TARTU ÜLIKOOL
BIOLOOGIA-GEOGRAAFIATEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
Bioinformaatika õppetool

Priit Adler

MikroRNA otsimise algoritmid

Bakalaureusetöö

Juhendaja PhD. Jaak Vilo

Tartu 2005

Sisukord

Lühendid

BLAST	Basic Local Alignment Search Tool	lokaalsel joondamisel põhinev järjestuse võrdlemist teostav algoritm
dsRNA	double stranded RNA	kaheahelaline RNA
EmiRSE		EMBL miRNA sihtmärkide ennustused
GO	Gene Ontology	geeni ontoloogia
HMMer	Profile Hidden Markov model	Profiili Peidetud Markovi mudel
miRNA	microRNA	mikroRNA
mRNA	messenger RNA	matriits RNA
ORF	Open Reading Frame	avatud lugemisraam
pre-mRNA	precursor mRNA	eellas mRNA, mRNA koos intronite ja eksonitega
PTGS	Posttranscriptional Gene Silencing	transkriptsioonijärgne geeni vaigistamine
RISC	RNA-induced Silencing Complex	RNA-põhjustatud vaigistamise kompleks
RNAi	RNAinterference	RNAinterferents
SD	Standard Deviation	standardhälve
siRNA	small interfering RNA	väike segav RNA
stRNA	short temporal RNA	lühike ajutine RNA
UTR	Untranslated Region	mittetransleeritav regioon

Sissejuhatus

Suur hulk rakus transkribeeritavatest geenidest ei kodeeri valku. Nende geenide produktiks on hoopis "mittekodeeriv" ehk "mitte-matriits" RNA molekul nagu näiteks ribosomaalne RNA ja transport RNA. Lisaks on veel märkimisväärne kogus mittekodeerivaid geene, mille funktsioon oli kuni eelmise sajandi lõpuni kindlalt määratlemata. Hiljutised avastused on näidanud, et paljudel nendest mittekodeerivatest geeniproduktidest on oluline roll geenieksprессiooni reguleerimisel.

Esimesena tuvastati nendest geenid, mis ekspresseerusid vaid kindlas organismi arengustaadiumis (Lee *jt.* 1993; Reinhart *jt.* 2000). Geenide produktiks olid lühikesed umbes 22 aluspaari pikkused RNA molekulid, millel oli regulatoorne võime. Selliseid RNA molekule hakati nimetama ajutisteks lühikesteks RNAdeks (*short temporal RNA, stRNA*). Seejärel tuvastati palju (üle saja) sarnaseid lühikesi RNA molekule, mida eristas stRNAst teistsugune, mitte arenguetapist sõltuv ekspressioonimuster (näit. koespetsiifiline). Seetõttu tuli kasutusele võtta uus mõiste.

MikroRNAks (*microRNA, miRNA*) hakati nimetama stRNAsid ja kõiki teisi lühikesi RNA molekule, millel olid stRNAle sarnased omadused, kuid tundmatu funktsioon (Lagos-Quintana *jt.* 2001). Viimase viie aasta jooksul on ilmunud palju töid, kus kirjeldatakse erinevate meetoditega tuvastatud miRNA geene. Siiski arvatakse, et paljud miRNAd on veel kirjeldamata nende keerulise ekspressiooni aja või koha tõttu.

MikroRNAde süsteemseks nimetamiseks ja katalogiseerimiseks on loodud andmebaas - miRNA Register (Griffiths-Jones 2004; Ambros *jt.* 2003). Seisuga aprill 2005 a. on loendatud 1650 mikroRNA geeni.

MikroRNAd on võimelised maha suruma geeni(de) avaldumist transkriptsioonijärgselt (*posttranscriptional gene silencing, PTGS*). Regulatsioon toimub järjestusspetsiifilise seondumisega matriits RNAle (*messenger RNA, mRNA*), mille tulemusel geeni translatsioon takerdub ajutiselt või mRNA lagundatakse.

MikroRNAde bioloogiast ja nende interaktsioonidest mRNAga rohkema info saamiseks on vaja teada, kuhu miRNAd mRNAl täpselt seonduvad. Teadmata kuhu miRNAd mRNAl võivad seonduvad, on raske planeerida eksperimentaalseid katseid

ning seetõttu on oluline võimalike seondumiste arvutuslik leidmine.

MikroRNAde suhteliselt lühike pikkus ja nende mittetäielik komplementaarsus mRNAga on põhjuseks, miks klassikalistest joondusmeetoditest üksi ei ole miRNA sihtmärkide ennustamisel suurt abi. Seetõttu tuleb sihtmärkide tuvastamiseks kasutada mitmeid omavahel kombineeritud meetodeid. Käesolevas töös on tutvustatud kolme esimesena avaldatud miRNA sihtmärkide tuvastamise algoritmi, tänu millele on omistatud bioloogiline roll juba mitmele miRNAle. Need on EmiRSE algoritm (Stark *jt.* 2003), miRanda algoritm (Enright *jt.* 2003) ja PicTar algoritm (Rajewsky & Socci 2004).

EmiRSE on Stark *jt.* poolt EMBLis (*European Molecular Biology Laboratory*) välja töötatud miRNAde sihtmärkide ennustamise algoritm (EmiRSE - Embl miRNA Sihtmärkide Ennustused). MiRanda algoritm on välja töötatud Enright *jt.* poolt Memoriaal Sloan-Kettering Vähiuuringute Keskuse juures Arvutusliku Bioloogia Keskuses New York'is. PicTar algoritm on välja töötatud New Yorki Ülikoolis Nikolaus Rajewsky ja Nicolas D. Socci poolt.

Käesoleva töö eesmärk on anda ülevaade nendest kolmest praeguseks publitseeritud algoritmist ja kirjeldada nende etapiviisilist lähenemist sihtmärgi otsingul. Töö viimases osas on teostatud ka kahe algoritmi võrdlus, kus 31-le juhuslikulult valitud miRNAle on ennustatud sihtmärgid mõlema algoritmiga. Algoritmides on sihtmärkide ennustamisel peale algsete joondamismeetodite kasutatud peamise täiendava infoallikana miRNA ja mRNA seondumise vabaenergiat. Uuritud on kui suure panuse annab üks või teine etapp algoritmi töös ning millest on põhjustatud saadud kahe algoritmi tulemuste erinevused.

Käesolev töö on jaotatud kolme peatükki. Kus esimses peatükis on kirjeldatud miRNAde elutsüklit, nende avastamislugu, ekspressiooni, küpsemist ja sihtmärgi järjestus-spetsiifilist vaigistamist. Teises peatükis kirjeldatakse igat miRNA sihtmärgi tuvastamise etappi kolme algoritmi näitel. Kolmandas peatükis on läbi viidud võrdlus kahe algoritmi (EmiRSE ja miRanda) tulemuste põhjal.

Käesolev töö on valminud Tartu Ülikooli Molekulaar- ja rakubioloogia instituudis bioinformaatika õppetoolis. Tahan tänada oma juhendajat Jaak Vilot motiveeriva ja samas kannatliku juhendamise eest ning teisi BIIT töörühma liikmeid toe, nõuannete ja toreda seltskonna eest.

Peatükk 1

MikroRNA

MikroRNAd (*microRNA, miRNA*) on hiljuti avastatud geeniproduktide klass. Need umbes 22 aluspaari (ap) pikkused endogeensed RNA järjestused mängivad olulist rolli nii loomade kui ka taimede geeniregulatsioonis. Regulatoorne roll seisneb nende seondumisel mRNAle järjestus-spetsiifiliselt, mille tulemusel represseeritakse translatsioon või lagundatakse mRNA.

Käesolevas peatükis antakse lühike ülevaade miRNAde avastamisest, esitatakse mõned tõestatud ja tõestamata teooriad, mis puudutavad miRNA geenide transkriptsiooni, miRNAde küpsemist ja sihtmärkgeenide translatsiooni repressiooni.

1.1 miRNA geenid

Nagu kogu RNA rakus, nii ka miRNA kodeeritakse DNA pealt. Esimene selline väike RNA molekul avastati Victor Ambrose laboris, kui kirjeldati *Caenorhabditis elegansi* vastse arengustaadiumide ajastamisel olulist geeni *lin-4* (Lee *jt.* 1993). Avastati, et *lin-4* ei kordeeri mitte valku, vaid geeni lõpp-produktiks on umbes 22 aluspaari pikkune RNA molekul. Teine samalaadne avastus tehti alles seitse aastat hiljem, kui kirjeldati *C. elegansi let-7* geen (Reinhart *jt.* 2000; Slack *jt.* 2000). *Let-7* on samuti seotud *C. elegansi* vastsestaadiumi vaheldumisega.

Juba sama aasta jooksul tuvastati *let-7* homoloogid ka inimese ja kärbse genoomidest, lisaks tuvastati *let-7* RNA inimesest, *Drosophilast* ja veel üheteistkümnest bilateraalse looma genoomist (Pasquinelli *jt.* 2000).

Natuke peale seda teatati juba rohkem kui sajast sarnasest väikesest mittekodeerivast RNA geenist (ligikaudu 20 kärbses, 30 inimeses ja 60 ussides) (Lagos-Quintana *jt.* 2001; Lau *jt.* 2001; Lee & Ambros 2001). Sarnaselt *lin-4* ja *let-7* geenide produktidele olid ka uued avastatud geeniproduktid lühikesed umbes 22 ap

pikkused endogeensed RNA molekulid. Kuid erinevalt *let-7* ja *lin-4* RNAdest ei ekspresseerunud leitud uued RNA molekulid kindlal ajahetkel, vaid hoopis mingis kindlas koetüübis. Kui siiani oli selliste RNA molekulidele viidatud kui lühikestele ajutistele RNAdele (*shost temporal RNA, stRNA*) (Pasquinelli *jt.* 2000), siis nüüd tuli kasutusele võtta uus mõiste - mikroRNA (miRNA). MikroRNAks hakati nimetama stRNAsid ja kõiki teisi lühikesi RNA molekule, millel olid sarnased omadused, kuid tundmatu funktsioon (Lagos-Quintana *jt.* 2001; Lau *jt.* 2001; Lee & Ambros 2001).

Seisuga aprill 2005 on loendatud 1650 miRNAd ja nende süsteemseks nimetamiseks ja katalogiseerimiseks on loodud andmebaas - miRNA Register[1] (Griffiths-Jones 2004; Ambros *jt.* 2003).

1.2 miRNA transkriptsioon

Sarnaselt *C. elegansi lin-4* ja *let-7* geenidele asuvad enamus miRNA geene suhteliselt kaugel siiani annoteeritud geenidest. MikroRNAd võivad olla klasterdunud ka suurematesse operonidesse (Joonis 1.1b). Selline miRNA geenide eraldatus viitab asjaolule, et neil on oma iseseisvad transkriptsiooniüksused (Lagos-Quintana *jt.* 2001; Lau *jt.* 2001; Lee & Ambros 2001). MikroRNA transkriptidele on iseloomulik pikenenud õlgaas[2] struktuur (Lagos-Quintana *jt.* 2001; Lau *jt.* 2001; Lee & Ambros 2001).

