Gde u genomu počinje replikacija DNK?

Bioinformatics Algorithms:
an Active Learning Approach
~Poglavlje 1~

Ćelije

- Svako živo biće se sastoji iz ćelija
- U svakoj ćeliji se neprestano dešavaju različiti procesi u kojima učestvuju sledeća hemijska jedinjenja:
  - Nukleinske kiseline
    - Dezoksiribonukleinska kiselina (DNK)
    - Ribonukleinska kiselina (RNK)
    - Sastoje se od azotnih baza (A,C,G,T/U)
  - Proteini
    - Nastaju na osnovu recepta koji je zapisan u DNK
    - Sastoje se od dvadeset esencijalnih aminokiselina

DNK

- DNK je nukleinska kiselina koja se nalazi u svakoj ćeliji živih bića
- DNK sadrži uputstva za razvoj organizama i pravilno funkcionisanje. Ova informacija se prenosi sa jedne na drugu ćeliju prilikom ćelijske deobe

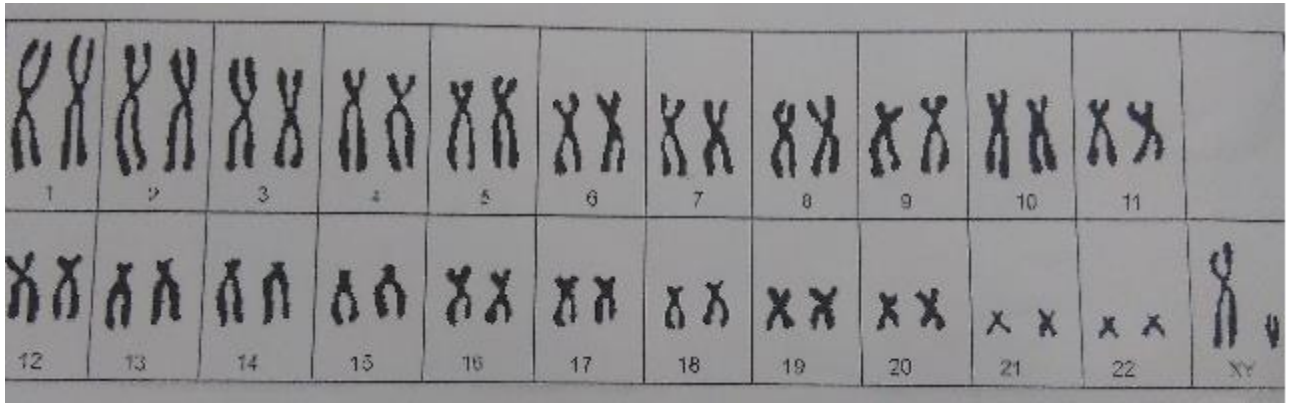
DNK

DNK

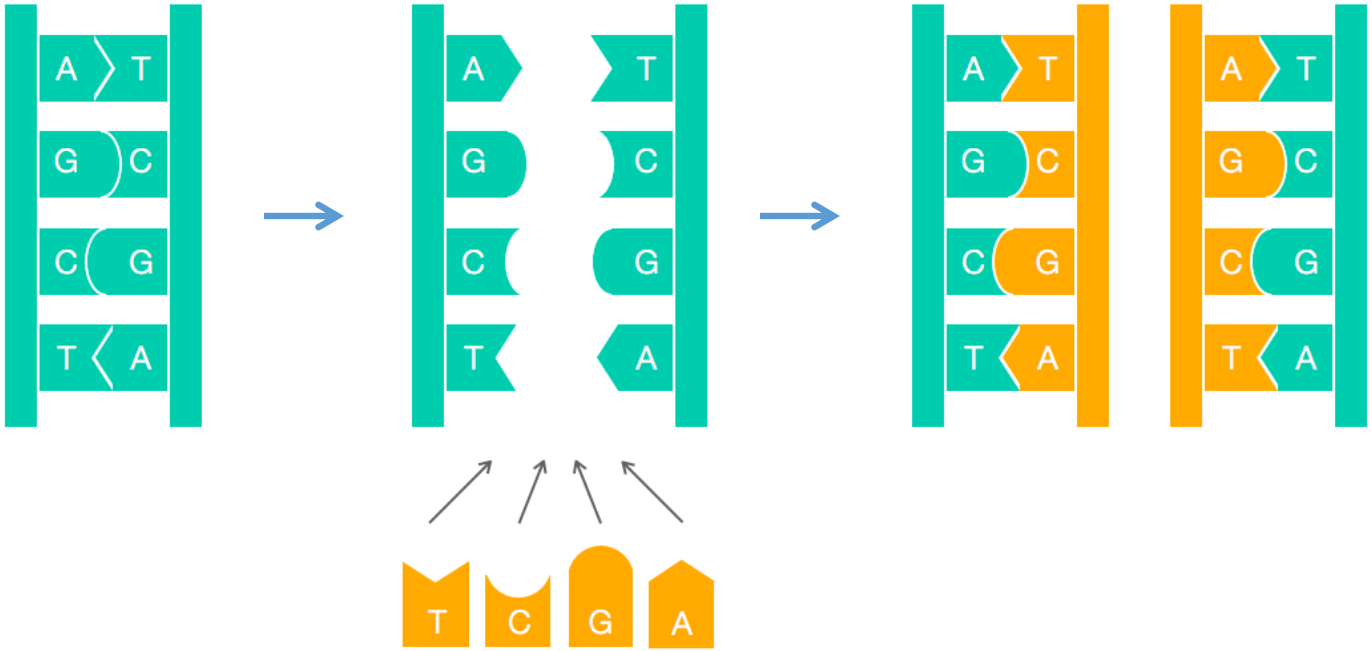
- DNK se sastoji iz dva lanca i svaki od njih je sastavljen od azotnih baza: adenin, timin, guanin i citozin (A, C, G, T); njih zovemo i nukleotidima
- Lanci DNK su međusobno spojeni i to tako da se veze uspostavljaju isključivo između A i T ili između G i C
- Tako, ako znamo sastav jednog lanca možemo zaključiti i sastav drugog lanca i zbog toga kažemo da su lanci DNK međusobno komplementarni
- Sa stanovišta računarstva, DNK posmatramo kao nisku nad azbukom {A,C,G,T}

