Bioinformatikai Algoritmusok 9. GY

Assembly és átfedések keresése

Lehotay-Kéry Péter
lkp@caesar.elte.hu

Forditotta: Nyíri Tamás
nytuaai@gmail.com
http://people.inf.elte.hu/nytuaai

(Ben Langmead diasora alapján)

Assembly

Read-ek

+

Referen tenom

Próbáljuk kirakni a puzzle-t a referencia kép nélkül.

Input DNS

Assembly

A teljes genom “shotgun” szekvenciálása a DNS replikálásával és darabokra vágásával kezdődik.

(Azért nevezik “Shotgun”-nak mert véletlenszerű darabokat kapunk a végén, mintha kilőttük volna egy shotgun-ból.)

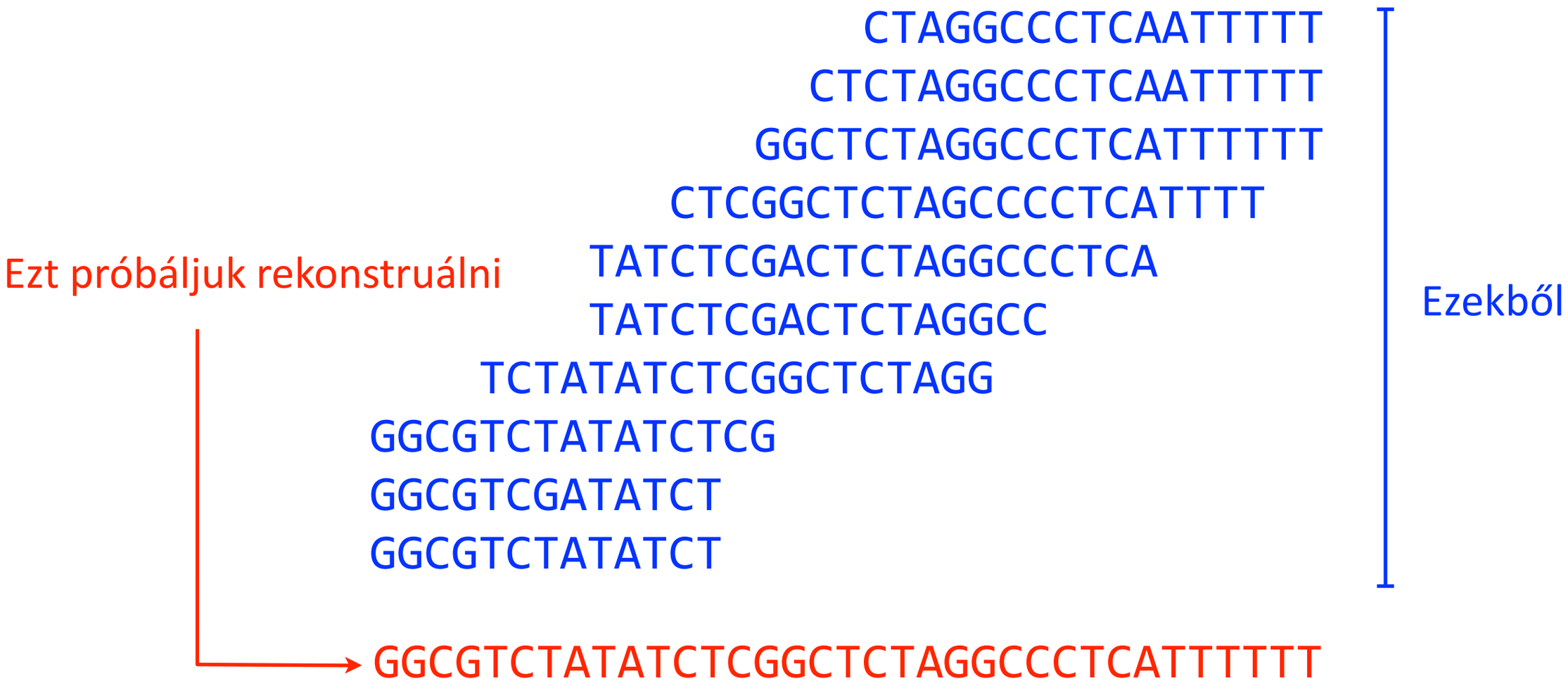
Input: GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

