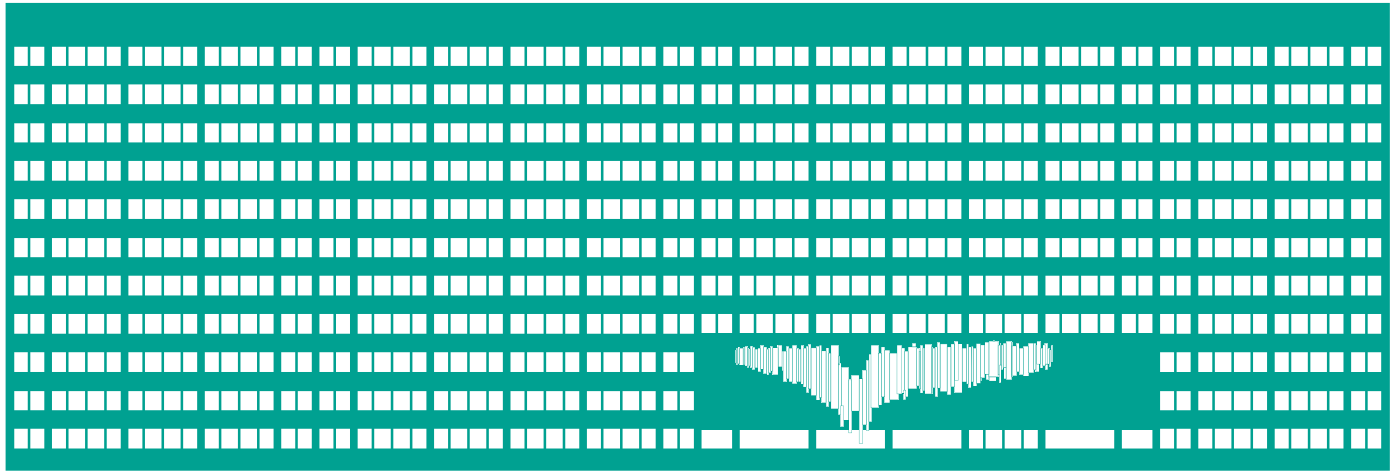
VŠB TECHNICKÁ UNIVERZITA OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA



www.vsb.cz

Algoritmy pro Bioinformatiku

Podobnost sekvencí

Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA443

michal.vasinek@vsb.cz

26. září 2023

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY

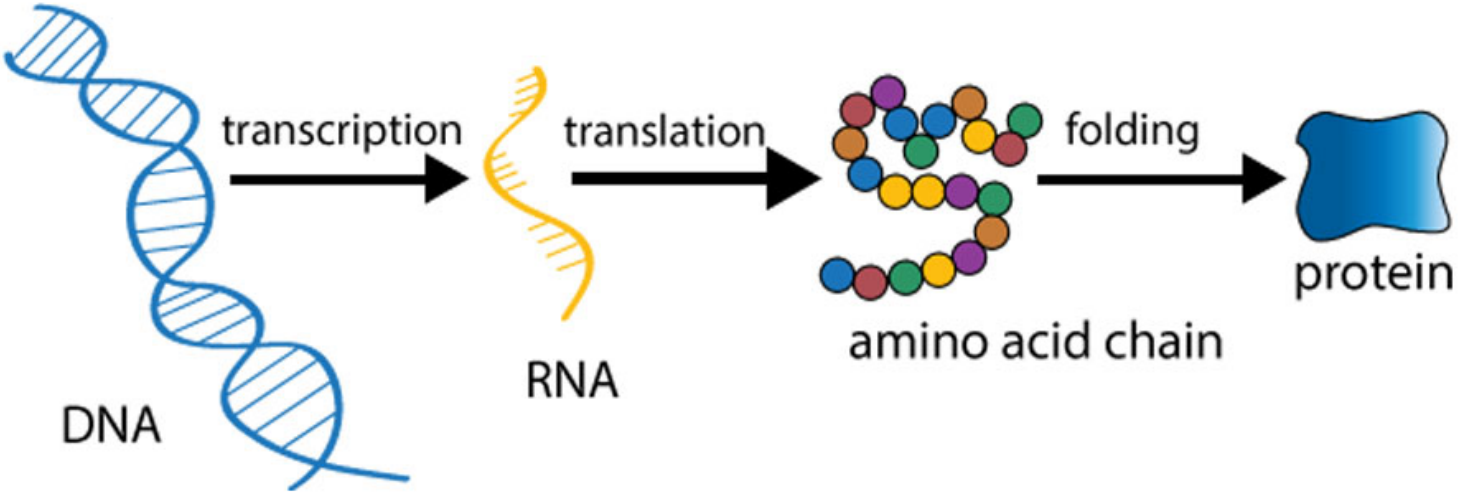

10011110010101111010101
ACCGTTACGGA CTTA
CATGGGTAGGGAGGGG
01101100110100111011

Přehled přednášky

- Biotechnologické nástroje:
  - Restrikční enzymy
  - Metoda shotgun
  - PCR reakce
- Měření podobnosti sekvencí:
  - Hammingova vzdálenost
  - Editační vzdálenost

Centrální Dogma

- Centrální Dogma nám říká, jakým způsobem obdržíme protein z konkrétního genu. Tomuto procesu se říká exprese genu.
- Exprese genu se skládá ze dvou kroků:
  - Transkripce: DNA $\rightarrow$ mRNA
  - Translace: mRNA $\rightarrow$ Protein
  - Následně dochází k po-translačním úpravám.
- Každá aminokyselina v proteinu je zakódována trojicí nukleotidů, tzv. codonem.

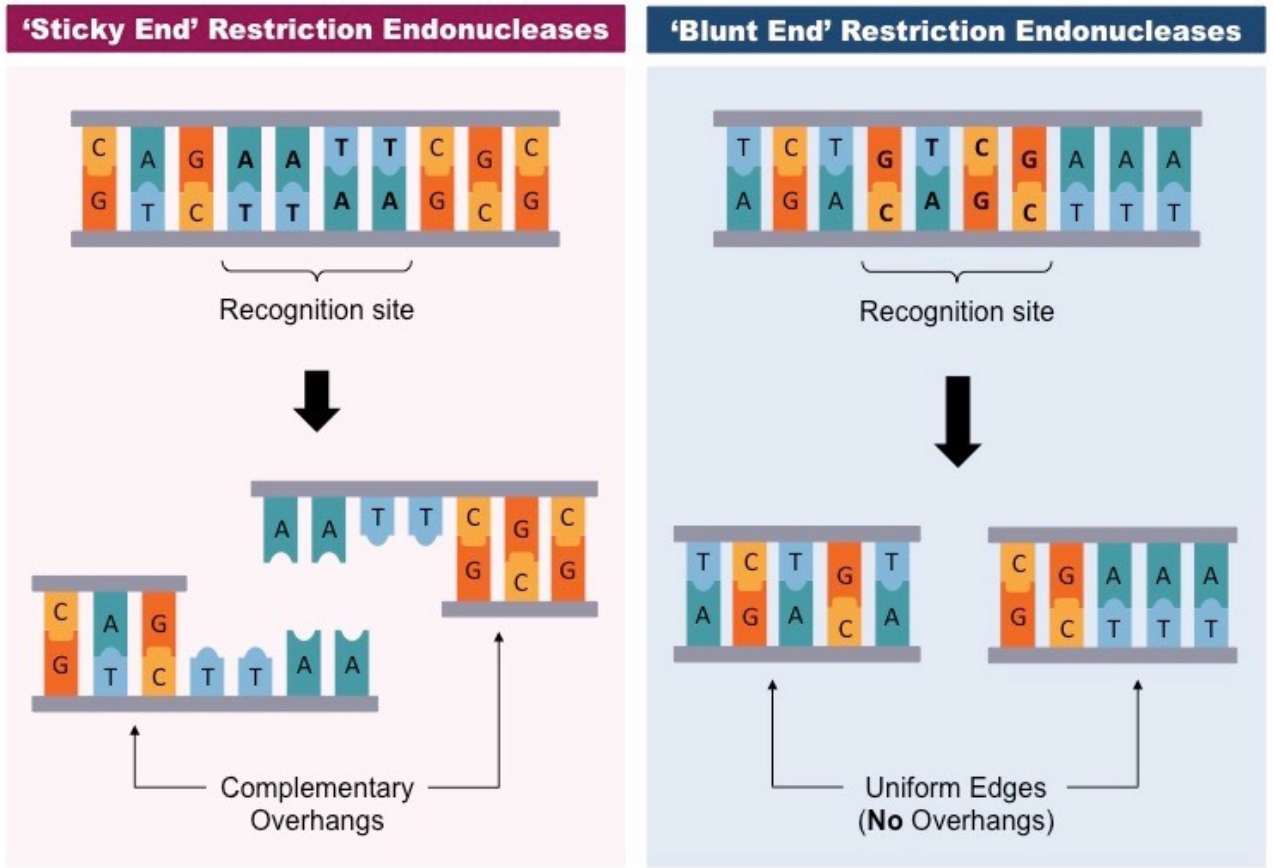


Biotechnologické nástroje

- Fragmentace DNA
  - Restrikční enzymy
  - Shotgun metoda
- Kopírování DNA
  - Klonování
  - PCR - Polymerase Chain Reaction
- Měření délky DNA
  - Gelová elektroforéza

Restrikční enzymy

- Restrikční enzym rozpoznává určitá místa v DNA, kde je schopen se navázat a DNA v tomto místě rozštěpit [1].



