Identifikácia barkódov v dátach nanopórového sekvenovania

autor: Bc. Adrián Goga
školiteľ: doc. Mgr. Tomáš Vinař, PhD.
konzultant: doc. Mgr. Bronislava Brejová, PhD.

Univerzita Komenského v Bratislave

16. júna 2020

Nanopórové DNA sekvenovanie

- DNA je nastrihaná na malé kúsky - čítania, ktoré sú po jednom sekvenované
- Čítania prechádzajú cez malý otvor - *nanopór*, na ktorom sa v pravidelných intervaloch meria prúd
- Rýchlosť prechádzania cez nanopór je variabilná
- Každé čítanie je tvorené postupnosťou hodnôt prúdu nameraných v nanopóre - signál
- Osekvenované čítania sa prekladajú do reťazcov nad abecedou $\{A, C, G, T\}$, avšak preklady sú chybové

Barkódovanie DNA čítaní

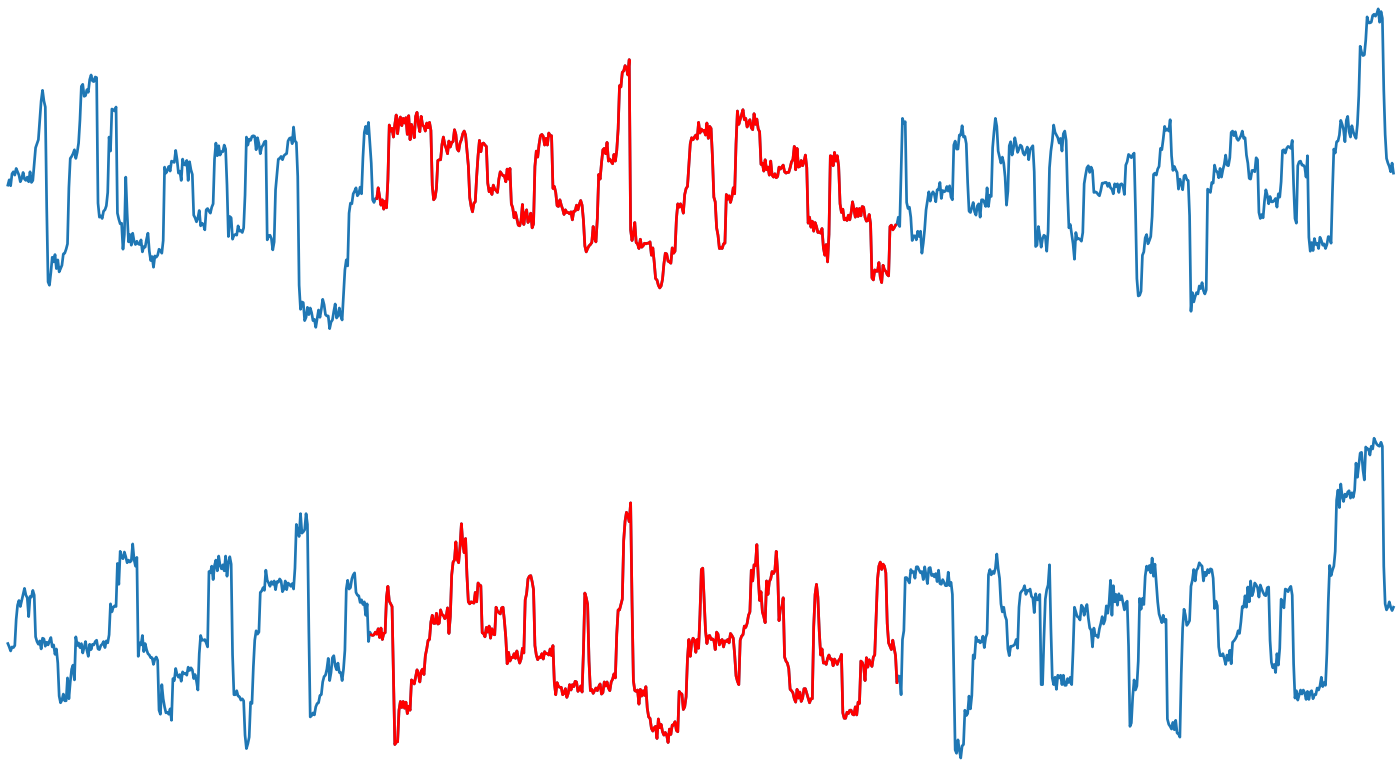
- Je výhodnejšie sekvenovať viacero DNA vzoriek v jednom behu sekvenátora
- Každé čítanie obsahuje krátku sekvenciu - tzv. **barkód**, ktorý slúži ako jednoznačný identifikátor vzorky
- Problém: chyby v preklade signálu barkódu $\Longrightarrow$ čítanie nemožno jednoznačne priradiť

