Seminář z bioinformatiky

Zarovnání sekvencí z bisulfitového sekvenování

Eva Dobešová

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 26 no. 15 2010, pages 1901–1902*
doi:10.1093/bioinformatics/btq291

Sequence analysis Advance Access publication June 18, 2010

An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System

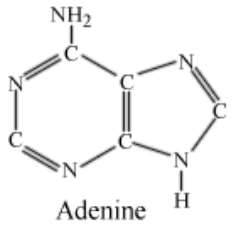
Brian D. Ondov[1,2], Charles Cochran[3], Mark Landers[4], Gavin D. Meredith[4], Miroslav Dudas[4] and Nicholas H. Bergman[1,2,*]

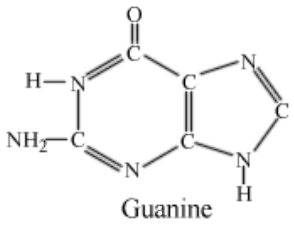
[1]National Biodefense Analysis and Countermeasures Center, 110 Thomas Johnson Drive, Frederick, MD 21702, [2]School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, GA 30332-0230, [3]Life Technologies, 850 Lincoln Centre Drive, Foster City, CA 94404 and [4]Invitrogen, a division of Life Technologies Corporation, Genetic Systems Business Unit, 5791 Van Allen Way, Carlsbad, CA 92008, USA

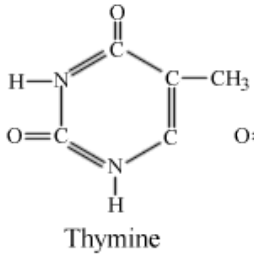
Associate Editor: Dmitrij Frishman

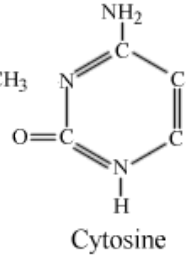
Úloha 5meC v DNA

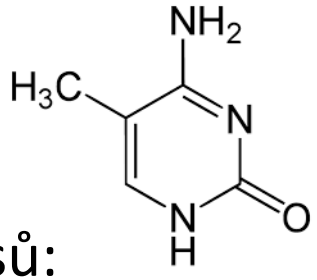
- DNA: A, G, T, C









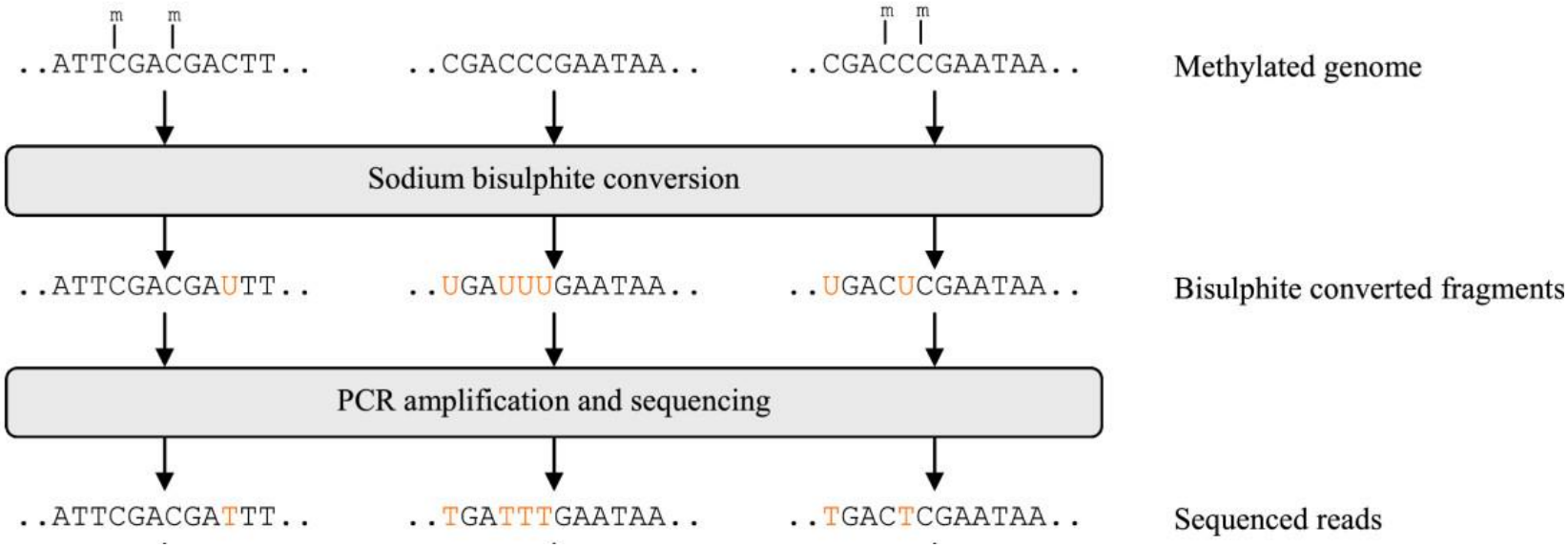


- 5. báze v DNA: 5-methylcytosin
  - Epigenetická modifikace ovlivňující spoustu procesů:
  - Embryonální vývoj, genomový imprinting, strukturu chromatinu, transkripci genů, umlčování transpozonů…
  - Abnormální methylace byla objevena u Alzheimerovy choroby a rakoviny.

- Jak stanovit 5meC v DNA?

Bisulfitové sekvenování

- Bisulfit: konvertuje nemethylované cytosiny na uracily (po PCR převedeny na thyminy), 5meC zůstává nezměněn.

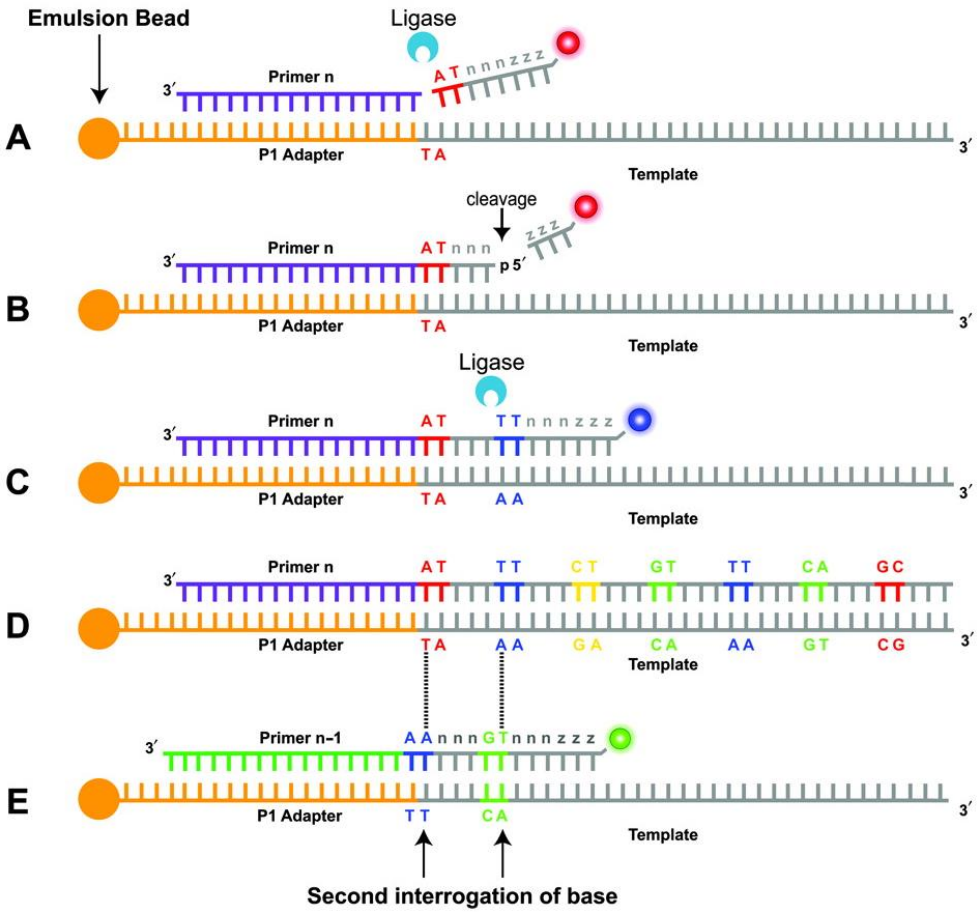


Hardcastle *PlantMethods* 2013, **9**:16

- Umožňuje zachytit methylaci cytosinů na úrovni jediného nukleotidu
- Dochází ke změně v sekvenci C-G → T-A

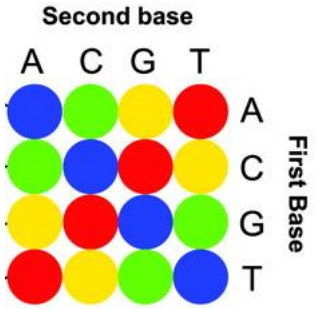
SOLiD

- Princip sekvenace:



- Dříve: délka 35 bp a přesnost 99,85 %
- Dnes: 85 bp a 99,99%

- Na rozdíl od metod 454 a Illumina využívá sekvenování na základě ligace (ale PCR reakce při přípravě knihovny)
- Využívá tzv. dvoubázové kódování:



- Každá báze je čtena dvakrát, možné rozeznání chyb od substitucí
- Výstup: color-space reads

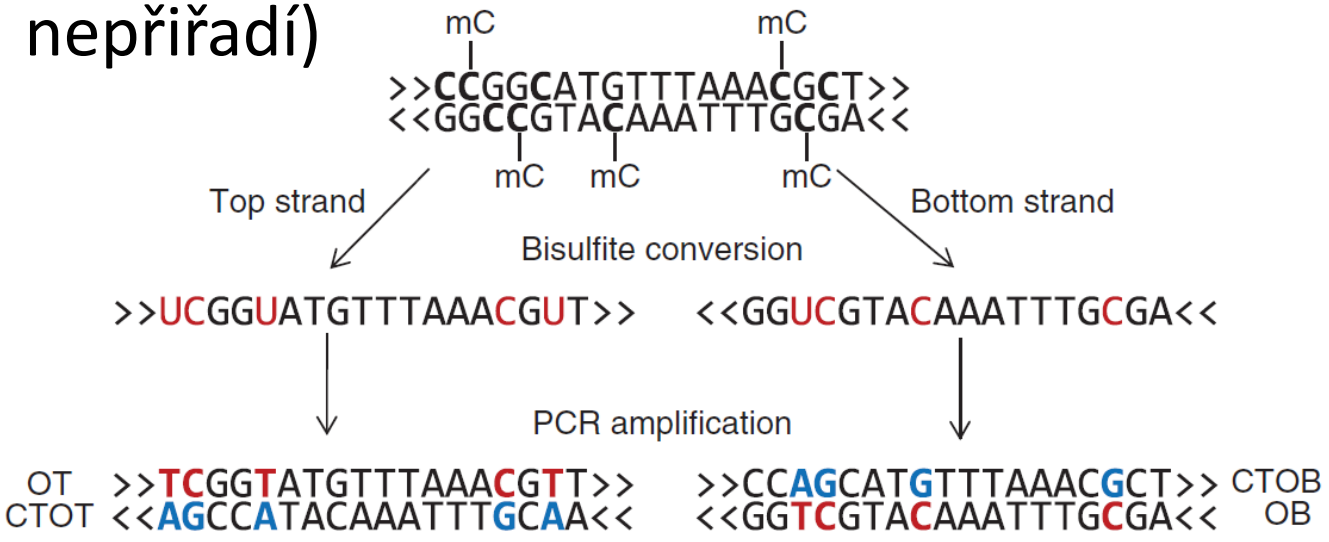
- Jak zarovnat sekvence, které obsahují hodně „mutací“?
  - těžko

Odkud obtížnost pramení:

- Sekvenované ready mají místo některých C → T, resp. místo G → A (na druhém vláknu)
- SOLiD – snadná detekce SNPs, ale bisulfitem indukované nukleotidové substituce (BINS) jsou časté a způsobují příliš mnoho nesprávných spojení (mismatch).

Přístupy zarovnání sekvencí

- Použít 4 referenční sekvence: původní a plně konvertovanou bisulfitem pro obě vlákna (Watson a Crick vlákno) → zarovnat tradičními nástroji pro SOLiD
  - Problém: ready obsahují jak methylované tak nemethylované cytosiny („napůl methylované“ se nepřiřadí)



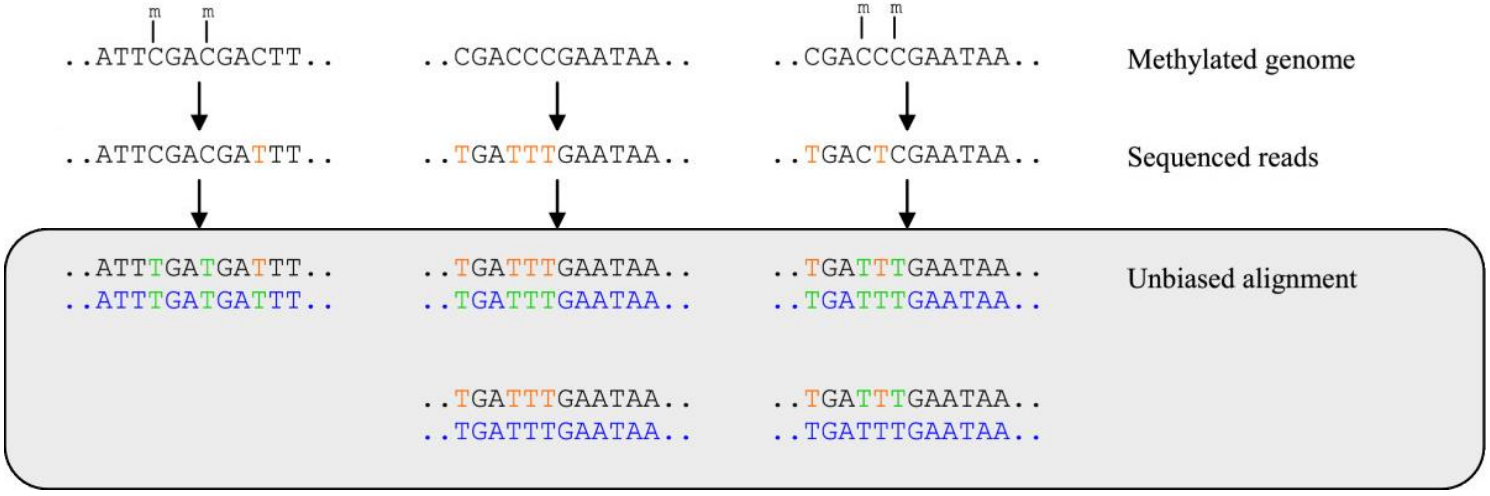
Přístupy zarovnání sekvencí

- Použít 4 referenční sekvence: původní a plně konvertovanou bisulfitem pro obě vlákna (Watson a Crick vlákno) → zarovnat tradičními nástroji pro SOLiD
  - Problém: ready obsahují jak methylované tak nemethylované cytosiny
- Převést SOLiD ready z dvounukleotidového kódování na řetězec nukleotidů a zarovnat existujícími nástroji, které jsou tolerantní k BINS (předpokládá absenci sekvenačních chyb).

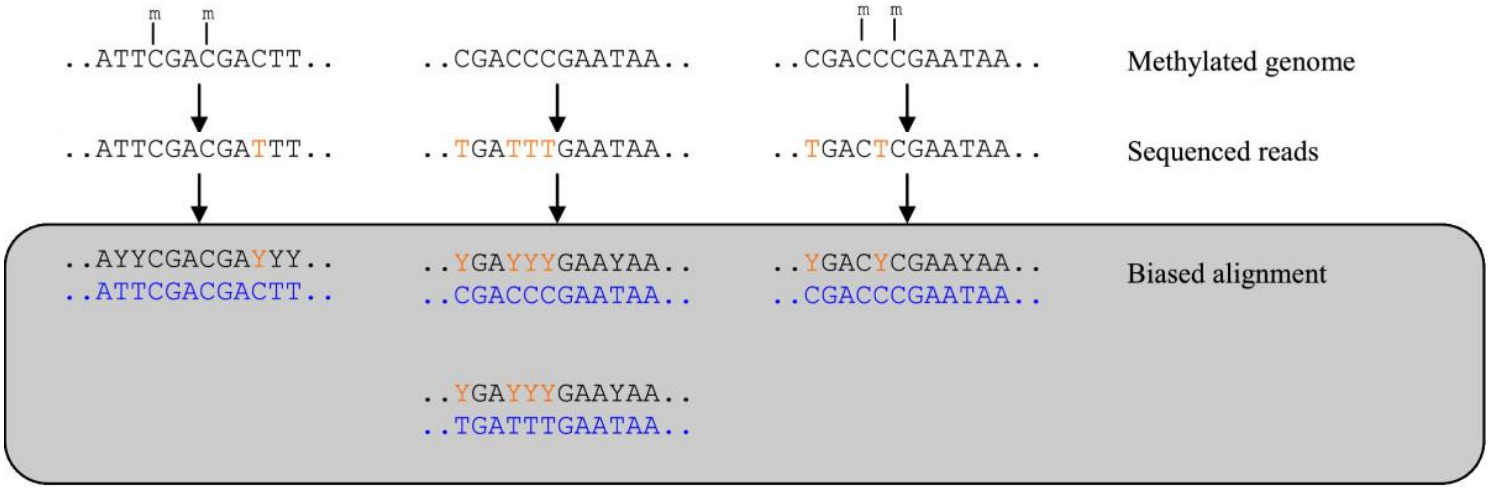
Zarovnání sekvencí vyjádřených jako řetězec nukleotidů

- Zarovnání BS-seq dat – menší množství informace dostupné pro zarovnání k referenčnímu genomu oproti klasické sekvenaci.
- 2 přístupy:
  - unbiased alignment (zarovnání bez chyb-mismatchů)
  - biased alignment

- Unbiased alignment (three-letter alignment):
  - *In silico* konverze všech cytosinů na thyminy, (sekvenované readech i referenční sekvence)
  - Výsledkem jsou sekvence obsahující pouze 3 báze (A,G,T), méně readů je přiřazeno jednoznačně
  - BS-Seeker,Bismark, MethylCoder, BRAT



- Biased alignment (Wildcard alignment):
  - Sekvenované thyminy jsou konvertovány na Y (C nebo T), resp. C v genomu → Y
  - Jednoznačnější přiřazení než u unbiased alignment, využívá maximální možnou informaci
  - BSMap, RMAP-BS



- Biased alignment (Wildcard alignment):
  - Sekvenované thyminy jsou konvertovány na Y (C nebo T), resp. C v genomu → Y
  - Jednoznačnější přiřazení než u unbiased alignment, využívá maximální možnou informaci
  - BSMap, RMAP-BS
  - Možné využít upravenou skórovací matici

	a	c	g	t
a	6	-18	-18	-18
c	-18	6	-18	3
g	-18	-18	6	-18
t	-18	-18	-18	3

Sloupce znázorňují báze v sekvenovaných readech, řádky báze v referenčním genomu.

Frith *Nucleic Acids Research* 2012, **40**:13

SOCS -B

- Algoritmus na zarovnání „barevných“ sekvencí, který dovoluje bisulfitem indukované nukleotidové substituce i sekvenační chyby.
- Spustitelný soubor ani zdrojový kód již nejsou dostupné
- Popis algoritmu: založený na Rabin(ově)-Karp(ově) alg.
  - Vytvoření hašovací tabulky (snižuje počet zarovnání)
  - Ohodnocení zarovnaní porovnáním sekvenovaných readů vyjádřených v barevném módu s „barevnou“ referenční sekvencí (obě fáze tolerantní k BINS i sekv. chybám)
    - Hashe jsou vypočítány na základě překladu „barevných“ readů do nukleotidové sekvence. Počítají se všechny 4 překlady.
  - Klíče upraveny tak, že berou C a T jako stejný symbol.
- Pro výpočet nejpravděpodobnějšího stavu každého mC využívá dynamické programování

Citace Scopus: 1x2011, 5x2012, 6x2013, ?2014

Rabinův-Karpův algoritmus (1987)

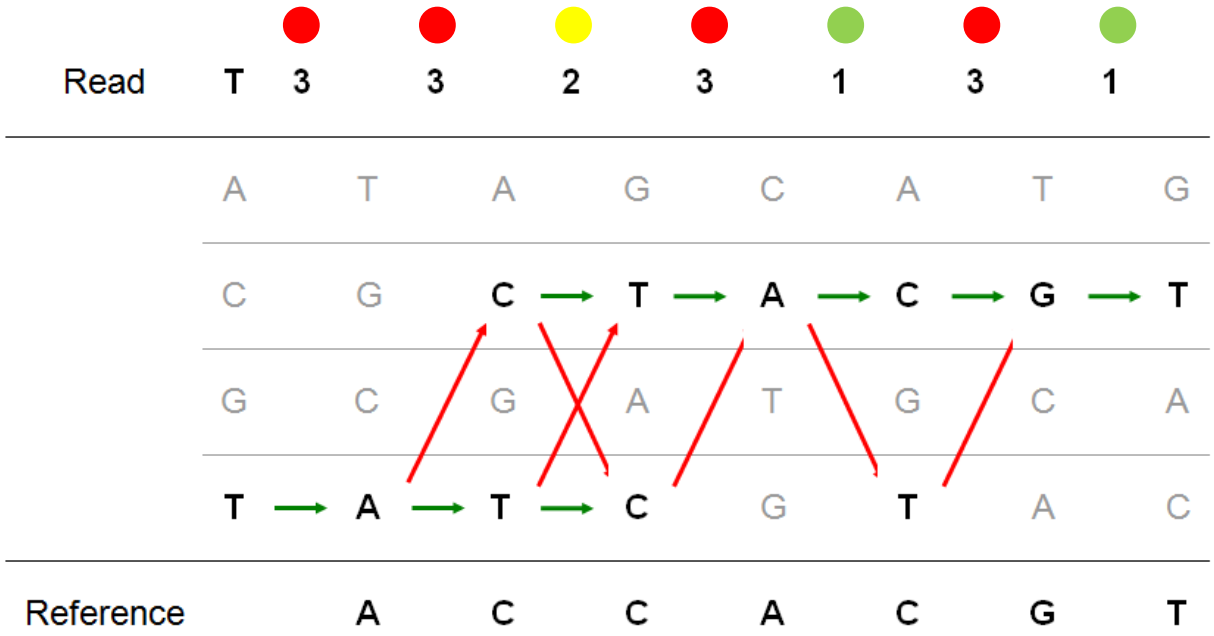
- Algoritmus pro vyhledávání v textu, využívající hashovací funkce.

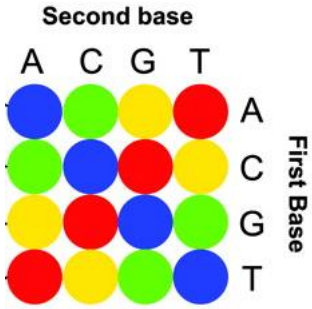
```
function RabinKarpSet(string s[1..n], set of string subs, m):
    set hsubs := emptySet
    for each sub in subs
        insert hash(sub[1..m]) into hsubs
    hs := hash(s[1..m])
    for i from 1 to n-m+1
        if hs ∈ hsubs and s[i..i+m-1] ∈ subs
            return i
        hs := hash(s[i+1..i+m])
    return not found
```

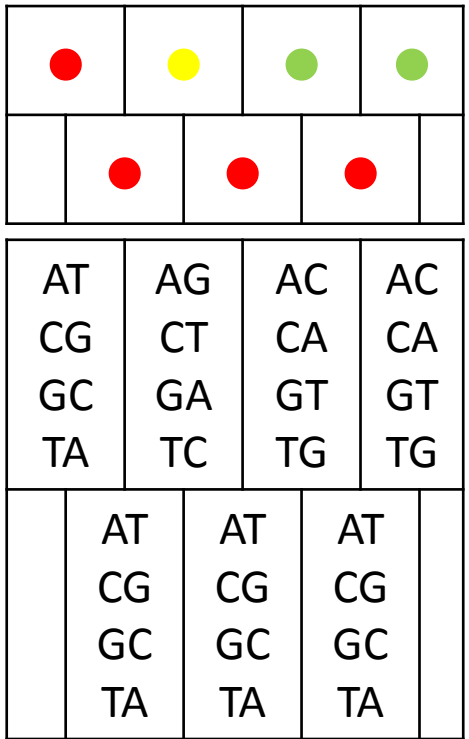
Wikipedia – Rabinův-Karpův algoritmus

Předpokládáme, že všechny podřetězce mají pevnou délku **m**

Příklad dynamické programovací tabulky (SOCS-B)







Tmavá písmena označují povolené stavy. První nukleotid musí odpovídat sekvenci adaptoru. Posun po zelených šipkách neovlivňuje skóre, červené šipky odpovídají sekvenačním chybám (3. a 4. nukleotid = CT, TC, TT – 1 sekvenační chyba). Pozice chyby je určena tam, kde měla barva nejnižší kvalitu.

Testování algoritmu

- Analyzováno bylo 54 705 478 readů z genomu *A. thaliana* (bisulf.)
- Kontrola: nástroj pro zarovnání poskytovaný Applied Biosystems (*mapreads*), referenční sekvence: plně konvertovaná Watson a Crick vlákna a nekonvertované Watson vlákno.
- SOCS-B: reference pouze nekonvertovaný genom.
- Algoritmus byl 2x citlivější u readů obsahujících 3 a méně chyb

Errors permitted	Mapreads (reads aligned)	SOCS-B (reads aligned)	SOCS-B increase factor
0	1 150 378	8 701 800	7.56
1	3 283 347	13 856 042	4.22
2	6 691 811	18 764 830	2.80
3	11 159 673	22 656 148	2.03

Technické parametry

- Doba běhu 30 h (Apple Mac Pro, dual 2.93 GHz Quad-Core Intel Xeon with hyper-threading, 32GB RAM)
- Více readů může být namapováno, pokud je povoleno více mismatch-ů, cena: delší doba výpočtu, nižší specificita.
- Možné distribuované zpracování.

Table 1 | Software tools for the analysis and interpretation of DNA methylation data

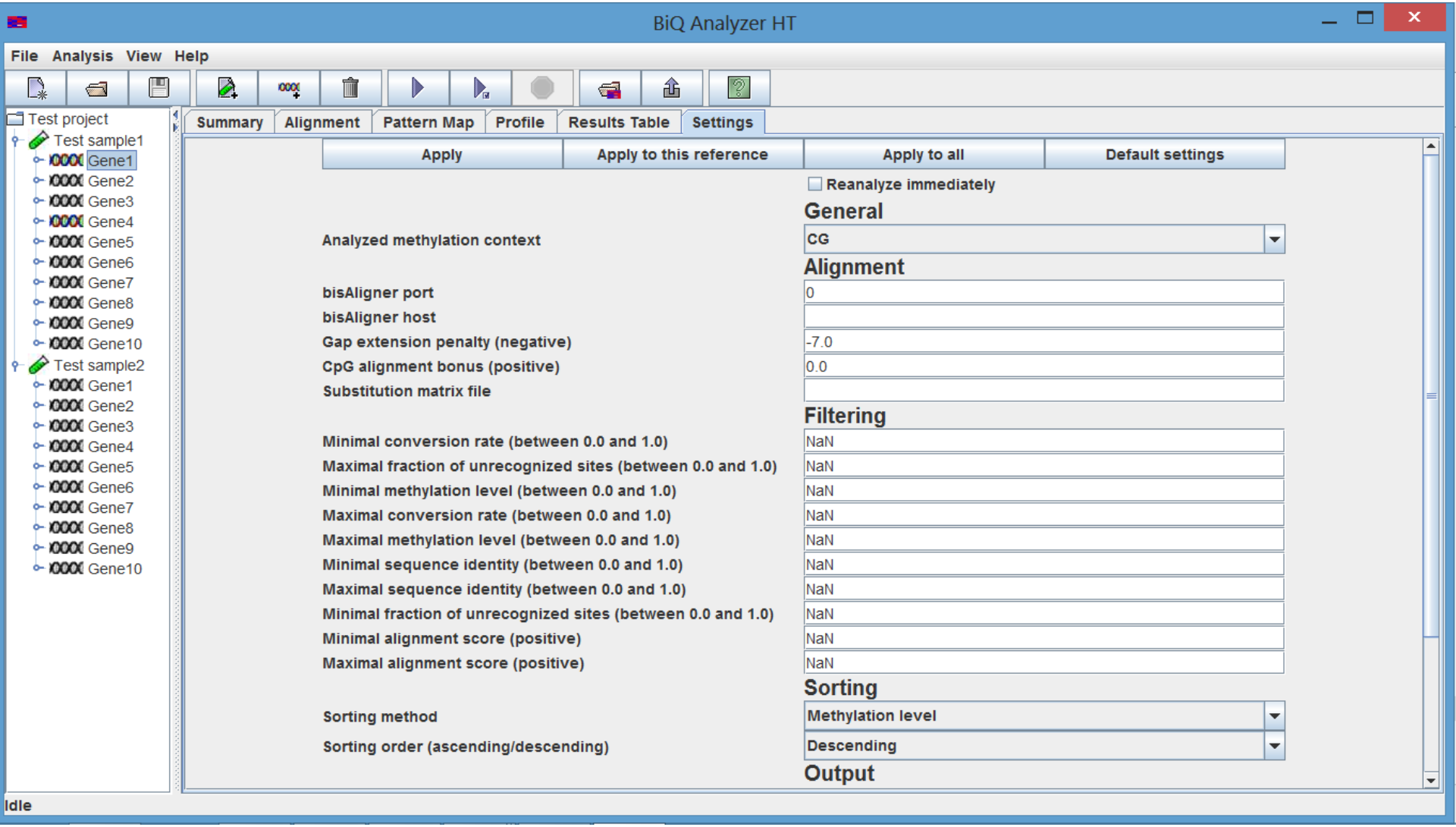
Software	Description	URL	Refs
Processing bisulphite-sequencing data			
B-SOLANA	Bisulphite aligner for processing bisulphite-sequencing data obtained in the two-base encoding of ABI SOLiD sequencers	http://code.google.com/p/bsolana	40
Bismark	Probably the most widely used three-letter bisulphite aligner; supports both Bowtie (fast, gap-free alignment) and Bowtie 2.0 (sensitive, gapped alignment)	http://www.bioinformatics.babraham.ac.uk/projects/bismark	28
Biz-SNP	Variant caller for inferring DNA methylation levels and genomic variants from bisulphite-sequencing reads that have been aligned by other tools	http://epigenome.usc.edu/publicationdata/bisznp2011	35
BRAT	Highly configurable and well-documented three-letter bisulphite aligner	http://compbio.cs.ucr.edu/brat	29,30
BS-Seeker	Basic three-letter bisulphite aligner based on Bowtie	http://pellegrini.mcdb.ucla.edu/BS_Seeker/BS_Seeker.html	31
BSMAP	Probably the most widely used wild-card bisulphite aligner	http://code.google.com/p/bsmap	21
GSNAP	Wild-card bisulphite aligner included in a widely used general-purpose alignment tool	http://share.gene.com/gmap	22
Last	Recent and well-validated wild-card bisulphite aligner included in a general-purpose alignment tool	http://last.cbrc.jp	23
MethylCoder	Three-letter bisulphite aligner that can be used with either Bowtie (high speed) or GSNAP (high sensitivity)	https://github.com/brentp/methylcode	32
Pash	Wild-card bisulphite aligner included in a general-purpose alignment tool	http://brl.bcm.tmc.edu/pash	24
RMAP	Wild-card bisulphite aligner included in a general-purpose alignment tool	http://www.cmb.usc.edu/people/andrewds/rmap	25
RRBSMAP	Variant of BSMAP that is specialized on reduced-representation bisulphite sequencing (RRBS) data	http://rrbsmap.computational-epigenetics.org	26
segemehl	Wild-card bisulphite aligner included in a general-purpose alignment tool	http://www.bioinf.uni-leipzig.de/Software/segemehl	27
Processing bisulphite microarray data			
ComBat	R script for correcting known or suspected batch effects using an empirical Bayes method	http://www.bu.edu/jlab/wp-assets/ComBat	52
Illumina BeadScan	Machine control and image processing software for Illumina Infinium microarray scanners	http://www.illumina.com/support/array/array_instruments/beadarray_reader.ilmn	
Illumina GenomeStudio	Graphical tool for data normalization, analysis and visualization of Illumina Infinium microarrays (and other genomic data types)	http://www.illumina.com/software/genomestudio_software.ilmn	
isva	R package for batch effect correction using an algorithm that is based on singular value decomposition	http://cran.r-project.org/web/packages/isva	50
methylumi	R/Bioconductor package for Infinium data normalization and general data handling	http://www.bioconductor.org/packages/release/bioc/html/methylumi.html	
minfi	R/Bioconductor package for Infinium data normalization, analysis and visualization	http://www.bioconductor.org/packages/release/bioc/html/minfi.html	
RnBeads	R package providing a software pipeline for Infinium data normalization, quality control, exploratory visualization and differentially methylated region (DMR) identification	http://rnbeads.computational-epigenetics.org	
SVA	R/Bioconductor package for correcting batch effects that are directly inferred from the data using surrogate variable estimation	http://www.bioconductor.org/packages/release/bioc/html/sva.html	53

Bock *Nature Reviews Genetics* 2012, **13**

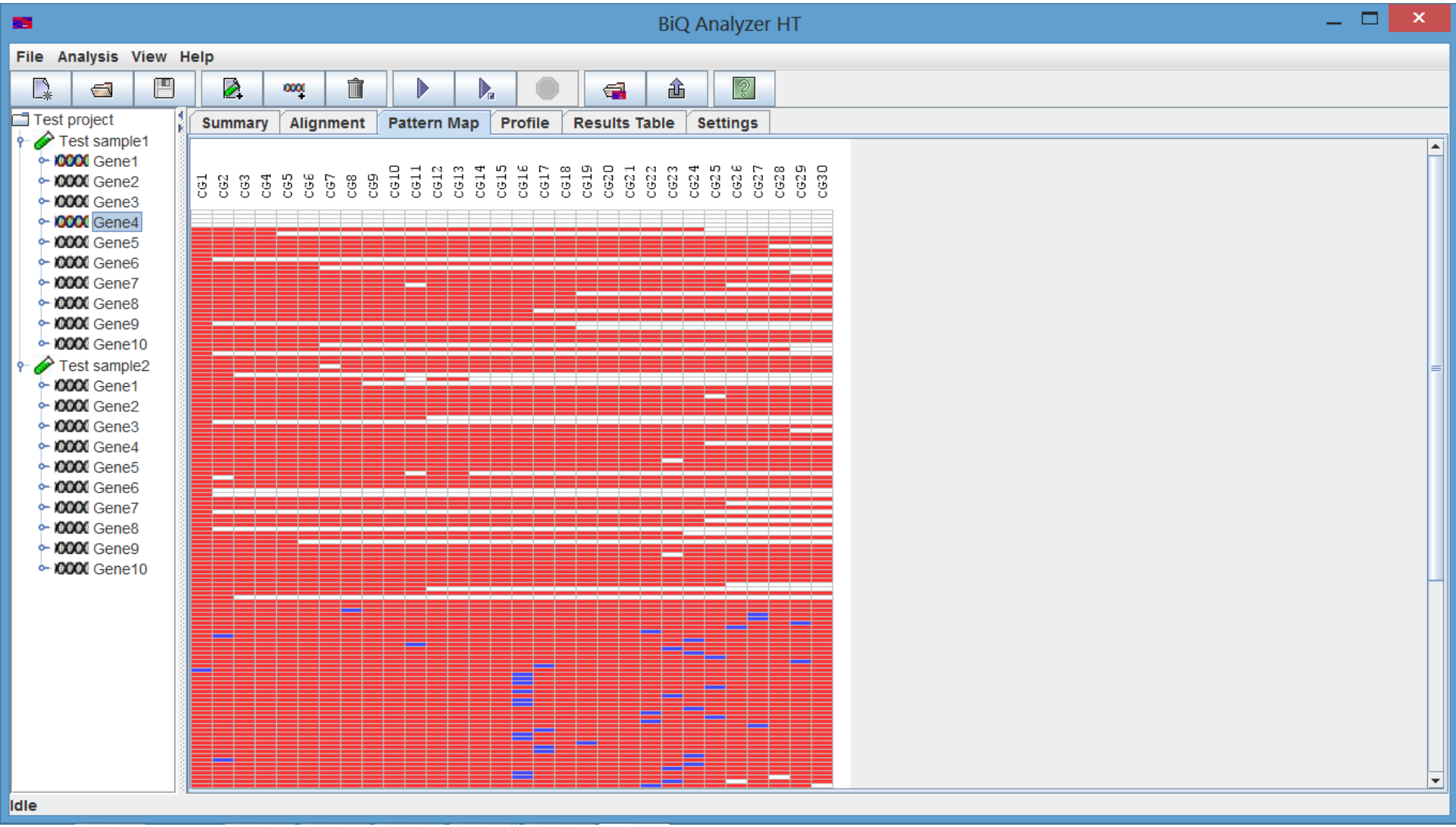
Praktický příklad: BiQ Analyzer HT

- Místně specifická analýza DNA methylace
- Wildcard aligner
- Vhodný pro zpracování dat z 454 sekvencí
- Zarovnávací algoritmus: Needleman-Wunsch (lokální zarovnání)
- Program založený na jazyku Java – možné spuštění na jakémkoli počítači
- Příjemné grafické rozhraní pro „biology“, možná volba parametrů a filtrování dat

Lutsik *Nucleic Acids Research*, 2011, 1–6


BiQ Analyzer HT
File Analysis View Help
Test project
Test sample1
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
Gene7
Gene8
Gene9
Gene10
Test sample2
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
Gene7
Gene8
Gene9
Gene10
Summary
Alignment
Pattern Map
Profile
Results Table
Settings
Apply
Apply to this reference
Apply to all
Default settings
Reanalyze immediately
General
Analyzed methylation context
CG
Alignment
bisAligner port
0
bisAligner host
Gap extension penalty (negative)
-7.0
CpG alignment bonus (positive)
0.0
Substitution matrix file
Filtering
Minimal conversion rate (between 0.0 and 1.0)
NaN
Maximal fraction of unrecognized sites (between 0.0 and 1.0)
NaN
Minimal methylation level (between 0.0 and 1.0)
NaN
Maximal conversion rate (between 0.0 and 1.0)
NaN
Maximal methylation level (between 0.0 and 1.0)
NaN
Minimal sequence identity (between 0.0 and 1.0)
NaN
Maximal sequence identity (between 0.0 and 1.0)
NaN
Minimal fraction of unrecognized sites (between 0.0 and 1.0)
NaN
Minimal alignment score (positive)
NaN
Maximal alignment score (positive)
NaN
Sorting
Sorting method
Methylation level
Sorting order (ascending/descending)
Descending
Output
Idle


BiQ Analyzer HT
File Analysis View Help
Test project
Test sample1
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
Gene7
Gene8
Gene9
Gene10
Test sample2
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
Gene7
Gene8
Gene9
Gene10
Summary
Alignment
Pattern Map
Profile
Results Table
Settings
Methylation summary per site
CG1 CG2 CG3 CG4 CG5 CG6 CG7 CG8 CG9 CG10 CG11 CG12 CG13 CG14 CG15 CG16 CG27 CG28 CG29 CG30 CG31 CG32 CG33 CG34 CG35 CG36 CG37 CG38 CG39 CG40 CG41 CG42 CG43 CG44 CG45 CG46 CG47 CG48 CG49 CG50 CG51 CG52 CG53 CG54 CG55
Idle


BiQ Analyzer HT
File
Analysis
View
Help
Test project
Test sample1
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
Gene7
Gene8
Gene9
Gene10
Test sample2
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
Gene7
Gene8
Gene9
Gene10
Summary
Alignment
Pattern Map
Profile
Results Table
Settings
CG1 CG2 CG3 CG4 CG5 CG6 CG7 CG8 CG9 CG10 CG11 CG12 CG13 CG14 CG15 CG16 CG17 CG18 CG19 CG20 CG21 CG22 CG23 CG24 CG25 CG26 CG27 CG28 CG29 CG30
Idle