Multiparametrické modelovanie mitochondriálnej DNA pre diagnostiku rakoviny

Prezentácia priebežnej práce na diplomovej práci

Bc. Ondrej Škorňák

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

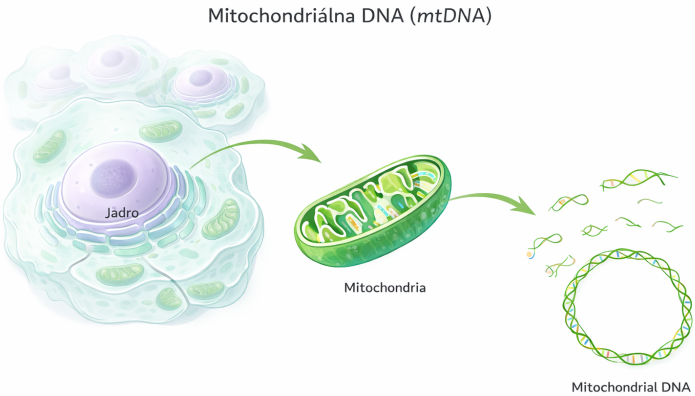
Vedúci práce: Mgr. Jaroslav Budiš, PhD.

Máj 2026

Obsah

Čo je mitochondriálna DNA

- mtDNA je „druhý genóm“ bunky
- nachádza sa v mitochondriách
- má 16 569 bp
- je kruhová, dvojvláknová
- obsahuje 37 génov
  - 13 pre proteíny OXPHOS
  - 22 tRNA
  - 2 rRNA



Zdroj: BioNumbers BNID 105470; Anderson et al., 1981; Chan, 2006.

Zastúpenie mtDNA v bunke

- **Jadrový genóm:** $\sim 6,3$ miliardy bp
- **1 molekula mtDNA:** 16 569 bp

Podiel z celkovej bunkovej DNA závisí od počtu kópií (typu bunky):

- **Krvná bunka:** $\sim$ 300 kópií $\rightarrow \sim 0{,}08\,\%$ celkovej DNA
- **Bežná somatická bunka:** $\sim$ 1 000 kópií $\rightarrow \sim 0{,}26\,\%$ celkovej DNA
- **Srdcový sval:** $\sim$ 4 000 – 6 000 kópií
- **Oocyt:** $\sim$ 100 000 kópií $\rightarrow \sim 20\,\%$ celkovej DNA

Význam pri rakovine

Aj v bežných bunkách tvorí mtDNA merateľnú časť DNA. V **rakovinových bunkách** sa počet kópií mtDNA (tzv. *copy number*) môže výrazne meniť nahor aj nadol podľa typu nádoru. Je to významný diagnostický marker.

Zdroj: NHGRI; BioNumbers BNID 105470; Shoubridge & Wai, 2007; Abd Radzak et al., 2022.

Úloha mtDNA v rakovinových bunkách

- mtDNA nie je len pasívny marker, ale **aktívna súčasť** nádoru
- Pozorované zmeny:
  - **mutácie** (chyby v sekvencii)
  - **heteroplazmia** (viac rôznych verzií naraz)
  - **počet kópií** (copy number)
- $\rightarrow$ mení sa „spracovanie energie“ bunky
- $\rightarrow$ podpora prežitia a rastu nádoru
- Efekt je špecifický pre daný typ rakoviny

Zdroj: Abd Radzak et al., 2022.

Voľne cirkulujúca mtDNA (ccf-mtDNA)

- Mitochondriálna DNA prítomná v krvnej plazme mimo buniek (cf-mtDNA / ccf-mtDNA).
- Do voľného obehu sa dostáva pri:
  - **apoptóze** (programovanej bunkovej smrti),
  - **nekróze** (nekontrolovanom odumieraní buniek),
  - **aktívnej sekrécii** (napr. v extracelulárnych vezikulách).
- Má odlišné **fragmentačné vlastnosti** než nukleárna cfDNA.
- **Diagnostický potenciál:** jej koncentrácia a vlastnosti môžu niesť informáciu o prítomnosti nádorového ochorenia.

Zdroj: van der Pol et al., 2023; Liu et al., 2024.

Tekutá biopsia

- **Tekutá biopsia** je diagnostická metóda využívajúca analýzu biologických tekutín (najčastejšie krvi) na detekciu nádorových biomarkerov.
- Predstavuje len **minimálne invazívny** zásah do ľudského tela.

