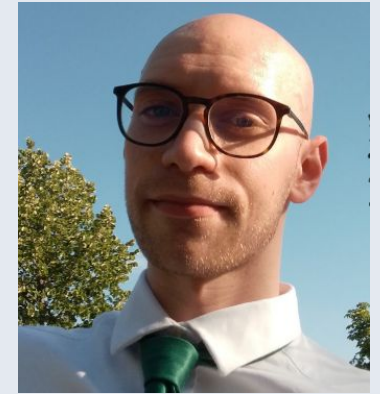
- HUN-REN Szegedi Biológiai Kutatóközpont, Biokémiai Intézet, tudományos munkatárs
- ELTE Eötvös Loránd Tudományegyetem, Genetikai Tanszék, docens

@ ari.eszter@brc.hu

☎ +36 70 380 9833

🌐 <https://genet.elte.hu/bioinformatika>

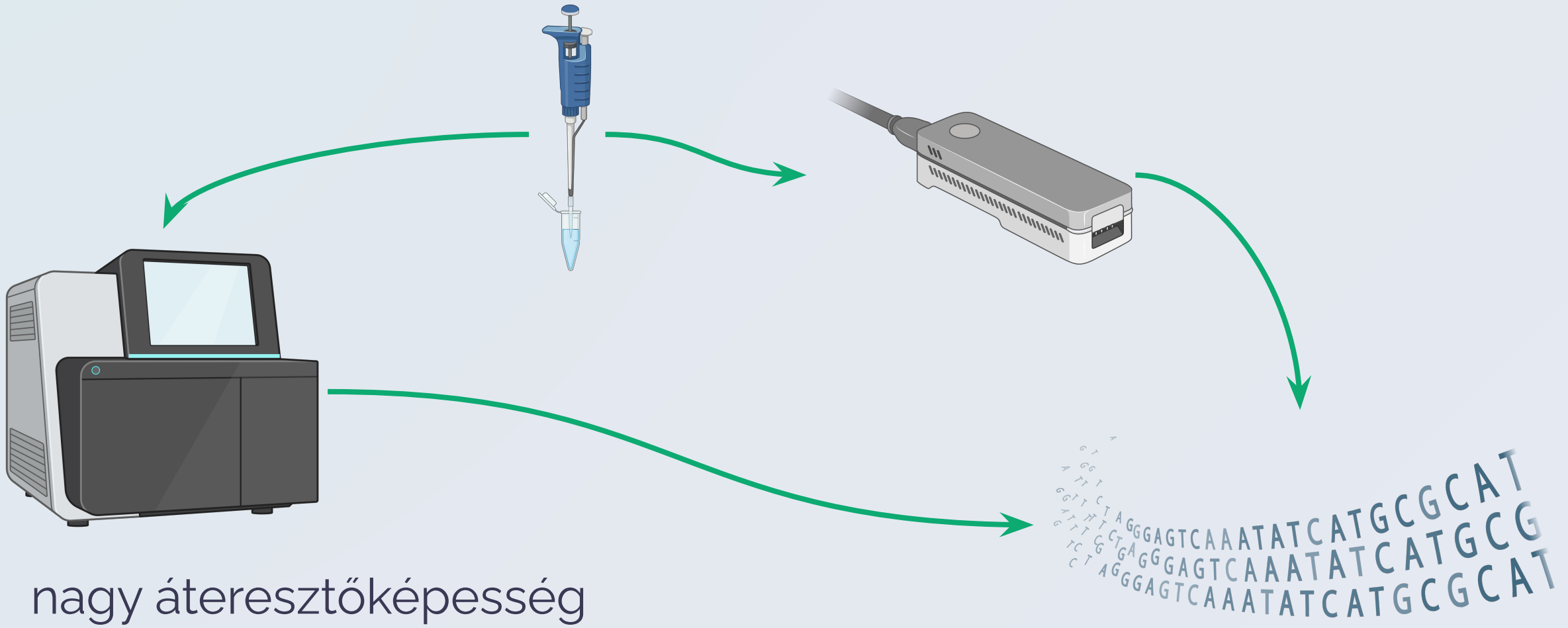
Kada Norbert



- HUN-REN Szegedi Biológiai Kutatóközpont, Biokémiai Intézet, szerződött hallgató
- ELTE Eötvös Loránd Tudományegyetem, biológus mester hallgató

