Algoritmi v bioinformatiki
Problem rekonstrukcije zaporedij

Martin Milanič
martin.milanic@upr.si
UP FAMNIT

9. maj 2025

Literatura za to poglavje:

- **Compeau-Pevzner**, Bioinformatics Algorithms: An Active Learning Approach, poglavje 3
  How do we Assemble Genomes?

- **Jones-Pevzner**, An Introduction to Bioinformatics Algorithms, poglavje 8
  Graph Algorithms

Spomnimo: nekaj osnov o grafih

Sprehod v grafu $G = (V, E)$ je zaporedje

$$x_0 e_1 x_1 \cdots e_\ell x_\ell \,,$$

kjer je $x_i \in V$ in $e_i = x_{i-1}x_i \in E$.

Opombe:

- $\ell =$ število povezav
- običajno pišemo kar $x_0 x_1 \cdots x_k$ (izpustimo torej $e_i$je) (ne za multigrafe)

Sled je sprehod, v katerem so vse povezave različne.

Pot je sprehod, v katerem so vse točke različne.

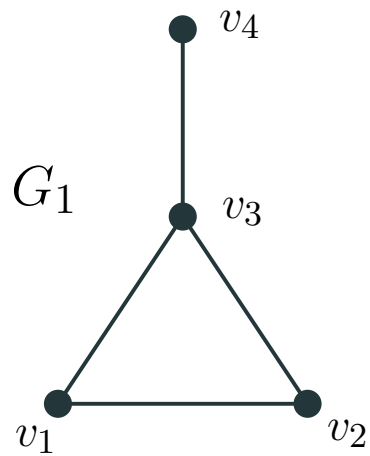
Sprehod je **sklenjen**, če je $x_\ell = x_0$.

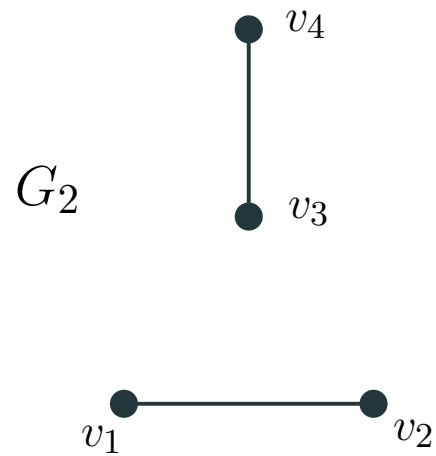
Cikel: taka sklenjena sled, da je $x_0 x_1 \cdots x_{\ell-1}$ pot.

Pri definicijah poti in ciklov v **digrafih** upoštevamo tudi usmerjenost povezav.

Graf je **povezan**, če za vsaki dve točki $u, v \in V(G)$ obstaja pot med njima.