Másolat:
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

Darabok:
GGCGTCTA TATCTCGG CTCTAGGCCCTC ATTTTTT
GGC GTCTATAT CTCGGCTCTAGGCCCTCA TTTTTT
GGCGTC TATATCT CGGCTCTAGGCCCT CATTTTTT
GGCGTCTAT ATCTCGGCTCTAG GCCCTCATTTTTT

Assembly

Feltesszük, hogy a szekvenciálás elegendően sok replikát állít elő, hogy majdnem minden genom pozícióra sok másolatunk keletkezik…



Assembly

...de nem tudjuk melyik honnan jött.

CTAGGCCCTCAATTTTT
GGCGTCTATATCT
CTCTAGGCCCTCAATTTTT
TCTATATCTCGGCTCTAGG
GGCTCTAGGCCCTCATTTTTT
CTCGGCTCTAGCCCCTCATTTT
TATCTCGACTCTAGGCCCTCA
GGCGTCGATATCT
TATCTCGACTCTAGGCC
GGCGTCTATATCTCG

Ezekből

Ezt próbáljuk rekonstruálni

GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

Assembly

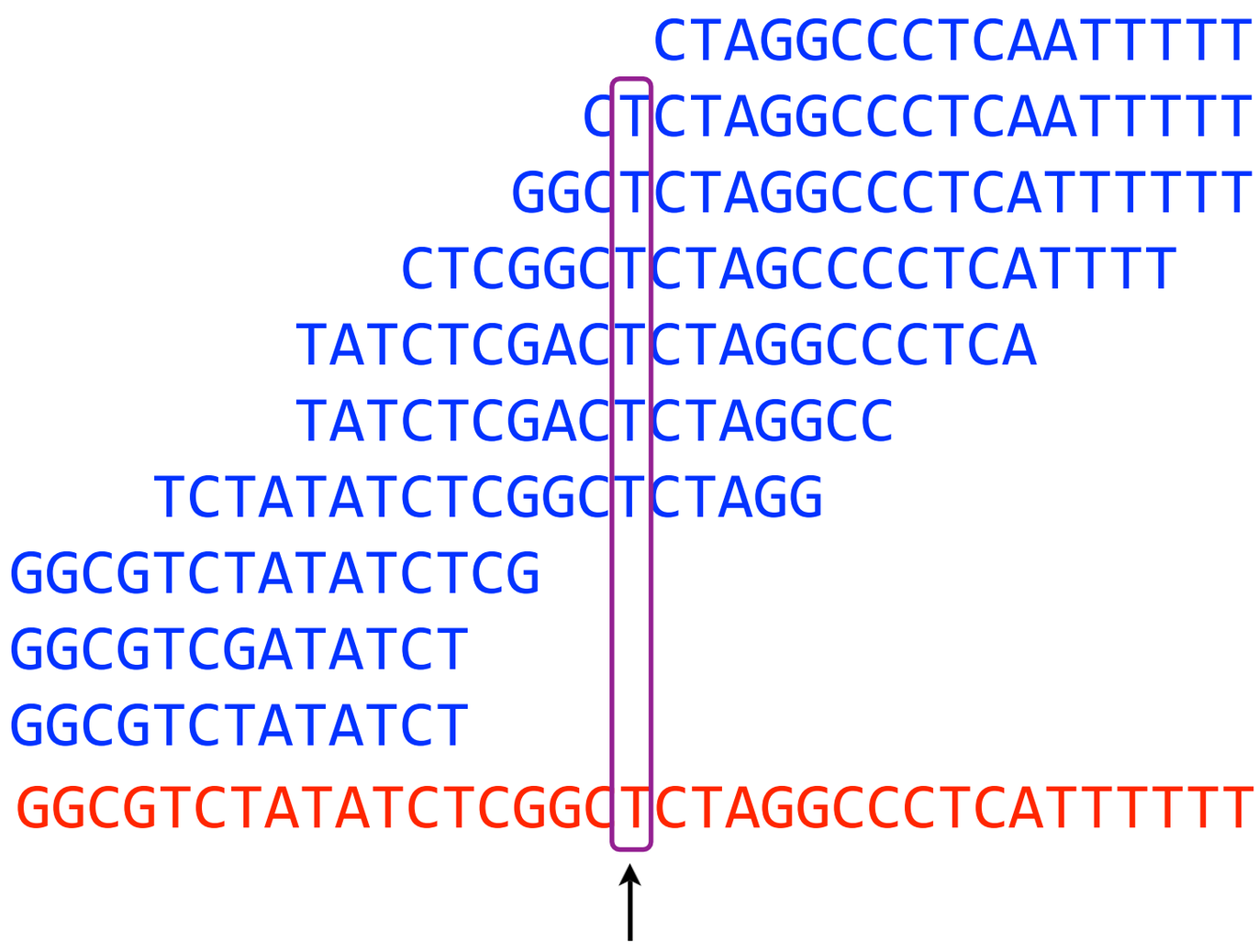
Átlagos lefedettség: Az eredeti gén minden pozíciójára számoljuk meg hány darab read tartalmazza azt a pozíciójú nukleotidot és átlagoljuk.

```
                                CTAGGCCCTCAATTTTT
                              CTCTAGGCCCTCAATTTTT
                            GGCTCTAGGCCCTCATTTTTT
                              CTCGGCTCTAGCCCCTCATTTT
                        TATCTCGACTCTAGGCCCTCA                 177 nukleotid
                        TATCTCGACTCTAGGCC
                     TCTATATCTCGGCTCTAGG
                GGCGTCTATATCTCG
                GGCGTCGATATCT
                GGCGTCTATATCT
                GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT            35 nukleotid
```

