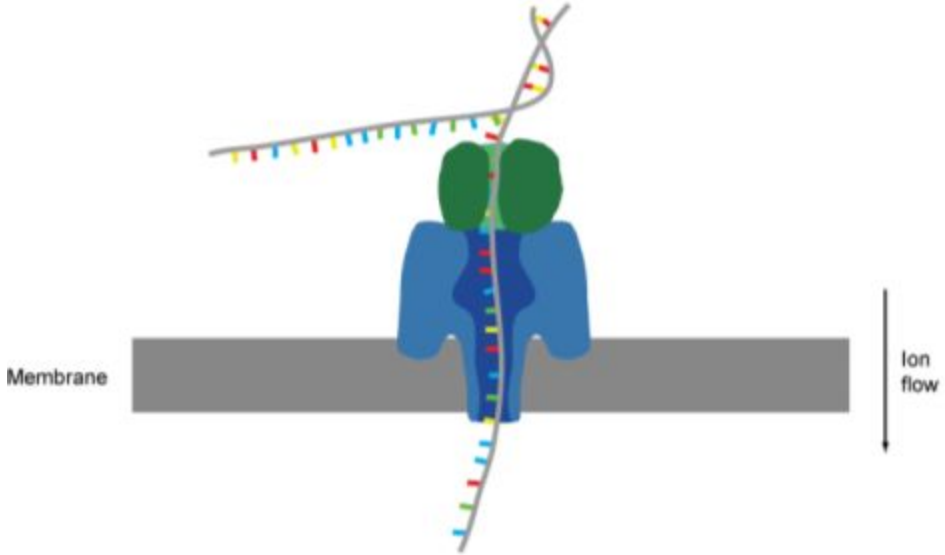
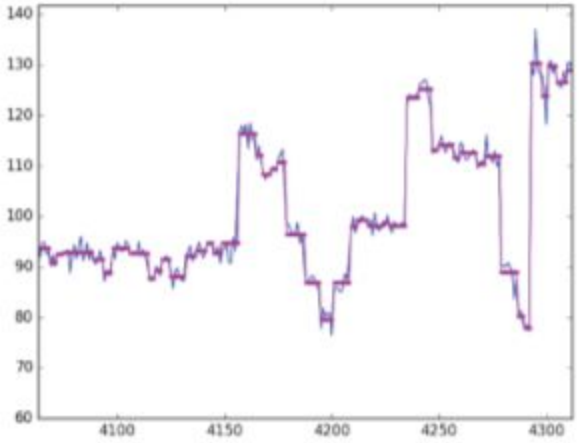
Zarovnávanie nanopórových čítaní ku grafom

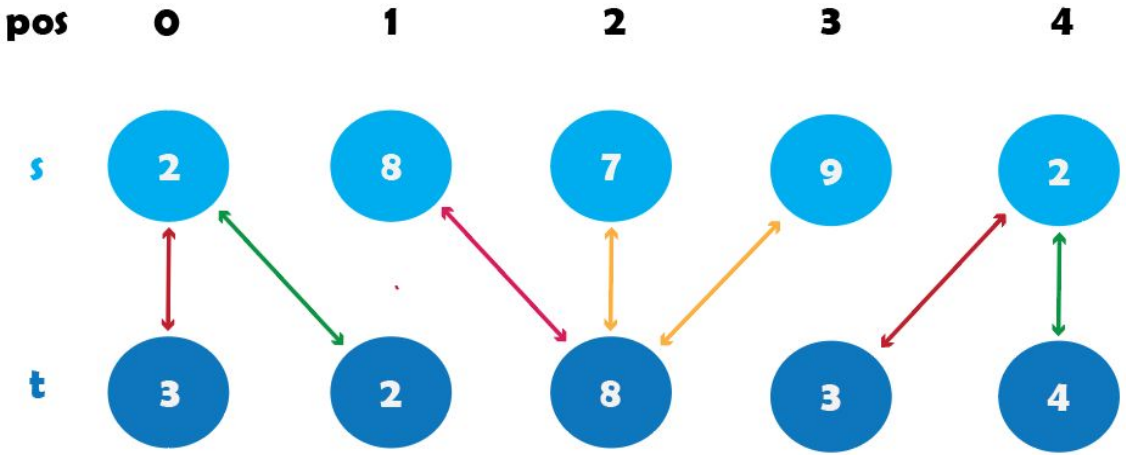
Názov (EN): Alignment of Nanopore Reads to Graphs
Názov (SK): Zarovnávanie nanopórových čítaní ku grafom
Študent: Roman Beňo
Školiteľ: doc. Mgr. Tomáš Vinař, PhD.