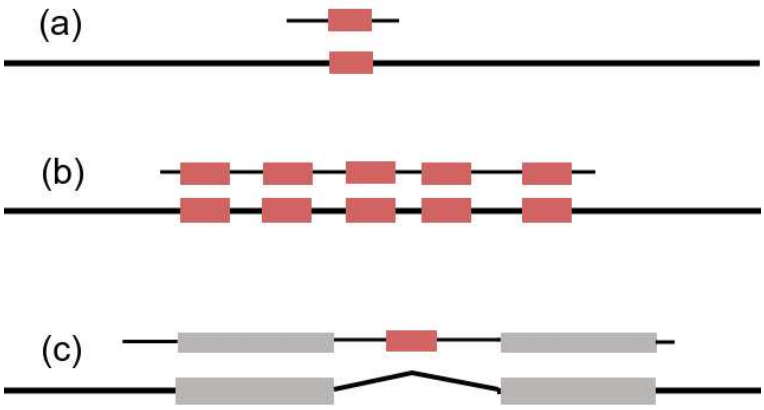
Samas on väike hulk miRNA geene (umbes neljandik inimese miRNA geenidest), mis asuvad pre-mRNA (eellas mRNA) intronites (Joonis 1.1c). Sellised miRNAd kasutavad arvatavasti "peremees" geeni promooterit ja nende regulatsioon on seotud vastava mRNA omaga (Aravin *jt.* 2003; Lagos-Quintana *jt.* 2003; Lai *jt.* 2003; Lim *jt.* 2003). Ülejäänud miRNA geenide transkriptsiooni initseerib arvatavasti nende oma promooter, aga ühtki primaarset transkripti pole siiani täielikult kirjeldatud. Siiski need primaarsed transkriptid (pri-miRNAd) (Lee *jt.* 2002) arvatakse olevat palju pikemad, kui konserveerunud õlgaas struktuurid, mille järgi hetkel miRNA geene defineeritakse. Seda, et pri-miRNA on pikem kui õlgaas struktuur toetab:

- hüpotees, et klasterdunud miRNAde õlgaas struktuurid võivad olla transkribeeritud ühelt ainsalt primaarselt transkriptilt (Lagos-Quintana *jt.* 2001; Lau *jt.* 2001)

- miRNAde ja andmebaasides olevate pikkade EST (*Expressed Sequence Tag*)

[1] http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml

[2] õlgaas - stem loop



Joonis 1.1: MikroRNA geenide võimalikud paigutused DNAl. **(a)** Üksik miRNA geen (punasega on märgitud õlgaas-strukuuri moodustav osa). **(b)** Viis miRNA geeni on moodustanud operoni (sama orientatsiooniga koos transkribeeritavad ja koos reguleeritavad geenide jadad). **(c)** MikroRNA geen asub valku kodeeriva geeni intronis (eksonid on märgitud hallina).

järjestuste omavaheline ülekattuvus (Lagos-Quintana *jt.* 2002; Aukerman & Sakai 2003)

- RT-PCR (*Reverse Transcriptase - Polymerase Chain Reaction*) eksperimentide tulemused, kus amplifitseeriti pri-miRNA fragmente, mis olid pikemad kui konserveerunud õlgaas strukuurid (Lee *jt.* 2002; Aravin *jt.* 2003).

Siiani on teadmata, milline RNA polümeraas miRNAde transkriptsiooni läbi viib. Kuigi enamus tõendeid viitab kaudselt, et seda viib läbi pol II (Ohler *jt.* 2004; Johnson *jt.* 2003; Johnston & Hobert 2003; Lagos-Quintana *jt.* 2002), siis on eksperimentaalseid tõid mis näitavad, et transkriptsiooni võib edukalt läbi viia ka pol III (Chen *jt.* 2004).

1.3 miRNA küpsemine

MikroRNA küpsemise (*i.k. maturation*) esimese etapina toimub pri-miRNA lõigustumine, mille tulemusel vabaneb umbes 60 - 70 ap pikkune õlgaas. Sellist vaheühendit nimetatakse eellas miRNAks või pre-miRNAks (Lee *jt.* 2002; Zeng & Cullen 2003).

Prekursor miRNA transporditakse eksportiin-5 transportvalgu vahendusel tuu-

mast tsütoplasmasse (Yi *jt.* 2003; Lund *jt.* 2004). Tsütoplasmas toimub järgmine lõigustumine, mille käigus vabastatakse pre-miRNA aas-struktuurist. Mõlema lõikamise eest vastutavad RNaas III tüüpi endonukelaasid: Drosha ja Dicer (Lee *jt.* 2003; Basyuk *jt.* 2003) (Joonis 1.2).

Tsütoplasmas, peale aasa eemaldamist Diceri poolt, jääb järele osaliselt komplementaarne kaheahelaline RNA dupleks. Dupleks koosneb küpsest miRNAst ja sellega osaliselt komplementaarsest vastasahelast, millele viidatakse kui miRNA* (Lau *jt.* 2001). Sellisele dupleksile on iseloomulik, et 5' ahela ots on fosforüleeritud ning 3' ahela ots on paar aluspaari pikem kui 5' ahela ots (Lee *jt.* 2003; Basyuk *jt.* 2003).

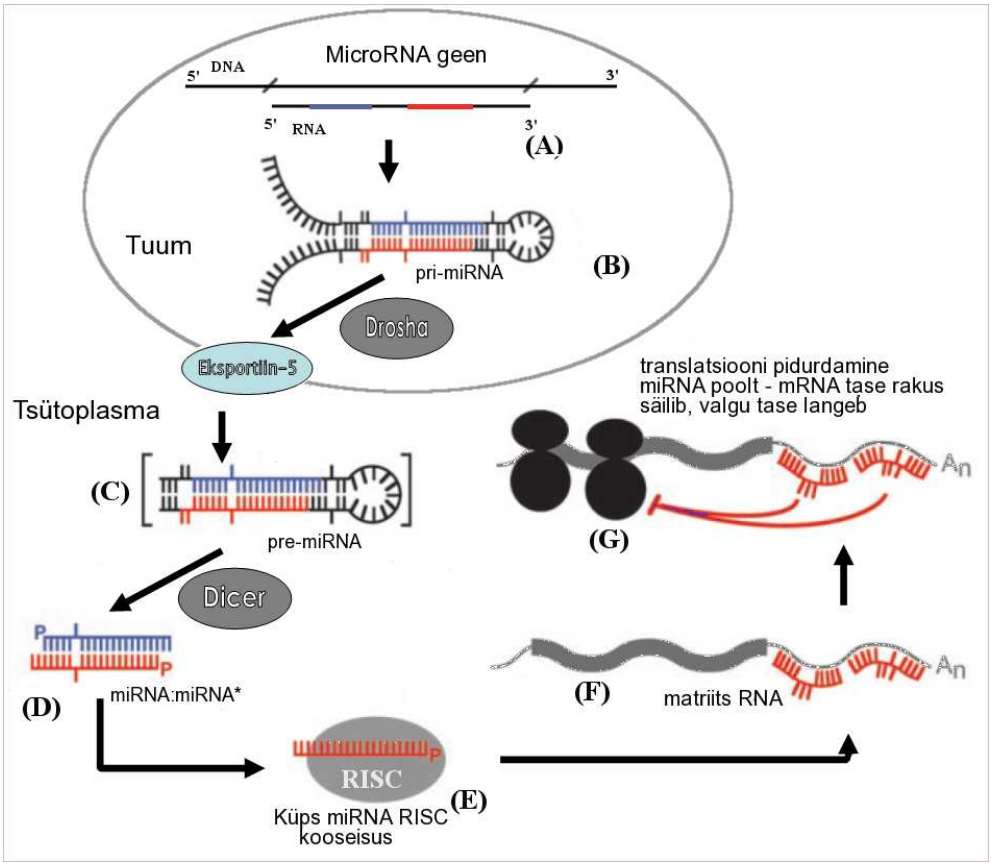
1.4 Translatsiooni repressioon

MikroRNA küpsemise käigus moodustunud dupleksi üks ahel liitub ribonukleoproteiini kompleksiga, mida tuntakse kui RNA-põhjustatud vaigistamise kompleks (*RNA-induced silencing comlex, RISC*). MikroRNA siseneb RISCi 5' ots ees. Kumb dupleksi ahelatest valitakse, sõltub dupleksi otste suhtelisest stabiilsusest: valitakse 5' ots mis on vabamalt (nõrgemini) seotud teise ahelaga (Khvorova *jt.* 2003; Schwarz *jt.* 2003). RISCiga mitteseondunud ahel lagundatakse. (Joonis 1.2)

RISC sai esmalt tuntuks RNAinterferentsi (*RNAinterference, RNAi*) põhjustava kompleksina, kus väikesed segavad RNAd (*small interfering RNA, siRNA*) põhjustavad peaaegu täieliku komplementaarsuse alusel sihtmärk mRNA endogeense lagundamise (Hammond *jt.* 2000). RNAi on peamiselt taimedes levinud geenide transkriptsioonijärgse vaigistamise (*PTGS*) ja võõr-RNA molekulide (viirused) vastu võitlemise mehhanism. MikroRNAd ja siRNAd on väga sarnased, ning nende käitumine erinevates tingimustes võib ka kattuda. Ainus erinevus on siRNA eellas-molekul, siRNA eellas-molekuliks on kaheahelaline RNA molekul (*double strand RNA, dsRNA*), mis omakorda võib pärineda eksogeensest dsRNAst, transposonist[3] või näiteks kaheahelalisest RNA viirusest. Tuletame meelde, et miRNA eellaseks on konserveerunud õlgaas struktuur, mis koosneb **ühest** voltunud ahelast.

On teada, et miRNA koos RISCiga seondub mRNA 3' mittekodeeriva regiooni (*Untranslated Region ,UTR*) järjestusele mittetäieliku komplementaarsuse alusel. Sellega põhjustatakse valgu translatsiooni mahasurumine, ilma mRNA enda stabiilsust mõjutamata. Kuid mehhanism, kuidas translatsiooni repressioon toimub on siiani detailselt kirjeldamata (Ambros 2004). Üks võimalus on, et miRNA takistab

[3]siin töös **RNA** vaheühendite abil genoomis ümberpositsineerimis võimet omavad järjestused



Joonis 1.2: Loomse miRNA elutsükkel. Erinevad miRNA olekud on tähistatud tähega (A - G). MikroRNA geen kodeeritakse DNAlt (A), primaarne transkript (pri-miRNA) moodustab pikenenud õlgaas struktuuri (B) (punane - miRNA, sinine - miRNA*). Pri-miRNAd töödeldakse kahe RNaas III endonukleaasiga: Drosha (raku tuumas) ja Dicer (tsütoplasmas). Peale esimest lõigustumist transporditakse saadud pre-miRNA (C) eksporiin-5 transportvalgu vahendusel tuumast tsütoplasmasse, kus toimub teine lõigustumine. Saadud miRNA:miRNA* dupleksi (D) üks ahel seondub RISCiga (E), teine ahel lagundatakse. RISCiga seotud miRNA seostub komplementaarsuse alusel mRNA 3'UTRile (F), mille kaudu takistatakse translatsioon seni teadmata mehhanismi abil (G).

nii valguprodukti kui ka mRNA vabanemist ribosoomilt, mistõttu mRNA tase rakus ei muutu, kuid funktsionaalse valgu tase langeb (Olsen 1999) (Joonis 1.2).

