Detekcia tandemových opakovaní v nanopórových dátach

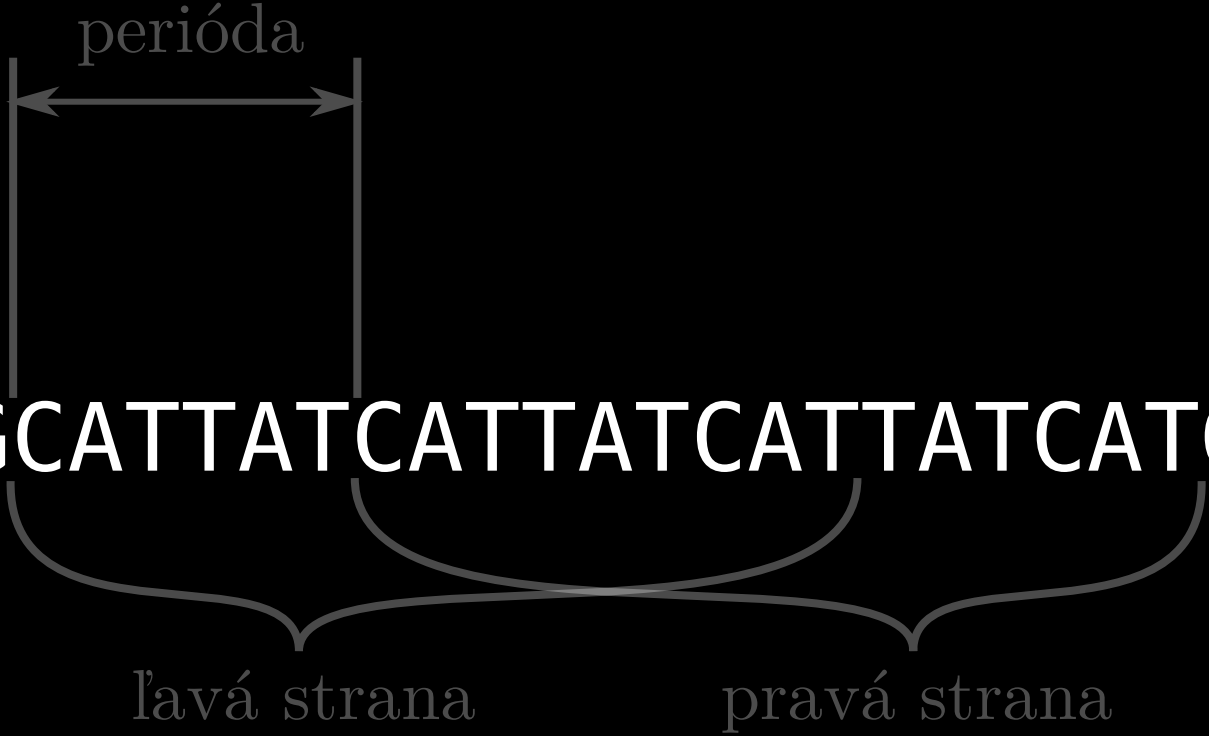
Eduard Batmendijn
Školiteľ: doc. Mgr. Tomáš Vinař, PhD.

18. júna 2020

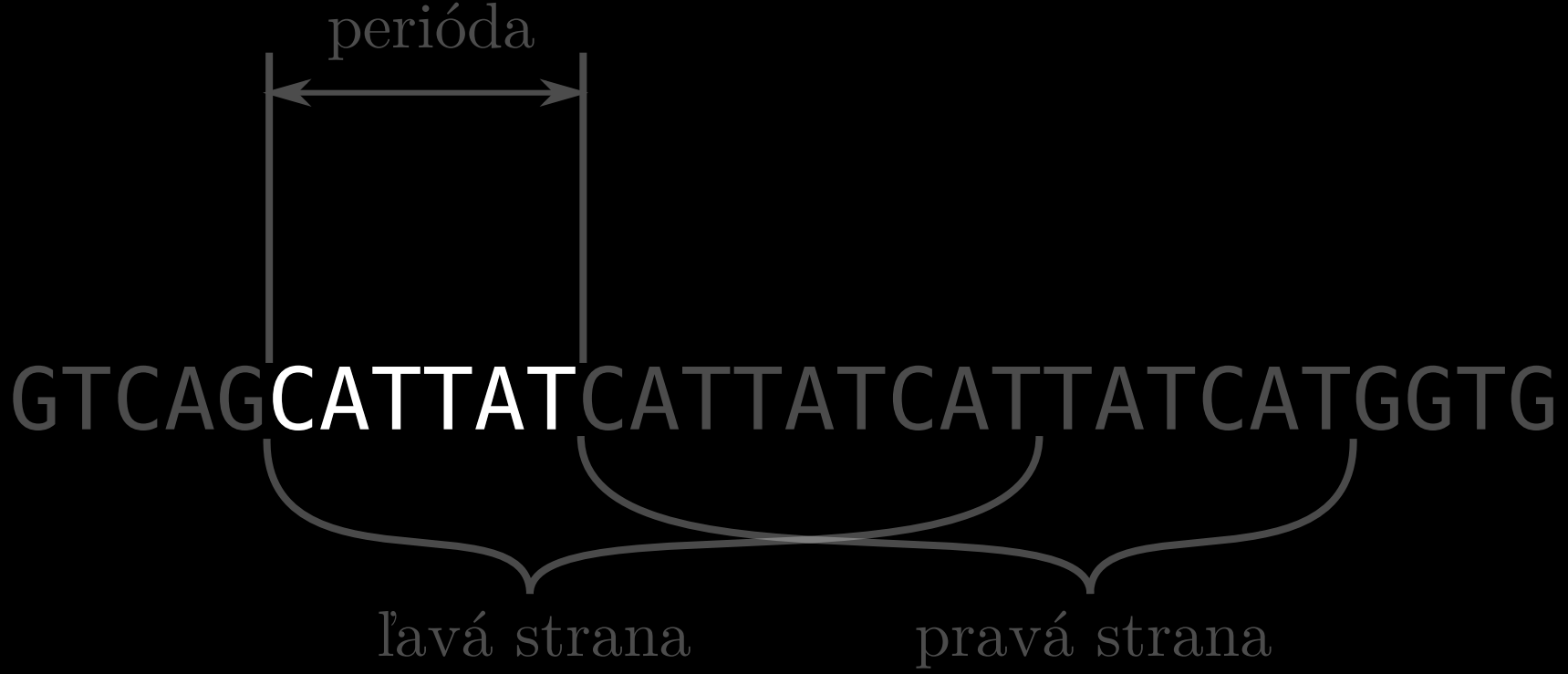
Detekcia tandemových opakovaní v nanopórových dátach