Átlagos lefedettség = 177 / 35 ≈ 7x

Assembly

Lefedettség adott helyen: Az eredeti genom egy bizonyos pozícióját hány read tartalmazza:



Assembly

Minél nagyobb a hasonlóság az egyik read vége és egy másik eleje között...

```
        TATCTCGACTCTAGGCC
        ||||||| |||||||
TCTATATCTCGGCTCTAGG
```

...annál valószínűbb hogy ezek az eredeti genom egymást fedő részeiről származnak:

```
        TATCTCGACTCTAGGCC
TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
```

Assembly

Tegyük fel hogy tényleg az eredeti genom egymást fedő részeiről származnak. Miért lehet bennük mégis eltérés?

TATCTCGACTCTAGGCC
||||||| |||||||
TCTATATCTCGGCTCTAGG

1. Szekvenciálási hiba
2. A kromoszóma öröklött másolataiban lévő különbség

   Pl. az emberek diploidok, azaz két másolattal rendelkeznek minden kromoszómából (apai és anyai). Ezek a másolatok különbözhetnek:

anyai read: TATCTCGACTCTAGGCC

apai read: TCTATATCTCGGCTCTAGG

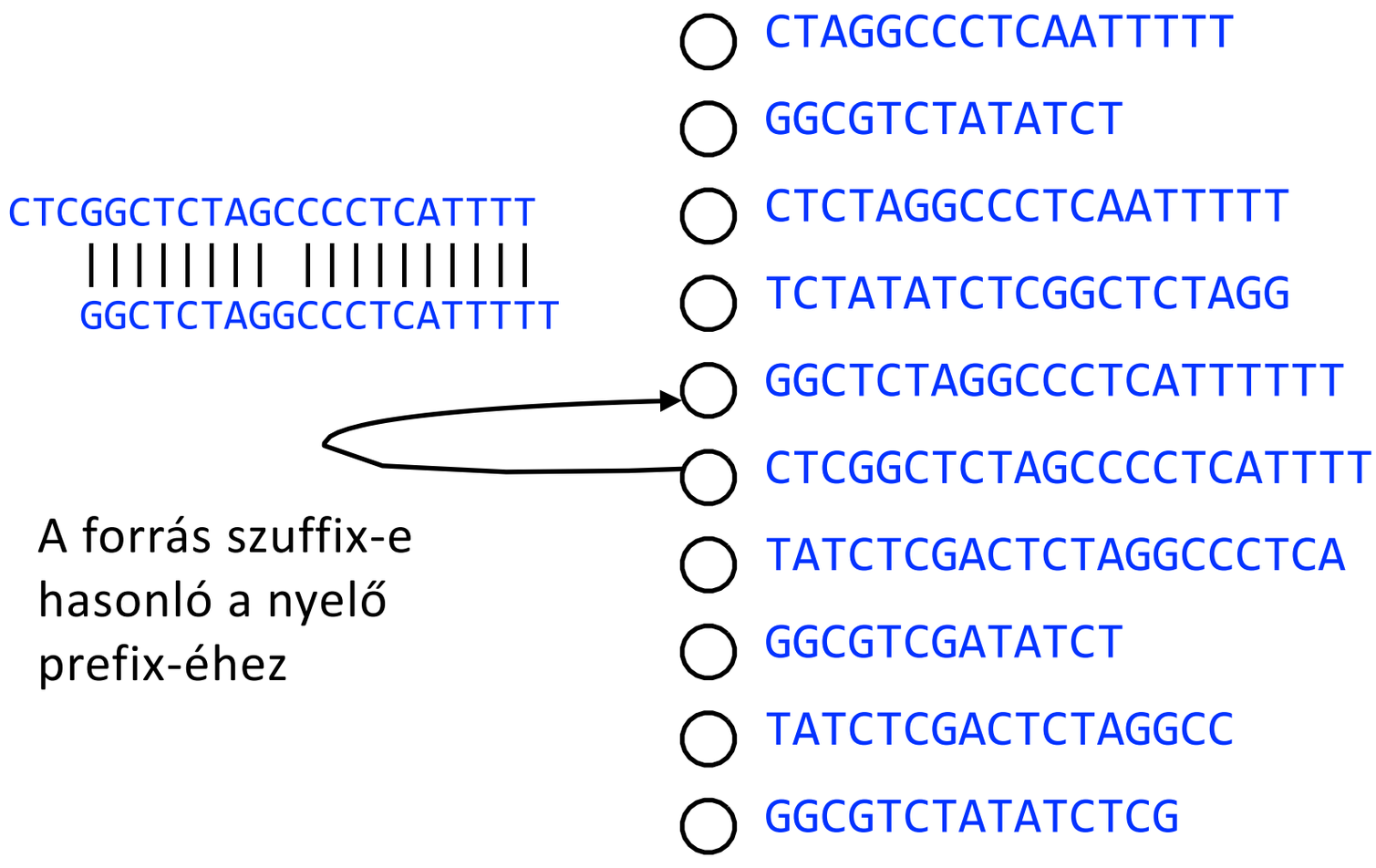
anyai szekvencia: TCTATATCTCGACTCTAGGCC

apai szekvencia: TCTATATCTCGGCTCTAGGCC

Az egyszerűség kedvéért ezt most ignoráljuk, de a valódi eszközöknek figyelembe kell venniük.

Átfedések

Megtalálhatjuk az összes átfedést úgy hogy készítünk egy irányított gráfot ahol a csúcsok a read-ek és az irányított élek az egymást fedő read-eket reprezentáló csúcsok között mennek.



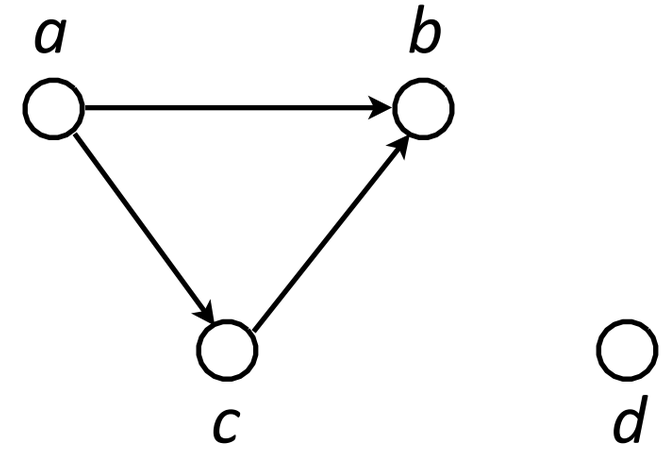
Irányított gráfok - ismétlés

Az irányított gráfot egy *G*(*V*, *E*) rendezett párral reprezentáljuk, ahol V a csúcsok halmaza és *E az irányított élek halmaza.*

Egy irányított él csúcsok rendezett párjaként reprezentálható.
Az első a forrás, a második a nyelő.

