Kako pronalazimo regulatorne motive u DNK?

Gruba sila i randomizovani algoritmi

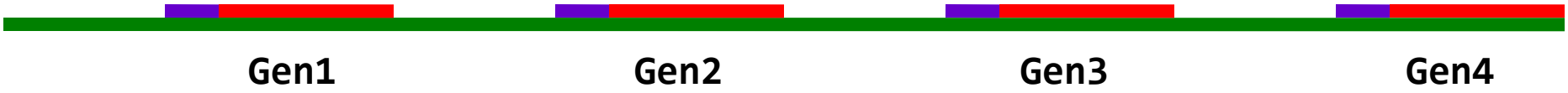
Bioinformatics Algorithms:
an Active Learning Approach
~Poglavlje 2~

Pregled

- **Regulatorni motivi**
  - **DNK kao izvor informacija**
  - Genska ekspresija i cirkadijalni ritam
  - Formulacija 1
  - Formulacija 2
  - Formulacija 3
  - Problem niske medijane
  - Pohlepna pretraga motiva
- Kako bacanje kockica pomaže u pronalaženju regulatornih motiva
  - Slučajna pretraga motiva
  - Gibsovo sempliranje
  - Pseudovrednosti

DNK kao izvor informacija

- Funkcionisanje svakog živog bića je omogućeno kroz fizičko-hemijske reakcije koje se dešavaju unutar ćelija
- Da bi došlo do odgovarajuće fizičko-hemijske reakcije, u ćeliji moraju nastati molekuli koji učestvuju u toj reakciji.
- Neki od molekula koji učestvuju u fizičko-hemijskim reakcijama u ćeliji su po sastavu *proteini* i aminokiseline od kojih se grade postoje u citoplazmi svake ćelije.
- ***Unutar DNK je zapisana informacija za stvaranje (sintezu) proteina u ćeliji.***
- Svaki segment DNK koji sadrži informaciju za sintezu jednog proteina predstavlja jedan **gen**


Gen1
Gen2
Gen3
Gen4

DNK kao izvor informacija

- U svakoj ćeliji jednog organizma se nalazi isti molekul DNK.
- Ipak, različite grupe ćelija imaju različite uloge u organizmu (nervne ćelije, ćelije kože, krvne ćelije, …).
- Kako ćelija uspeva da izvrši različite uloge (tj. da sintetiše baš onaj protein koji je potreban za neku fizičko-hemijsku reakciju) kada je izvor informacija (DNK tj. skup gena) isti u svakoj ćeliji jednog organizma?

DNK kao izvor informacija

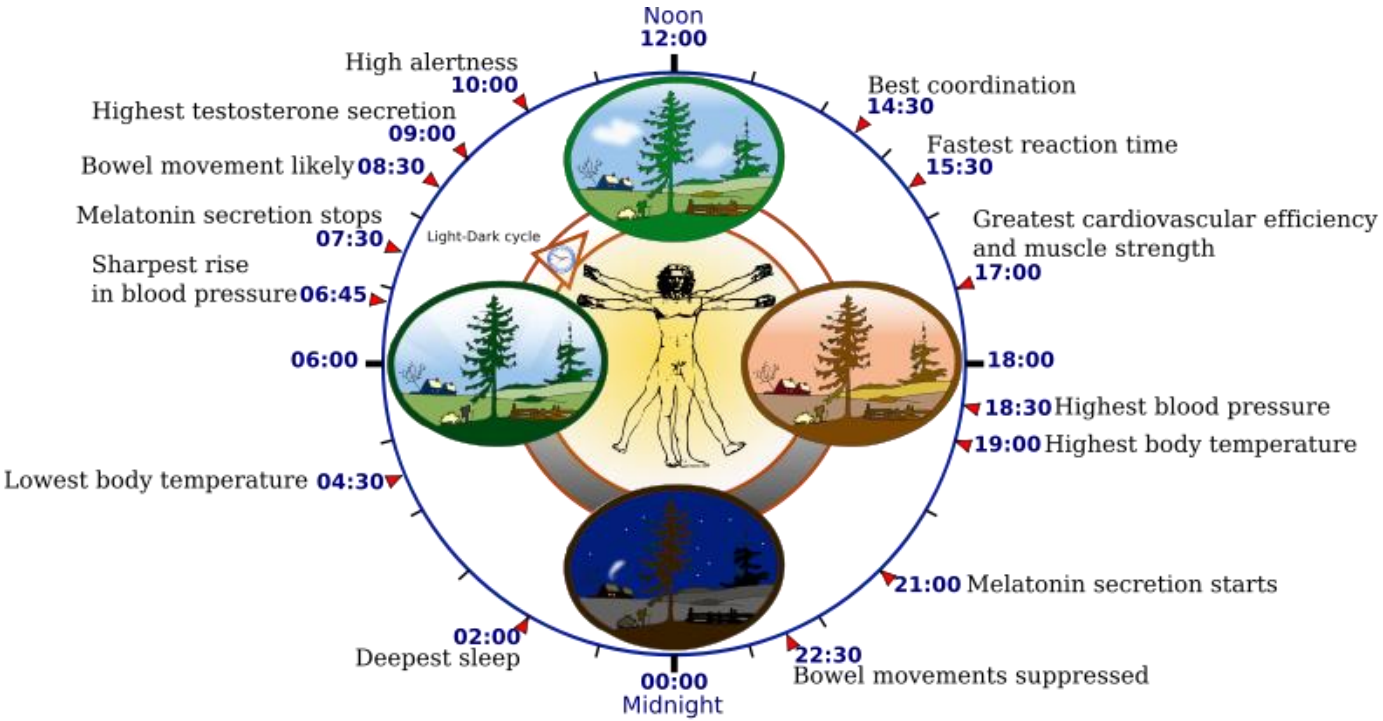
- Ne koristi svaka ćelija informacije iz svih gena istovremeno – neke informacije se koriste u jednom trenutku, neke u nekom drugom
- Uz to, neke ćelije će neke gene koristiti nekad, dok neke druge ćelije te iste gene neće koristiti nikad, iako ih imaju na raspolaganju (tj. sadrže istu DNK)
- Zato kažemo da neki gen u ćeliji nekad može *da se ispoljava* (da dolazi do njegove ***ekspresije*** odnosno na osnovu informacije koju gen nosi dolazi do sinteze proteina) a nekad *da se ne ispoljava*

Pregled

- **Regulatorni motivi**
  - **DNK kao izvor informacija**
  - **Genska ekspresija i cirkadijalni ritam**
  - Formulacija 1
  - Formulacija 2
  - Formulacija 3
  - Problem niske medijane
  - Pohlepna pretraga motiva
- Kako bacanje kockica pomaže u pronalaženju regulatornih motiva
  - Slučajna pretraga motiva
  - Gibsovo sempliranje
  - Pseudovrednosti

Cirkadijalni ritam

Dnevni ritam u funkcionisanju svakog živog bića



Kako svaka ćelija u našem organizmu zna koliko je sati?

Kako biljke znaju da li je noć ili dan?

Biljke moraju da promene gensku ekspresiju za preko 1000 gena (za fotosintezu, za cvetanje, za otpornost na mraz, ...) pri prelasku iz dana u noć

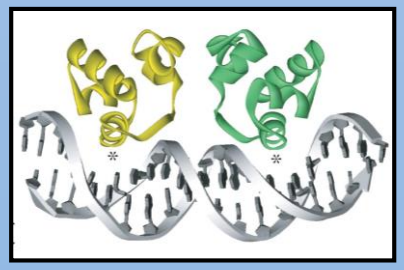
Koji molekuli govore genima da promene ekspresiju u milijardama ćelija?

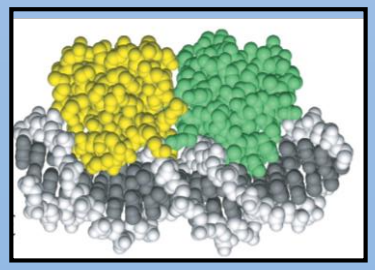
Neki kaktusi cvetaju samo noću

Suncokret se okreće prema Suncu

Transkripcija i transkripcioni faktori

- **Transkripcija gena** predstavlja prvi korak sinteze proteina.
- Početak ovog procesa iniciraju posebni proteini zaduženi za regulaciju transkripcije, takozvani **transkripcioni faktori**.
- Oni omogućavaju da započne transkripcija gena tako što se vezuju za kratke DNK fragmente (**mesta vezivanja transkripcionih faktora**) koji se nalaze *ispred* gena.
- animacija





Kako biljke znaju da li je noć ili dan?

Biljke moraju da promene gensku ekspresiju za preko 1000 gena (za fotosintezu, za cvetanje, za otpornost na mraz, ...) pri prelasku iz dana u noć

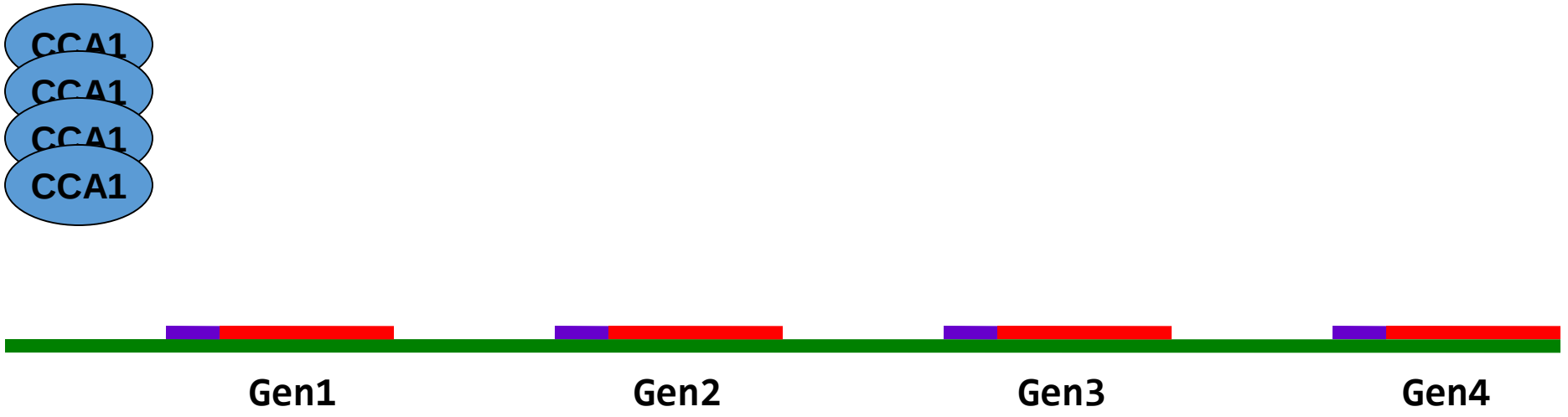
Koji molekuli govore genima da promene ekspresiju u milijardama ćelija?

Transkripcioni faktori.
U slučaju biljaka
to su CCA1, LCY i TOC1.