DNK

- DNK je organizovana u strukure koje se nazivaju hromozomi
- Celokupni DNK sadržaj jednog organizma čini njegov genom
- Različiti organizmi imaju različit broj hromozoma
- Koliko je dugačka DNK kod čoveka?
- Kako se savija DNK (animacija)

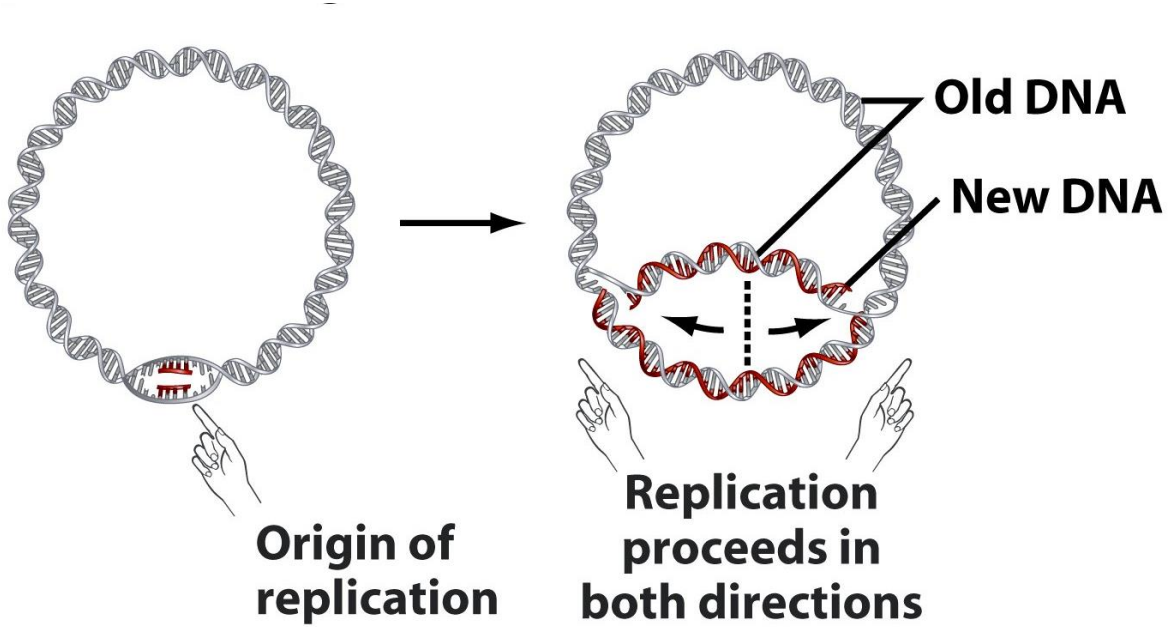


Replikacija genoma u ćeliji



- Pre nego što započne deobu, ćelija kreira kopiju svog genoma
- animacija

Početak replikacije



- Replikacija počinje u delu DNK koji nazivamo *početni region replikacije* (oriC, skraćeno od *origin of chromosomal replication*)
- Razmatraćemo samo bakterije koje imaju najčešće jedan hromozom kružnog oblika

Pregled

- **Pretraga za skrivenim porukama u početnom regionu replikacije**
  - **Šta je skrivena poruka u početnom regionu replikacije?**
  - Neke skrivene poruke su manje očekivane od drugih
  - Grupe skrivenih poruka
- **Pretraga za početnim regionom replikacije u genomu**
  - Iskrivljeni dijagrami
  - Nalaženje čestih reči sa propustima

Nalaženje početnog regiona replikacije

Problem nalaženja *oriC* : Nalaženje *oriC* u genomu.

- **Ulaz.** Genom.
- **Izlaz.** Lokacija *oriC* u genomu.

OK – hajde da isečemo ovaj DNK fragment. Da li genom može da se replicira bez njega?

Ovo nije dobro definisan problem!

Kako ćelija prepozna oriC?

Početni region replikacije *Vibrio cholerae* (≈500 nukleotida):

```
atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagatgatcaagagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttacggaagaatgatcaagctgctgctcttgatcatcgtttc
```

Ovde se mora nalaziti **skrivena poruka** koja govori ćeliji da započne replikaciju baš na ovom mestu.