A csúcsot grafikusan körrel reprezentáljuk.

A két csúcs között futó élet pedig nyílként mely az egyik körből a másikba mutat.



$V = \{ a, b, c, d \}$

$E = \{ (a, b), (a, c), (c, b) \}$

Forrás Nyelő

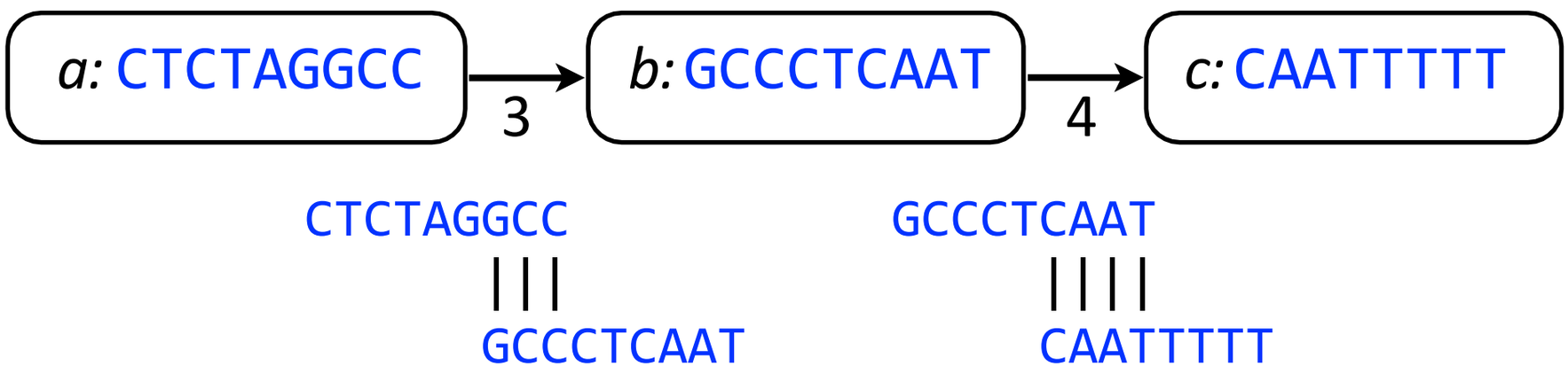
Átfedési gráf

Alul: átfedési gráf, ahol az átfedés egy legalább 3 karakter hosszú szuffix/prefix illeszkedés.

A csúcsok read-ek, az irányított élek átfedések a forrás szuffix-e és a nyelő prefix-e között.

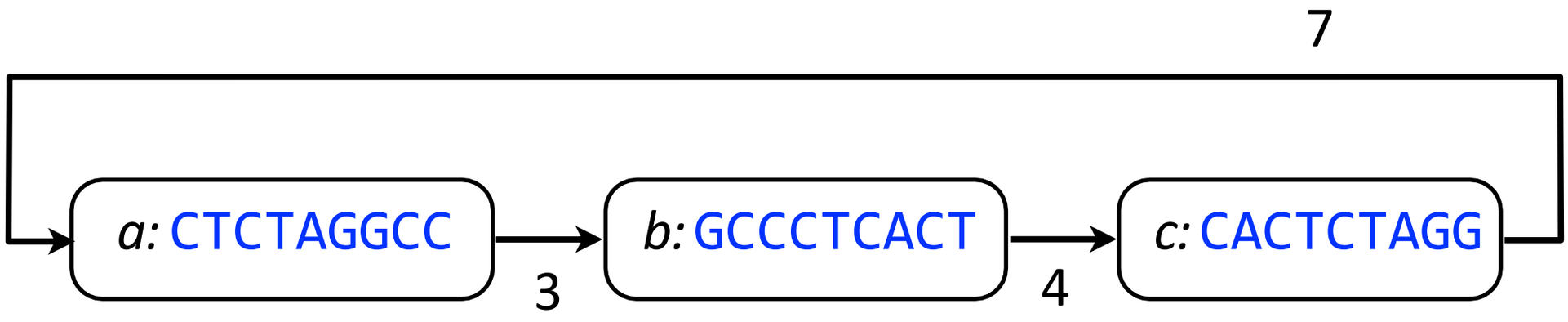
Csúcsok (read-ek): { *a:* CTCTAGGCC, *b:* GCCCTCAAT, *c:* CAATTTTT }