Výhody

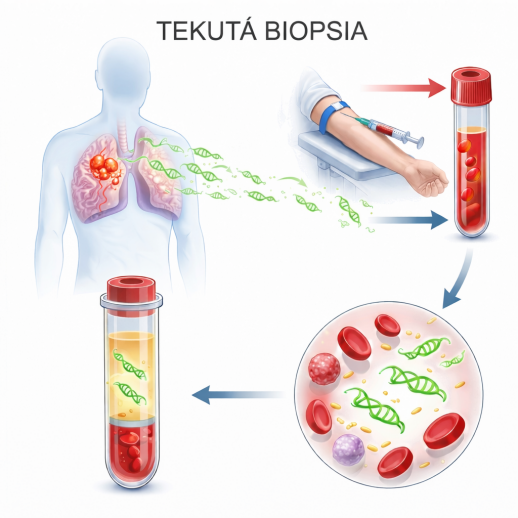
- znížené riziko pre pacienta
- možnosť opakovaného odberu (monitorovanie v čase)
- zachytáva heterogenitu nádoru

Nevýhody

- veľmi nízka koncentrácia biomarkerov (cfDNA/cf-mtDNA)
- analytická a technologická náročnosť detekcie
- zložitejšia štandardizácia procesov

Zdroj: Ma et al., 2024.

Tekutá biopsia



Zdroj: Ma et al., 2024.

Obsah

Kontext a cieľ diplomovej práce

- **Nadväznosť na bakalársku prácu:**
  - Predchádzajúca analýza celkovej cfDNA z tekutej biopsie pomocou modelov strojového učenia.
  - Diplomová práca mení zameranie špecificky na **mitochondriálnu DNA (mtDNA)**.
- **Vstupné dáta:**
  - Sekvenačné dáta z tekutej biopsie.
  - Z celkovej zmesi DNA sú už **vyfiltrované iba fragmenty mtDNA**.
- **Základná myšlienka:**
  - Nesledovať fragmenty len izolovane, ale prísť s **multiparametrickým prístupom** k ich analýze.

Spracovanie dát a metodika

- Každý fragment mtDNA posudzujeme cez **viacero vlastností súčasne**:
  - dĺžka fragmentu,
  - prítomnosť a typ mutácií,
  - poloha v mitochondriálnom genóme.

- **Nástroje na analýzu:**
  - **Štatistické metódy** – hľadanie a potvrdzovanie rozdielov medzi skupinami.
  - **Metódy strojového učenia** – tvorba robustného predikčného modelu.

Hlavný cieľ práce

Vytvoriť multiparametrický model, ktorý pomôže identifikovať **fragmenty s vyšším onkologickým rizikom** a umožní tak lepšie **rozlíšiť pacientov od kontrolných jedincov**.

Zdroj: Liu et al., 2024.

Dostupné dáta (interné kohorty)

Kohorta PreveLynch:

- $n = 976$ vzoriek plazmy (cfDNA)
- 753 zdravých kontrol
- 223 pacientov s kolorektálnym karcinómom

Kohorta GenoScan:

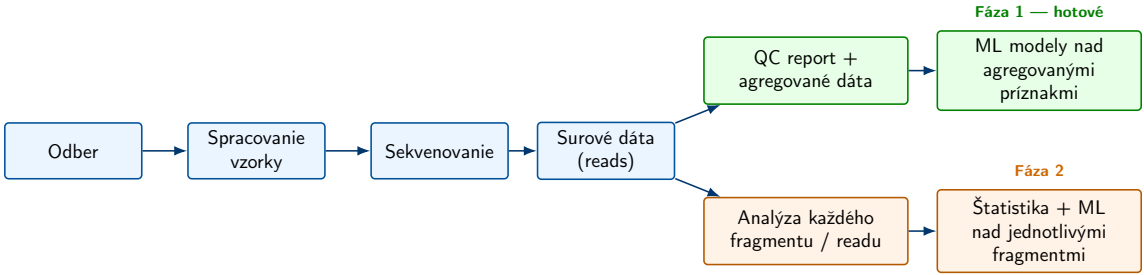
- $n = 453$ vzoriek plazmy (cfDNA)
- 380 zdravých kontrol
- 73 pacientov s rakovinou prostaty

Špecifiká a výzvy:

- Silná nevyváženosť tried ($\sim 3:1$ kontroly:pacienti).
- Niektorí pacienti s viacerými odbermi v čase $\Rightarrow$ nutné zoskupovanie na úrovni pacienta.
- Rôzne typy rakoviny $\Rightarrow$ možnosť hľadať efekty špecifické pre typ nádoru.
- **Nové dáta:** Prístup k 3 ďalším menším datasetom (každý má 200 vzoriek).