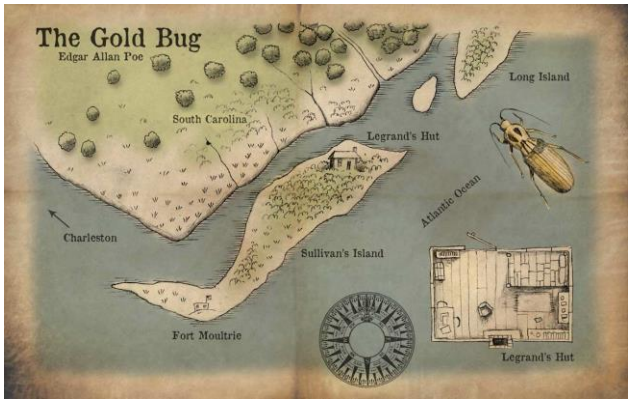
Problem skrivene poruke

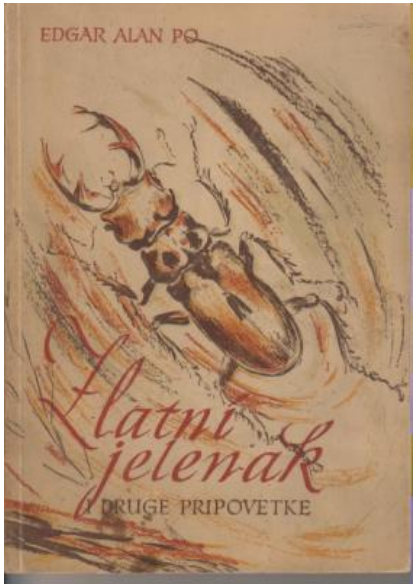
Problem skrivene poruke. Naći skrivenu poruku u niski karaktera.

- **Ulaz.** Niska *Text* (koja predstavlja region početka replikacije).
- **Izlaz.** Skrivena poruka u niski *Text*.

Pojam skrivene poruke nije precizno definisan.

Dešifrovanje





53++!305))6*;4826)4+.)4+);806*;48!8`60))85;]8*:+*8!83(88)5*!;46(;88*96*?;8)*+(;485);5*!2:*+(;4956*2(5*4)8`8*;4069285);)6!8)4++;1(+9;48081;8:8+1;48!85;4)485!528806*81(+9;48;(88;4(+?34;48)4+;161;:188;+?;

Zašto se “;48” tako često pojavljuje?

Nagoveštaj: Poruka je na engleskom jeziku.

53++!305))6*;4826)4+.)4+);806*48
!8`60))85;]8*:+*8!83(88)5*!46(88
96?;8)*+(;485);5*!2:*+(;4956*2
(5*4)8`8*;4069285);)6!8)4++;1(+9
;48081;8:8+1;48!85;4)485
528806*81(+9;48;(88;4(+?34;48)4+
;161;:188;+?;

“THE” je najčešća reč u engleskom jeziku

53++!305))6*THE26)4+.)4+)806*THE
!8`60))85;]8*:+*8!83(88)5*!;46(;
88*96*?;8)*+(THE5);5*!2:*+(;4956
*2(5*4)8`8*;4069285);)6!8)4++;1(
+9THE081;8:8+1THE!85;4)485!52880
6*81(+9THE;(88;4(+?34THE)4+;161;
:188;+?;

53++!305))6*THE26)H+.)H+)806*THE
!E`60))E5;]E*:+*E!E3(EE)5*!TH6(T
EE*96*?;E)*+(THE5)T5*!2:*+(TH956
*2(5*H)E`E*TH0692E5)T)6!E)H++T1(
+9THE0E1TE:E+1THE!E5T4)HE5!5288 0
6*E1(+9THET(EETH(+?34THE)H+T161T
:1EET+?T

Problem skrivene poruke

Problem skrivene poruke. Naći skrivenu poruku u niski karaktera.

- **Ulaz.** Niska *Text* (koja predstavlja region početka replikacije).
- **Izlaz.** Skrivena poruka u niski *Text*.

Pojam skrivene poruke nije precizno definisan.

Nagoveštaj: U različitim biološkim niskama, neke reči se pojavljuju iznenađujuće često u kratkim delovima genoma

na primer, **AATTT** je niska dužine 5 (5-gram) koja se pojavljuje iznenađujuće često u sledećem tekstu:

ACA**AATTT**GCAT**AATTT**CGGGA**AATTT**CCT

Problem čestih reči

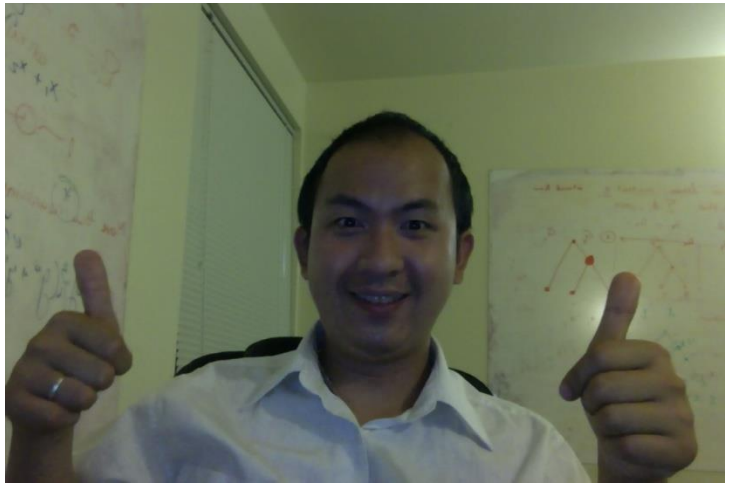
Problem čestih reči. Pronaći najčešće *k*-grame u niski karaktera.

- **Ulaz.** Niska *Text* i ceo broj *k*.
- **Izlaz.** Svi **najčešći *k*-grami** u niski *Text*.

Problem čestih reči

Problem čestih reči. Pronaći najčešće *k*-grame u niski karaktera.