Peatükk 2

Algoritmid

Eelmises peatükis tutvustati miRNA elutsüklit, alates genoomsest eellasest lõpetades mRNAle seostunud küpse miRNAga. Loomsete miRNA sihtmärkide leidmisega genoomist on seotud mitmed probleemid. Nagu juba eespool mainitud, ei seondu loomsed miRNAd 100 % komplemetaarsusega oma sihtmärgile (nagu seda teevad mõningad taimede miRNAd/siRNAd (Rhoades *jt.* 2002)). Tõenäoliselt mängib ka RISC miRNA seondumisel mRNAle olulist rolli, mis seab ka osaliselt komplementaarse sihtmärgi tõesuse küsimärgi alla. Et nendest keerukustest üle saada, on välja töötatud mitmeid erinevaid mitmeetapilisi miRNA sihtmärkide otsimise algoritme. Järgnevalt kirjeldatakse kolme sellist algoritmi.

MikroRNA sihtmärkide tuvastamine jaguneb enamasti kolme etappi.

Esmalt koostatakse andmebaasid, kuhu kuuluvad uuritavad genoomsed (enamasti geenide 3'UTR) järjestused ja otsitavad miRNA järjestused.

Teine etapp on kindlate reeglite alusel komplementaarsuse otsing miRNA ja UTR järjestuste vahel. Kõigis kirjeldatud töödes lähtutakse samast bioloogilisest eripärast, milleks on miRNA 5' - 3' komplementaarsuse asümeetria. Eksperimentaalselt tõestatud miRNA sihtmärkidel esineb miRNA 5' otsal suhteliselt parem komlementaarsus sihtmärgiga kui miRNA 3' otsal (Joonised 2.2 ja 2.3). Sellest tingitult pööravad ka kirjeldatavad meetodid miRNA 5' otsa komplementaarsusele oluliselt suuremat rõhku, samas 3' otsaga arvestatakse vähem või üldse mitte.

Ainult komplemetaarsuse otsing võib siiski anda palju valepositiivseid tulemusi. Põhjuseks joondamisalgoritmide poolt otsitavate järjestuste lühike pikkus (mis võib olla alla poole miRNA pikkusest). Seega on oluline kasutada mingisugust täiendavat filtrit, mis kompenseeriks sellise ebatäpsuse. Sellise filtrina kasutatakse tekkindud miRNA:mRNA dupleksi termodünaamilist vabaenergiat (ΔG). Eksperimetaalselt tõestatud miRNA:mRNA dupleksitel on madalam vabaenergia väärtus, kui enamus mistahes kahe juhusliku järjestuse vahel (Rajewsky & Socci 2004; Stark *jt.* 2003).

Allkirjeldatud meedodid ei piirdu siiski vaid ΔG arvutamisega, sest ka ΔG üksi ei anna piisavalt usaldusväärset tulemust. Probleemi lahendamiseks on kasutatud erinevaid statistilisi lähenemisi, mis on ära toodud algoritmi juures.

Umbes 10 % selgrootutes leitud miRNAdest on säilinud ka imetajates. Seega on tõenäoline, et nende geenide regulatoorne mehhanism on samuti liikidevaheliselt konseveerunud. Kuna miRNAsid sisaldavad liigid on eraldatud miljonite aastate pikkuse evolutsiooniga, siis on märkimisväärne, et paljud umbes 22 aluspaari pikkused miRNAd on siiski väga sarnase järjestusega. Selline järjestuse evolutsiooni puudumine võib tuleneda sellest, et igal miRNAl on rohkem kui üks sihtmärk, mis muudab miRNA ja sihtmärkjärjestuse aluspaarilise koos-evolutsiooni väga ebatõenäoliseks (Enright *jt.* 2003). Lähtudes sellest teadmisest kasutatakse erinevate genoomide UTR järjestuste joondamist. Joondamine aitab leida rohkem konserveerunud regioone ja seega usaldusväärsemaid piirkondi tegelike miRNAde avastamiseks.

Piltliku ülevaate saamiseks algoritmi etappidest, on näitena toodud MiRanda algoritmi graafiline illustratsioon joonisel 2.1.

2.1 Algandmete leidmine, andmebaaside koostamine

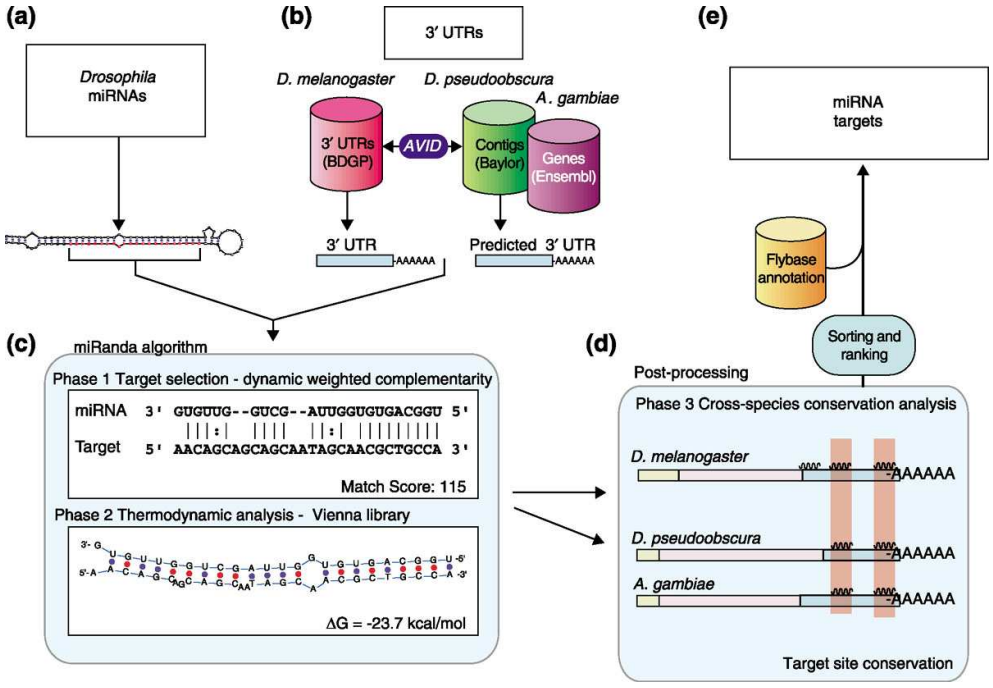
Kaks uurimisrühma (Stark *jt.* 2003; Enright *jt.* 2003) on oma UTR andmebaasi koostanud kolme genoomi geenide 3'UTRide joonduse baasil. Kasutatud genoomid on *Drosophila melanogaster, Drosophila pseudoobscura* ja *Anopheles gambiae.* Kolmas töörühm (Rajewsky & Socci 2004) on joondatud ainult *D. melanogaster* ja *D. pseudoobscura* geenide 3'UTRid.

Nendest kolmest genoomist kõige paremini kaardistatud on *D. melanogasteri* genoom, kus on olemas ka kõikide teadaolevate geenide 3'UTRid. *D. melanogasteri* 3'UTR järjestused saadi kõigil juhtudel Berkeley *Drosophila* Genoomi Projektist[1] (*Berkeley Drosophila Genome Project, BDGP*). Võrdlevad *D. pseudoobscura* geenide 3'UTR järjestused saadi kahel juhul Baylor College of Medicine *D. pseudoobscura* genoomi projektist ja kolmandal juhul Berkeley genoomibrauserist[2] (Bray *jt.* 2003; Couronne *jt.* 2003), kus on olemas kahe genoomi joondus.

A. gambiae geenide järjestused otsiti Ensembl andmebaasist (Hubbard *jt.* 2002).

[1] http://www.fruitfly.org/annot/release3.html

[2] pipeline.lbl.gov/pseudo/



Joonis 2.1: MiRanda algoritmi analüüsi sammud. Andmebaasid, mis koosnevad (a) miRNAdest ja (b) 3' UTRidest, töödeldakse kõigepealt (c) miRanda algoritmiga. Algoritm otsib komplemetaarseid järjestusi miRNAde ja 3' UTRide vahel, kasutades dünaamilist progammeerimise joondust (*Phase 1*) ja arvutab termodünaamilise vaba energia (*Phase 2*). (d) Seejärel töödeldakse kõik andmed, filtreerides kõigepealt välja tulemused, mis ei näita mingisugust konserveeruvust võrreldes *D. pseudoobscura* ja *A. gambiae* genoomiga (*Phase 3*). Järele jäänud tulemused sorteeritakse ja järjestatakse skoori alusel. (e) Lõpuks annoteeritakse kõik miRNA sihtmärkgeenide ennustused, kasutades FlyBase andmebaasi ning salvestatakse edasisteks uuringuteks (Enright *jt.* 2003).

2.1.1 EmiRSE

2.1.1.1 UTR andmebaas

D. melanogasteri geenide 3'UTR järjestustest valiti need, mis on pikemad kui 50 ap. Sama geeni erinevate transkriptide korduvad 3'UTR järjestused jäeti andmebaasist välja. *D. pseudoobscura* ortoloogsete geenide 3'UTR järjestuste kindlaks tegemiseks loeti *D. melanogasteri* avatud lugemisraami (*open reading frame, ORF*) viimased 50 aminohapet, mis joondati *D. pseudoobscura* genoomile. Joondamiseks kasutati tBLASTn (*translating Basic Local Alignment Search Tool*) algoritmi ($E \leq 10^{-}5$). *A. gambiae* ortoloogide leidmiseks kasutati sama protokolli madalamate nõudmistega ($E \leq 0{,}05$). Usaldusväärsed *D. pseudoobscura* UTR järjestused leiti umbes kahele kolmanikule *D. melanogasteri* 10196-st geenist, keskmise järjestuse konserveeruvusega 22 %. Kuna sellest kahest kolmandikust olid vähem kui pooled

ortoloogid leitavad ka *A. gambiae* genoomist, siis konserveerust viimases ei peetud oluliseks ja seda arvestati lihtsalt kui täiendavat tõendit sihtmärgi tegelikkusest.

2.1.1.2 Ennustatavad miRNAd

Töös on kasutatud 74-ja miRNA-Registris (*miRNA-Register*) (Griffiths-Jones 2004) saada olevat *D. melanogasteri* miRNAd.