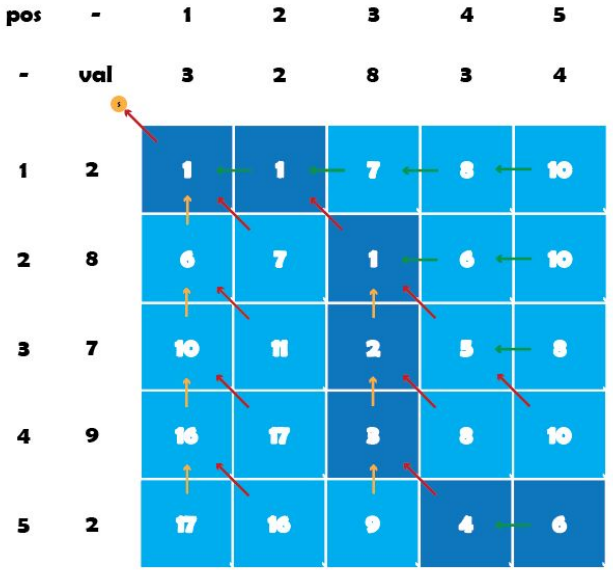
I. Stručný popis problematiky



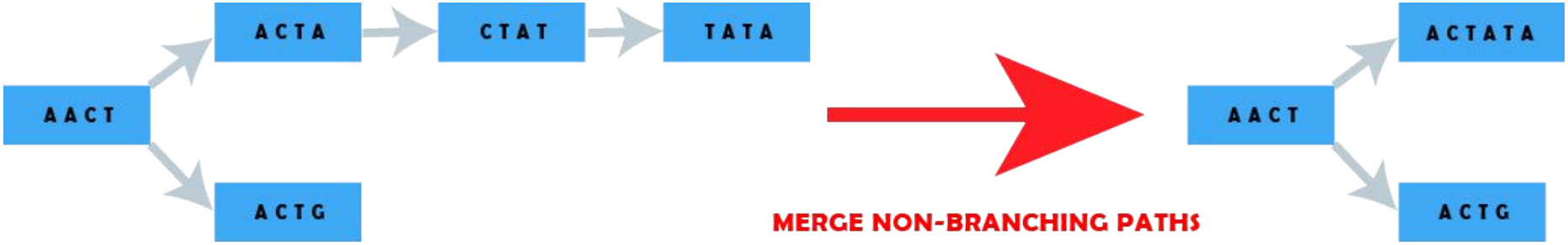


Nanopórové sekvenovanie, MinION, squiggle (“čmáranec”), špecifiká výstupných dát, base-calling