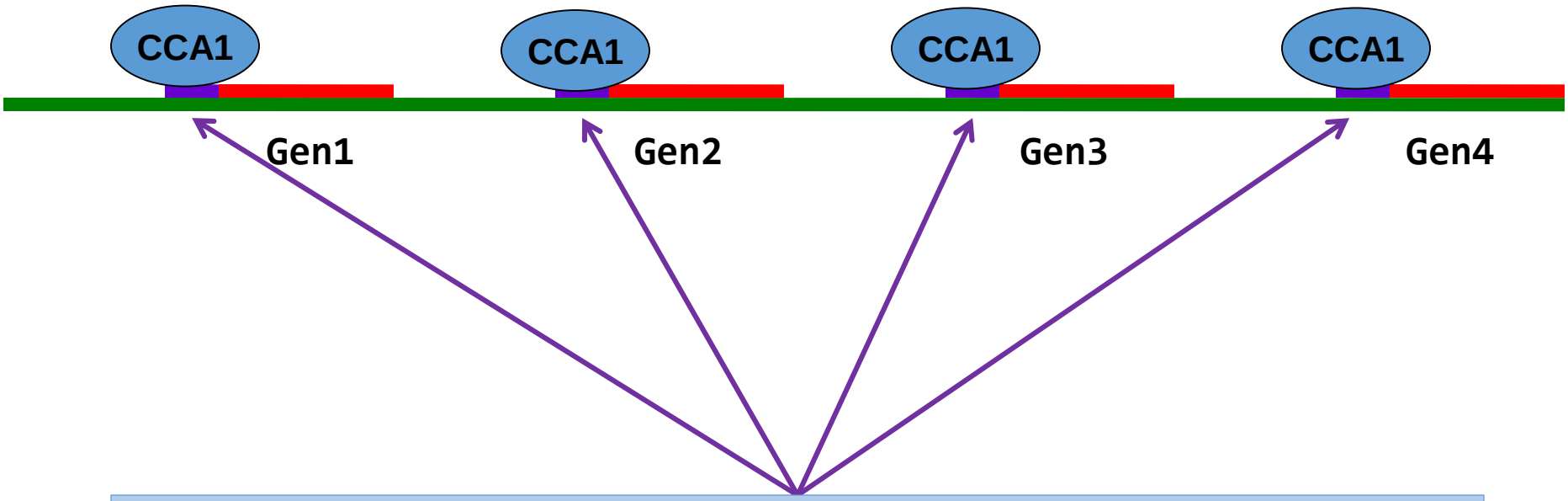
Neki kaktusi
cvetaju samo noću

Suncokret
se okreće prema Suncu

Transkripcioni faktori se vezuju za svoja mesta vezivanja na DNK



Kako CCA1 zna gde je njegovo mesto vezivanja?



U ovim regionima mora da postoji neka skrivena poruka koja govori CCA1 **gde** da se veže.

Mesta vezivanja transkripcionih faktora

Skrivena poruka (*regulatorni motiv* **AAAAAATCT**): mesto vezivanja transkripcionog faktora CCA1

Regulatorni motivi

cagt**ATAAAGTCT**actgatgcaacctgactcatgacgaggaa

Gen1

agtcgactgactt**AACAAATCT**cggatcgattcgtccgagga

Gen2

cgtcagctctgtcgggattcgccccgtattc**AAAAAAGCT**ac

Gen3

accgtctcc**ACAAAACCT**gctcgtgccactgatgcaacctga

Gen4

- Regulatorni motivi predstavljaju delove DNK koji se nalaze ispred gena
- Za prikazane gene se vezuje *isti* transkripcioni faktor (CCA1) pa stoga kažemo da su ovi geni *koregulisani*
- Ipak, regulatorni motivi kod koregulisanih gena *nisu isti* ali jesu *slični*

Gde se kriju skrivene poruke?

cagtataaagtctactgatgcaacctgactcatgacgaggaa

Gen1

agtcgactgacttaacaaatctcggatcgattcgtccgagga

Gen2

cgtcagctctgtcgggattcgccccgtattcaaaaaagcac

Gen3

accgtctccacaaaacctgctcgtgccactgatgcaacctga

Gen4

- Kao i ostali delovi DNK, i regulatorni motivi mogu da *mutiraju* i da nakon manjeg broja mutacija zadrže svoja svojstva
- Kako da lociramo regulatorne motive bez znanja kako oni treba da izgledaju?

Pregled

- **Regulatorni motivi**
  - **DNK kao izvor informacija**
  - **Genska ekspresija i cirkadijalni ritam**
  - **Formulacija 1**
  - Formulacija 2
  - Formulacija 3
  - Problem niske medijane
  - Pohlepna pretraga motiva
- Kako bacanje kockica pomaže u pronalaženju regulatornih motiva
  - Slučajna pretraga motiva
  - Gibsovo sempliranje
  - Pseudovrednosti

Formulacija problema

- Šta imamo: kolekciju nukleotidnih sekvenci koje se nalaze ispred koregulisanih gena
- Šta želimo: da unutar te kolekcije pronađemo podsekvence (motive) koji su međusobno slični
- Ovaj biološki problem se naziva *problemom nalaženja motiva*
- Ovom problemu se može pristupiti na različite načine i stoga je poznato više njegovih *formulacija*

Formulacija 1

- Pretpostavimo da se kolekcija sekvenci koju imamo na raspolaganju može kreirati na sledeći način:
  - na slučajan način generišemo t nukleotidnih sekvenci dužine n
  - u svaku sekvencu *ubacimo* šablon dužine k na proizvoljnu poziciju (dužina svake niske ostaje n)
  - svaki *ubačeni šablon* mutiramo na najviše k proizvoljnih pozicija

Generišemo na slučajan način deset nukleotidnih sekvenci

atgaccgggatactgataccgtatttggcctaggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatactgggcataaggtaca

tgagtatccctgggatgacttttgggaacactatagtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgaccttgtaagtgttttccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatggcccacttagtccacttatag

gtcaatcatgttcttgtgaatggatttttaactgagggcatagaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtactgatggaaactttcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttggtttcgaaaatgctctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatttcaacgtatgccgaaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttctgggtactgatagca

Ubacimo šablon **AAAAAAAGGGGGGG** na slučajno odabrane pozicije

atgaccgggatactgat**AAAAAAAAGGGGGGG**ggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaata**AAAAAAAAGGGGGGG**a

tgagtatccctgggatgactt**AAAAAAAAGGGGGGG**tgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatg**AAAAAAAAGGGGGGG**tccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaat**AAAAAAAAGGGGGGG**cttatag

gtcaatcatgttcttgtgaatggattt**AAAAAAAAGGGGGGG**gaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgt**AAAAAAAAGGGGGGG**caattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagtt**AAAAAAAAGGGGGGG**ctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcat**AAAAAAAAGGGGGGG**accgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagctt**AAAAAAAAGGGGGGG**a

Mutiramo svaki ubačeni šablon **AAAAAAAGGGGGGG** na 4 proizvoljno odabrane pozicije

atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa

tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag

gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa

za datu kolekciju, **AAAAAAGGGGGGG** je ***(k,d)-motiv***: *k*-gram koji se pojavljuje u svakoj sekvenci sa najviše *d* razlika

atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa

tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag

gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa

(k,d)-motiv: *k*-gram koji se pojavljuje u svakoj sekvenci sa najviše *d* razlika. U datoj kolekciji se *(k,d)-motiv* **AAAAAAGGGGGGG** pojavljuje kroz sledeće instance (motive):

AgAAgAAAGGttGGG
cAAtAAAAcGGcGGG
...
ActAAAAAGGaGcGG

***(k,d)-motiv* se ne mora pojaviti ni u jednoj sekvenci!**

Formulacija 1

Problem nalaženja motiva, formulacija 1.
Pronalaženje *(k,d)*-motiva u skupu niski.

- **Ulaz:** Skup niski *Dna* i celi brojevi *k* (dužina motiva) i *d* (maksimalni broj razlika).
- **Output:** Svi *(k,d)*-motivi u skupu *Dna*.

Napomena:

- Jedna kolekcija niski može imati više *(k,d)*-motiva – rešenje podrazumeva pronalaženje svih takvih niski

Formulacija 1

- Da bi se uporedili različiti algoritmi za rešavanje ovog problema, generisana je *benchmark* kolekcija podataka nad kojom su svi algoritmi testirani: 10 nukleotidnih niski dužine 600, *k=15*, *d=4*, traženi *(k,d)*-motiv je AAAAAAAAGGGGGGG

Pretraga grubom silom

- Za svaki k-gram ispitamo da li je (k,d)-motiv za dati skup niski

Koliko potencijalnih k-motiva postoji u nukleotidnoj niski? 4^k

Da li ima smisla da ispitamo svih 4^k k-grama?

- Svaki (k,d)-motiv se može razlikovati na najviše d pozicija od njegove instance u nekoj od sekvenci skupa *Dna*
- Stoga, možemo suziti prostor pretrage tako što ćemo generisati sve **takve** k-grame i proveriti da li su oni (k,d)-motivi

Rešenje 1: ispitivanje svih mogućih *(k,d)*-motiva za datu kolekciju

```
MotifEnumeration(Dna, k, d)
for each k-mer a in Dna
    • generate all possible k-mers a' differing from a
      by at most d mutations
    • for each such k-mer a'
        if a' is a (k,d)-mer in each sequence in Dna
            output a'
```

- Spor za velike *k* i *d*

- Da li će ovaj algoritam raditi nad realnim podacima?
  - Pretpostavlja se da svaka sekvenca iz skupa *Dna* sadrži instancu *(k,d)*-motiva. Ovaj uslov ne važi uvek. Na primer, u skupu sekvenci mogu se naći geni koji nisu pod kontrolom odgovarajućih transkripcionih faktora

Rešenje 2: ispitivanje *(k,d)*-motiva formiranih na osnovu parova podsekvenci

atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa

tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag

gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

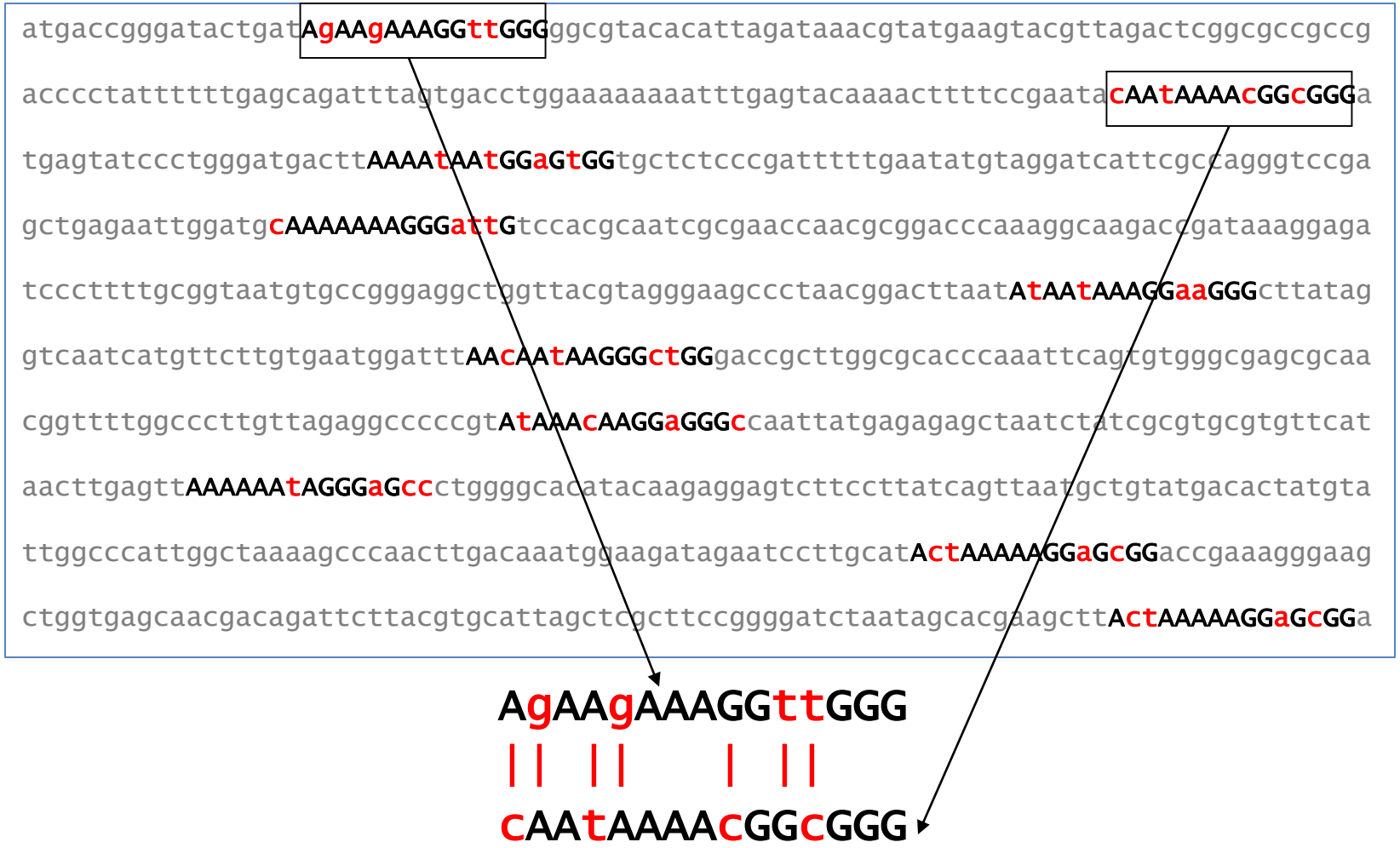
cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa

Rešenje 2: ispitivanje *(k,d)*-motiva formiranih na osnovu parova podsekvenci



Rešenje 2: ispitivanje *(k,d)*-motiva formiranih na osnovu parova podsekvenci

Zašto poređenje po parovima nije dobro?