Ciele práce

- Väčšina existujúcich nástrojov pracuje s preloženými sekvenciami ($20 - 25\%$ nepriradených čítaní kvôli chybám v preklade)
- Deepbinner (Wick a spol., 2018) - hlboká konvolučná neurónová sieť klasifikujúca surový signál (iba $\approx 5\%$ čítaní nepriradených, potrebuje veľa trénovacích dát)
- Náš cieľ: použiť **učenie bez učiteľa**
- Neuvažujeme znalosť sekvencií barkódov
- Budeme pracovať so surovým signálom pre čo najvyššiu presnosť

Návrh metódy učenia bez učiteľa

1. Porovnať dvojicu čítaní na základe podobnosti ich signálov v oblastiach, v ktorých sa nachádzajú barkódy, napr:

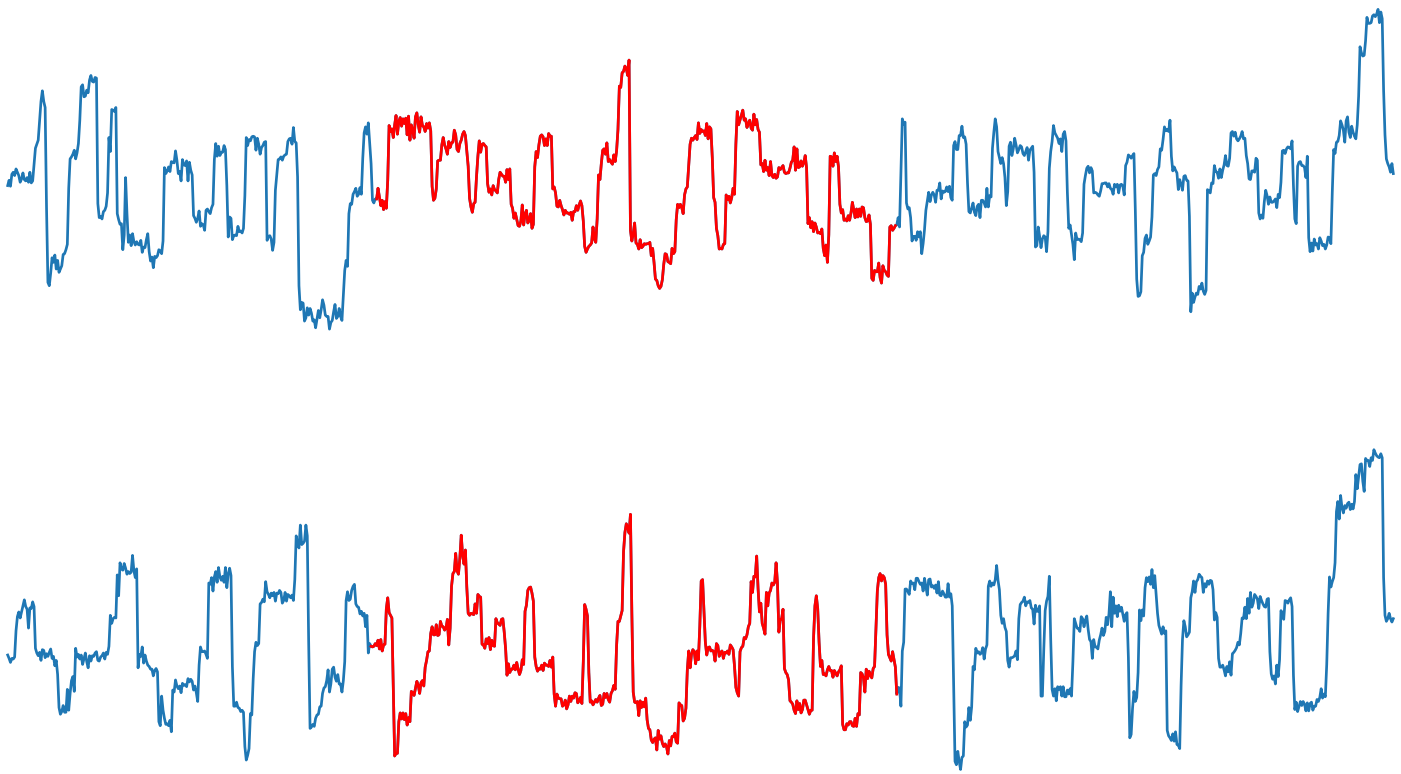


2. Pomocou tejto podobnosti nájsť zhluky čítaní, ktoré zodpovedajú jednotlivým sekvenovaným vzorkám

Meranie podobnosti signálov

Skombinovali sme myšlienky dvoch známych algoritmov:

- Smith-Waterman pre lokálne zarovnávanie reťazcov (Smith a Waterman, 1981)
- Dynamic Time Warping pre globálne zarovnávanie signálov (Sankoff a Kruskal, 1983)



Meranie podobnosti signálov

Nový algoritmus sme nazvali **L**ocal **D**ynamic **T**ime **W**arping.

$$\mathbf{LDTW}[i,j] = \max \begin{cases} \mathbf{LDTW}[i,j-1] \ + \ s(x_i, y_j) \\ \mathbf{LDTW}[i-1,j] \ + \ s(x_i, y_j) \\ 0 \end{cases} \tag{1}$$

- Pre dva signály s dĺžkami n, m vieme $LDTW$ podobnosť spočítať v čase $O(nm)$
- Dôležitým komponentom je skórovacia funkcia $s(\cdot, \cdot)$
- Požadujeme: ak X, Y obsahujú rovnaký barkód, tak $s(X, Y) >> 0$, inak $s(X, Y) << 0$

Skórovanie pomerom vierohodností

Nech p_{xy} je pravdepodobnosť, že hodnoty x a y vidíme zarovnané a nech p_x je pravdepodobnosť výskytu hodnoty x.

$$s(x, y) = \log\left(\frac{p_{xy}}{p_x p_y}\right) \tag{2}$$

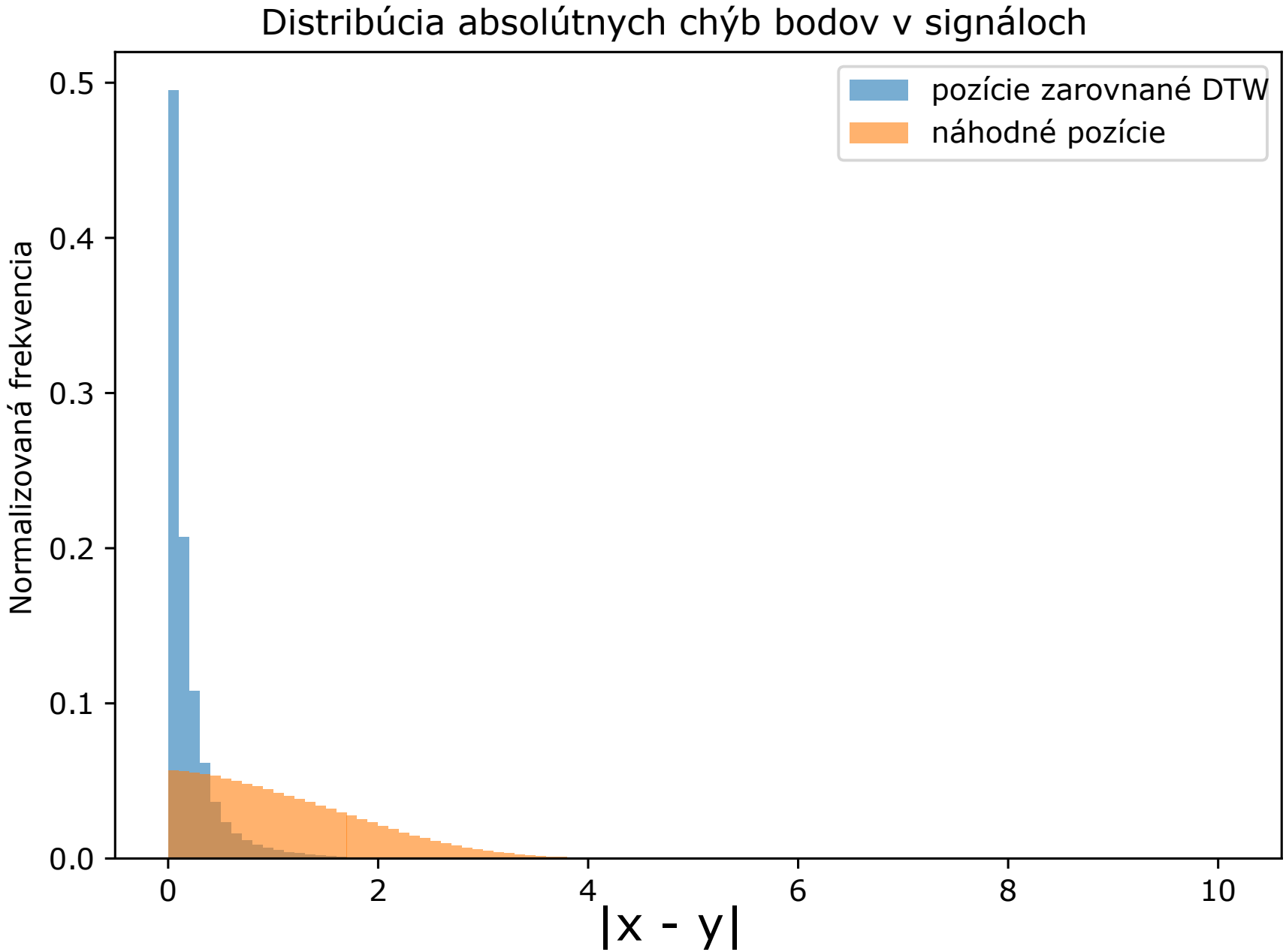
$$s(X, Y) = \sum_{i=1}^{n} \log\left(\frac{p_{x_i,y_i}}{p_{x_i} p_{y_i}}\right) = \log\left(\prod_{i=1}^{n} \frac{p_{x_i,y_i}}{p_{x_i} p_{y_i}}\right) =$$