Zgled: graf G_1 je povezan, graf G_2 pa ne



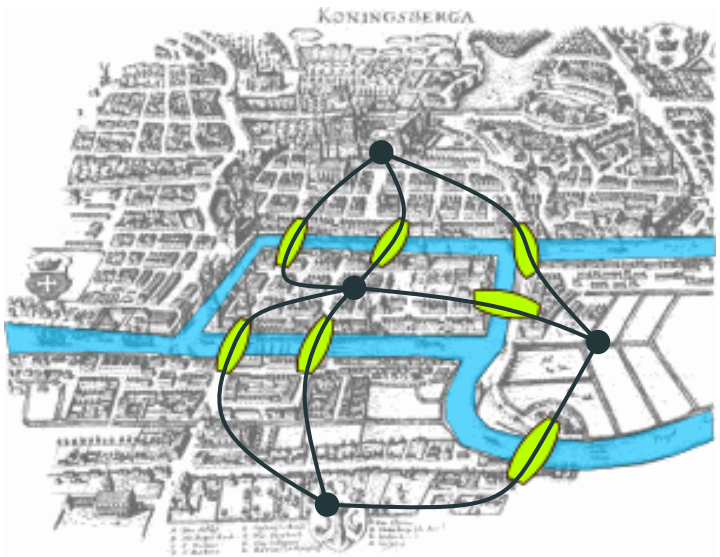


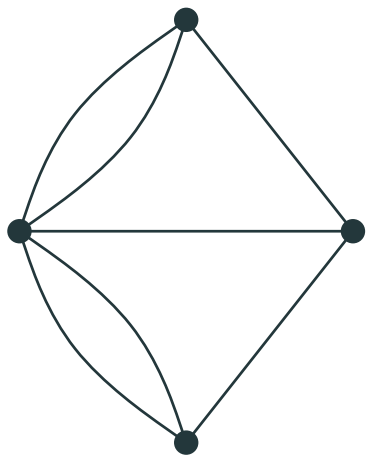
Eulerjeva sled v grafu je sled, ki gre skozi vsako povezavo.

(Ker se povezave v sledi ne morejo ponoviti, gre skozi vsako povezavo natanko enkrat.)

Eulerjev obhod je sklenjena Eulerjeva sled.

Izvor: 7 mostov Königsberga





(slika prirejena po sliki iz wikipedije)

Izrek (Euler, 1736)

(i) *Multigraf G brez izoliranih točk ima Eulerjev obhod*
$\Leftrightarrow$ *G je povezan in ima vse točke sode stopnje.*

(ii) *Multigraf G brez izoliranih točk ima nesklenjeno Eulerjevo sled*
$\Leftrightarrow$ *G je povezan in ima natanko dve točki lihe stopnje.*

Očitna potrebna pogoja sta torej tudi zadostna.

hamiltonska pot: pot, ki gre skozi vse točke grafa

hamiltonski cikel: cikel, ki gre skozi vse točke grafa

Ali lahko podamo kakšen preprost potreben in zadosten pogoj za obstoj hamiltonskega cikla v grafu?

Najverjetneje ne, saj je odločitveni problem, ali dan graf ima hamiltonski cikel **NP-poln**.

Enako velja za hamiltonske poti.

Za **digrafe** so definicije Eulerjevih sledi in obhodov in hamiltonskih poti in ciklov točno takšne, kot bi pričakovali.

Analog Eulerjevega izreka za digrafe je naslednji.

Digraf D je **šibko povezan**, če je neusmerjen graf, ki ga iz digrafa D dobimo tako, da pozabimo na usmerjenosti povezav, povezan.

Izrek

Digraf D brez izoliranih točk ima Eulerjev obhod

$\Leftrightarrow$ *D je šibko povezan in vhodna in izhodna stopnja vsake točke sta enaki.*

Sekvenciranje

Sekvenciranje = določitev zaporedja nukleotidov v dani molekuli DNA / RNA oz. zaporedja aminokislin v danem proteinu.

Na voljo je veliko biotehnoloških metod sekvenciranja:

- Sanger-Nicklen-Coulson (Cambridge, Velika Britanija) 1977 / Maxam–Gilbert (Cambridge, MA, ZDA) 1977
- DNA mikromreže (sekvenciranje s hibridizacijo)
- Nova generacija visokozmogljivostnega sekvenciranja
  454, Illumina, ABI (Applied Biosystems), Pacific Biosciences ...

Metode sekvenciranja pogosto najprej določijo **množico podzaporedij neznanega zaporedja**, ki se med seboj prekrivajo.

Pojavi se problem, **kako ugotoviti celotno zaporedje**.

Ena možnost je, da vsa podzaporedja staknemo drugega za drugim.

- Na ta način dobimo dolgo, neuporabno zaporedje, ki ne upošteva prekrivanj med podzaporedji.

Bolj smiselno se zdi, da poskusimo najti **čim krajše** zaporedje, ki vsebuje vsa dana zaporedja kot podzaporedja.

- Ta *problem najkrajšega nadzaporedja* je žal NP-**težek**.

Poglejmo si še en pristop.

Dano je zaporedje s nad abecedo Σ in naravno število ℓ.

ℓ-spekter zaporedja s je multimnožica vseh podzaporedij zaporedja s dolžine ℓ.

Označimo ga s *Spekter*(s, ℓ).