AgAAgAAAGGttGGG *ubačena instanca 1*

| | || 4 razlike

AAAAAAAAGGGGGGG + *šablon*

| | | | 4 razlike

cAAtAAAAcGGGGGc *ubačena instanca 2*

Ovako pronađene ubačene instance mogu imati do **4 + 4 = 8 mutacija u 15-gramima!**

Rešenje 2: ispitivanje *(k,d)*-motiva formiranih na osnovu parova podsekvenci

- Koliko se često pojavljuju parovi 15-grama koji se razlikuju na najviše 8 pozicija?
- Na *benchmark* kolekciji ovim pristupom je pronađeno *nekoliko hiljada* parova k-grama koji su se razlikovali na manje od 8 pozicija

Formulacija 1

- Formulacija 1 predstavlja jednu apstrakciju biološkog problema nalaženja motiva
- Ipak, ona ima svoja ograničenja, pre svega zbog toga što ne uzima u obzir situacije kada neka sekvenca *ne sadrži* instancu (k,d)-motiva
- Kolekcije sekvenci koje se dobijaju iz laboratorija su šumovite (često se za neku sekvencu pogrešno pretpostavi da sadrži gen koji reguliše odgovarajući transkripcioni faktor) i u realnim primenama su ovakve situacije česte

Pregled

- **Regulatorni motivi**
  - **DNK kao izvor informacija**
  - **Genska ekspresija i cirkadijalni ritam**
  - **Formulacija 1**
  - **Formulacija 2**
  - Formulacija 3
  - Problem niske medijane
  - Pohlepna pretraga motiva
- Kako bacanje kockica pomaže u pronalaženju regulatornih motiva
  - Slučajna pretraga motiva
  - Gibsovo sempliranje
  - Pseudovrednosti

Reformulacija problema

- Pogodniji pristup problemu nalaženja motiva bi bila da se na neki način *ocene* pojedinačne instance (k,d)-motiva
- Jedna instanca motiva je *bolja* od druge instance ako je u nekom smislu *sličnija* (k,d)-motivu
- Ipak, (k,d)-motiv nam nije poznat pa ga ne možemo porediti sa pojedinačnim instancama
- Zbog toga ima smisla za datu kolekciju sekvenci iz svake od njih odabrati po jedan k-gram i dobijeni skup k-grama na neki način *oceniti*
- Ocena skupa k-grama treba da opisuje koliko su k-grami *međusobno slični*

Formulacija 2

- Pretpostavimo da je za datu kolekciju sekvenci iz svake sekvence na neki način izabran po jedan k-gram (potencijalni motiv) i da su svi tako odabrani k-grami objedinjeni u kolekciji *Motifs*

Motifs

T	**C**	**G**	**G**	**G**	**G**	G	**T**	**T**	**T**	T	T
C	**C**	**G**	**G**	T	**G**	**A**	C	**T**	**T**	A	**C**
A	**C**	**G**	**G**	**G**	**G**	**A**	**T**	**T**	**T**	T	**C**
T	T	**G**	**G**	**G**	**G**	**A**	C	**T**	**T**	T	T
A	A	**G**	**G**	**G**	**G**	**A**	C	**T**	**T**	**C**	**C**
T	T	**G**	**G**	**G**	**G**	**A**	C	**T**	**T**	**C**	**C**
T	**C**	**G**	**G**	**G**	**G**	**A**	**T**	**T**	C	A	T
T	**C**	**G**	**G**	**G**	**G**	**A**	**T**	**T**	C	**C**	T
T	A	**G**	**G**	**G**	**G**	**A**	A	C	**T**	A	**C**
T	**C**	**G**	**G**	**G**	T	**A**	**T**	A	A	**C**	**C**

- Kako bismo mogli da ocenimo ovu kolekciju motiva?

Formulacija 2

- *Konsenzus niska* za kolekciju *Motifs* je sastavljena od najzastupljenijih nukleotida u svakoj koloni kolekcije
- Ako kolekcija *Motifs* predstavlja kolekciju motiva, tada je *Consensus(Motifs) idealan motiv*

Motifs	T	C	G	G	G	G	g	T	T	T	t	t
	c	C	G	G	t	G	A	c	T	T	a	C
	a	C	G	G	G	G	A	T	T	T	t	C
	T	t	G	G	G	G	A	c	T	T	t	t
	a	a	G	G	G	G	A	c	T	T	C	C
	T	t	G	G	G	G	A	c	T	T	C	C
	T	C	G	G	G	G	A	T	T	c	a	t
	T	C	G	G	G	G	A	T	T	c	C	t
	T	a	G	G	G	G	A	a	c	T	a	C
	T	C	G	G	G	t	A	T	a	a	C	C
Consensus(Motifs)	T	C	G	G	G	G	A	T	T	T	C	C

Formulacija 2

- *Skor* za kolekciju *Motifs* je zbir brojeva nukleotida u svakoj koloni koji su različiti od najzastupljenijeg nukleotida
- Koja je najmanja a koja najveća vrednost skora?

	T	C	G	G	G	G	g	T	T	T	t	t
	c	C	G	G	t	G	A	c	T	T	a	C
	a	C	G	G	G	G	A	T	T	T	t	C
	T	t	G	G	G	G	A	c	T	T	t	t
Motifs	a	a	G	G	G	G	A	c	T	T	C	C
	T	t	G	G	G	G	A	c	T	T	C	C
	T	C	G	G	G	G	A	T	T	c	a	t
	T	C	G	G	G	G	A	T	T	c	C	t
	T	a	G	G	G	G	A	a	c	T	a	C
	T	C	G	G	G	t	A	T	a	a	C	C

Score(Motifs) 3+ 4+ 0+ 0+ 1+ 1+ 1+ 5+ 2+ 3+ 6+ 4=30

Formulacija 2

Problem pronalaženja motiva, formulacija 2. Za dati skup sekvenci, naći skup *k*-grama (po jedan iz svake sekvence) sa *minimalnim skorom* među svim mogućim skupovima *k*-grama iz datog skupa sekvenci.

- **Ulaz:** Skup sekvenci *Dna* i ceo broj *k*.
- **Izlaz:** Skup *k*-grama *Motifs*, po jedan iz svake sekvence skupa *Dna*, tako da je vrednost *Score(Motifs)* minimalna.

Rešenje

- Rešenje primenom grube sile je previše sporo

Pretraga motiva grubom silom

- t – broj niski, n – dužina svake niske
- $(n-k+1)^t$ načina da formiramo matricu motiva

($n-k+1$ načina da da odaberemo po jedan k-gram iz svake od t sekvenci)

- za računanje skora matrice neophodno je $k \cdot t$ koraka
- $k << n$ (u praksi, k je mnogo manje od n)

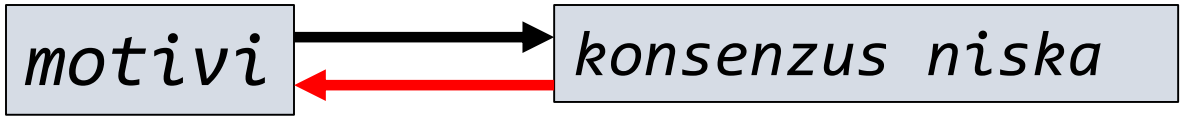
vremenska složenost: $O(n^t \cdot k \cdot t)$ – previše sporo!

- Potrebno je reformulisati problem i predložiti novo rešenje

Pregled

- **Regulatorni motivi**
  - **DNK kao izvor informacija**
  - **Genska ekspresija i cirkadijalni ritam**
  - **Formulacija 1**
  - **Formulacija 2**
  - **Formulacija 3**
  - Problem niske medijane
  - Pohlepna pretraga motiva
- Kako bacanje kockica pomaže u pronalaženju regulatornih motiva
  - Slučajna pretraga motiva
  - Gibsovo sempliranje
  - Pseudovrednosti

Formulacija 3



- U rešenju primenom grube sile smo ispitivali sve moguće kolekcije motiva i za svaku određivali konsenzus nisku
- U novom rešenju, ideja je obrnuta – ispitaćemo sve moguće niske, za svaku ćemo pretpostaviti da je konsenzus niska i u datoj kolekciji sekvenci ćemo pronaći kolekciju motiva koja joj *najviše odgovara*
- Da bismo izmerili *koliko* jednoj niski odgovara jedna kolekcija motiva, potrebna nam je funkcija koja to ocenjuje

Formulacija 3

	T	C	G	G	G	G	g	T	T	T	t	t	3
	c	C	G	G	t	G	A	c	T	T	a	C	4
	a	C	G	G	G	G	A	T	T	T	t	C	2
	T	t	G	G	G	G	A	c	T	T	t	t	4
Motifs	a	a	G	G	G	G	A	c	T	T	C	C	3
	T	t	G	G	G	G	A	c	T	T	C	C	2
	T	C	G	G	G	G	A	T	T	c	a	t	3
	T	C	G	G	G	G	A	T	T	c	C	t	2
	T	a	G	G	G	G	A	a	c	T	a	C	4
	T	C	G	G	G	t	A	T	a	a	C	C	3
													=30
pattern	T	C	G	G	G	G	A	T	T	T	C	C	

Hamingovo rastojanje od niske *pattern* do svakog motiva *i*:

$HammingDistance(pattern, Motif_i)$

Hamingovo rastojanje između dva *k*-grama je broj pozicija na kojima se razlikuju.

GATTCTCA
|| |
GACGCTGA

$d(GATTCTCA, GACGCTGA) = 3$

Formulacija 3

Hamingovim rastojanjem definisano je rastojanje između *dve niske karaktera*

Kako možemo definisati rastojanje između jedne niske (*k-mer*) i jedne kolekcije niski (*Motifs={Motif$_1$, …, Motif$_t$})?*

Uvodimo *d*(*k-mer*, *Motifs*) kao sumu Hamingovih rastojanja između *k*-grama i svakog motiva *Motif$_i$*:

$$d(k\text{-}mer, Motifs)=\sum_{i=1,t} HammingDistance(k\text{-}mer, Motif_i)$$

Šta je *d(Consensus(Motifs), Motifs)?*

Score(Motifs) = d(Consensus(Motifs), Motifs)

Motifs	T	C	G	G	G	G	g	T	T	T	t	t	3
	c	C	G	G	t	G	A	c	T	T	a	C	4
	a	C	G	G	G	G	A	T	T	T	t	C	2
	T	t	G	G	G	G	A	c	T	T	t	t	4
	a	a	G	G	G	G	A	c	T	T	C	C	3
	T	t	G	G	G	G	A	c	T	T	C	C	2
	T	C	G	G	G	G	A	T	T	c	a	t	3
	T	C	G	G	G	G	A	T	T	c	C	t	2
	T	a	G	G	G	G	A	a	c	T	a	C	4
	T	C	G	G	G	t	A	T	a	a	C	C	3
													=30
Consensus(Motifs)	T	C	G	G	G	G	A	T	T	T	C	C	
Score(Motifs)	3+	4+	0+	0+	1+	1+	1+	5+	2+	3+	6+	4=30	

#nepopularnih simbola (po redovima)

Hamingovo rastojanje od niske *Consensus* do svakog motiva *i*:

$d(Consensus, Motif_i)$

Najpopularniji simbol po koloni

#nepopularnih simbola (po kolonama)

Score(Motifs)=
nepopularnih simbola po kolonama =
nepopularnih simbola po redovima =
d(Consensus(Motifs), Motifs)

Formulacija 3

Problem pronalaženja motiva, formulacija 2. Naći skup *k*-grama *Motifs*, po jedan iz svake sekvence iz skupa sekvenci *Dna*, takav da je *Score(Motifs)* minimalan.

$$Score(Motifs) = d(Consensus(Motifs), Motifs)$$

Problem pronalaženja motiva, formulacija 3. Naći *k*-gram *Pattern* i skup *k*-grama *Motifs*, po jedan iz svake sekvence iz skupa sekvenci *Dna*, takav da je *d(Pattern, Motifs)* minimalan.