kada.norbert@gmail.com

DNS szekvenálás: Technológia



DNS szekvenálás: Adatmennyiség

- Elérhető *nagy áteresztőképességű* szekvenálási adatok mennyisége:
 - feldolgozott: 3,17 millió összeszerelt genom ([NCBI Genome](#))
 - “feldolgozatlan”: 38 millió elérhető ([NCBI SRA](#))
- Mindegyikhez tartozik **metaadat** → [NCBI BioSample](#) adatbázis

Biosample bejegyzés példa

Pathogen: clinical or host-associated sample from Escherichia coli		
Identifiers	BioSample: SAMN53123345; Sample name: HBH_654H_18; SRA: SRS27016153	
Organism	Escherichia coli cellular organisms; Bacteria; Pseudomonadati; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia	
Package	Pathogen: clinical or host-associated; version 1.0	
Attributes	Isolate	HBH_654H_18
	collected by	Habib Bourguiba Hospital
	collection date	2018
	geographic location	Tunisia: Sfax
	host	Homo sapiens
	host disease	HAI
	isolation source	wound
	latitude and longitude	34.740325 N 10.748221 E
BioProject	PRJNA1358714 Retrieve all samples from this project	
Submission	Laboratoire de Microbiologie, CHU Habib Bourguiba, Basma MENIF; 2025-11-09	

Big data, Top-down megközelítés → metaadat probléma

- Olyan kutatási témákon dolgozunk, melyek **nagy mennyiségű**, 10-100 000 genomszekvenálási **adat** feldolgozásával kezdődnek.
- Meg kell találnunk a megfelelő szekvenciákat a 40 milliós adatbázisban
- **Probléma: nem egységesek metaadatok**
 - de még a kulcsok sem (pl. "*treatment*", "*treatement*")
 - → Egyszerű keresés

Leggyakoribb

genotype
WT
wild type genotype
wild type
missing
Wild type
wild-type
wildtype
Wild-type
wt
Wildtype



A nagy adatbankok egyelőre nem egységesítik a metaadatokat (forráshiány)

**National Library of Medicine**
National Center for Biotechnology Information

Log in

i Notice

Because of a lapse in government funding, the information on this website may not be up to date, transactions submitted via the website may not be processed, and the agency may not be able to respond to inquiries until appropriations are enacted. The NIH Clinical Center (the research hospital of NIH) is open. For more details about its operating status, please visit cc.nih.gov. Updates regarding government operating status and resumption of normal operations can be found at opm.gov.

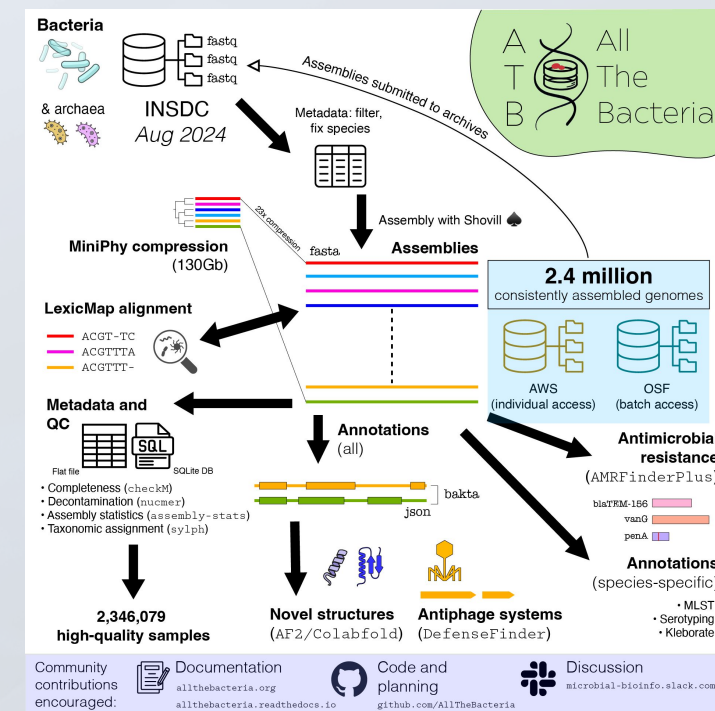
**EMBL**

About EMBL > [EMBL Communications](#) > [News](#) Related: [Press office](#) [All EMBL sites](#)