[1] https://www.youtube.com/watch?v=Ik_Pxht1LM0

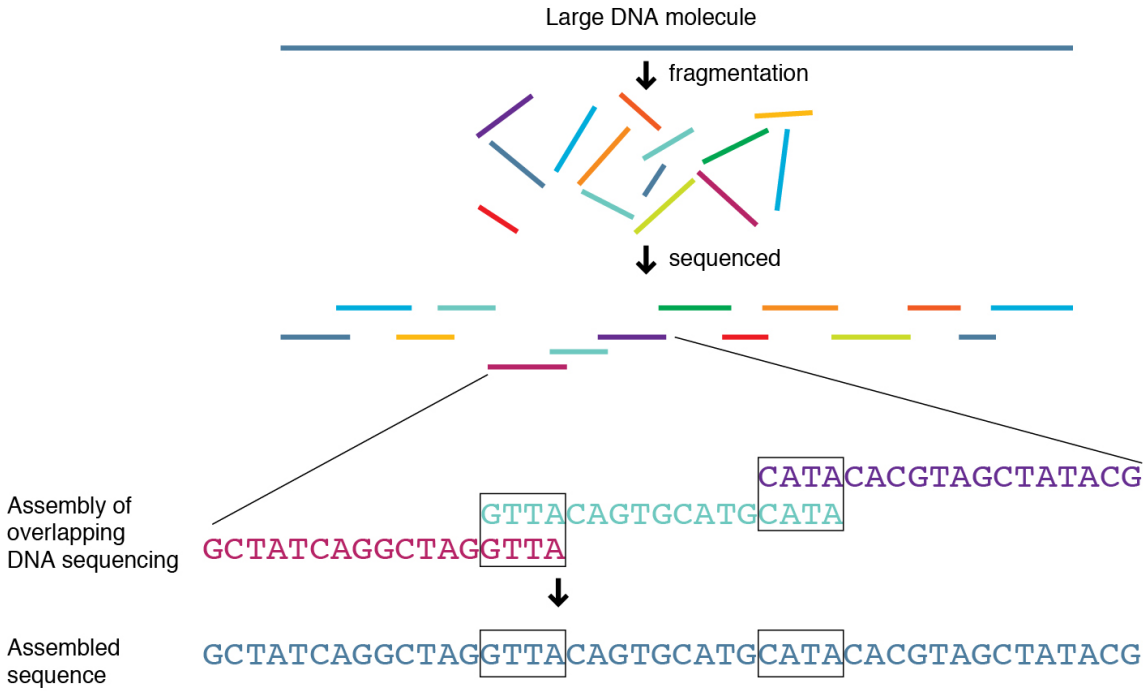
Restrikční enzymy

- Objeveny v roce 1968.
- Výskyt u bakterií a archeabakterií.
- Přirozeně se restrikční enzymy používají k přestřížení cizí DNA => předcházení infekcí.
- EcoRI přestřihává DNA v místech, kde se vyskytuje sekvence GAATTC. Je svým vlastním reverzně komplementárním palindromem.
- V současnosti známe přes 3000 restrikčních enzymů.
- Využití: optické mapování.[2]

[2] https://www.youtube.com/watch?v=S2ng6glu04I

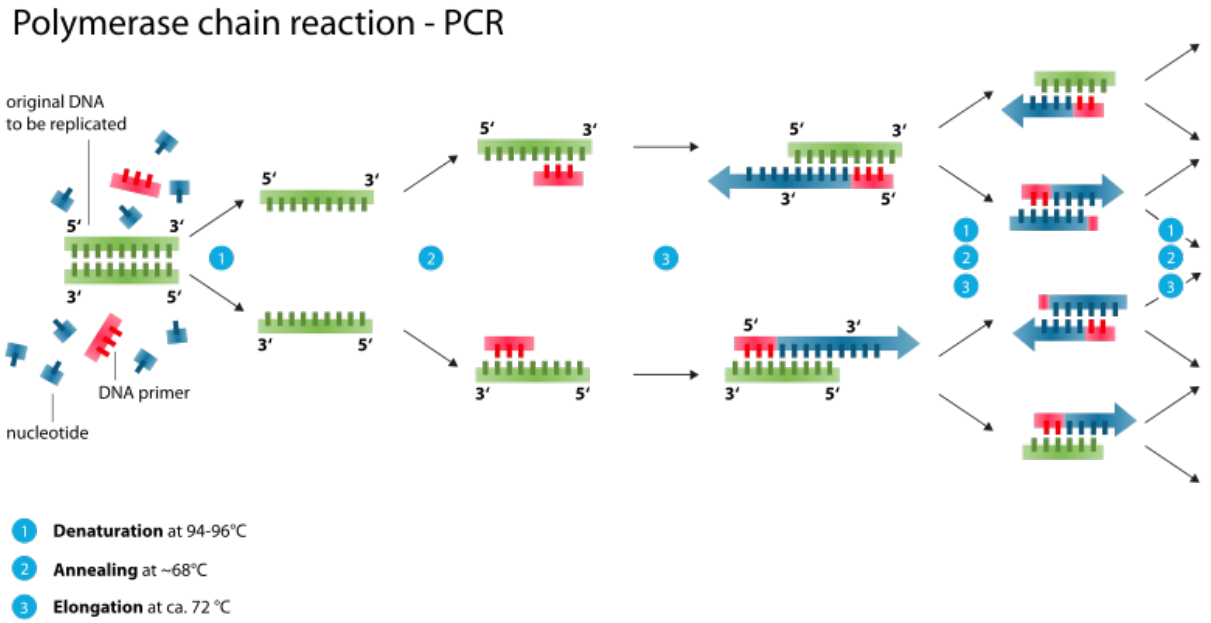
Shotgun Metoda

- Náhodná fragmentace DNA na krátké úseky.
- Celý proces proveden několikrát pro zajištění překrývajících se úseků.
- Metoda se používá pro sekvenování celého genomu.
- Krátké úseky pospojovány pomocí algoritmů tzv. de novo sestavení genomu.