2.1.2 MiRanda

2.1.2.1 UTR andmebaas

D. melanogasteri kõikide (14287 transkripti 9805 geeni kohta) 3'UTR järjestuste peptiid- ja nukleotiidjärjestusi kasutades otsiti *D. pseudoobscura* genoomist 2000 ap pikkuseid oletatavaid UTR regioone. Otsimiseks kasutati vastavalt tBLASTn ja BLASTn algoritmi. Saadud regioonid joondati *D. melanogasteri* UTRide järgi kasutades AVID (Bray *jt.* 2003) joondusprogrammi. Viimaks, kasutades sama joondust, lühendati saadud kandidaadid. Tulemuseks olid 12416 *D. pseudoobscura* 3'UTRi (mis vastab 8282 geenile). *A. gambiae* geenide 3'UTR järjestuste tuvastamiseks loeti iga geeni viimasest eksonist 2000 ap allavoolu. Ortoloogia kontrollimiseks joondati *A. gambiae* ja *D. melanogasteri* peptiidjärjestused, kasutades BLASTp algoritmi. Kokku leiti niimoodi 9823 *A. gambiae* geenide UTRi, mis vastasid *D. melanogasteri* geenide UTRidele.

2.1.2.2 Ennustatavad miRNAd

Töös kasutati 73-e unikaalset *D. melanogasteri* miRNAd, mis olid saadaval miRNA-Registris.

2.1.3 PicTar

2.1.3.1 UTR andmebaas

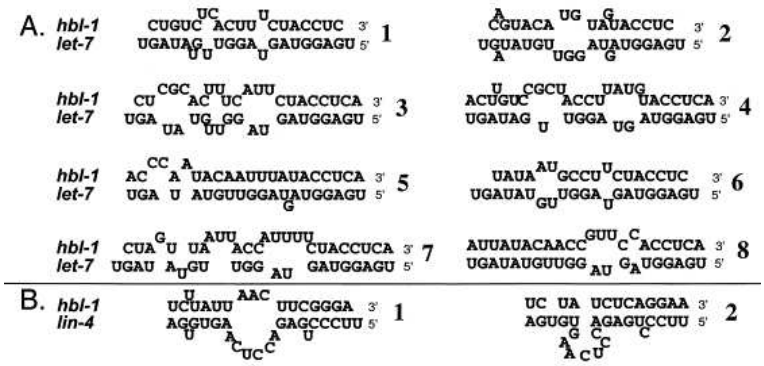
D. melanogasteri geenide 3'UTR järjestutele vastavad *D. pseudoobscura* järjestused otsiti Berkeley genoomibrauserit kasutades, kus on olemas kahe genoomi joondus. *D. pseudoobscura* regioonid, mis olid kõrgeima protsentuaalse kattuvusega, eraldati. Lisaks teostati kontroll veendumaks, et geeni kordeeriv järjestus asuks samas regioonis. Homoloogi ei leitud kahele geenile. Keskmine kattuvus genoomide lõikes oli 52 %.

2.1.3.2 Drosophila 74-ne miRNA geenide hulk

MikroRNAd mida kasutati selles töös saadi kahest allikast. Esimene hulk, 62 miRNAd saadi eksperimentaalse töö tulmustest (Aravin *jt.* 2003), kus klooniti *D. melanogasteri* väikeseid RNA järjestusi. Teine hulk saadi ühest eelnevast miRNA sihtmärkgeenide tuvastamise arvutusliku meedodi tulemustest (Lai *jt.* 2003). Viimane hulk tuvastati otsides lühikesi konserveerunud järjestusi *D. melanogasteri* ja *D. pseudoobscura* genoomsetest järjestustest. Kasutati lühikesi konserveerunud järjestusi, millel on pikenenud õlgaas-stuktuur ja antud raami piires esineb teatud lahkenvus kaha liigi vahel. *D. melanogasteri* ja *D. pseudoobscura* saadud hulgad joondati BLASTiga teineteise vastu ja leiti 12 kattuvat miRNAd (miR-274, miR-219, miR-276a, miR-33, miR-280, miR281a, miR-282, miR-248, miR-263a, miR-289, miR-287, miR-288). Need 12 järjestust lisati esimesele hulgale.

2.1.3.3 Treeninghulk

Moodustati treeninghulk, kuhu kuulus 25 eksperimentaalselt tõestatud *C. elegans'i lin-4* ja *let-7* sihtmärkjärjestust. Nendele järjestustele konstrueeriti vastavad genoomsed 30 nukleotiidi pikkused järjestused. Valitutest 10 paari on järgmised: kaheksa paari hbl/let-7 ja kaks paari hbl/lin-4 (Joonis 2.2).



Joonis 2.2: Joonisel on näha *let-7* ja *lin-4* vastavad seondumisjärjestused *C. elgegansi hbl-1* geeni 3'UTRil.

2.1.3.4 Taustsüsteem

Taustsüsteemi loomiseks kasutati positsioon-iseseisvate näitajate põhjal loodud järjestusi, mille genereerimiseks kasutatavad taustsüsteemi sagedused olid $p_A = 0.34$, $p_C = 0.19$, $p_G = 0.18$ ja $p_U = 0.29$. Sellised sagedused on kooskõlas ka treeninghulka

kuuluvate *C. elegans'i* geenide 3'UTRde koosseisuga ja klapivad baassagedustega, mis kehtivad kõigil teadaolevatel *D. melanogaster'i* geenide 3'UTRel. Sarnased baassagedused esinevad ka kui vaadelda *D. pseudoobscura* geenide 3'UTRe.

2.2 Sihtmärgi otsing

Sihtmärgi otsingu all mõistetakse käesolevas töös miRNA või selle alamjärjestuse komplementaarsuse otsimist andmebaasist (UTR andmebaasid) joondamisalgoritmide abil. Saadud tulemused on esmaseks valimiks miRNA sihtmärkide ennustamisel.

2.2.1 EmiRSE

EmiRSE (Stark *jt.* 2003) algoritm on kaheetapiline lähenemine miRNA sihtmärgi otsingule, mis seob endas tundliku järjestuse otsimist andmebaasist ja RNA voltumise algoritmi. Viimasega hinnatakse tekkinud miRNA ja sihtmärk mRNA vahelise RNA dupleksi kvaliteeti. Kvaliteet on tekkinud dupleksi vabaenergia.

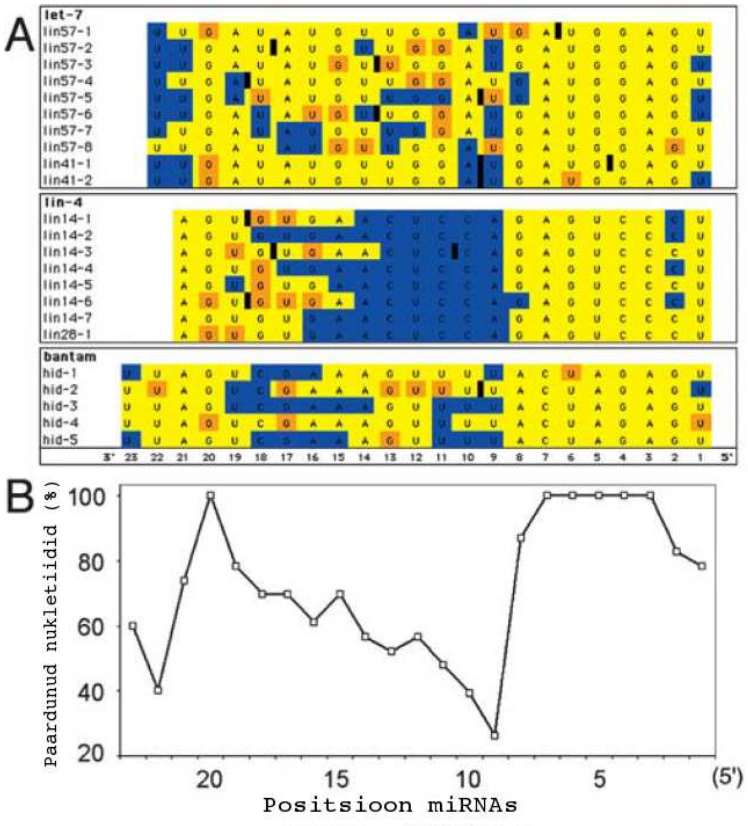
2.2.1.1 Skriinimise strateegia

Võrreldi eksperimetaalselt tõestatud miRNAde (*lin-4, let-7* ja *bantam*) sihtmärkjärjestusi, et leida neis ühiseid jooni. Kõigil uuritud sihtmärkjärjestustel oli suhteliselt parem komplementaarsus miRNA 5' otas, mujal sarnasusi ei täheldatud (joonised 2.3A ja 2.3B). MikroRNA 5' otsa esimese 8 ap sees esinesid vaid mõned mittesobivused või G:U aluspaarid.

Joondamiseks kasutati **Profiili Peidetud Markovi mudeli algoritmi** (*Profile Hidden Markov model, HMMer*) (Eddy 1996; 1998), et otsida miRNA 5' otsast esimesele 8 ap vastavaid komplementaarseid järjestusi, lubades G:U "võnkuvaid" (*i.k. "wobble"*) aluspaare. Kus võimalik, otsiti vastavad järjestused ka *D. pseudoobscura* geenide 3'UTR järjestusest. Arvesse võeti järjestused mõlemast genoomist, kuigi regioon väljaspool leitud järjestust (5' otast 8 ap) võib erineda kahe organismi vahel, viies erinevusteni järgnevas etapis (järjestus pikendatakse miRNAga ühepikkuseks, ning lisatakse veel 5 nukleotiidi).

2.2.1.2 HMMer treenimine

Iga miRNA kohta koostati kaks HMMer profiili. Kuna HMMer otsib järjestust DNAlt, siis profiilid sisaldavad miRNAga komplementaarseid järjestusi. MikroRNA



Joonis 2.3: (A) *let-7, lin-4* ja *bantam* miRNA komplementaarsus teadaolevate sihtmärkide näitel. Kollane - traditsiooniline paardumine. Oranž - G:U "võnkuv" paar. Sinine - mittesobivus. Must - ling sihtmärk järjestuses. Lingu moodustav osa sihtmärk järjestusest pole välja toodud. Järjestused on näidatud miRNA pikkusega. (B) Sama joonise (A) osa kvantitatsioon. Selline võrdlus näitab, et 5' miRNA ots omab alati head komplemetaarsust sihtmärgiga ja lubab järeldada, et otsides kaheksat esimest miRNA liiget komplementaarsuse alusel, leitakse kõik teadaolevad sihtmärgid (Stark *jt.* 2003).

esimene profiil sisaldab 5 täpset koopiat miRNA 5' otsa esimesest 8-st aluspaarist. Teine sisaldab lisaks veel 5 koopiat, kus C on asendatud T-ga ja A on asendatud G-ga, lubamaks G:U aluspaare.

