TARTU ÜLIKOOL

Uudiseid k-meride abil bakterite leidmisest

CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers (2015)

Rachid Ounit, Steve Wanamaker, Timothy J. Close and Stefano Lonardi

Ja

CoMeta: Classification of Metagenomes Using k-mers (2015)

Jolanta Kawulok and Sebastian Deorowicz

Bioinformaatika ajakirjaklubi

8. Oktoober 2015

Märt Roosaare

CLARK

- CLAssifier based on Reduced K-mers
- Määrab iga lugemi eraldi (nagu Kraken)
- Suur töökiirus – väidetavalt kiirem kui Kraken
- Suudab määrata viiruseid, baktereid ja ilmselt ka muud (olenevalt andmebaasi ülesehitusest), aga mõeldud eelkõige bakterite detekteerimiseks
- Mõeldud töötama eelkõige liigi/perekonna tasemel (tase tuleb ise määrata alguses)
- Ei kasuta juhtpuud määramisprotsessis

CLARK - andmebaas

- Võimalik kasutajal ise luua, defineerides eraldi iga „märklaua“ ning nende jaotuse mingi kindla taksonoomia piires
- Andmebaasi ei saa alla laadida, ehitatakse algul üles (2700+ RefSeq bakterit üle 8h)
- Igast märklauast tehakse k-meride list ja iga k-meri kohta märgitakse märklauad, kus nad esinevad ning esinemiste arv
- Viimase sammuna visatakse minema kõik k-merid mis esinevad rohkem kui ühes märklauas

TARTU ÜLIKOOL
1632

CLARK - märklauad

/storage8/bakter1.fasta bacillus
/storage8/bakter2.fasta bacillus
/storage8/bakter3.fasta clostridium
/storage8/bakter4.fasta bacillus
/storage8/bakter5.fasta clostridium
/storage8/bakter6.fasta bacillus

.txt fail kus on FASTA aadress ning kasutaja poolt määratud taksonoomiline klass (ei ole NCBI taksonoomiast sõltuv)

Automaatselt ehitatakse andmebaas mitte taksonite nimede, vaid ID-numbritega

CLARK - töötõhimõte

1. Sisse antud lugemid tehakse k-merideks
2. Otsitakse kõik k-merid lugemist, mis esinevad andmebaasis
3. Lugem määratakse sellele taksonile, mille k-mere esineb kõige rohkem, juhtpuud ei kasutata ning seega saab määrata ainult üht taksonoomilist taset
4. Määramisele antakse ka usaldusväärtus, valemiga: conf = $h_1/(h_1+h_2)$ | h = k-meri esinemise arv

CLARK – erinevad variandid

1. CLARK – L (light) – analüüsib vaid osa k-meeridest (iga 27-meri kohta jäetakse 4 järgmist mitte ülekattuvat vahele)
2. CLARK – e (express) – analüüsitakse samuti vähendatud hulka k-mere (ainult mitte ülekattuvad) ning määramine tehakse esimese k-meri järgi mis andmebaasist vaste saab
3. CLARK põhiprogramm kiirendatud kujul – kui üks andmebaasis olev organism on rohkem kui poolte võimalike k-meridega matchi saanud määratakse kohe ja lõpetatakse edasine analüüs

CLARK-L ja CLARK-e väga sarnane miniKrakenile

TARTU ÜLIKOOL
1632

CLARK – võrdlus NBC ja Krakeniga

Table 2 Species-level classification accuracy and speed of CLARK, KRAKEN, and NBC for four simulated metagenomes

	HiSeq			MiSeq			simBA-5		
	Prec	*Sens*	*Speed*	*Prec*	*Sens*	*Speed*	*Prec*	*Sens*	*Speed*
NBC (*k*=15)	68.67	68.70	0.008	68.33	68.33	0.007	91.74	91.74	0.007
CLARK (*k*=20)	69.44	61.46	272	70.72	62.45	239	91.32	82.48	269
KRAKEN (*k*=31)	74.00	53.49	2,332	77.72	58.72	1,361	92.99	78.70	1,976
CLARK (*k*=31)	86.74	58.59	3,011	89.49	61.84	1,566	98.85	76.80	2,855
KRAKEN-Q (*k*=31)	75.88	50.78	6,224	78.07	53.68	5,308	92.67	74.39	7,023
CLARK-*E* (*k*=31)	90.08	55.18	30,976	94.31	58.36	24,029	98.92	66.02	24,996
CLARK-*I* (*k*=27)	85.35	53.95	1,676	85.89	64.91	904	85.55	46.28	1,702

Precision and sensitivity are expressed as percentages, while speed is expressed in 10^3 reads per minute for NBC, KRAKEN, and CLARK on the classification of "HiSeq", "MiSeq", "sim
datasets against the 1473 species-level targets, in single-threaded mode.

CLARK on tundlikum ja kiirem kui Kraken. NBC jääb mängust välja, sest on liiga aeglane – praktiliselt ei ole võimalik kasutada

CLARK – võrdlus ja tulemused

1. Raske aru saada, mis andmebaasiga Krakenit kasutati (default andmebaasis on ka arhed ja viirused) – kas andmebaasid on võrdsed?
2. Kui k-meride pikkused on samad, siis peaks CLARK ja Kraken saama liigi kohta (Kraken liigi ja alamate taksonite) sama palju k-mere – kuidas saab Krakeni juhtpuu abil määramine anda halvema tulemuse kui CLARKi ilma juhtpuuta?
3. Krakeni tulemused erinevad üsnagi palju sõltuvalt kasutajast (vähemalt järgmise artikli andmetes)
4. CLARK tundub vägagi nagu Krakeni lihtsustatud variant, mis ei näita head terviklikku pilti (paljusid lugemeid ei suudeta ilmselt määrata liigi tasemel ja CLARK jätab need määramata, mõjutades seega üldpilti, kui kasutajat huvitab proovi koostis)
5. Andmebaas väga hästi seadistatav, võib olla nii kasulik (erinevad uued rakendused – resistentsusgeenide otsimine) kui kahjulik (suurt bakterite andmebaasi keerulisem teha)

CLARK – praktiline kasutamine

1. Väga hea lihtne alla laadida ja installida
2. Skriptid andmebaasi ja taksonoomia ehitamiseks-allatõmbamiseks NCBI andmebaasidest – programm ei leia neid hiljem üles ja tõmbab uuesti
3. Väljund .csv formaadis iga lugemi kohta, edasine küljendamine on keeruline (ei saanud täpselt aru), tõenäoliselt lihtsam ise skript kirjutada
4. Väljundist ei ole automaatselt aru saada, mis organismidega tegu (ainult NCBI tax ID numbrid)

TARTU ÜLIKOOL 1632

CLARK – testimine MAMBA-s

Sisendiks 12,3 Gbp lugemeid 10-st eri bakteriliigist:

Chromobacterium violaceum ATCC 12472 (gb: AE016825.1), Corynebacterium glutamicum ATCC 13032 (gb: BA000036.3), Escherichia coli ATCC 8739 (gb: CP000946.1), Escherichia coli W (gb: CP002185.1), Klebsiella pneumoniae KCTC 2242 (gb: CP002910.1), Polaromonas naphthalenivorans CJ2 (gb: CP000529.1), Pseudomonas stutzeri ATCC 17588 (gb: CP002881.1), Roseobacter denitrificans OCh 114 (gb: CP000362.1), and Staphylococcus epidermidis ATCC 12228 (gb: AE015929.1)

Faili analüüsimise ajad:

1. StrainSeeker – 41 minutit (tänu vähesele liikide hulgale vaja vähe puu harusid läbida)
2. Kraken – 1h 42min
3. CLARK – 1h 40m

Krakeni ja StrainSeekeri tulemused ühtisid suurel määral, CLARKi tulemusi ei saanud mõistlikult analüüsida, aga ilmselt sarnased

Kui vaid ühe taseme järgi määrata, siis iga lugem, kus esineb lähedase tüve k-mer vea tõttu, määratakse valesti (ja neid on päris palju eriti suures proovis)

CoMeta

- Classification of Metagenomes
- Määrab iga lugemi eraldi (nagu Kraken)
- Töökiirus kordades aeglasem kui Krakenil
- Suudab määrata viiruseid, baktereid ja ilmselt ka muud (olenevalt andmebaasi ülesehitusest), aga mõeldud eelkõige bakterite detekteerimiseks
- Kasutab juhtpuud määramisprotsessis, kuid määrab iteratiivselt (kõigepealt hõimkond ja edasi järjest täpsemaks)

TARTU ÜLIKOOL
1632

CoMeta - uuendused

Olulisim uuendus iga lugemi skoorimine

- Skoorimisel ei loeta mitte k-meride arvu, mis leitakse andmebaasist, vaid mittekattuvate nukleotiidide arvu k-merides, mis andmebaasis vaste annavad

S: A A T C G G G C C A T C C C
x x
R: T A T C G G C C C A T C C C

Reference sequence (S) AATCGGGCCATCCC	Query read (R) TATCGGCCCATCCC	
Sorted k-mers in D_i	k-mers	ξ
AATCG	TATCG	0
ATCCC	**ATCGG**	**5**
ATCGG	TCGGC	5
CATCC	CGGCC	5
CCATC	GGCCC	5
CGGGC	GCCCA	5
GCCAT	CCCAT	5
GGCCA	**CCATC**	**10**
GGGCC	**CATCC**	**11**
TCGGG	**ATCCC**	**12**

TARTU ÜLIKOOL 1632

CoMeta – võrdlus (täpsus ja tundlikkus)

Table 4. Comparison of programs for various level classification using Illumina reads.

Programs	HiSeq 92 bp			MiSeq 156 bp		
	Sensitivity	Precision	Classified	Sensitivity	Precision	Classified
PHYLUM						
LMAT *kFull*	89.89	99.74	90.12	88.23	99.47	88.70
MiniKraken[a]	65.34	99.79	65.48	75.88	99.93	75.93
CoMeta *micDb*	81.64	98.97	82.49	86.71	99.11	87.49
CLASS						
LMAT *kFull*	88.06	99.66	88.36	85.79	99.65	86.09
MiniKraken[a]	65.16	99.65	65.39	75.73	99.91	75.80
CoMeta *micDb*	80.87	98.14	82.40	86.34	98.83	87.36
ORDER						
LMAT *kFull*	86.48	99.80	86.65	81.00	99.63	81.30
MiniKraken[a]	64.89	99.51	65.21	75.52	99.87	75.62
CoMeta *micDb*	80.34	97.73	82.21	85.39	98.01	87.12
FAMILY						
LMAT *kFull*	84.96	99.79	85.14	79.40	99.72	79.62
MiniKraken[a]	64.75	99.46	65.10	75.43	99.81	75.57
CoMeta *micDb*	80.13	97.61	82.09	85.05	97.76	87.00
GENUS						
LMAT *kFull*	84.74	99.80	84.91	73.75	99.53	74.10
MiniKraken[a]	64.54	99.45	64.90	71.95	98.04	73.39
MiniKraken[b]	66.12	99.44	—	67.95	97.41	—
Kraken[b]	77.15	99.20	—	73.46	94.71	—
Kraken-GB[b]	93.75	99.51	—	86.23	98.48	—
CoMeta *micDb*	79.82	97.44	81.92	77.50	90.83	85.32

[a]—The results of the program are counted by ourselves.

[b]—The results of the program are taken from the Wood–Salzberg' paper [51].

TARTU ÜLIKOOL 1632

CoMeta – võrdlus (kiirus ja RAM)

Table 5. Comparison of RAM memory usage and CPU times.

Program	FACS 269bp	MetaPhyler 300bp	CARMA 265bp	PhyloPythia 961bp	HiSeq 92bp	MiSeq 156bp
CPU Runtime (minutes)						
CARMA[a]	290880	77340	74950	360107	—	—
MEGAN[a]	288020	72060	72010	351060	—	—
MetaPhyler[a]	10	20	2	28	—	—
MG-RAST[a]	60	10080	20160	12960	—	—
LMAT *kML*	36(60[b])	58	43	348	—	—
LMAT *kFull*	54(93[b])	213	38	772	15	33
MiniKraken	—	1.22	1.07	2.95	1.3	1.2
CoMeta *allDb*	41(76[b])	14	28	144	—	—
CoMeta *micDb* (ph)	—	9	14	35	8	9
CoMeta *micDb* (ge)	—	—	—	79	42	68
Memory Usage (Megabytes of RAM)						
CARMA[a]	100	100	100	120	—	—
MEGAN[a]	1024	1024	1024	1410	—	—
MetaPhyler[a]	5734	5734	5734	5734	—	—
MG-RAST[a]	—	—	—	—	—	—
LMAT *kML*	17000(17284[b])	17019	2128	13311	—	—
LMAT *kFull*	9295(9481[b])	13247	13286	15092	5807	12392
MiniKraken	—	4098	3210	4100	1317	1449
CoMeta *allDb*	71260(71903[b])	70743	71313	69508	—	—
CoMeta *micDb*	—	19552	19320	19552	10297	17689

CoMeta – võrdlus Krakeniga

- Skoorimisel arvestada vaid mittekattuvaid k-mere/nukleotiide tundub nutikas, kuid aeganõudev
- Eraldi taksonoomiliste tasemete analüüs aeganõudev – andmebaas on väiksem, kuid tuleb mitu korda iga lugemit analüüsida
- CPU kasutus – kasutab palju rohkem mälu ning võtab OLULISELT rohkem aega kui Kraken (minutid vs mõni tund) – ei ole kasutatav kiireks diagnostikaks
- Kasutati palju suuremat andmebaasi kui miniKrakeni puhul (võrreldav Kraken-GB) – tundlikkuse erinevused andmebaasist tingitud

TARTU ÜLIKOOL

KÜSIMUSED ?