Élek (átfedések): { (*a*, *b*), (*b*, *c*) }



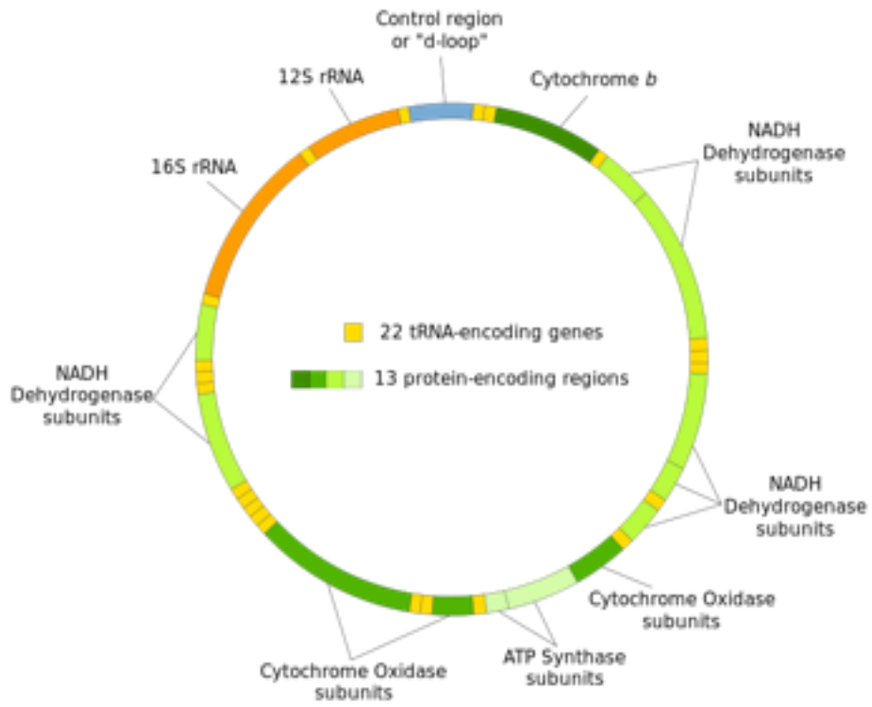
Átfedési gráf

Az átfedési gráf tartalmazhat köröket is. *A kör egy olyan út melynek az első és az utolsó* csúcsa megegyezik.



Kör keletkezhet amiatt hogy DNS string maga is körkörös. (pl. a baktérium genomok gyakran körkörösek, illetve a mitokondriális DNS-ek is azok.

Körök keletkezhetnek a DNS-ben lévő ismétlődő részek miatt is, amint látni fogjuk később.



Átfedések keresése

Hogyan építjük fel az átfedési gráfot?

Miből áll egy átfedés?

Egyelőre nevezzük „átfedés”-nek azt amikor X egy ≥ l hosszú szuffix-e k helyen pontosan egyezik Y ugyanolyan hosszú prefixével adott k számra.

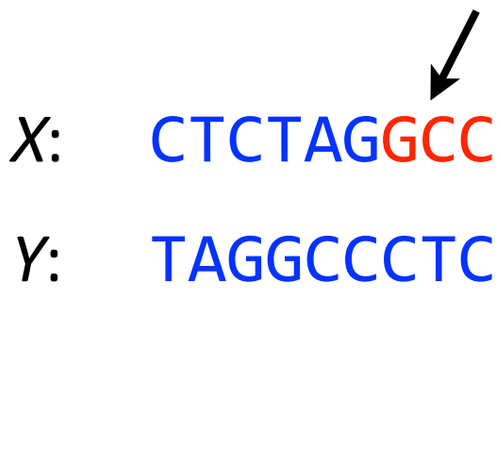
Átfedések keresése

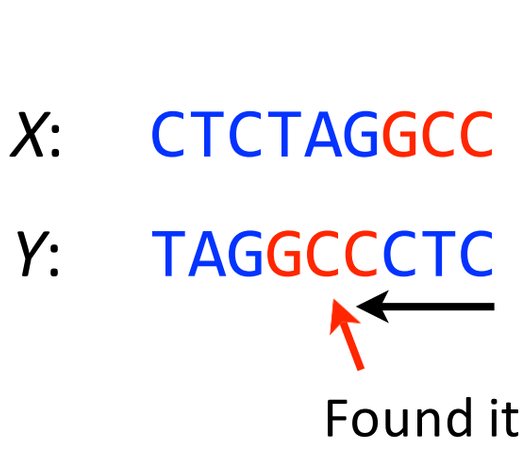
Átfedés: X egy l hosszú szuffix-e egyezik Y egy l hosszú prefixével ahol l adott.