Umesto minimizacije funkcije po jednoj promenljivoj (*Score(Motifs)*), minimizujemo funkciju po dve promenljive (*d(Pattern, Motifs)*). Zar to nije teži problem?

Formulacija 3

Umesto minimizacije funkcije po jednoj promenljivoj (*Score(Motifs)*), minimizujemo funkciju po dve promenljive (*d(Pattern, Motifs)*). Zar to nije teži problem?

Promenljive *Pattern* i *Motifs* nisu nezavisne. Prilikom minimizacije funkcije *d*, nećemo uzimati u obzir sve moguće kolekcije *Motifs*, već samo one koje *najviše odgovaraju* datoj niski *Pattern*

Primer kolekcije *Motifs* za kolekciju niski *Dna*, za nisku AAA kao *Pattern*

Dna

ttaccttAAC
gATAtctgtc
ACGgcgttcg
ccctAAAgag
cgtcAGAggt

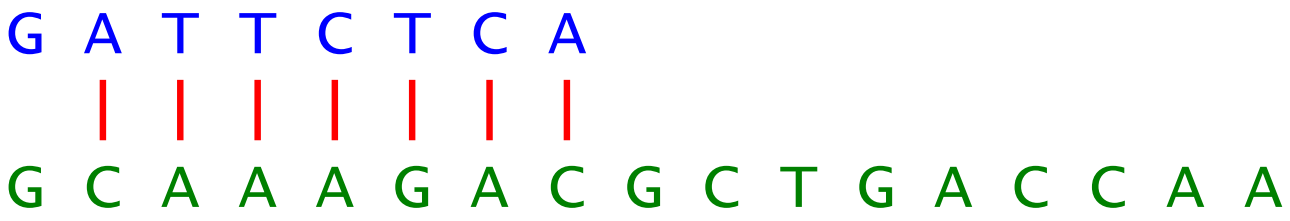
Kako da formalizujemo računanje skora između niske *Pattern* i kolekcije niski *Dna* pri čemu su niske iz kolekcije *Dna* različite dužine od niske *Pattern*?

Rastojanje između niski

- Ako su niske *iste* dužine, primenjujemo Hamingovo rastojanje
- Ako su niske *različite* dužine, potrebno je odabrati rastojanje koje bolje prikazuje razliku između ovih niski

Rastojanje između *k*-grama i (duže) niske

Rastojanje *d*(GATTCTCA, GCAAAGACGCTGACCAA) = ?



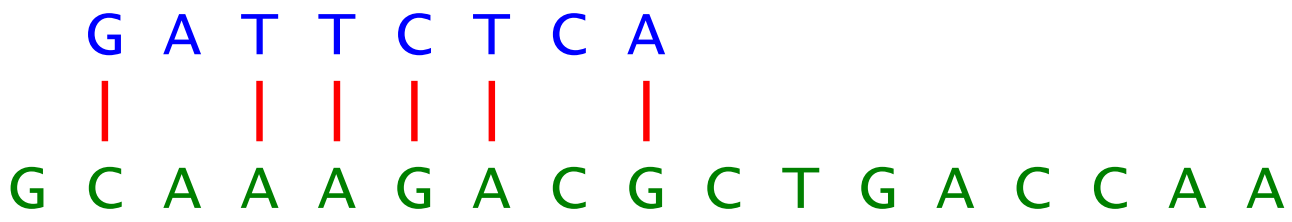
Rastojanje: **7**

d(*Pattern*, *String*):

najmanje Hamingovo rastojenje između *k*-grama *Pattern* i svih *k*-grama u niski *String*

Rastojanje između *k*-grama i (duže) niske

d(GATTCTCA, GCAAAGACGCTGACCAA) = ?



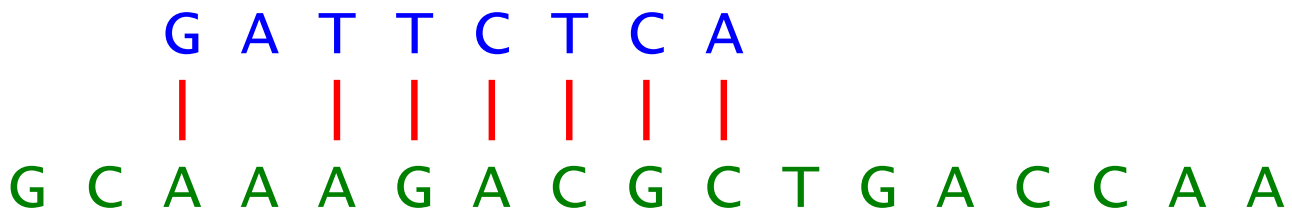
Rastojanje: 7 **6**

d(*Pattern*, *String*):

najmanje Hamingovo rastojenje između *k*-grama *Pattern* i svih *k*-grama u niski *String*

Rastojanje između *k*-grama i (duže) niske

d(GATTCTCA, GCAAAGACGCTGACCAA) = ?



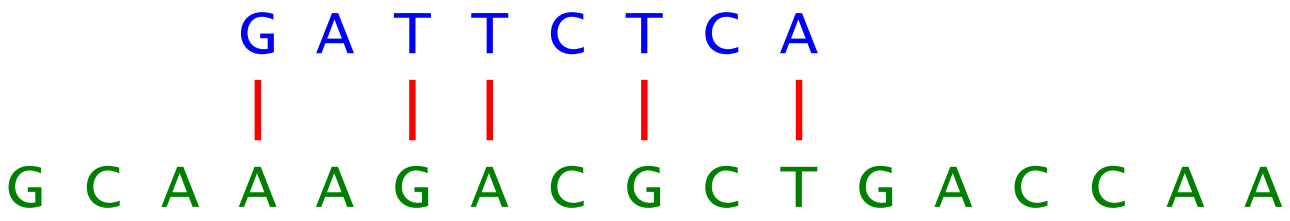
Rastojanje: 7 6 **7**

d(*Pattern*, *String*):

najmanje Hamingovo rastojenje između *k*-grama *Pattern* i svih *k*-grama u niski *String*

Rastojanje između *k*-grama i (duže) niske

d(GATTCTCA, GCAAAGACGCTGACCAA) = ?



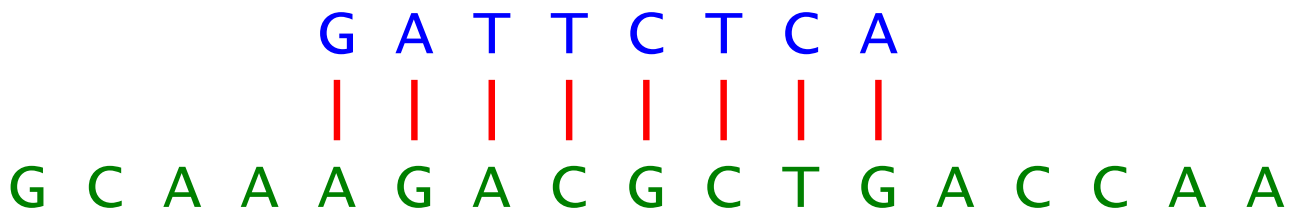
Rastojanje: 7 6 7 **5**

d(*Pattern*, *String*):

najmanje Hamingovo rastojenje između *k*-grama *Pattern* i svih *k*-grama u niski *String*

Rastojanje između *k*-grama i (duže) niske

d(GATTCTCA, GCAAAGACGCTGACCAA) = ?



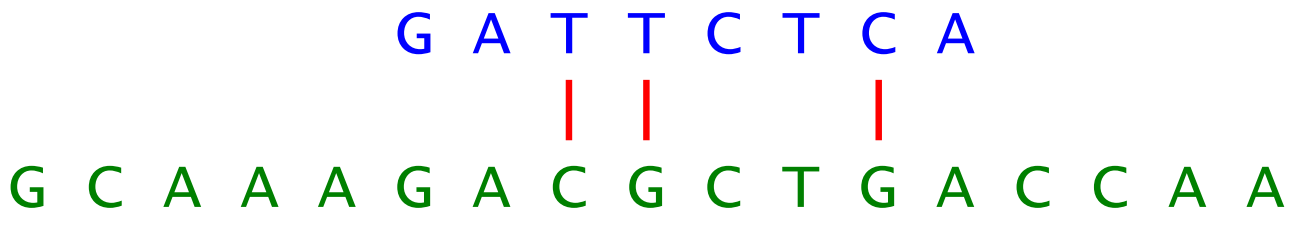
Rastojanje: 7 6 7 5 **8**

d(*Pattern*, *String*):

najmanje Hamingovo rastojenje između *k*-grama *Pattern* i svih *k*-grama u niski *String*

Rastojanje između *k*-grama i (duže) niske

d(GATTCTCA, GCAAAGACGCTGACCAA) = ?



Rastojanje: 7 6 7 5 8 **3**

d(*Pattern*, *String*):

najmanje Hamingovo rastojenje između *k*-grama *Pattern* i svih *k*-grama u niski *String*

Rastojanje između *k*-grama i (duže) niske

d(GATTCTCA, GCAAAGACGCTGACCAA) = ?

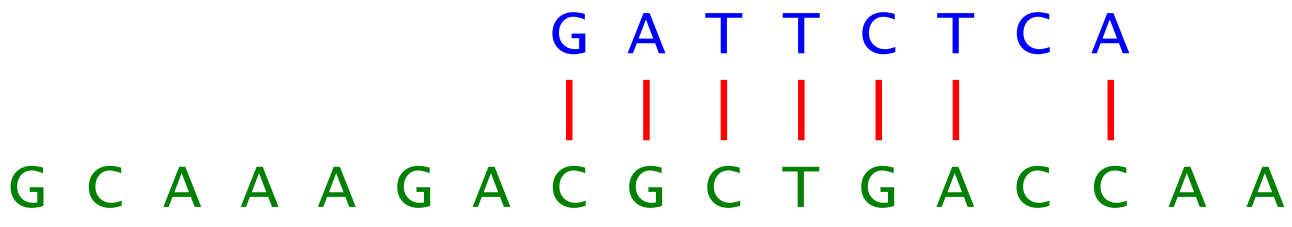
Rastojanje: 7 6 7 5 8 3 **8**

d(*Pattern*, *String*):

najmanje Hamingovo rastojenje između *k*-grama *Pattern* i svih *k*-grama u niski *String*

Rastojanje između *k*-grama i (duže) niske

d(GATTCTCA, GCAAAGACGCTGACCAA) = ?



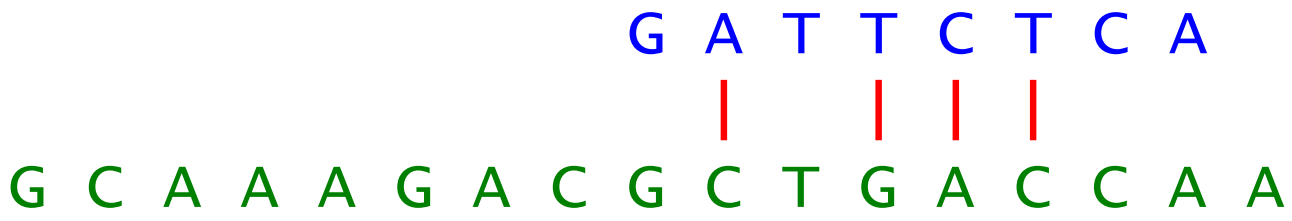
Rastojanje: 7 6 7 5 8 3 8 **7**

d(*Pattern*, *String*):

najmanje Hamingovo rastojenje između *k*-grama *Pattern* i svih *k*-grama u niski *String*

Rastojanje između *k*-grama i (duže) niske

d(GATTCTCA, GCAAAGACGCTGACCAA) = ?