2.2.2 MiRanda

2.2.2.1 Skanneerimise algoritm

MiRanda (Enright *jt.* 2003) algoritm kasutab esmasel skanneerimisel **Smith-Waterman** (Smith & Waterman 1981) algoritmile sarnast joondust, kuid erinevalt viimasest ei hinda sarnasust (A-A / U-U), vaid komplementaarsust (A=U / G≡C). Antud skoorimaatriks lubab ka G:U "võnkuvaid" aluspaare, mis on olulised täpse

miRNA:mRNA dupleksi tuvastamisel (Wuchty *jt.* 1999). Komplementaarsuse parameetrid on järgnevad:

G≡C	->	+5
A=U	->	+5
G:U	->	+2
kõik teised paarid	->	-3

Algoritm kasutab afiinsuskaristust: gapi (insertsioon või deletsioon) avamine -8 ja gapi pikendamine -2. Arvestades teadaolevate sihtmärkide võrdlustulemusi, korrutatakse esimeses 11-s positsioonis olevad komplementaarsuse skoorid (+ / -) kaalufaktoriga (antud töös 2,0), iseloomustamaks miRNA:mRNA komplemntaaruse 5' - 3' asümmeetriat. Viimaks rakendades nelja empiirilist reeglit (positsioone loetakse miRNA 5' otsast):

- mitte ühtegi ebasobivust (*i.k. mismatch*) positsioonides 2 kuni 4
- vähem kui 5 ebasobivust positsioonides 3 kuni 12
- vähemalt üks ebasobivus positsioonides 9 kuni L-5 (kus L on järjestuse (miRNA) pikkus)
- vähem kui kaks ebasobivust viimases viies positsioonis.

Saadud skoor summeeritakse üle kõigi järjestuse positsioonide ja leitakse kõik mittekattuvad hübridisatsiooni järjestused kahanevas järjekorras komplementaarsusskoori alusel kuni mingi piirväärtuseni (vaikeväärtus = 80).

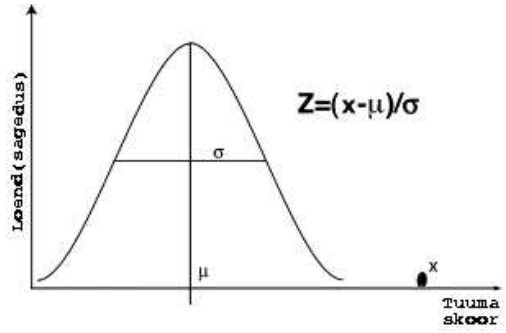
```
Read Sequence:dme-miR-7 (23 nt)
Read Sequence:CG6494-RB-u3  type=three_prime_UTR; loc=3L:8652371..8653186; ID=CG6494-RB-u3; name=h-RB-u3;
db_xref=FlyBase:FBgn0001168; release=r4.1; species=dmel; len=816(816 nt)
=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
Performing Scan: dme-miR-7 vs CG6494-RB-u3
=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-

   Forward:     Score: 117.000000  Q:1 to 24  R:441 to 466 Align Len (25) (76.00%) (80.00%)

   Query:    3' UGUUGUUU--UAGUGAUCAGAAGGU 5'
                |||:||||  | ||  |||||||||
   Ref:      5' ACAGCAAATCAGCAAAAGTCTTCCA 3'

   Energy:  -25.240000 kCal/Mol

Scores for this hit:
>dme-miR-7      CG6494-RB-u3    117.00  -25.24  0.00    1 24    441 466 25      76.00%  80.00%

Score for this Scan:
Seq1,Seq2,Tot Score,Tot Energy,Max Score,Max Energy,Strand,Len1,Len2,Positions
>>dme-miR-7     CG6494-RB-u3    117.00  -25.24  117.00  -25.24  1       23      816     440
Complete
```

Joonis 2.4: Näide MiRanda programmi väljundist. Joonisel on näha skoor, miRNA järjestus, komplementaarsus mRNAga, vabaenergia. Tulemuste rida eraldi konkreetse sihtmärkjärjestuse kohta ja tulemuste rida antud geeni kohta kokku.

2.2.3 PicTar

2.2.3.1 "Siduv tuum"

Teadaolevate miRNA sihtmärkjärjestuste hoolikal uurimisel ilmnesid enamus juhtudel GC-rikkad järjestikused miRNAga paarduvad järjestused - "siduv tuum" (*i.k. "binding nucleus"*). Sellele põhinevalt disainiti **lihtne hindamise skeem**, mis tuvastaks sellise "siduva tuuma". Selle "tuuma" skoor on järjestikuste aluspaaride (A=U, G≡C ja G:U) kaalutud summa. Need kolm parameetrit otsiti sellised, mis eristaksid kõige paremini treeninghulga skoore juhuslikest taustskooridest. Nende väärtuste summa on Z skoor (joonis 2.5).



Joonis 2.5: "Siduva tuuma" skooride hindamine. Tausta skoori jaotus on väljendatud oma keskmise (μ) ja laiusega (hälve). Treeninghulga keskmine (x) on paremal. Z skoor, Z = (x - μ)/δ , näitab kui palju on sihtmärk taustast erinev (Rajewsky & Socci 2004).

Parimad parameetrite kaalutud väärtused olid $w_{GC} = 5$, $w_{AU} = 2$ ja $w_{GU} = 0$. Nende parameetitega saavutatakse kõige parem eristuvus treeninghulga ja tausta vahel. Kuid siiski tuleb määrata ka piirväärtus, millest allapoole jäävaid järjestuste skoore enam sihtmärkidena ei loeta.

2.3 Struktuuri analüüs

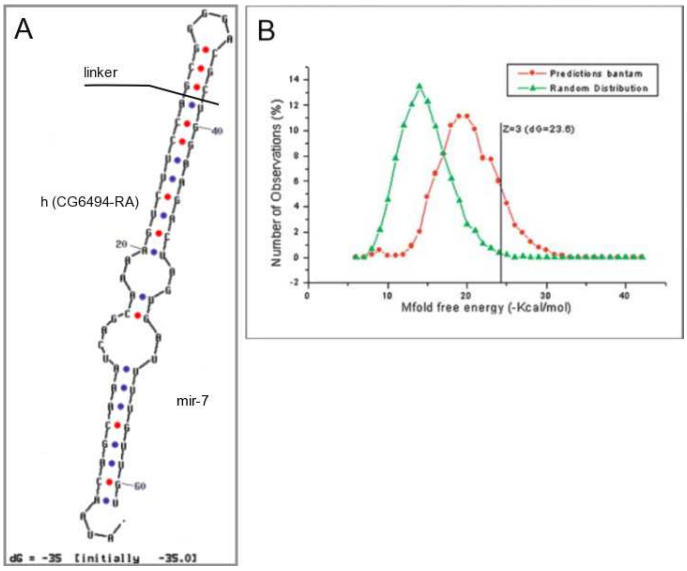
Struktuuri analüüsi all mõistetakse käseolevas töös miRNA ja mRNA vaheliste interaktsioonide hindamist seondumise vabaenergia näol. Lisaks võidakse kasutada ka statistilist töötlust või sihtmärkide konserveeruvust lähedastes genoomides. Saadud tulemused on iga algoritmi puhul lõplikuks miRNA sihtmärkide kogumiks.

2.3.1 EmiRSE

2.3.1.1 Mfold

Leitud sihtmärkjärjestused pikendati miRNAga võrdseteks lisades veel viis nukleotiidi, et lubada võimalikke linge (*i.k. bulge*). Seejärel hinnati järjestuste võimet luua keemiliselt soodsaid miRNA:mRNA duplekseid kasutades Mfoldi. Mfold on programm RNA voltumise ennustamiseks, mis sisaldab endas andmeid tuntud RNA struktuuridest koos nende termodünaamiliste parameetritega. Näiteks seosed, mis on olulised RNA paardumise vabaenergia arvutamiseks (Mathews *jt.* 1999; Zuker 2003).

Mfold vajab sisendina vaid ühte lineaarset järjestust. Sisendi saamiseks liideti iga ennustatud sihtmärkjärjestus temale vastava miRNAga standardse juuksenõela-struktuuri-moodustava linker järjestuse (GCGGGGACGC) abil. Näide Mfold kasutajaliidese väljundist on joonisel 2.6A.



Joonis 2.6: (A) Graafiline representatsioon Mfoldi väljundist miR-7 miRNA ja h(CG6494) (*hairy*) geeni 3'UTR vahel (Z = 6.29, ΔG = -35.0, sihtmärk asub 439 aluspaari allavoolu stop koodonist). Mfoldi kasutates on oluline siduda miRNA ja sihtmärk järjestus kokku juuksenõela moodustava linker-järjestusega. (B) 10000 juhusliku saidi (rohelised) ja ennustatud *bantam* miRNA saitide (punased) protsentuaalne jaotumine Mfold poolt arvutatud vabaenergia alusel. x-teljel on igale saidile eraldi Mfoldi poolt arvutatud ΔG (Stark *jt.* 2003).

2.3.1.2 Z-väärtus

Mfoldiga arvutati seondumise vabaenergia (ΔG) igale ennustatud sihtmärgile, mis võimaldas ΔG alusel sihtmärgid järjestada. Siiski tuleb arvesse võtta, et miRNA:mRNA vabaeneriga oleneb miRNA pikkusest ja dinukleotiidide komositsioonist (seondumise vabaenergia oleneb oluliselt sellest, millised nukleotiidid asuvad järjestikku). Seega ei ole võimalik eristada tegelikke saite juhuslikest, kasutades ainult ΔG või võrreldes erinevaid miRNAsid. Eristamaks juhuslikke sihtmärke[3] tegelikest, arvutati Z-väärtus, ehk mittejuhuslikkuse mõõt. Juhuslike sihtmärkide keskmine vabaeneriga võeti Z = 0 (Z = standardhälve (*Standard Deviation, SD*) üle tausta keskmise skoori). Joonisel 2.6B on toodud *bantam* miRNA sihtmärkide voltumisenergia jaotus võrreldes 10 000 juhuslikult valitud sihtmärkjärjestusega.

Enamus eksperimentaalselt kinnitatud sihtmärkgeenide 3'UTR järjestustes on rohkem kui üks miRNA seondumiskoht. Z-väärtus võimaldab ühes UTR järjestuses esinevad erinevad skoorid summeerida, kasutades vaid neid skoore, mis taustast erinevad ($Z \geq 3$). Selline lähenemine ei ole võimalik, kasutades ainult ΔG üksinda, sest isegi keskmisel juhuslikul ennustusel on hea vabaenergia näitaja. Mitme juhusliku sihtmärgi summa UTRis võib olla kõrgem kui üksikul õigel ennustusel (Joonis 2.6B). Z-väärtuse lävendiks valiti $Z \geq 3$, sest voltumis energiaga rohkem kui 3 SD sammu üle keskmise ilmneb vaid 0.3 % juhuslikest sihtmärkidest. Z-väärtuse tõstmine suurendab tõenäosust, et ennustus on tõene. Samas suureneb risk jätta välja tõelised ennustused, millel on madalam voltumise vabaenergia (Stark *jt.* 2003).