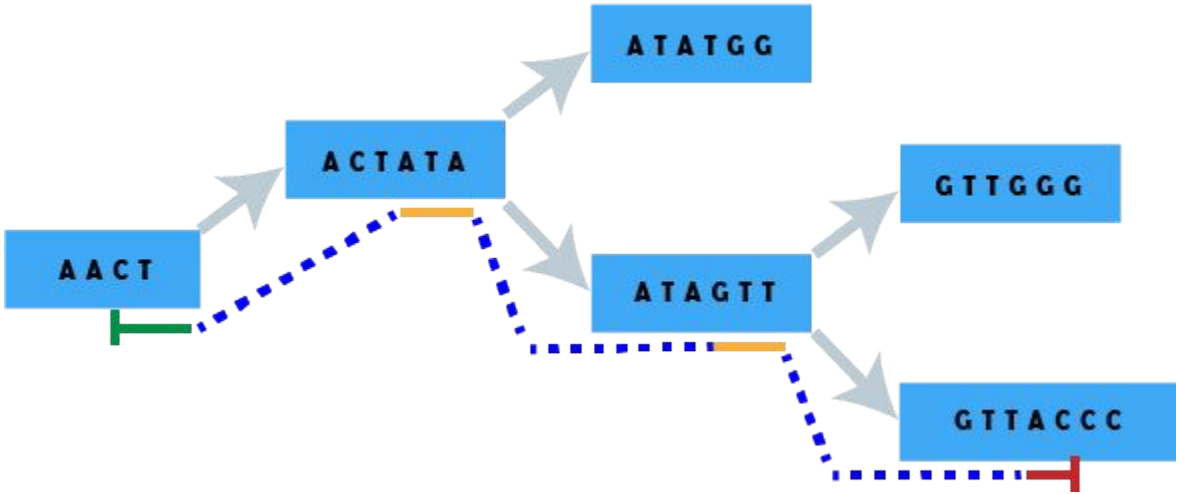


Dynamic time warping - penalizácia/cena (penalty), funkcia na výpočet ceny (cost function), minimum warping path

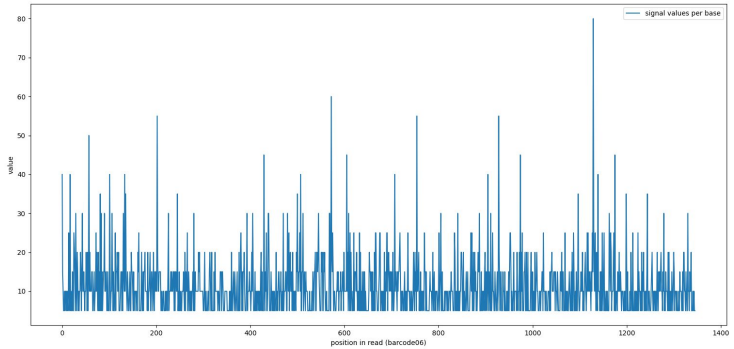


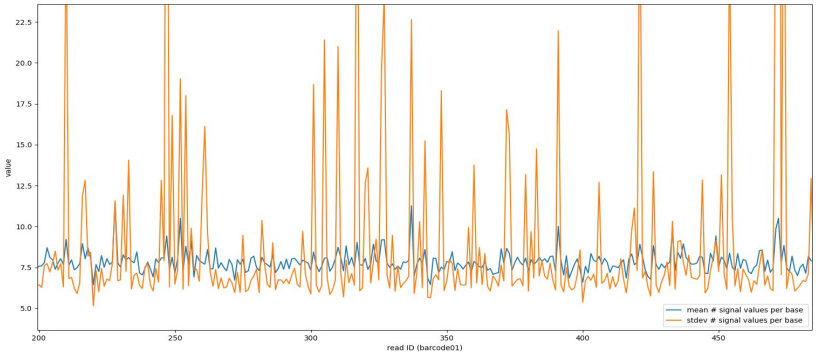
reprezentácia sekvencii v pamäti,
De Bruijnov graf,
hľadanie sledu,
aplikácia

