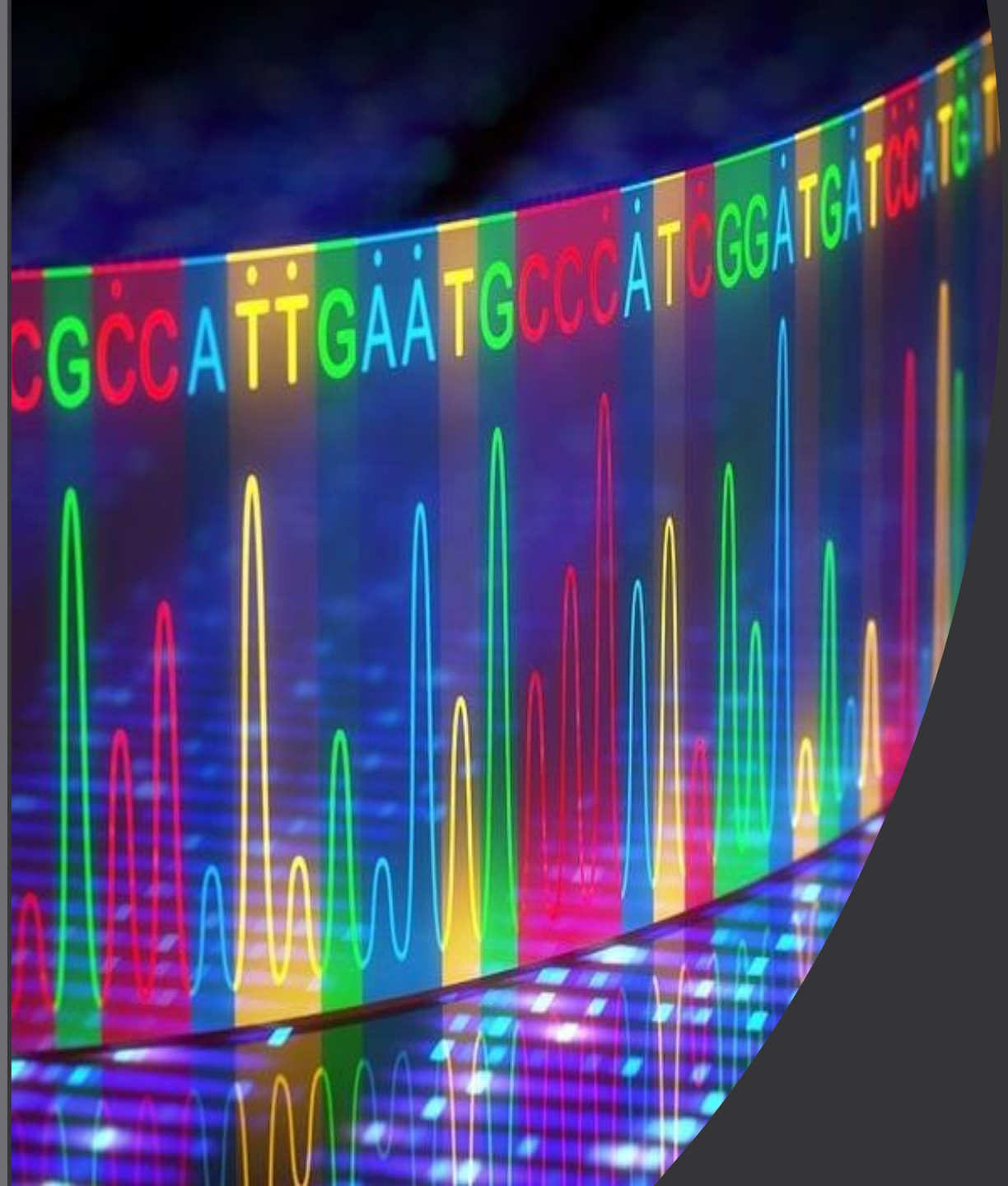


Restrikční mapy & Transkripční motivy

Programování v bioinformatice

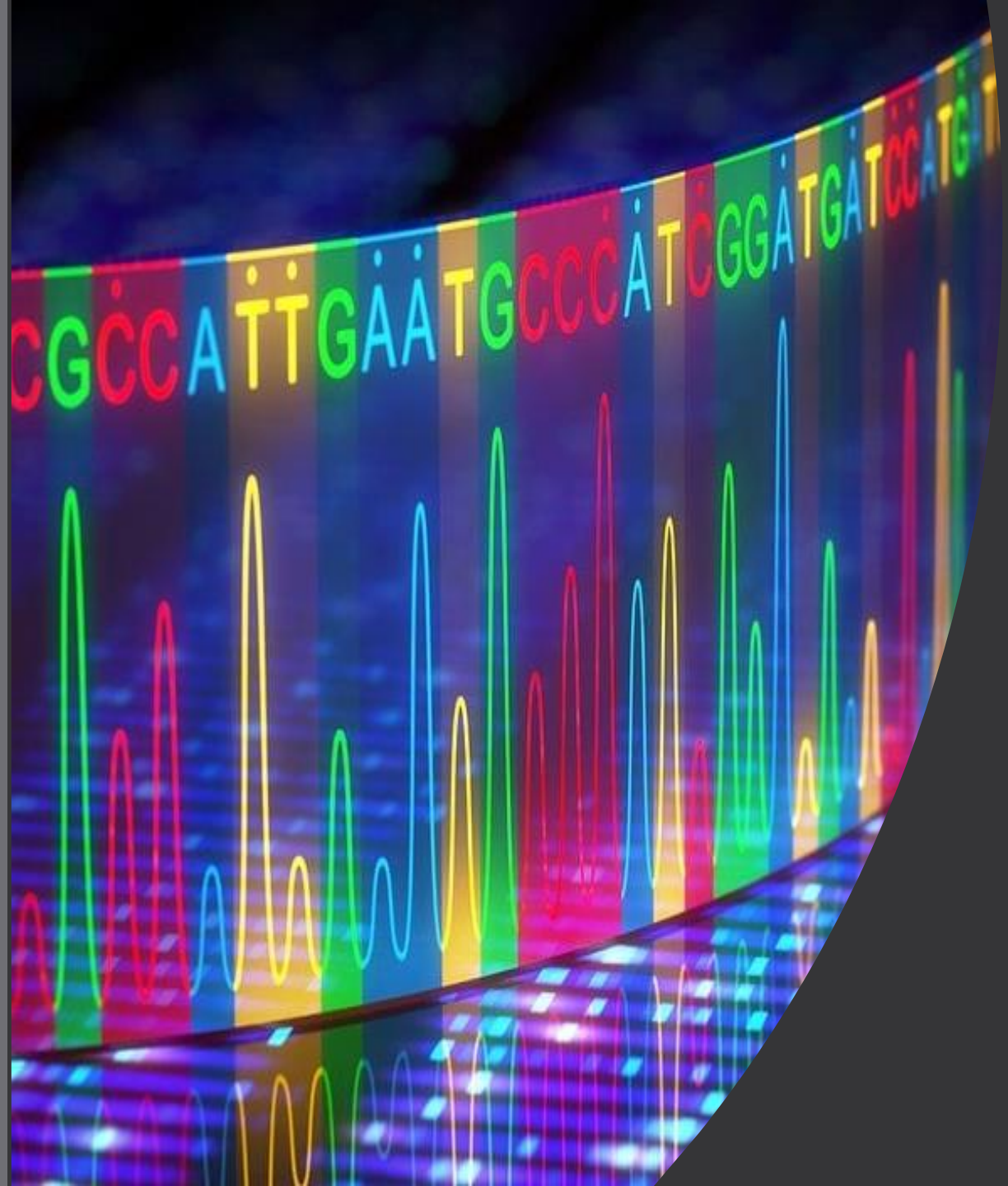
MPC – PRG 2021/2022

Vyučující:
Ing. Kateřina Jurečková (garant)
Ing. et Ing. Jana Schwarzerová, MSc

Opakování – příprava na TEST (SOUHRN)

1. & 2. TÝDEN	Typy Algoritmů Regulární výraz Vývojové diagramy Výpočetní náročnost
3. TÝDEN	Rekurze a iterace
4. TÝDEN	Dynamické programování Hirschbergův algoritmus
5. TÝDEN	Restrikční a transkripční motivy Restrikční mapy

Půlsemestrální test slouží k ověření znalostí a pochopení učiva z první poloviny semestru. Bude složen, jak z přednášek tak ze cvičení. Můžete se těšit na teoretické otázky, ale i na příklady!
Za půlsemestrální test můžete získat max. 30 bodů.