PCR[3]

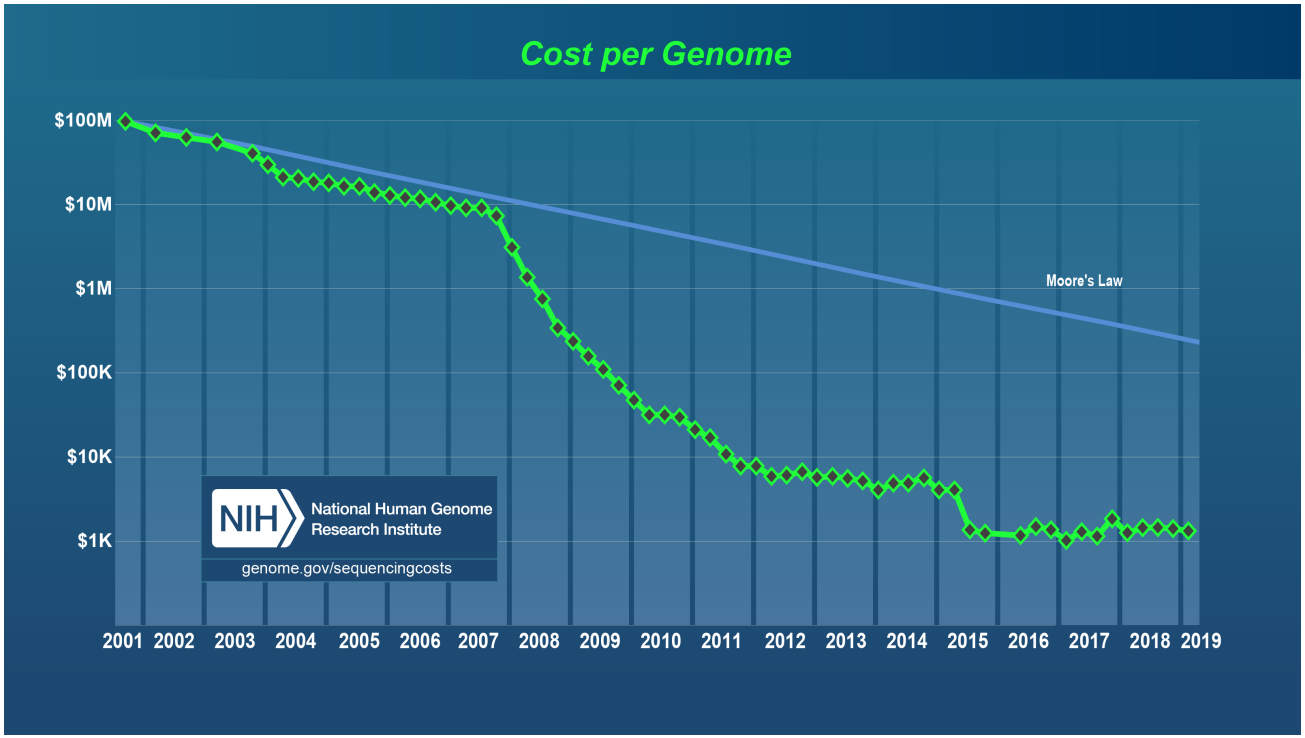
- PCR - polymerase chain reaction
- Vynalezeno v roce 1984.
- Umožňuje extrémně rychlou replikaci vybraných úseků DNA.
- Zjišťování mutací ve vybraných oblastech DNA.



[3] https://www.youtube.com/watch?v=wBrNbbAIAFo

Human Genome Project

- Lidský genom přečten a publikován 2000/2001.
- Stále zpřesňován. Poslední verze 2022, T2T-CHM13.



Měření podobnosti sekvencí

K čemu se podobnost využívá

- Určování biologických funkcí různých genů. Vycházíme z předpokladu, že dva geny mající podobnou sekvenci mají i podobnou funkci.
- Nalezení evoluční vzdálenosti mezi dvěma organismy.[4] [5]
- Sestavení genomu.
- PCR - amplikonové sekvenování - přiřazení sekvence ke správnému místu v referenční sekvenci.

[4] Dva lidé se liší přibližně v 0,6% DNA

[5] Člověk a šimpanz sdílí 99% DNA

Vzdálenost dvou řetězců

- Hammingova vzdálenost:
  - Dva řetězce X a Y stejné délky.
  - Hammingova vzdálenost je počet rozdílných symbolů.

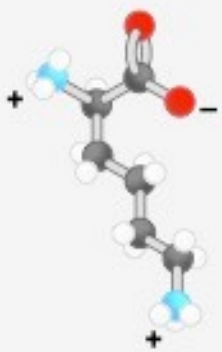
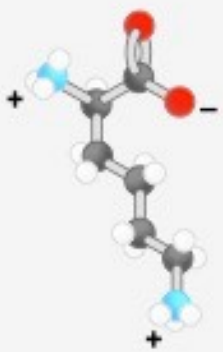
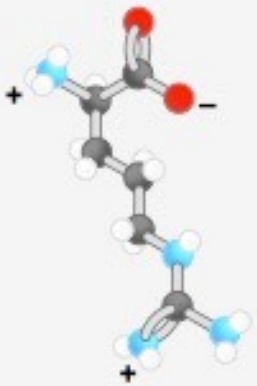
$$A = \texttt{acgtacgt}$$
$$B = \texttt{aggtcgat}$$

acgtacgt

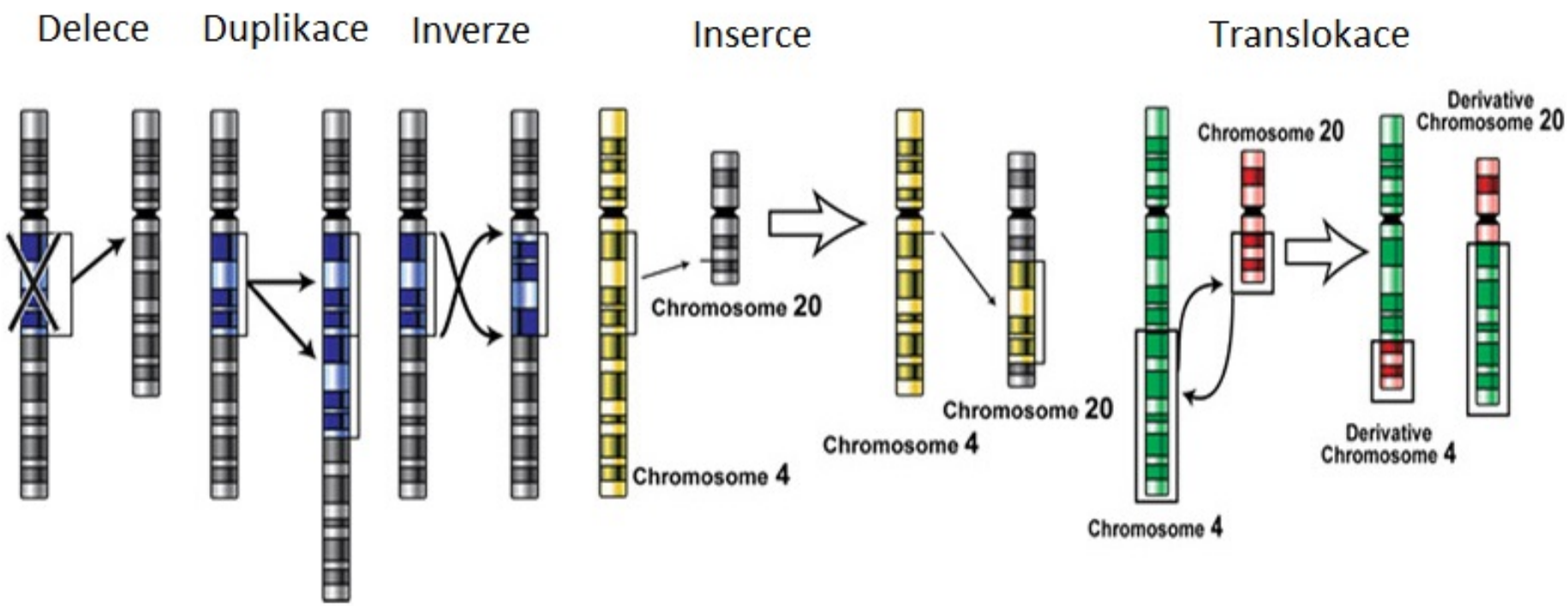
aggtcgat

  - Hammingova vzdálenost $d_H(A, B) = 4$.
- Editační vzdálenost - viz. další slidy

Typy mutací

	No mutation	Point mutations		
		Silent	Missense	Nonsense
DNA	TTC	TTT	TCC	ATC
mRNA	AAG	AAA	AGG	UAG
Protein	Lys	Lys	Arg	STOP
				

Další typy mutací



Editační vzdálenost

- Mějme dva řetězce A a B, pomoci kolika následujících operací dokážeme z A obdržet B:
  - Nahrazení písmene za jiné písmeno - R.
  - Vynechání písmene, delece - D.
  - Vložení písmene, inserce - I.
  - Shoda písmem - M.
- Editační vzdálenost $\leq$ Hammingova vzdálenost.

$$A = \texttt{acgtacgt}$$

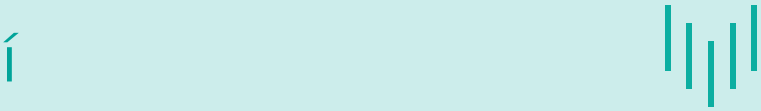
$$B = \texttt{aggtcgat}$$

acgtacg_t

aggt_cgat

editační vzdálenost $d_E(A, B) = 3$

Editační vzdálenost - ocenění operací



- Každá operace s řetězcem M, R, I, D má přířazenu svou cenu.
- Cíl : nalézt nejmenší počet operací resp. minimální editační vzdálenost.