Obsah

Plán práce: workflow analýzy



Dve úrovne analýzy v diplomovej práci

Fáza 1 — agregované dáta

- pracujem s **globálnymi charakteristikami celej vzorky**
- zo surových dát extrahujem napr.:
  - chybovosť
  - priemernú dĺžku fragmentov
  - % čítaní mapovaných na mtDNA
  - pokrytie mitochondriálneho genómu
- nad týmito príznakmi trénujem **ML modely**
- Táto časť je už hotová

Fáza 2 — jednotlivé fragmenty

- idem **hlbšie než na úroveň celej vzorky**
- budem analyzovať **každý mtDNA read samostatne**
- pri každom fragmente sledujem:
  - dĺžku
  - mutácie
  - polohu v mtDNA genóme
- využijem **štatistiku + strojové učenie**

Hlavná myšlienka

Fáza 1 hľadá signál na úrovni **celej vzorky**. Fáza 2 hľadá signál priamo na úrovni **jednotlivých mtDNA fragmentov**, čo je jadro diplomovej práce.

Obsah

Experimenty: porovnané panely

Tri sady príznakov

- **Core12:** pôvodný ručne kurátorovaný panel
- **All36:** plná V4 sada; výkonový benchmark
- **Strict Core 8:** biologicky prísny panel bez raw countov, MAPQ a error-rate metrík

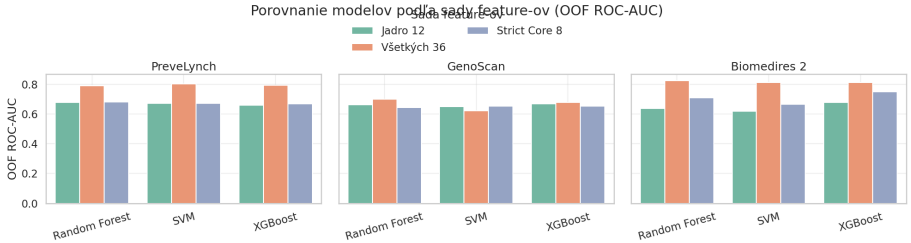
Evaluácia

- 3 kohorty: PreveLynch, GenoScan, Biomedires 2
- modely: Random Forest, SVM, XGBoost
- **StratifiedGroupKFold** podľa pacienta
- metriky: OOF ROC-AUC, PR-AUC, bootstrap 95 % CI

Kľúčová otázka

Ostane signál zachovaný aj po odstránení technicky rizikových a redundantných premenných?