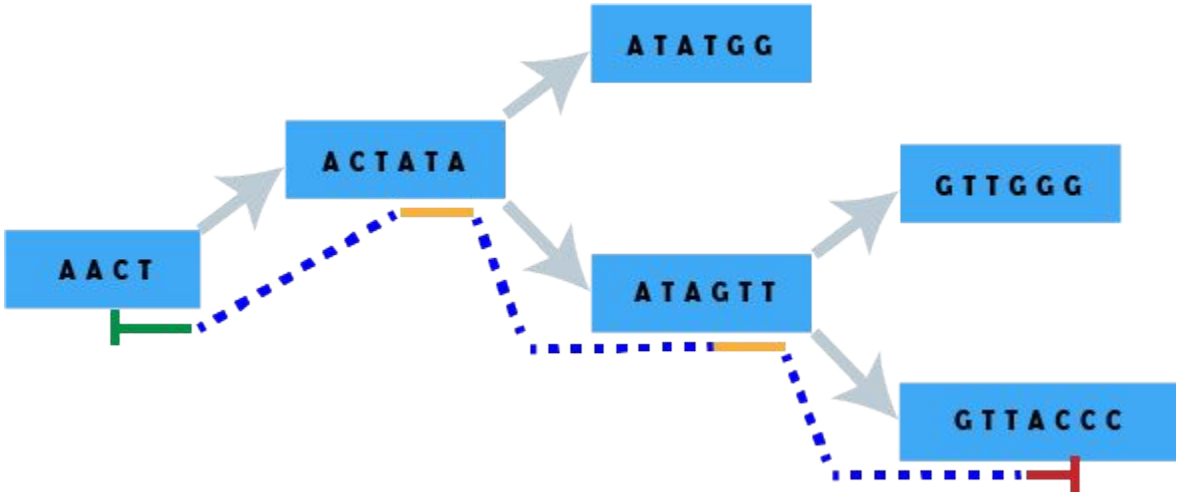
II. Postup



Každý vrchol obsahuje “falošný” umelo vygenerovaný squiggle
Pórový model (mean)
Počet hodnôt per k-mer

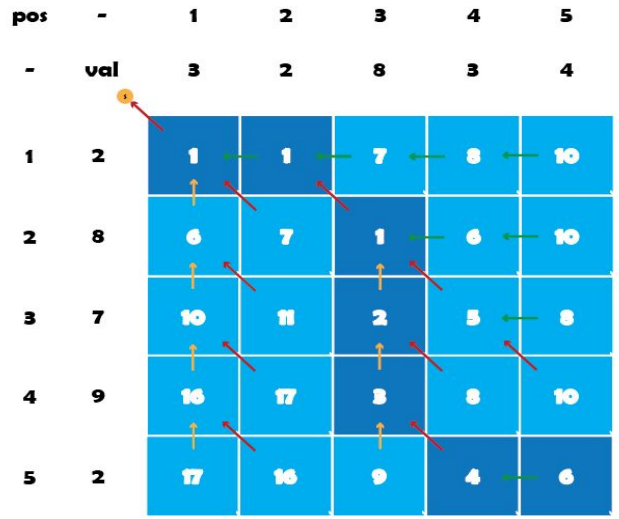






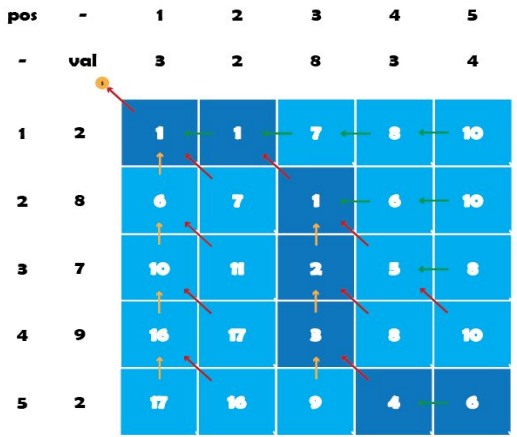
“Hlava”, “telo”, “chvost” hľadanej cesty
Prechody grafom
Práca s maticou:

1. Pokrok v rámci vstupného readu
2. Lacnejšie hodnoty pre “hlavu”
3. Posledný riadok a stĺpec sú významné



Hľadanie počiatočných vrcholov cesty
Skórovacia funkcia pre posúdenie, či by vrchol bol dobrou “hlavou”
Najlepšie 1% vrcholov (zaokrúhlené nahor)

```
...
double read_aligned_multip = N * 3.0 / (read_position * 1.0);
double read_pos_fraction = read_position * 1.0 / N;
double score = (penalty+read_pos_fraction) * multip * read_aligned_multip *
read_aligned_multip;
return score;
```



Rozširovanie cesty

```
struct AlignResult {
    double penalty = 0;
    int read_pos = 0;
    long long path_tag = -1;
    int len = 0;
};
```

DTW vo vnútri vrcholu - pole vstupných a výstupných “situácii”