	_	a	c	g	t
_		1	1	1	1
a	1	0	1	1	1
c	1	1	0	1	1
g	1	1	1	0	1
t	1	1	1	1	0

acgtacg_t

aggt_cgat

$$0 + 1 + 0 + 0 + 1 + 0 + 0 + 1 + 0 = 3$$

Fundamentální problémy

- Existence nerozhodnutelných posloupností operací.

$$A = acgtac$$
$$B = actgac$$

acgtac	acgt_ac	ac_gtac
actgac	ac_tgac	actg_ac

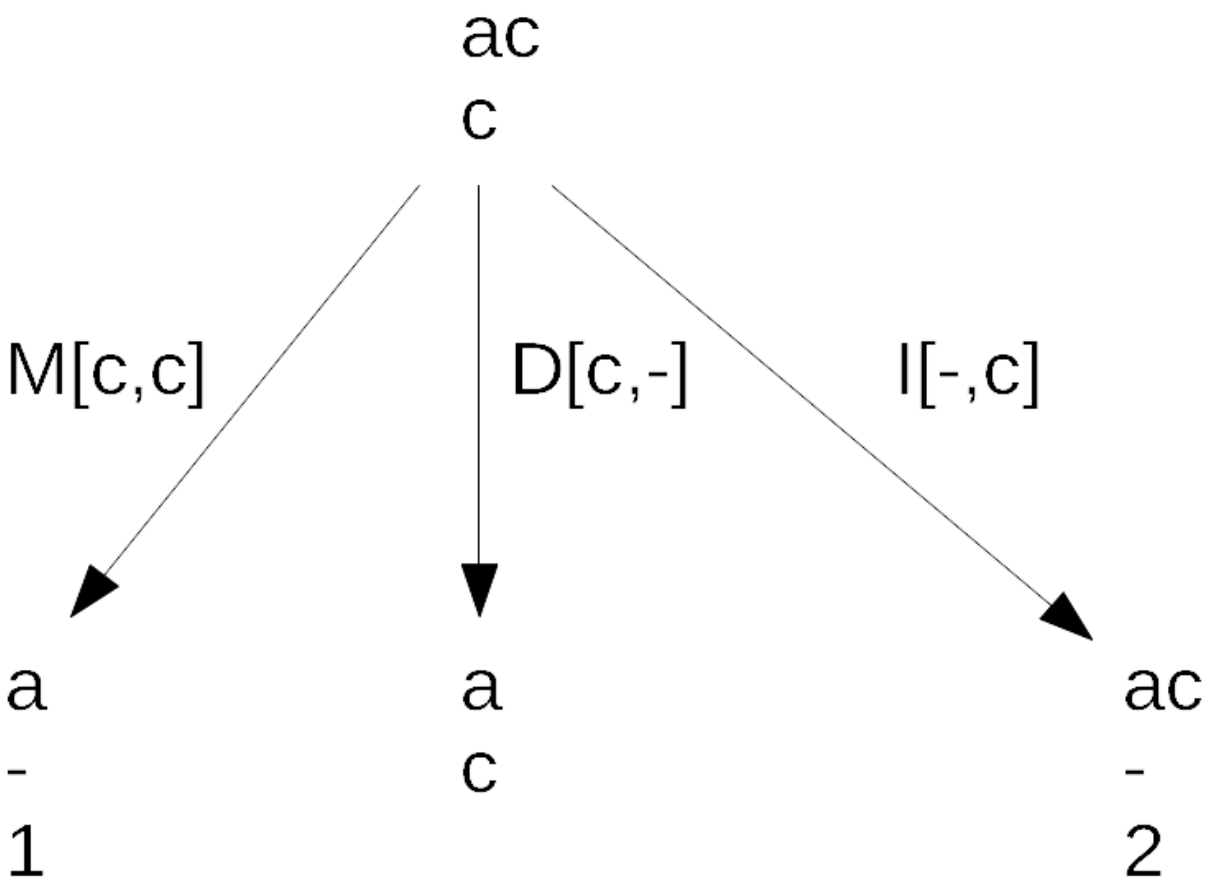
- Všechny alternativy mají stejnou editační vzdálenost.

Výpočet editační vzdálenosti

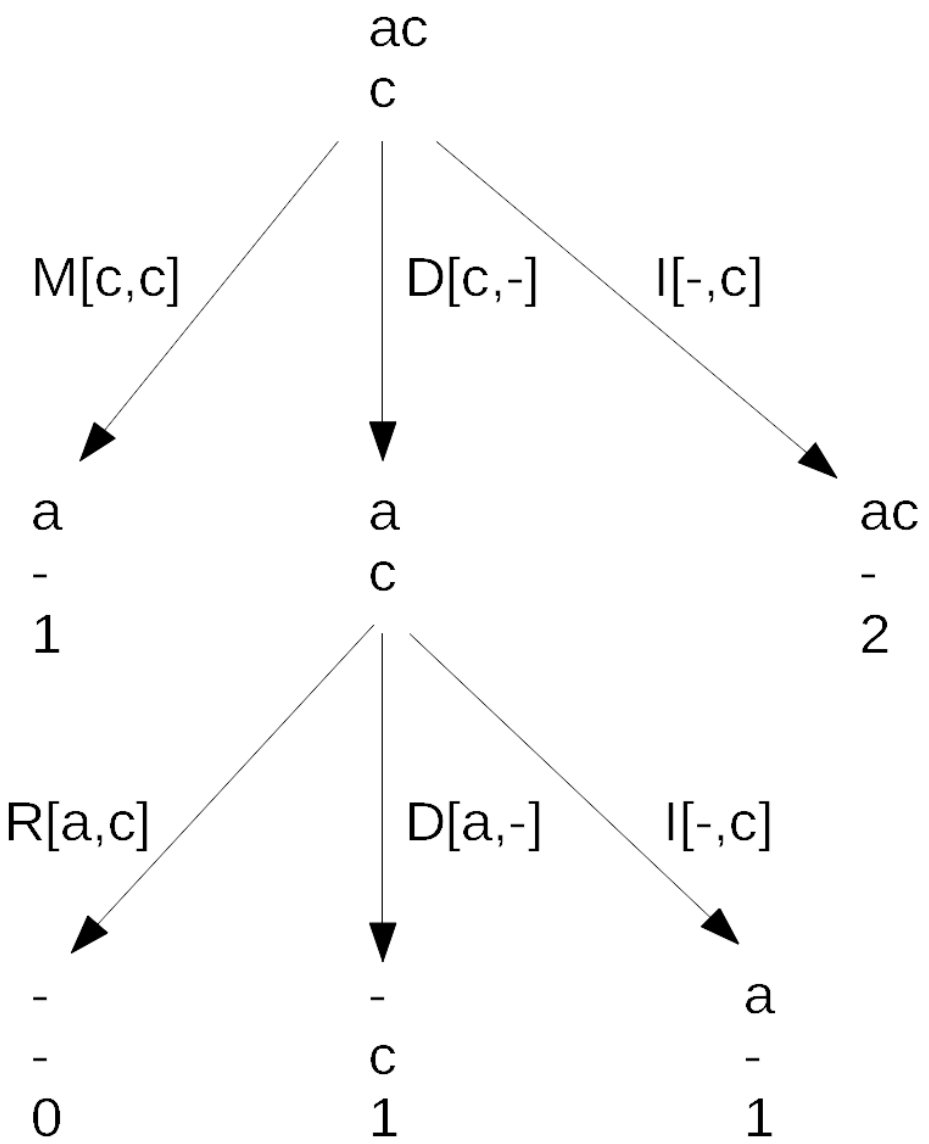
- Uvažujme editační vzdálenost mezi dvěma řetězci s_1 a s_2, nechť $|s_1| > 0$ a $|s_2| = 0$. Jaká je jejich editační vzdálenost? $d_E = |s_1|$
- Nechť i, j jsou dva prefixy řetězců s_1 a s_2.
- $d_E[i, j]$ je potom editační vzdálenost mezi dvěma prefixy. $d_E[i, 0] = i$ a $d_E[0, j] = j$.

