Unikaalsete ja populaarsete oligote selekteerimine (EST andmebaasist)

13.10.2004

ESTs

ESTs – expressed sequence tags
ehk osaline järjestus juhuslikust cDNA kloonist, mis on saadud mRNAde sekveneerimisel

tavaliselt 200-700 bp pikad

Kasutatakse:

geenide ekspressioon ja funktsioon

transkribeeritud regioonide uurimine

valkude indentifitseerimine

ESTs

http://www.ncbi.nlm.nih.gov/dbEST/

Number of public entries: 23,970,155

sekveneerimine vs EST kasutamine (nisu, oder, rukis ~$5x10^9$ bp)

Uurimisobjekt:
ODER (*Horedeum vulgare*)

~400000 ESTs

~310000 BAC klooni (6.3-fold genome coverage)

Kodeeriv ala (12% genoomist) koondunud ~600Mb sisse

Unikaalne ja populaarne

Unikaalne oligo – oligo, mis leidub täispikkuses 1 EST järjestuses ja ei leidu osaliselt ja täispikkuses üheski teises

nö markerjärjestus

Populaarne oligo – oligo, mis leidub täispikkuses või osaliselt suurimas arvus ESTides (ei pea esinema kõigis)

Kloonid, mis sisaldavad ekspresseeritud geene

Olemasoloevad meetodid

Põhiline probleem: mõeldud väikeste mahtude jaoks

Enamus meetodeid kb andmemahud

TEIRESIAS (Rigoutsos & Floratos, 1998)
28 Mb andmehulk ja 1GB RAM ei andnud tulemusi mälu puudusel

OLIGOARRAY (Rouillard *et. al.*, 2002)
BLAST –F –F –S 1
Mfold – sekundaarstruktuuride leidmine

6342 pärmi geeni -> ~1 päev 700MHz

Unikaalne oligo

l-mer - string (l pikkusega)
d – maksimaalne arv mismatch'e oligote vahel (x,y)
seed – identne osa stringi sees (alamstring)

1. l-mer jagatakse *seed*'ideks t = [d / 2] + 1
2. iga *seed*'i pikkuseks on q = [l / t]

Hamming distance definition: *The number of bits which differ between two binary strings. More formally, the distance between two strings A and B is* $\sum | A_i - B_i |$.

Unikaalne oligo

FAAS1

4^q - erinevate *seed*'ide arv kokku (-> hash tabel)

Iga isend tabelis sisaldab positsioone, kus ta EST'ides asub.
Grupisiseselt ei erine üks seed teisest rohkem kui 1 mismatch'iga

FAAS2

- Võrreldakse kõiki teisi seed'de, mis omavad ainult 1 mismatch'i käesoleva isendiga
- Pikendatakse *flanking* piirkondi ümber *seed*'ide (vasak-parem kokku)
- Kui H(x,y) <= d, märgitakse vastavad l-mer'id mitteunikaalseteks
- d –1 (*flanking* järjestuste sees)

Unikaalne oligo

Sisendandmed:
46 145 ESTs
28475017 bp

maskeeriti eelnevalt poly(A) ja poly(T) kordused!