Restrikční mapy & Transkripční motivy

Programování v bioinformatice

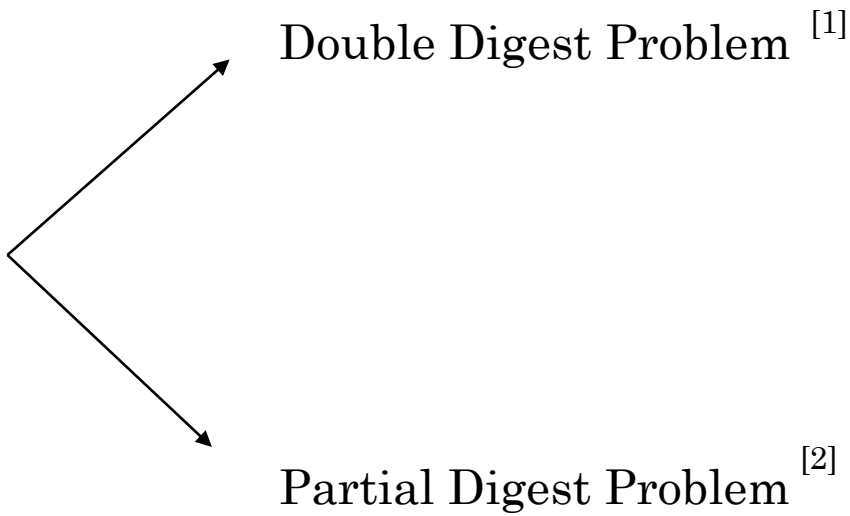
MPC – PRG 2021/2022

Vyučující:
Ing. Kateřina Jurečková (garant)
Ing. et Ing. Jana Schwarzerová, MSc

Restrikční mapy – TEORIE

- Restrikční mapování = postup, podle kterého probíhá sestavování restrikční mapy genomu.
- Restrikční mapa = forma fyzikální mapy DNA, schematicky znázorňující polohy restrikčních míst na její molekule. Vzdálenosti mezi jednotlivými místy se udávají v počtech nukleotidů
- Restrikční místo = místo na sekvenci, ve které probíhá štěpení dvouřetězcové DNA katalyzované restriktázou
- Význam restrikčního mapování:
  - základní krok při charakterizaci DNA
  - výchozí bod pro sekvenovací techniky a genově inženýrské postupy
  - metoda genetické analýzy, která dovoluje zaznamenat různé změny v DNA, neboť restrikční místo může hrát důležitou úlohu genetické značky

Praktická část cvičení



[1] SUR-KOLAY, Susmita, et al. The double digest problem: Finding all solutions. *International journal of bioinformatics research and applications*, 2009, 5.5: 570-592.
[2] SKIENA, Steven S.; SUNDARAM, Gopalakrishnan. A partial digest approach to restriction site mapping. *Bulletin of Mathematical Biology*, 1994, 56.2: 275-294.

Double Digest Problem

- Slouží k fyzickému mapování DNA
- Rozděluje dlouhé struktury DNA pomocí dvou enzymů ve třech experimentech:
  1. Enzym A
  2. Enzym B
  3. Enzym A a B dohromady
- Výstupem experimentu jsou 3 fragmenty, o určité délce, která lze identifikovat pomocí elektroforézy

Double Digest Problem

- Postup:

1) Zadané fragmenty
2) Uspořádání fragmentů
3) Vytvořím mapu pozic
4) Sloučím pozici
5) Postupná diference
6) Setřídění porovnání s X_{AB}
7) Reverze řešení

Příklad viz pracovní list

Partial Digest Problem

- Slouží k fyzickému mapování DNA pomocí jednoho restrikčního enzymu při různých reakčních časech

  - Postup:

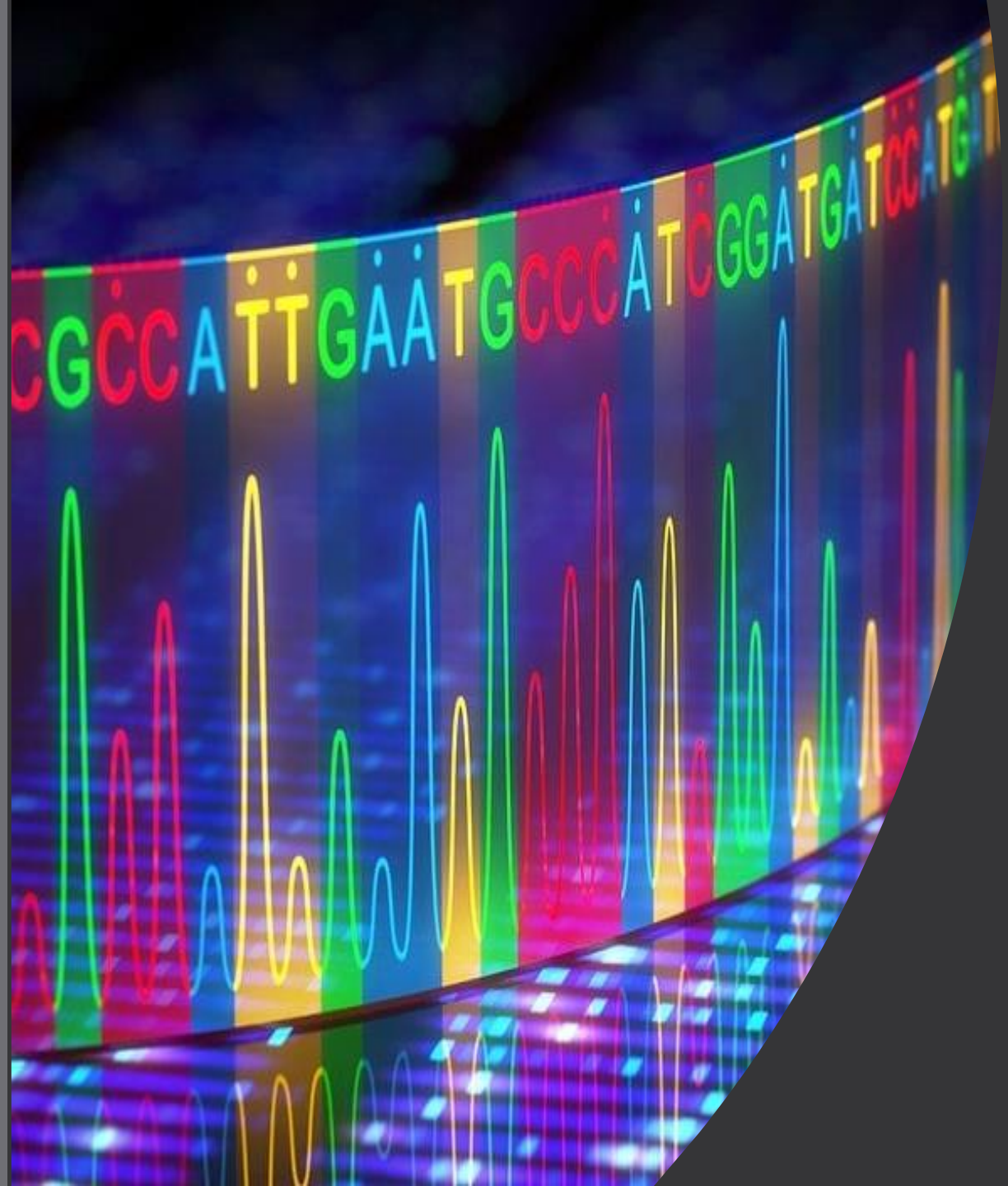
1) Zadaný vektor délek fragmentů ΔX
2) Výpočet počtu prvků v X
3) Vyhledání maxima ΔX
4) Odstranění prvku
5) Opakování bodu 3 a 4
6) Kontrola správnosti: všechny diference, setřídění = ΔX

Příklad viz pracovní list

Souhrn úkolů

1. **Pracovní list**
2. **Programování:**
    1) **Úkol:** V R naprogramujte funkci pro brute-force algoritmus DDP pro jedno možné uspořádání fragmentů. Následně upravte pro všechny možné uspořádání fragmentů.
    2) **Úkol:** V R implementujte rekurzivní algoritmus pro PDP (Partial Digest Problem) podle následujícího pseudokódu:

Přestávka



Restrikční mapy & Transkripční motivy

Programování v bioinformatice

MPC – PRG 2021/2022

Vyučující:
Ing. Kateřina Jurečková (garant)
Ing. et Ing. Jana Schwarzerová, MSc

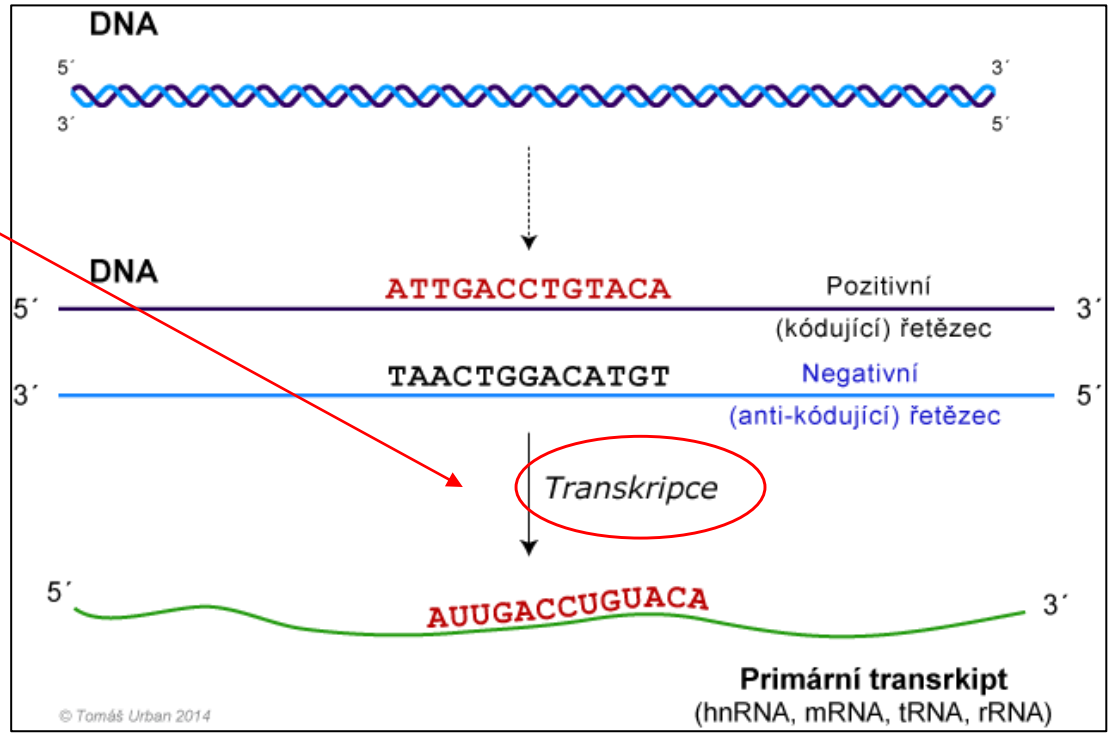
Opakování

- **Potřebná teorie – viz PŘEDNÁŠKA 5 TÝDEN**

Transkripce
– přepis DNA do RNA

Transkripční faktor
– specifický protein
– navazuje se na transkripční motiv
– zahájení a regulace transkripce

Transkripční motiv
– krátký úsek DNA (5-20 nukleotidů)
– výskyt možný v obou vláknech DNA
– časté opakování v rámci genu