Detekcia tandemových opakovaní v nanopórových dátach

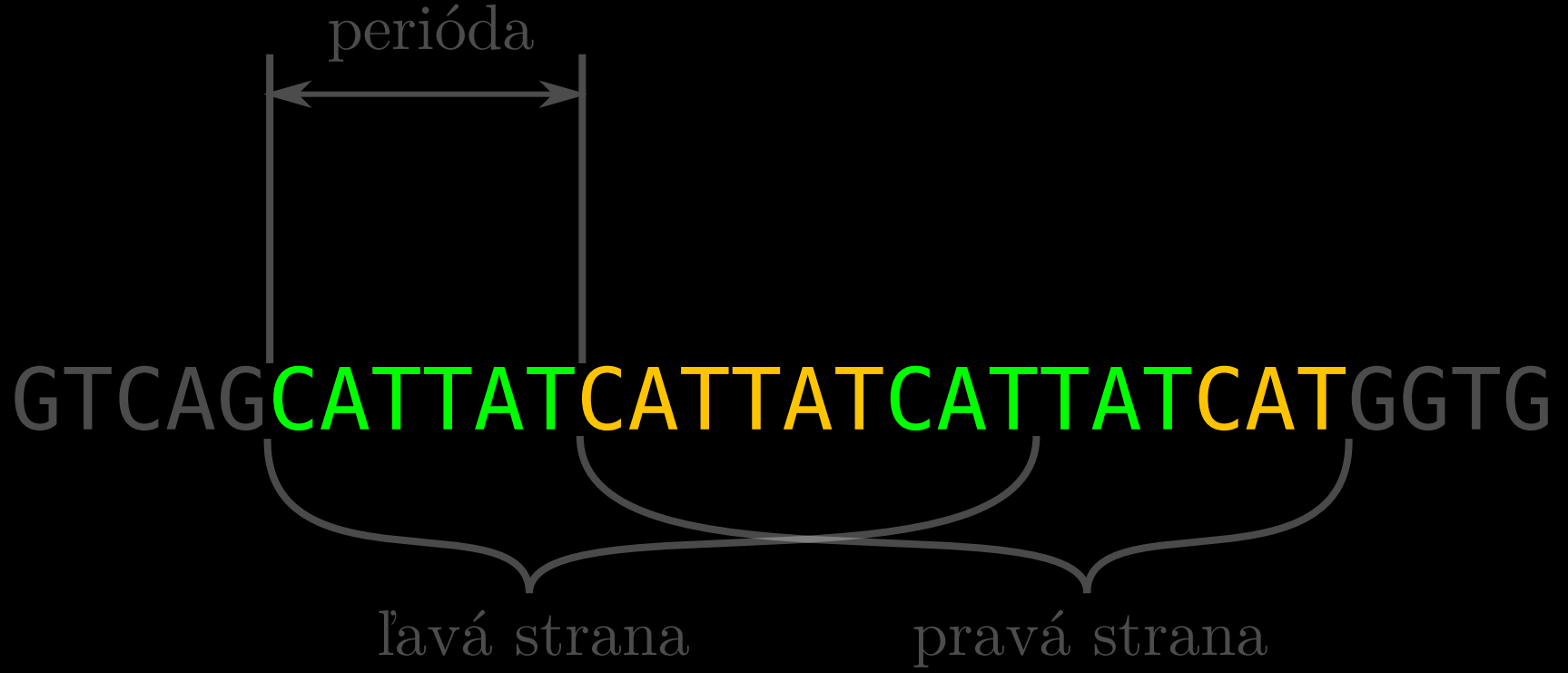
Tandemové opakovanie



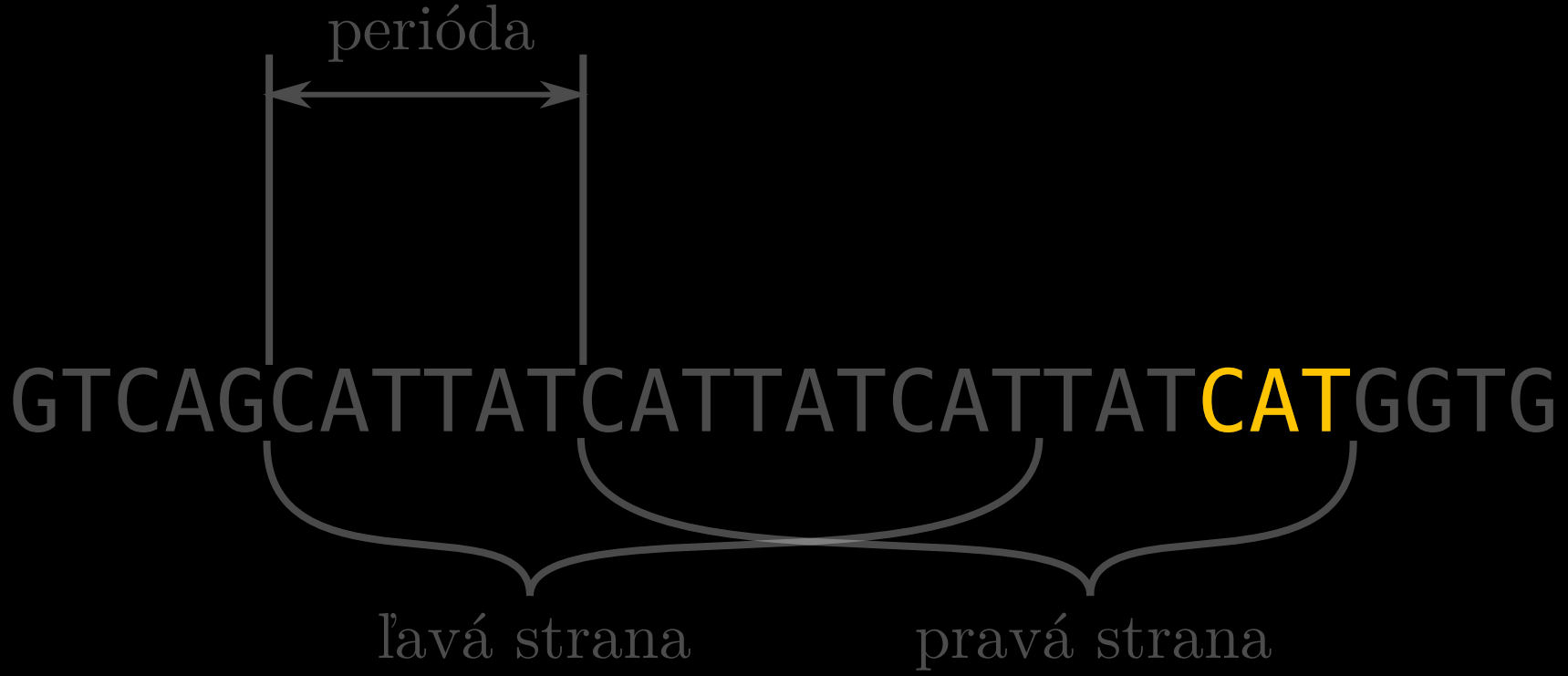
Tandemové opakovanie



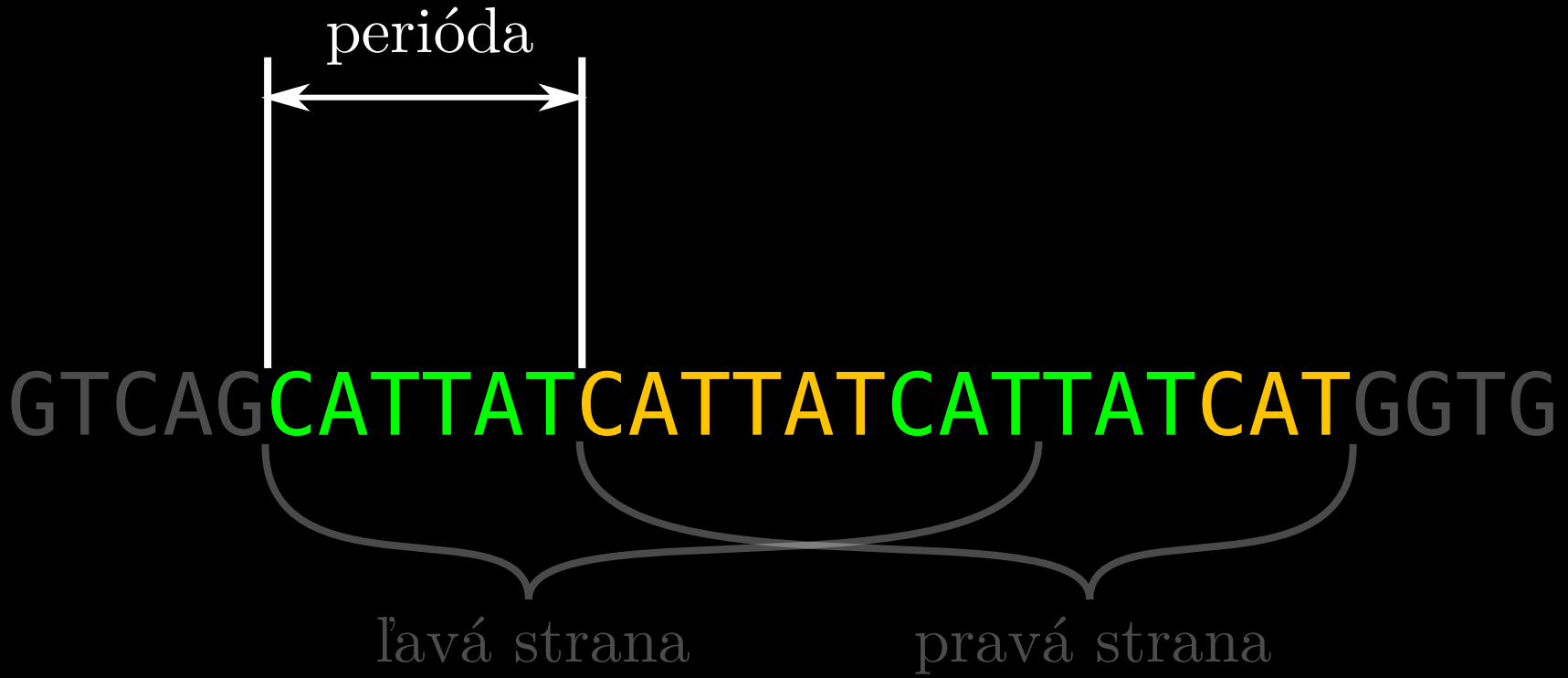
Tandemové opakovanie



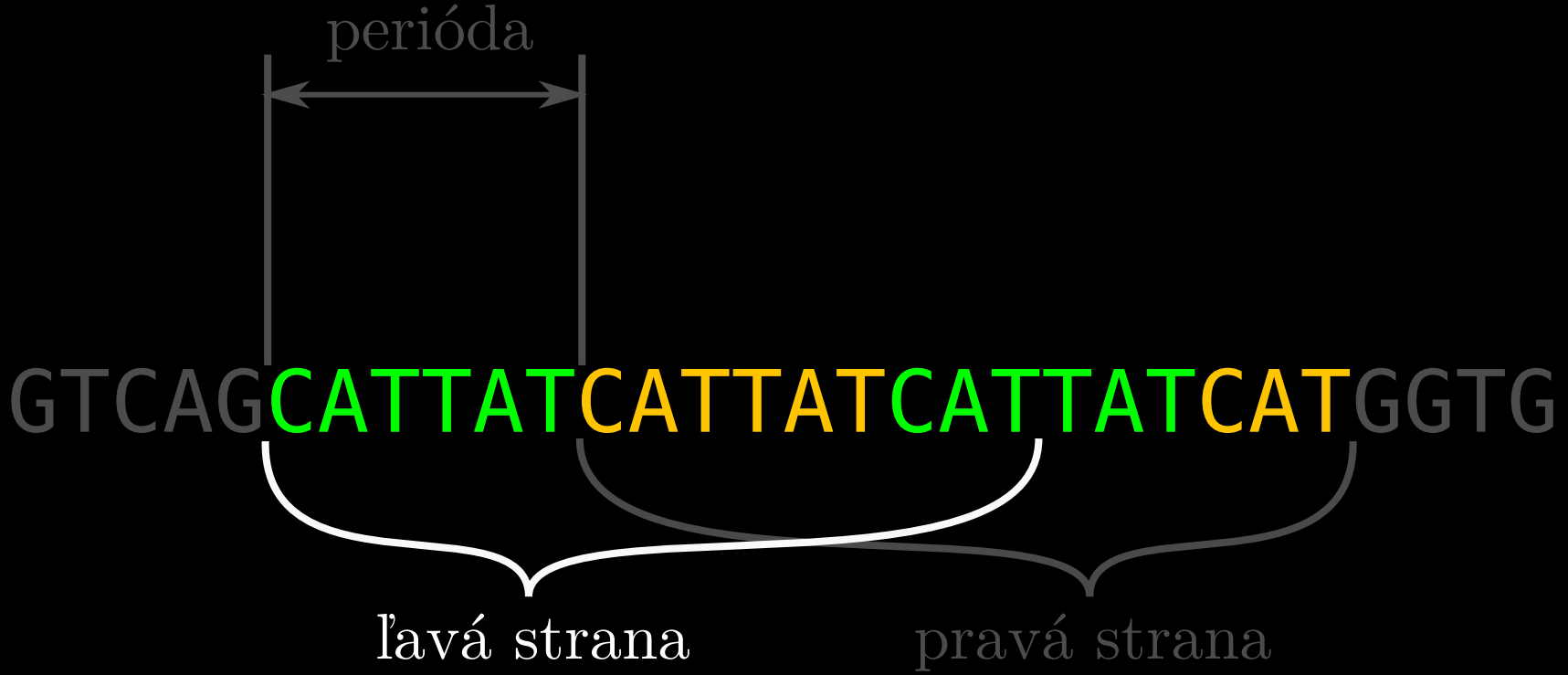
Tandemové opakovanie



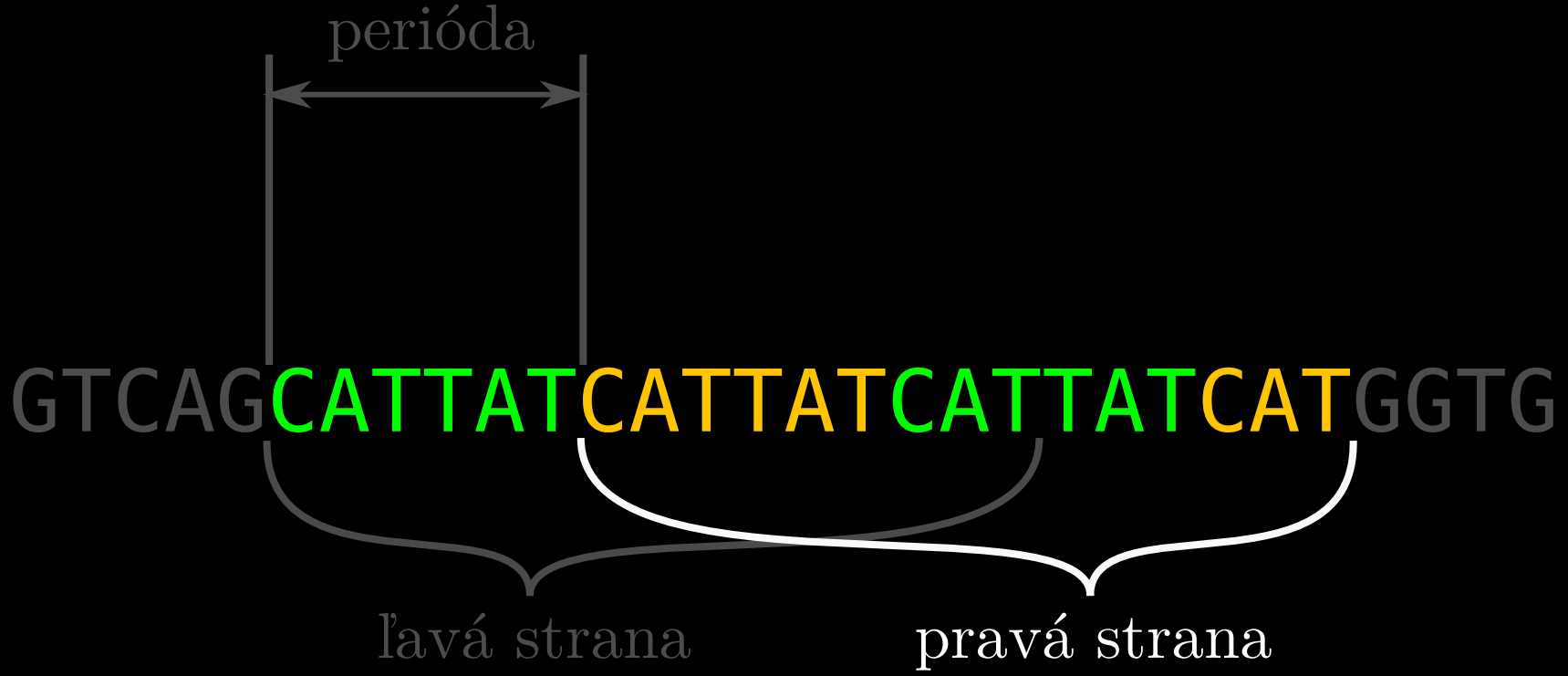
Tandemové opakovanie



Tandemové opakovanie



Tandemové opakovanie



Nedokonalé tandemové opakovanie

GTCAGCATTATCATGATCATTATCATGGTG

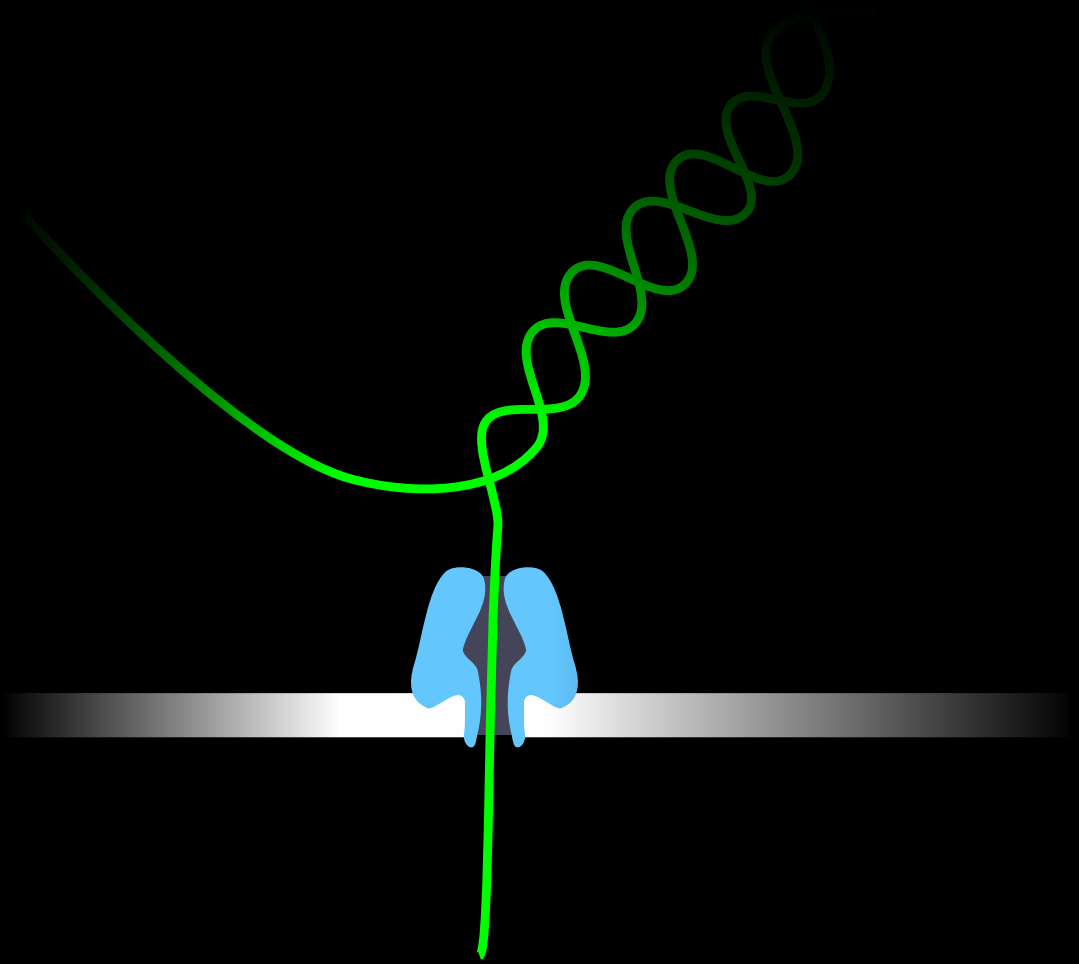
Nedokonalé tandemové opakovanie

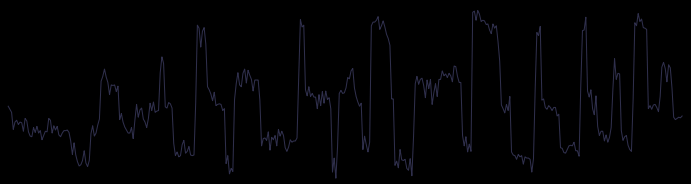
GTCAGCATTATCATGATCATTATCATGGTG

Detekcia tandemových opakovaní v nanopórových dátach

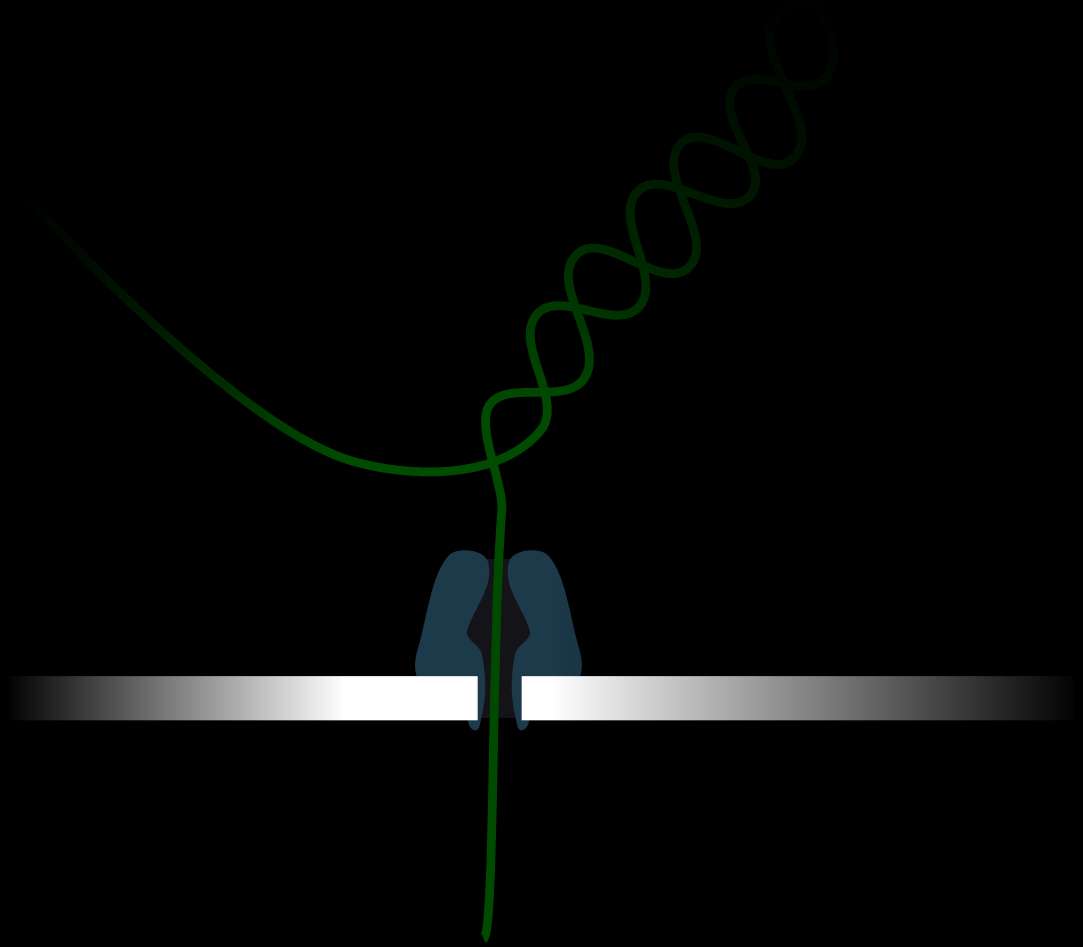
Detekcia tandemových opakovaní v nanopórových dátach