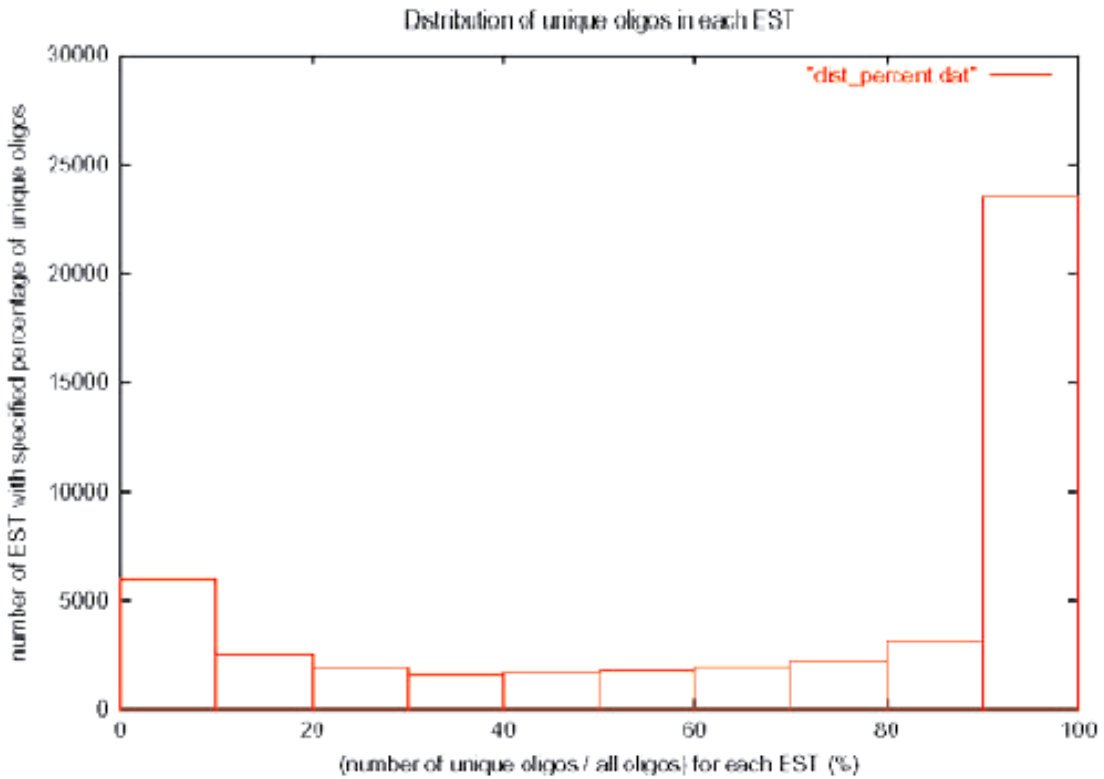
Parameetrid:
l = 33, d = 5, q = 11

Vastavalt parameetritele:
4^{11} erinevad *seed*'i

Iga oligo jaotatakse 3'e gruppi pikkusega q

Unikaalne oligo

# of occurrences	# of seeds
0	242399
1-9	3063288
10-19	708745
20-29	120698
30-39	31637
40-49	11908
50-5049	15629



13430 EST'i sisldasid ainult unikaalseid 33'meere

2159 EST'i ei sisaldanud ühtegi unikaalset 33'meeri

2h 26m, 200MB mälu

Populaarne oligo

l – oligo pikkus
c – südamiku pikkus
d – maksimaalne arv mismatche oligote vahel (x,y)
T_c – minimaalne arv südamikke

l = 8, c = 5, d = 1, T_c = 3

Populaarne oligo

>EST0
TGGAGTCCTCGGACACGATCACATCGACAATGTGAA
GGCGA
>EST1
GTGAAGGAGGTAGATCAAATAGAGCCTGCCCTAAAA
AGGCAGCTTATAATCTCCACTGCT
>EST2
TCCGACTACTGCACCCCGAGCGGATCACACAATGGAA
GGCCCGTGCGC
>EST3
GTGAAGGAGGTAGATACTCGTATACGATCACTGCCTA
AAAAGGCAGCTTATAATCTCCATATCGCTG

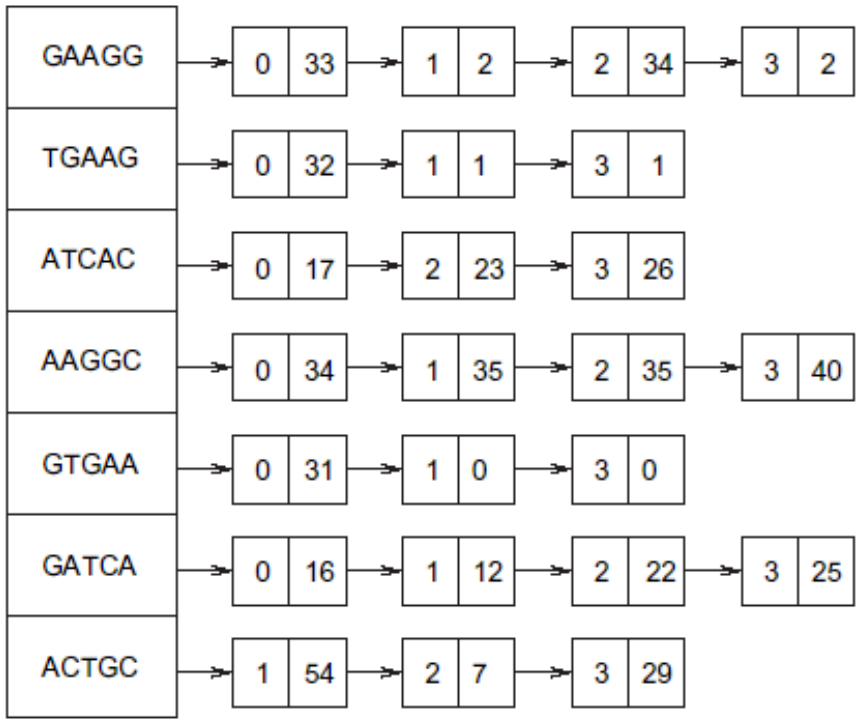


Fig. 3. The example dataset ($l = 8, c = 5, d = 1, T_c = 3$). The table of popular cores.

FAAS1

- 7 populaarset südamikku 148st võimalikust
- iga isend tabelis näitab südamiku positsiooni vastavas EST järjestuses

Populaarne oligo

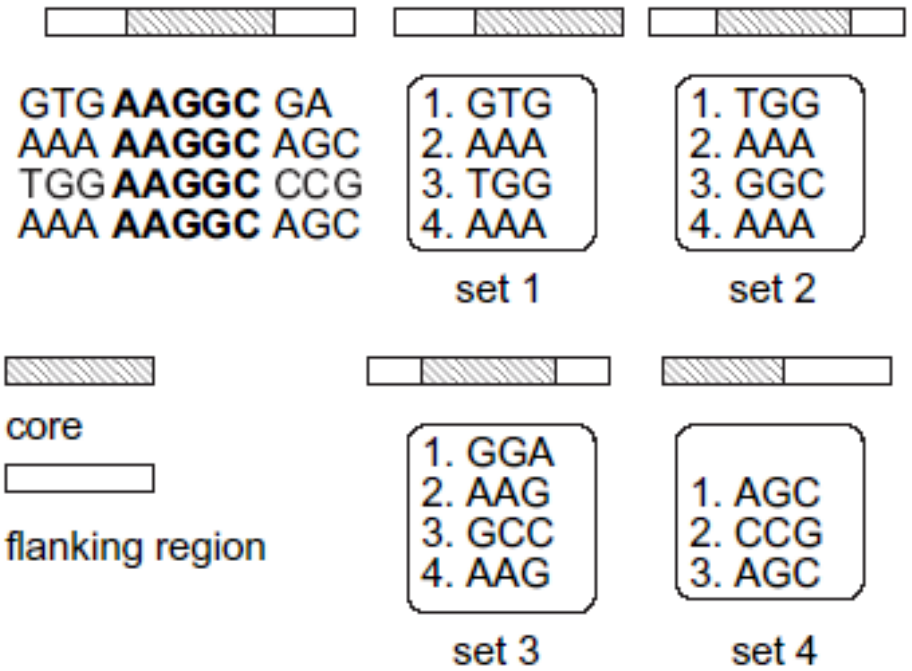


Fig. 4. Collecting flanking regions for the core. There are four sets of flanking regions for AAGGC.

FAAS2

- Kogutakse kokku *flanking* järjestused 7 südamiku ümber
- 4 komplekti stringe (vasak ja parem ühendatud)

Populaarne oligo

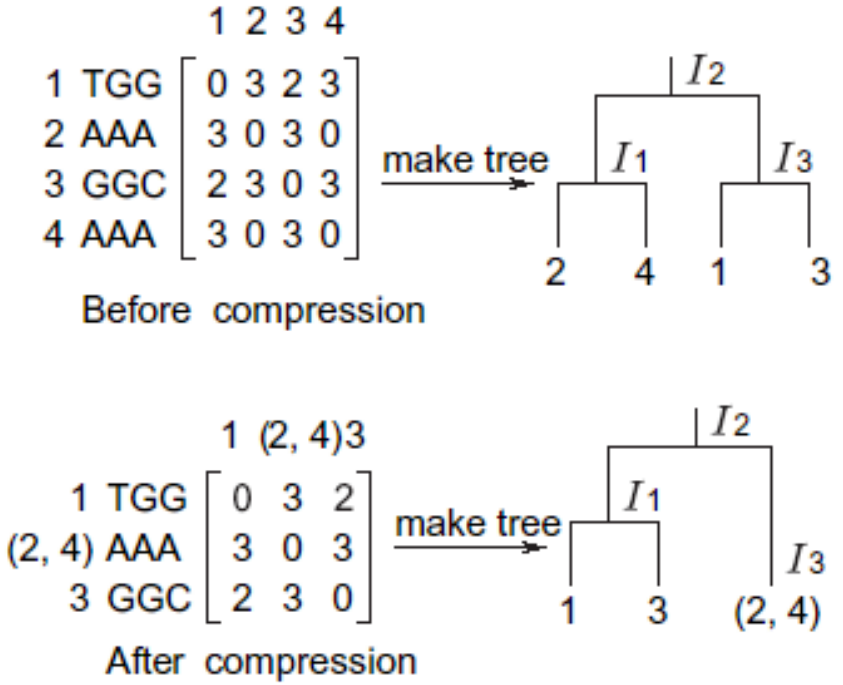


Fig. 5. UPGMA tree construction for set 2 of the core AAGGC.

FAAS3

- *Flanking* järjestused klasterdatakse hierarhiliselt (UPGMA)
- "Puu" ehitamise ajal - kui *Hamming distance* "lehtede" vahel on 0, identsed stringid kombineeritakse üheks

Populaarne oligo

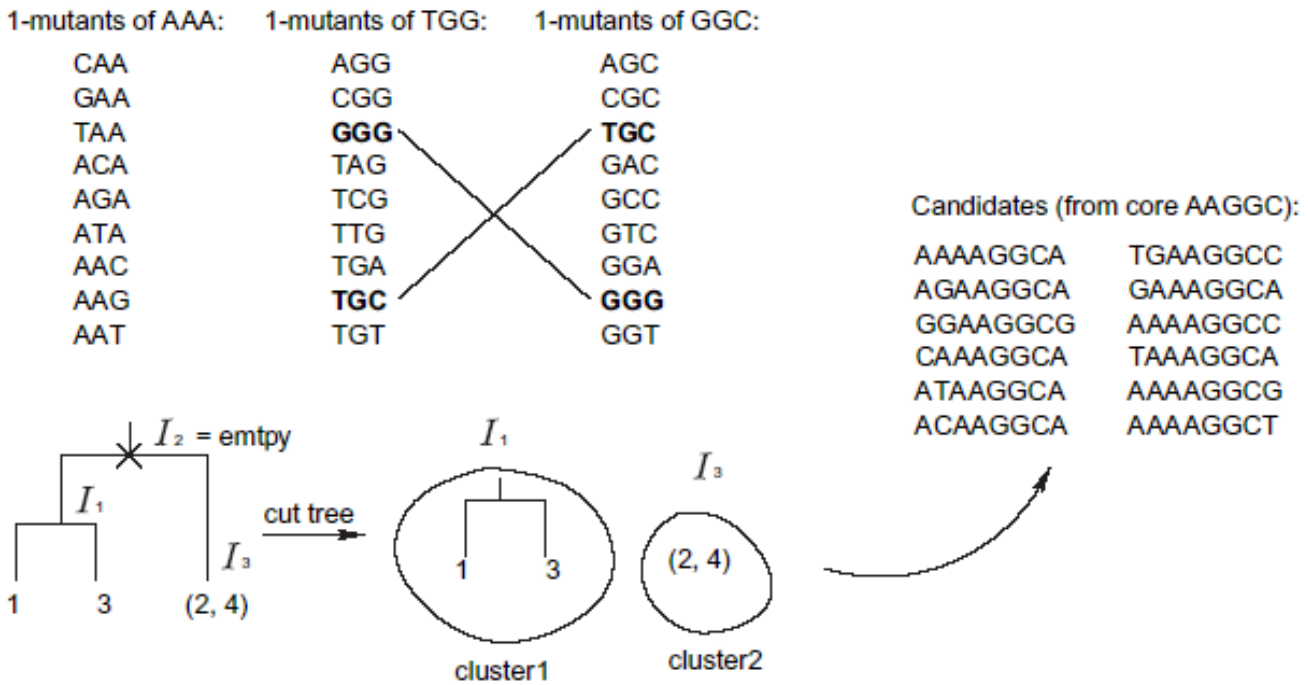


Fig. 6. Clustering set 2 of the core AAGGC.

Kuna AAA oli 2s korduses I_3 sees, lisatakse kõik variandid kandidaatide hulka

Populaarne oligo

FAAS4

- Järele jäänud kandidaadid sorteeritakse ja eemaldatakse duplikaadid.
- Lisatakse teistelt südamikkudelt leitud kanditaatidele.

PUHASTUSFAAS

- GC sisaldus
- Poly(A), poly(T) ja lihtsamad kordused

KOMPRESS (greedy set coverage algorithm)

- Valitakse välja oligod, mis ületavad etteantud läve
- Võetakse kõige enam EST’e kattev oligo.
- Eemaldatakse kõik kaetud EST’id valikust
- Uuendatakse ülejäänud oligote EST-seostumiste arvu
- Korratakse eelnevat, kuni ühtegi EST’i pole järgi jäänud

Populaarne oligo

Colors	Number of cores
1	22 523 412
2–10	21 28 677
11–20	5148
21–30	1131
31–40	492
41–50	346
51–60	242
61–70	77
71–80	34
81–90	29
91–100	43
101–176	19

The left column is the range of the number of colors. The right column is the number of cores with a certain number of color.