Egyszerű ötlet: keressük meg Y prefixében X k-hosszú szuffix-ét. Terjesszük ki a találatokat balra hogy biztosítsuk hogy az Y egész prefix-e egyezik.

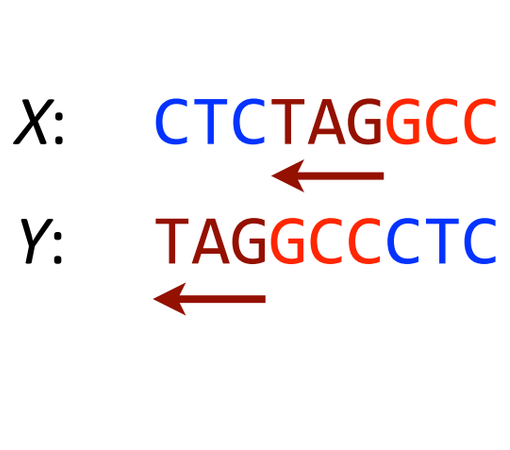
Legyen $k = 3$

Keressünk erre Y-ban, jobbról balra



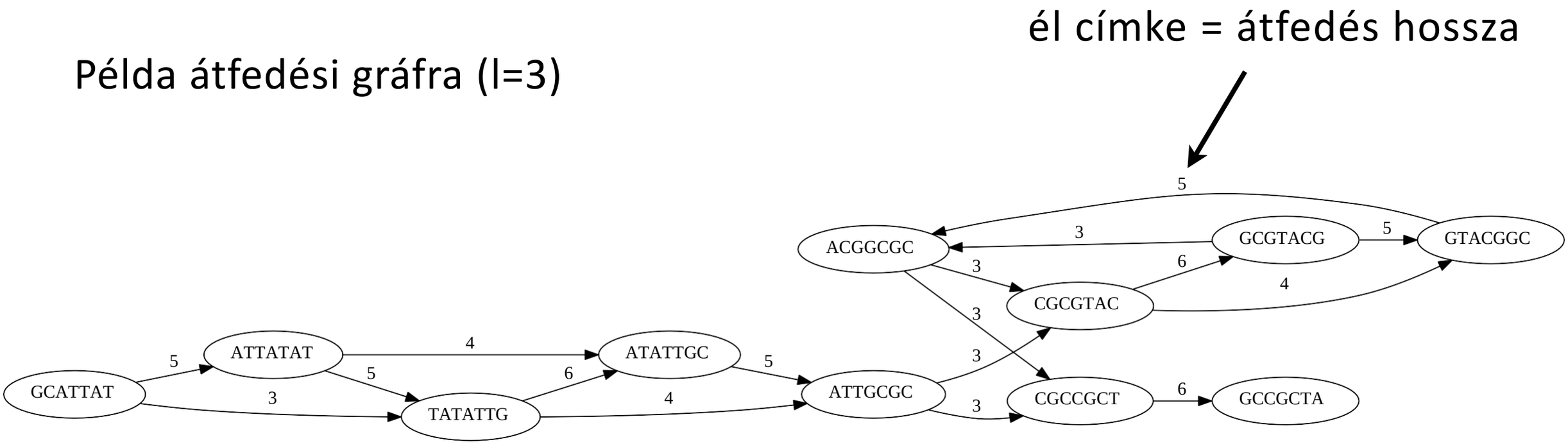


Terjesszük ki balra; ebben az esetben megerősítjük, hogy Y 6 hosszú prefix-e egyezik X ugyanolyan hosszú szuffix-évél.



Átfedések keresése

Példa átfedési gráfra (l=3)



Eredeti string: GCATTATATATTGCGCGTACGGCGCCGCTACA