Sestavování DNA

Bioinformatika

Tomáš Martínek

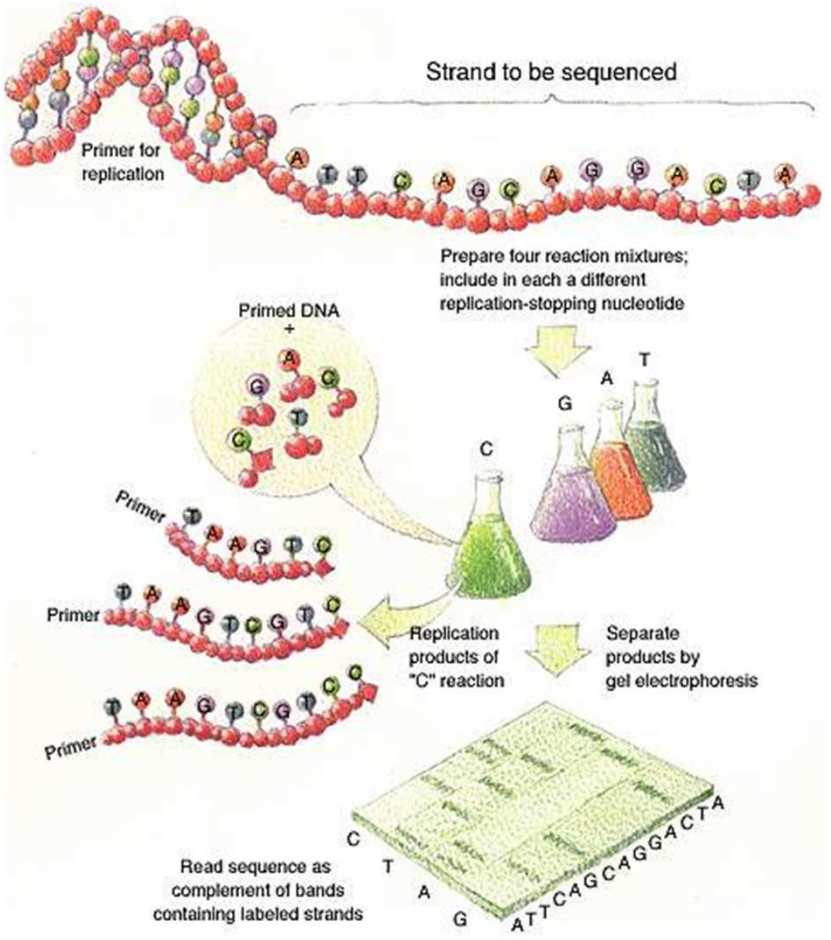
martinto@fit.vutbr.cz

Osnova

- Sangerova metoda
- Sekvenování rozsáhlých genomů
  - BAC by BAC
  - Shotgun
- Sestavování fragmentů
  - Overlap
  - Layout
  - Consensus
- Shrnutí

Sangerova metoda

- Využívá princip replikace DNA
- Nobelova cena 1980
- Postup:
  1. Analyzovaný fragment DNA se namnoží a společně s primerem se vloží do 4 oddělených zkumavek s roztokem obsahujícím volné nukleotidy A,C,G,T
  2. Do jednotlivých zkumavek se navíc vloží nukleotidy, upravené tak, aby se na ně již nemohly vázat další nukleotidy - deoxynukleotidy ddA, ddC, ddG, ddT
  3. Při chemické reakci se v jednotlivých zkumavkách vytvoří všechny možné sekvence končící na daný znak - např. ACGTAAGCTA v roztoku s ddT vytvoří fragmenty ACG a AAGCTAAGC
  4. Sekvence se seřadí podle velikosti a přečtou z elektroforézního gelu
- Dnes lze tímto způsobem číst fragmenty DNA dlouhé pouze cca 500-700 bp



Korekce výsledků z Elektroforeogramu

- Výsledky z Elektroforeogramu musí být dále filtrovány a korigovány
- Korekční metoda PHRED [1990s Phil Green]
- Každý znak získané sekvence je ohodnocen skórem, které odpovídá pravděpodobnosti, že daný znak byl správně nasekvenován
- Hodnoty skóre:

Skóre	Pr. chyby	Přesnost
10	1 z 10	90%
20	1 z 100	99%
30	1 z 1000	99.9%
40	1 z 10k	99.99%
50	1 z 100k	99.999%

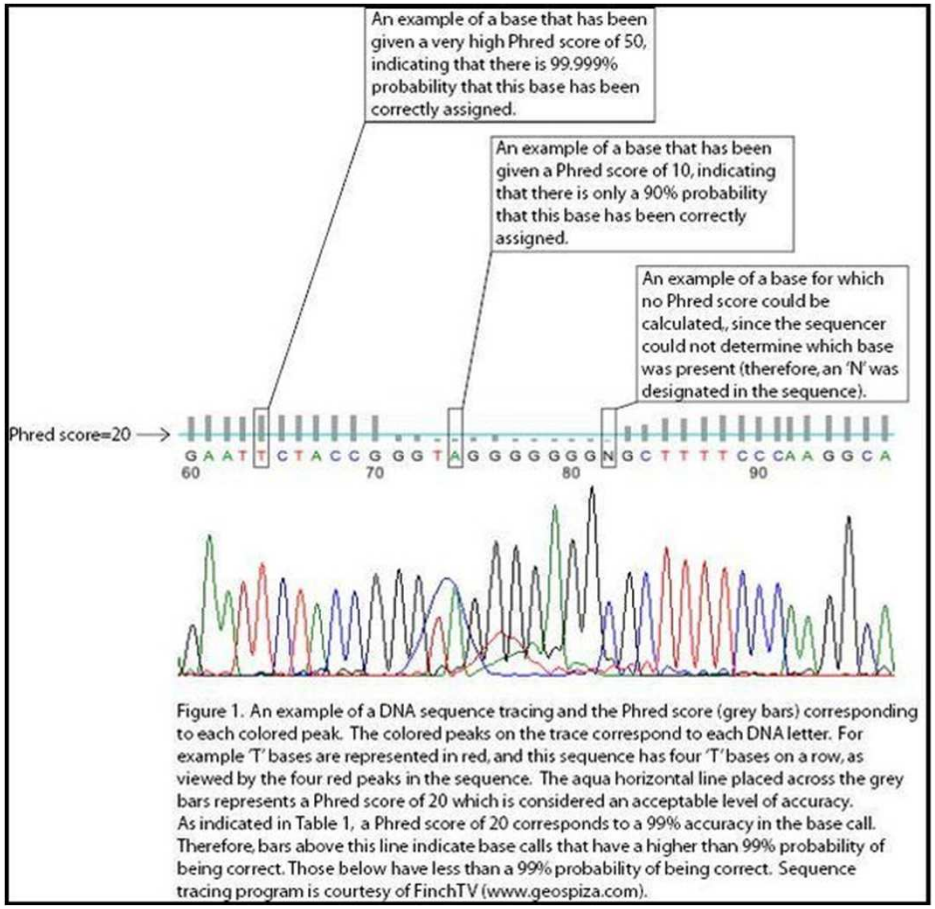
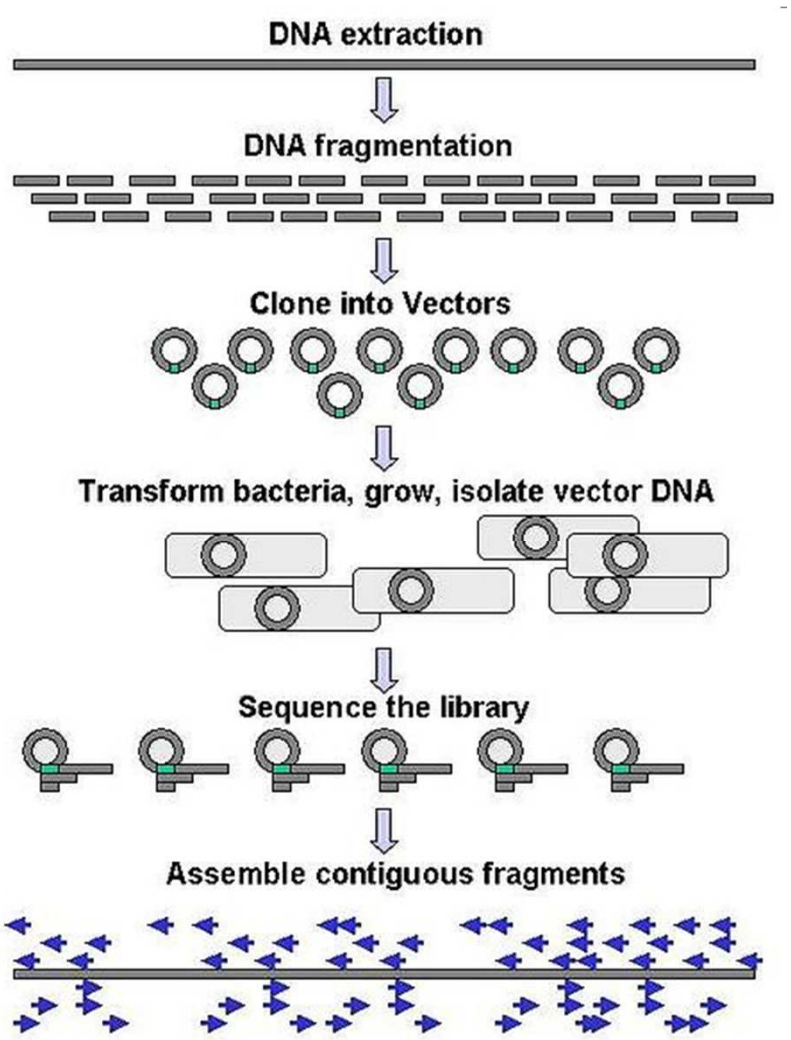


Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy. As indicated in Table 1, a Phred score of 20 corresponds to a 99% accuracy in the base call. Therefore, bars above this line indicate base calls that have a higher than 99% probability of being correct. Those below have less than a 99% probability of being correct. Sequence tracing program is courtesy of FinchTV (www.geospiza.com).

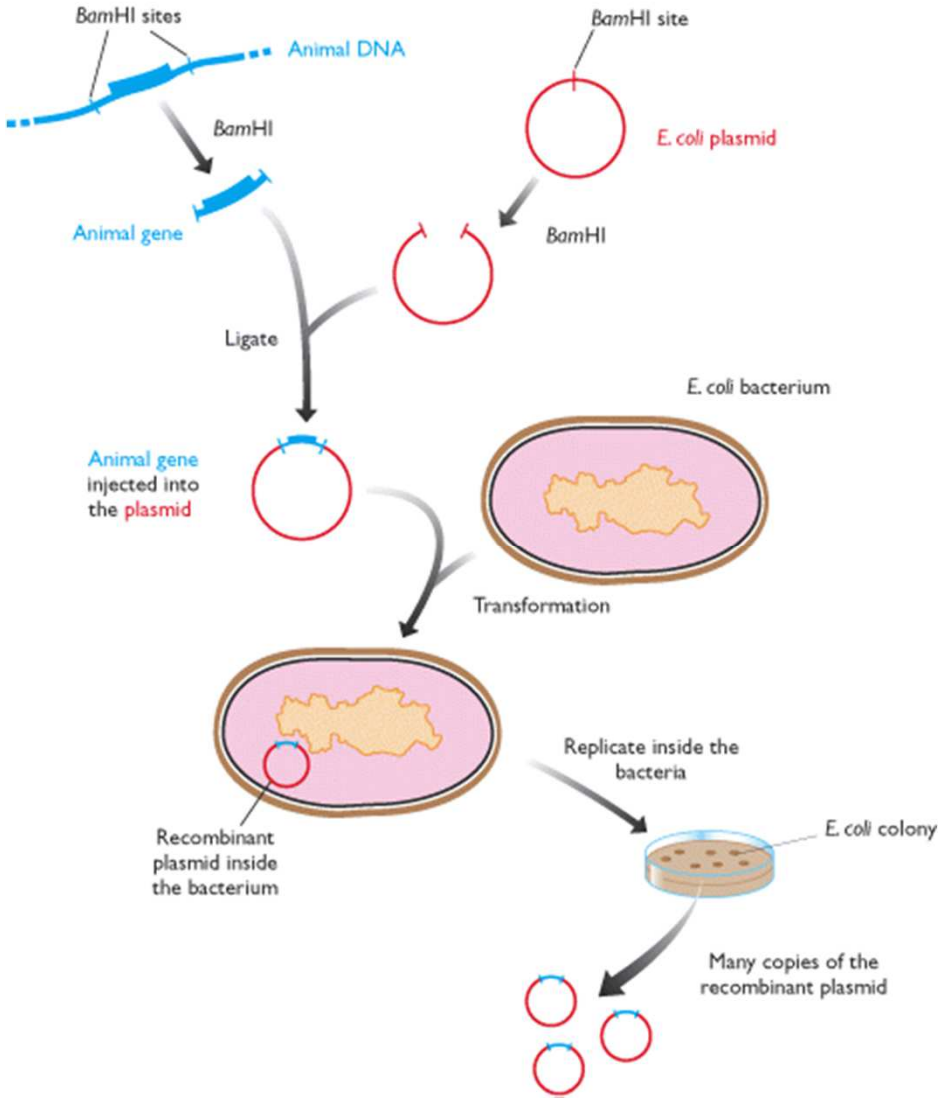
Sekvenování rozsáhlých genomů

- Jak pomocí 500 bp úseků nasekvenujeme celý genom?
- Postup:
  1. Rozdělení původní DNA do překrývajících se fragmentů (cca 200k bp)
  2. Namnožení fragmentů skrze hostitelské bakterie (BAC)
  3. Vytvoření knihovny fragmentů (BACů) pokrývající celý genom a jejich samostatné sekvenování - BAC-by-BAC metoda
  4. Sestavení překrývajících se fragmentů do jedné sekvence celého genomu



Klonování DNA

- Cílem je namnožit daný úsek DNA s použitím hostitelské bakterie
- Postup:
  1. Příprava rekombinantní molekuly DNA a hostitelské buňky s kružnicovou molekulou DNA (např. plazmid bakterie E. coli)
  2. Přenos rekombinantní DNA do plazmidu hostitelské buňky na přesně známou pozici
  3. Replikace modifikované molekuly plazmidu uvnitř hostitelské buňky
  4. Selekce klonů obsahujících rekombinantní DNA a analýza klonované DNA



Klonování DNA

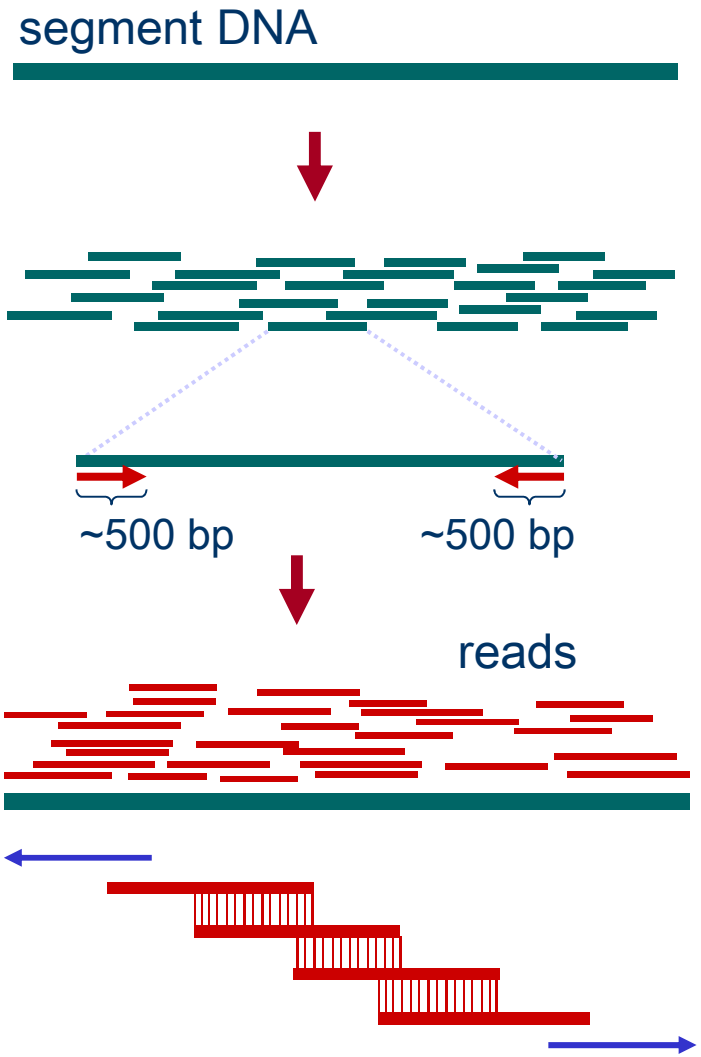
- Závislost velikosti fragmentu na použitém vektoru

VEKTOR	**Velikost fragmentu (bp)**
Plazmid	2,000 - 10,000
Cosmid	40,000
BAC (Bacterial Artificial Chromosome)	70,000 - 300,000
YAC (Yeast Artificial Chromosome)	> 300,000 (používáno zřídka)

- Jak ale nasekvenovat BACy o velikosti 200k bp, když Sangerova metoda dokáže přečíst pouze fragmenty o velikosti 500 bp?

Shotgun sekvenování

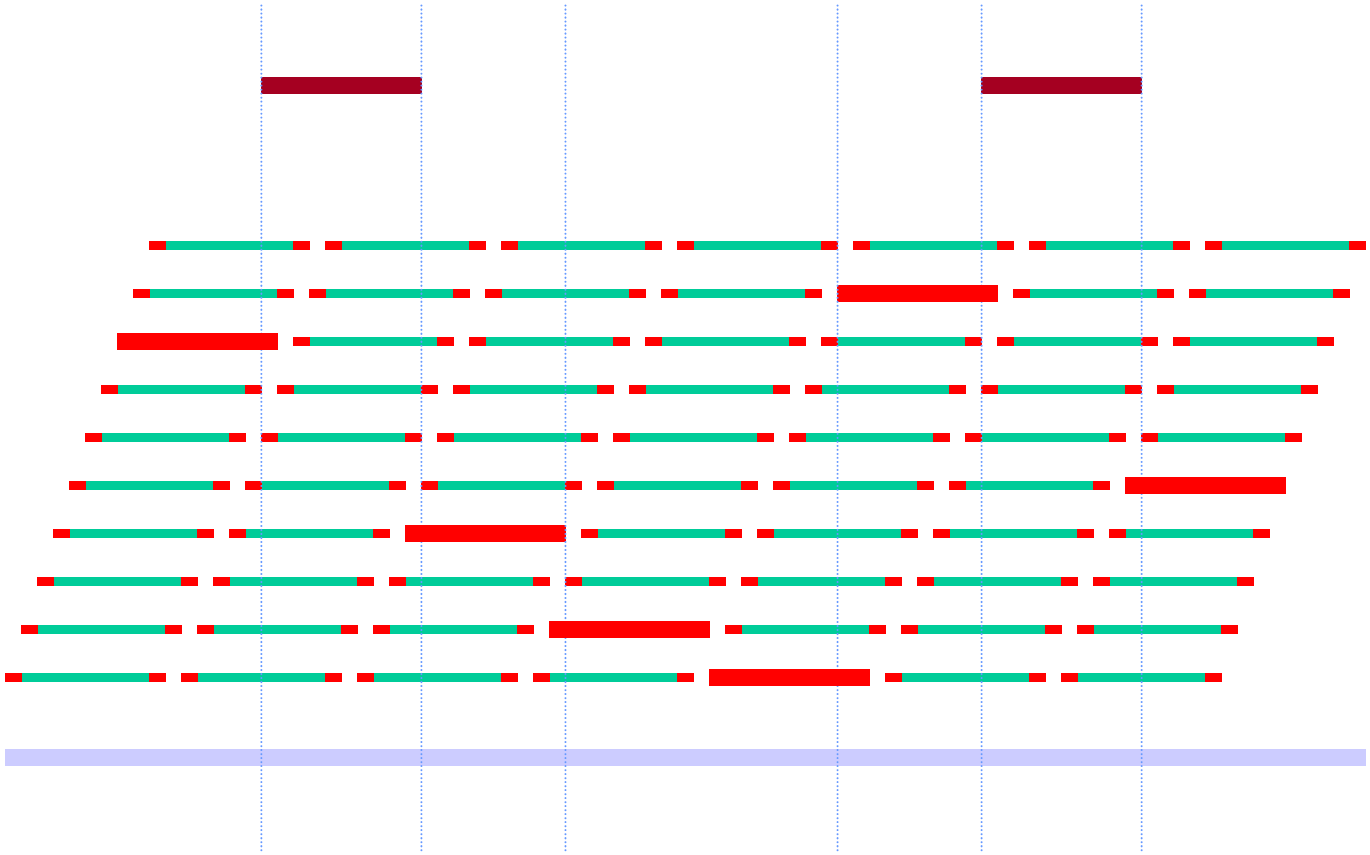
- Postup:
  1. Vstupní sekvence je náhodně rozsekána na mnoho překrývajících se fragmentů o délce cca 2k bp
  2. Namnožení fragmentů probíhá skrze plazmidy (kratší délka, jednodušší)
  3. Z každého konce fragmentů je Sangerovou metodou přečteno cca 500 bp (read)
  4. Na základě překryvů je vestaven původní segment
- Poznámky:
  - 500 bp z obou konců fragmentu tvoří dvojici s přibližně známou velikostí mezery – výrazně zjednodušuje bod 4. [1997 Weber, Mayers]
- Jak ale poskládám jednotlivé BACy k sobě?



Walking metoda

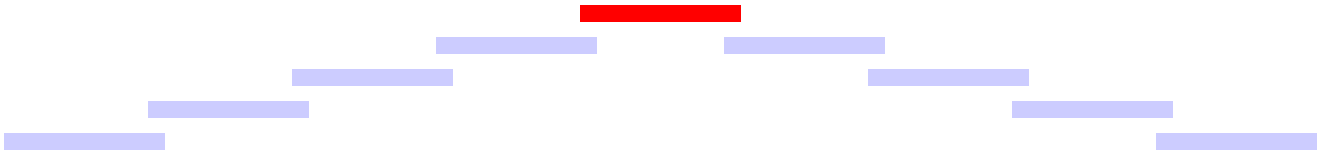
- Postup:
  1. Vytvoří se rozsáhlá (redundantní) knihovna BACů, kde každý BAC bude mít nasekvenovány oba své konce (cca 500 bp)
  2. Náhodně vybraný BAC se nasekvenuje metodou shotgun
  3. Obsah nasekvenovaného BACu se porovná se všemi nasekvenovanými konci BACů v knihovně a nalezne se dvojice BACů, které rozšiřují nasekvenovaný BAC z leva i zprava
  4. Rozšiřující BACy se nasekvenují a metoda pokračuje od bodu 3.

Walking metoda



Walking metoda

- Výsledné schéma:



- Výhody:
  - Nedochází k více-násobnému sekvenování stejných úseků
- Nevýhody:
  - Mnoho sekvenčních cyklů
  - Jeden cyklus trvá cca 1-2 měsíce
  - Pro nasekvenování genomu savců je potřeba cca 15.000 cyklů
- Řešení:
  - Spuštění výpočtu z několika míst paralelně => snížení počtu cyklů
  - Redundance sekvenování stoupne cca na 20%

BAC-by-BAC vs. Shotgun

1997

Gene Myers

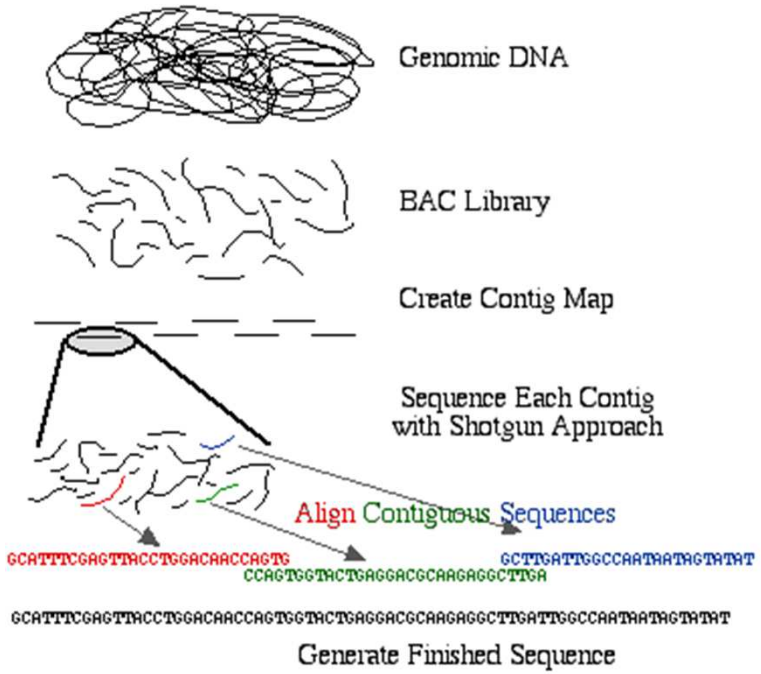
Let's sequence the human genome with the shotgun strategy

That is impossible, and a bad idea anyway

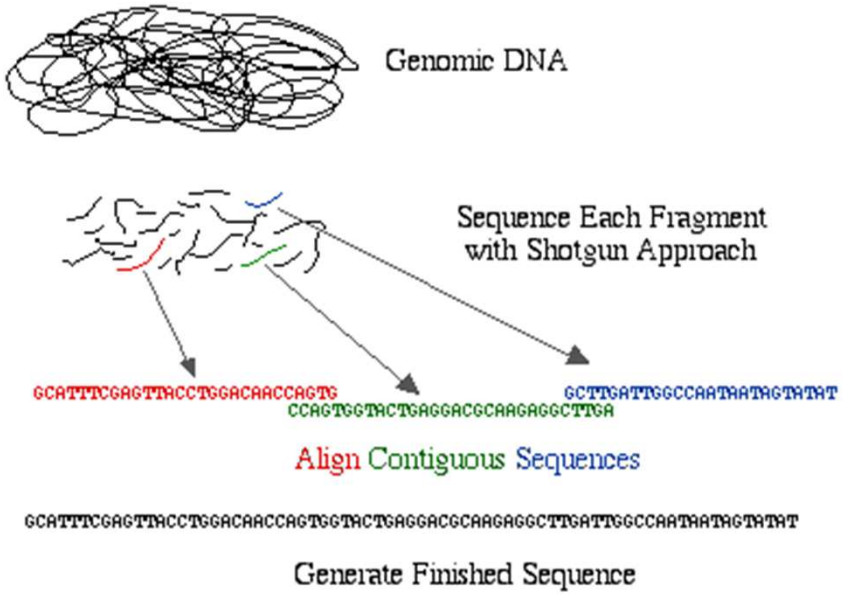
Phil Green

BAC-by-BAC vs. Shotgun

- BAC-by-BAC
  - Vetší odolnost proti chybám
  - Pomalejší a drahé z pohledu experimetů
  - Jednodušší z pohledu výpočtu



- Shotgun
  - Je vynechán krok vytváření BAC knihovny
  - Vetší náchylnost k chybám ⇒ pokrytí původního genomu musí být výrazně vyšší 10-12x
  - Rychlejší z pohledu experimentů
  - Vyšší náročnost na výpočetní výkon



Výpočet pokrytí

- Vstupní parametry:
  - G Délka DNA segmentu
  - l Délka readu
  - N Počet readů
- Výpočet pokrytí
  - C = N.l/G
- Jaké pokrytí C je dostatečné?

- Lander-Waterman model
  - předpokládá uniformní rozložení readů

$$P\{X = i\} = \frac{e^{-C}\lambda^{i}}{i!}$$

  - pravděpodobnost, že jakákoliv báze není sekvenována je $P_0=e^{-C}$
  - celková délka mezer je Ge^{-C}
  - počet mezer je Ne^{-C}

Výpočet pokrytí

- Pravděpodobnost, že jakákoliv báze není sekvenována je $P_0=e^{-C}$

C	$P_0=e^{-C}$	Nesekvenován P_0	Sekvenován $(1-P_0)$
1	0.37	37%	63%
2	0.135	13.5%	87.5%
3	0.05	5%	95%
4	0.018	1.8%	98.2%
5	0.0067	0.6%	99.4%
6	0.0025	0.25%	99.75%
7	0.0009	0.09%	99.91%
8	0.0003	0.03%	99.97%
9	0.0001	0.01%	99.99%
10	0.000045	0.005%	99.995%

Sestavování fragmentů

Shortest Superstring Problém (SSP)

- Vstup:
  - Množina řetězců M = {s_1, s_2, ..., s_n}
- Výstup:
  - Cílem je nalézt takový minimální řetězec S (super string), který obsahuje všechny řetězce z množiny M jako své podřetězce
- Příklad:
  - M = {AAA, AAC, ACA, ACC, CAA, CAC, CCA, CCC}
  - S = AAACCCACAA
- Jak získat nejkratší super string?

Shortest Superstring Problém (SSP)

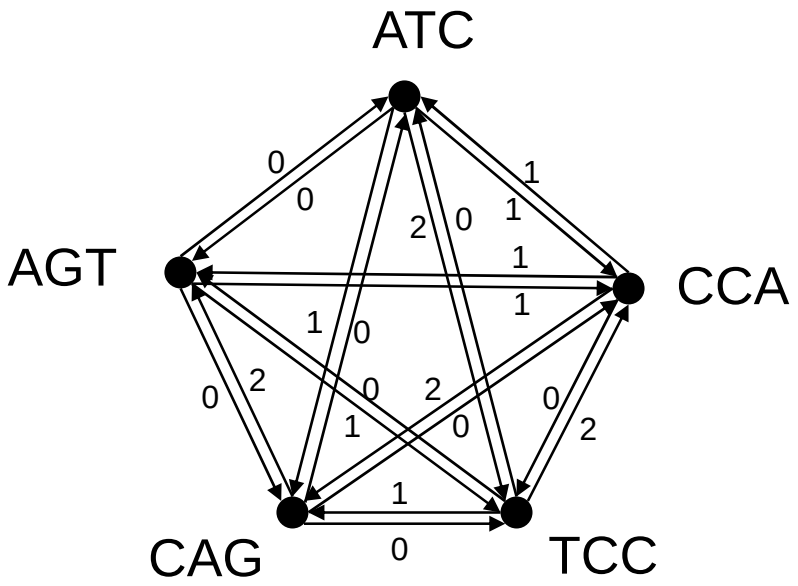
- Definice operace překrytí dvou řetězců s_i a s_j
  - *overlap*(s_i,s_j) je definována jako délka nejdelšího prefixu s_j, který se shoduje se sufixem s_i
- Příklad:
  - s_i = aaaggcatcaaatctaaaggcatcaaa
  - s_j = aaaggcatcaaacctgatggaatcaaa

```
aaaggcatcaaatctaaaggcatcaaa
                          aaaggcatcaaacctgatggaatcaaa
aaaggcatcaaatctaaaggcatcaaa
                    aaaggcatcaaacctgatggaatcaaa
```

  - *overlap*(s_i,s_j) = 12

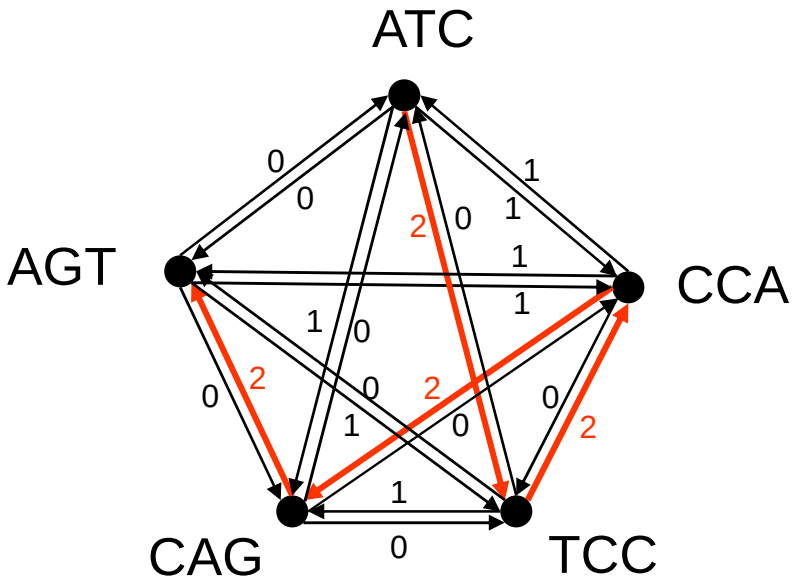
Shortest Superstring Problém (SSP)

- Sestavení grafu G kde:
  - Vrcholy tvoří jednotlivé řetězce z množiny *M*
  - Hrany propojují každý vrchol s každým (řetězce s_i a s_j) a jsou ohodnoceny hodnotou *overlap*(s_i, s_j)
- Příklad:
  - M = { ATC, CCA, CAG, TCC, AGT }



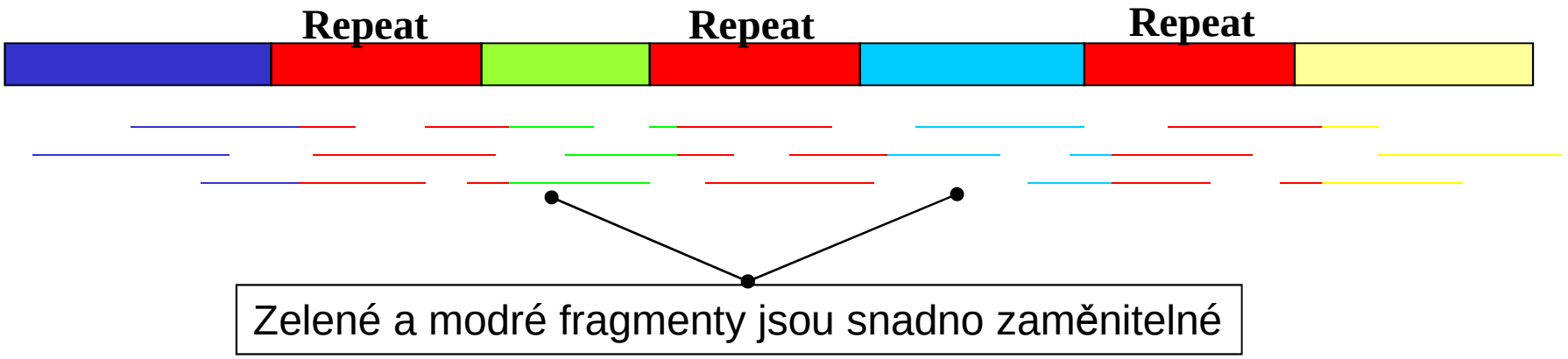
Shortest Superstring Problém (SSP)

- Řešení problému:
  - Nejkratší super string je odpovídá cestě v grafu G, která projde všemi vrcholy právě jednou a současně dosáhne v součtu ohodnocení hran nejvyššího skóre
- Příklad:
  - M = { ATC, CCA, CAG, TCC, AGT }
  - S = ATCCAGT
- Ekvivalentní problém obchodního cestujícího
- NP-úplný problém



Shortest Superstring Problém (SSP)

- Existuje heuristika s 2.66-aproximací tj. nejhorší případ je vzdálen 2.66 krát od optimálního řešení [Armen and Stein, 1996]
- Další problémy SPP:
  - neuvažují se chyby sekvenování (1-3%)
  - není známo, z které dvojice vláken fragment je
  - hlavním problémem jsou ale opakování, které se vyskytují velmi často



Opakování v DNA

- **Low-Complexity DNA**
  - např. ATATATATACATA...
- **Microsatellite repeats**
  - $(a_1 \ldots a_k)^N$ kde k ~ 3-6
  - např CAGCAGTAGCAGCACCAG
- **Transposons/retrotransposons**
  - **SINE** - Short Interspersed Nuclear Elements
    - např., *Alu*: ~300 bp dlouhé, 10^6 kopií, (5-15% změn)
  - **LINE** - Long Interspersed Nuclear Elements
    - ~500 - 5,000 bp dlouhé, 200,000 kopií
  - **LTR retroposons -** Long Terminal Repeats (~700 bp) na každém konci
- **Gene Families**
  - důležité geny jsou několikrát duplikovány
- **Segmental duplications**
  - velmi dlouhé a velmi podobné segmenty

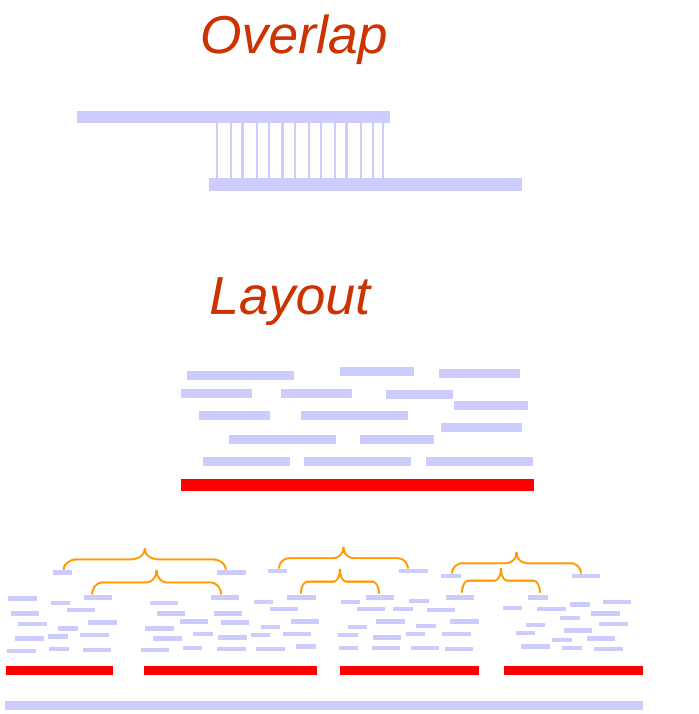
Více než 50% lidského genomu jsou opakování!!

Opakování

- Triazzle puzzle
  - vypadá jednoduše, ale obsahuje opakování
  - často se stane, že vám zbudou 2 kostky, které není kde vložit, zatímco ostatní sedí perfektně
  - www.triazzle.com

Sestavování fragmentů DNA

- Sestavování rozděleno do tří hlavních fází:
  1. Hledání překryvů (Overlap)
  2. Spojování fragmentů do contigů a superconfigů (Layout)
  3. Oprava chyb (Consensus)
- Existující nástroje
  - 1. generace: neuvažovala páry readů - Phrap, CAP, ...
  - 2. generace: počítá s páry - Arachne (free), CA, Euler



Hledání překryvů (Overlap)

- Vstupní data:
  - Párové ready
  - Každý znak v readu je ohodnocen skórem podle metody PHRED
  - Skóre vyjadřuje pravděpodobnost, že byl daný znak správně nasekvenován
  - Konce readů s velmi nízkým skóre jsou ořezány
- Výstup:
  - Cílem je nalézt fragmenty s největším překryvem – vyskytující se v původní DNA vyskytují blízko sebe
- Poznámky:
  - Je nezbytné uvažovat 1-3% chyb, proto se pro hledání často používá dynamické programování
  - Problémy způsobují hlavně opakování

Hledání překryvů (Overlap)

- Triviální algoritmus:
  - Porovná všechny dvojice segmentů mezi sebou
  - Časová složitost: $O(n^2)$
- Efektivnější algoritmus:
  - Dostatečně překrývající se segmenty s velkou pravděpodobností sdílejí určité úseky velikosti *k* bez chyb
  - *k* je obvykle voleno jako 24
- Postup:
  1. Vytvoření seznamu všech k-tic ze všech segmentů společně s číslem segmentu
  2. Seřazení seznamu podle k-tic *O(n.log(n))*
  3. Segmenty, které mají výrazný překryv budou v seřazeném seznamu u sebe

Hledání překryvů (Overlap)

- Příklad:
  - Vstupní fragmenty:
    1. TAATAT
    2. GTCTGA
    3. TATAAA

Seznam trojic:	Seřazený seznam:
(TAA,1)	(AAA,3)
(AAT,1)	(AAT,1)
(ATA,1)	(ATA,1)
(TAT,1)	(ATA,3)
(GTC,2)	(CTG,2)
(TCT,2)	(GTC,2)
(CTG,2)	(TAA,1)
(TGA,2)	(TAA,3)
(TAT,3)	(TAT,1)
(ATA,3)	(TAT,3)
(TAA,3)	(TCT,2)
(AAA,3)	(TGA,2)

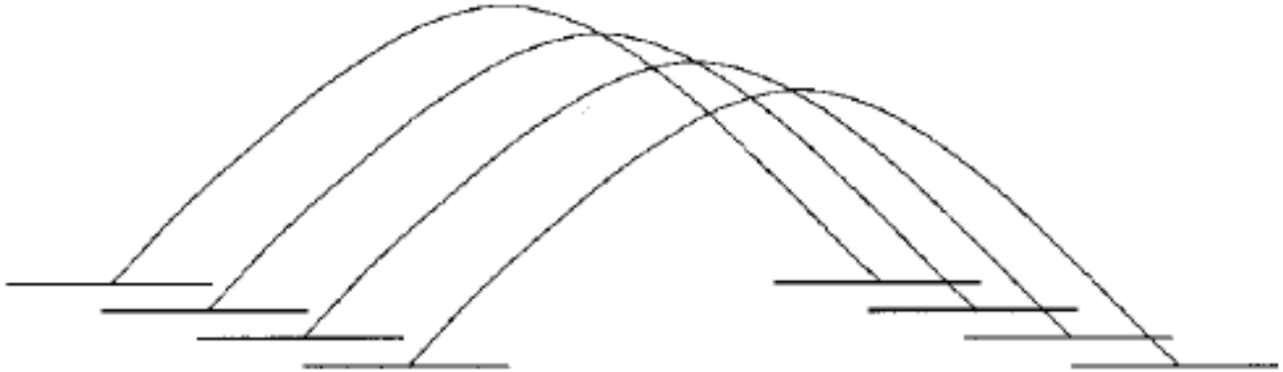
Hledání překryvů (Overlap)

- Identifikace dvojic fragmentů a jejich rozšíření na obě strany
  - jsou brány v úvahu pouze dvojice s podobností >95%
  - pro rozšiřování se používá dynamické programování

- Problém: Opakování generují extrémně mnoho k-tic
  - k-tice, které se vyskytují N-krát vyžadují N^2 dodatečných porovnání
  - Příklad: ALU sekvence se vyskytují 10^6 krát a vyžadují 10^{12} porovnání
  - Řešení: k-tice s velmi vysokou četností jsou redukovány

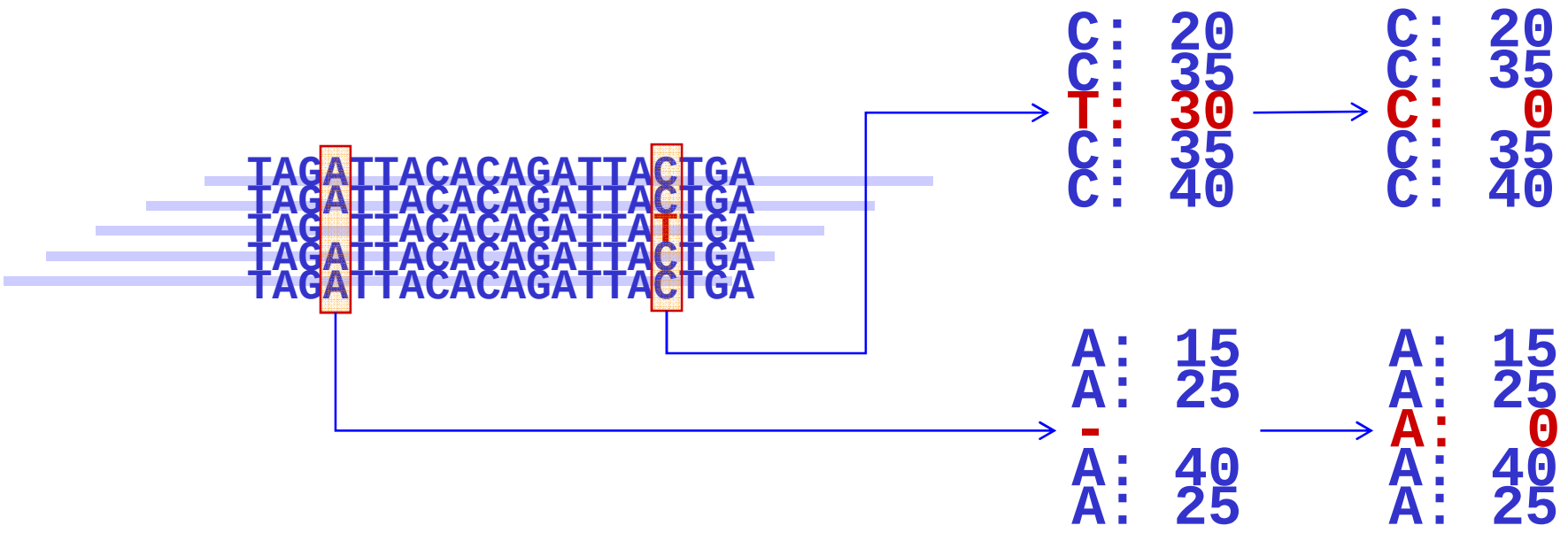
Hledání překryvů (Overlap)

- Identifikace dvojic fragmentů (pokračování)
  - Každá dvojice překrývajících se segmentů je ohodnocena skórem
  - Toto skóre závisí jednak na kvalitě jednotlivých znaků, ale také na výskytu chyb
  - Navíc se kontroluje, zda odpovídají druhé páry readů
  - Většinou souhlasí, až na případy dlouhých opakování, kde párované ready neodpovídají a program tyto vazby vyloučí



Hledání překryvů (Overlap)

- Korekce chyb překryvů
  - Vytvoří se vícenásobného lokální zarovnání
  - Možné výskyty chyb se zkorigují podle znaků, které převažují
  - Korekční proces bere v úvahu i skóre jednotlivých znaků



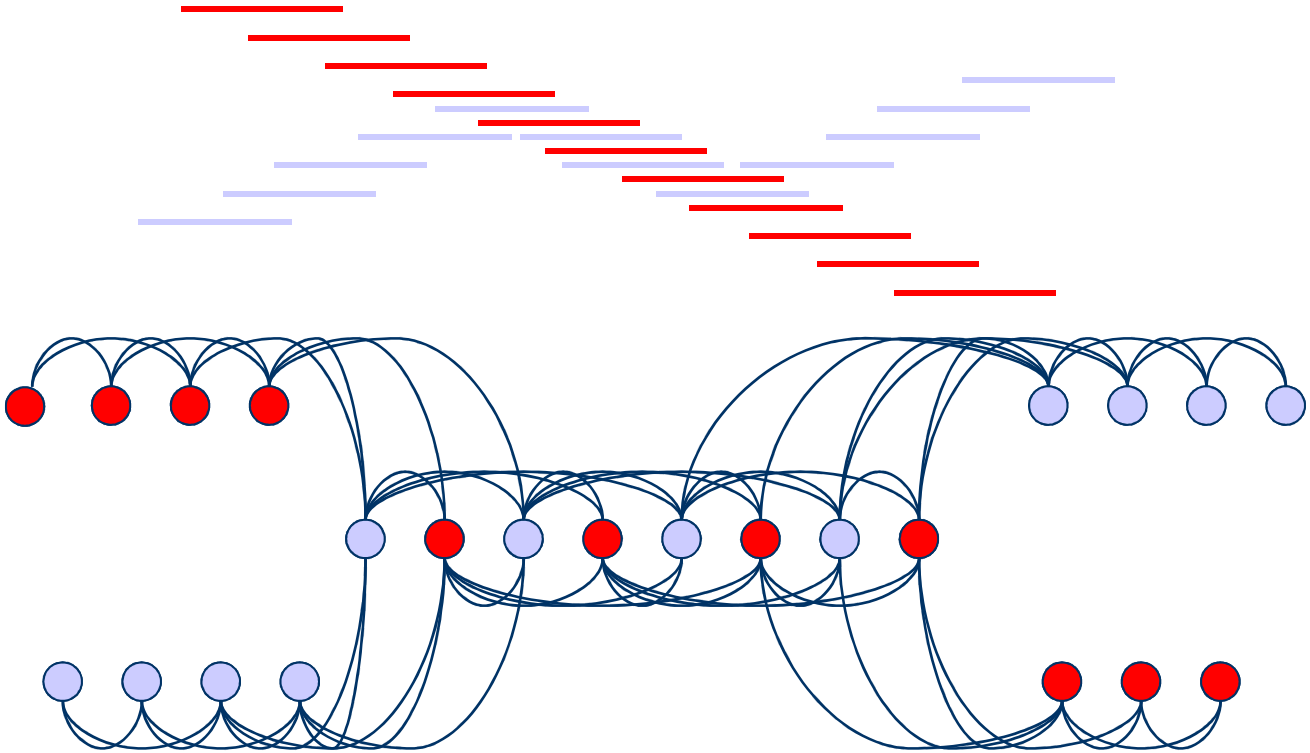
Spojování fragmentů do contigů (Layout)

- Cílem je spojit překrývající se fragmenty (ready) do větších částí (contigů) a dále ještě větších částí (supercontigů)
  - Velikost contigu je cca 30kb
  - Bez opakování je to přímočarý proces, opakování celý proces výrazně komplikují
- Používané řešení:
  1. Fáze: skrýt opakující se sekvence a soustředit se pouze na unikátní úseky - contigy
  2. Fáze: propojení více contigů do větších úseků (supercontigů) a doplnění mezer z fragmentů tvořících opakování
- Poznámky:
  - Po první fázi zůstává cca 20% genomu nesestaveno
  - Vyžaduje větší úsilí ve druhé fázi výpočtu – je nezbytné využít propracované heuristiky

Spojování fragmentů do contigů (Layout)

- Sestavení grafů překrývajících se fragmentů
  - Vrcholy: jednotlivé ready $r_1,...,r_n$
  - Hrany: spojují ready, které se překrývají

    overlap(ri, rj, posun, orientace, skóre)

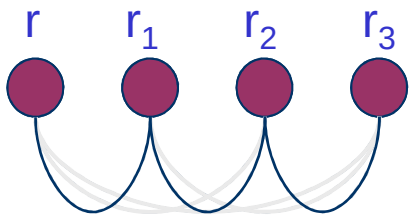


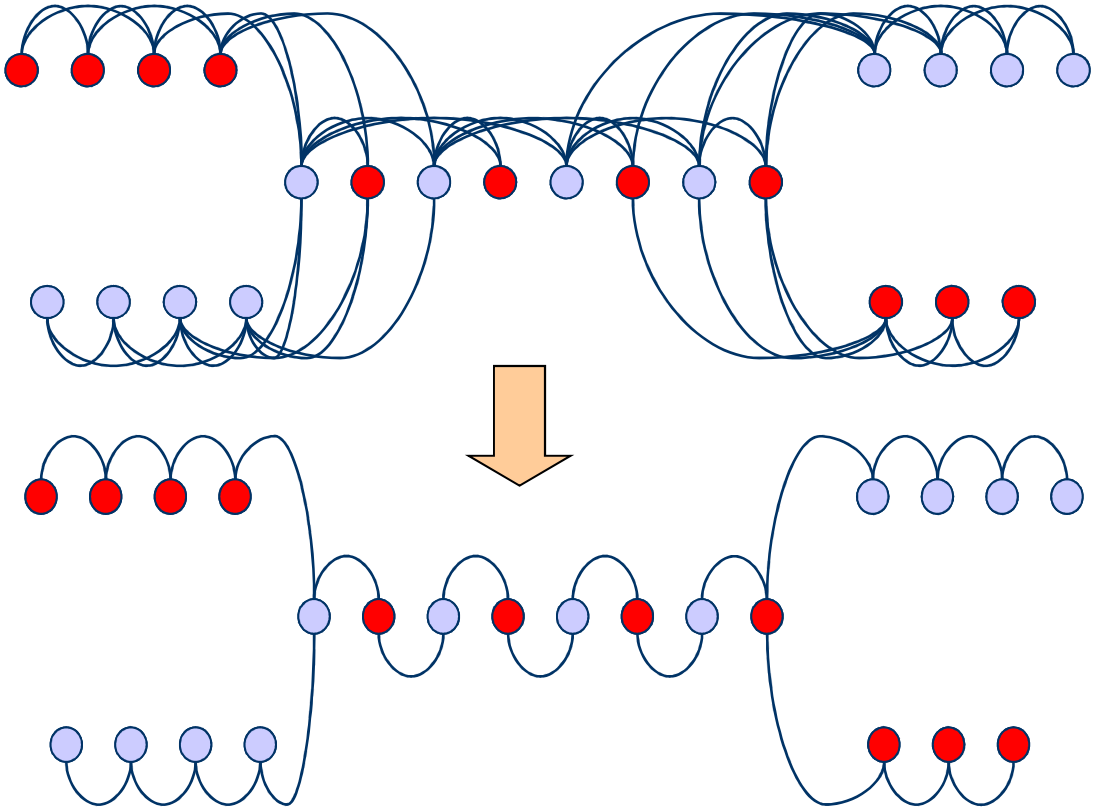
Příklad readů pocházejících z různých částí genomu (modrý a červený) a obsahují opakování

Poznámka: samozřejmě neznáme skutečné zabarvení jednotlivých uzlů

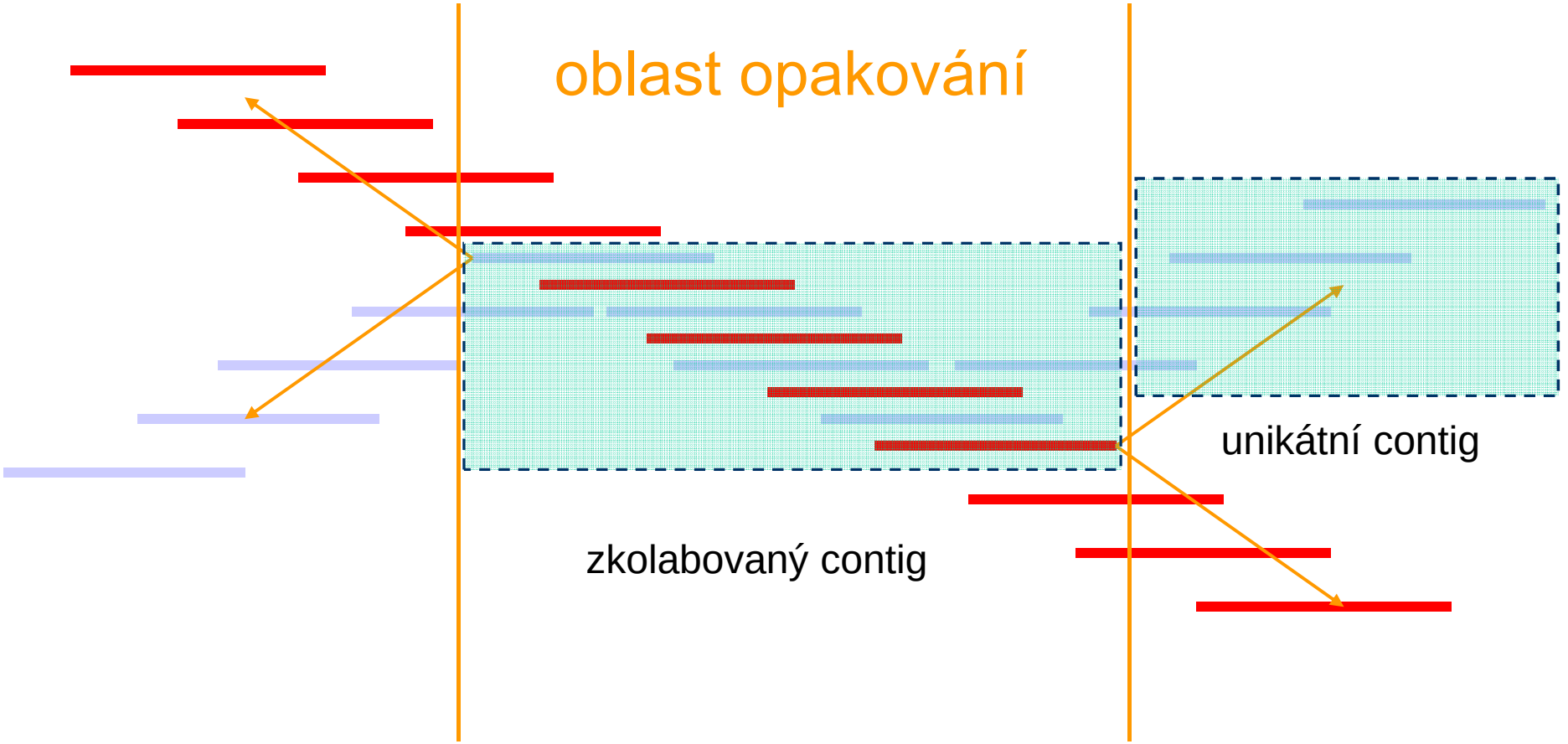
Spojení fragmentů do contigů

- Odstraň tranzitivní překryvy
  - Jestliže read *r* překrývá ready r_1, r_2, a r_1 překrývá r_2, potom (r, r_2) může být odvozeno z (r, r_1) a (r_1, r_2)



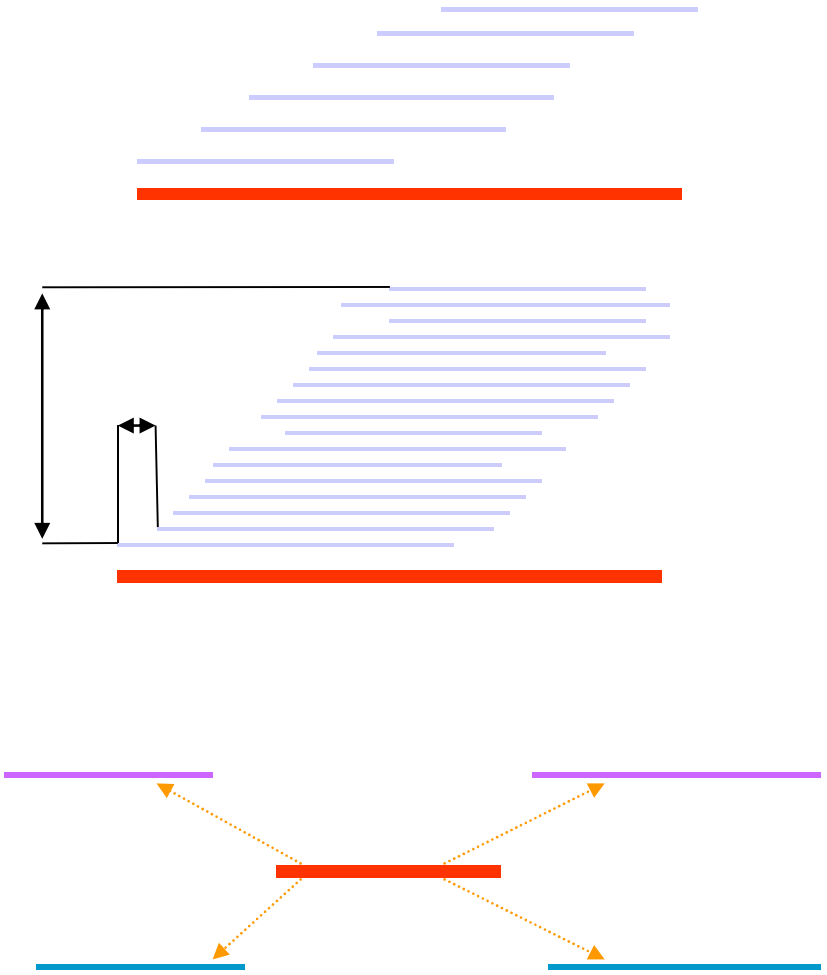


Spojování fragmentů do contigů (Layout)



- 1. Fáze: propojují jednotlivé ready až po potencionální hranice opakující se oblasti

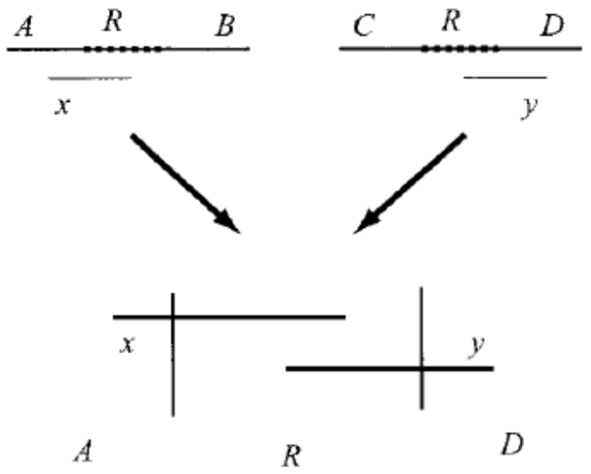
Spojování fragmentů do contigů (Layout)

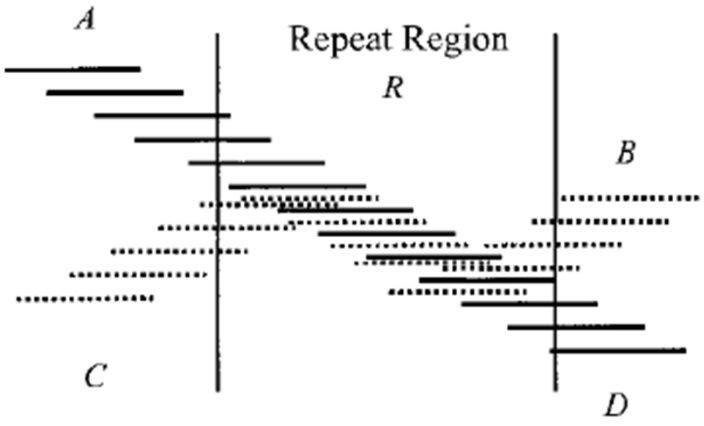


- Contigy obsahující opakování jsou obvykle rozpoznány dvěma faktory:
  1. Mají výrazně vyšší pokrytí překryvů a současně hustotu readů
  2. Obsahují vazby na více contigů, které se nepřekrývají

Spojování fragmentů do contigů (Layout)

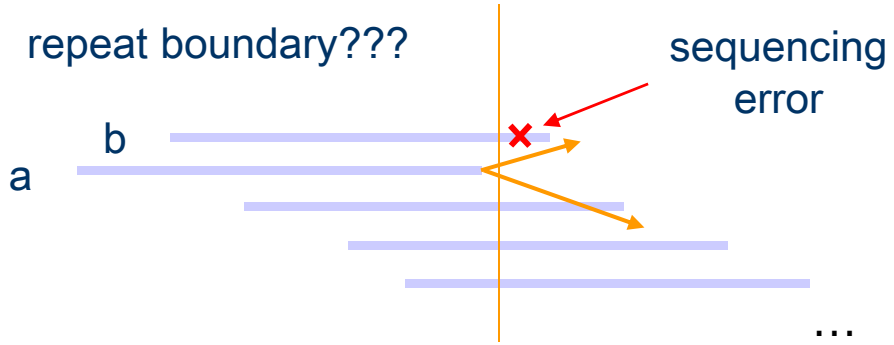
- Detekce hranic contigů obsahujících opakování
  - x se překrývá s R a R se překrývá s y, ale x a y se nepřekrývají - toto je základní pravidlo pro detekci opakujících se sekcí
  - obecně pokud všichni sousedi napravo od R se navzájem překrývají, potom mohou být spolehlivě spojeny do contigů, v opačném případě se může jednat o opakování

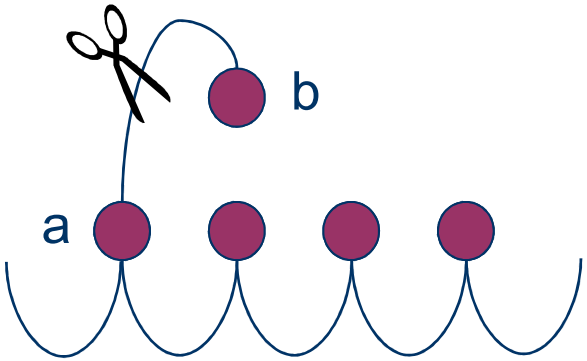




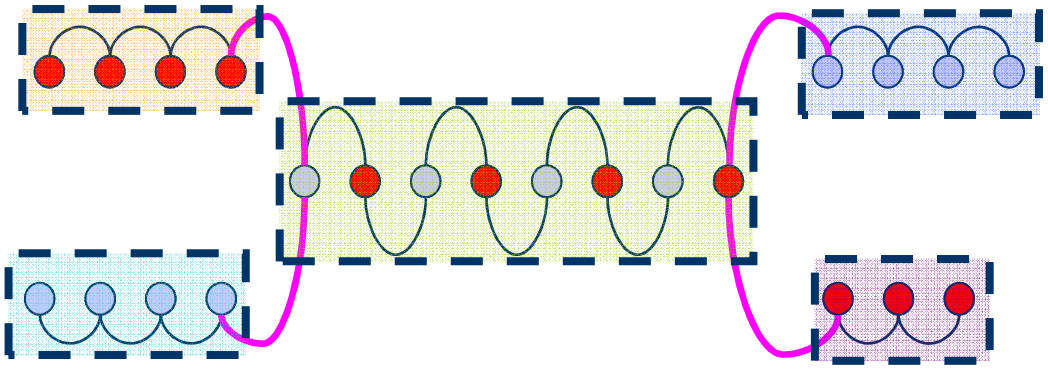
Spojení fragmentů do contigů

- Spojení unikátních úseků do contigů a ošetření možných chyb vzniklých při sekvenování

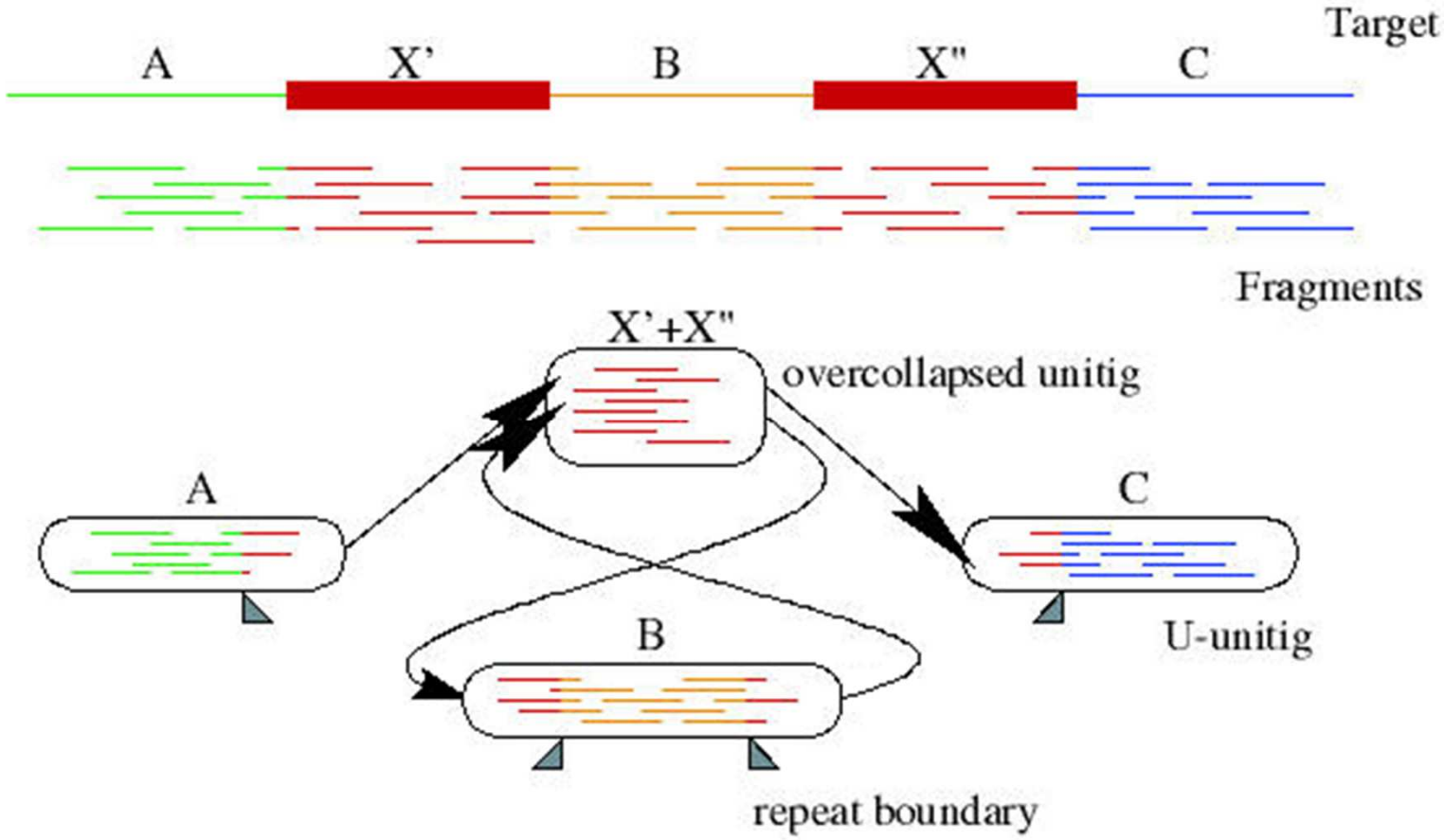




- Vytvořené contigy tvoří opět graf
  - rozlišujeme dvě skupiny contigů:
    1. obsahující unikátní sekvenci
    2. obsahující opakování

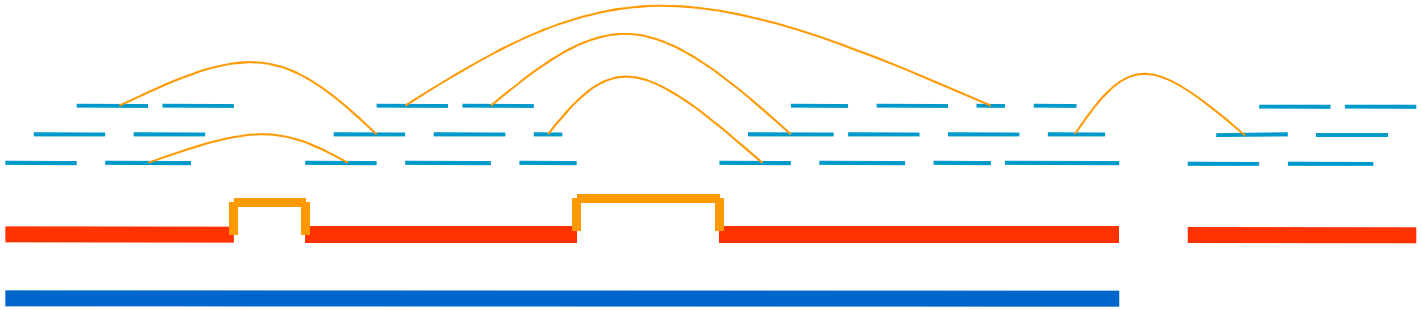


Graf po sestavení contigů



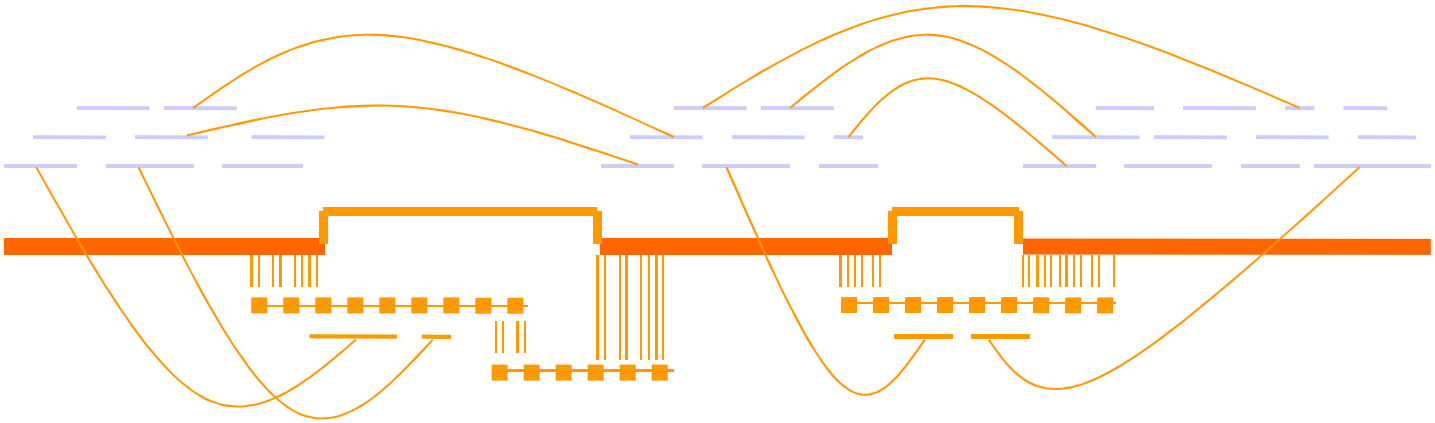
Spojení contigů do supercontigů

- 2. Fáze: spojování contigů do supercontigů
  - Spojování contigů probíhá na základě párových vazeb
  - Unikátní contigy jsou spojeny pokud obsahují dva a více vazeb
  - Heuristika seřazuje dvojice contigů do prioritní fronty, která preferuje:
    1. vetší počet spojů mezi contigy
    2. menší mezeru mezi contigy
  - Na základě párových readů se také identifikuje orientace contigů



Spojení contigů do supercontigů

- Vyplňování mezer mezi supercontigy
  - Mezery vznikají ve dvou případech
    1. Výskyt opakování
    2. Oblasti s malým pokrytím při shotgun sekvenování
  - Mezery jsou vyplňovány především na základě párových vazeb doplněním fragmentů z contigů obsahujících opakování



Oprava chyb (Consensus)

- Na základě výsledného vytvoření supercontigů se vytvoří vícenásobné zarovnání spárovaných readů a provedou se opravy podobně jako tomu bylo u hledání překryvů - tentokrát už ve výsledné sekvenci se správným pořadím readů

Historie sekvenování

1976	MS2 (RNA virus) 40 kB
1988	projekt sekvenování lidského genomu (15 roků)
1995	H. influenzae 2 MB, shotgun (TIGR)
1996	S. cerevisiae (pivní kvasinka) 10 MB, bac-by-bac (NIH)
1998	C. elegans (háďátko) 100 MB, bac-by-bac (Wellcome Trust)
1998	Celera: lidský genom do tří let!
2000	D. melanogaster (vinná muška) 180 MB, shotgun (Celera, Berkeley)
2001-3	2x lidský genom 3 GB (NIH, Celera)
2002-7	Myš, potkan, slepice, šimpanz, pes, makak
2007	Watsonův a Venterův genom (454)

Budoucnost sekvenování

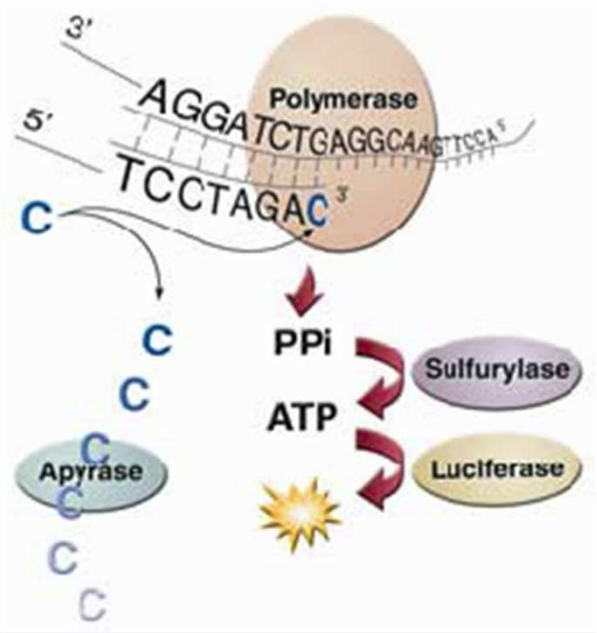
- Zrychlení
  - 1. Lidský genom: 15 let
  - Současná rychlost: řádově týdny
- Zlevnění
  - 1. Lidský genom: $3 miliardy
  - Současná cena: $10 000
- Personální genomika
  - Každý člověk bude mít svůj genom
  - Široké použití v oblasti medicíny
- Sekvenování nových organizmů
  - Sekvenování velké části žijících organizmů
  - Rekonstrukce genů předků
  - Nalezení všech funkčních elementů genů

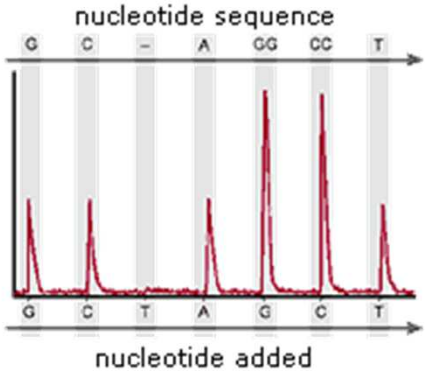
$1000 Genom

- Cílem projektu: získat lidský genom za cenu $1000 během několika dní
- Začátek projektu: 2006
- Dokončení projektu: konec 2009 (stále pokračuje)
- Nové technologie ve vývoji:
  - Pyrosekvenování
  - SMRT (Single Molecule Real-Time DNA Sequencing)
  - Nanopore sekvenování

Pyrosekvenování

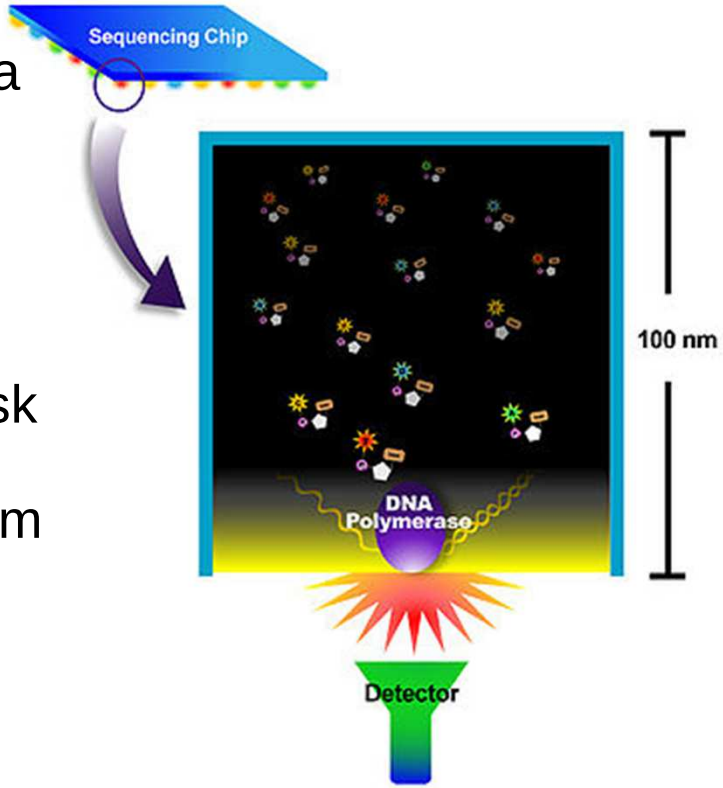
- Pricip:
  - Základem jsou 4 enzymy: DNA polymeráza, apyráza, luciferáza a ATP sulfuryláza
  - Postupně jsou vkládány nukleotidy dATP, dGTP, dCTP a dTTP
  - Pokud se nukleotid naváže, potom se uvolní pyrofosfát, který ve spojení s luciférázou vytvoří světelný záblesk
  - Intenzita záblesku určuje počet navázaných nukleotidů za sebou
- Rychlost: báze/90s, možnost paralelizace
- Velikost readu: cca 400 bází
- Technologické problémy s dalším zlepšováním metody





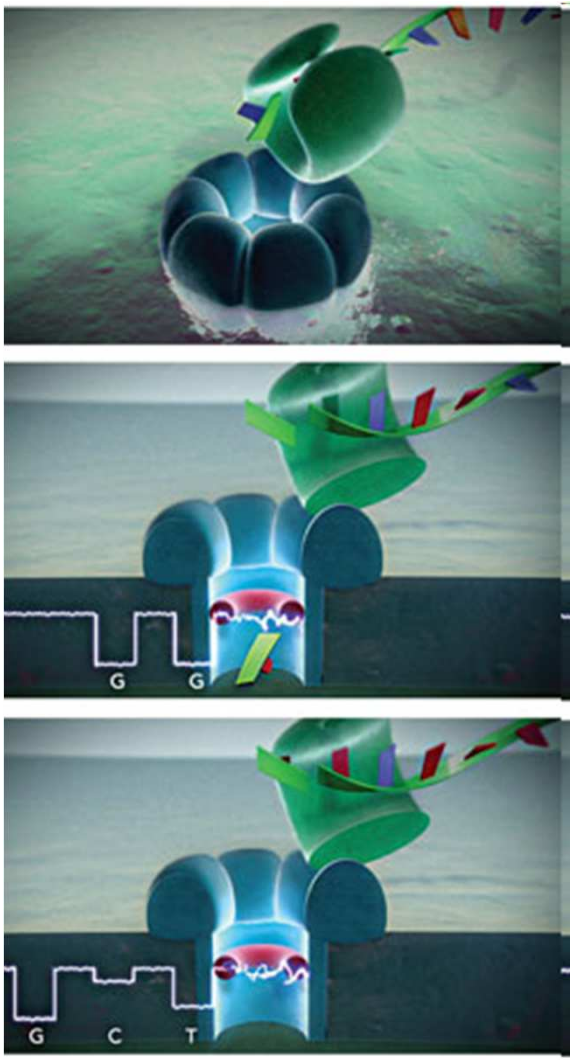
SMRT (Single Molecule Real-Time DNA Sequencing)

- Pacific Biosciences
- Princip:
  - Využívá DNA Polymerázu, ktera je schopna replikovat celou molekulu DNA v rámci minut
  - Fluorescenčně obarvené nukleotidy: každý typ nukleotidu jinou barvičkou.
  - Jakmile se nukleotid naváže na vlákno barvivo se uvolní a vytvoří intenzivní záblesk
  - Celý proces probíhá v nádobce o šířce 70nm, uzavřené sklíčkem a snímané citlivým senzorem
- Výhody:
  - Vytváří velmi dlouhé ready
  - Rychlost: 10 bází za sekundu, paralelizace
  - Doba sekvenování DNA v řádu minut, celý proces sestavování v rámci hodiny
  - Předpokládaná cena: $100



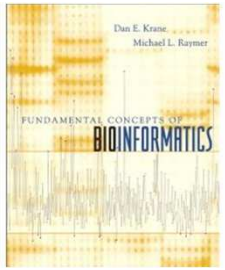
Nanopore sekvenování

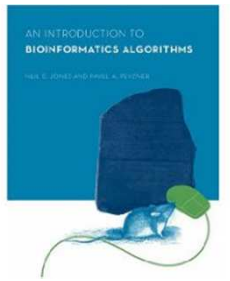
- Oxford Nanopore Technologies
- Princip:
  - Molekula DNA prochází přes úzké místo (nanopore) o velikosti jednotek nanometrů
  - Nanopore měří velikost nápoje, která je pro jednotlivé nukleotidy různá
- Výhody:
  - Není potřeba PCR (lze sekvenovat i nepatrné množství DNA)
  - Velmi vysoká délka readu (sta-tisíce bází)
  - Velmi nízká cena
- Stále se řeší problémy s:
  - Fyzická realizace nanopore
  - Citlivosti sondy pro měření nápoje



Literatura

- Dan K. Krane, Michael L. Raymer: ***Fundamental Concepts of Bioinformatics***, ISBN: 0-8053-4633-3, Benjamin Cummings 2003.
- Neil C. Jones, Pavel A. Pevzner, ***An Introduction to Bioinformatics Algorithms***, ISBN-10: 0262101068, The MIT Press, 2004





Konec

Děkuji za pozornost