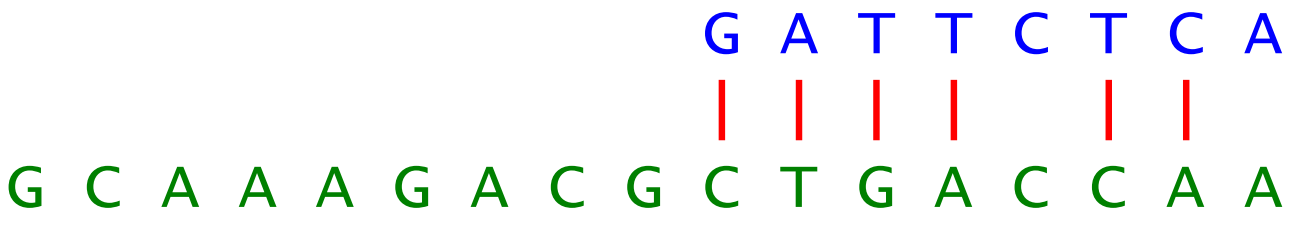
Rastojanje: 7 6 7 5 8 3 8 7 **4**

d(*Pattern*, *String*):

najmanje Hamingovo rastojenje između *k*-grama *Pattern* i svih *k*-grama u niski *String*

Rastojanje između *k*-grama i (duže) niske

d(GATTCTCA, GCAAAGACGCTGACCAA) = ?



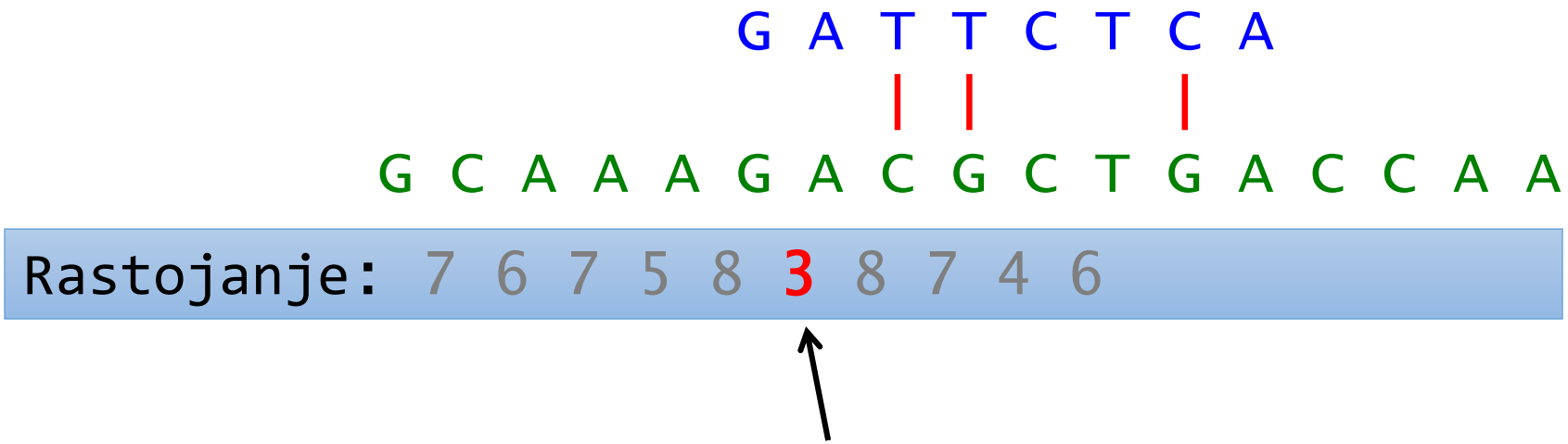
Rastojanje: 7 6 7 5 8 3 8 7 4 **6**

d(*Pattern*, *String*):

najmanje Hamingovo rastojenje između *k*-grama *Pattern* i svih *k*-grama u niski *String*

Rastojanje između *k*-grama i (duže) niske

d(GATTCTCA, GCAAAGACGCTGACCAA) = 3



d(*Pattern*, *String*):

najmanje Hamingovo rastojenje između *k*-grama *Pattern* i svih *k*-grama u niski *String*

Rastojanje između k-grama i skupa (dužih) niski

Rastojanje između *k*-grama *k-mer* i skupa niski *Dna* = {Dna_1, … , DNA_t}:

$$d(k\text{-}mer, Dna) = \sum_{i=1,\ldots,t} d(k\text{-}mer, Dna_i)$$

Pattern = AAA

ttaccttAAc 1
gAtAtctgtc 1
Acggcgttcg 2
ccctAAAgag 0
cgtcAgAggt 1

d(AAA, *Dna*)=5

Rastojanje između k-grama i skupa (dužih) niski

Problem pronalaženja motiva, formulacija 3. Naći *k*-gram *Pattern* i skup *k*-grama *Motifs*, po jedan iz svake sekvence iz skupa sekvenci *Dna*, takav da je *d(Pattern, Motifs)* minimalan.

Pattern = AAA

ttaccttAAc 1
gAtAtctgtc 1
Acggcgttcg 2
ccctAAAgag 0
cgtcAgAggt 1

d(AAA, *Dna*)=5

Za fiksiranu nisku *Pattern* se može odrediti kolekcija motiva *Motifs* iz kolekcije niski *Dna* za koju *d(Pattern, Motifs)* ima najmanju vrednost i to je ista vrednost koju ima i *d(Pattern, Dna)*.

Rastojanje između k-grama i skupa (dužih) niski

Problem pronalaženja motiva, formulacija 3. Naći *k*-gram *Pattern* i skup *k*-grama *Motifs*, po jedan iz svake sekvence iz skupa sekvenci *Dna*, takav da je *d(Pattern, Motifs)* minimalan.

Pattern = AAA

ttaccttAAc 1
gAtAtctgtc 1
Acggcgttcg 2
ccctAAAgag 0
cgtcAgAggt 1

d(AAA, *Dna*)=5

To znači da prilikom minimizacije skora *d(Pattern, Motifs)* ne vršimo minimizaciju po dve promenljive (*Pattern* i *Motifs*) već samo po jednoj promenljivoj *Pattern* dok kolekciju *Motifs* određujemo na osnovu promenljive *Pattern*

Rastojanje između k-grama i skupa (dužih) niski

Problem pronalaženja motiva, formulacija 3. Naći *k*-gram *Pattern* i skup *k*-grama *Motifs*, po jedan iz svake sekvence iz skupa sekvenci *Dna*, takav da je *d(Pattern, Motifs)* minimalan.

Pattern = AAA

ttaccttAAc 1
gAtAtctgtc 1
Acggcgttcg 2
ccctAAAgag 0
cgtcAgAggt 1

d(AAA, *Dna*)=5

Rešavanje ovog problema podrazumeva računanje skora *d(Pattern, Dna)* za sve moguće niske *Pattern* i određivanje one niske *Pattern* za koje je *d(Pattern, Dna)* minimalno.

Ovakva niska se naziva ***niska medijana***.

Pregled

- **Regulatorni motivi**
  - **DNK kao izvor informacija**
  - **Genska ekspresija i cirkadijalni ritam**
  - **Formulacija 1**
  - **Formulacija 2**
  - **Formulacija 3**
  - **Problem niske medijane**
  - Pohlepna pretraga motiva
- Kako bacanje kockica pomaže u pronalaženju regulatornih motiva
  - Slučajna pretraga motiva
  - Gibsovo sempliranje
  - Pseudovrednosti

Problem niske medijane

Problem niske medijane. Pronaći nisku medijanu.

- **Ulaz:** Skup sekvenci *Dna* i ceo broj *k*.
- **Izlaz:** *k*-gram *k-mer* koji minimizuje rastojanje *d(k-mer, Dna)*

```
MedianString(Dna, k)
    best-k-mer ← AAA · · · AA
    for each k-mer from AAA · · · AA to TTT · · · TT
        if d(k-mer, Dna) < d (best-k-mer, Dna)
            best-k-mer ← k-mer
    return(best-k-mer)
```

Vremenska složenost:

- $4^k \cdot n \cdot t \cdot k$ (za skup sekvenci *Dna* sa t sekvenci dužine n).
- za izračunavanje *d(k-mer, Dna)* potrebno je izračunati *d(k-mer, Dna$_i$)* t puta, po jednom *za svaku sekvencu*
- za izračunavanje *d(k-mer, Dna$_i$)* potrebno je izvršiti $n-k+1$ poređenja niski dužine k što je ukupno $(n-k+1) \cdot k$ poređenja karaktera ~ $n \cdot k$ zbog $k \ll n$

Problem niske medijane

Problem niske medijane. Pronaći nisku medijanu.

- **Ulaz:** Skup sekvenci *Dna* i ceo broj *k*.
- **Izlaz:** *k*-gram *k-mer* koji minimizuje rastojanje *d(k-mer, Dna)*

```
MedianString(Dna, k)
    best-k-mer ← AAA · · · AA
    for each k-mer from AAA · · · AA to TTT · · · TT
        if d(k-mer, Dna) < d (best-k-mer, Dna)
            best-k-mer ← k-mer
    return(best-k-mer)
```

Vremenska složenost:

- $4^k \cdot n \cdot t \cdot k$ (za skup sekvenci *Dna* sa t sekvenci dužine n).
- za izračunavanje *d(k-mer, Dna)* potrebno je izračunati *d(k-mer, Dna$_i$)* t puta, po jednom *za svaku sekvencu*
- za izračunavanje *d(k-mer, Dna$_i$)* potrebno je izvršiti $n-k+1$ poređenja niski dužine k što je ukupno $(n-k+1) \cdot k$ poređenja karaktera ~ $n \cdot k$ zbog $k \ll n$

Problem pronalaženja motiva = problem niske medijane

Problem pronalaženja motiva
versus
Problem niske medijane

Vremenska složenost:
$n^t \cdot k \cdot t$

Vremenska složenost:
$4^k \cdot n \cdot t \cdot k$

Iako je **MedianString** algoritam mnogo brži od **BruteForceMotifSearch**, i dalje je previše spor za velike k.

MedianString na *benchmark* skupu

- Razmotrimo kako se MedianString ponaša na *benchmark* skupu gde k iznosi 15 a ubačeni šablon je **AAAAAAAAGGGGGGG**
- Ispostavlja se da je MedianString algoritam prespor za k=15
- Kada je testiran na istom skupu za k=13, kao resenje vraca nisku **AAAAAtAGaGGGG** čiji je skor 29

Algoritam	K	Resenje	skor
MedianString	13	**AAAAAtAGaGGGG**	29
MedianString	15	previse spor	

Pregled

- **Regulatorni motivi**
  - **DNK kao izvor informacija**
  - **Genska ekspresija i cirkadijalni ritam**
  - **Formulacija 1**
  - **Formulacija 2**
  - **Formulacija 3**
  - **Problem niske medijane**
  - **Pohlepna pretraga motiva**
- Kako bacanje kockica pomaže u pronalaženju regulatornih motiva
  - Slučajna pretraga motiva
  - Kako bakterije hiberniraju?
  - Gibsovo sempliranje
  - Pseudovrednosti

Od skupa motiva do profilne matrice

Motifs

T	C	G	G	G	G	g	T	T	T	t	t
c	C	G	G	t	G	A	c	T	T	a	C
a	C	G	G	G	G	A	T	T	T	t	C
T	t	G	G	G	G	A	c	T	T	t	t
a	a	G	G	G	G	A	c	T	T	C	C
T	t	G	G	G	G	A	c	T	T	C	C
T	C	G	G	G	G	A	T	T	c	a	t
T	C	G	G	G	G	A	T	T	c	C	t
T	a	G	G	G	G	A	a	c	T	a	C
T	C	G	G	G	T	A	T	a	a	C	C

Count(Motifs)

A:	2	2	0	0	0	0	9	1	1	1	3	0
C:	1	6	0	0	0	0	0	4	1	2	4	6
G:	0	0	10	10	9	9	1	0	0	0	0	0
T:	7	2	0	0	1	1	0	5	8	7	3	4

broj nukleotida *i* u koloni *j*

Od skupa motiva do profilne matrice

Motifs

T	C	G	G	G	G	g	T	T	T	t	t
c	C	G	G	t	G	A	c	T	T	a	C
a	C	G	G	G	G	A	T	T	T	t	C
T	t	G	G	G	G	A	c	T	T	t	t
a	a	G	G	G	G	A	c	T	T	C	C
T	t	G	G	G	G	A	c	T	T	C	C
T	C	G	G	G	G	A	T	T	c	a	t
T	C	G	G	G	G	A	T	T	c	C	t
T	a	G	G	G	G	A	a	c	T	a	C
T	C	G	G	G	T	A	T	a	a	C	C

Profile(Motifs)

A:	.2	.2	0	0	0	0	.9	.1	.1	.1	.3	0
C:	.1	.6	0	0	0	0	0	.4	.1	.2	.4	.6
G:	0	0	1	1	.9	.9	.1	0	0	0	0	0
T:	.7	.2	0	0	.1	.1	0	.5	.8	.7	.3	.4

frekvencija nukleotida *i* u koloni *j*

Svaka kolona matrice *Profile* predstavlja otežanu četvorostranu kockicu kod koje su strane označene sa A, C, G i T. Stoga, profilna sekvenca predstavlja niz od *k* kockica.