- **Ulaz.** Niska *Text* i ceo broj *k*.
- **Izlaz.** Svi **najčešći *k*-grami** u niski *Text*.



Kažemo da je *k*-gram ***Pattern*** **najčešći *k*-gram** u tekstu ako se nijedan drugi *k*-gram ne pojavljuje više puta nego *Pattern.*

AATTT je najčešći 5-gram u sledećem tekstu:

ACA**AATTT**GCAT**AATTT**CGGGA**AATTT**CCT

Da li problem čestih reči ima biološki smisao?

Problem čestih reči. Pronaći najčešće *k*-grame u niski karaktera.

- **Ulaz.** Niska *Text* i ceo broj *k*.
- **Izlaz.** Svi **najčešći *k*-grami** u niski *Text*.

Replikaciju DNK u ćeliji vrši enzim DNK polimeraza, a ceo proces započinje protein ***DnkA***.

DnkA se vezuje za kratke segmente unutar regiona početka replikacije, dužine obično 9 nukleotida, koji se nazivaju DnkA boksovi.

DnkA boks je skrivena poruka koja ukazuje DnkA gde da se veže za DNK.

U regionu početka replikacije postoji više DnkA boksova.

Koja je vremenska složenost algoritma za rešavanje problema čestih reči?

Problem čestih reči. Pronaći najčešće *k*-grame u niski karaktera.

- **Ulaz.** Niska *Text* i ceo broj *k*.
- **Izlaz.** Svi **najčešći *k*-grami** u niski *Text*.

- $|Text|^2 \cdot k$
- $|Text|$

Na vežbama i u nastavku: kako **naivni i spor** algoritam vremenske složenosti $|Text|^2 \cdot k$ može biti modifikovan u **brzi** algoritam složenosti $|Text|$

Pregled

- **Pretraga za skrivenim porukama u početnom regionu replikacije**
  - Šta je skrivena poruka u početnom regionu replikacije?
  - **Neke skrivene poruke su manje očekivane od drugih**
  - Grupe skrivenih poruka
- **Pretraga za početnim regionom replikacije u genomu**
  - Iskrivljeni dijagrami
  - Nalaženje čestih reči sa propustima

*Ori*C bakterije Vibrio Cholerae

```
atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaacctgagtgg
atgacatcaagataggtcgttgtatctccttcctctcgtactctcatgaccacggaaagatgatcaag
agaggatgatttcttggccatatcgcaatgaatacttgtgacttgtgcttccaattgacatcttcagc
gccatattgcgctggccaaggtgacggagcgggattacgaaagcatgatcatggctgttgttctgttt
atcttgttttgactgagacttgttaggatagacggtttttcatcactgactagccaaagccttactct
gcctgacatcgaccgtaaattgataatgaatttacatgcttccgcgacgatttacctcttgatcatcg
atccgattgaagatcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatca
tgtttccttaaccctctatttttacggaagaatgatcaagctgctgctcttgatcatcgtttc
```

- Kod bakterija je *ori*C obično dužine nekoliko stotina nukleotida

Od svih najčešćih reči, koja predstavlja skrivenu poruku?

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaacctgagtgg
atgacatcaagataggtcgttgtatctccttcctctcgtactctcatgaccacggaaag**ATGATCAAG**
agaggatgatttcttggccatatcgcaatgaatacttgtgacttgtgcttccaattgacatcttcagc
gccatattgcgctggccaaggtgacggagcgggattacgaaagcatgatcatggctgttgttctgttt
atcttgttttgactgagacttgttaggatagacggtttttcatcactgactagccaaagccttactct
gcctgacatcgaccgtaaattgataatgaatttacatgcttccgcgacgatttacct**CTTGATCAT**cg
atccgattgaagatcttcaattgttaattctcttgcctcgactcatagccatgatgagct**CTTGATCA**
Tgtttccttaaccctctatttttacggaaga**ATGATCAAG**ctgctgct**CTTGATCAT**cgtttc

Najčešći 9-grami u ovom *oriC* (svi su se pojavili 3 puta):
ATGATCAAG, **CTTGATCAT**, **TCTTGGATCA**, **CTCTTGATC**

Koliko je verovatno da se u oko 500 nukleotida 9-gram pojavi 3 ili više puta? 1/1300

Skrivena poruka pronađena

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaacctgagtgg
atgacatcaagataggtcgttgtatctccttcctctcgtactctcatgaccacggaaag**ATGATCAAG**
agaggatgatttcttggccatatcgcaatgaatacttgtgacttgtgcttccaattgacatcttcagc
gccatattgcgctggccaaggtgacggagcgggattacgaaagcatgatcatggctgttgttctgttt
atcttgttttgactgagacttgttaggatagacggttttcatcactgactagccaaagccttactct
gcctgacatcgaccgtaaattgataatgaatttacatgcttccgcgacgatttacct**CTTGATCAT**cg
atccgattgaagatcttcaattgttaattctcttgcctcgactcatagccatgatgagct**CTTGATCA**
Tgtttccttaaccctctatttttacggaaga**ATGATCAAG**ctgctgct**CTTGATCAT**cgtttc

ATGATCAAG

||||||||| ove niske su obrnuto komplementarne

TACTAGTTC (***DnkA*** se može vezati za obe)

Iznenađujuće je da se 9-gram pojavi 6 ili više puta (računajući i obrnute komplemente) u kratkom segmentu od ≈ 500 nukleotida => to su DnkA boksovi