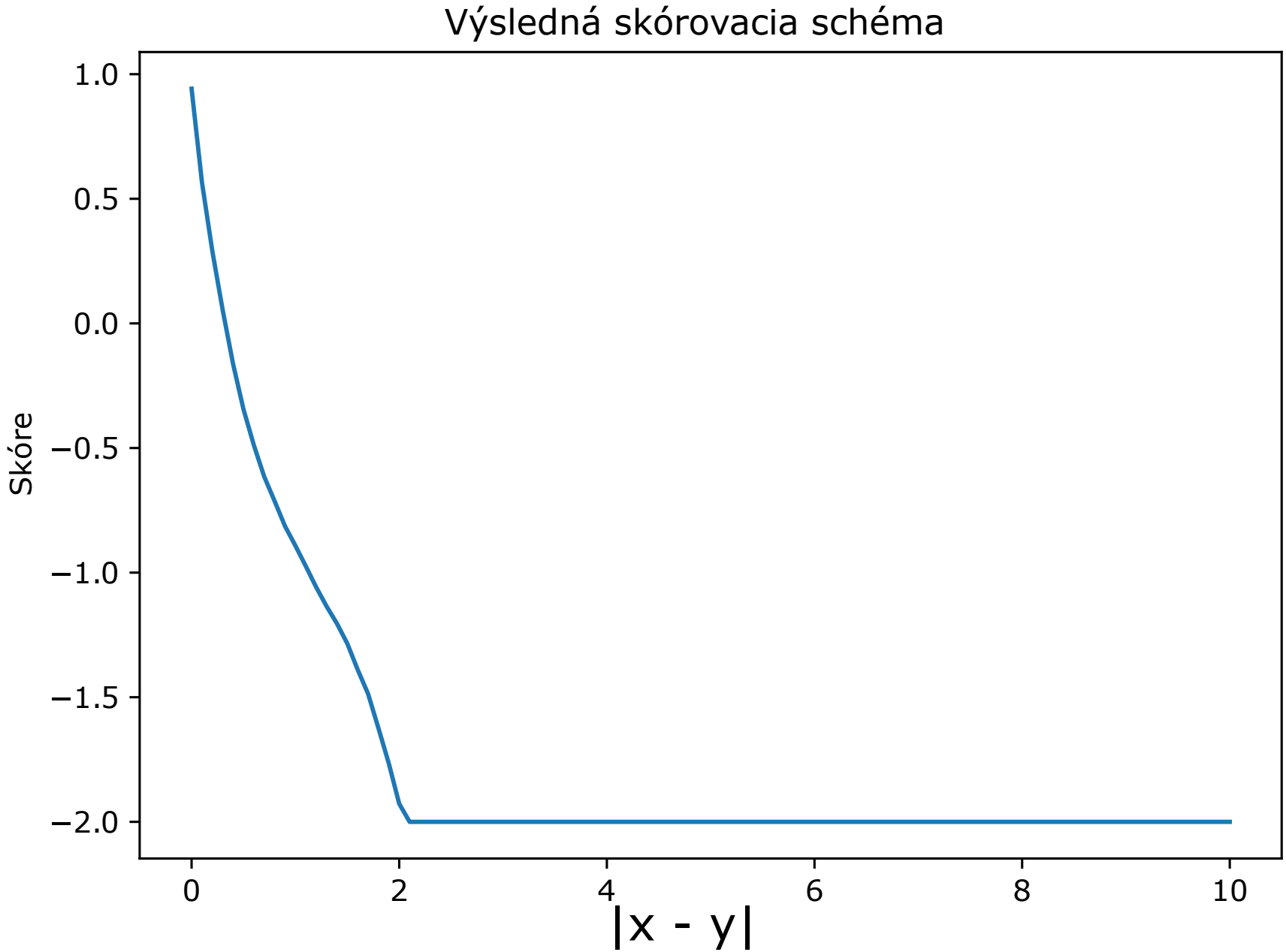
$$= \log\left(\frac{\prod_{i=1}^{n} p_{x_i,y_i}}{\prod_{i=1}^{n} p_{x_i} \prod_{i=1}^{n} p_{y_i}}\right) = \log\left(\frac{P(X, Y \mid \text{X a Y spolu súvisia})}{P(X, Y \mid \text{X a Y spolu } \mathbf{ne}\text{súvisia})}\right) \tag{3}$$