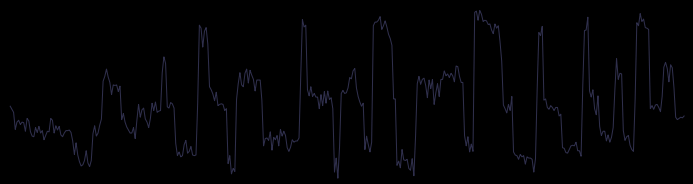
Nanopórové sekvenovanie



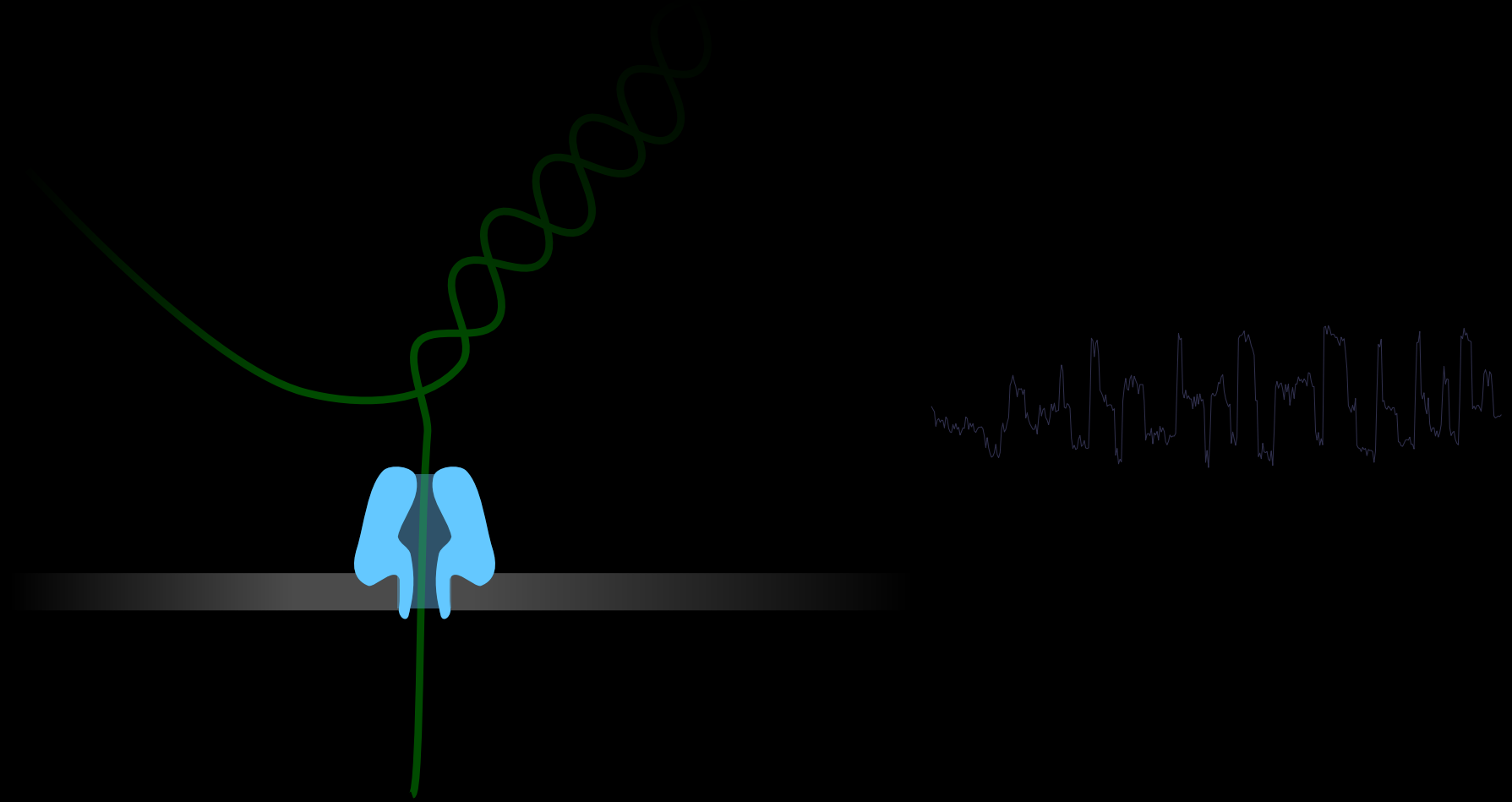


Nanopórové sekvenovanie

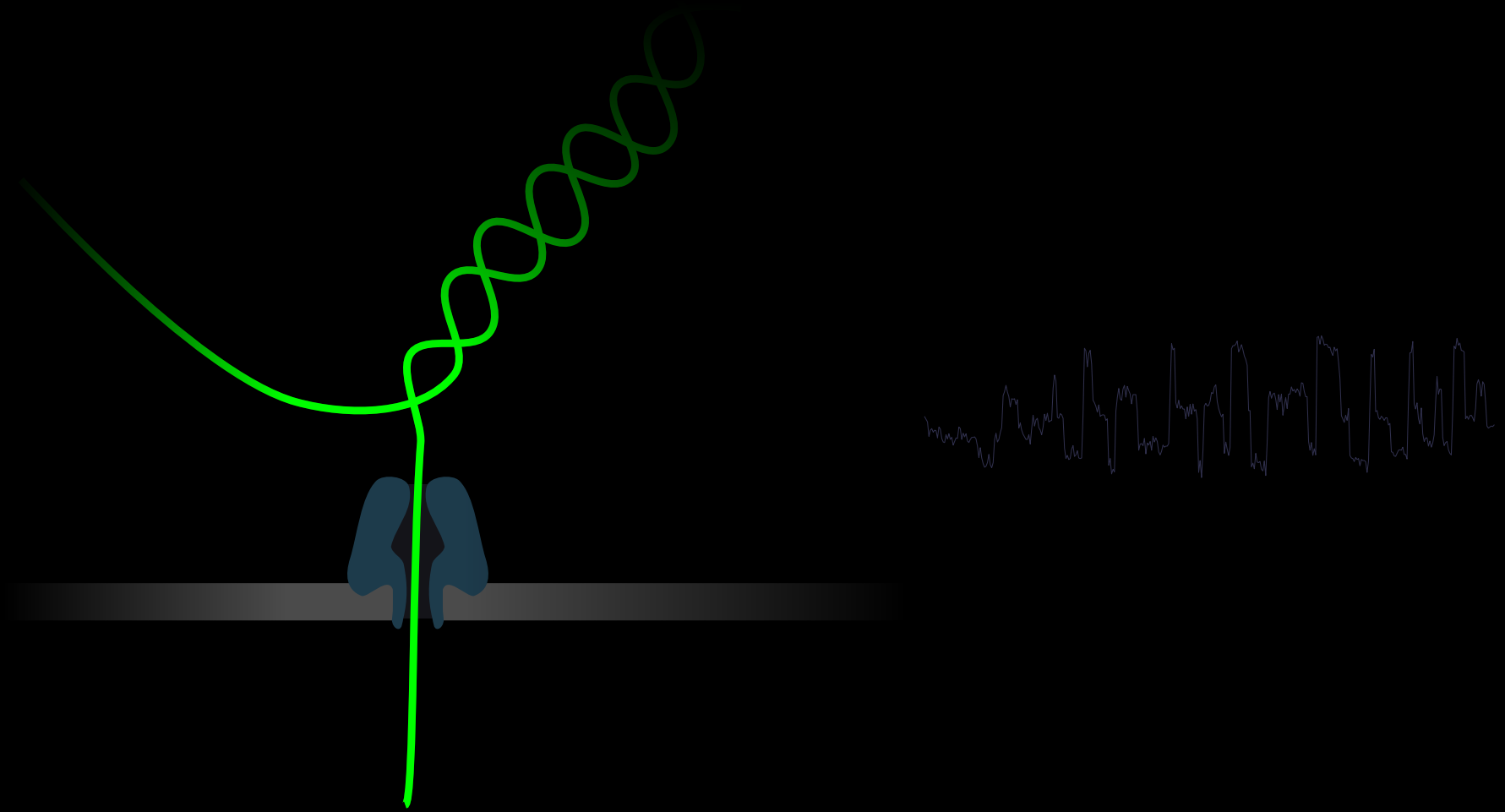




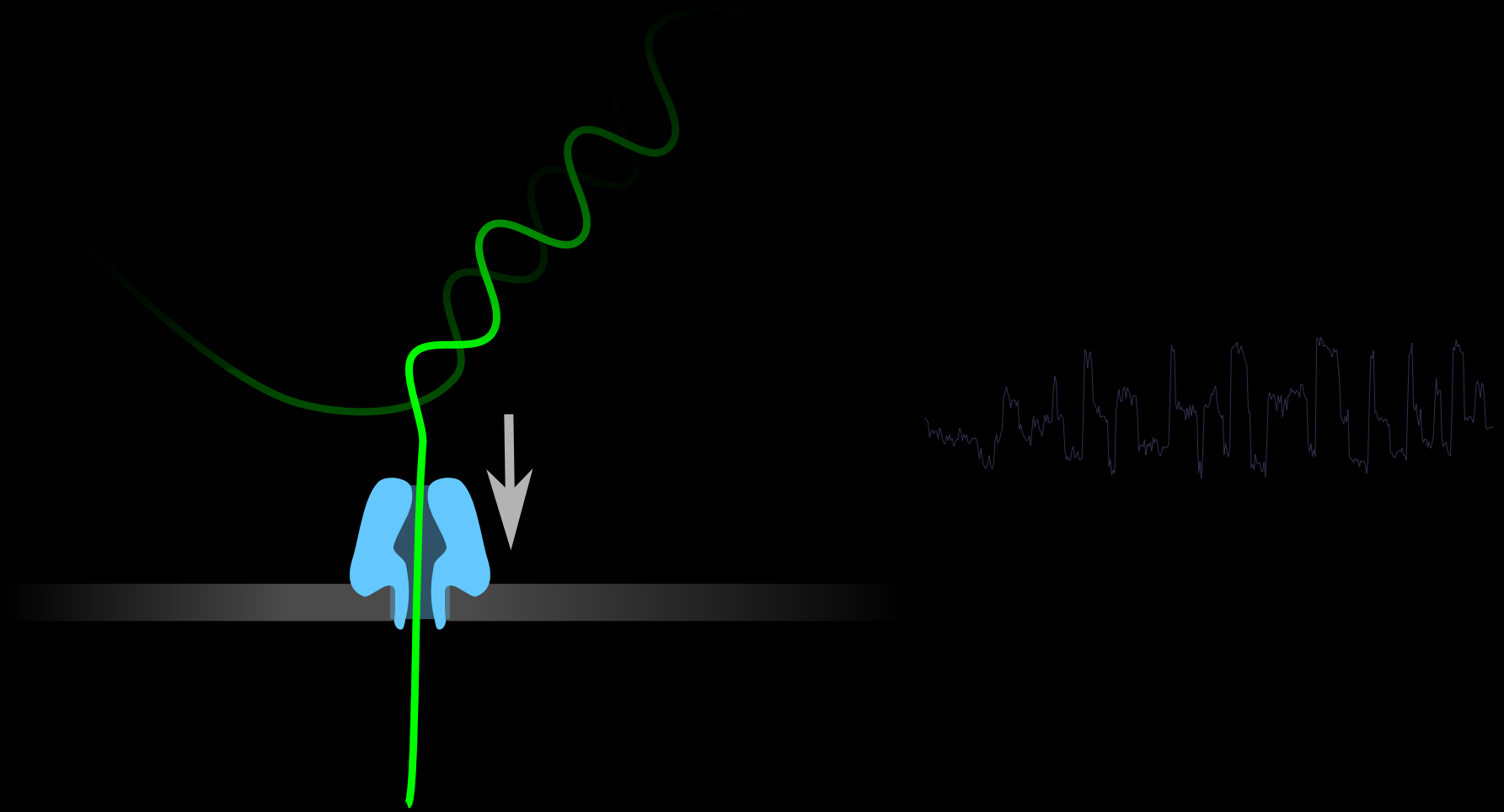
Nanopórové sekvenovanie



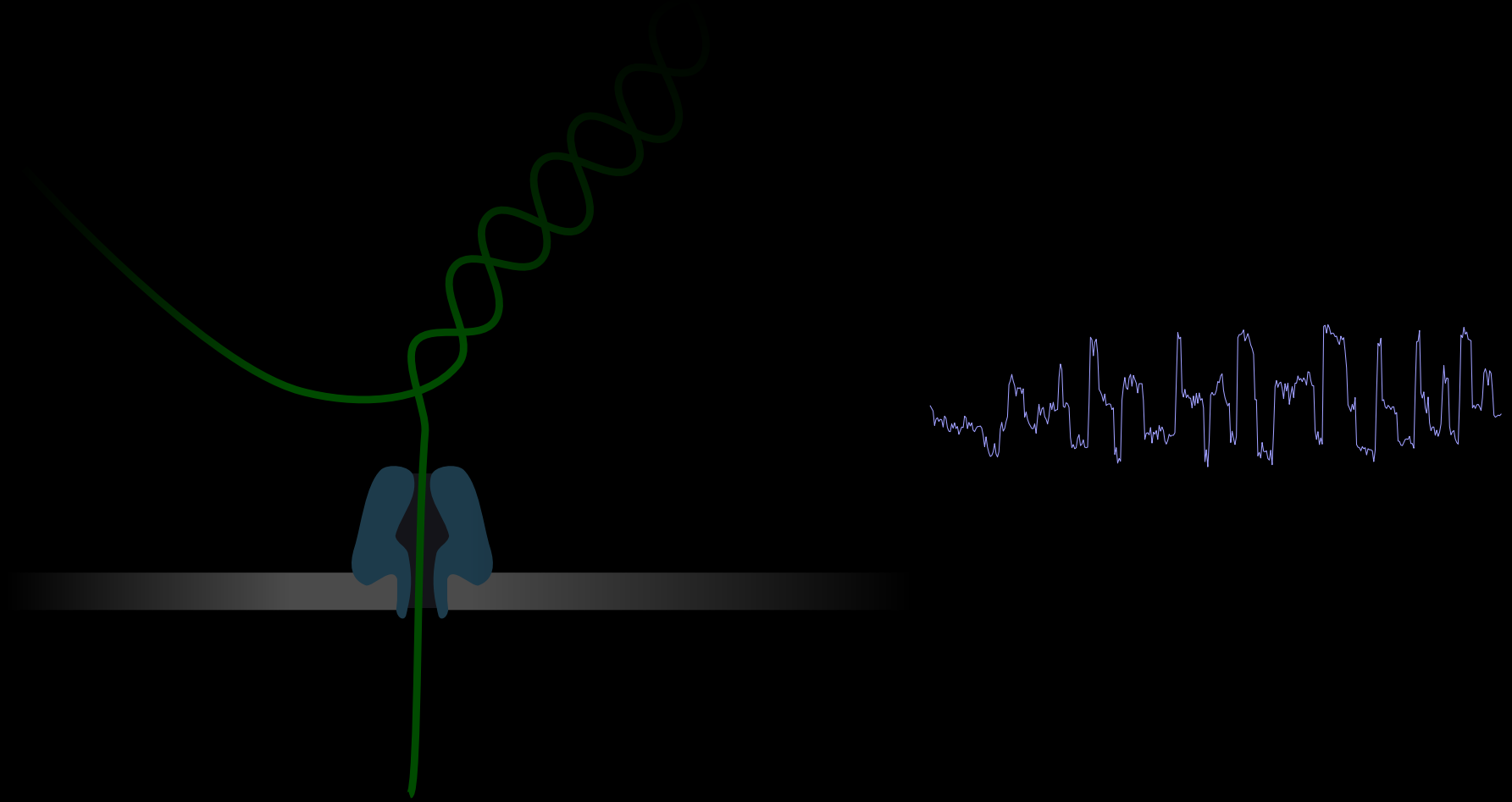
Nanopórové sekvenovanie



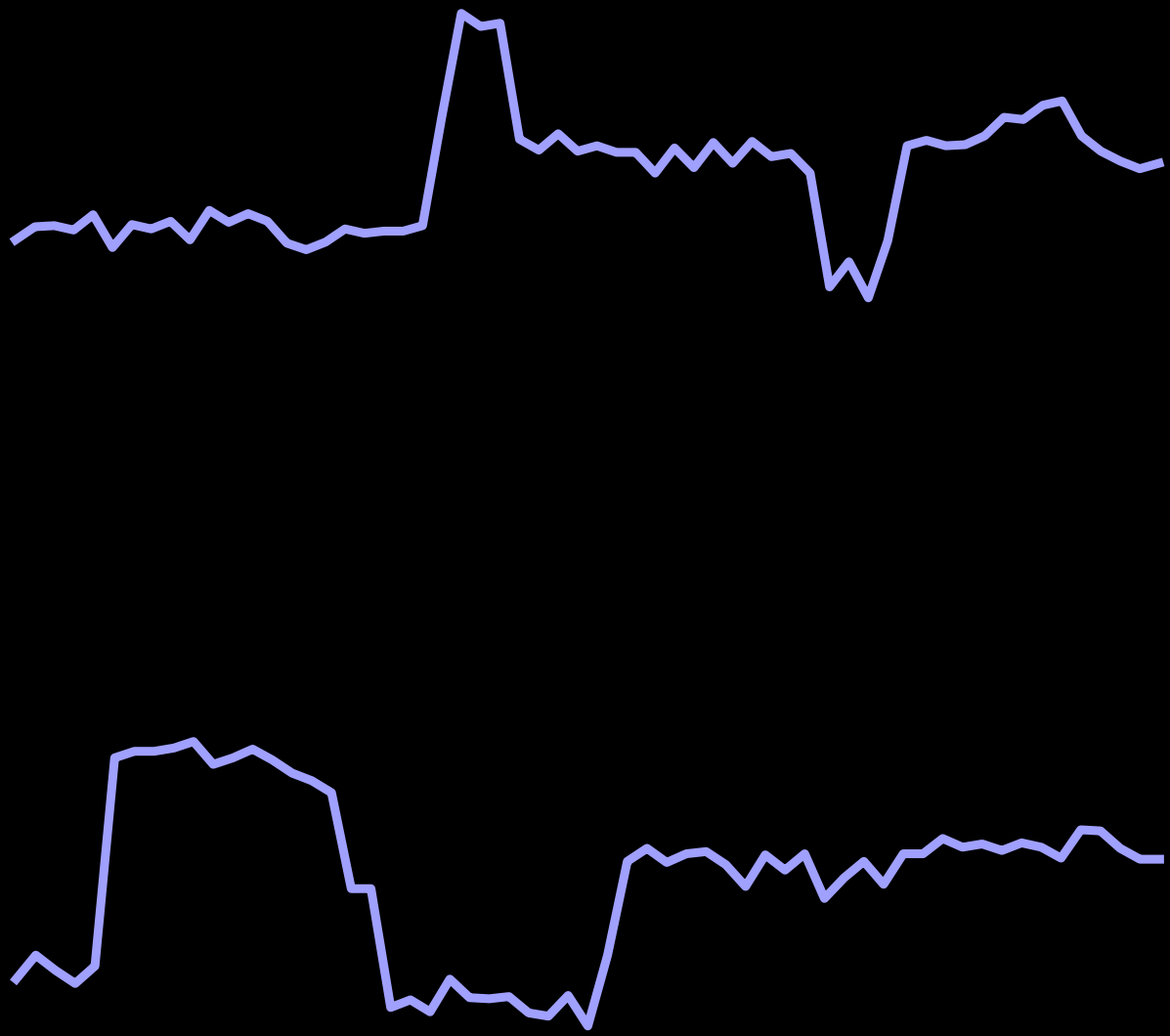
Nanopórové sekvenovanie



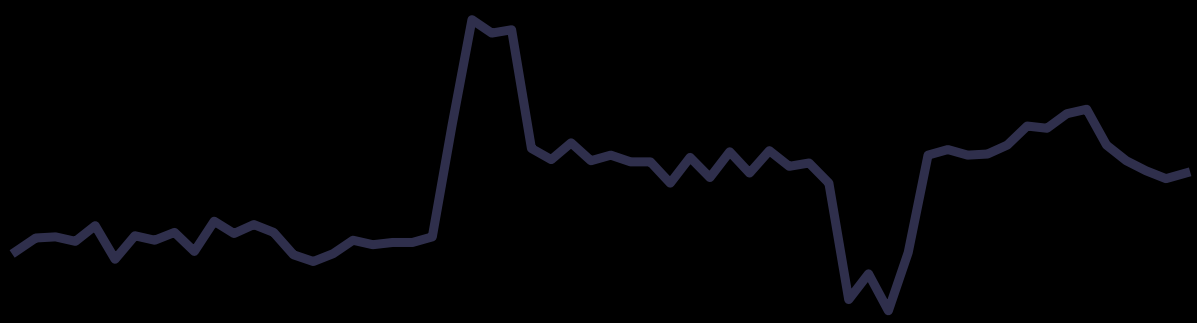
Nanopórové sekvenovanie



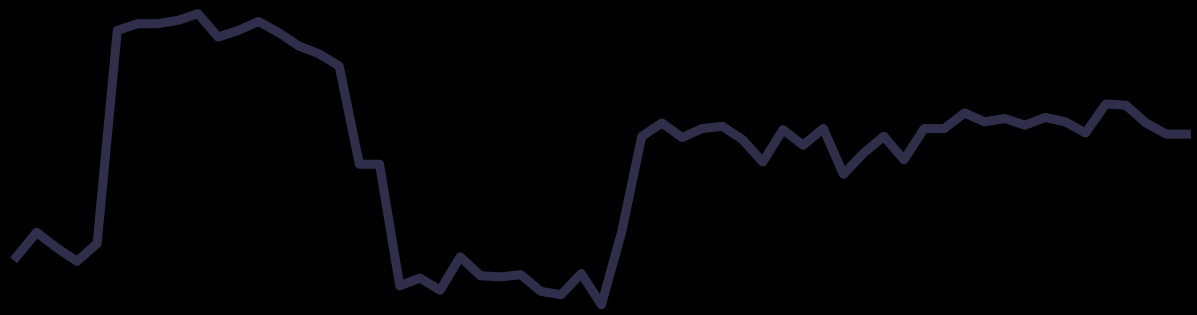
Nerovnomerná rýchlosť



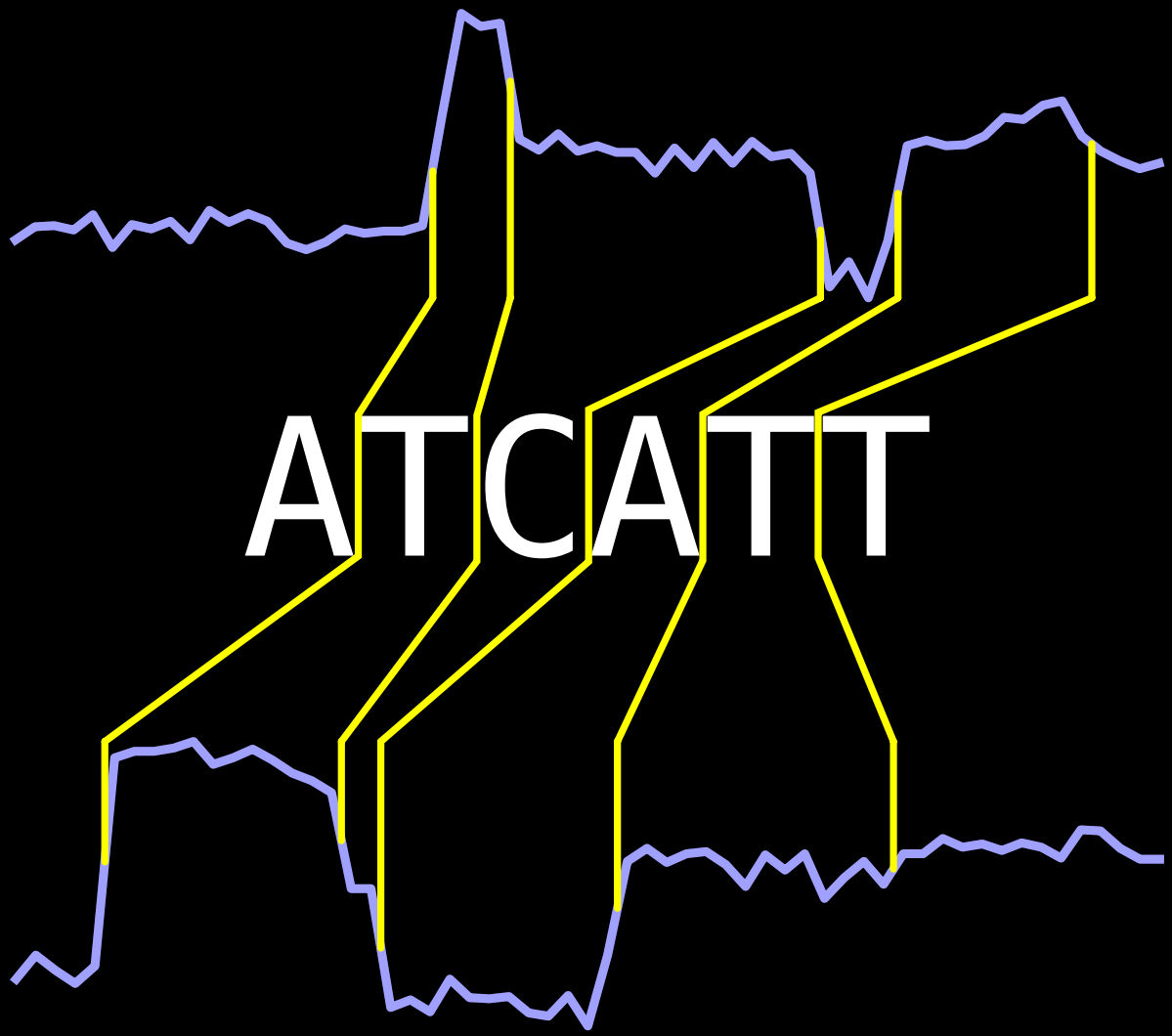
Nerovnomerná rýchlosť



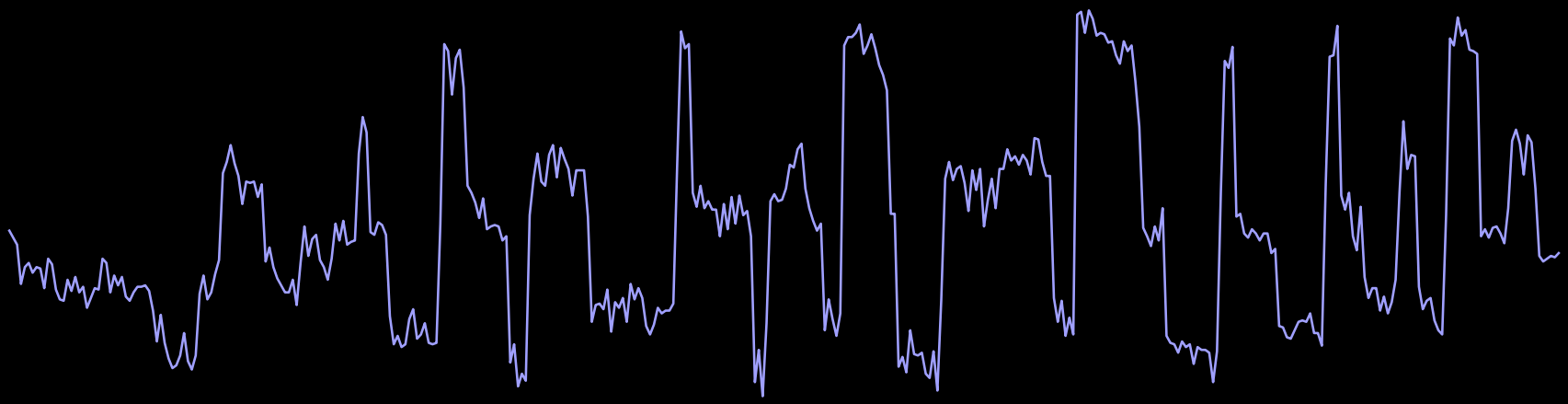
ATCATT



Nerovnomerná rýchlosť

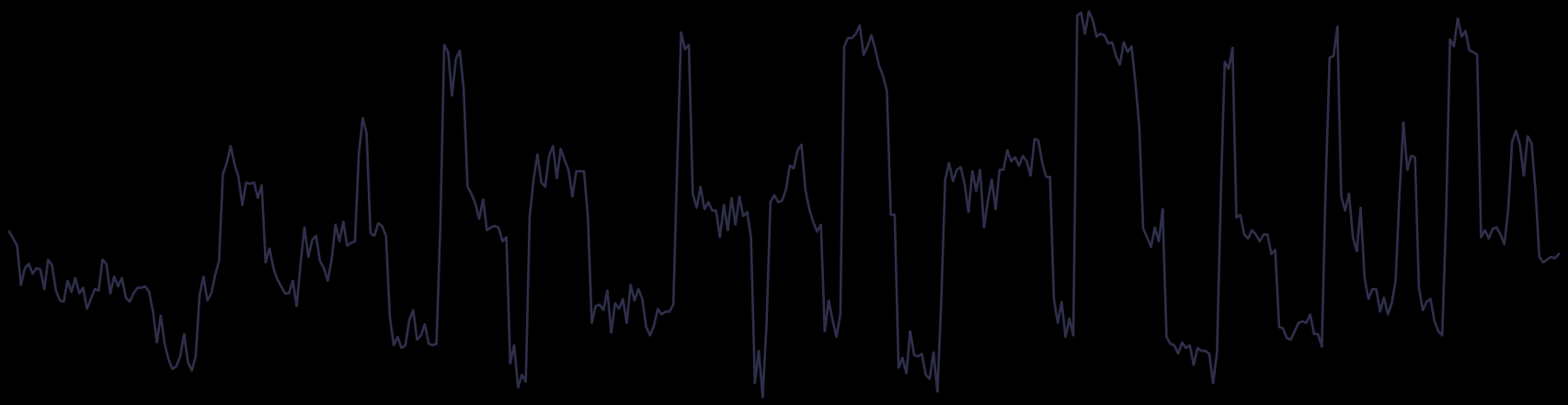


Určovanie báz (basecalling)



AAAAAAACACTTTCATCATTATCA
TTATCATTATCATCATCATATCTA

Určovanie báz (basecalling)

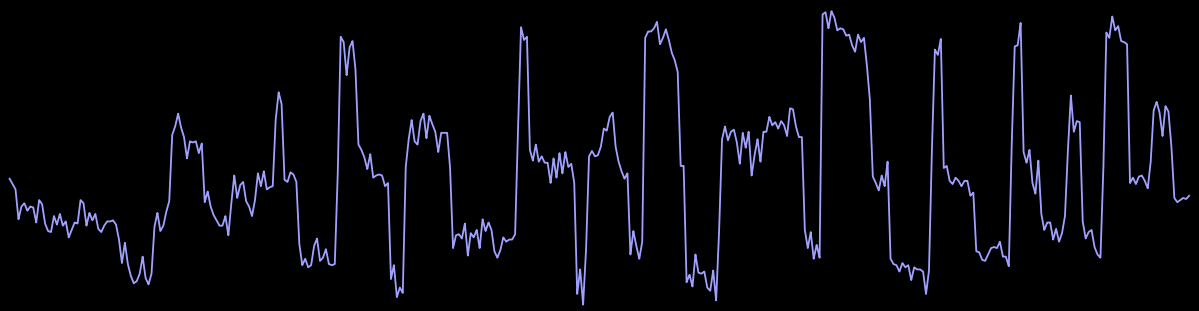


AAAAAAACACTTTCATCATTATCA
TTATCATTATCATCATCATATCTA

85% – 90%

Riešenie

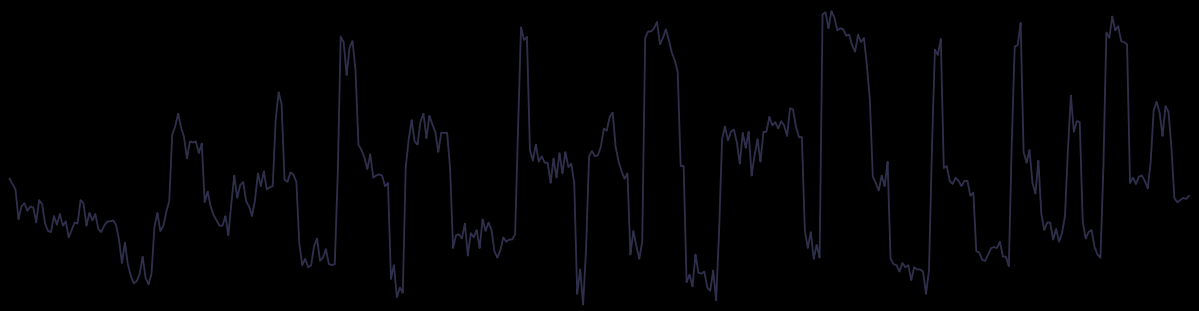
Naivná detekcia



AAAAAAACACTTTCATCATTATCA
TTATCATTATCATCATCATATCTA

AAAAAAACACTTTCATCATTATCA
TTATCATTATCATCATCATATCTA