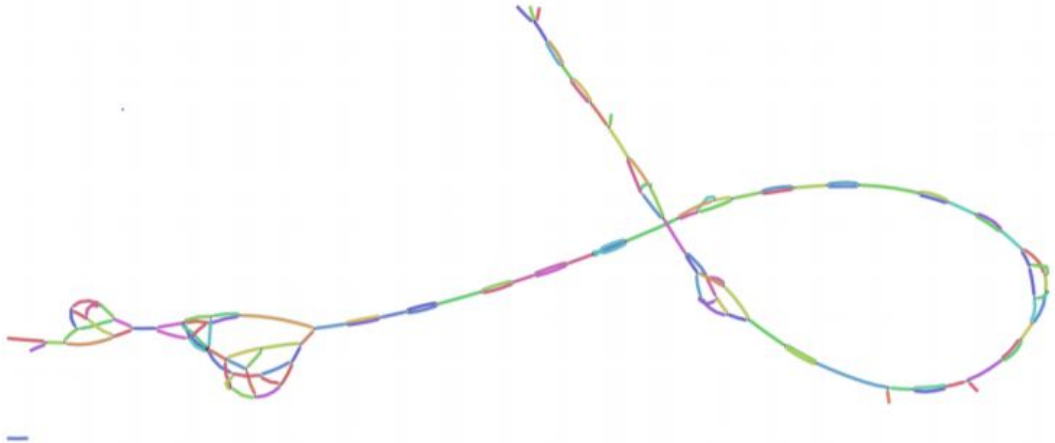
Kontrola celkovej “situácie”

```
(penalty*multip) / (read_position*1.0);
```

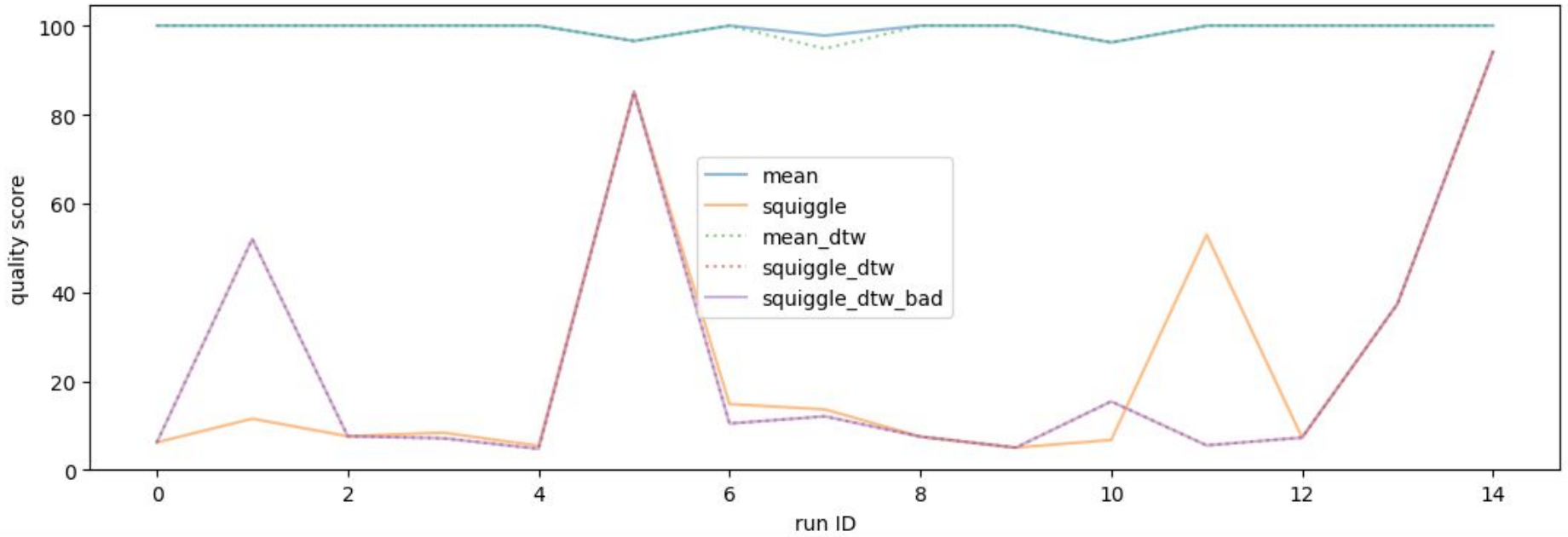
III. Experiment

	Dataset 1	Dataset 2
Vertices	148	159
Edges	200	255
Total sequence length	4209	2984

Fusobacterium nucleatum

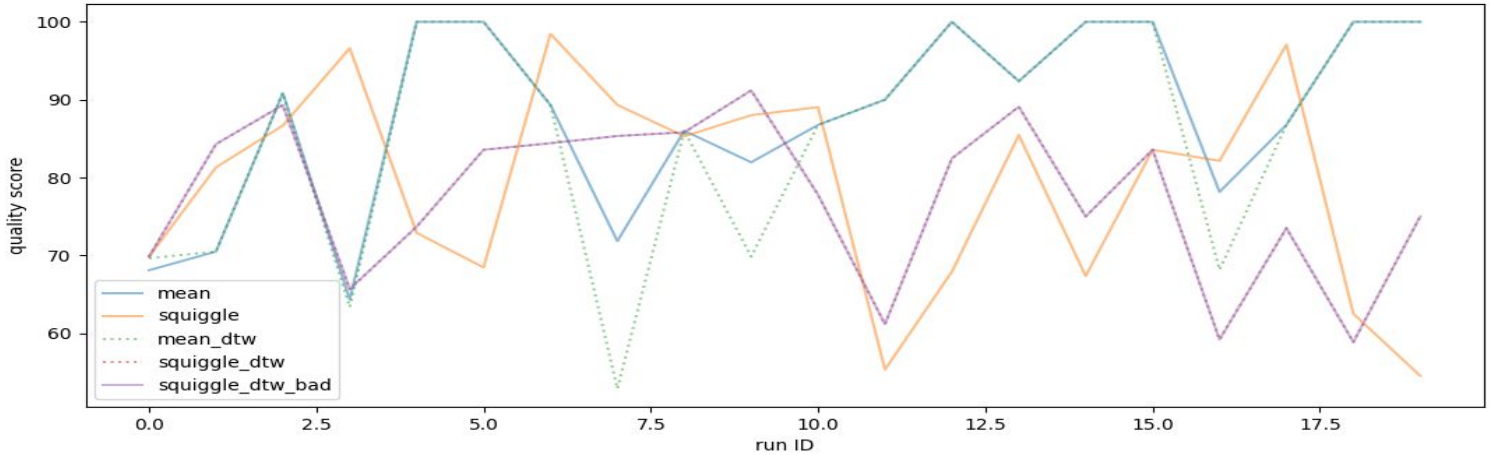


Generovanie input squiggle + referenčný squiggle
Cost function v DTW (eu, p2, pc)
Úprava penalty podľa dĺžky warping path (yes, no)