Zgled:

$s = \texttt{ACCGTTGTTGA}$, $\ell = 4$

$$\mathit{Spekter}(s, \ell) = \{\texttt{ACCG}, \texttt{CCGT}, \texttt{CGTT}, \texttt{GTTG}, \texttt{TTGT},$$

$$\texttt{TGTT}, \texttt{GTTG}, \texttt{TTGA}\}\,.$$

(Podzaporedje GTTG se pojavi dvakrat, torej ga tudi v *Spekter*(s, ℓ) vključimo dvakrat.)

Problem rekonstrukcije zaporedja

Za dano DNA molekulo je spekter njenega zaporedja mogoče določiti eksperimentalno, s pomočjo DNA mikromrež in hibridizacije.

(Sestavimo mikromrežo z vsemi možnimi 4^ℓ ℓ-tericami nad abecedo $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.)

Formuliramo lahko torej naslednji problem:

Problem rekonstrukcije zaporedja:

Podatki:	Naravno število ℓ, multimnožica S (ki predstavlja multimnožico vseh ℓ-podzaporedij (neznanega) zaporedja s).
Naloga:	Izračunaj zaporedje s, za katerega velja $\mathit{Spekter}(s, \ell) = S$.

Problem rekonstrukcije zaporedja:

Podatki:	Naravno število ℓ, multimnožica S (ki predstavlja multimnožico vseh ℓ-podzaporedij (neznanega) zaporedja s).
Naloga:	Izračunaj zaporedje s, za katerega velja $\mathit{Spekter}(s, \ell) = S$.

Ta problem je poseben primer NP-težkega problema iskanja najkrajšega nadzaporedja.

V nasprotju z njim pa je učinkovito rešljiv!

Problem rekonstrukcije zaporedja kot problem hamiltonske poti

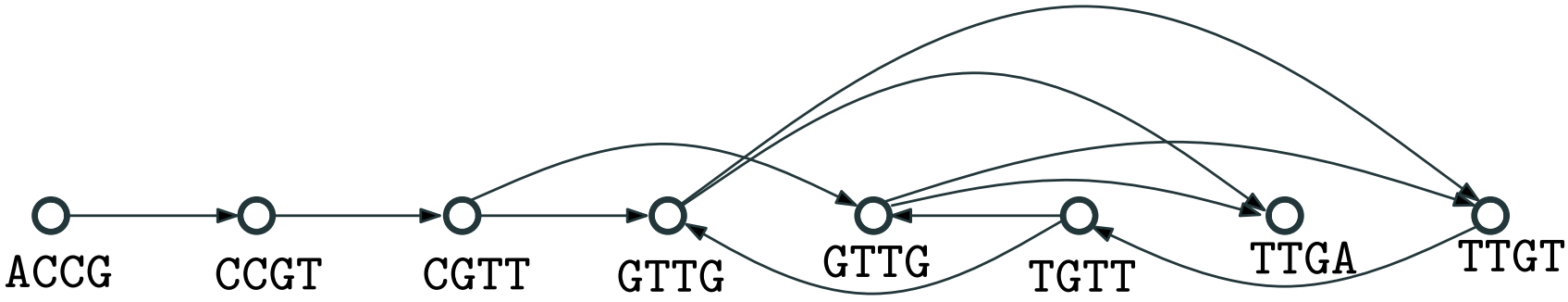
Problem rekonstrukcije zaporedja lahko modeliramo kot problem hamiltonske poti:

Za dane vhodne podatke (ℓ, S) problema rekonstrukcije zaporedij konstruiramo digraf $D = (V, E)$, kjer je:

- $V = S$
- $(p, q) \in E$ natanko tedaj, ko se p in q **prekrivata**, tj.,
  ko je zadnjih $\ell - 1$ znakov zaporedja p
  enakih prvim $\ell - 1$ znakov zaporedja q.

p : ACCGGT
q : CCGGTT

Zgled: $\ell = 4$, $S = \{\texttt{ACCG}, \texttt{CCGT}, \texttt{CGTT}, \texttt{GTTG}, \texttt{GTTG}, \texttt{TGTT}, \texttt{TTGA}, \texttt{TTGT}\}$.



Zaporedje $s = \texttt{ACCGTTGTTGA}$ ustreza naslednji poti v digrafu D:

$$\texttt{ACCG} \to \texttt{CCGT} \to \texttt{CGTT} \to \texttt{GTTG} \to \texttt{TTGT} \to \texttt{TGTT} \to \texttt{GTTG} \to \texttt{TTGA}\,.$$

To pa je pot, ki vsako točko v grafu obišče natanko enkrat – hamiltonska pot.

Torej je dovolj, če v konstruiranem grafu poiščemo hamiltonsko pot.

Ta pristop pa ima nekatere pomanjkljivosti:

- Hamiltonskih poti je lahko več.
- Problem hamiltonske poti je v splošnem NP-**težek**.

Problem rekonstrukcije zaporedja
problem Eulerjeve sledi

Problem rekonstrukcije zaporedja pa lahko modeliramo tudi kot problem Eulerjeve sledi.

Za dane vhodne podatke (ℓ, S) problema rekonstrukcije zaporedij

konstruiramo digraf $D = (V, E)$ (v katerem so možne večkratne povezave), kjer je:

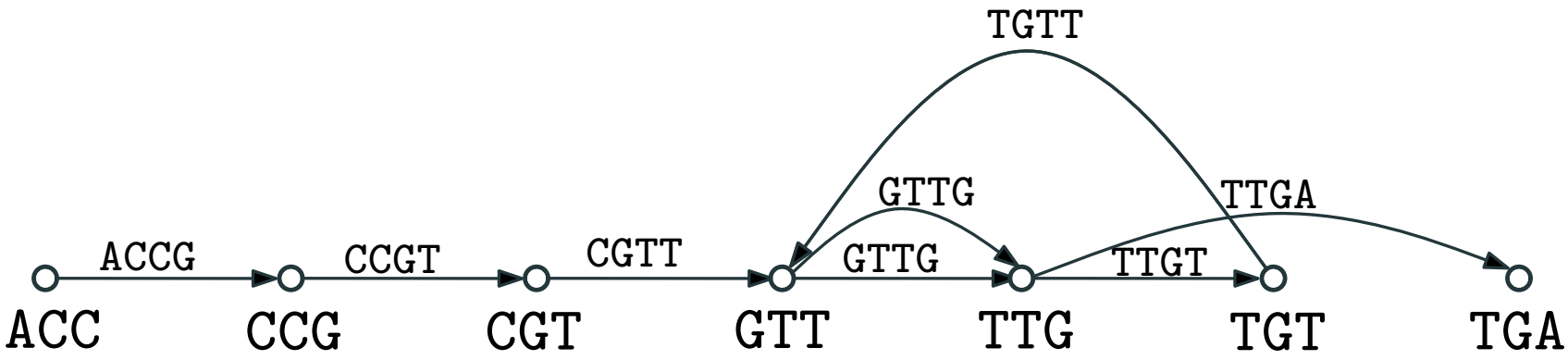
- $V =$ množica vseh zaporedij dolžine $\ell - 1$, ki nastopajo kot podzaporedja zaporedij v multimnožici S,
- $(p, q) \in E$ natanko tedaj, ko obstaja tako zaporedje $s \in S$, da je *prvih $\ell - 1$ znakov zaporedja s enakih p,*
  zadnjih $\ell - 1$ znakov zaporedja s pa je enakih q.

  Povezavo označimo z ℓ-terico s.

Vsako zaporedje $s \in S$ zdaj predstavlja povezavo in ne točke.

Zgled:

$\ell = 4,\ S = \{\mathtt{ACCG}, \mathtt{CCGT}, \mathtt{CGTT}, \mathtt{GTTG}, \mathtt{GTTG}, \mathtt{TGTT}, \mathtt{TTGA}, \mathtt{TTGT}\}$.



Zaporedje $s = \mathtt{ACCGTTGTTGA}$ ustreza naslednjemu sprehodu v grafu D:

$$\mathtt{ACC} \to \mathtt{CCG} \to \mathtt{CGT} \to \mathtt{GTT} \to \mathtt{TTG} \to \mathtt{TGT} \to \mathtt{GTT} \to \mathtt{TTG} \to \mathtt{TGA}\,.$$

To pa je sprehod, ki vsebuje vsako povezavo v digrafu natanko enkrat – Eulerjeva sled.

Torej je dovolj, če v konstruiranem digrafu poiščemo Eulerjevo sled.

Problem neenoličnosti sicer ostane, lahko pa Eulerjevo sled konstruiramo v polinomskem času s preprostim algoritmom.

Spomnimo: karakterizacija Eulerjevih digrafov

Točka $v \in V$ je **uravnotežena**, če je $d^+(v) = d^-(v)$.

Izrek (Euler, 1736)

Digraf brez izoliranih točk vsebuje Eulerjev obhod

$\Leftrightarrow$ *D je šibko povezan in so vse njegove točke uravnotežene.*

Dokaz:

Če D vsebuje Eulerjev obhod, potem je šibko povezan, saj lahko iz poljubne točke do poljubne druge pridemo vzdolž Eulerjevega obhoda.

Jasno je tudi, da so vse njegove točke uravnotežene. Obhod namreč v vsako točko pride natanko tolikokrat, kot iz točke odide.

Tudi obratno velja:

Če je D šibko povezan in so vse njegove točke uravnotežene, potem lahko Eulerjev obhod zgradimo z naslednjim postopkom:

- Izberi poljubno točko v in tvori maksimalno sled, ki se začne v v.
- Iz predpostavke o uravnoteženosti sledi, da se zadnja povezava na sledi konča v točki v.

  Je torej sklenjena sled, recimo ji C_1.
- Poišči točko w na C_1, ki ima še kakšne povezave, ki jih dosedaj nismo uporabili.
- Ponovi postopek na točki w. Spet dobimo sklenjeno sled, recimo C_2.
- Združi sledi C_1 in C_2 v eno samo sklenjeno sled C.
- Postopek ponavljaj, dokler obstaja še kakšna neporabljena povezava.

Algoritem je mogoče implementirati tako, da teče v času $\mathcal{O}(|V| + |E|)$. □

Točka $v \in V$ je **skoraj uravnotežena**, če je $|d^+(v) - d^-(v)| = 1$.

Izrek (Euler, 1736)

Digraf D brez izoliranih točk vsebuje nesklenjeno Eulerjevo sled

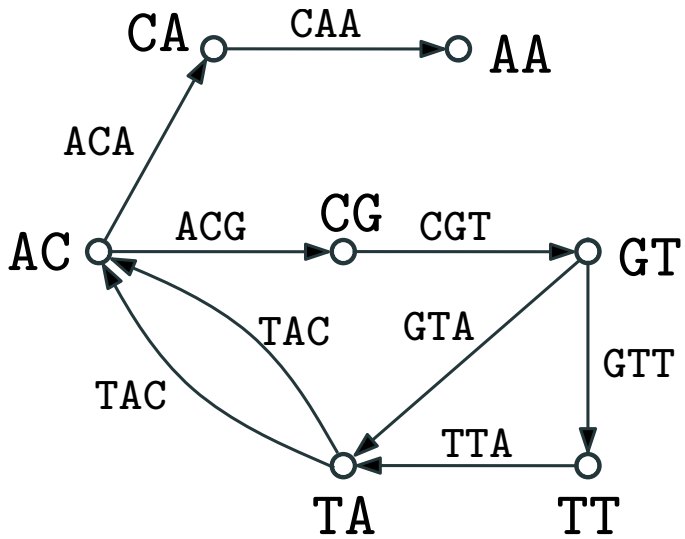
$\Leftrightarrow$ *D je šibko povezan in ima natanko dve skoraj uravnoteženi točki, vse ostale točke pa so uravnotežene.*

Podobno kot Eulerjeve obhode lahko poiščemo tudi nesklenjene Eulerjeve sledi.

- Dodamo povezavo od u do v, kjer je u edina točka, za katero velja $d^+(u) = d^-(u) - 1$, v pa edina točka, za katero velja $d^+(v) = d^-(v) + 1$.
- V dobljenem digrafu (kjer so vse točke uravnoteženen) poiščemo Eulerjev obhod C.
- Del obhoda C od v do u tvori nesklenjeno Eulerjevo sled v digrafu D.

Zgled:

$\ell = 3,\ S = \{\mathtt{ACG}, \mathtt{TAC}, \mathtt{GTA}, \mathtt{CAA}, \mathtt{TTA}, \mathtt{TAC}, \mathtt{GTT}, \mathtt{ACA}, \mathtt{CGT}\}$.



Graf vsebuje natanko dve Eulerjevi sledi.

Sledi ustrezata zaporedjema GTTACGTACAA in GTACGTTACAA.

Nekaj opomb o metodah sestavljanja genoma

Zgornji modeli so zelo poenostavljene abstrakcije problema združevanja odčitkov DNA zaporedja v celotno zaporedje.

- V praksi prihaja do napak. 1–3% nukleotidov je napačno sekvenciranih.
- DNA molekula tvori dvojno vijačnico; vnaprej ni znano, s katere strani pride kateri odčitek (zaporedje).

Še največji problem pa predstavljajo **večkratne pojavitve** podzaporedij.

- Zaporedje *Alu*, dolgo okrog 300bp, se v zaporedju človeškega genoma pojavi več kot milijonkrat.

  Neko drugo zaporedje *LINE*, dolgo okrog 1000bp, pa okrog 200.000-krat.

- Okrog 25% vseh genov se na genomu pojavi večkrat.

Problem večkratnih pojavitev bi bil rešen, če bi lahko sekvencirali odčitke, daljše od podzaporedij, ki se pojavijo večkrat.

Večina sodobnih tehnik sekvenciranja omogoča odčitke dolžine nekaj sto bp.

James Weber in Gene Myers iz podjetja Celera Genomics

(podjetje, ki je leta 2001 objavilo osnutek DNA zaporedja človeškega genoma)

sta razvila tehniko **sekvenciranja z mate-pair odčitki**:

- izberemo število L, večje od dolžine večine podzaporedij, ki v človeškem genomu nastopi večkrat,
- namesto celotnega podzaporedja dolžine L sekvenciramo samo začetno in končno podzaporedje,
- dobimo odčitka s in t, ki tvorita par ("mate pair")
  (na genomu sta približno L enot narazen).

Prednost tega pristopa: verjetnost, da bosta s in t vsebovana v dolgem ponavljajočem se zaporedju DNA, je majhna:

vsaj eden od obeh očitkov bo (zelo verjetno) na enolično določenem mestu v celotnem zaporedju.

Pred približno 10 leti je podjetje **Pacific Biosciences** razvilo tehnologijo sekvenciranja, ki proizvede odčitke dolžine $10.000 - 15.000$bp.

- Tehnologija je bila uporabljena za določitev genoma gorile.
- Je relativno počasna in (zaenkrat) zelo draga.
  Uporablja se predvsem v kombinaciji z ostalimi tehnologijami.

Dandanes imata najmočnejšo prisotnost v sekvenciranjih z dolgimi odčitki podjetji *Oxford Nanopore Technologies* in *PacBio*.

Obe tehnologiji rutinsko ustvarjata odčitke dolžine nekaj 10.000 bp. ONT gre lahko celo do nekaj 100.000 bp.

Leta 2018 so bili odčitki ONT uporabljeni za sestavljanje centromere kromosoma, kar zaradi ponavljanja še nikoli ni bilo storjeno s kratkimi odčitki.

Sekvenciranje s hibridizacijo na mikromrežih danes ni tako pomembno.

V projektu sestavljanja genoma **de novo** (to je, ko je treba zaporedje, ki predstavlja genom, sestaviti "iz nič" z uporabo odčitkov, brez uporabe referenčnega genoma) prihajajo odčitki iz platform sekvenciranja naslednje generacije (NGS).

Za kvalitetne rešitve se pogosto uporablja hibridni pristop, ki združuje kratke in natančne odčitke podjetja Illumina ter dolge, vendar manj natančne, odčitke podjetij Pacific Biosciences ali ONT.

Temeljni modeli hamiltonskih poti in Eulerjevih sledi so še vedno aktualni, v praksi pa so sodobni asemblerji očitno veliko bolj zapleteni, predvsem zato, ker se morajo spopasti z napakami pri odčitkih.

Glejte na primer naslednji članek za nekaj tipičnih sodobnih asemblerjev, prilagojenih dolgim odčitkom:

`https://academic.oup.com/bioinformatics/article/32/14/2103/1742895`