Opakování

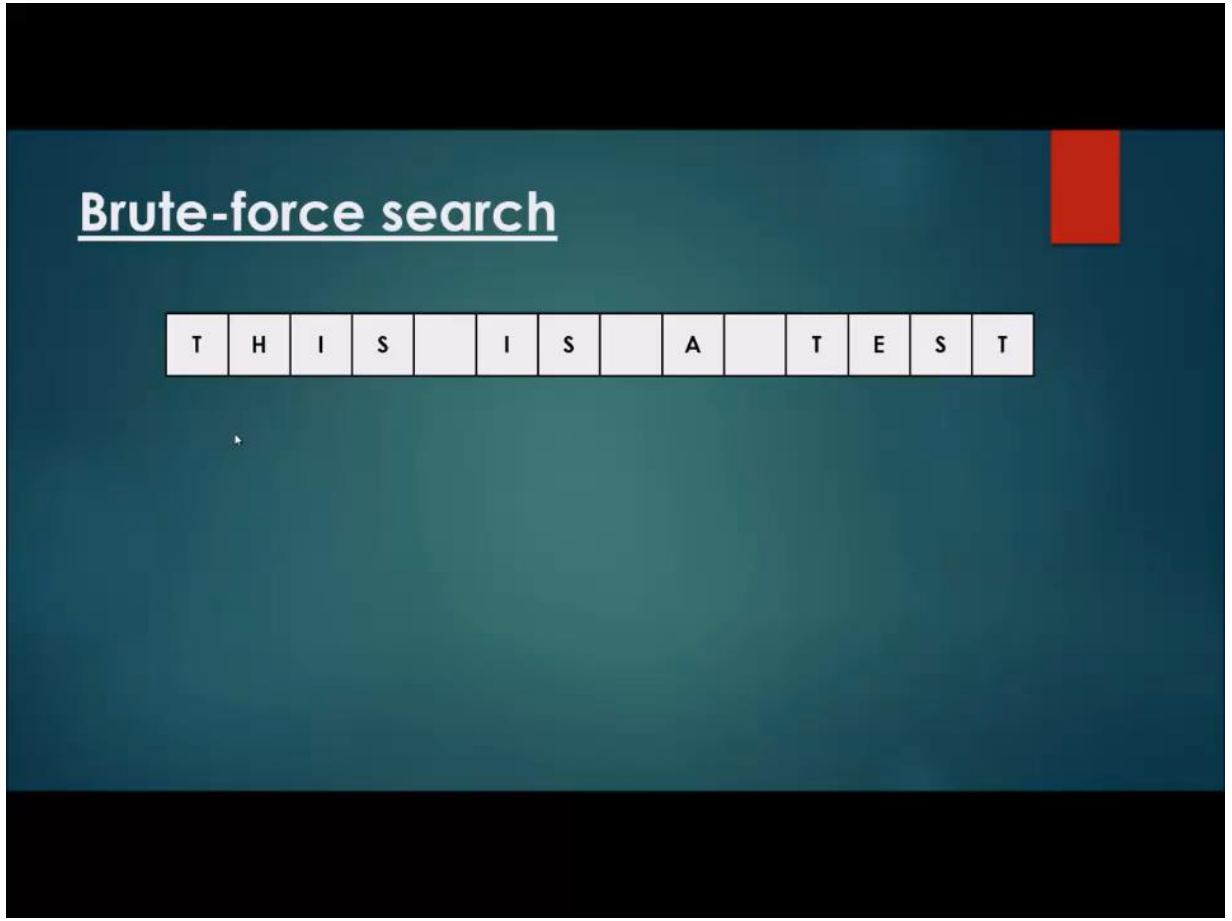
- **Potřebná teorie – viz PŘEDNÁŠKA 5 TÝDEN**

Dvě skupiny algoritmů:

1) Znakové (word-based, string-based),
– založené na kompletní enumeraci oligonukleotidových četností výskytu
– často exhaustive search algoritmy
=> globálně optimální výsledek

2) Pravděpodobnostní – využívají modely parametrů
– princip maximální věrohodnosti (maximum-likelihood), Bayesovskou podmíněnou pravděpodobnost. Lze využít i strojové učení (genetické algoritmy, neuronové sítě).

Opakování – Brute-force search



Brute force Motif Search

1. Funkce **Score**
2. Funkce **NextLeaf**
3. Funkce **BFMotifSearch**
4. Funkce **NextVertex**
5. Funkce **SimpleMotifSearch**

Kniha 1, kapitola 4

Přednáška – Restrikční a transkripční motivy

1. Score

Score = function (Sek, s, L)

- Sek = soubor sekvencí DNA (*např. Zkušební sekvence pro Skore*)
- s = vektor počátečních pozic motivů
- L = délka motivu
- Výstup: bestScore, blok