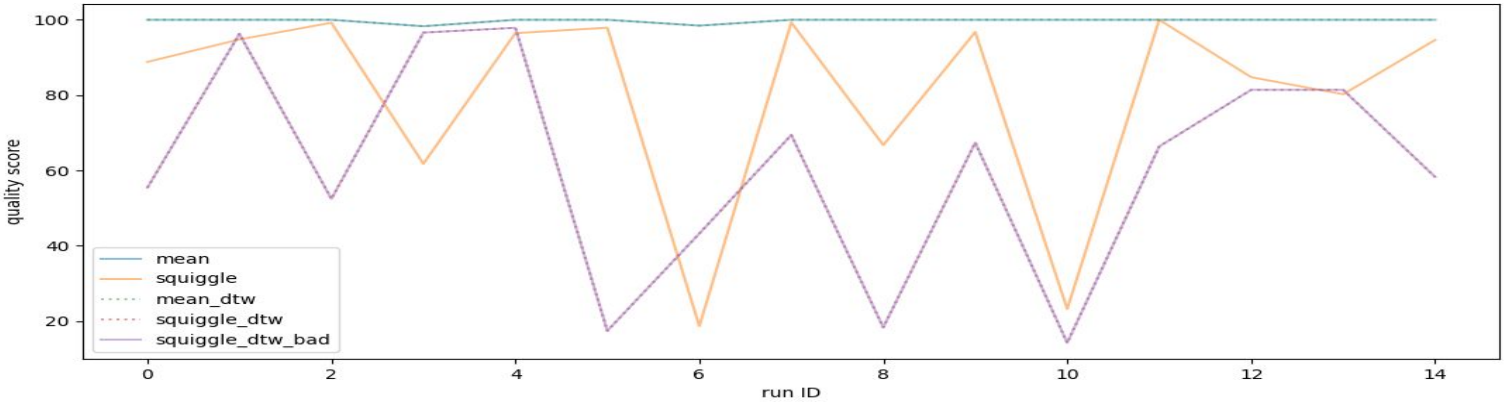


eu-l-1

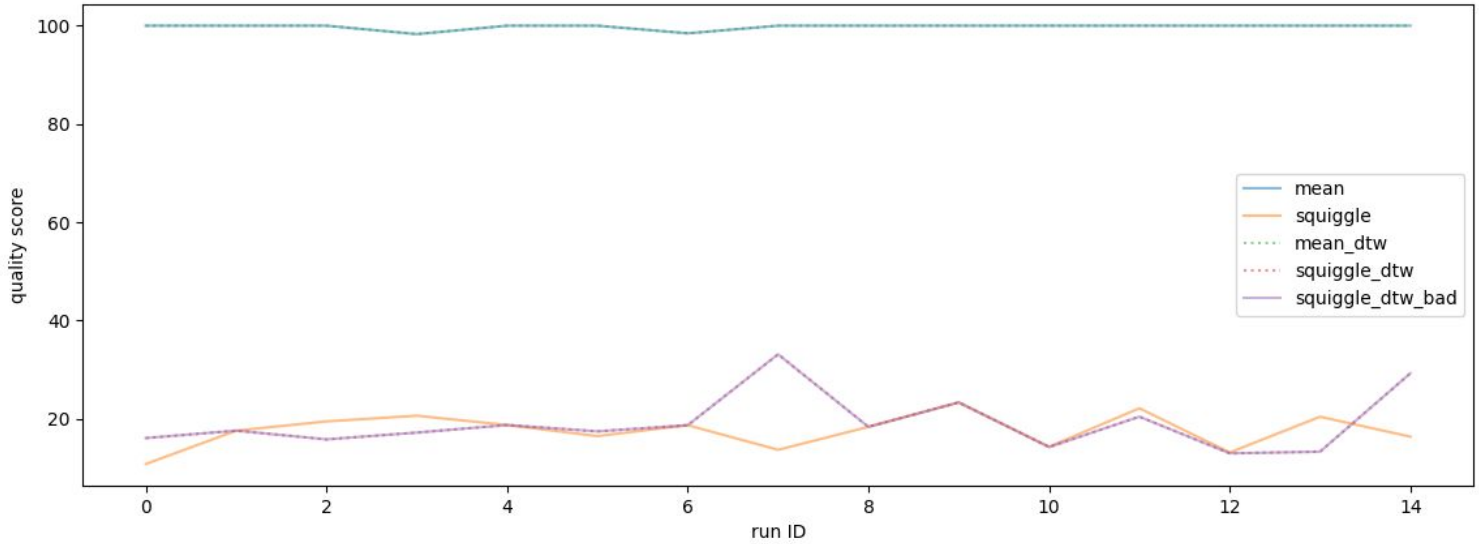
eu-l-r


quality score
run ID
mean
squiggle
mean_dtw
squiggle_dtw
squiggle_dtw_bad

p2-n-r


quality score
run ID
mean
squiggle
mean_dtw
squiggle_dtw
squiggle_dtw_bad

pc-n-r



Výkon aj kvalita sa líšia v závislosti od topológie grafu
pc a eu “zlé funkcie” - p2 trochu náprava - l viac náprava
Všetky squiggle rany majú podobné výsledky - počty generovaných hodnôt v referencii nie sú podstatné (diskriminácia všetkých vs. len niekoho)

IV. Súhrn a námety na ďalšie štúdium

Experimentálny postup, začať s vymysleným pekným readom, zhoršovať ho
Výsledky sú zlé

Prečo je prvý dataset taký neúspešný?

Topológia skutočného grafu ??
Úprava vyhľadávania, zložitosť, veľké datasety

Veľa šumu

Ďakujem za pozornosť