Editační vzdálenost - rekurzivní postup výpočtu

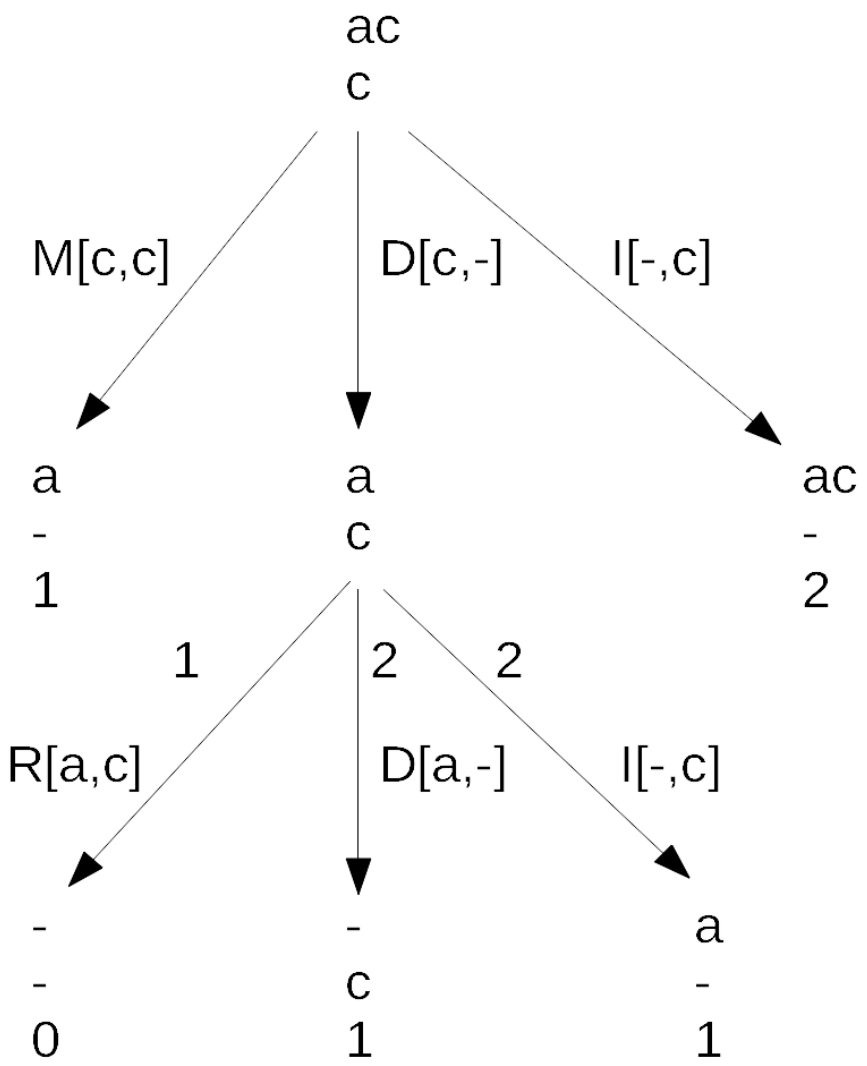
- V příkladu je rekurzivní postup aplikován od konce sekvencí. Na směru však nezáleží.
- $op[s_1[i], s_2[j]]$ - op - jedna z operací R,M,I,D, s_1, s_2 jsou porovnávané sekvence.



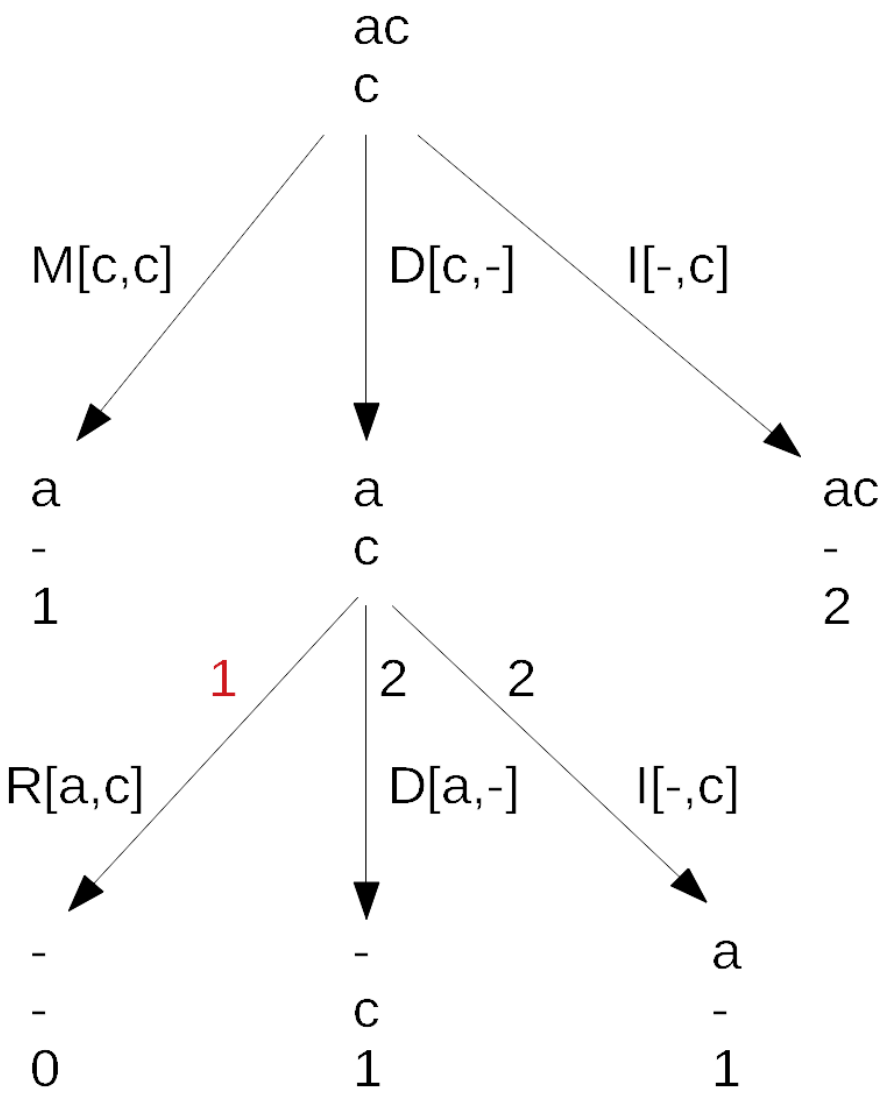
Editační vzdálenost - rekurzivní postup výpočtu