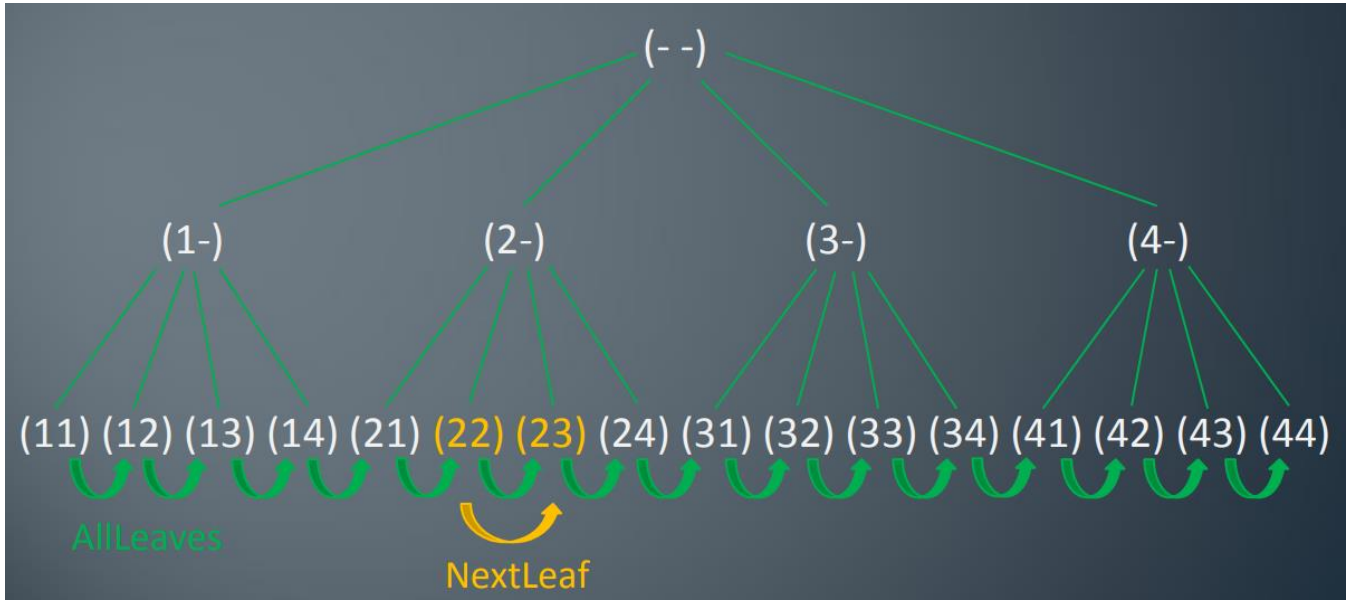
- Pro libovolný vektor *s* z *DNA* se počítá frekvenční profil a celkové skóre vektoru je suma maximálních frekvencí ve sloupcích profilu.

$$Score(s, DNA) = \sum_{i=1}^{l} \max_{k \in \{A,C,G,T\}} count(k, i)$$

```
A T C C G T A
G T G C A T A
A A G C G T A
A T G C G T G

3 1 0 0 1 0 3
0 0 1 4 0 0 0
1 0 3 0 3 0 1
0 3 0 0 0 4 0

konsenzus   A T G C G T A

skóre       3+3+3+4+3+4+3 = 23
```

2. NextLeaf

```
NextLeaf(a, L, k)
1  for i ← L to 1
2      if a_i < k
3          a_i ← a_i + 1
4          return a
5      a_i ← 1
6  return a
```



a = rep (1, L)
L = počet sekvencí
k = n – l + 1
n = délka sekvence
l = délka motivu

3. BFMotifSearch

BFMotifSearch (Sek, t, n, l)