Odhad potrebných pravdepodobností

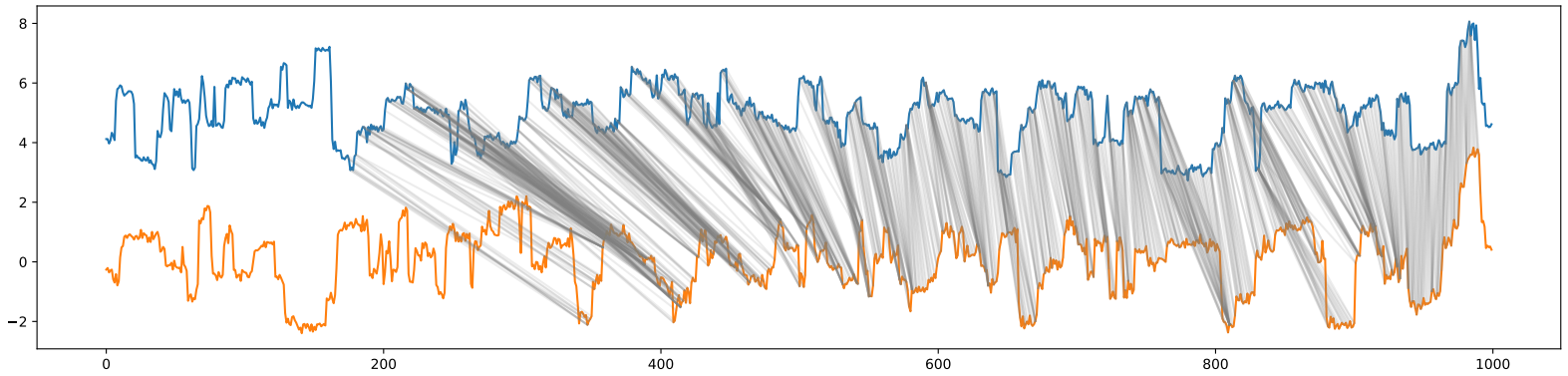
Nech $s(x, y)$ závisí iba od $|x - y|$. Potom $s(x, y)$ odhadneme nasledovne:

$$s(x,y) = \log\left(\frac{P\big(|x-y| \text{ v správne zarovnaných barkódoch}\big)}{P\big(|x-y| \text{ v nezávislých signáloch}\big)}\right) \approx$$
$$\approx \log\left(\frac{P\big(|x-y| \text{ v DTW zarovnaniach barkódov}\big)}{P\big(|x-y| \text{ v nezávislých signáloch}\big)}\right) \quad (4)$$

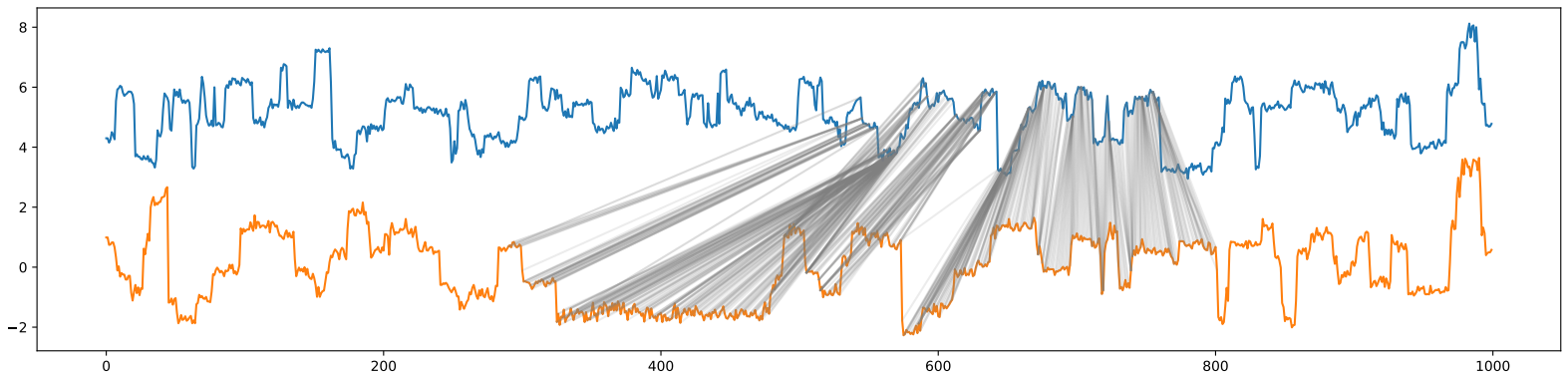

Distribúcia absolútnych chýb bodov v signáloch
pozície zarovnané DTW
náhodné pozície
Normalizovaná frekvencia
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
|x - y|