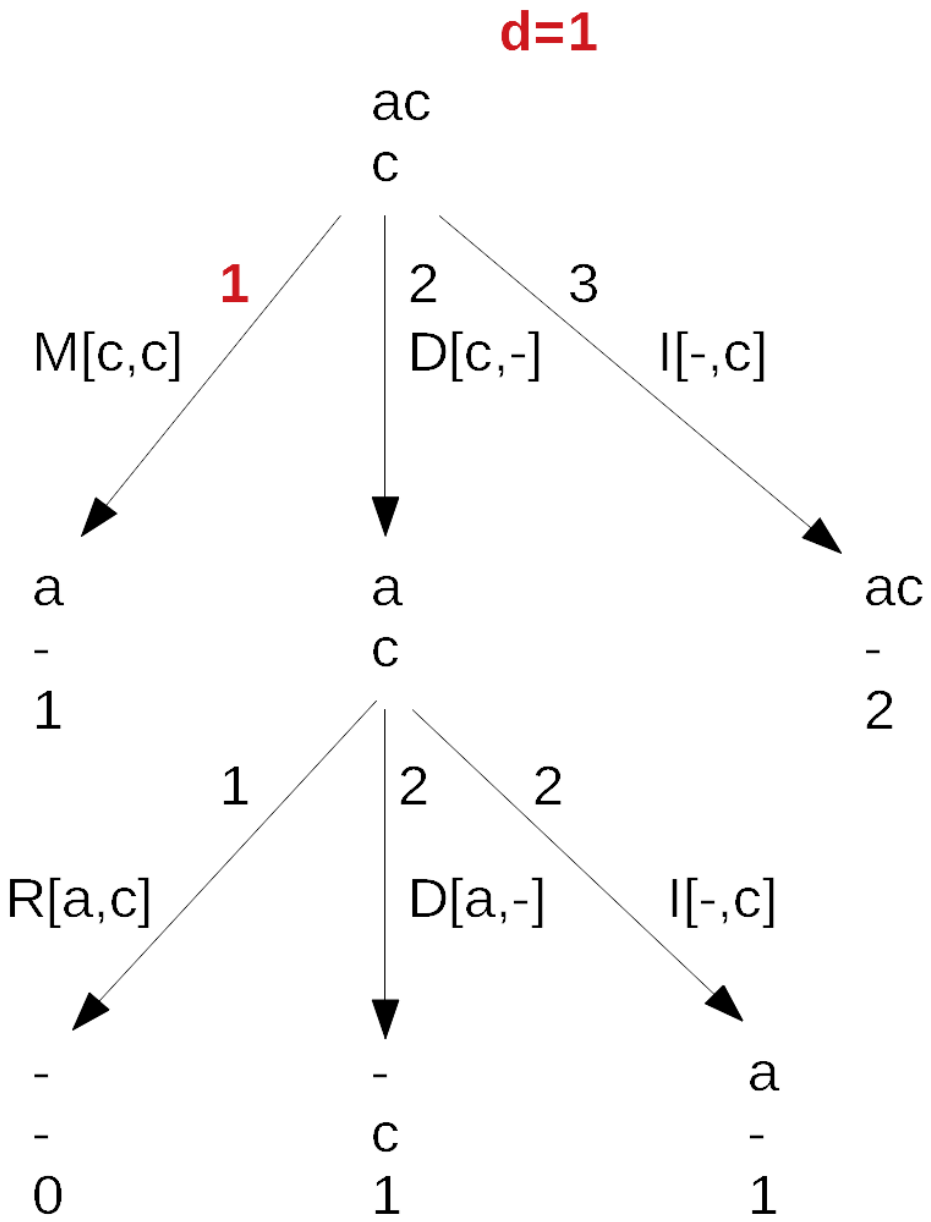
Editační vzdálenost - rekurzivní postup výpočtu



Editační vzdálenost - rekurzivní postup výpočtu



Editační vzdálenost - rekurzivní postup výpočtu



Editační vzdálenost - formalizace

- Z definice okrajových podmínek:

$$d_E[0, i] = i$$

$$d_E[j, 0] = j$$

- Předpoklad, že poslední znak v řetězcích je buďto shodou, neshodou nebo insercí:

$$d_E[i, j] = min \begin{cases} d_E[i-1, j] + 1 \\ d_E[i, j-1] + 1 \\ d_E[i-1, j-1] + \delta(s_1[i-1], s_2[j-1]) \end{cases}$$

- Funkce, která nám pomáhá rozlišit mezi shodou a neshodou:

$$\delta(a, b) = 0 \text{ pokud } a = b, \text{ jinak } \delta(a, b) = 1$$

- Rekurentní předpis => rekurzivní funkce.

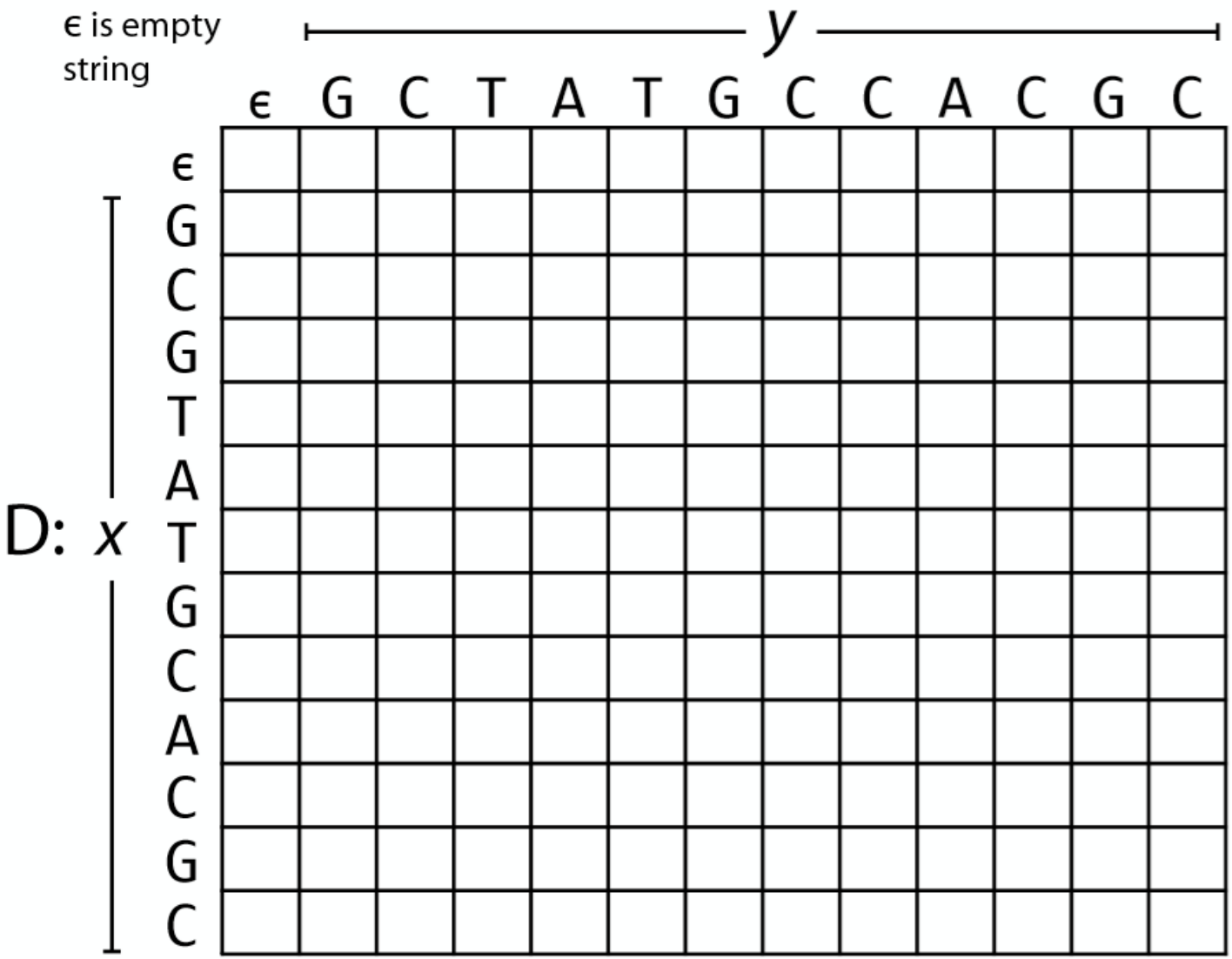
Editační vzdálenost - dynamické programování

Mějme dva řetězce:

$$y = \texttt{GCTATGCCACGC}$$
$$x = \texttt{GCGTATGCACGC}$$

- Vyplňujeme tabulku D, podle definic z předchozího slidu.
- Každá buňka tabulky obsahuje vzdálenost mezi dvěma prefixy řetězců x a y.