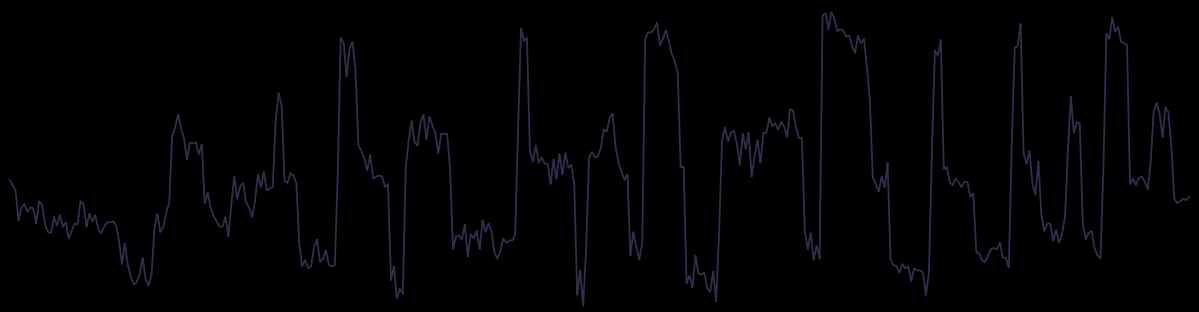
Naivná detekcia



AAAAAAACACTTTCATCATTATCA
TTATCATTATCATCATCATATCTA

AAAAAAACACTTTCATCATTATCA
TTATCATTATCATCATCATATCTA

Naivná detekcia



AAAAAAACACTTTCATCATTATCA
TTATCATTATCATCATCATATCTA

AAAAAAACACTTTCATCATTATCA
TTATCATTATCATCATCATATCTA

Naivná detekcia

- Chybovosť prekladu
- Tvrdá predpoveď
- Nápad 1: detekcia priamo v signáli
- Nápad 2: čiastočné predspracovanie basecallerom

Naivná detekcia

- Chybovosť prekladu
- Tvrdá predpoveď
- Nápad 1: detekcia priamo v signáli
- Nápad 2: čiastočné predspracovanie basecallerom

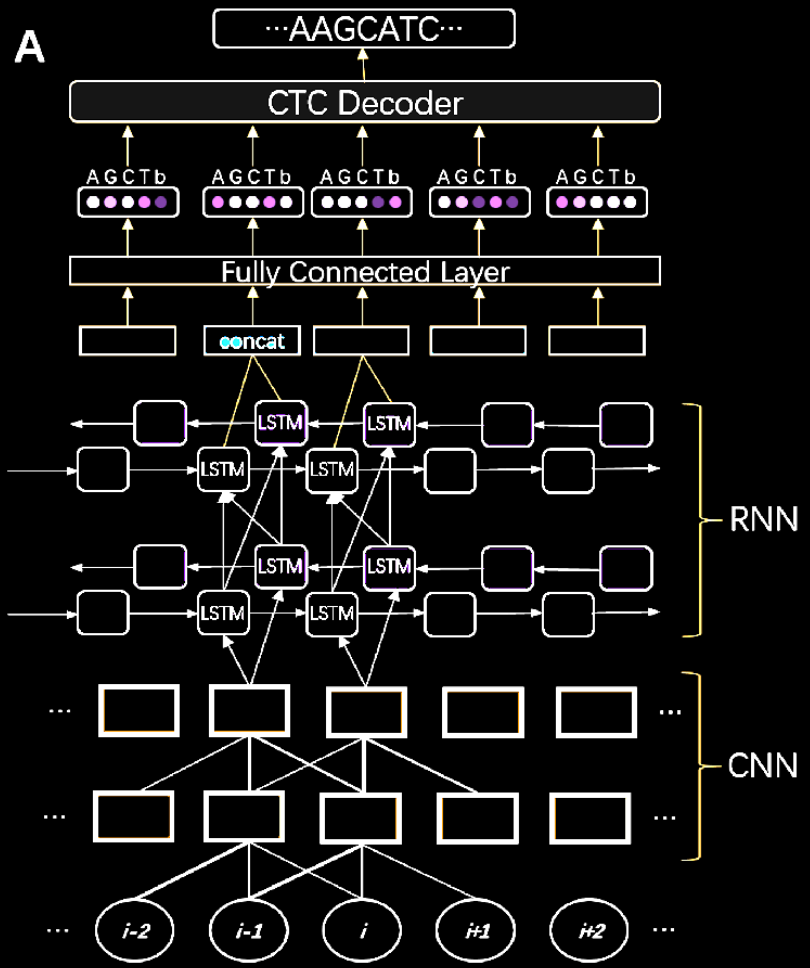
Naivná detekcia

- Chybovosť prekladu
- Tvrdá predpoveď
- Nápad 1: detekcia priamo v signáli
- Nápad 2: čiastočné predspracovanie basecallerom

Naivná detekcia

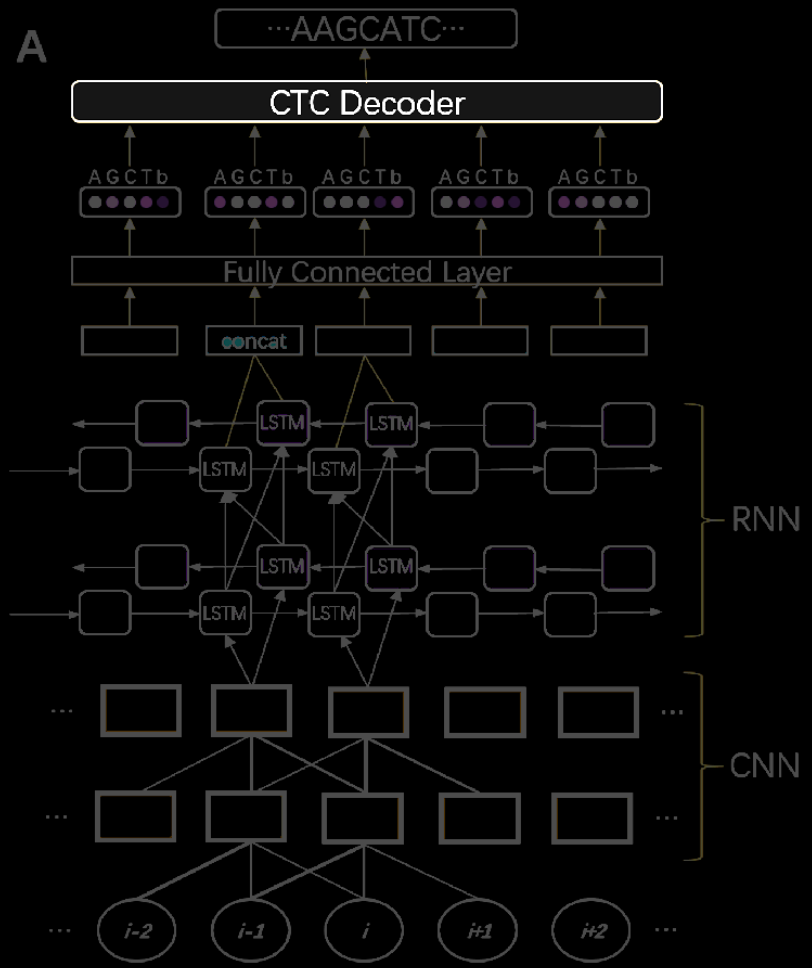
- Chybovosť prekladu
- Tvrdá predpoveď
- Nápad 1: detekcia priamo v signáli
- Nápad 2: čiastočné predspracovanie basecallerom

Chiron (Teng et al. 2018)



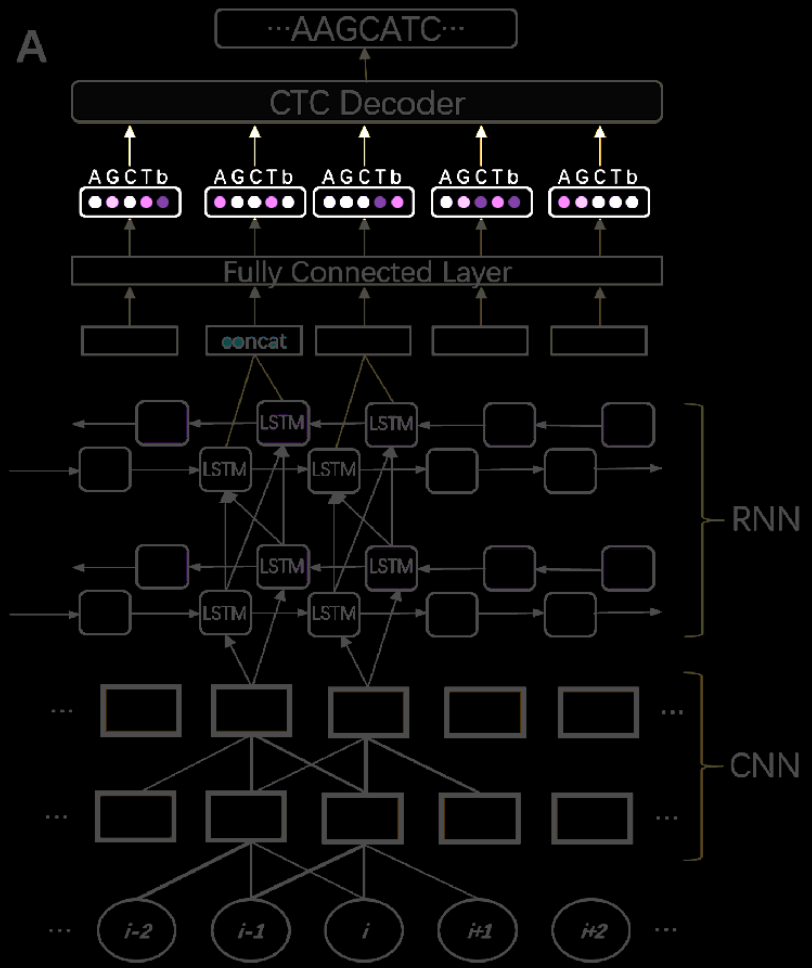
Connectionist Temporal Classification

Chiron (Teng et al. 2018)