Výsledná skórovacia schéma
Skóre
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0
2
4
6
8
10
|x - y|

LDTW ako miera podobnosti



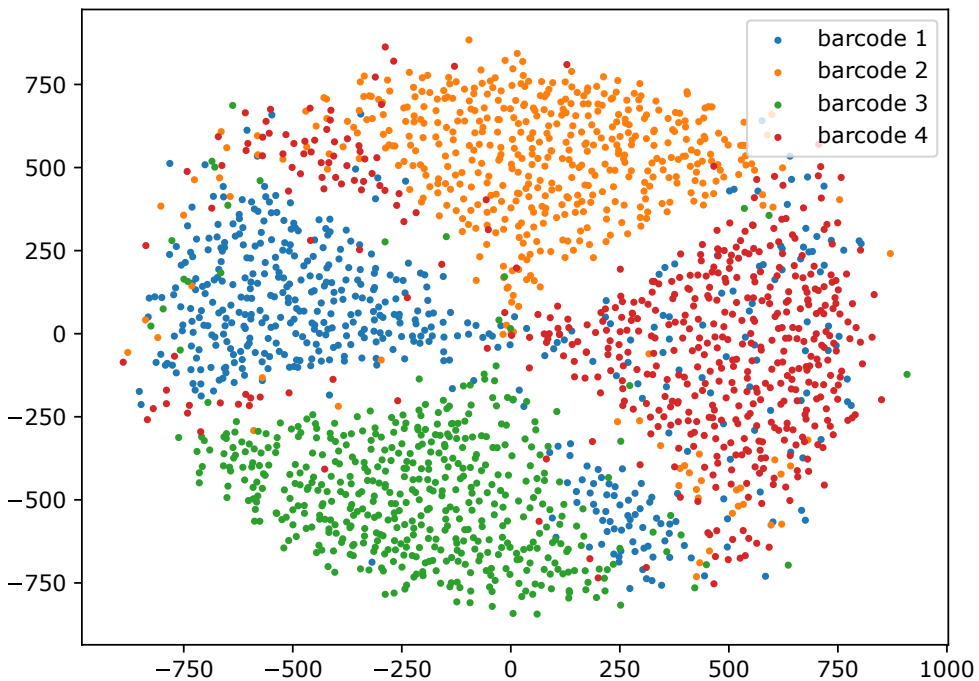
Obr.: Rovnaké barkódy, vysoké skóre zarovnania.



Obr.: Rôzne barkódy, nízke skóre zarovnania.

Nasledujúca fáza

- Vieme merať podobnosť čítaní
- Nasledujúca fáza: zhlukovanie čítaní na základe podobnosti signálov



Obr.: Vizualizácia LDTW podobnosti pomocou viacrozmerného škálovania (Kruskal, 1964).