Tabel 2.1: MikroRNA sihtmärkide hulk

MikroRNA	Tulemus peale esmast skaneerimist	Tulemusi $Z \geq 3$
miR-7	14247	1051
miR-283	7973	475
miR-210	6361	864

2.3.2 MiRanda

2.3.2.1 Vienna 1.3

Selleks, et hinnata ennustatud dupleksite termodünaamilisi omadusi, kasutab algoritm voltumisreegleid, mis pärinevad Vienna 1.3 RNA sekundaarse struktuuri

[3]juhuslik sihtmärk - UTR andmebaasist jusulikult valitud järjestus, pikkusega miRNA pluss 5

programmeerimisteegist (RNAlib) (Wuchty *jt.* 1999). Kasutatud laiendatud termodünaamilised parameetrid (Schneider & Sander 1996) on arvutuslikult töömahukamad kui algne skaneering aga lubab potentsiaalseid sihtmärkjärjestusi hinnata vastavalt nende voltumisenergiale. Sarnaselt Mfoldile ühildatakse miRNA ja vastav mRNA järjestus üheks järjestuseks kasutades kaheksa aluspaari pikkust linkerit, mis sisaldab mittepaarduvaid 'X' aluseid.

Tabel 2.2: MikroRNA sihtmärkide hulk

MikroRNA	Tulemus peale esmast skaneerimist	Tulemus peale ΔG arvutamist
miR-7	16451	242
miR-283	9822	6
miR-210	9408	1755

2.3.2.2 Sihtmärkide konserveeruvus

Kõik miRNA järjestused skaneeritakse *D. melanogasteri*, *D. pseudoobscura* ja *A. gambiae* geenide 3'UTR järjestuste vastu. Sihtmärk tuvastakse kui algse Smith-Waterman'i hübridisatsiooni joonduse skoor on S $\geq$ 80 ja dupleksi struktuuri miinimumenergia on Δ G $\leq$ -14 kcal/mol. Iga sobivus miRNA ja geeni 3'UTR järjestuse vahel hinnatakse vastavalt kogu energia (*total energy*) ja kogu skoori (*total score*) alusel. MikroRNA sihtmärk loetakse konserveerunuks *D. pseudoobscuras* ja *A. gambiaes*, kui *D. melanogasteri* geeni 3'UTRis ennustatud sihtmärkjärjestusele ekvivalentne järjestus on leitav ortoloogses *D. pseudoobscura* või *A. gambiae* geeni 3'UTRi samas positsioonis. Sihtmärgid on ekvivalentsed kui *D. melanogasteri* ja *D. pseudoobscura* järjestused on rohkem kui 80 % identsed ning *A. gambiae* puhul rohkem kui 60 % identsed (Enright *jt.* 2003). Kõik skaneerimise tulemused on seejärel reastatud leitud konserveerunud sihtmärkjärjestuste alusel ja salvestatud edasisteks uuringuteks.

2.3.3 PicTar

PicTar algoritmi teine etapp sisaldab *in silico*[4] miRNA ja mRNA hübridisatsiooni, milleks kasutatakse **Mfoldi** (Zuker 2003). "Siduva tuuma" pikkus on tavaliselt 6 - 8 aluspaari, seega alla poole terve miRNA pikkusest. Eristuvuse parandamiseks

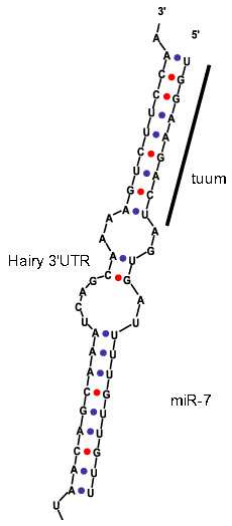
[4]in silico - arvutis

hübridiseeriti ka ülejäänud miRNA mRNAga. Seega peale "tuuma" leidmist eraldati 40 aluspaari pikkune "aken" ja hübridiseeriti see miRNAga. Erinevalt esimese meetodis kirjeldatud lähenemisest, seotakse käesoleval juhul kaks uuritavat RNA järjestust omavahel tehisliku linker segmendiga, mis koosneb mittenukleotiididest (L). Näiteks kui üks järjestus on 5' -ACGTACGT- 3' ja teine järjestus on 5' -GCATGCAT- 3', siis liidetud järjestus näeks välja 5' -ACGTACGTLLLGCATGCAT- 3'. Viimane on vajalik, sest lihtsalt kahe RNA järjestuse liitmine linkerjäägiga ja seejärel Mfoldi vabaenergia arvutamine võib anda mitte usaldusväärseid tulemusi, sest linkerjääke koheldakse nagu sisemisi lingjärjestusi, mis annavad vaba energia arvutamisel väärad tulemusi (Rajewsky & Socci 2004).

Järgmine etapp on ära hoida igasugune paardumine kahe originaalse järjestuse siseselt. See saavutati Mfold programmile spetsiaalse konfiguratsiooni faili lisamisega, kus on keelatud paardumine mingis kindlas järjestuse vahemikus.

Mõningatel miRNA:mRNA dupleksitel treeninghulgas oli silmapaistvalt madal vaba energia (vähem kui -27 kcal/mol). Seega vaba energia skoori võib kasutada märkimaks väljapaistvaid kandidaate, kuid selle meetodi põhirõhk lasub siiski "tuuma" skooril.

Mfold programm kasutati temperatuuri sätetega 20^o C ja vaikeväärtustega teistel juhtudel[5].



Joonis 2.7: Näide *D. melanogaster'i* geeni *hairy* seondumine temale vastava miRNA *mir-7-ga*. Joonisel on näha mRNA:miRNA dupleks sellisena, nagu Mfold seda ennustas. Vabaenergia on antud saidil -30.6 kcal/mol ja "siduva tuuma" skoor on 30. "Siduv tuum" asub miRNA 5' otsas. Sihtmärksait asub 438 aluspaari allavoolu stop koodonist. (Rajewsky & Socci 2004)

[5]http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form1-2.3.cgi

Peatükk 3

Algoritmide võrdlus

3.1 Kasutatud andmed ja programmid

Käesoleva töö raames kasutati andmebaasina *D. melanogasteri* geenide 3'UTR järjestusi, mida on võimalik alla laadida Berkeley *Drosophila* Genoomi Projekti veebilehelt[1]. Võrreldes eelnevalt kirjeldatud töödega kasutati algoritmide võrdlemiseks sama organismi uuemat genoomi annotatsiooni (r4.1). Sarnaselt Stark *jt.* grupiga (Stark *jt.* 2003) jäeti andmebaasist välja erinevate transkriptide korduvad UTR järjestused ja UTR järjestused, mis olid lühemad kui 50 ap. Saadud andmed salvestati Fasta formaadis.

Algoritmide võrdluseks valiti miRNA-Registry andmebaasist kõigi *D. melanogasteri* miRNAde seast juhslikult 31 miRNA järjestust (let-7, miR-11, miR-133, miR-210, miR-280, miR-303, miR-311, miR-317, miR-3, miR-79, miR-9a, miR-100, miR-125, miR-14, miR-219, miR-283, miR-305, miR-313, miR-318, miR-4, miR-7, miR-10, miR-12, miR-1, miR-275, miR-286, miR-309, miR-316, miR-33, miR-5, miR-87).

EmiRSE algoritm kasutab mitut programmi eraldi: HMMersearch (Eddy 1998), Mfold (Zuker 2003) (isiklik kontakt). Algoritmide võrdluseks kasutati veebis saadaval olevaid tulemusi.

MiRanda algoritmi võrdluseks kasutati allalaetavat programmi – miRanda[2].

PicTar'i programmi versioon on hetkel väljatöötamisel. Seetõttu ei ole võimalik saada selle meetodiga võrreldavaid tulemusi ja seepärast jäi edaspidisest võrdlusest kõrvale (v.a. Mfold vabaenergia).

[1] http://www.fruitfly.org (aprill 2005 a.)

[2] http://www.microrna.org/miranda.html

3.1.0.1 Võrdlus

Antud tööd võrreldi EmiRSE Alexander Stark'i *jt.* saadud tulemusi MiRanda programmi poolt ennustatud tulemustega. MiRanda programmi piirväärtustena kasutati vabaenergia $\Delta G \leq$ -20 kcal/mol ja skoor $S \geq 80$. Vabaenergia -20 kcal/mol on programmi poolt antav vaikeväärtus. Skoor $S \geq 80$ on artiklis (Stark *jt.* 2003) pakutud vaikeväärtus, kuigi programm ise pakub vaikeväärtuseks skoor ≥ 50. Skooride selline lahknevus jääb autorile mõistetamatuks. EmiRSE algoritmi tulemused saadi veebist[3]. Valitud miRNAde puhul järjestati ennustused Z(me) väärtuse[4] (vt. peatükk 2.3.1.2) järgi ja valiti geenid, mille puhul $Z \geq 3$.

Kõigile valitud 31-le miRNAle leiti ennustatavad sihtmärkjärjestused, kasutades miRanda programmi. Eraldi faili salvestati kõik geenid, milles esines üks või rohkem sihtmärkjärjestust üle miRanda piirväärtuste. Saadud geenide hulka võrreldi vastavate EmiRSE algoritmiga saadud geenide hulgaga. Tulemused on esitatud tabeli kujul lisas 1. (Tabel 3.1.) ja graafikuna (Joonis 3.1)

3.2 Tulemused

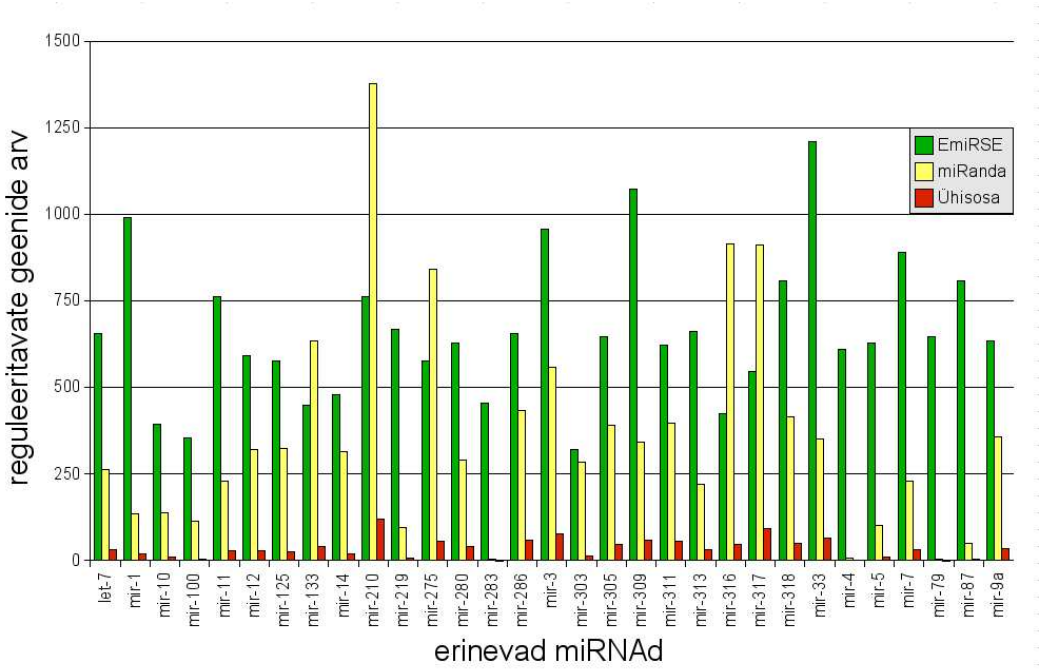
MiRanda leidis eelpool kirjeldatud parameetritega 31-le miRNAle kokku 11070 sihtmärkgeeni, mille hulgas oli 4503 erinevat geeni. Keskmiselt leiti 2,4 miRNAd geeni kohta. EmiRSA tulemustest vastas eelpool esitatud parameetritele kokku 20515 sihtmärkgeeni, mille hulgas oli 5314 erinevat geeni. Keskmiselt leiti 3,8 miRNAd geeni kohta. Erinevate algoritmide lõikes oli kattuvus 1745 geeni. Selline number on märkimisväärne arvestades sellega, et igale miRNAle eraldi ennustatud sihtmärkgeenide kattuvus oli vaid 10 % lähtudes miRanda ja 5 % lähtudes EmiRSE ennustatud geenide hulgast.