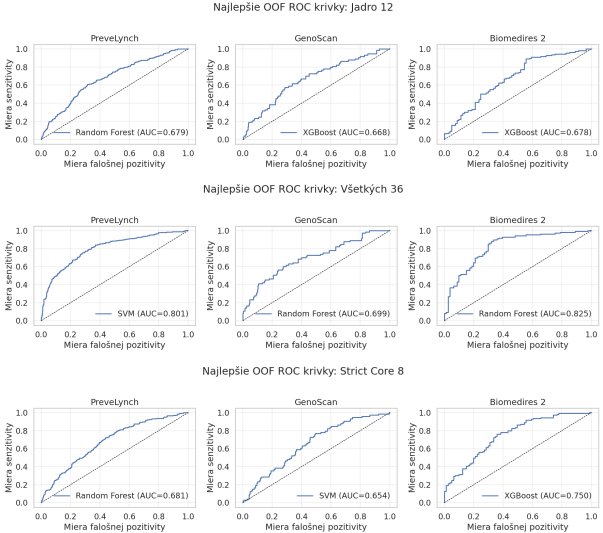
Výsledky — výkon panelov



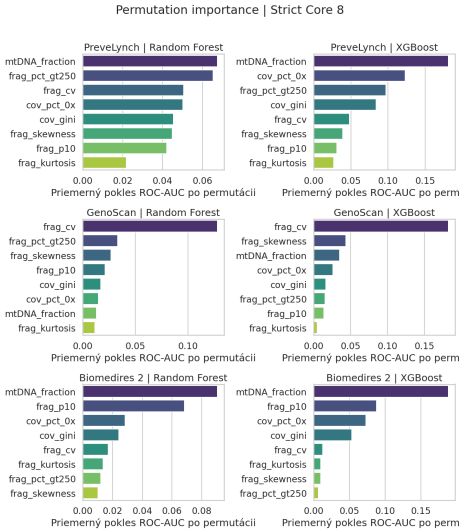
Kohorta	Core12	Strict Core 8	All36
PreveLynch	0,679 / 0,389	0,681 / 0,386	**0,801 / 0,571**
GenoScan	0,668 / 0,284	0,654 / 0,253	**0,699 / 0,388**
Biomedires 2	0,678 / 0,721	0,750 / 0,796	**0,825 / 0,847**

Hodnoty sú OOF ROC-AUC / PR-AUC najlepšieho modelu v danej kohorte a sade príznakov.

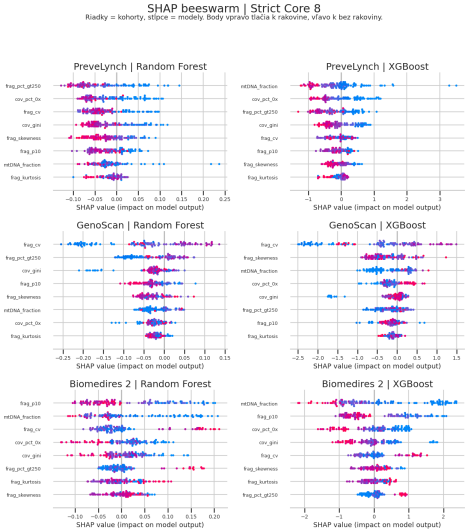
Výsledky — ROC krivky najlepších modelov



Výsledky — Dôležitosť príznakov



Výsledky — Smer efektu



Zdroj: Lundberg & Lee, 2017.

Interpretácia experimentov

- **All36** dosahuje najvyšší výkon, ale významnú časť signálu tvoria read-depth a technické QC premenné.
- **Strict Core 8** zachováva výkon na úrovni Core12; v Biomedires 2 je výrazne lepší ako Core12 (0,750 vs. 0,678 ROC-AUC).
- Najstabilnejšie biologicky čitateľné bloky sú **mtDNA podiel**, **coverage breadth/nerovnomernosť** a **fragmentačné metriky**.
- Rozdiel medzi All36 a Strict Core 8 je dôležitý sanity check: časť výkonu môže byť technický alebo kohortový signál.

Záver pre ďalšiu prácu

Ako hlavný interpretovateľný panel je vhodnejší **Strict Core 8**; All36 ponechať ako horný výkonový benchmark.

Referencie

1. Anderson, S. et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature*. BioNumbers BNID 105470: bionumbers.hms.harvard.edu.
2. National Human Genome Research Institute. Base Pair. genome.gov.
3. Shoubridge, E. A. & Wai, T. (2007). Mitochondrial DNA and the Mammalian Oocyte. *Current Topics in Developmental Biology*, 77, 87–111. doi:10.1016/S0070-2153(06)77004-1.
4. Abd Radzak, S. M. et al. (2022). Insights regarding mitochondrial DNA copy number alterations in human cancer. *International Journal of Molecular Medicine*, 50, 104. doi:10.3892/ijmm.2022.5160.
5. van der Pol, Y. et al. (2023). The landscape of cell-free mitochondrial DNA in liquid biopsy for cancer detection. *Genome Biology*, 24, 229. doi:10.1186/s13059-023-03074-w.
6. Liu, Y. et al. (2024). Aberrant fragmentomic features of circulating cell-free mitochondrial DNA as novel biomarkers for multi-cancer detection. *EMBO Molecular Medicine*, 16, 3169–3183. doi:10.1038/s44321-024-00163-6.
7. Ma, L. et al. (2024). Liquid biopsy in cancer: current status, challenges and future prospects. *Signal Transduction and Targeted Therapy*, 9, 336. doi:10.1038/s41392-024-02021-w.
8. Lundberg, S. M. & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *NeurIPS*. arXiv:1705.07874.

Priestor na otázky