Od skupa motiva do profilne matrice

Motifs

T	C	G	G	G	G	g	T	T	T	t	t
c	C	G	G	t	G	A	c	T	T	a	C
a	C	G	G	G	G	A	T	T	T	t	C
T	t	G	G	G	G	A	c	T	T	t	t
a	a	G	G	G	G	A	c	T	T	C	C
T	t	G	G	G	G	A	c	T	T	C	C
T	C	G	G	G	G	A	T	T	c	a	t
T	C	G	G	G	G	A	T	T	c	C	t
T	a	G	G	G	G	A	a	c	T	a	C
T	C	G	G	G	T	A	T	a	a	C	C

Profile(Motifs)

A:	.2	.2	0	0	0	0	.9	.1	.1	.1	.3	0
C:	.1	.6	0	0	0	0	0	.4	.1	.2	.4	.6
G:	0	0	1	1	.9	.9	.1	0	0	0	0	0
T:	.7	.2	0	0	.1	.1	0	.5	.8	.7	.3	.4

frekvencija nukleotida *i* u koloni *j*

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

Svaka kolona matrice *Profile* predstavlja otežanu četvorostranu kockicu kod koje su strane označene sa A, C, G i T. Stoga, profilna sekvenca predstavlja niz od *k* kockica.

Koja je verovatnoća da k bacanja ovakvih kockica generišu datu nisku?

Računanje skora *k*-grama sa profilnom matricom *Profile*

Neka je data profilna matrica *Profile*:

A	1/2	7/8	3/8	0	1/8	0
C	1/8	0	1/2	5/8	3/8	0
T	1/8	1/8	0	0	1/4	7/8
G	1/4	0	1/8	3/8	1/4	1/8

Verovatnoća konsenzus niske:

$Pr(AAACCT|Profile) = ???$

Računanje skora *k*-grama sa profilnom matricom *Profile*

Neka je data profilna matrica *Profile*:

A	**1/2**	**7/8**	**3/8**	0	1/8	0
C	1/8	0	1/2	**5/8**	**3/8**	0
T	1/8	1/8	0	0	1/4	**7/8**
G	1/4	0	1/8	3/8	1/4	1/8

Verovatnoća konsenzus niske:

$$Pr(AAACCT|Profile) = 1/2 \times 7/8 \times 3/8 \times 5/8 \times 3/8 \times 7/8 = 0.033646$$

Računanje skora *k*-grama sa profilnom matricom *Profile*

Neka je data profilna matrica *Profile*:

A	**1/2**	7/8	**3/8**	0	**1/8**	0
C	1/8	0	1/2	**5/8**	3/8	0
T	1/8	**1/8**	0	0	1/4	7/8
G	1/4	0	1/8	3/8	1/4	**1/8**

Verovatnoća druge niske:

Pr(AAACCT|*Profile*) =

1/2 x 7/8 x 3/8 x 5/8 x 3/8 x 7/8 = 0.033646

Pr(**ATACAG** |*Profile*) =

1/2 x **1/8** x **3/8** x **5/8** x **1/8** x **1/8** = 0.001602

Računanje skora *k*-grama sa profilnom matricom *Profile*

Neka je data profilna matrica *Profile*:

A	**1/2**	7/8	**3/8**	0	**1/8**	0
C	1/8	0	1/2	**5/8**	3/8	0
T	1/8	**1/8**	0	0	1/4	7/8
G	1/4	0	1/8	3/8	1/4	**1/8**

Verovatnoća druge niske:

$$Pr(AAACCT|Profile) = 1/2 \times 7/8 \times 3/8 \times 5/8 \times 3/8 \times 7/8 = 0.033646$$

Što je *k*-gram *k-mer* bliži konsenzus niski, to je verovatnoća Pr(*k*-mer| *Profile*) veća.

$$1/8 \times 1/8 = 0.001602$$

Koji je najverovatniji 6-gram za datu profilnu matricu *Profile* u niski CTATAAACCTTACAT?

A	**1/2**	**7/8**	**3/8**	0	1/8	0
C	1/8	0	1/2	**5/8**	**3/8**	0
T	1/8	1/8	0	0	1/4	**7/8**
G	1/4	0	1/8	3/8	1/4	1/8

Najverovatniji *k*-gram u sekvenci za datu profilnu matricu: onaj *k*-gram sa najvećom verovatnoćom Pr(*k-mer* | *Profile*) od svih k-grama u sekvenci.

6-mer	**Pr (6-mer|*Profile*)**	
CTATAAACCTTACAT	1/8 x 1/8 x 3/8 x 0 x 1/8 x 0	0
C**TATAAA**CCTTACAT	1/2 x 7/8 x 0 x 0 x 1/8 x 0	0
CT**ATAAAC**CTTACAT	1/2 x 1/8 x 3/8 x 0 x 1/8 x 0	0
CTA**TAAACC**TTACAT	1/8 x 7/8 x 3/8 x 0 x 3/8 x 0	0
CTATAAACCTTACAT	1/2 x 7/8 x 3/8 x 5/8 x 3/8 x 7/8	**.0336**
CTATA**AACCTT**ACAT	1/2 x 7/8 x 1/2 x 5/8 x 1/4 x 7/8	.0299
CTATAA**ACCTTA**CAT	1/2 x 0 x 1/2 x 0 1/4 x 0	0
CTATAAA**CCTTAC**AT	1/8 x 0 x 0 x 0 x 0 x 1/8 x 0	0
CTATAAAC**CTTACA**T	1/8 x 1/8 x 0 x 0 x 3/8 x 0	0
CTATAAACC**TTACAT**	1/8 x 1/8 x 3/8 x 5/8 x 1/8 x 7/8	.0004

```
GREEDYMOTIFSEARCH(Dna, k, t)
    BestMotifs ← motif matrix formed by first k-mers in each string from Dna
    for each k-mer Motif in the first string from Dna
        Motif_1 ← Motif
        for i = 2 to t
            form Profile from motifs Motif_1, ..., Motif_{i−1}
            Motif_i ← Profile-most probable k-mer in the i-th string in Dna
        Motifs ← (Motif_1, ..., Motif_t)
        if SCORE(Motifs) < SCORE(BestMotifs)
            BestMotifs ← Motifs
    return BestMotifs
```

Primer rada **GreedyMotifSearch**

Dna sa ugrađenim (4,1)-motivom ACGT

ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT

Inicijalizujemo *BestMotifs* (podebljano) i *BestScore*

ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT

U svakom koraku, fiksiramo jedan k-gram **prve** niske, Motif_1

1. **ttAC**CTtaac
2. t**tACC**Ttaac
3. tt**ACCT**taac
4. ttA**CCTt**aac
5. ttAC**CTta**ac
6. ttACC**Ttaa**c
7. ttACCT**taac**

Imaćemo ukupno 7 kandidata za najbolju kolekciju motiva

Motif_1 će biti prvi motiv kod svih kandidata. Ostale motive biramo na sledeći način:

GreedyMotifSearch

Kolekcija

ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT

Motifs
ttAC

Profile(Motifs)

A:	0	0	1	0
C:	0	0	0	1
G:	0	0	0	0
T:	1	1	0	0

ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT

Odredimo najverovatniji k-gram iz **druge** niske
gATGTctgtc
na osnovu *Profile(Motifs)*

gATGTctgtc
g**ATGT**ctgtc
gA**TGTc**tgtc
gAT**GTct**gtc
gATG**Tctg**tc
gATGT**ctgt**c
gATGTc**tgtc**

Motifs
ttAC
gATG

Ažuriramo *Profile(Motifs)*

A:	0	1/2	1/2	0
C:	0	0	0	1/2
G:	1/2	0	0	1/2
T:	1/2	1/2	1/2	0

GreedyMotifSearch

Kolekcija

ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT

Motifs
ttAC
gATG

Profile(*Motifs*)

A: 0 1/2 1/2 0
C: 0 0 0 1/2
G: 1/2 0 0 1/2
T: 1/2 1/2 1/2 0

ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT

Odredimo najverovatniji k-gram iz **treće** niske
ccgGCGTtag
na osnovu *Profile*(*Motifs*)

ccgGCGTtag
c**cgGC**GTtag
cc**gGCG**Ttag
ccg**GCGT**tag
ccgG**CGTt**ag
ccgGC**GTta**g
ccgGCG**Ttag**

Motifs
ttAC
gATG
ccgG

Ažuriramo *Profile*(*Motifs*)

A: 0 1/3 1/3 0
C: 1/3 1/3 0 1/3
G: 1/3 0 1/3 2/3
T: 1/3 1/3 1/3 0

GreedyMotifSearch

Kolekcija

ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT

Motifs
ttAC
gATG
ccgG

Profile(Motifs)

A:	0	1/3	1/3	0
C:	1/3	1/3	0	1/3
G:	1/3	0	1/3	2/3
T:	1/3	1/3	1/3	0

ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT

Odredimo najverovatniji k-gram iz **četvrte** niske
cactaACGAg
na osnovu *Profile(Motifs)*

cactaACGAg
c**acta**ACGAg
ca**ctaA**CGAg
cac**taAC**GAg
cact**aACG**Ag
cacta**ACGA**g
cactaA**CGAg**

Motifs
ttAC
gATG
ccgG
cact

Ažuriramo *Profile(Motifs)*

A:	0	1/2	1/4	0
C:	1/2	1/4	1/4	1/4
G:	1/4	0	1/4	1/2
T:	1/4	1/4	1/4	1/4

GreedyMotifSearch

Kolekcija

ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT

Motifs
ttAC
gATG
ccgG
cact

Profile(*Motifs*)

A:	0	1/2	1/4	0
C:	1/2	1/4	1/4	1/4
G:	1/4	0	1/4	1/2
T:	1/4	1/4	1/4	1/4

ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT

Odredimo najverovatniji k-gram iz **pete** niske
cgtcagAGGT
na osnovu *Profile*(*Motifs*)

cgtcagAGGT
c**gtca**gAGGT
cg**tcag**AGGT
cgt**cagA**GGT
cgtc**agAG**GT
cgtca**gAGG**T
cgtcag**AGGT**

Motifs
ttAC
gATG
ccgG
cact
cgtc

Izračunamo skor za kolekciju motiva koju smo odredili

Ako je manji od *BestScore*, ažuriramo *BestMotifs* i *BestScore*

Ponovimo postupak za ostale k-grame prve niske

GreedyMotif vs MedianString

- *GreedyMotif* je brži i može da pronađe motive dužine k=15, dok je MedianString postajao prespor za k>13
- Međutim, veća brzina donosi manju tačnost: pronalazi gtAAAtAgaGatGtG sa skorom 58 (ukupno rastojanje 58 od ostalih motiva), dok je traženi kanonski motiv AAAAAAAAGGGGGGG
- Može se u velikoj meri poboljšati primenom Laplasovog pravila (predstavljeno u nastavku) tako da pronađe AAAAAtAgaGGGGtt sa skorom 41
- Da li možemo naći još bolji algoritam?