3.2.1 Komplementaarsustest

Esimese etapina mõlemas algoritmis toimub komplementaarsusotsing kasutades kindlaid heuristilisi reegleid, nende rakendamiseks kasutatakse aga erinevaid algoritme. MiRanda algoritm kasutab Smith-Waterman algoritmi edasiarendust (vt. peatükk 2.2.2.1).

EmiRSE meedodis kasutatakse esialgseks joondamiseks HMMer algoritmi. HMMer algoritmi abil otsitakse andmebaasist ainult miRNA 5' otsa esimele **8-le** aluspaarile vastavat järjestust. Kuigi ka miRanda premeerib miRNA 5' otsa (esimesed 11 ap)

[3] http://www.russell.embl.de/miRNAs/ (aprill 2005 a.)
[4] Z(me) - Z väärtus *D. melanogasteri* puhul



Joonis 3.1: EmiRSA ja miRanda algoritmide tulemuste kattuvus. Iga miRNA kohta on loendatud kokku geenid, mida nad võiks reguleerida (ei ole vahet tehtud kas geenis on üks või rohkem seondumiskohta). Rohelisega on märgitud EmiRSA algoritmiga saadud geenid ($Z \geq 3$), Kollasega on märgitud miRanda programmi poolt ennustatud geenid ($S \geq 80$ ja $\Delta G \leq$ -20 kcal/mol). Punasega on märgitud kahe algoritmi geenide kogumikkude ühisosa.

komplementaarust kaalufaktori abil, võib hea komplementaarsus miRNA 3' otsaga kompenseerida puudujäägid 5' otsas.

3.2.2 Vabaenergia arvutamine

Nagu eespool mainitud, ei piisa usadlusväärse miRNA sihtmärgi ennustamiseks vaid joondusmeetodidest. Seepärast on miRNA sihtmärkide ennustamise algoritmid sidunud endas sihtmärkide joondamise ja joondusel tekkinud dupleksi vabaenergia arvutamise. Sarnaselt joondamisele kasutatavad mõlemad võrreldavad algoritmid ka vabaenergia arvutamiseks erinevaid algoritme. Samuti konstrueeritakse arvutamiseks vajalikud miRNA-linker-mRNA järjestused erinevalt.

Järgnevalt võrreldakse kahe erineva algoritmiga saadud vabaenergia väärtusi, mis on arvutatud miRNA miR-11 ja tema sihtmärkjärjestuste vahel. Antud võrdluses kasutatavad sihtmärjärjestused on ennustatud miRanda programmi abil.

Võrdluses on kasutatud vabaenergiaväärtusi, mis on arvutatud neljale andmehulgale:

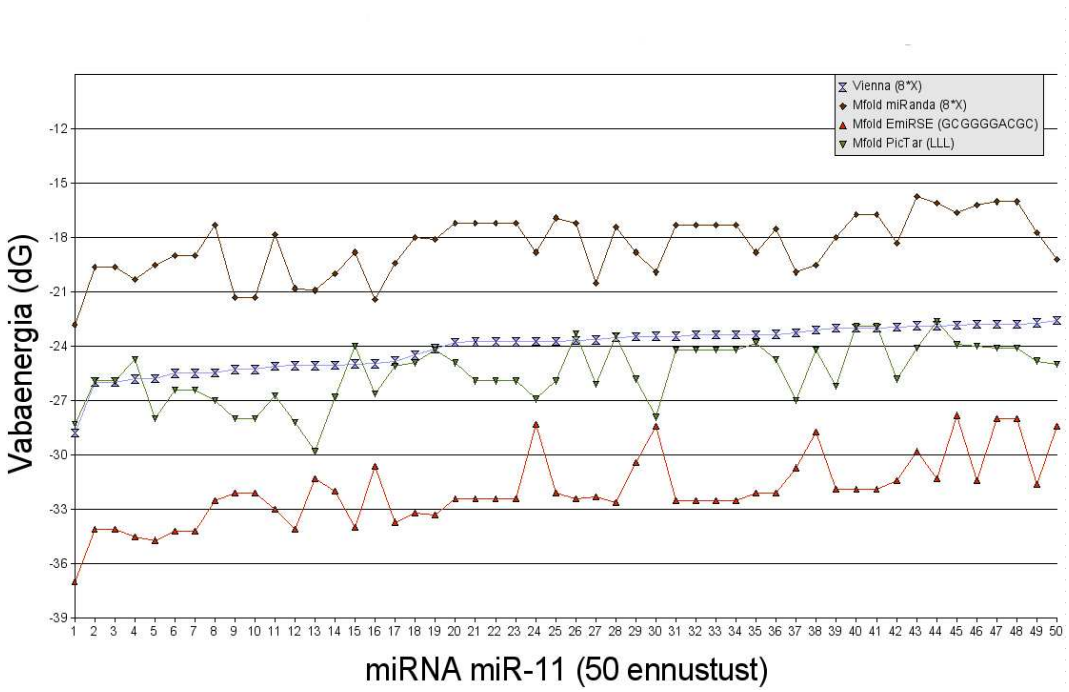
- miRanda programmi poolt arvutatud vabaenergia väärtused (Vienna 1.3, RNA-lib, linker - XXXXXXXX)
- miRanda algoritmi poolt kasutatava linkerjärjestusega arvutatud vabaenergia väärtused (Mfold[5], linker - XXXXXXXX)
- EmiRSE algoritmi poolt kasutatava linkerjärjestusega arvutatud vabaenergia väärtused (Mfold, linker - GCGGGGACGC)
- PicTar algoritmi poolt kasutatava linkerjärjestusega arvutatud vabaenergia väärtused (Mfold, temp 20^o, linker - LLL)

Arvutatud väärtuste põhjal koostati graafik (Joonis 3.2). Võrdlus näitab, et ühele ja samale järjestusele arvutatud vabaenergia (sama linkeriga arvutatud väärtused) erineb, kui kasutada erinevat algoritmi ning see erinevus ei ole konstantne (sinine ja pruun joonisel 3.2). Võrdluse teise poole eesmärk oli selgitada, kui palju mõjutab vabaeneriga arvutamist kasutatav linkerjärjestus. Tulemused on toodud samal graafikul (Joonis 3.2, pruun, roheline ja punane).

Selgus, et ka erineva linkerjärjestuse kasutamine mõjutab vabaenergia arvutamise tulemusi. EmiRSE algoritmi poolt kasutatav standardne õlgaas-struktuuri moodustav linkerjärjestus (GCGGGGACGC) annab kõigil juhtudel soodsama vabaenergia kui tehislinkerid (XXXXXXXX, LLL). Mfoldi vabaenergia ainult GCGGGGACGC struktuurile annab väärtuseks -5.6 kcal/mol. Saadud vabaenergia on väiksem kui kesmine erinevus tehislinkeritega arvutatud vabaenergiatest. Teistele linkerjärjestustele iseseisvat vabaenergiat arvutada ei ole võimalik, sest need ei moodusta iseseisev struktuuri.

Võiks arvata, et kahte tehislinkeriga saadud andmehulkade (pruun ja roheline joonisel 3.2) erinevuse põhjustab peamiselt temperatuuride erinevus. Tehtud kontrolltest näitas, et kuigi temperatuuri alandamine põhjustab küll soodsamat vabaenergiat, siis ei piisa sellest antud erinevuse kirjeldamiseks. Kontrolltestid erinevate tehislinkeri pikkustega näitasid, et peamine erinevus peitub hoopis selles. Seega vabaenetgia erinevuse annab kokku kahe suuruse: tehislinkeri pikkuse ja temperatuuride erinevuse (vahekord pole teada) muut. Siiski on arvutatavad energiad kõigil tingimustel piisavalt erinevad, et head korrelatsiooni ühegi kahe joonduse vahel ei esine.

[5]http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form3.cgi

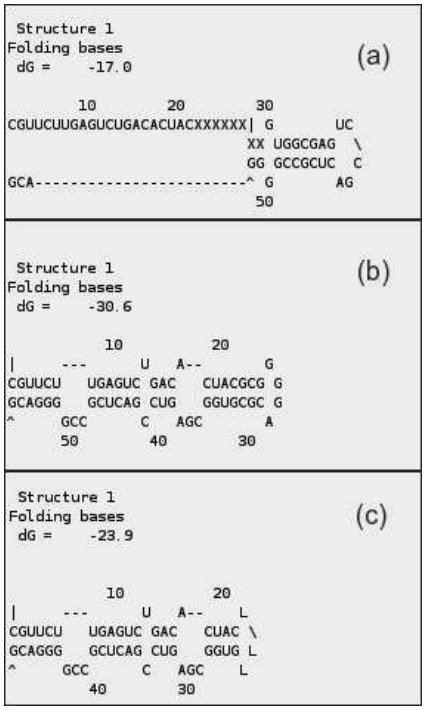


Joonis 3.2: Vabaenergia võrdlus ja linkerjärjestuse võrdlus. Erinevad vabaenergiad on joondatud vastavalt Vienna 1.3 RNAlib poolt arvutatud vabaenerigale (sinine), mis olid ka võrdluse algandmeteks. Mfoldiga on arvutatud vabaenergia kasutades sama linkerit, mida kasutas miRanda algoritm (tumepruun). Seejärel arvutati vabaenergia kasutades linkerit, mida kasutati EmiRSE algoritmi poolt (punane). Viimasena arvutati vabaenergia kasutades linkerit, mida kasutas PicTar algoritm (roheline) (PicTar vabaeneriga on arvutatud temperatuurisätetega t = 20^o C ja vaikeväärtustega teistel juhtudel. Teistel arvutustel on kasutatud temperatuurisätteid t = 37^o C).