Skrivene poruke kod bakterije *Thermotoga petrophila*

```
aactctatacctccttttgtcgaatttgtgtgatttatagagaaaatcttattaactgaaactaa
aatggtaggtttggtggtaggttttgtgtacattttgtagtatctgatttttaattacataccgta
tattgtattaaattgacgaacaattgcatggaattgaatatatgcaaaacaaacctaccaccaaac
tctgtattgaccattttaggacaacttcagggtggtaggtttctgaagctctcatcaatagactat
tttagtctttacaaacaatattaccgttcagattcaagattctacaacgctgttttaatgggcgtt
gcagaaaacttaccacctaaaatccagtatccaagccgatttcagagaaacctaccacttacctac
cacttacctaccacccgggtggtaagttgcagacattattaaaaacctcatcagaagcttgttcaa
aaatttcaatactcgaaacctaccacctgcgtcccctattatttactactactaataatagcagta
taattgatctgaaaagaggtggtaaaaaa
```

Poruke **ATGATCAAG** i **CTTGATCAT** iz Vibrio Cholerae se nijednom ne pojavljuju u ovom *ori*C

Najčešće reči u ovom *ori*C:
AACCTACCA, ACCTACCAC, GGTAGGTTT, TGGTAGGTT, AAACCTACC, CCTACCACC

Različiti genomi → različite skrivene poruke (*DnkA* boksovi)

Skrivene poruke kod bakterije *Thermotoga petrophila*

aactctatacctccttttttgtcgaatttgtgtgatttatagagaaaatcttattaactgaaactaa
aatggtaggtttGGTGGTAGGttttgtgtacattttgtagtatctgatttttaattacataccgta
tattgtattaaattgacgaacaattgcatggaattgaatatatgcaaaacaaaCCTACCACCaaac
tctgtattgaccattttaggacaacttcagGGTGGTAGGtttctgaagctctcatcaatagactat
tttagtctttacaaacaatattaccgttcagattcaagattctacaacgctgttttaatgggcgtt
gcagaaaacttaccacctaaaatccagtatccaagccgatttcagagaaacctaccacttacctac
cactaCCTACCACCcgggtggtaagttgcagacattattaaaaacctcatcagaagcttgttcaa
aaatttcaatactcgaaaCCTACCACCtgcgtcccctattatttactactactaataatagcagta
taattgatctgaaaagaggtggtaaaaaa

CCTACCACC
||||||||| kandidati za skrivene poruke.
GGATGGTGG

Naučili smo da pronađemo skrivene poruke **ako je *oriC* dat**, ali ne znamo da pronađemo *oriC* u genomu.

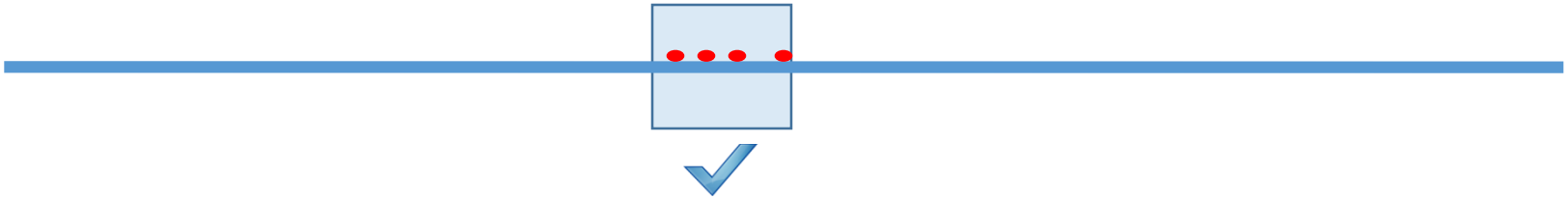
Pregled

- **Pretraga za skrivenim porukama u početnom regionu replikacije**
  - Šta je skrivena poruka u početnom regionu replikacije?
  - Neke skrivene poruke su manje očekivane od drugih
  - **Grupe skrivenih poruka**
- **Pretraga za početnim regionom replikacije u genomu**
  - Iskrivljeni dijagrami
  - Nalaženje čestih reči sa propustima

Pronalaženje početnog regiona replikacije

Prethodni problem: ako je **poznat** *oriC* (prozor unutar DNK dužine oko 500 nukleotida), naći **česte reči** u *oriC* koje predstavljaju kandidate za *DnkA* boksove.

početni region replikacije → **česte reči**



Pronalaženje početnog regiona replikacije

Prethodni problem: ako je **poznat** *oriC* (prozor unutar DNK dužine oko 500 nukleotida), naći **česte reči** u *oriC* koje predstavljaju kandidate za *DnkA* boksove.

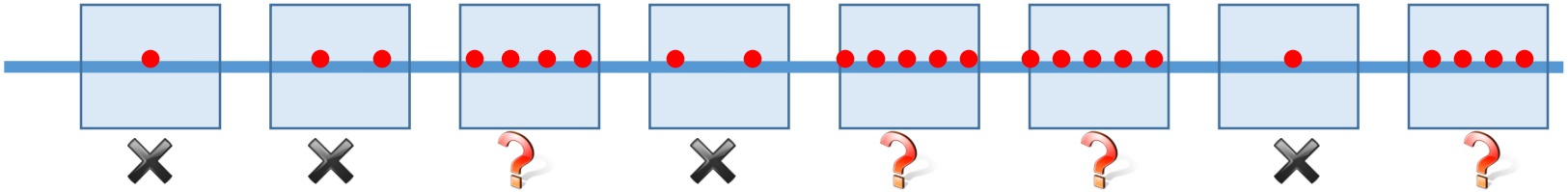
početni region replikacije → **česte reči**

Šta ako ne znamo gde se *ori*C nalazi u genomu?

Pronalaženje početnog regiona replikacije

Prethodni problem: ako je **poznat** *oriC* (prozor unutar DNK dužine oko 500 nukleotida), naći **česte reči** u *oriC* koje predstavljaju kandidate za *DnkA* boksove.

početni region replikacije → **česte reči**



Novi problem: naći česte reči u **svim** prozorima unutar genoma. Prozori koji sadrže **grupe** čestih reči predstavljaju kandidate za početni region replikacije.

česte reči → **početni region replikacije**

Kako definišemo grupu čestih reči?

Formalno: Kažemo da *k*-gram formira (***L***, ***t***)-**grupu** unutar teksta *Genome* ako postoji **kratak** interval (dužine ***L***) teksta *Genome* u kom se *k*-gram pojavljuje **više puta** (bar ***t***).

Problem pronalaženja grupa. Naći *k*-grame koji formiraju grupe unutar niske karaktera.

- **Ulaz.** Niska *Genome* i celi brojevi *k* (dužina podniske), ***L*** (dužina prozora) i ***t*** (broj podniski u grupi).
- **Izlaz.** Svi *k*-grami koji formiraju *(**L**, **t**)*-**grupe** u niski *Genome*.

U genomu bakterije *E.coli* postoji 1904 različitih 9-grama koji formiraju (500,3)-grupe. Koji od njih ukazuje na početni region replikacije?

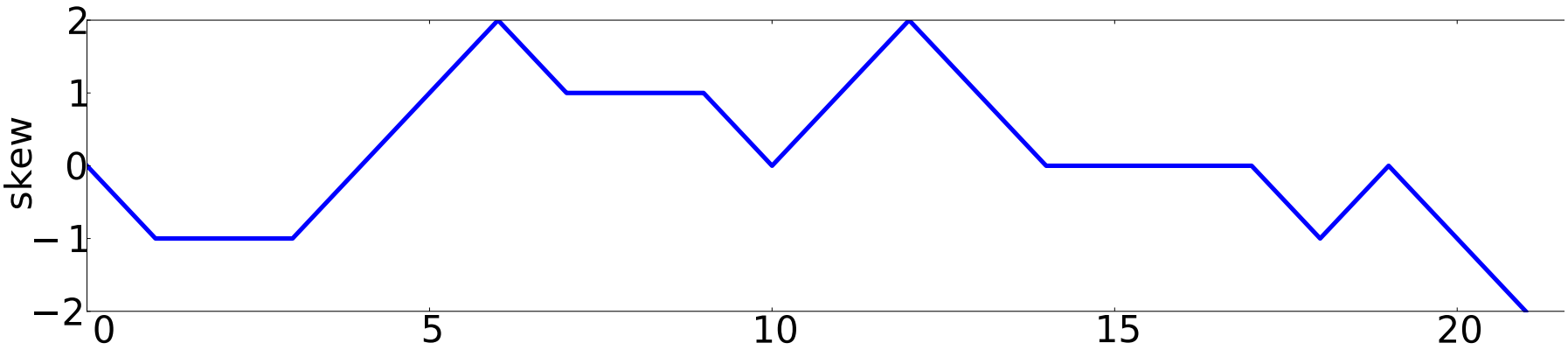
Pregled

- **Pretraga za skrivenim porukama u početnom regionu replikacije**
  - Šta je skrivena poruka u početnom regionu replikacije?
  - Neke skrivene poruke su manje očekivane od drugih
  - Grupe skrivenih poruka
- **Pretraga za početnim regionom replikacije u genomu**
  - **Iskrivljeni dijagrami**
  - Nalaženje čestih reči sa propustima