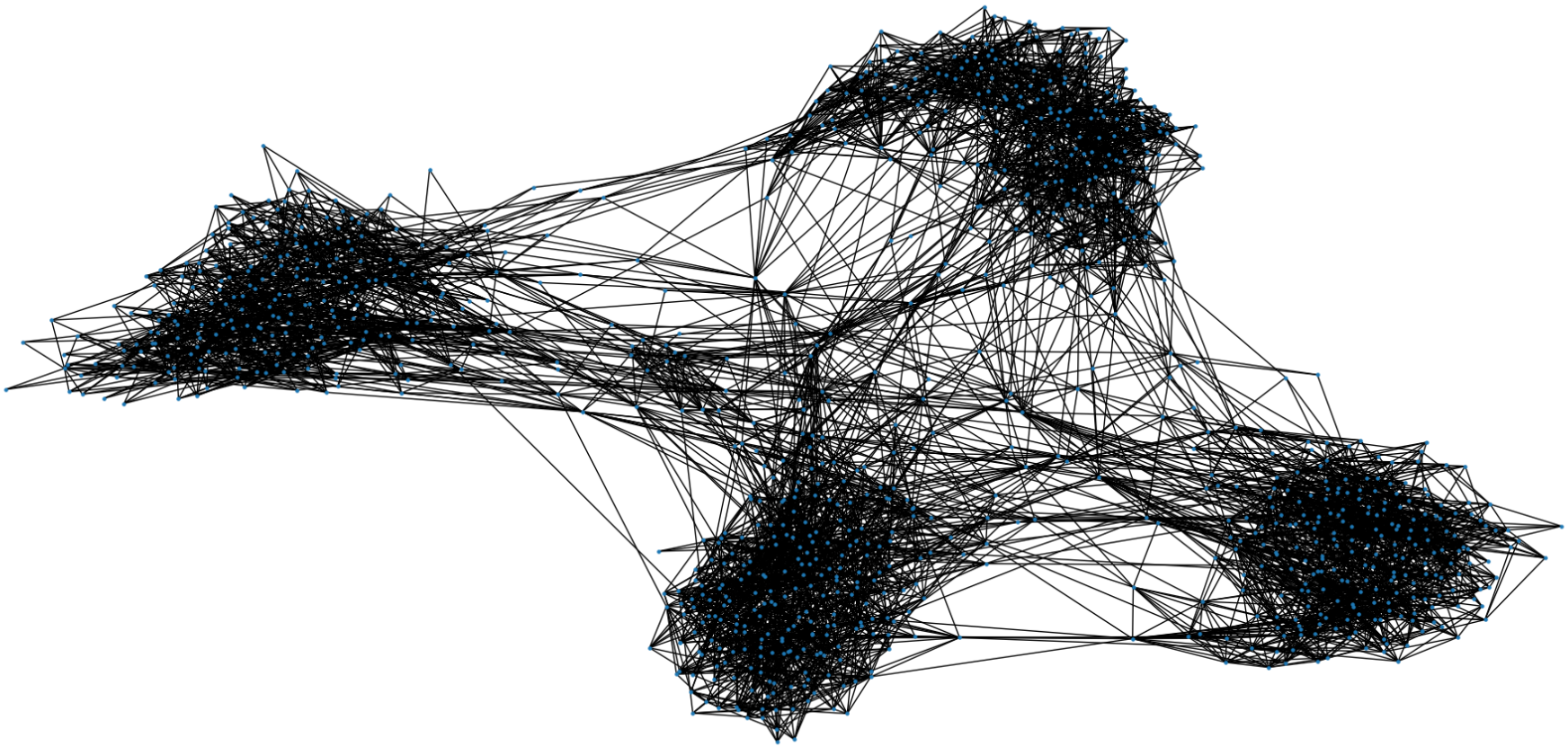
Spektrálne zhlukovanie

Základná štruktúra algoritmov spektrálneho zhlukovania (von Luxburg, 2007):

1. Z matice podobností S si skonštruujeme váhovaný graf G
2. Laplacian grafu G: $L = D - S$ ($d_{ii} = \sum_j s_{ij}$ a 0 mimo diag.)
3. Vypočítame prvých k vlastných vektorov L a uložíme ich do riadkov matice $U_{k \times n}$
4. Použijeme štandardný zhlukovací algoritmus (napr. $k-$means) na stĺpce matice U

Viacero formulácií, my sme použili variant tzv. *normalizovaného* spektrálne zuhlukovanie (Shi a Malik, 2000).

Spektrálne zhlukovanie - predspracovanie vstupu



Obr.: Príklad grafu k najbližších susedov ($k = 20$) pre 4 barkódy.

Zhlukovanie veľkého množstva čítaní

Aplikovať spektrálne zhlukovanie na celú množinu (niekoľkých miliónov) čítaní by bolo pomalé, preto náš algoritmus postupuje nasledovne:

1. Výber malej náhodnej vzorky čítaní $\mathcal{A}$
2. Výpočet **LDTW** podobností pre všetky dvojice z $\mathcal{A}$
3. Spektrálne zhlukovanie na $\mathcal{A}$
4. Výber malého počtu reprezentantov z každého zhluku
5. Každému ďalšiemu čítaniu je priradená značka na základe podobností k reprezentantom

Dosiahnuté výsledky

Testovanie na čítaniach s barkódmi, na ktorých náš model nebol trénovaný.

Experiment	Barkódy	Presnosť(%)	Označkovaných čítaní(%)
1	5, 6, 7	98.77	94.59
2	5, 6, 7, 8	88.69	66.82
3	5, 7, 9, 12	97.41	91.91
4	8, 9	98.95	93.26
5	5, 6, 11, 12	98.60	95.18
6	11, 12	98.50	92.81

Experiment: učenie s učiteľom

- Dovolili sme nášmu modelu pozrieť sa na niekoľko správne označkovaných čítaní - učenie s učiteľom
- Porovnateľné výsledky s Deepbinnerom
- Rýchlejší a **interpretovateľný** model

	Presnosť (%)	Označkovaných (%)
naša metóda	98.39	92.84
Deepbinner (neurónová sieť)	98.41	94.84
Albacore (preložené sekvencie)	97.27	81.10