Algoritam	K	Resenje	skor
MedianString	13	**AAAAAtAGaGGGG**	29
MedianString	15	previse spor	
GreedyMotif	15	**gtAAAtAgaGatGtG**	58
GreedyMotif Laplace	15	**AAAAAtAgaGGGGtt**	41

Pregled

- **Regulatorni motivi**
  - **DNK kao izvor informacija**
  - **Genska ekspresija i cirkadijalni ritam**
  - **Formulacija 1**
  - **Formulacija 2**
  - **Formulacija 3**
  - **Problem niske medijane**
  - **Pohlepna pretraga motiva**
- **Kako bacanje kockica pomaže u pronalaženju regulatornih motiva**
  - **Slučajna pretraga motiva**
  - Gibsovo sempliranje
  - Pseudovrednosti

Od motiva do profila

Za dati skup motiva *Motifs*, možemo konstruisati profilnu matricu

Profile(*Motifs*)

Motifs

T	C	G	G	G	G	g	T	T	T	t	t
c	C	G	G	t	G	A	c	T	T	a	C
a	C	G	G	G	G	A	T	T	T	t	C
T	t	G	G	G	G	A	c	T	T	t	t
a	a	G	G	G	G	A	c	T	T	C	C
T	t	G	G	G	G	A	c	T	T	C	C
T	C	G	G	G	G	A	T	T	c	a	t
T	C	G	G	G	G	A	T	T	c	C	t
T	a	G	G	G	G	A	a	c	T	a	C
T	C	G	G	G	T	A	T	a	a	C	C

Profile(Motifs)

A:	.2	.2	0	0	0	0	.9	.1	.1	.1	.3	0
C:	.1	.6	0	0	0	0	0	.4	.1	.2	.4	.6
G:	0	0	1	1	.9	.9	.1	0	0	0	0	0
T:	.7	.2	0	0	.1	.1	0	.5	.8	.7	.3	.4

Od motiva do profila do motiva

Za dati skup motiva *Motifs*, možemo konstruisati profilnu matricu

Profile(*Motifs*)

Za datu profilnu sekvencu *Profile* i skup sekvenci *Dna*, možemo konstruisati skup motiva

Motifs(*Profile*, *Dna*)

kao skup najverovatnijih *k*-grama za datu profilnu sekvencu *Profile* u svakoj sekvenci iz *Dna*.

Od motiva do profila do motiva

Za datu profilnu sekvencu *Profile* i skup sekvenci *Dna*, možemo konstruisati skup motiva *Motifs*(*Profile*, *Dna*) kao skup najverovatnijih *k*-grama za datu profilnu sekvencu *Profile* u svakoj sekvenci iz *Dna*.

6-mer	**Pr (6-mer|*Profile*)**	
CTATAAACCTTACAT	1/8 x 1/8 x 3/8 x 0 x 1/8 x 0	0
CTATAAACCTTACAT	1/2 x 7/8 x 0 x 0 x 1/8 x 0	0
CTATAAACCTTACAT	1/2 x 1/8 x 3/8 x 0 x 1/8 x 0	0
CTATAAACCTTACAT	1/8 x 7/8 x 3/8 x 0 x 3/8 x 0	0
CTATAAACCTTACAT	1/2 x 7/8 x 3/8 x 5/8 x 3/8 x 7/8	**.0336**
CTATAAACCTTACAT	1/2 x 7/8 x 1/2 x 5/8 x 1/4 x 7/8	.0299
CTATAAACCTTACAT	1/2 x 0 x 1/2 x 0 1/4 x 0	0
CTATAAACCTTACAT	1/8 x 0 x 0 x 0 x 0 x 1/8 x 0	0
CTATAAACCTTACAT	1/8 x 1/8 x 0 x 0 x 3/8 x 0	0
CTATAAACCTTACAT	1/8 x 1/8 x 3/8 x 5/8 x 1/8 x 7/8	.0004

Od motiva do profila do motiva do profila do…

Za dati skup motiva *Motifs*, možemo konstruisati profilnu matricu

Profile(*Motifs*)

Za datu profilnu sekvencu *Profile* i skup sekvenci *Dna*, možemo konstruisati skup motiva

Motifs(*Profile*, *Dna*)

kao skup najverovatnijih *k*-grama za datu profilnu sekvencu *Profile* u svakoj sekvenci iz *Dna*.

Iteriramo!

Motifs(*Profile*(*Motifs*(*Profile*(*Motifs*),*Dna*)),*Dna*)),*Dna*)…

Randomizovana pretraga motiva

```
RandomizedMotifSearch(Dna, k, t)
  randomly select k-mers Motifs = (Motif_1, … ,Motif_t) in each string from DNA
  bestMotifs ← Motifs
  while forever
    Profile ← Profile(Motifs)
    Motifs ← Motifs(Profile, Dna)
    if Score(Motifs) < Score(bestMotifs)
      bestMotifs ← Motifs
    else
      return(bestMotifs)
```

Primer rada RandomizedMotifSearch

Dna sa ugrađenim (4,1)-motivom ACGT	Odaberemo na slučajan način *Motifs* (podebljano)	*Motifs*	*Profile*(*Motifs*)

ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT

ttACCT**taac**
gAT**GTct**gtc
ccgGCGTtag
c**acta**ACGAg
cgtcag**AGGT**

t	a	a	c
G	T	c	t
c	c	g	G
a	c	t	a
A	G	G	T

A:	2/5	1/5	1/5	1/5
C:	1/5	2/5	1/5	1/5
G:	1/5	1/5	2/5	1/5
T:	1/5	1/5	1/5	2/5

.0016/ttAC .0016/tACC **.0128**/ACCT .0064/CCT**t** .0016/C**tta** .0016/T**taa** .0016/**taac**

.0016/gAT**G** **.0128**/AT**GT** .0016/T**GTc** .0032/**GTct** .0032/**Tct**g .0032/**ct**gt .0016/**t**gtc

.0064/**ccgG** .0036/**cgG**C .0016/**gG**CG **.0128**/**G**CGT .0032/CGTt .0016/Gtta .0016/Ttag

.0032/c**act** .0064/**acta** .0016/**cta**A .0016/**ta**AC .0032/**a**ACG **.0128**/ACGA .0016/CGAg

.0016/cgtc .0016/gtca .0016/tcag .0032/cag**A** .0032/ag**AG** .0032/g**AGG** **.0128**/**AGGT**

tt**ACCT**taac
g**ATGT**ctgtc
ccg**GCGT**tag
cacta**ACGA**g
cgtcag**AGGT**

Motifs (*Profile* (*Motifs*), *Dna*)

Kako slučajno odabrani motivi mogu da nas dovedu do traženog rešenja?

- U SubtleMotifProblemu, imamo 10 niski dužine 600 generisane na slučajan način u koje su ubačeni motivi dužine 15.
- Koja je verovatnoća da ćemo slučajnim izborom u jednom pogoditi motiv?
  - 1/(600-15+1) = 1/586
- A koja je verovatnoća da ćemo promašiti?
  - 1-1/586
- A u celoj kolekciji?
  - (1-1/586)^10 ~ 0.983
- Ako pokrenemo 1000 puta?
  - 983 puta ćemo promašiti a u preostalih 17 ćemo pogoditi bar nekog
- Zašto je važno da pogodimo bar jednog? Ako ne pogodimo motiv, pogodićemo deo niske generisan na slučajan način gde nam je svaki nukleotid generisan sa jednakom verovatnoćom na svakoj poziciji (0.25) i možemo očekivati uniformnu profilnu matricu (sve vrednosti 0.25) koja nas nigde ne usmerava
- Ako pogodimo bar jedan motiv, vrednosti u matrici neće biti sasvim slučajne i matrica će usmeravati našu pretragu

RandomizedMotifSearch vs GreedyMotifSearch

- Randomizovane algoritme možemo pokretati više puta i uzeti najbolje rešenje
- Za 100 000 pokretanja, vraća kolekciju motiva sa konsenzus niskom AAAAAAAAacaGGGG čiji je skor 43 (dosta dobar rezultat, za GibbsMotifSearch je 41)
- RandomizedMotifSearch takođe može da se koristi za pronalaženje dužih motiva
- Još jedno poboljšanje: GibbsMotifSearch

Algoritam	K	Resenje	skor
MedianString	13	**AAAAAtAGaGGGG**	29
MedianString	15	previse spor	
GreedyMotif	15	**gtAAAtAgaGatGtG**	58
GreedyMotif Laplace	15	**AAAAAtAgaGGGGtt**	41
RandomizedMotif	15	**AAAAAAAAacaGGGG**	43

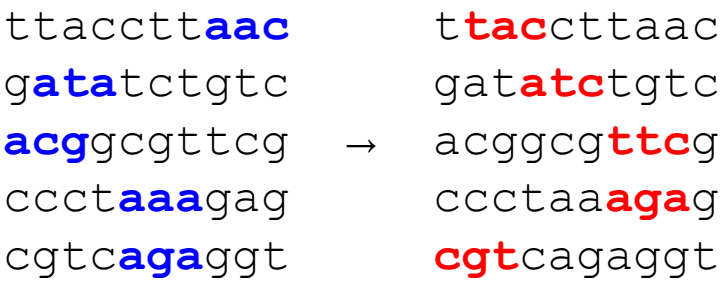
Pregled

- **Regulatorni motivi**
  - **DNK kao izvor informacija**
  - **Genska ekspresija i cirkadijalni ritam**
  - **Formulacija 1**
  - **Formulacija 2**
  - **Formulacija 3**
  - **Problem niske medijane**
  - **Pohlepna pretraga motiva**
- **Kako bacanje kockica pomaže u pronalaženju regulatornih motiva**
  - **Slučajna pretraga motiva**
  - **Gibsovo sempliranje**
  - Pseudovrednosti

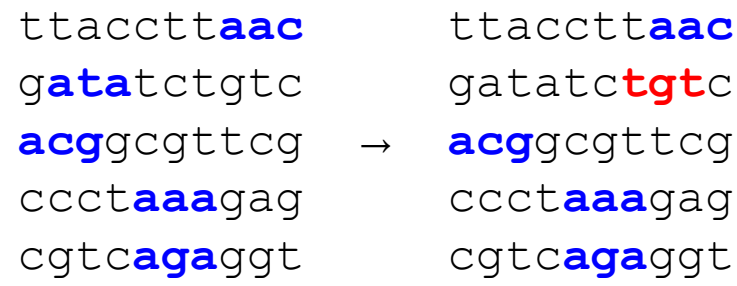
Gibsovo sempliranje

RandomizedMotifSearch može da menja sve *k*-grame u jednoj iteraciji i tako može da odbaci potencijalno tačan motiv.

Gibsovo sempliranje menja jedan *k*-gram u svakoj iteraciji.