Editační vzdálenost - dynamické programování



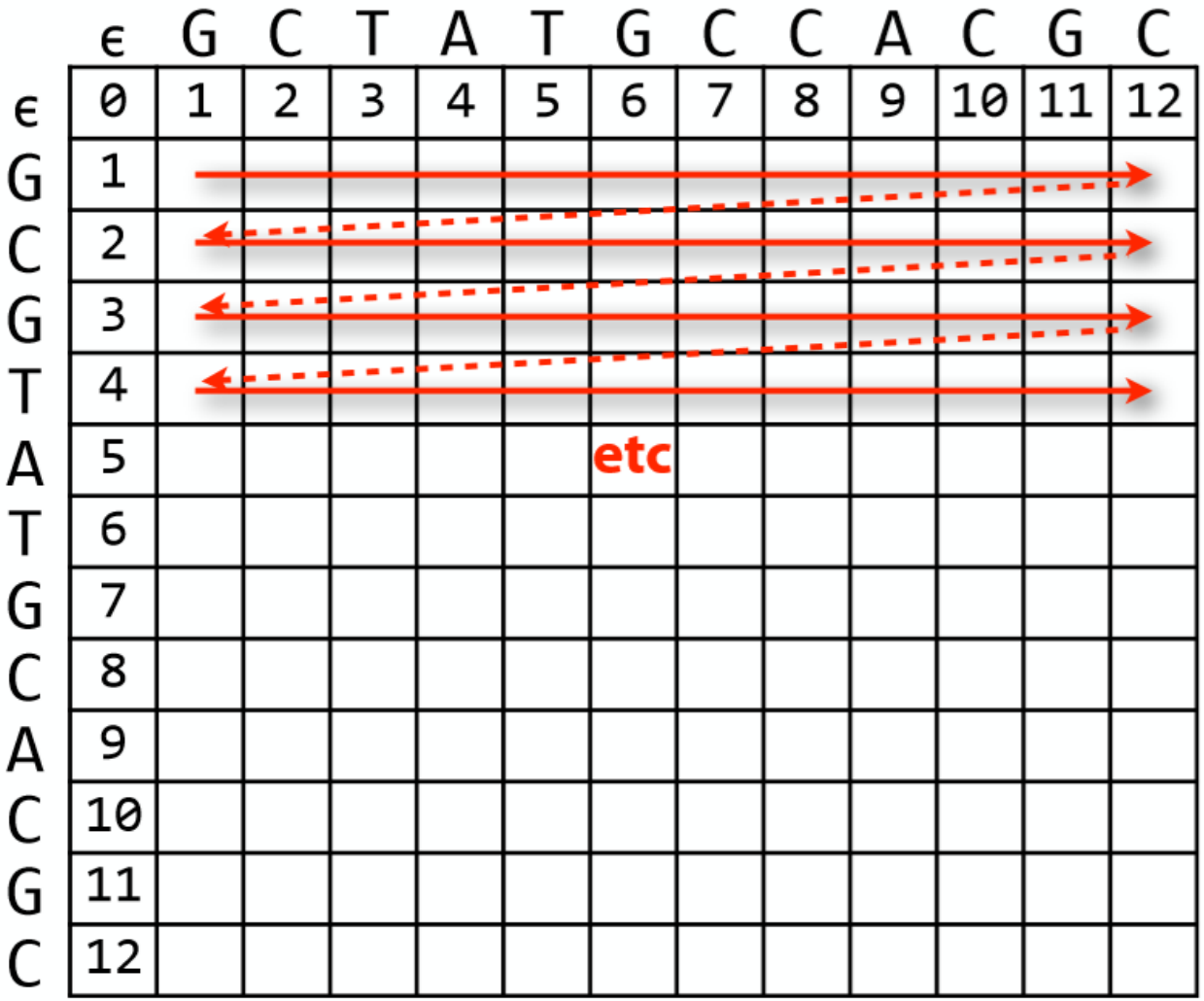
Editační vzdálenost - dynamické programování

Doplníme hodnoty $D[0, i] = i$ a $D[j, 0] = j$

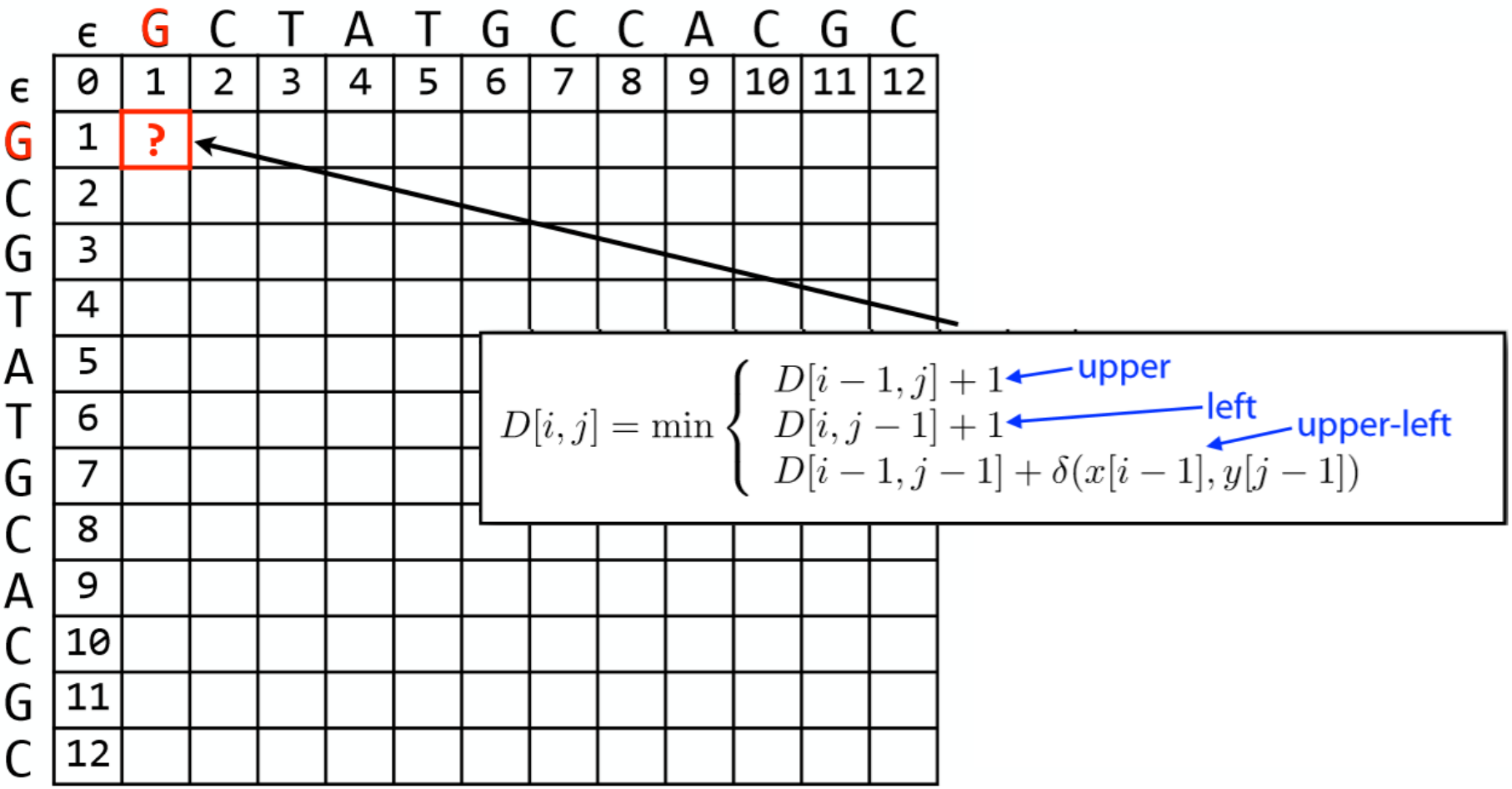
		ϵ	G	C	T	A	T	G	C	C	A	C	G	C
		0	1	2	3	4	5	6	7	8	9	10	11	12
ϵ	0													
G	1													
C	2													
G	3													
T	4													
A	5													
T	6													
G	7													
C	8													
A	9													
C	10													
G	11													
C	12													

Editační vzdálenost - dynamické programování

Dle vyznačeného směru dopočítáme do tabulky hodnoty:



Editační vzdálenost - dynamické programování



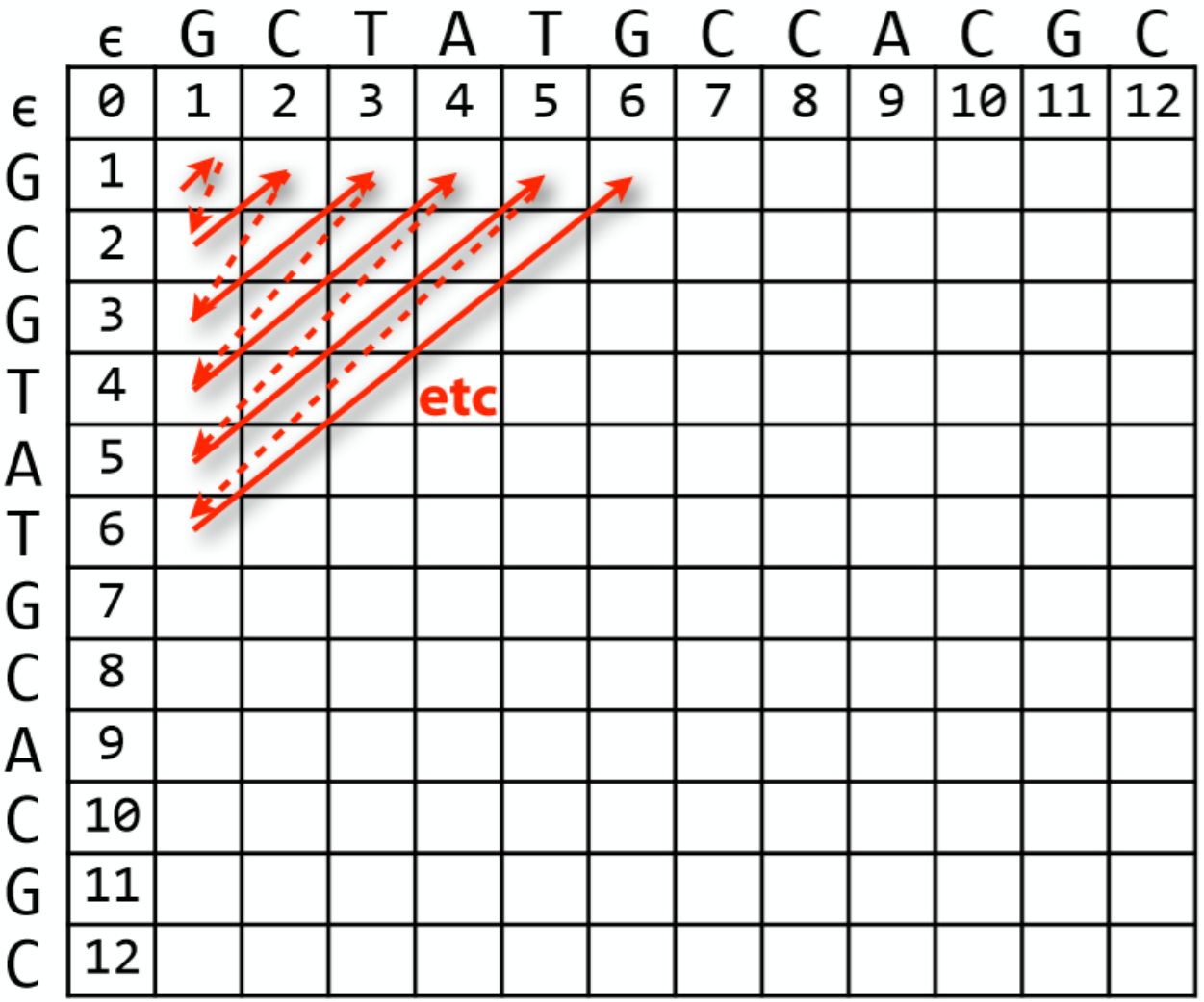
Editační vzdálenost - dynamické programování

		ϵ	G	C	T	A	T	G	C	C	A	C	G	C
	ϵ	0	1	2	3	4	5	6	7	8	9	10	11	12
G		1	0	1	2	3	4	5	6	7	8	9	10	11
C		2	1	0	1	2	3	4	5	6	7	8	9	10
G		3	2	1	1	2	3	3	4	5	6	7	8	9
T		4	3	2	1	2	2	3	4	5	6	7	8	9
A		5	4	3	2	1	2	3	4	5	5	6	7	8
T		6	5	4	3	2	1	2	3	4	5	6	7	8
G		7	6	5	4	3	2	1	2	3	4	5	6	7
C		8	7	6	5	4	3	2	1	2	3	4	5	6
A		9	8	7	6	5	4	3	2	2	2	3	4	5
C		10	9	8	7	6	5	4	3	2	3	2	3	4
G		11	10	9	8	7	6	5	4	3	3	3	2	3
C		12	11	10	9	8	7	6	5	4	4	3	3	2

← Editační vzdálenost

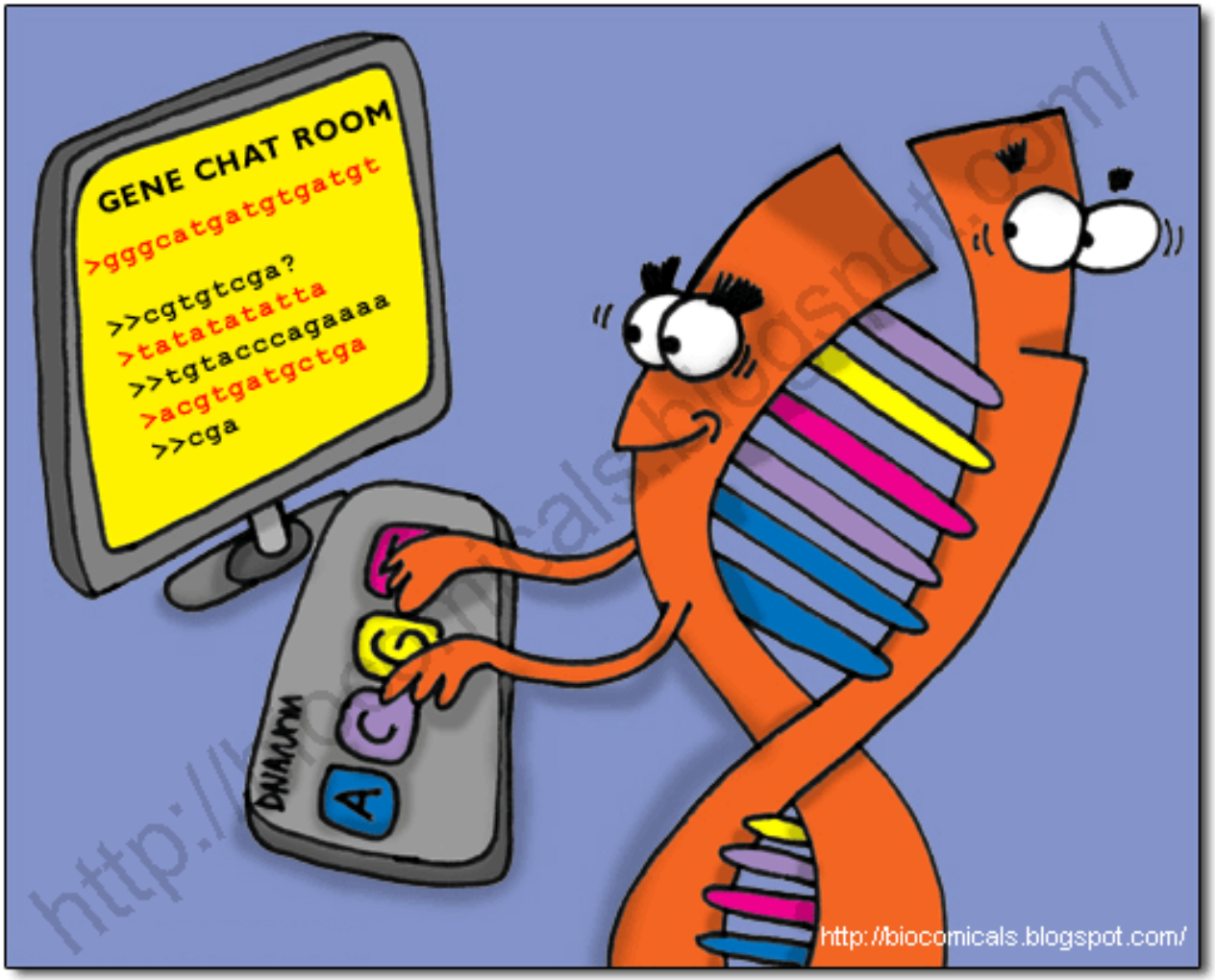
Editační vzdálenost - dynamické programování

Tabulku lze vyplňovat více způsoby:



Editační vzdálenost - shrnutí

- Vyplňujeme tabulku o m řádcích a n sloupcích.
- Editační vzdálenost nalezneme v pravém dolním políčku tabulky D.
- Prostorová a časová náročnost je $O(mn)$.
- Využití k určení podobnosti dvou sekvencí.


GENE CHAT ROOM
>gggcatgatgtgatgt
>>cgtgtcga?
>tatatatatta
>>tgtacccagaaaa
>acgtgatgctga
>>cga
A
G
http://biocomicals.blogspot.com/

Příště uvidíte

- Jak se zarovnávají sekvence vůči referenční sekvenci.
- Příklady algoritmů.
- Jak se modifikuje tabulka pro ohodnocení změn a proč.

DÄŻkuji za pozornost

Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA443

michal.vasinek@vsb.cz

26. září 2023

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY