Tandeemsete korduste leidmise programme…

Triinu Kõressaar

10.11.2004

Kordusjärjestused (1):

-Hajuskordusjärjestused

SINE
LINE
LTR
transposoonid
pseudogeenid
segmentide duplikatsioonid

Kordusjärjestused (2):

- Tandeemsed kordused
  1. Mikrosatelliidid (SSR)
     - Bloki suurus: 100-300bp
     - Korduste suurus: 1-4bp (1-13bp)
     - Lokalisatsioon: üle genoomi
     - Nt STR-id
  2. Minisatelliidid
     - Bloki suurus: 0,1-20kb
     - Korduste suurus: 6-64bp
     - Lokalisatsioon: telomeerides ja nende läheduses
     - Tüüpiliselt heksameeridena
     - Nt VNTR
  3. Satelliidid
     - Bloki suurus: 100kb-1Mb
     - Korduste suurus: 5-171bp
     - Lokalisatsioon: Tsentromeerid ja heterokromatiin
     - Satelliit 1,2,3, alfa-beeta-gamma satelliidid

Tandeemsete korduste leidmise tähtsus

-polümorfsed
-mutatsioonide hotspotid

Kasutatakse järgnevates valdkondades:

-üle genoomsete markeritena populatsioonigeneetikas,
-geneetilise ahelduse analüüsil
-molekulaarse evolutsiooni sündmuste ja edasiviivate jõudude uurimisel,
-kohtumeditsiinis
-indiviidide identifitseerimisel (isaduse tuvastamisel)
-geneetiliste haiguste kaardistamine ja uurimisel (oluline osa neuroloogiliste haiguste põhjustamisel, nt trinukleotiidsete ekspansioonide haigused)
-kasvajate arengu uurimine
-geeniregulatsiooni uurimisel (seonduvad valkudega, nt bakteritel osalevad kuuma-shoki poolt indutseeritud ekspressioonimehhanismides)
-kromatiini struktuuri uurimisel (mõjutavad kromatiini struktuuri)
-jt

Programmid tandeemsete korduste leidmiseks

Üldised korduste leidmise programmid

REPuter (Kurtz et al., 2001)

veeb: http://bibiserv.techfak.uni-bielefeld.de/reputer/

Kordused: võimaldab leida erineva pikkusega kordusi,
leiab nii täpsed kui degeneratiivsed kordused

Sihtmärk DNA: võimaldab leida suhteliselt pikkadest DNA järjestustest kordusi

Algoritm: Kasutab sufikspuud

Mäluvajadus, tööaeg: lineaarne genoomi pikkusega ja väljundi suurusega

Kood: tasuta mitte-kasumit taotlevatele kasutajatele

Platvorm: erinevad UNIXi versioonid

FORRepeats (Lefebvre et al., 2003)

veeb: **http://al.jalix.org/FORRepeats/**

Algoritm: heuristiline meetod leidmaks kordusi kogu genoomist

Üldised korduste leidmise programmid

SRF – Spectral Repeat Finder (Sharma et al., 2004)

veeb: http://www2.imtech.res.in/raghava/srf/

Kordused: võimaldab leida nii tandeemseid kordusi kui ka hajuskordusi

Sihtmärk DNA: võimaldab leida suhteliselt pikkadest DNA järjestustest kordusi

Algoritm: kasutab Fourieri transformatsiooni

Tööaeg: O(n2)

Kood: programmi võimalik kasutada läbi veebi

Paljud teised: BLAST, MegaBLAST, MUMer …

Spetsiifiliselt tandeemsete korduste leidmisele orienteeriutd programmid

Tandeemseid kordusi võib leida lähtudes kahest algoritmist:

a) Sõnaraamtu-põhine
- -Kasutatakse teadaolevate motiivide sõnaraamatut
- -Sihtmärkjärjestusest otsitakse motiive
- -Leitakse täpsed (exact) tandeemsed kordused

Näide:
TROLL – Tandem Repeat Occurrence Locator

Spetsiifiliselt tandeemsete korduste leidmisele orienteeriutd programmid

b) mudeli-põhine

- Defineeritakse mudel ehk konsensus tandeemse korduse jaoks
- Otsitakse sihtmärkjärjestusest regioonid, mis vastavad definitsioonile

Kahesuguseid järjestusi:

- täpsed kordusjärjestused (exact-repeats)
- degeneratiivsed järjestused

Näide:

TRF – Tandem Repeats Finder

String – Search for Tandem Repeats IN Genomes

Sputnik (Abajin, 1994 http://espressosoftware.com/pages/sputnik.jsp)

Programmid tandeemsete korduste leidmiseks

TROLL – Tandem Repeat Occurrence Locator (Castelo et al., 2002)

veeb: http://finder.sourceforge.net/

Kordused: võimaldab leida SSR-e

Sihtmärk DNA: võimaldab leida suhteliselt pikkadest DNA järjestustest kordusi

Algoritm: kasutab Aho-Corasick algoritmi (keyword-tree), programmile antakse ette korduvaid mustreid sisaldav fail

Tööaeg: lineaarse keerukusega

Kood: vabavara, käsurea programm

Platvorm: Linux

+ kiirem kui programmid, mis ei vaja kasutaja käest tandeemsete korduste järjestusi
leiab tandeemsete korduste täpsed vasted

Programmid tandeemsete korduste leidmiseks

TRF – Tandem Repeat Finder (Benson, 1999)

veeb: http://c3.biomath.mssm.edu/trf.html

Kordused: võimaldab leida erineva pikkusega kordusi,
leiab nii täpsed kui degeneratiivsed kordused

Sihtmärk DNA: sihtmärkjärjestuste pikkus on limiteeritud 5Mb

Algoritm: Kasutab heuristikat, kaks etappi

-*detekteerimine* –kasutatakse statistilisi kriteeriume (Pm ja Pi), et leida kanditaat tandeemseid järjestusi

-*analüüsimine* – kanditaatide seast püütakse leida tandeemseid kordusi

Tööaeg: aeg kasvab lineaarselt järjestuse pikkuse kasvamisega

Kood: vabavara

Platvorm: Windows, Solaris, Linux, UNIX jt

\+ ei vaja kasutaja käest tandeemseid järjestusi,
pole piirangut detekteeritava korduse pikkuse kohta
insertsioone ja deletsioone käsitletakse erinevalt

Programmid tandeemsete korduste leidmiseks

STRING – Search for Tandem Repeats IN Genome (Parisi et al., 2003)

veeb: http://www.caspur.it/~castri/STRING/

Kordused: võimaldab leida erineva pikkusega kordusi,

Sihtmärk DNA: võimaldab leida suhteliselt pikkadest DNA järjestustest kordusi

Algoritm: Kasutab heuristikat, dünaamilist programmeerimist, defineeritakse exact-TR ja inexact-TR

Tööaeg: polünomiaalne

Kood: vabavara

Platvorm: UNIX, Linux, McOS

Programmid tandeemsete korduste leidmiseks

STAR – Search for Tandem Approximate Repeats (Delgrange et al., 2004)

veeb: http://atgc.lirmm.fr/star/

Kordused: võimaldab leida SSR-e

Sihtmärk DNA: võimaldab leida suhteliselt pikkadest DNA järjestustest kordusi

Algoritm:

motiiv ->tandeemsed duplikatsioonid->ETR -> mutatsioonid ->ATR

kasutatakse Minimum Description Length (MDL) kriteeriumi, mis väljendab seda, kas antud ETR-ist on vaadeldav ATR tekkinud

Tööaeg: n(log n)

Kood: source code available on request, saab kasutada ka veebipõhist

Platvorm: Linux, SunOS, Mac OS, Windows

Mreps (Kolpakov et al., 2003) - http://mreps.loria.fr/

heuristiline algoritm, leiab efektiivselt mikrosatelliitidest kuni satelliitideni, ei limiteeri otsitava järjestuse suurust, võimaldab leida ‘loose’ kordusi, mis omavad omavahel suuremat varieerumist

Kokkuvõtteks:

-Tandeemsete korduste leidmiseks on palju erinevaid meetodeid-programme

-Enamus programmidest kasutavad heuristilist algoritmi, mis võimaldab efektiivselt leida mutatsioonide hotspotiks olevaid tandeemseid kordusi

-Keeruline on defineerida suurusi, mis kirjeldavad ühte tandeemset kordust (st et ATR1 ja ATR2 on mõlemad tekkinud ETR-ist)

-Suurtest genoomsetest järjestustest korduste otsimine ressursimahukas

-Kõik programmid pole vabavaralised

KASUTATUD KIRJANDUS

- Finishing the euchromatic sequence of the human genome. Nature. 2004 Oct 21;431(7011):931-45.
- Initial sequencing and analysis of the human genome. Lander ES, Linton LM, Birren B et al. Nature. 2001 Feb 15;409(6822):860-921.
- REPuter: fast computation of maximal repeats in complete genomes. Kurtz S, Schleiermacher C. Bioinformatics. 1999 May;15(5):426-7.
- Triplet repeats in human genome: distribution and their association with genes and other genomic regions.
  Subramanian S, Madgula VM, George R, Mishra RK, Pandit MW, Kumar CS, Singh L. Bioinformatics. 2003 Mar 22;19(5):549-52.
- FORRepeats: detects repeats on entire chromosomes and between genomes. Lefebvre A, Lecroq T, Dauchel H, Alexandre J. Bioinformatics. 2003 Feb 12;19(3):319-26.
- Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. Sharma D, Issac B, Raghava GP, Ramaswamy R. Bioinformatics. 2004 Jun 12;20(9):1405-12. Epub 2004 Feb 19.
- TROLL--tandem repeat occurrence locator. Castelo AT, Martins W, Gao GR. Bioinformatics. 2002 Apr;18(4):634-6.
- Tandem repeats finder: a program to analyze DNA sequences. Benson G. Nucleic Acids Res. 1999 Jan 15;27(2):573-80.
- STRING: finding tandem repeats in DNA sequences. Parisi V, De Fonzo V, Aluffi-Pentini F. Bioinformatics. 2003 Sep 22;19(14):1733-8.
- STAR: an algorithm to Search for Tandem Approximate Repeats. Delgrange O, Rivals E. Bioinformatics. 2004 Nov 1;20(16):2812-20. Epub 2004 Jun 04.
- mreps: Efficient and flexible detection of tandem repeats in DNA. Kolpakov R, Bana G, Kucherov G. Nucleic Acids Res. 2003 Jul 1;31(13):3672-8.