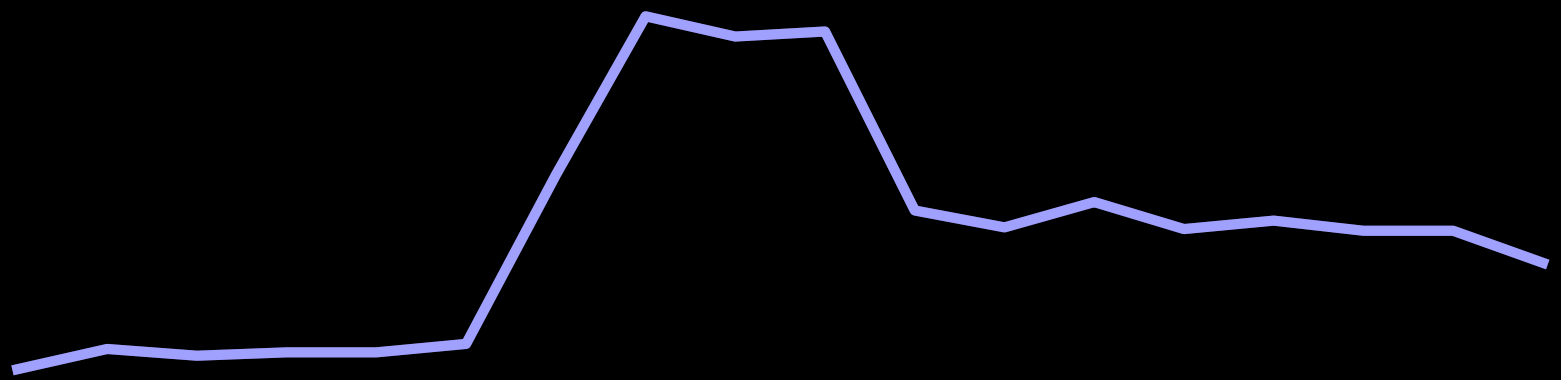
Connectionist Temporal Classification

Chiron (Teng et al. 2018)



Connectionist Temporal Classification

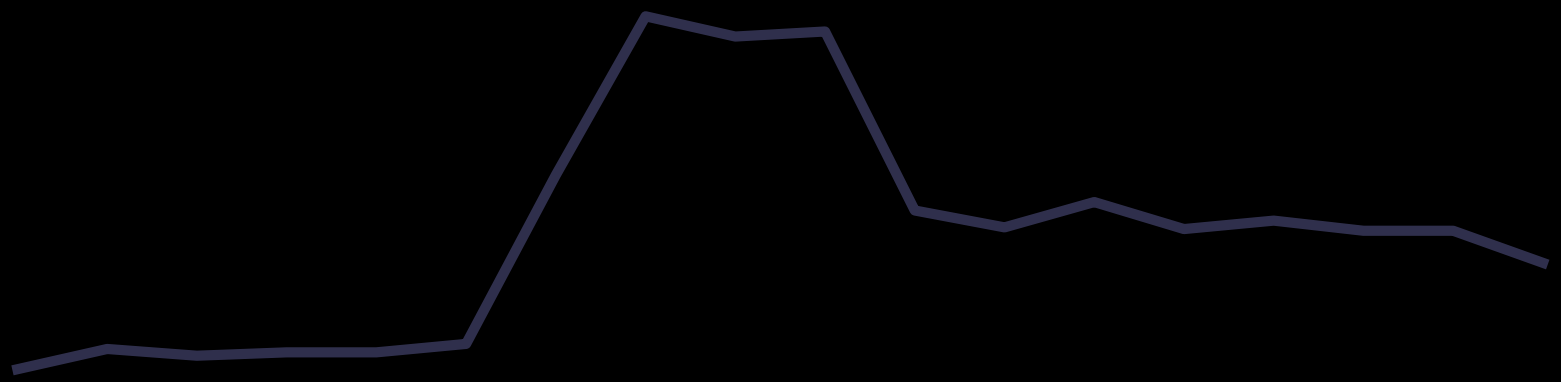
CTC anotácie



- - A A A - - T T C C - - - - - -

A T C

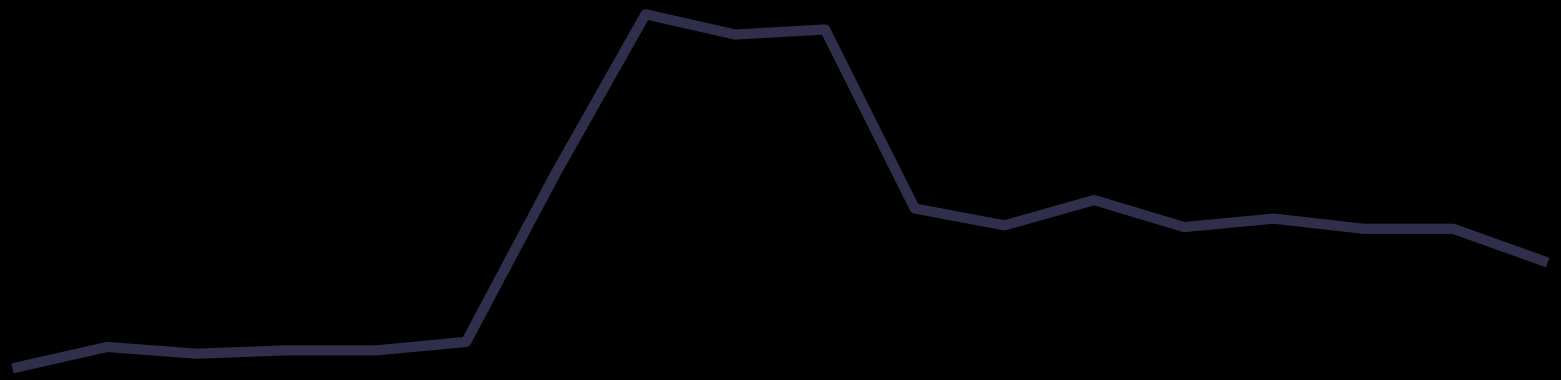
CTC anotácie



- -AAA- -TTCCC- - - - - -

A T C

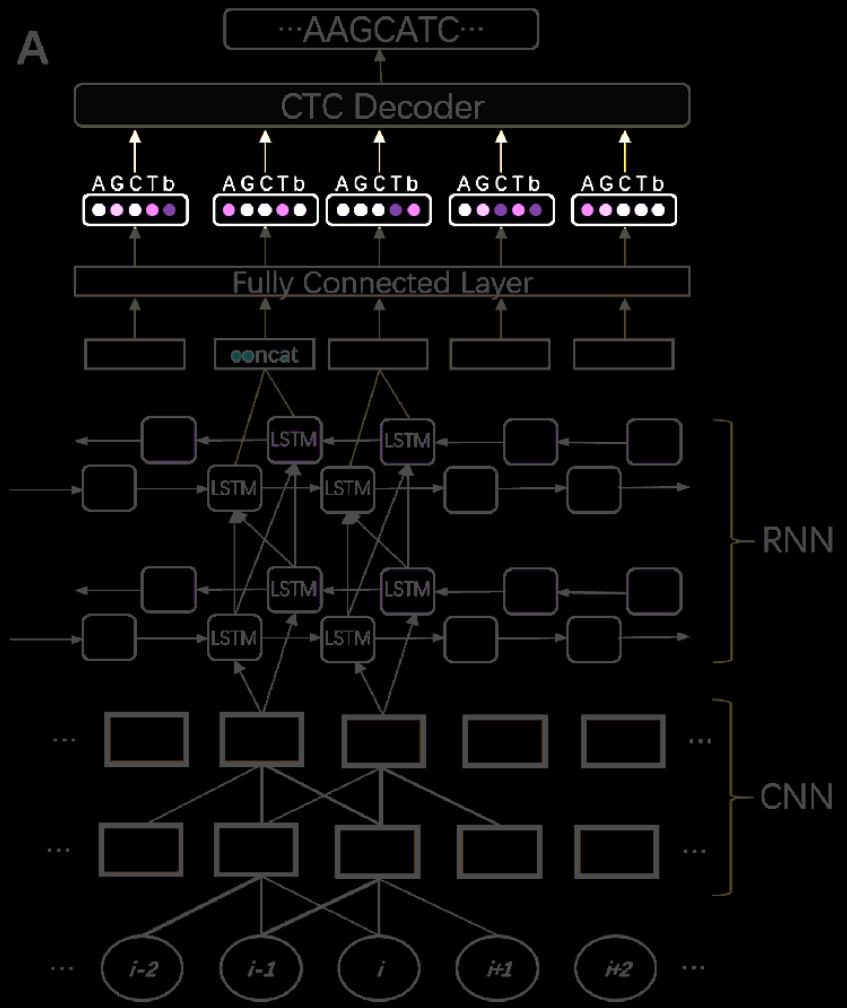
CTC anotácie



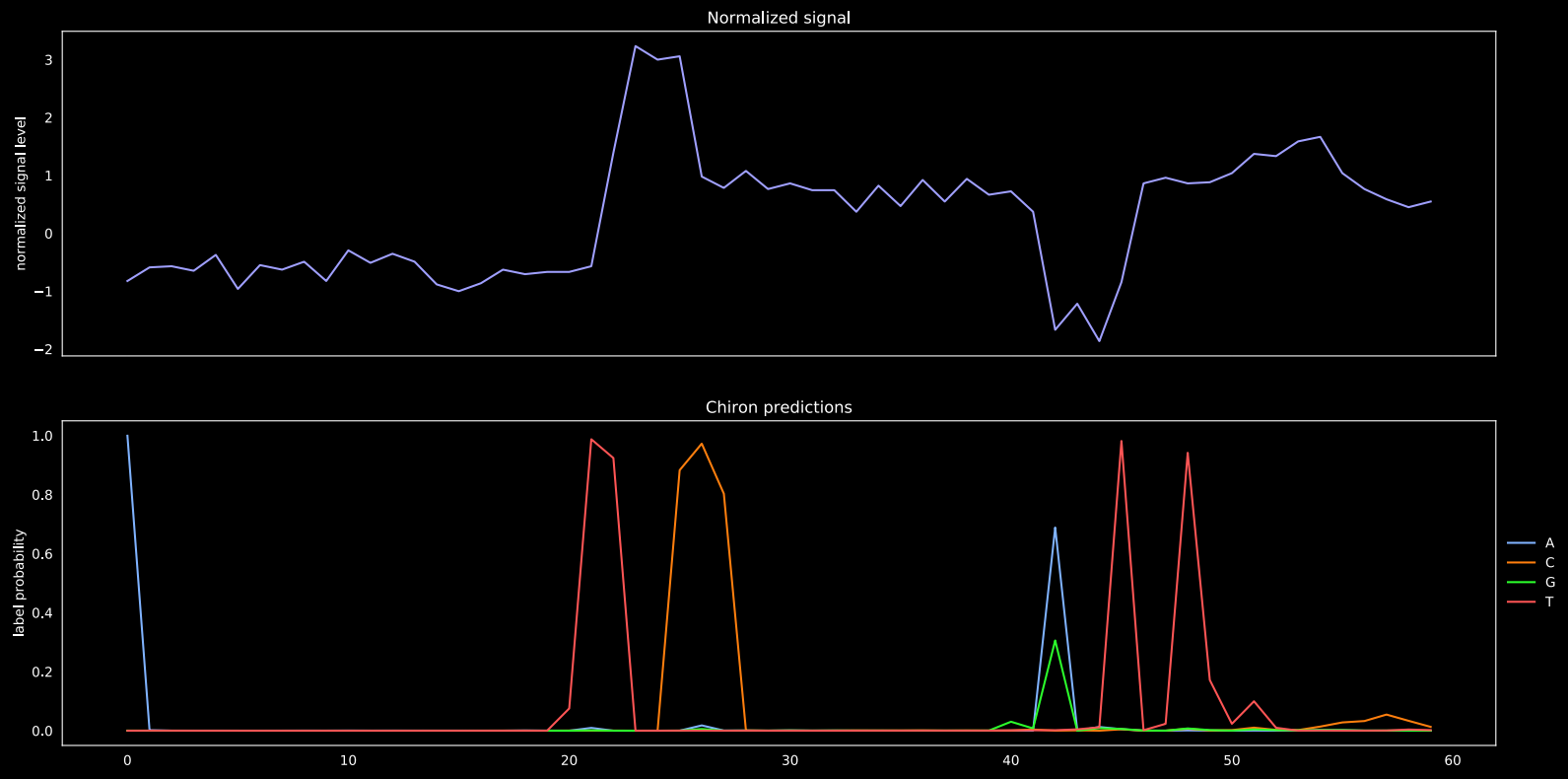
- - A A A - - T T C C - - - - - - -

A T C

CTC predpovede



CTC predpovede



CTC dekódovanie

```
--AAA--TTCCC------
A-----T-C---------
AAAAAATTTCCCCCCCCC
ATC---------------
```

ATC

```
--AAA--GGGGG------
-A-----G----------
AAAAAAGGGGGGGGGGG
AG----------------
```

AG

CTC dekódovanie

```
--AAA--TTCCC------
A-----T-C---------
AAAAAATTTCCCCCCCCC
ATC---------------
```

ATC

```
--AAA--GGGGG------
-A-----G----------
AAAAAAGGGGGGGGGGG
AG----------------
```

AG

CTC dekódovanie

```
--AAA--TTCCC------
A-----T-C---------
AAAAAATTTCCCCCCCCC
ATC---------------
```

ATC

```
--AAA--GGGGG------
-A-----G----------
AAAAAAGGGGGGGGGGG
AG----------------
```

AG

Detekcia v CTC predpovediach

- Tantan (Frith 2011)
- Skrytý Markovovský model (HMM)
- Rozšírenie z nukleotidov na CTC predpovede

Detekcia v CTC predpovediach

- Tantan (Frith 2011)
- Skrytý Markovovský model (HMM)
- Rozšírenie z nukleotidov na CTC predpovede

Detekcia v CTC predpovediach

- Tantan (Frith 2011)
- Skrytý Markovovský model (HMM)
- Rozšírenie z nukleotidov na CTC predpovede

Rozširovanie HMM

- HMM pre nukleotidy
- blanky
- nedokonalosti
- zlučovanie rovnakých označení

Rozširovanie HMM

- HMM pre nukleotidy
- blanky
- nedokonalosti
- zlučovanie rovnakých označení

Rozširovanie HMM

- HMM pre nukleotidy
- blanky
- nedokonalosti
- zlučovanie rovnakých označení

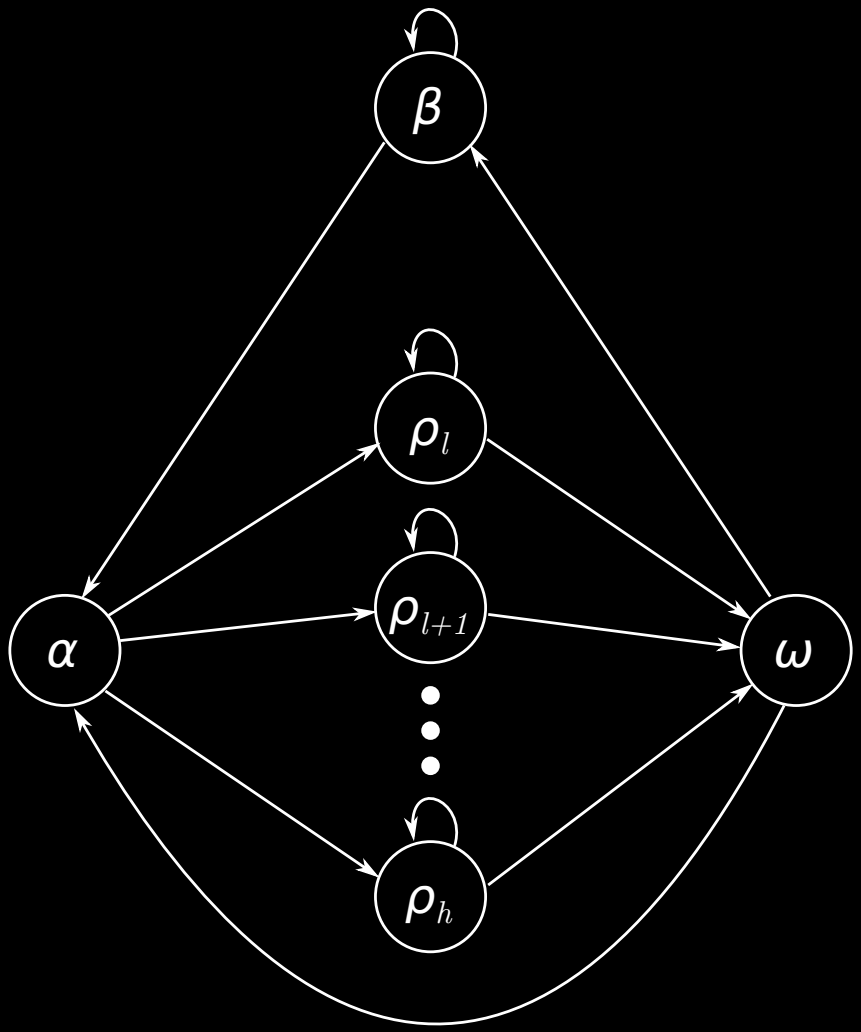
Rozširovanie HMM

- HMM pre nukleotidy
- blanky
- nedokonalosti
- zlučovanie rovnakých označení

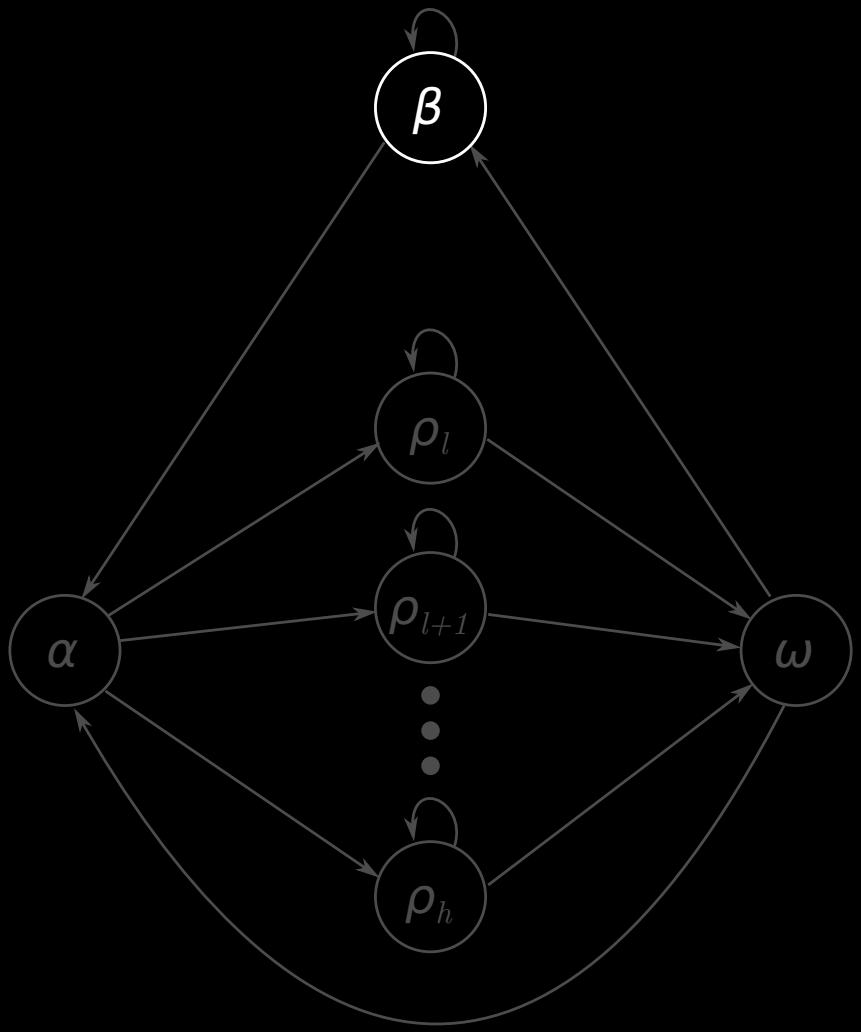
Rozširovanie HMM

- HMM pre nukleotidy
- blanky
- nedokonalosti
- zlučovanie rovnakých označení

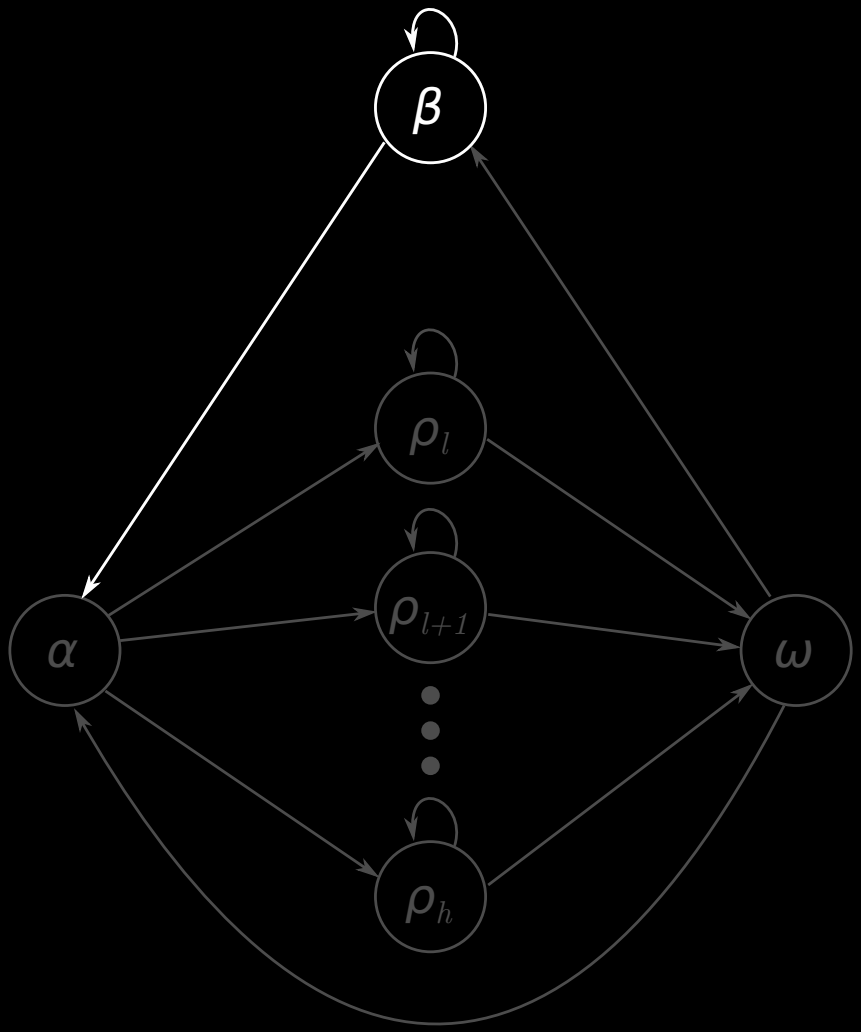
HMM pre nukleotidy



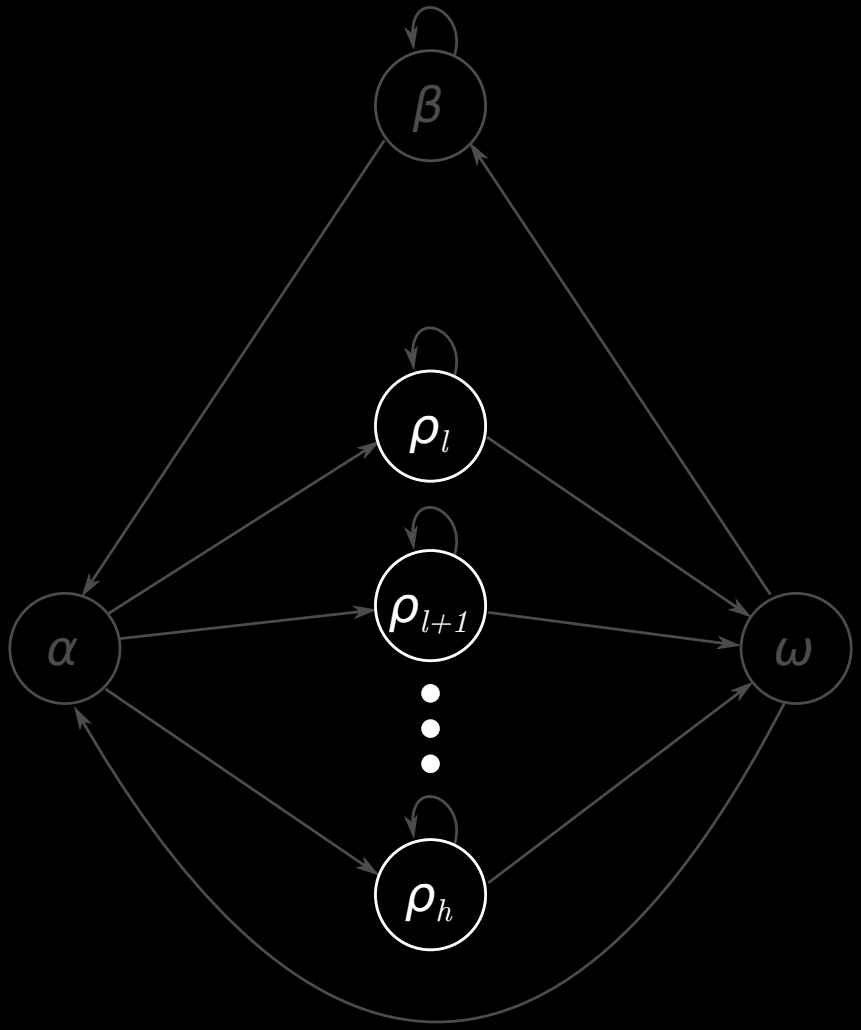
HMM pre nukleotidy



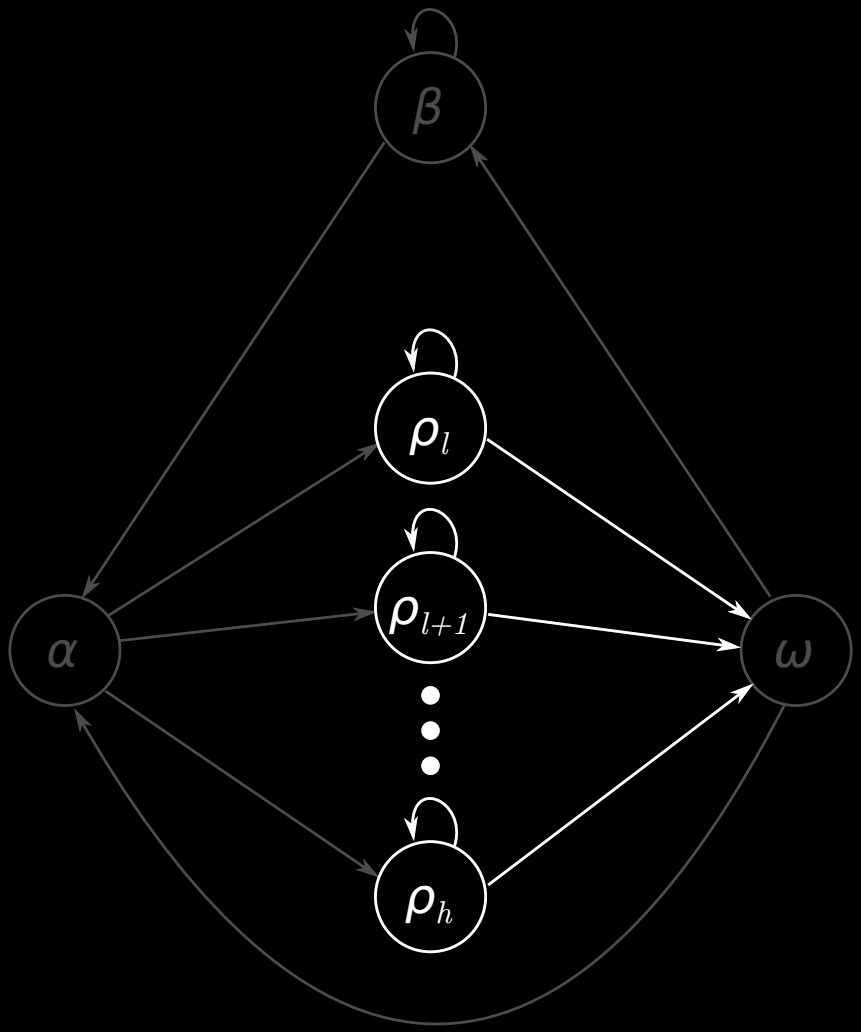
HMM pre nukleotidy



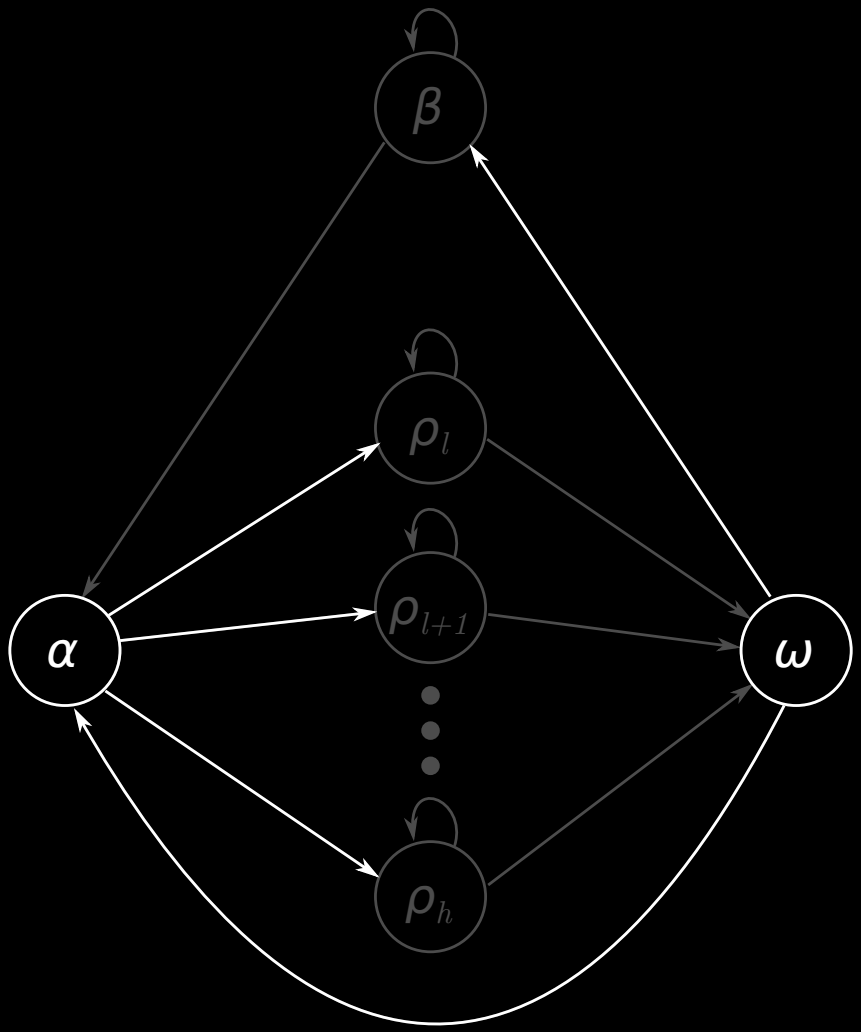
HMM pre nukleotidy



HMM pre nukleotidy

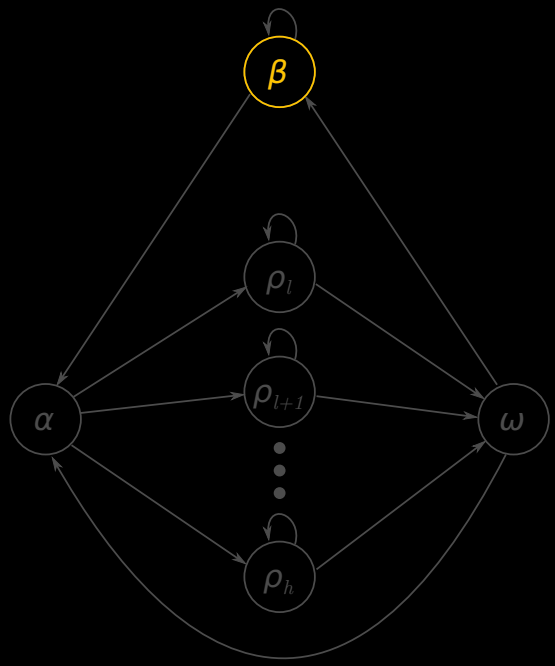


HMM pre nukleotidy



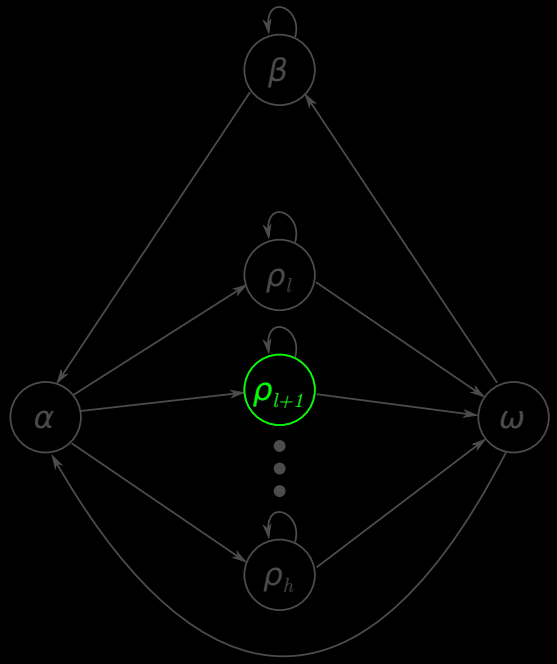
HMM pre nukleotidy

GGTCATCATCATCAA



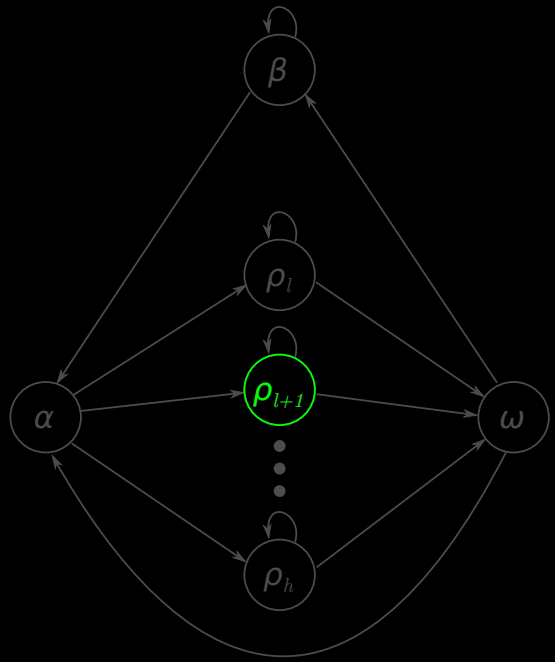
HMM pre nukleotidy

GGTCATCATCATCAA



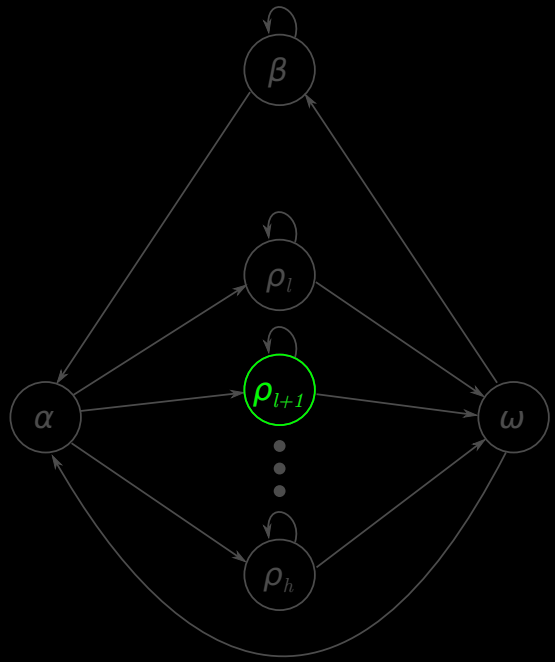
HMM pre nukleotidy

GGTCATCATCATCAA



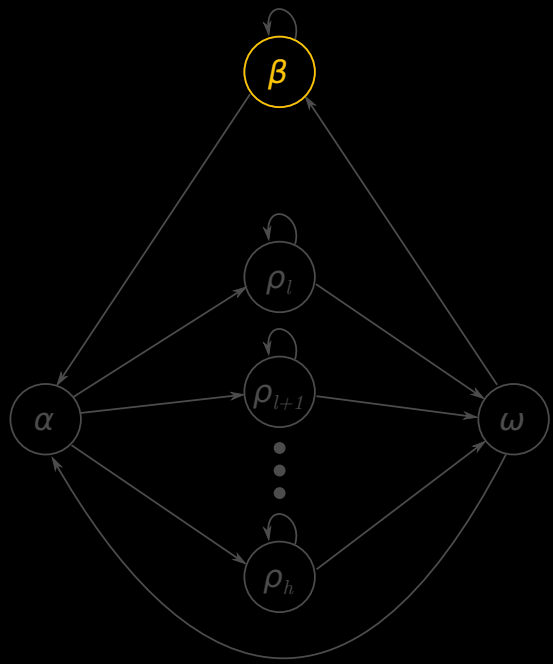
HMM pre nukleotidy

GGTCATCATCATCAA



HMM pre nukleotidy

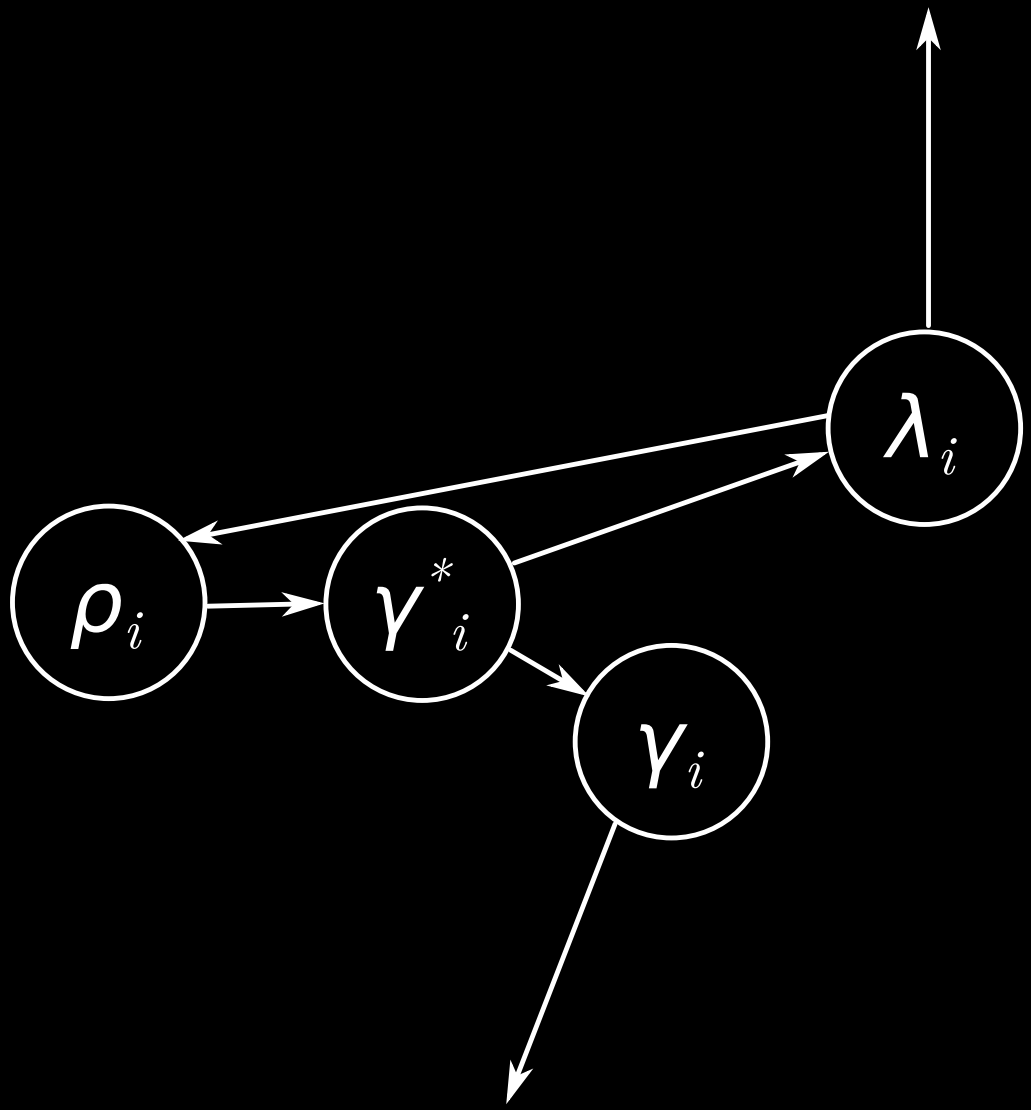
GGTCATCATCATCAA



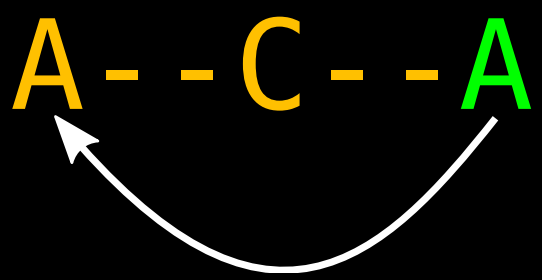
Rozširovanie HMM

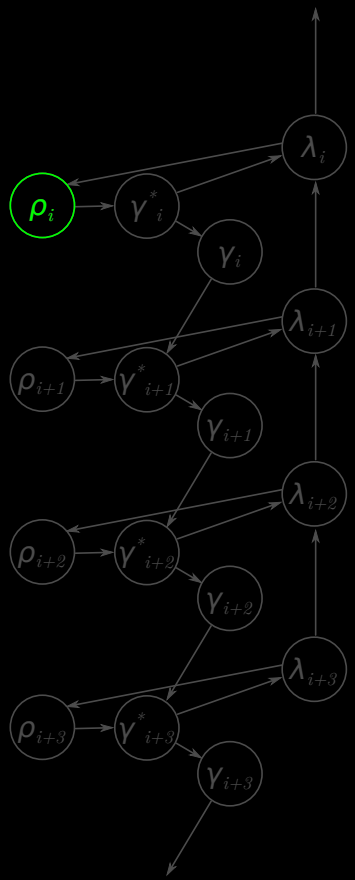
- HMM pre nukleotidy
- blanky
- nedokonalosti
- zlučovanie rovnakých označení

Blanky

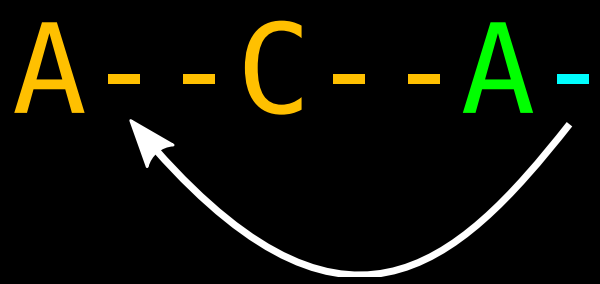


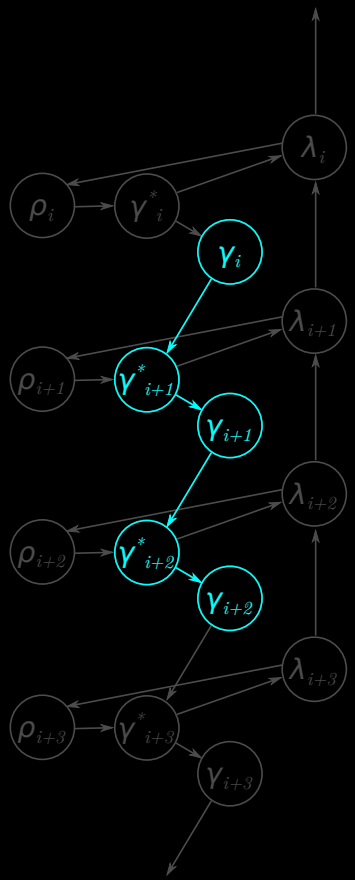
Blanky


A - - C - - A

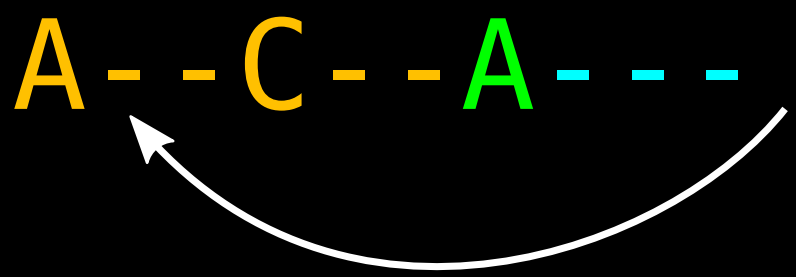

λ_i
ρ_i
γ^*_i
γ_i
λ_{i+1}
ρ_{i+1}
γ^*_{i+1}
γ_{i+1}
λ_{i+2}
ρ_{i+2}
γ^*_{i+2}
γ_{i+2}
λ_{i+3}
ρ_{i+3}
γ^*_{i+3}
γ_{i+3}

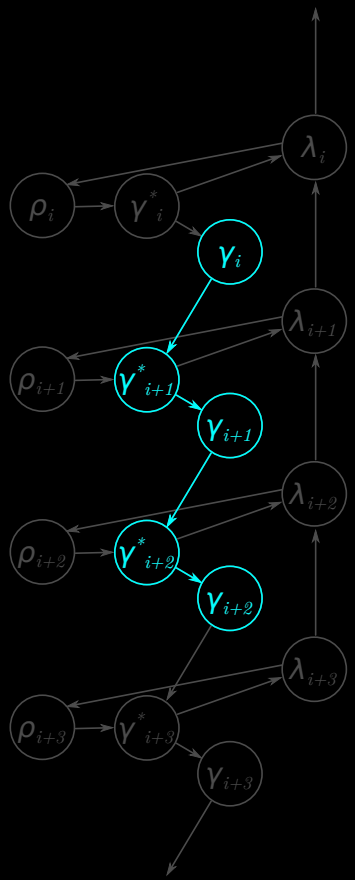
Blanky


A - - C - - A -

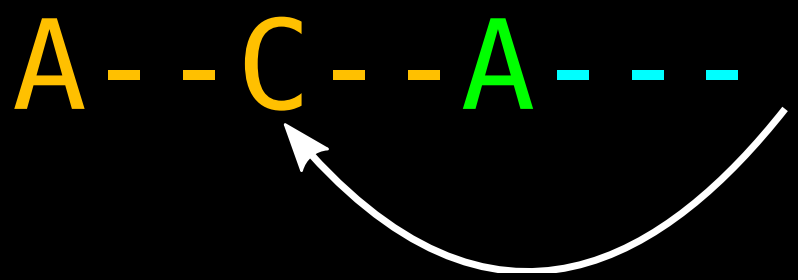

λ_i
ρ_i
γ^*_i
γ_i
λ_{i+1}
ρ_{i+1}
γ^*_{i+1}
γ_{i+1}
λ_{i+2}
ρ_{i+2}
γ^*_{i+2}
γ_{i+2}
λ_{i+3}
ρ_{i+3}
γ^*_{i+3}
γ_{i+3}

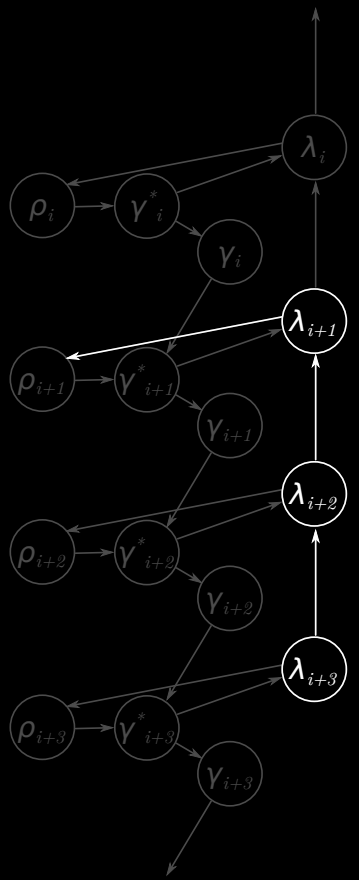
Blanky


A - - C - - A - - -


λ_i
ρ_i
γ^*_i
γ_i
λ_{i+1}
ρ_{i+1}
γ^*_{i+1}
γ_{i+1}
λ_{i+2}
ρ_{i+2}
γ^*_{i+2}
γ_{i+2}
λ_{i+3}
ρ_{i+3}
γ^*_{i+3}
γ_{i+3}

Blanky


A - - C - - A - - -


λ_i
ρ_i
γ^*_i
γ_i
λ_{i+1}
ρ_{i+1}
γ^*_{i+1}
γ_{i+1}
λ_{i+2}
ρ_{i+2}
γ^*_{i+2}
γ_{i+2}
λ_{i+3}
ρ_{i+3}
γ^*_{i+3}
γ_{i+3}

Rozširovanie HMM

- HMM pre nukleotidy
- blanky
- nedokonalosti
- zlučovanie rovnakých označení

Nedokonalosti

- zmenené nukleotidy
- emisná distribúcia v ρ
- inzercie
- emisná distribúcia v γ
- delécie
- prechody z λ

Nedokonalosti

- zmenené nukleotidy
- emisná distribúcia v ρ
- inzercie
- emisná distribúcia v γ
- delécie
- prechody z λ

Nedokonalosti

- zmenené nukleotidy
- emisná distribúcia v ρ
- inzercie
- emisná distribúcia v γ
- delécie
- prechody z λ

Nedokonalosti

- zmenené nukleotidy
- emisná distribúcia v ρ
- inzercie
- emisná distribúcia v γ
- delécie
- prechody z λ

Nedokonalosti

- zmenené nukleotidy
- emisná distribúcia v ρ
- inzercie
- emisná distribúcia v γ
- delécie
- prechody z λ

Nedokonalosti

- zmenené nukleotidy
- emisná distribúcia v ρ
- inzercie
- emisná distribúcia v γ
- delécie
- prechody z λ

Rozširovanie HMM

- HMM pre nukleotidy
- blanky
- nedokonalosti
- zlučovanie rovnakých označení

Zlučovanie rovnakých označení

- modelovanie: zhruba 15-krát viac stavov
- ignorovanie: nižšia presnosť
- 

- Heuristika: lokálne maximá

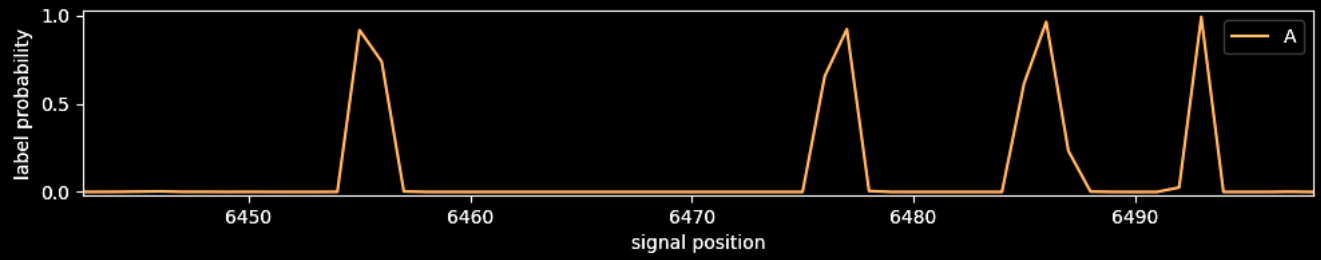
Zlučovanie rovnakých označení

- modelovanie: zhruba 15-krát viac stavov
- ignorovanie: nižšia presnosť

- Heuristika: lokálne maximá

Zlučovanie rovnakých označení

- modelovanie: zhruba 15-krát viac stavov
- ignorovanie: nižšia presnosť



- Heuristika: lokálne maximá

Zlučovanie rovnakých označení

- modelovanie: zhruba 15-krát viac stavov
- ignorovanie: nižšia presnosť

- Heuristika: lokálne maximá

Vyhodnotenie

Datasety

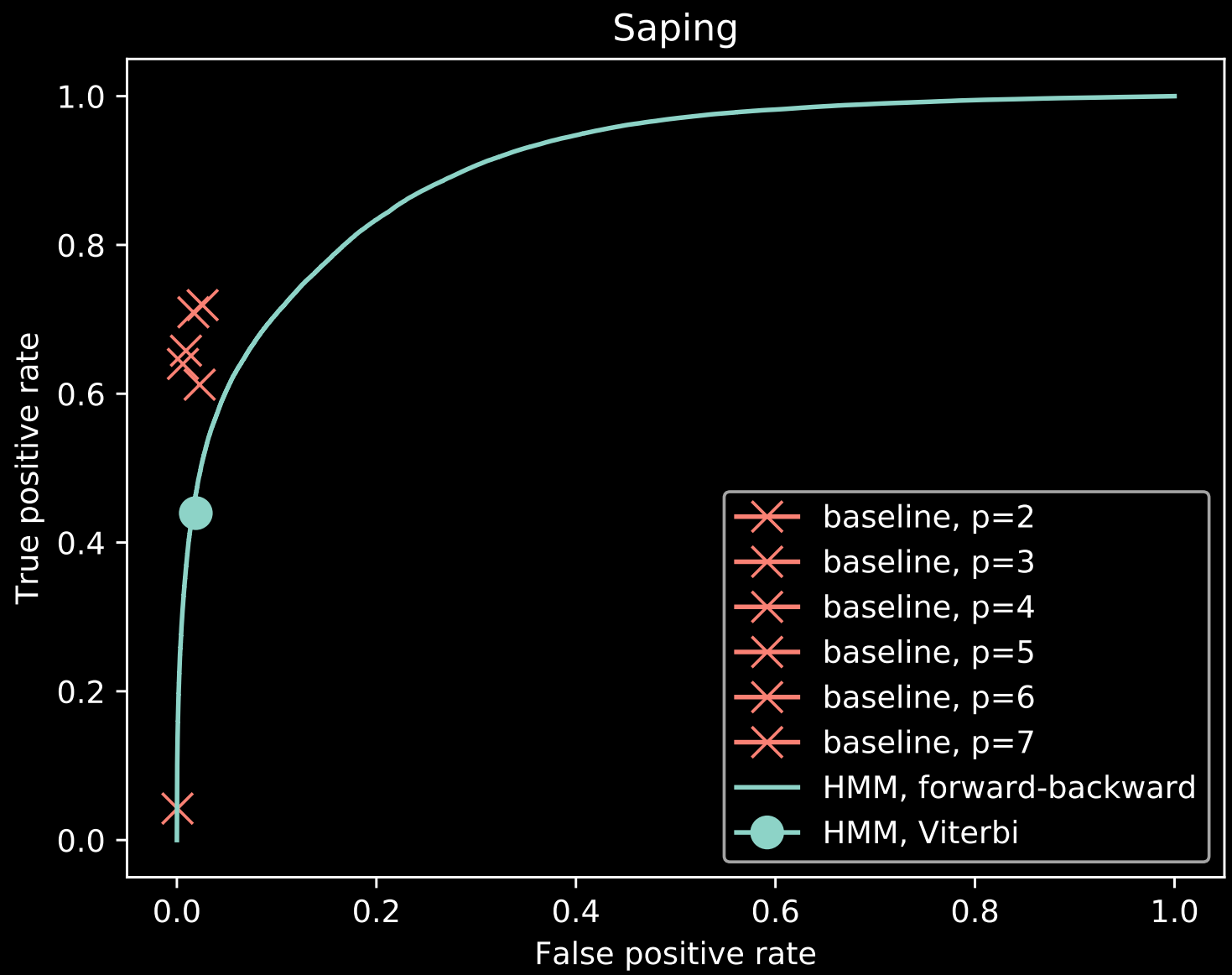
Name	**Reads**	**Repeats ratio**	**Average repeat period**
Saping	100	7.61%	5.84
Jamang	100	18.07%	63.72