Iskrivljeni (skew) dijagram

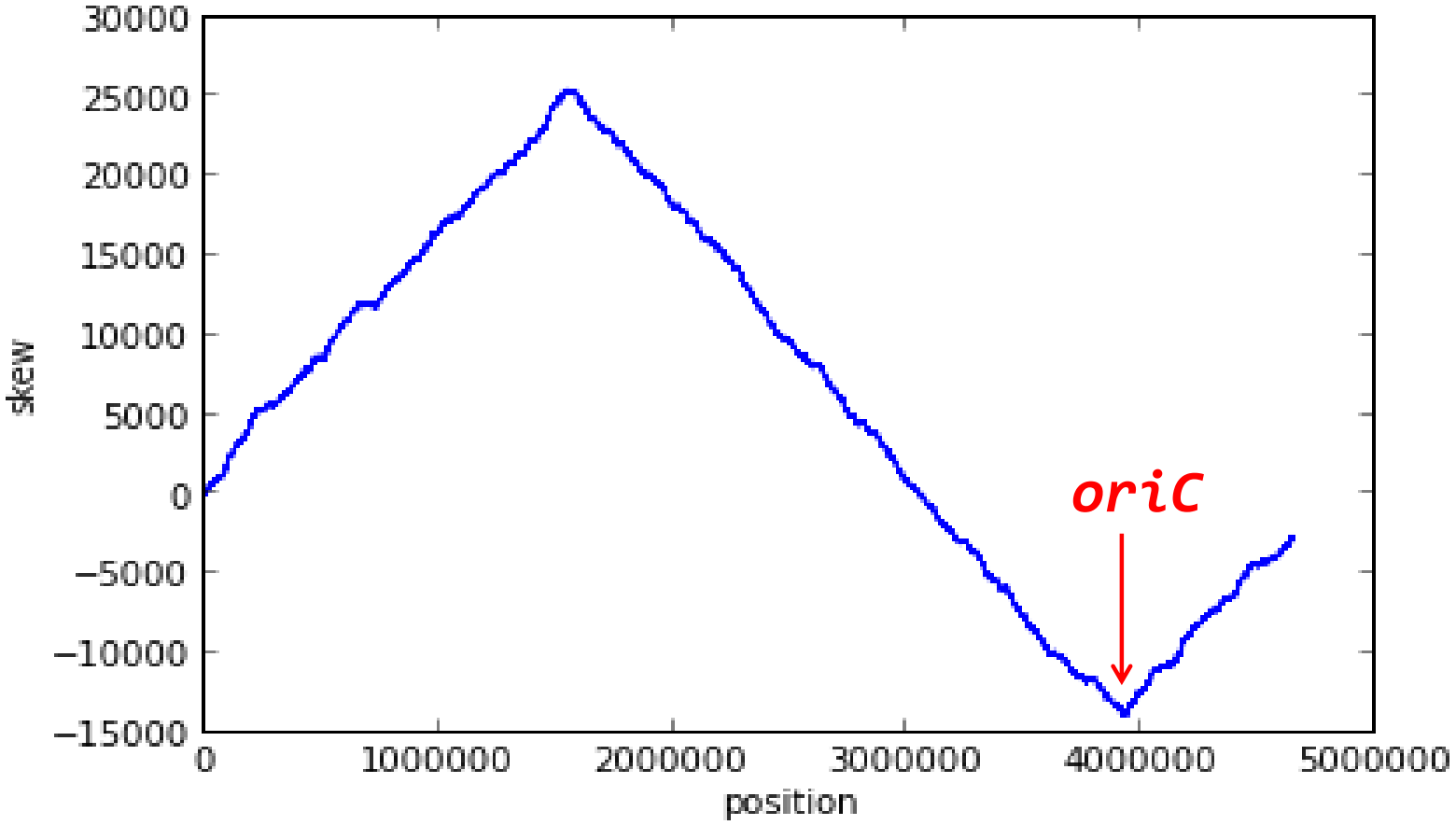
Skew(k): #G - #C za prvih *k* nukleotida teksta *Genome*.

Skew dijagram: Grafik funkcije *Skew*(*k*)

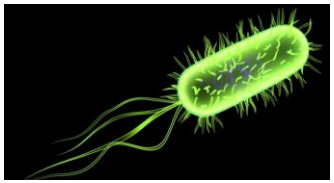


CATGGGCATCGGCCATACGCC

Iskrivljeni dijagram bakterije *E. Coli*: Gde je početni region replikacije?



Našli smo početni region replikacije za E.Coli ALI …



Minimalna vrednost iz Skew dijagrama ukazuje na ovaj region u *E. coli*:

```
aatgatgatgacgtcaaaaggatccggataaaacatggtgattgcctcgcataacgcggta
tgaaaatggattgaagcccgggccgtggattctactcaactttgtcggcttgagaaagacc
tgggatcctgggtattaaaaagaagatctatttatttagagatctgttctattgtgatctc
ttattaggatcgcactgccctgtggataacaaggatccggcttttaagatcaacaacctgg
aaaggatcattaactgtgaatgatcggtgatcctggaccgtataagctgggatcagaatga
ggggttatacacaactcaaaaactgaacaacagttgttctttggataactaccggttgatc
caagcttcctgacagagttatccacagtagatcgcacgatctgtatacttatttgagtaaa
ttaacccacgatcccagccattcttctgccggatcttccggaatgtcgtgatcaagaatgt
tgatcttcagtg
```

U ovom regionu **nema** čestih 9-grama (koji se pojavljuju 3 ili više puta)!

Pregled

- **Pretraga za skrivenim porukama u početnom regionu replikacije**
  - Šta je skrivena poruka u početnom regionu replikacije?
  - Neke skrivene poruke su manje očekivane od drugih
  - Grupe skrivenih poruka
- **Pretraga za početnim regionom replikacije u genomu**
  - Iskrivljeni dijagrami
  - **Nalaženje čestih reči sa propustima**

Reči koje liče na česte reči

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaag**ATGATCAAG**agaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacct**CTTGATCAT**cgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagct**CTTGATCAT**gtt
tccttaaccctctatttttacggaaga**ATGATCAAG**ctgctgct**CTTGATCAT**cgtttc

oriC kod *Vibrio cholerae* ima 6 *DnkA* boksova – **možemo li naći još reči koje liče na njih?**

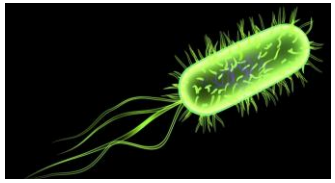
Reči koje liče na česte reči

oriC kod *Vibrio cholerae* sadrži ATGATCAAC i CATGATCAT, koje se od *DnkA* boksova ATGATCAAG/CTTGATCAT razlikuju na jednoj poziciji:

atca**ATGATCAAC**gtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaag**ATGATCAAG**agaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaag**CATGATCAT**ggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacct**CTTGATCAT**cgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagct**CTTGATCAT**gtt
tccttaaccctctatttttacggaaga**ATGATCAAG**ctgctgct**CTTGATCAT**cgtttc

- **Problem čestih reči sa propustima.** Pronaći najčešće *k*-grame sa propustima u niski karaktera.
- **Ulaz.** Niska *Text* i celi brojevi *k* i *d*.
- **Izlaz.** Svi **najčešći *k*-grami** sa najviše *d* propusta u niski *Text*.