Populaarne oligo

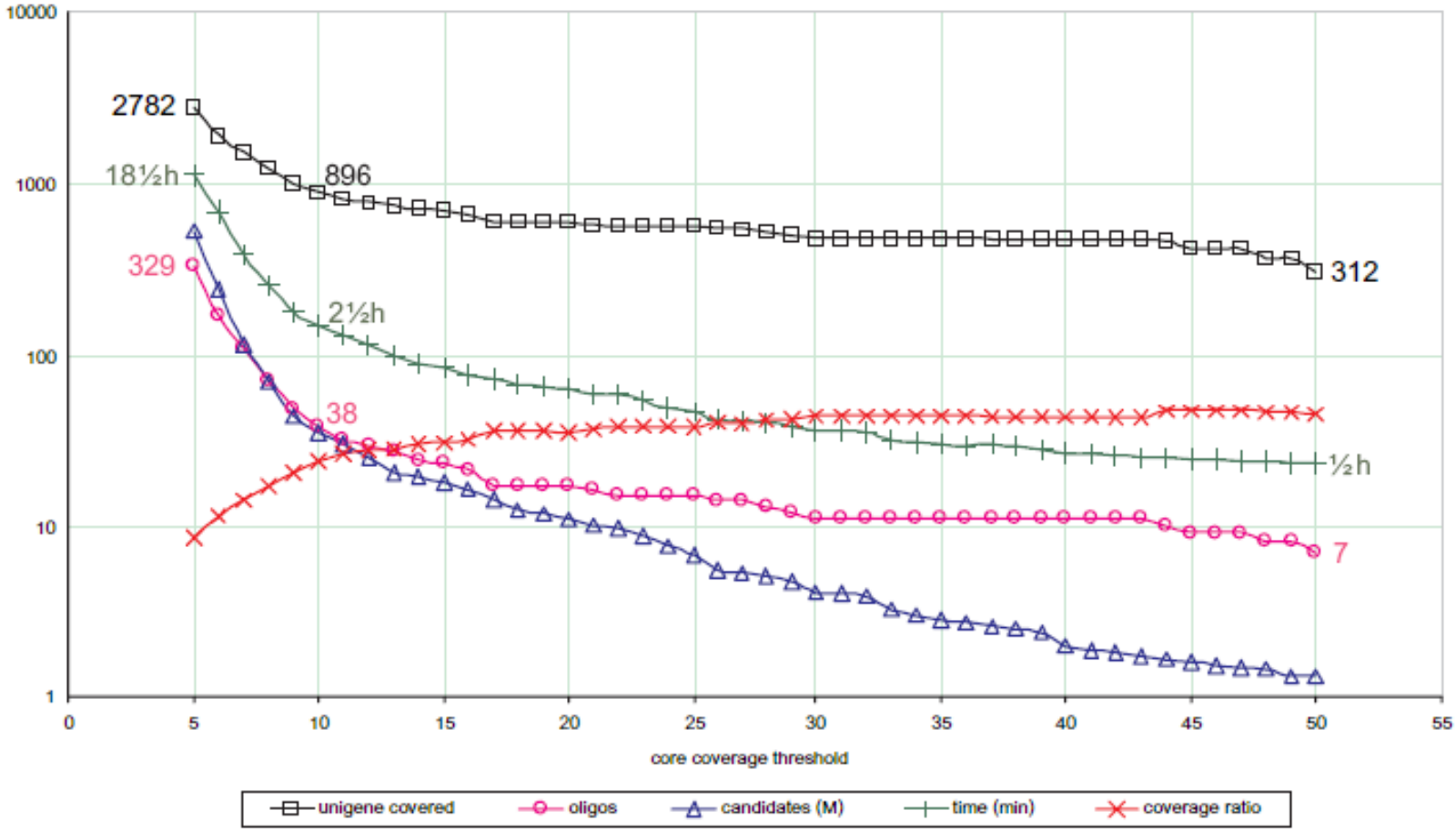


Fig. 8. Results of running the algorithm on the Barley dataset. Shown are the number of candidates generated by the algorithm (in millions), the number of ESTs covered, the final number of popular oligos, the coverage ratio and the time taken by the algorithm (for different choices of T_c).

Zheng J, Close TJ, Jiang T, Lonardi S.
"Efficient selection of unique and popular oligos for large EST databases."
Bioinformatics. 2004 Sep 1;20(13):2101-12.

Rouillard JM, Herbert CJ, Zuker M.
"OligoArray: genome-scale oligonucleotide design for microarrays."
Bioinformatics. 2002 Mar;18(3):486-7.