 **Oana Stroe**

20 March 2024
People & Perspectives

Sense makers: how standards are enabling data reuse

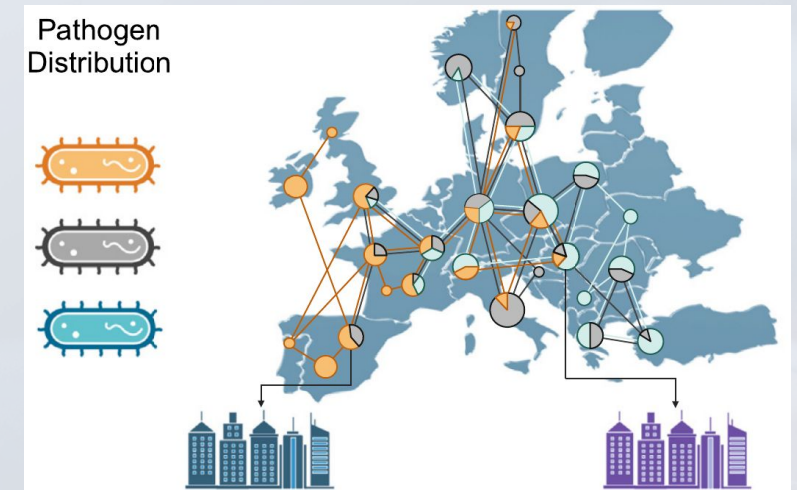


régi vs. újabban bevitt adatok

Érdekes projektek → metaadat probléma

Milyen sebességgel terjednek **időben** és **térben** az egyes kórokozó típusok?

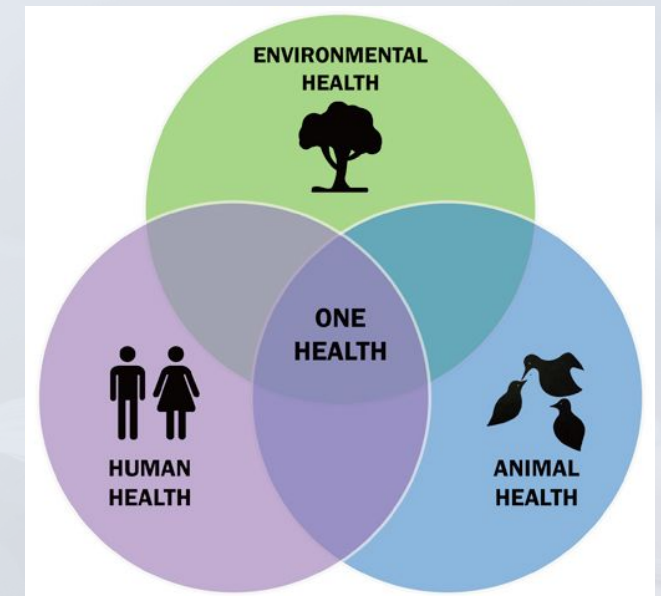
- Nem egységesített a **mintavételi időpont** és **földrajzi hely** leírása
 - *Mintavételi időpont*: csak **30%-ban követi az ISO formátumot** (yyyy-mm-dd)
 - *Országnevek*: ISO3 kódok helyett szabadon beírt, gyakran helyesírási hibás



Koncz,M. et al. (2024) *Cell*, 187, 5901

Érdekes projektek → metaadat probléma

- Van-e összefüggés egy baktérium típus
 - **gazdaszervezetei** (pl. ember, sertés),
 - **élettere**i (mintavételi forrás, pl. vér, szennyvíz)
 - és *antibiotikum rezisztenciája* között?
- De nem egységesített a **gazda** és a **mintavételi forrás** leírása
 - pl: **VÉR**: BLOOD CULTURE, Blood culture, Blood-culture, HUS patient, Peripheral blood, blood, blood culture, blood culture from newborn with neonatal bacterimia, blood culture of newborn with neonatal bacterimia, blood from patient with bacteremia, blood sample, blood sample of individual with bacteremia, blood/urine, bloodstream, newborn blood culture, patient blood, patient with hemolytic uremic syndrome (HUS), sepsis, vasculated blood of an human, whole blood



