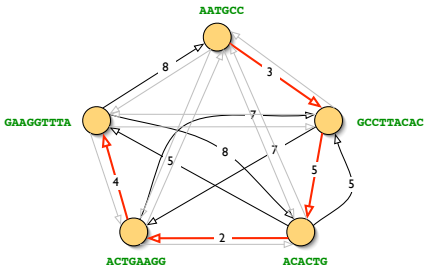
11. De novo assembly

A GENOM ÖSSZERAKÁS (*assembly*) problémája az, hogy véletlenül mintavételezett DNS fragmensekből állapítsuk meg a kromoszómák szekvenciáját. Tipikusan a mintaszekvenciák párosítva vannak.

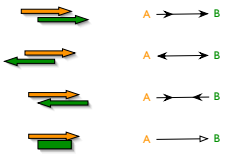
```
s1: AATGCC.....................GGATTC
s2:    GCCTTACAC.....................AGGATTC
s3:         ACACTG.................TCAAG
s4:           ACTGAAGG.....................ATTC
s5:              GAAGGTTTA.....................CGGACC
--------------------
B : AATGCCTTACACTGAAGGTTTA........GGATTCAAGGATTC..CGGACC
```

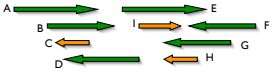
A klasszikus algoritmikai megközelítés az *overlap-layout-consensus* lépéseket követi. Az overlap fázisban meghatározzuk az egyes read-ek közötti átfedéseket. A layout lépésben az átfedő readek sorrendjét állapítjuk meg, és az egyben kiolvasható régiókat (*contig*ok). A consensus lépésben meghatározzuk az ismeretlen genom szekvenciát az egy pozícióba eső bázisok alapján (lényegében többségi szavazással).

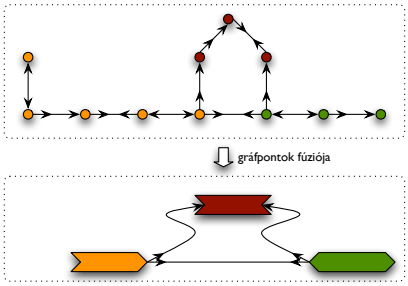
Overlap. Az összes read pár illesztése helyett heurisztikus illesztést érdemes elvégezni. A CAP3 program például a read-ek konkatenált szekvenciáján épít egy k-mer index táblázatot, és aztán egyesével illeszti a readeket (seed-and-extend módszerrel) a konkatenált „referenciához". A további lépéshez leglényegesebb információ az, hogy milyen mértékben és milyen orientálással fedi át egymást két read.



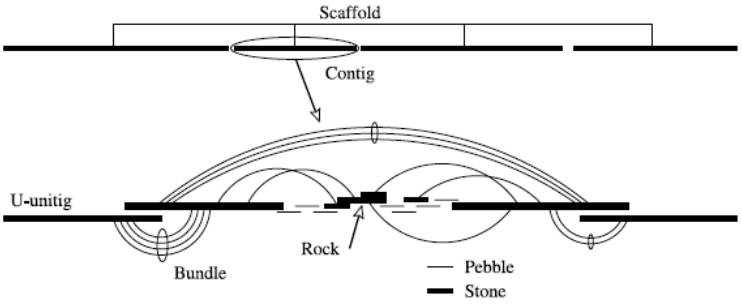
Layout. A layout problémája megfogalmazható gráfok segítségével. Minden read-nek megfeleltetve egy gráfpontot, az élek súlyozhatóak az illesztésből „kilógó" rész hosszával. Ekkor egy út a gráfban megfelel egy összerakott genomszekvenciának. (Ez nem a legjobb absztrakció, de érzékelteti, hogy az overlap-layout-consensus módszer lényegében Hamilton utas optimalizálási problémákat próbál megoldani.) A forward-reverse szálon illeszkedés kezeléséhez felvehetünk kétszeresen irányított éleket. Az így keletkező gráf az élek orientáltságának és az illesztések kompatibilitása alapján tovább egyszerüsíthető, lokális lépésekben.







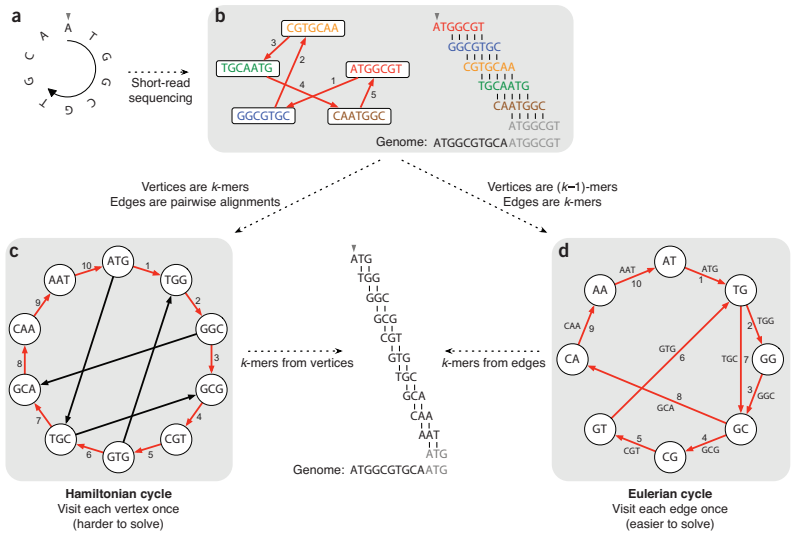
Scaffolding. Egy *scaffold* (váz / állványzat) contigok sorozata, egymáshoz képest ismert orientáltsággal és távolsággal.



1. ábra. Contigok összekötése a Celera assemblerben. Páros readek összekapcsolhatnak contigokat vagy a contig melletti rész szekvenciáját segítenek meghatározni. [Myers++]

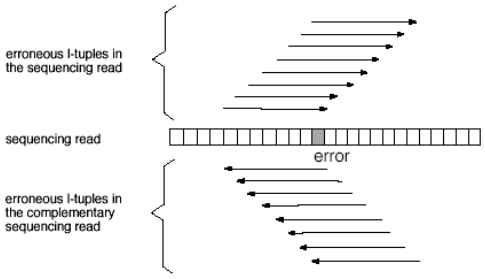
Szekvencia spektrum

***k*-mer spektrum**: a szekvenciában megtalálható k-merek halmaza. Az egyik első alternatív szekvenáló módszer: Sequencing-by-hybridization. Hibridizációval megállapítható, hogy egy adott k-mert tartalmaz-e a szekvencia: egy chipre ültetve az összes k-mert, a szekvencia teljes spektruma megállapítható. Assembly a spektrumból? De Bruijn gráffal! Ebben a gráfban a k-mereket az élekhez rendeljük, és a gráfpontok a $k-1$ hoszzú szavak. Így a gráfban Euler utat kell keresni az assembly-hez, ami lineáris időben megtehető.

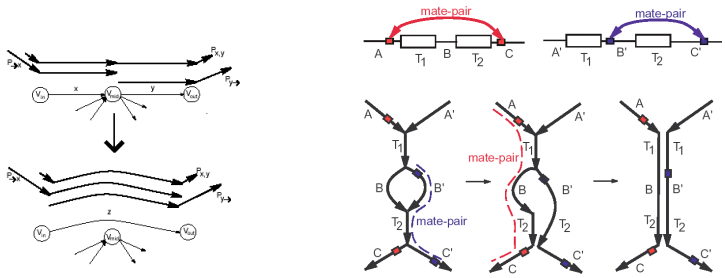


2. ábra. Két gráf absztrakció genom összerakásra. A de Bruijn gráfos megközelítés gyorsabb algoritmushoz vezet. [Compeau++]

Nagyobb genomokra $k > 20$ szokásos választással, a readeket először felbontjuk a k-merekre, amin megépítjük a de Bruijn gráfot. E megközelítés legnagyobb nehézsége az, hogy a szekvenálási hibák miatt sok „árva” k-mer keletkezhet, ami a gráfpontok számát nagyon megnövelheti. Ezért egy első lépésben hibakorrekcióra van szükség (ritka k-mer $\Rightarrow$ egy nukleotid megváltoztatásával gyakori k-merhez jutunk?).

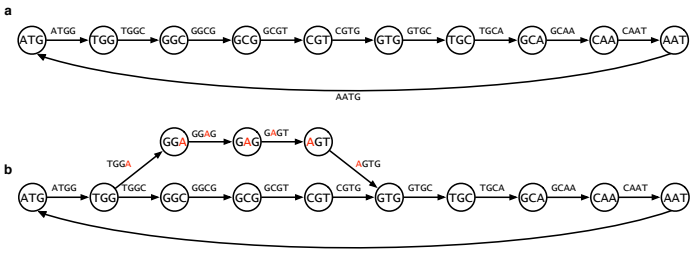


A *read*-eket és a párokat megőrizzük mint rövid ismert utakat a gráfban. Lokális gráftranszformációkkal (új élek, pontok) lépésenként lehet „kibogozni" a gráfot.



3. ábra. Tipikus gráf-egyszerűsítő lépések read-ek (vastag nyilak) egymás utáni *k*-merjei és párok imsert kapcsolatai alapján. [Pevzner & Tang]

Buborékok. Hibás read egy buborékot csinál a gráfban: ritka *k*-merekből álló út. Ezt gráfbejárással fel lehet fedezni (Tour Bus Correction: amikor másodszor ugyanahhoz a ponthoz érünk a bejárásban, akkor meg lehet keresni a közös őst és illeszteni a két útnak megfelelő szekvencia között).



4. ábra. Buborék hibás read miatt (pirossal jelölt 'A').

Hivatkozások

Phillip E. C. Compeau, Pavel A. Pevzner, and Glenn Tessler. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29: 987–991, 2011. DOI: 10.1038/nbt.2023.

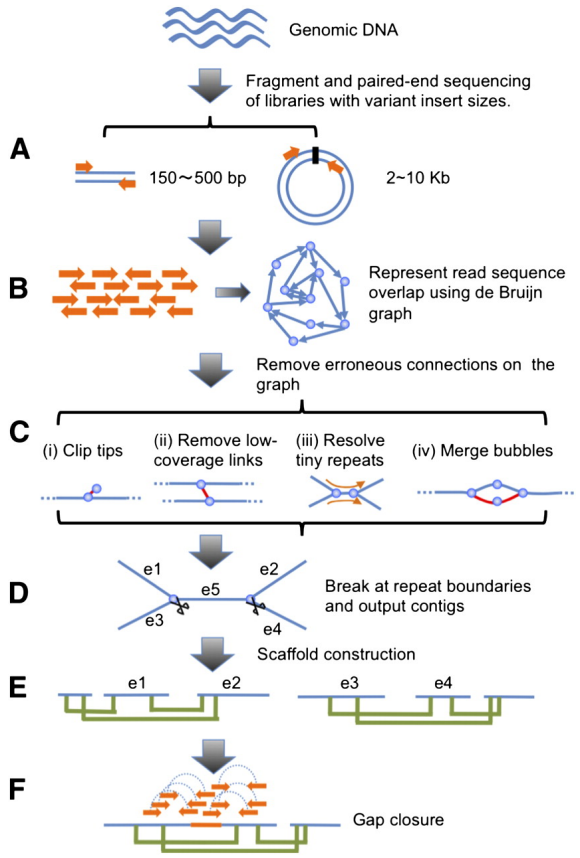
Ruiqiang Li, Hongmei Zhu, Jue Ruan, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20:265–272, 2010. DOI: 10.1101/gr.097261.109.

Eugene W. Myers et al. A whole-genome assembly of Drosophila. *Science*, 287:2196–2204, 2000. DOI: 10.1126/science.287.5461.2196.

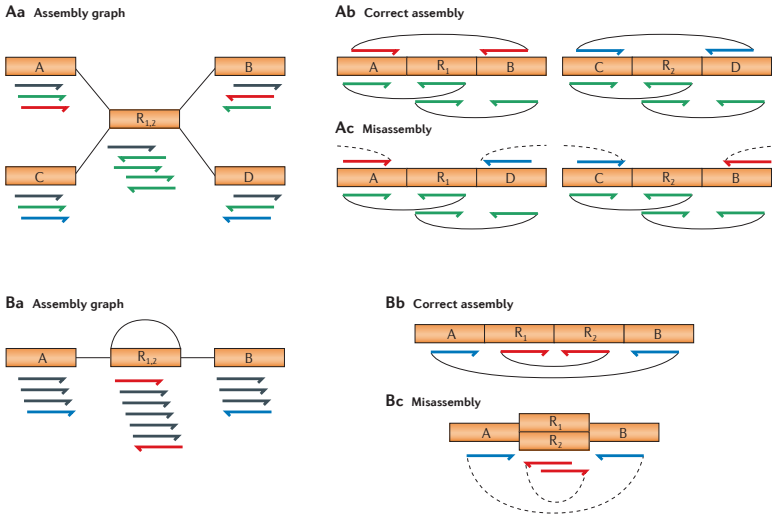
Niranjan Nagarajan and Mihai Pop. Sequence assembly demystified. *Nature Reviews Genetics*, 14:157–167, 2013. DOI: 10.1038/nrg3367. ✱ **Kötelező olvasmány!**

Pavel A. Pevzner and Haixu Tang. Fragment assembly with double-barreled data. *Bioinformatics*, 17:S225–S233, 2001. DOI: 10.1093/bioinformatics/17.suppl_1.S225.

Todd J. Treangen and Steven L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13:36–46, 2012. DOI: 10.1038/nrg3117.



5. ábra. A SOAPdenovo program lépései. Az egyik legfontosabb fázis a hibák javítása a gráf bejárásakor (**C**). [Li++]



6. ábra. Ismétlődő régiók (*repeat*-ek) akár tandem, akár egymástól messze nehezítik az assemblyt. Ez a gráfban nem feloldható „csomót" jelent: nem egyértelmű, hogy merre kell mennie az utaknak. A repeateket csak úgy lehet feloldani, ha van az egész régión átívelő párunk. [Treangen & Salzberg]