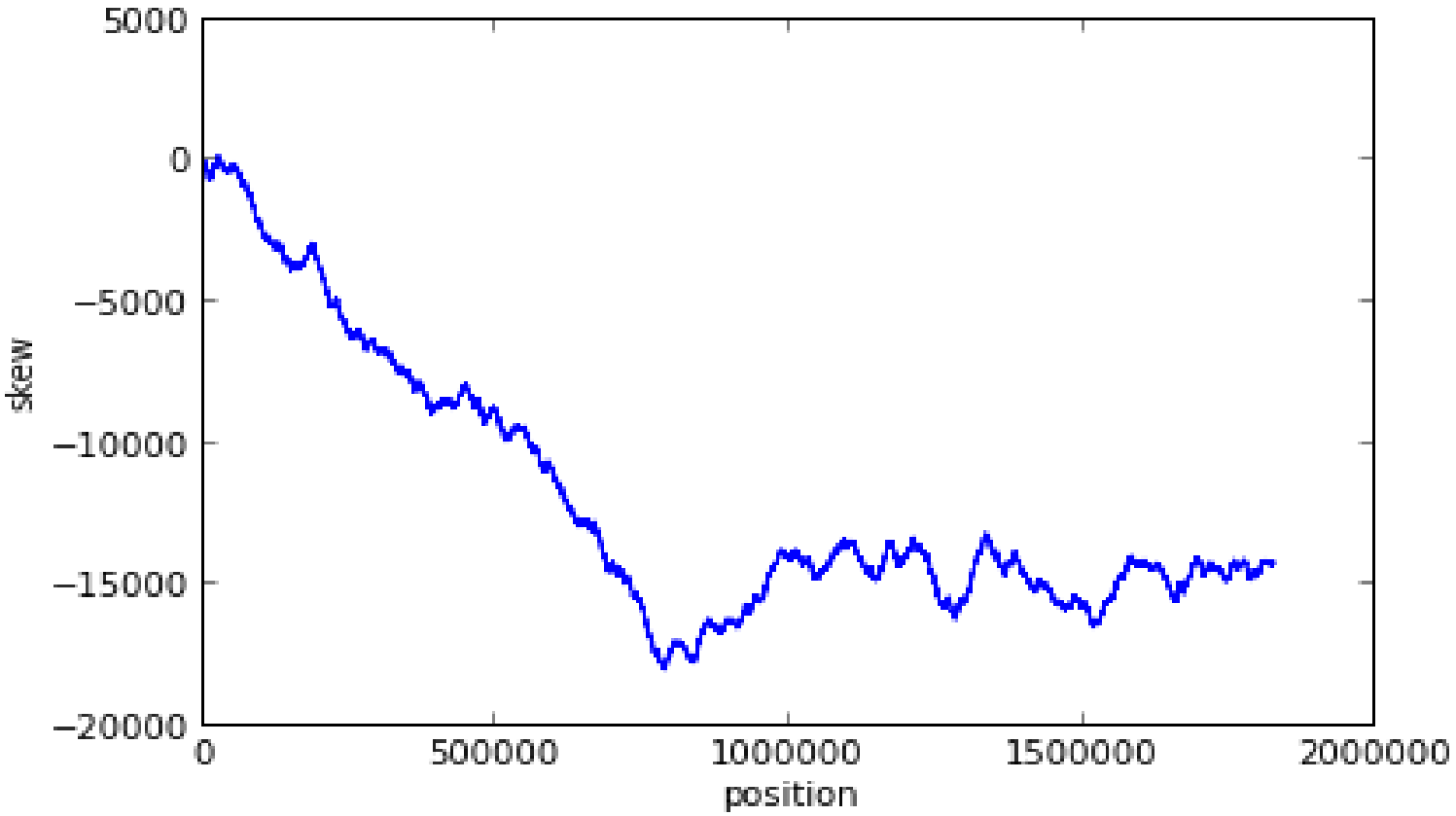
DnkA boksovi kod *E. Coli*



Česti 9-grami (sa jednim propustom i obrnutim komplementima) u *oriC* bakterije *E. coli*

aatgatgatgacgtcaaaaggatccggataaaacatggtgattgcctcgcataacgcgg
tatgaaaatggattgaagcccgggccgtggattctactcaactttgtcggcttgagaaa
gacctgggatcctgggtattaaaaagaagatctatttatttagagatctgttctattgt
gatctcttattaggatcgcactgcccTGTGGATAAcaaggatccggcttttaagatcaa
caacctggaaaggatcattaactgtgaatgatcggtgatcctggaccgtataagctggg
atcagaatgaggggTTATACACAactcaaaaactgaacaacagttgttcTTTGGATAAC
taccggttgatccaagcttcctgacagagTTATCCACAgtagatcgcacgatctgtata
cttatttgagtaaattaacccacgatcccagccattcttctgccggatcttccggaatg
tcgtgatcaagaatgttgatcttcagtg

- Skew dijagram je često kompleksniji nego u slučaju *E. coli*.



Skew dijagram bakterije *Thermotoga petrophila*

- Slajdovi pokrivaju poglavlje 1 knjige *Bioinformatics Algorithms: an Active Learning Approach*
- Sadržaj slajdova je preuzet sa zvaničnih prezentacija autora uz njihovu dozvolu i dodatno prilagođen