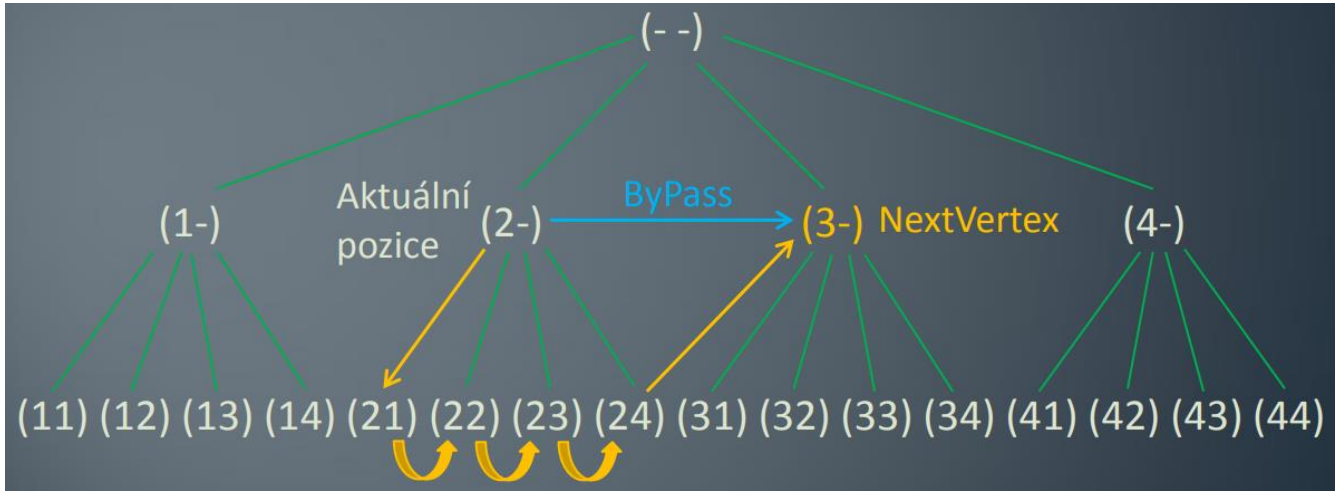
```
1  s ← (1, 1, . . . , 1)
2  bestScore ← Score(s, DNA)
3  while forever
4      s ← NextLeaf(s, t, n − l + 1)
5      if Score(s, DNA) > bestScore
6          bestScore ← Score(s, DNA)
7          bestMotif ← (s_1, s_2, . . . , s_t)
8      if s = (1, 1, . . . , 1)
9          return bestMotif
```

Sek = soubor sekvencí DNA (*např. Zkušební sekvence pro BFMotifSearch*)
l = délka motivu
t = počet sekvencí
n = délka jedné sekvence

4. NextVertex

```
NextVertex(a, i, L, k)
1  if i < L
2        a_{i+1} ← 1
3        return (a, i + 1)
4  else
5        for j ← L to 1
6              if a_j < k
7                    a_j ← a_j + 1
8                    return (a, j)
9  return (a, 0)
```

–> Přednáška 3 – slajd 55 až 58



5. SimpleMotifSearch

```
SIMPLEMOTIFSEARCH(DNA, t, n, l)
 1  s ← (1, ..., 1)
 2  bestScore ← 0
 3  i ← 1
 4  while i > 0
 5      if i < t
 6          (s, i) ← NEXTVERTEX(s, i, t, n − l + 1)
 7      else
 8          if Score(s, DNA) > bestScore
 9              bestScore ← Score(s, DNA)
10              bestMotif ← (s_1, s_2, ..., s_t)
11          (s, i) ← NEXTVERTEX(s, i, t, n − l + 1)
12  return bestMotif
```

Souhrn úkolů – implementujte Brute force Motif Search

Brute force Motif Search

1. Funkce **Score**
2. Funkce **NextLeaf**
3. Funkce **BFMotifSearch**
4. Funkce **NextVertex**
5. Funkce **SimpleMotifSearch**

Kniha 1, kapitola 4

Přednáška – Restrikční a transkripční motivy

Jste u konce! Děkuji za Váš čas!

Jana Schwarzerová
xschwa16@stud.feec.vutbr.cz

Reference & Další studijní materiály pro doplnění

- SUR-KOLAY, Susmita, et al. The double digest problem: Finding all solutions. *International journal of bioinformatics research and applications*, 2009, 5.5: 570-592.
- SKIENA, Steven S.; SUNDARAM, Gopalakrishnan. A partial digest approach to restriction site mapping. *Bulletin of Mathematical Biology*, 1994, 56.2: 275-294.
- https://is.muni.cz/el/1431/jaro2009/Bi4035/um/3B_Konstrukce_restrikcnich_map.pdf
- https://ocw.mit.edu/courses/mathematics/18-417-introduction-to-computational-molecular-biology-fall-2004/lecture-notes/lecture_03.pdf