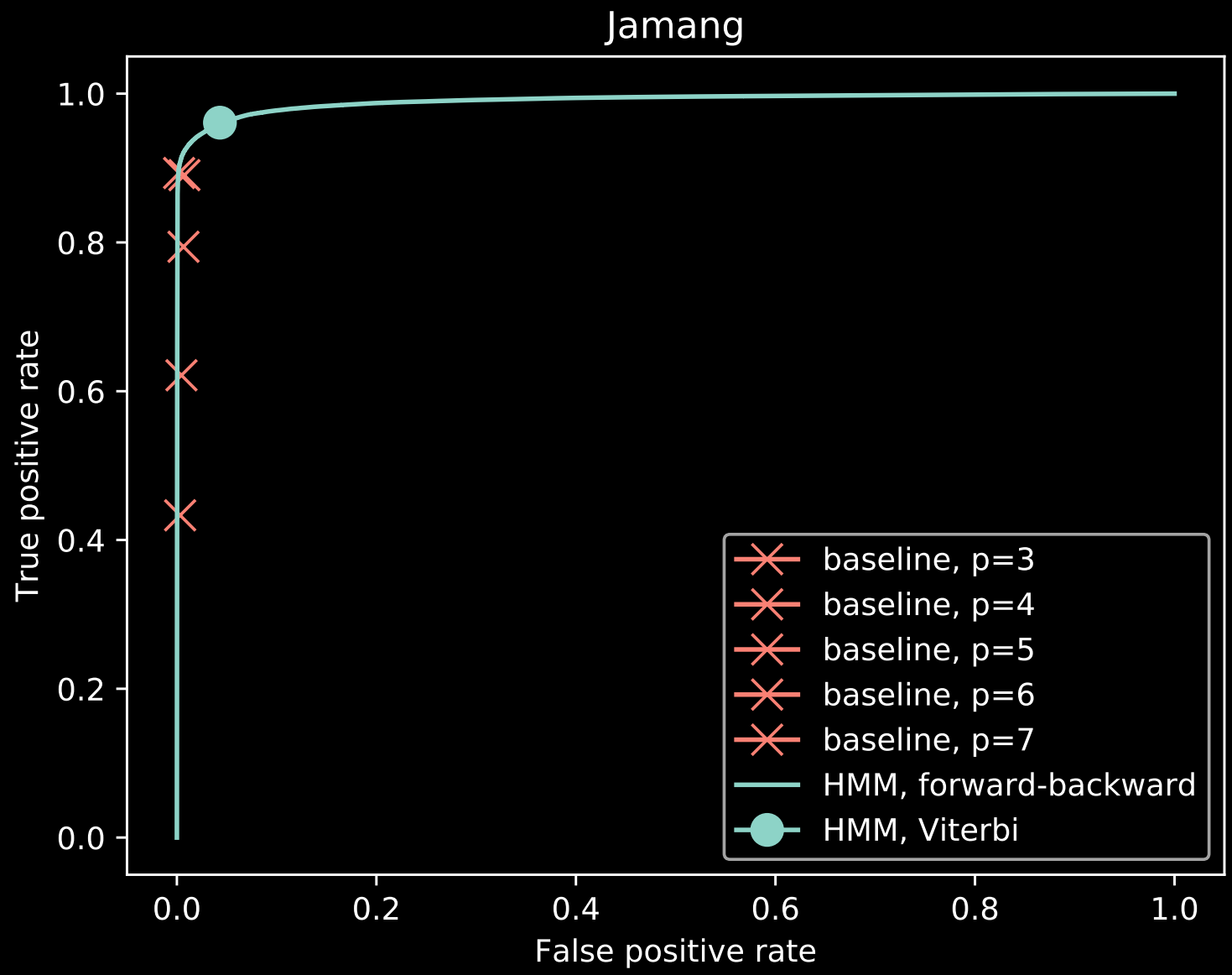
Metriky

- True positive rate
- False positive rate

Metriky

- True positive rate
- False positive rate


Saping
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
baseline, p=2
baseline, p=3
baseline, p=4
baseline, p=5
baseline, p=6
baseline, p=7
HMM, forward-backward
HMM, Viterbi


Jamang
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
baseline, p=3
baseline, p=4
baseline, p=5
baseline, p=6
baseline, p=7
HMM, forward-backward
HMM, Viterbi

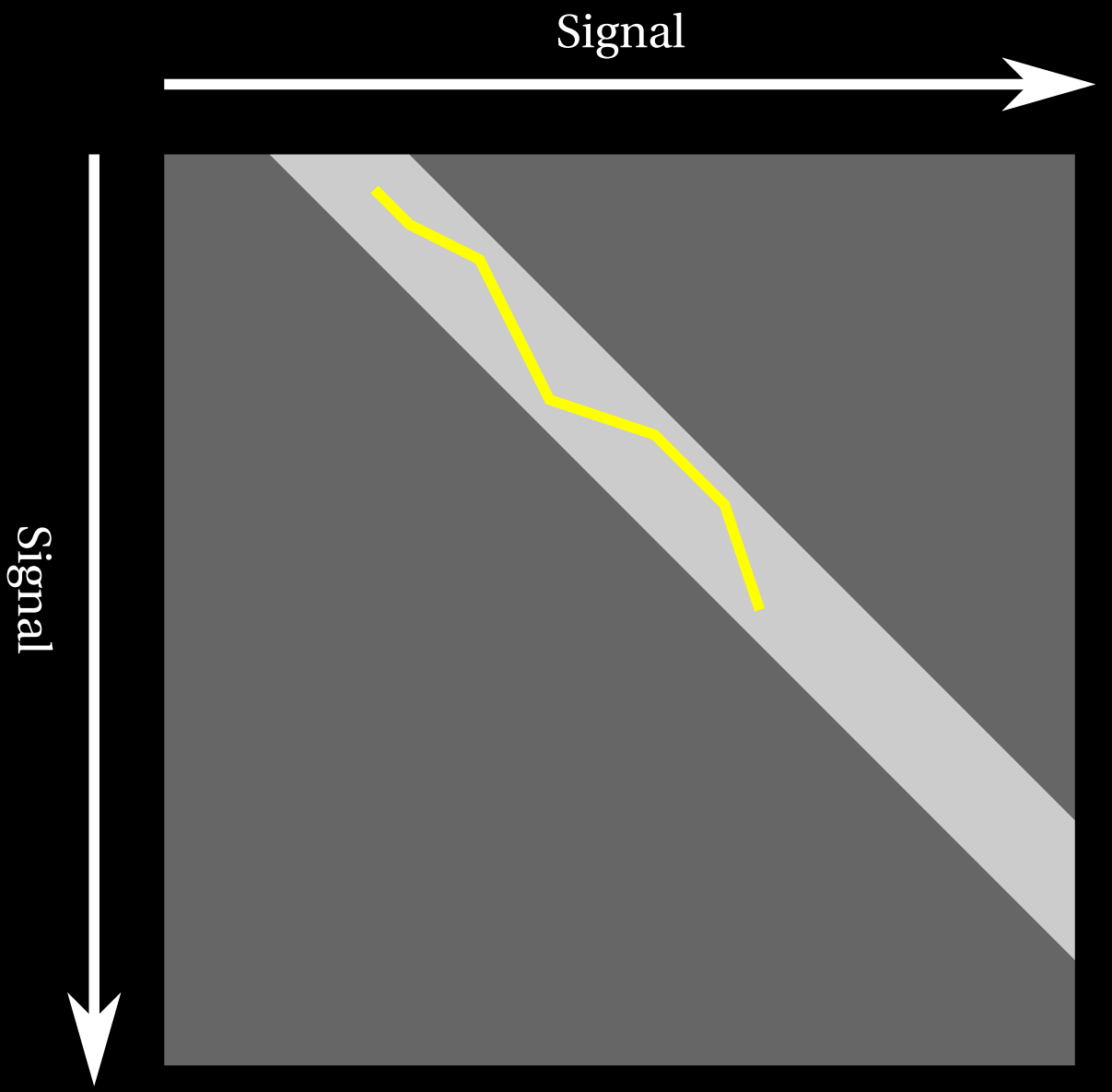
Použité obrázky

H. Teng et al. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. GigaScience, 7(5), 04 2018. giy037.

Otázka 1

- Akú funkciu optimalizuje algoritmus z kapitoly 3?
- Dala by sa upraviť tak, aby bolo možné použiť efektivnejší algoritmus?

Otázka 1



Otázka 1

- Akú funkciu optimalizuje algoritmus z kapitoly 3?
- Dala by sa upraviť tak, aby bolo možné použiť efektivnejší algoritmus?

Otázka 1

- $s_1 s_2 \ldots s_n$
- $t_1 t_2 \ldots t_m$
- $\begin{matrix} s_{i_1} & s_{i_2} & \ldots & s_{i_l} \\ t_{j_1} & t_{j_2} & \ldots & t_{j_l} \end{matrix}$
- $i_k \leq i_{k+1} \leq i_k + 1$
- $l \geq m, n$
- $\frac{b(n+m)}{2} - \sum_{k=1}^{l} (s_{i_k} - t_{i_k})^2$
- $b =$ bonus

Otázka 1

- $s_1 s_2 \ldots s_n$
- $t_1 t_2 \ldots t_m$
- $\begin{matrix} s_{i_1} & s_{i_2} & \ldots & s_{i_l} \\ t_{j_1} & t_{j_2} & \ldots & t_{j_l} \end{matrix}$
- $i_k \leq i_{k+1} \leq i_k + 1$
- $l \geq m, n$
- $\frac{b(n+m)}{2} - \sum_{k=1}^{l} (s_{i_k} - t_{i_k})^2$
- $b = \text{bonus}$

Otázka 1

- $s_1 s_2 \ldots s_n$
- $t_1 t_2 \ldots t_m$
- $$\begin{matrix} s_{i_1} & s_{i_2} & \ldots & s_{i_l} \\ t_{j_1} & t_{j_2} & \ldots & t_{j_l} \end{matrix}$$
- $i_k \leq i_{k+1} \leq i_k + 1$
- $l \geq m, n$
- $\frac{b(n+m)}{2} - \sum_{k=1}^{l} (s_{i_k} - t_{i_k})^2$
- $b = \text{bonus}$

Otázka 1

- $s_1 s_2 \ldots s_n$
- $t_1 t_2 \ldots t_m$
- $\begin{matrix} s_{i_1} & s_{i_2} & \ldots & s_{i_l} \\ t_{j_1} & t_{j_2} & \ldots & t_{j_l} \end{matrix}$
- $i_k \leq i_{k+1} \leq i_k + 1$
- $l \geq m, n$
- $\frac{b(n+m)}{2} - \sum_{k=1}^{l} (s_{i_k} - t_{i_k})^2$
- $b = \text{bonus}$

Otázka 1

- $s_1 s_2 \ldots s_n$
- $t_1 t_2 \ldots t_m$
- $\begin{matrix} s_{i_1} & s_{i_2} & \ldots & s_{i_l} \\ t_{j_1} & t_{j_2} & \ldots & t_{j_l} \end{matrix}$
- $i_k \leq i_{k+1} \leq i_k + 1$
- $l \geq m, n$
- $\frac{b(n+m)}{2} - \sum_{k=1}^{l} (s_{i_k} - t_{i_k})^2$
- $b = \text{bonus}$

Otázka 1

- $s_1 s_2 \ldots s_n$
- $t_1 t_2 \ldots t_m$
- $\begin{matrix} s_{i_1} & s_{i_2} & \ldots & s_{i_l} \\ t_{j_1} & t_{j_2} & \ldots & t_{j_l} \end{matrix}$
- $i_k \leq i_{k+1} \leq i_k + 1$
- $l \geq m, n$
- $\frac{b(n+m)}{2} - \sum_{k=1}^{l} (s_{i_k} - t_{i_k})^2$
- $b =$ bonus

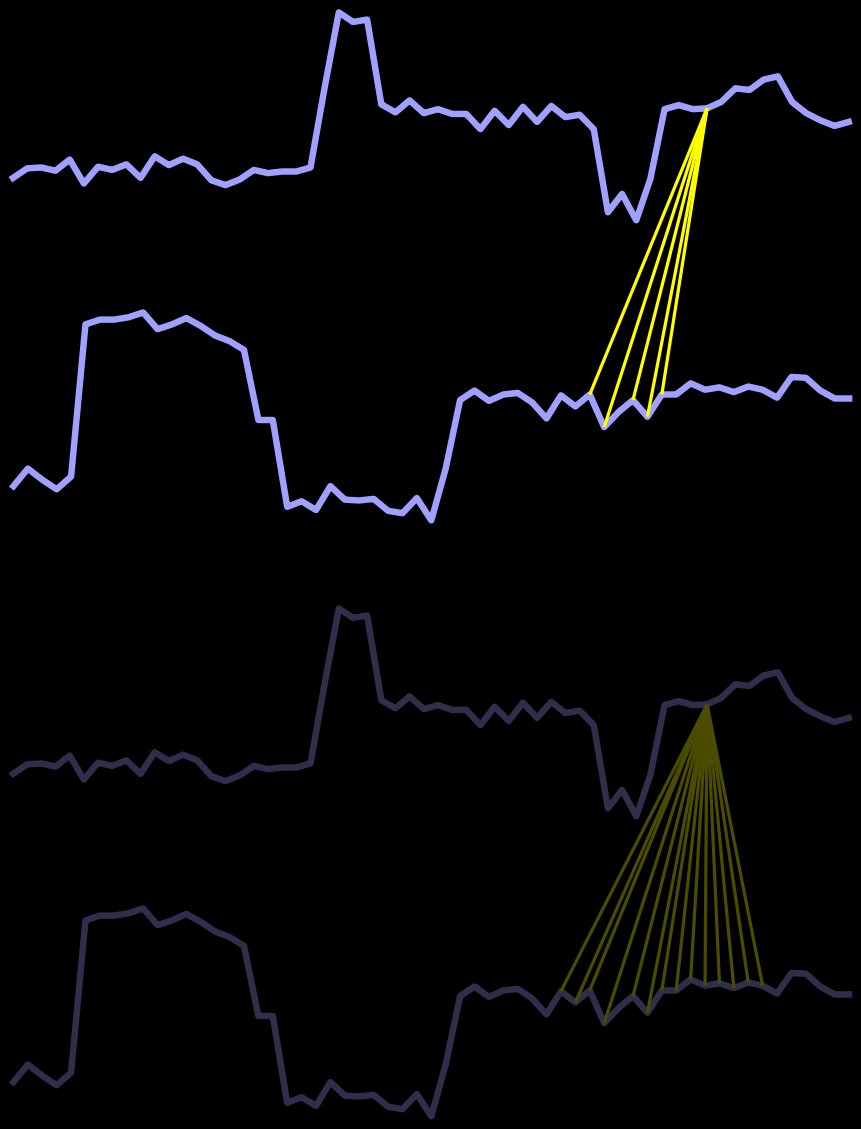
Otázka 1

- $s_1 s_2 \ldots s_n$
- $t_1 t_2 \ldots t_m$
- $\begin{matrix} s_{i_1} & s_{i_2} & \ldots & s_{i_l} \\ t_{j_1} & t_{j_2} & \ldots & t_{j_l} \end{matrix}$
- $i_k \leq i_{k+1} \leq i_k + 1$
- $l \geq m, n$
- $\frac{b(n+m)}{2} - \sum_{k=1}^{l} (s_{i_k} - t_{i_k})^2$
- $b =$ bonus

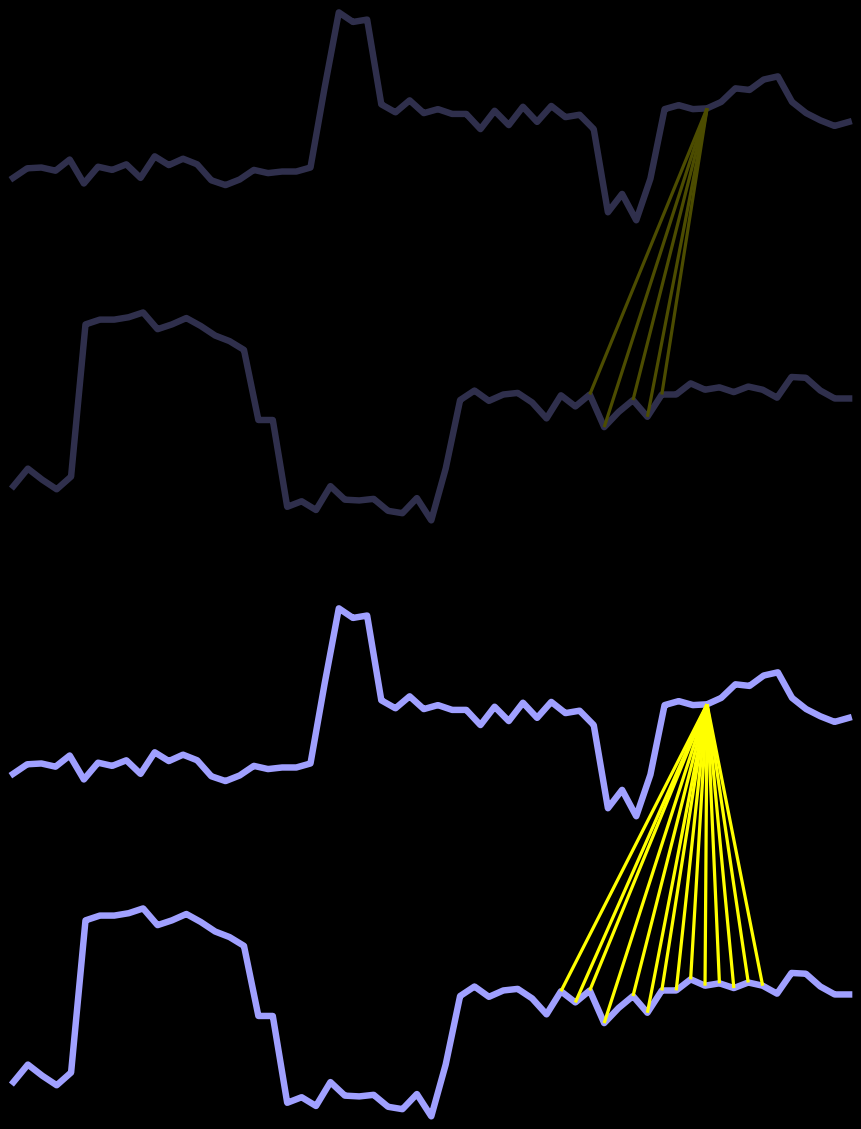
Otázka 1

- Akú funkciu optimalizuje algoritmus z kapitoly 3?
- Dala by sa upraviť tak, aby bolo možné použiť efektivnejší algoritmus?

Otázka 1



Otázka 1



Otázka 2

- Alternatívny prístup z konca časti 4.2.5 s použitím preprocessovania a jednoduchého modelu nebol až tak horší ako plný model. Vedeli by ste si predstaviť presun ďalších častí plného modelu do preprocessingu? Ak áno, tak ktoré?
- Dal by sa tiež použiť algoritmus z kapitoly 3?

Otázka 2

- Alternatívny prístup z konca časti 4.2.5 s použitím preprocessovania a jednoduchého modelu nebol až tak horší ako plný model. Vedeli by ste si predstaviť presun ďalších častí plného modelu do preprocessingu? Ak áno, tak ktoré?
- Dal by sa tiež použiť algoritmus z kapitoly 3?