RandomizedMotifSearch
(može da promeni **sve** *k*-grame u 1 iteraciji)



GibbsSampler
(menja **jedan** *k*-gram u 1 iteraciji)

Primer rada **GibbsSampler**

Slučajno odabrati skup motiva ***Motifs*** (**podebljano**)

ttACCT**taac**
gAT**GTct**gtc
ccgGCGTtag
c**acta**ACGAg
cgtcag**AGGT**

Slučajno odabrati *k*-gram iz ***Motifs*** i ukloniti ga

ttACCT**taac**
gAT**GTct**gtc
----
c**acta**ACGAg
cgtcag**AGGT**

Motifs nakon uklanjanja

t	**a**	**a**	**c**
G	**T**	**c**	**t**
----------			
a	**c**	**t**	**a**
A	**G**	**G**	**T**

Izaberemo novi k-gram iz obrisane niske kao onaj sa najvećom verovatnoćom

Count matrica

A:	**2**	1	1	1
C:	**0**	1	1	1
G:	1	1	1	**0**
T:	1	1	1	**2**

Izračunamo verovatnoće za sve k-grame iz obrisane niske

Profile matrica

A:	2/4	1/4	1/4	1/4
C:	**0**	1/4	1/4	1/4
G:	1/4	1/4	1/4	**0**
T:	1/4	1/4	1/4	2/4

0 (**ccgG**) **0** (**cgG**C) **0** (**gGCG**) **1/128** (**GCGT**) **0** (CGTt) **1/256** (GTta) **0** (Ttag)

Bacimo 7-stranu kockicu otežanu verovatnoćama *k*-grama

Primer rada **GibbsSampler**

Slučajno odabrati skup motiva ***Motifs*** (**podebljano**)

ttACCT**taac**
gAT**GTct**gtc
ccg**GCGT**tag
c**acta**ACGAg
cgtcag**AGGT**

Slučajno odabrati *k*-gram iz ***Motifs*** i ukloniti ga

ttACCT**taac**
gAT**GTct**gtc
---
c**acta**ACGAg
cgtcag**AGGT**

Motifs nakon uklanjanja

t	**a**	**a**	**c**
G	**T**	**c**	**t**
---	---	---	---
a	**c**	**t**	**a**
A	**G**	**G**	**T**

Izaberemo novi k-gram iz obrisane niske kao onaj sa najvećom verovatnoćom

Count matrica

A:	**2**	1	1	1
C:	**0**	1	1	1
G:	1	1	1	**0**
T:	1	1	1	**2**

Izračunamo verovatnoće za sve k-grame iz obrisane niske

Profile matrica

A:	2/4	1/4	1/4	1/4
C:	**0**	1/4	1/4	1/4
G:	1/4	1/4	1/4	**0**
T:	1/4	1/4	1/4	2/4

0 (**ccgG**) **0** (**cgGC**) **0** (**gGCG**) **1/128** (**GCGT**) **0** (CGTt) **1/256** (GTta) **0** (Ttag)

Bacimo 7-stranu kockicu otežanu verovatnoćama *k*-grama

Primer rada **GibbsSampler**

Slučajno odabrati skup motiva ***Motifs*** (**podebljano**)

ttACCT**taac**
gAT**GTct**gtc
ccg**GCGT**tag
c**acta**ACGAg
cgtcag**AGGT**

Slučajno odabrati *k*-gram iz ***Motifs*** i ukloniti ga

ttACCT**taac**
gAT**GTct**gtc
---
c**acta**ACGAg
cgtcag**AGGT**

Motifs nakon uklanjanja

t	**a**	**a**	**c**
G	**T**	**c**	**t**
---	---	---	---
a	**c**	**t**	**a**
A	**G**	**G**	**T**

iteriramo

Izaberemo obrisane sa najveć

1
1
0
2

Izračunamo verovatnoće za sve k-grame iz obrisane niske

Profile matrica

A:	2/4	1/4	1/4	1/4
C:	**0**	1/4	1/4	1/4
G:	1/4	1/4	1/4	**0**
T:	1/4	1/4	1/4	2/4

0 (**ccgG**) **0** (**cgGC**) **0** (**gGCG**) **1/128** (**GCGT**) **0** (CGTt) **1/256** (GTta) **0** (Ttag)

Bacimo 7-stranu kockicu otežanu verovatnoćama *k*-grama

Primer rada **GibbsSampler**

Slučajno odabrati skup motiva ***Motifs*** (**podebljano**)

ttACCT**taac**
gAT**GTct**gtc
ccgGCGTtag
c**acta**ACGAg
cgtcag**AGGT**

Slučajno odabrati *k*-gram iz ***Motifs*** i ukloniti ga

ttACCT**taac**
gAT**GTct**gtc
----
c**acta**ACGAg
cgtcag**AGGT**

Motifs nakon uklanjanja

t	**a**	**a**	**c**
G	**T**	**c**	**t**
----------			
a	**c**	**t**	**a**
A	**G**	**G**	**T**

Izaberemo novi k-gram iz obrisane niske kao onaj sa najvećom verovatnoćom

Count matrica

A:	**2**	1	1	1
C:	**0**	1	1	1
G:	1	1	1	**0**
T:	1	1	1	**2**

Izračunamo verovatnoće za sve k-grame iz obrisane niske

Profile matrica

A:	2/4	1/4	1/4	1/4
C:	**0**	1/4	1/4	1/4
G:	1/4	1/4	1/4	**0**
T:	1/4	1/4	1/4	2/4

Ima li smisla bacati dvostranu kockicu?

Bacimo 7-stranu kockicu otežanu verovatnoćama *k*-grama

Pregled

- **Regulatorni motivi**
  - **DNK kao izvor informacija**
  - **Genska ekspresija i cirkadijalni ritam**
  - **Formulacija 1**
  - **Formulacija 2**
  - **Formulacija 3**
  - **Problem niske medijane**
  - **Pohlepna pretraga motiva**
- **Kako bacanje kockica pomaže u pronalaženju regulatornih motiva**
  - **Slučajna pretraga motiva**
  - **Gibsovo sempliranje**
  - **Pseudovrednosti**

Laplasovo pravilo

U malim skupovima podataka uvek postoji šansa da se događaj koji je moguć neće dogoditi (npr nule u *Count* matrici).

Randomizovani algoritmi uvode **pseudovrednosti** koji povećavaju verovatnoće retkih događaja i eliminišu frekvencije jednake nuli zabeležene na osnovu iskustva

Ako ponovimo eksperiment koji može da rezultuje uspehom ili neuspehom ***n*** puta i dobijemo ***s*** puta uspeh, koja je verovatnoća da će i sledeći eksperiment imati uspešan ishod?

Pseudovrednosti

Ako su X_1, ..., X_{n+1} uslovno nezavisne slučajne logičke promenljive (neuspeh 0, uspeh 1), tada:

$$\Pr(X_{n+1}=1 \mid X_1+\ldots+X_n=s\ )=s/n$$

Pseudovrednosti

Ako su X_1, ..., X_{n+1} uslovno nezavisne slučajne logičke promenljive (neuspeh 0, uspeh 1), tada:

$$\Pr(X_{n+1}=1 \mid X_1+\ldots+X_n=s)=(s+1)/(n+2)$$

S obzirom da znamo da su dva ishoda moguća (uspeh i neuspeh), to znači da smo obavili $n+2$ eksperimenta umesto pretpostavljenih n.

Obavljeno je $n+2$ eksperimenta sa $s+1$ uspešnim ishodom. Dodate vrednosti su poznate pod nazivom **pseudovrednosti**.

Slučajno odabrati skup motiva ***Motifs*** (**podebljano**)

ttACCT**taac**
gAT**GTct**gtc
ccgGCGTtag
c**acta**ACGAg
cgtcag**AGGT**

Slučajno odabrati *k*-gram iz ***Motifs*** i ukloniti ga

ttACCT**taac**
gAT**GTct**gtc
---
c**acta**ACGAg
cgtcag**AGGT**

Motifs nakon uklanjanja

t	**a**	**a**	**c**
G	**T**	**c**	**t**
a	**c**	**t**	**a**
A	**G**	**G**	**T**

Izaberemo novi k-gram iz obrisane niske kao onaj sa najvećom verovatnoćom

Count matrica

A:	**2**	1	1	1
C:	**0**	1	1	1
G:	1	1	1	**0**
T:	1	1	1	**2**

Izračunamo verovatnoće za sve k-grame iz obrisane niske

Profile matrica

A:	2/4	1/4	1/4	1/4
C:	**0**	1/4	1/4	1/4
G:	1/4	1/4	1/4	**0**
T:	1/4	1/4	1/4	2/4

0 (**ccgG**) **0** (**cgG**C) **0** (**gGCG**) **1/128** (**GCGT**) **0** (CGTt) **1/256** (GTta) **0** (Ttag)

Slučajno odabrati skup motiva ***Motifs*** (**podebljano**)

ttACCT**taac**
gAT**GTct**gtc
ccg**GCGT**tag
c**acta**ACGAg
cgtcag**AGGT**

Slučajno odabrati *k*-gram iz ***Motifs*** i ukloniti ga

ttACCT**taac**
gAT**GTct**gtc
---
c**acta**ACGAg
cgtcag**AGGT**

Motifs nakon uklanjanja

t	**a**	**a**	**c**
G	**T**	**c**	**t**
---	---	---	---
a	**c**	**t**	**a**
A	**G**	**G**	**T**

Izaberemo novi k-gram iz obrisane niske kao onaj sa najvećom verovatnoćom

ažuriramo *Count* matricu sa pseudovrednostima

A:	**2+1**	1+1	1+1	1+1
C:	**0+1**	1+1	1+1	1+1
G:	1+1	1+1	1+1	**0+1**
T:	1+1	1+1	1+1	**2+1**

Ponovo izračunamo verovatnoće za sve 4-grame iz obrisane niske

ažuriramo profilnu matricu *Profile*

A:	3/8	2/8	2/8	2/8
C:	1/8	2/8	2/8	2/8
G:	2/8	2/8	2/8	1/8
T:	2/8	2/8	2/8	3/8

1/1024 (ccgG) 2/1024 (cgGC) 2/1024 (gGCG) **6/1024 (GCGT)** 3/1024 (CGTt) 4/1024 (GTta) 4/1024 (Ttag)

Bacamo sedmostranu kockicu umesto dvostranu

Slučajno odabrati skup motiva ***Motifs*** (**podebljano**)

ttACCT**taac**
gAT**GTct**gtc
ccg**GCGT**tag
c**acta**ACGAg
cgtcag**AGGT**

Slučajno odabrati *k*-gram iz ***Motifs*** i ukloniti ga

----
gAT**GTct**gtc
ccg**GCGT**tag
c**acta**ACGAg
cgtcag**AGGT**

Motifs nakon uklanjanja

G T c t
G C G T
a c t a
A G G T

Izaberemo novi k-gram iz obrisane niske kao onaj sa najvećom verovatnoćom

ažuriramo *Count* matricu sa pseudovrednostima

A:	**2+1**	1+1	1+1	1+1
C:	**0+1**	1+1	1+1	1+1
G:	1+1	1+1	1+1	**0+1**
T:	1+1	1+1	1+1	**2+1**

Ponovo izračunamo verovatnoće za sve 4-grame iz obrisane niske

ažuriramo profilnu matricu *Profile*

A:	3/8	2/8	2/8	2/8
C:	1/8	2/8	2/8	2/8
G:	2/8	2/8	2/8	1/8
T:	2/8	2/8	2/8	3/8

1/1024 (ccgG) 2/1024 (cgGC) 2/1024 (gGCG) **6/1024 (GCGT)** 3/1024 (CGTt) 4/1024 (GTta) 4/1024 (Ttag)

Bacamo sedmostranu kockicu umesto dvostranu

GibbsMotifSearch vs ostali

- GibbsMotifSearch pronalazi kolekciju motiva sa konsenzus niskom AAAAAAgAGGGGGGt čiji je skor 38 (bolji od ostalih)
- Mana: pošto pretražuje samo mali podskup od svih rešenja, može se dogoditi da se zaglavi u lokalnom minimumu
- Zbog toga je dobro uzeti u obzir veći broj izvršavanja

Algoritam	K	Resenje	skor
MedianString	13	**AAAAAtAGaGGGG**	29
MedianString	15	previse spor	
GreedyMotif	15	**gtAAAtAgaGatGtG**	58
GreedyMotif Laplace	15	**AAAAAtAgaGGGGtt**	41
RandomizedMotif	15	**AAAAAAAAacaGGGG**	43
GibbsMotif	15	**AAAAAAgAGGGGGGt**	38

Koji pristup odabrati?

- Nema pravila
- Nekada će biti bolji jedan, nekada drugi
- Analiza u sekciji „How Does Tuberculosis Hibernate to Hide from Antibiotics“

- Slajdovi pokrivaju poglavlje 2 knjige *Bioinformatics Algorithms: an Active Learning Approach*
- Sadržaj slajdova je preuzet sa zvaničnih prezentacija autora i dodatno prilagođen