Metaadat sztenderdizálás

- **Szabály alapú** (*regex*)
 - kulcsszavak keresése (szintaktikai szabályokkal)
 - **problémák**: elgépelések, negatív kontextus, **nyelvi változatosság**
- **Gépi tanulás**
 - **probléma**:
 - erősen feladat-specifikus modellek → **rugalmasság**

Megoldás: Nagy nyelvi modellek?



Nagy nyelvi modellek (LLM)

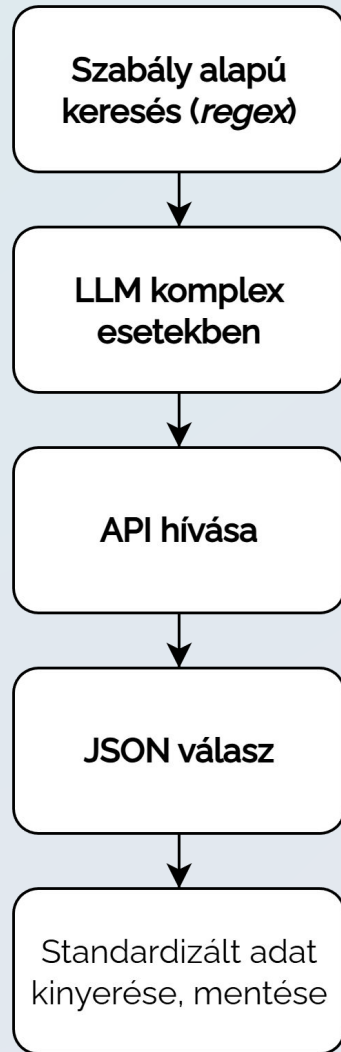
1. Hatalmas mennyiségű, **ember által írt szövegen tanulnak**
2. A legtöbb metaadat az **inkonzisztenciák ellenére is az ember számára értelmezhető**

↳ képesek “értelmezni” (prediktálni) olyan szöveget, amit **szabályok alapján működő programok nem & általánosíthatóak**

Tudományos munka - fontos paraméterek:

- **temperature** = 0 → valószínűség
- **fixált seed** → reprodukálhatóság

Munkafolyamat



Feladatonként különböző, de pl. dátumoknál elegendő önmagában.

Pl. mintavételi forrás (*K. pneumoniae* esetében **1700 különböző érték** → **10-15 kategória**)

HUN-REN Cloud GenAI szolgáltatásának szerverei

Nagyon ritka a hibás output!



Prompt engineering

Mintavételi forrás

Analyze the following bacterial isolation source and classify it into one of the categories.

Categories:

- laboratory
- hospital
- ..etc..

Respond ONLY with a **single JSON object** containing the chosen category.

Example: `{{"category": "urinary"}}`

The original value is: "{value}".

The host is: "{host}".



Eredmények

Mintavételi forrás

Nagy nyelvi modellel feldolgozott példák

Metaadat	LLM kategória
freshwater steam sediment	environmental
bronchial washing	respiratory
Abdominal fluiduid	gastrointenstinal
penile lesion	wound
CL-BLD	blood
Pre-washed mixed salad, import product	food



LLM-ek használata - tanulságok

- Sok adatsor feldolgozása egyszerre problémás → **Lassú**
- A felelősség a kutatóé → **Ellenőrzés**

De **páratlan képességű, automatizálható** módszer!



Meta-analízisok elősegítése

Köszönetnyilvánítás

Papp Balázs
HUN-REN SZBK



Grézal Gábor
HUN-REN SZBK



Köszönjük a **technikai segítséget**: *GenAI4Science*: Szabó Gergő és Márton István

Támogatónk: *HUN-REN Támogatott Kutatócsoportok Programja*: TKCS-2024/66

A projekt megvalósítói:



Köszönjük a figyelmet!

Bármilyen kérdés?