Ďalšia práca

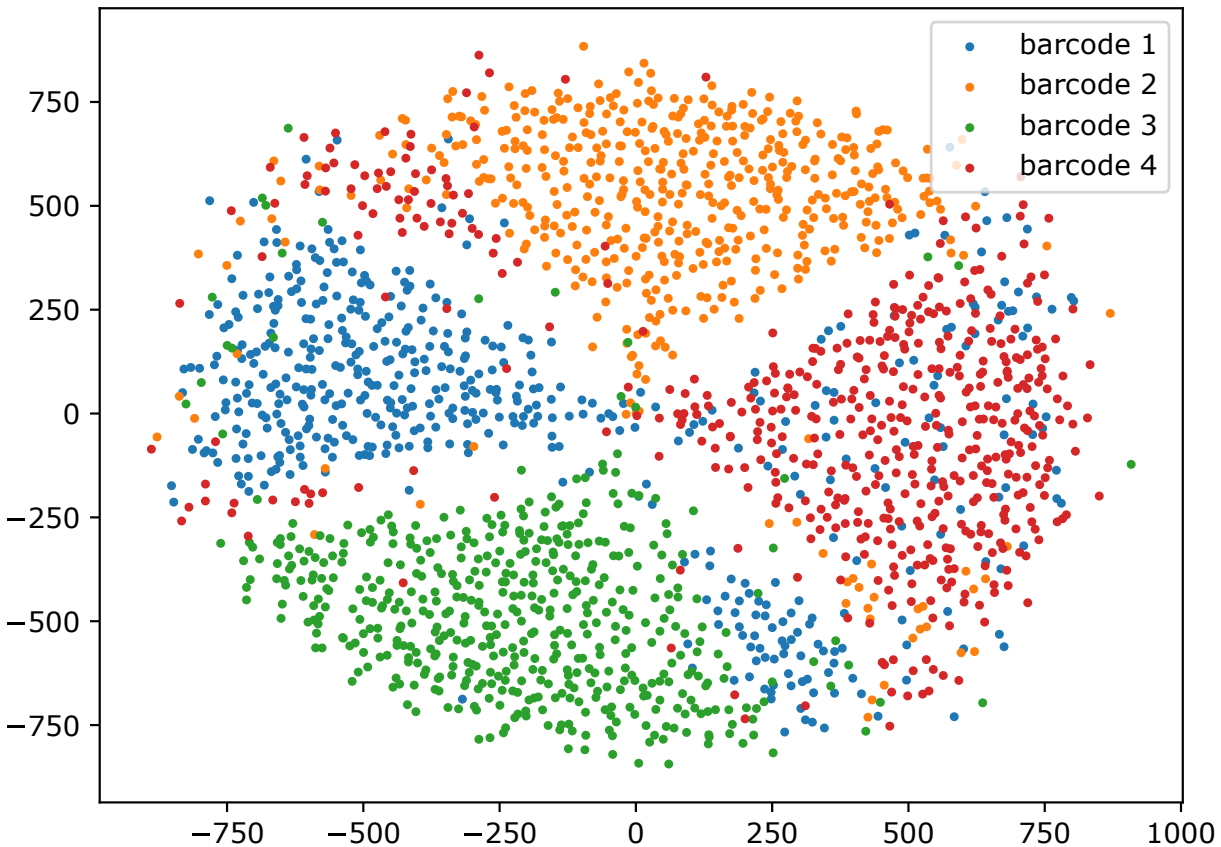
- Paralelné výpočty na grafickej karte
- Zložitejšia skórovacia funkcia
- Uvažovať históriu signálu
- Zhlukovanie nevyvážených dát (napr. pre spektrálne zhlukovanie Qian a spol., 2013)

Koniec

Ďakujem všetkým za pozornosť.

Otázka č.1

"Chýba mi uvedenie použitia tejto sady [Native Barcoding Kit] a jej výhod a nevýhod hneď v úvode." - Ing. Matej Lexa, PhD. (oponent)



Obr.: Vizualizácia LDTW podobnosti pomocou viacrozmerného škálovania.

Otázka č.2

"Kód v Pythone používa niektoré externé rozširujúce balíčky, ktoré nie sú nikde v texte spomenuté alebo citované. Príkladom môže byť balík HDF5PY..." - Ing. Matej Lexa, PhD. (oponent)

Použité balíčky/knižnice:

- H5PY (`https://github.com/h5py/h5py`)
- scikit-learn (`https://github.com/scikit-learn/scikit-learn`)
- matplotlib (`https://github.com/matplotlib/matplotlib`)
- seaborn (`https://github.com/mwaskom/seaborn`)
- numpy (`https://github.com/numpy/numpy`)