3.2.3 Sihtmärkide statistiline töötlus

miRanda algoritmi peamine sihtmärkjärjestuse valideerimine lähtub nende konserveeruvusest lähedastes genoomides (*D. melanogaster, D. pseudoobscura* ja *A. gambiae*). Ennustus loetakse konserveerunuks (tõeseks), kui ta esineb kõigis genoomides ortoloogse geeni 3'UTR järjestuse samas positsioonis.

EmiRSE algoritm arvutab igale miRNAle oma taustsüsteemi, nö. mittejuhuslikkuse mõõdu. Mfoldi vabaenergia arvutatakse UTR andmebaasist juhuslikult valitud 10000-le miRNA pluss viis pikkusega järjestuse ja antud miRNA vahel. Saadud andmekogumile arvutatakse keskmine skoor (Z = 0) ja standardhälve. Mittejuhuslikeks loetakse miRNAle ennustatud sihtmärkjärjestused, mille Z $\geq$ 3. Liikidevahelist konserveeruvust loetakse kui lisatõendit sihtmärgi tõelisusest.



Joonis 3.3: Mfoldi tekstiväljundid erinevate linkerjärjestustega. (**a**) Mfoldi poolt ennustatud paardumine 8 ap pikkuse mittenukelotiidse linkerjärjestuse, (**b**) standardse juuksenõela-struktuuri moodustava linkerjärjestuse ja (**c**) lühikese mittenukleotiidse linkerjärjestuse abil.

Antud peatükis võrreldi kahe algoritmiga (EmiRSE, miRanda) saadud 31-le miRNAle ennustatud sihtmärke. Tulemustest lähtub, et ülekattuvus kahe meetodi vahel on väike. See on põhjustatud kahest peamisest erinevusest. Algoritmid kasutavad miRNA esmaseks joondamiseks erinevaid alamalgoritme, milleks on vastavalt HMMer (EmiRSE) ja Smith-Watermani edasiarendus (miRanda). Siiski ei selgita see nii väikest ülekattuvust, sest isegi kui miRanda vajab komplementaarsust ka miRNA 3' otsal, siis peaks ikkagi kõik miRanda algoritmi poolt ennsutatavad sihtmärgid esinema ka EmiRSE ennustuste seas. MiRanda algoritm võimaldab siiski gappe ka miRNA 5' otsal, aga seda juhul, kui 3' ots näitab head komplementaarsust. Selline erinevus kirjeldab aga ainult ühte osa mittekattuvusest.

Teine oluline erinevus tuleneb kasutatavast vabaenergia arvutamise algoritmist. Nagu näitas võrdluse teine pool, mängib vabaenergia arvutamisel suurt rolli, millist linkerjärjestust ja millise pikkusega kasutatakse (Joonised 3.2 ja 3.3). On raske hinnata, milline võrreldud lähenemistest õige on. Head joondust ühegi kahe hulga vahel ei tekkinud, kuid antud töö autori meelest on kõige perspektiivikam lähenemine lühikese mittenukleotiididest koosneva linkerjärjestuse kasutamine, sest standardne juuksenõela-struktuuri-moodustav linker võib lähtudes dinukleotiidide

kompositsioonist anda lisaenergiat miRNA:mRNA dupleksile, mis võib viia väärate tulemusteni. Samas liiga pika mittenukleotiididest linkeri kasutamine ei sunni kahte järjestust omavahel kindlates piirides paarduma (Joonis 3.3 (a)).

Kokkuvõte

MikroRNAd (microRNA, miRNA) on hiljuti avastatud geeniproduktide klass. Need umbes 22 aluspaari pikkused endogeensed RNA järjestused mängivad olulist rolli nii loomade kui ka taimede geeniregulatsioonis. MikroRNAd võivad seonduda mRNAle järjestus-spetsiifiliselt, mille tulemusel represseeritakse translatsioon või lagundatakse mRNA.

Mõiste – mikroRNA võeti kasutusele alles 2001 aastal, kui uusi avastatud lühikesi RNA molekule ei olnud võimalik enam paigutada seni tuntud rühmadesse (siRNA, stRNA jne.). MikroRNAde edasises uurimises on väga olulisel kohal bioinformaatilised meetodid, mille põhirõhk on võimalike miRNA sihtmärkide tuvastmine genoomsest järjestusest.

Töös käsitletakse mikroRNA sihtmärkide tuvastmisega seotud probleeme. Töö esimeses pooles on antud ülevaade miRNAdest, nende avastamise ajaloost ja uurimise plahvatuslikust arengust kuni praeguse hetkeni.

Kirjeldatud on kolme miRNA sihtmärkide ennustamise algoritmi: EmiRSE, miRanda ja PicTar. Valitud algoritmid on omalaadsete seas esimesed ja nad ka avaldati peaaegu üheaegselt. Tänu nende ennustustele on bioloogiline roll tuvastatud mitmele miRNAle.

Meetodite arendamise peamiseks takistuseks on seni ebapiisav eksperimentaalselt tõestatud sihtmärkide olemasolu. Nende algoritmide avaldamise kuupäevaks olid eksperimetaalselt tõestatud sihtmärgid teada vaid kolmele miRNAle.

Vaatamata kasinale treeninghulgale on kasutatud meetodid osutunud küllalt läbinägelikeks, et tagada hea stardiplatvorm järgmistele samalaadsetele algoritmidele.

Käesoleva töö ülesehituses on lähtutud algoritmide etapiviisilisest lähenemisest: andmebaaside ja treeninghulkade loomine, esmane järjestuse otsing komplementaarsuse alusel ja tekkinud dupleksi kvaliteedi hindamine termodünaamiliste reeglite põhjal. Kirjeldatakse igat etappi eraldi. Igas etapis on kõrvuti kolm algoritmi. Sellist ülesehitust kasutatakse rõhutamaks erienvate etappide olulisust ja ka parema võrdlusmomendi saamiseks.

Töö viimases osas on läbi viidud kahe algoritmiga (EmiRSE ja miRanda) saadud

tulemuste võrdlus. Saadud võrdluse tulemuste põhjal võib järeldada, et ülekattuvus kahe hulga vahel on väga väike. Ühisosa kahe algoritmiga ennustatud miRNA sihtmärkide hulgast oli keskmiselt vaid $\sim$7%.

Kuigi algoritmid on sarnased oma ülesehituse ja töö prinsiibi poolest, näitab selline tulemuste lahknevus, et meetodeid miRNA sihtmärkide ennustamiseks tuleb veel edasi arendada. Uute algoritmide loomisele aitab kaasa ka järjepidev miRNA sihtmärkide tuvastamine.

MicroRNA targeting algorithms

Priit Adler

Summary

MicroRNAs (miRNAs) are short non-coding RNAs that regulate gene expression in plants and animals. Although their biological importance has become clear, remains less well understood how they recognize and regulate target genes. In this paper we give a brief overview about miRNAs and their targeting algorithms.

In the first part we describe miRNAs and their biology. They haven't been discovered until recently, but the research is very intensive and new knowledge has been added every day. Up to present time the miRNAs themselves have been described quite well. Nevertheless the finding of miRNA targets remains a problem. In the second part of the work we examine three simultaneously published algorithms on this subject.

Algorithms under discussion are EmiRSE algorithm by Stark *et al.*, miRanda algorithm by Enright *et al.* and PicTar algorithm by Nikolaus Rajewsky and Nicholas Socci. These algorithms were chosen due to the fact that they were published simultaneously and that they were the very first of their kind.

The main obsticle for further development of miRNA targeting algorithms is the lack of valid data about miRNA targets.

Nevertheless these algorithms are worth of starting platform for future research.

The first aim of this work is to emphasize the importance of a step by step prediction of miRNA targets. Brief overview is given of each step in those algorithms. Those steps are gathering data, inquiry based on complementarity and calculation of free energy.

In the third part there is a comparison carried out about EmiRSE and miRanda. The experimental data was compared and analyzed.

As a result of this comparison, it could be argued that the match between two set of data is very small.

Although the algorithms are quite similar, the differences in results show that the methods for miRNA targeting require further development.

Viited

Ambros, V.; Bartel, B.; Bartel, D.; Burge, C.; Carrington, J.; Chen, X.; Dreyfuss, G.; Eddy, S.; Griffiths-Jones, S.; Marshall, M.; Matzke, M.; Ruvkun, G.; and Tuschl, T. 2003. A uniform system for microRNA annotation. *RNA* 9(3):227–279.

Ambros, V. 2004. The functions of animal microRNAs. *Nature* 431(7006):350–355.

Aravin, A.; Lagos-Quintana, M.; Yalcin, A.; Zavolan, M.; Marks, D.; Snyder, B.; Gaasterland, T.; Meyer, J.; and Tuschl, T. 2003. The small RNA profile during *Drosophila melanogaster* development. *Dev Cell* 5(2):337–350.

Aukerman, M., and Sakai, H. 2003. Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell.* 15(11):2730–2741.

Basyuk, E.; Suavet, F.; Doglio, A.; Bordonne, R.; and Bertrand, E. 2003. Human let-7 stem-loop precursors harbor features of RNase III cleavage products. *Nucleic Acids Res.* 31(22):6593–6597.

Bray, N.; Dubchak, I.; and Pachter, L. 2003. AVID: A global alignment program. *Genome Res.* 13:97–102.

Chen, C.; Li, L.; Lodish, H.; and Bartel, D. 2004. MicroRNAs modulate hematopoietic lineage differentiation. *Science* 303(5654):83–86.

Couronne, O.; Poliakov, A.; Bray, N.; Ishkhanov, T.; Ryaboy, D.; Rubin, E.; Pachter, L.; and Dubchak, I. 2003. Strategies and tools for whole-genome alignments. *Genome Res.* 13(1):73–80.

Eddy, S. 1996. Hidden Markov models. *Curr Opin Struct Biol.* 6(3):361–365.

Eddy, S. 1998. Profile Hidden Markov models. *Bioinformatics* 14(9):755–763.

Enright, A.; John, B.; Gaul, U.; Tuschl, T.; Sander, C.; and Marks, D. 2003. MicroRNA targets in *Drosophila*. *Genome Biology* 5(1):R1.

Griffiths-Jones, S. 2004. The microRNA Registry. *NAR* 32:109–111.

Hammond, S.; Bernstein, E.; Beach, D.; and Hannon, G. 2000. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* 404(6775):293–296.

Hubbard, T.; Barker, D.; Birney, E.; Cameron, G.; Chen, Y.; Clark, L.; Cox, T.; Cuff, J.; Curwen, V.; Down, T.; Durbin, R.; Eyras, E.; Gilbert, J.; Hammond, M.; Huminiecki, L.; Kasprzyk, A.; Lehvaslaiho, H.; Lijnzaad, P.; Melsopp, C.; Mongin, E.; Pettett, R.; Pocock, M.; Potter, S.; Rust, A.; Schmidt, E.; Searle, S.; Slater, G.; Smith, J.; Spooner, W.; Stabenau, A.; Stalker, J.; Stupka, E.; Ureta-Vidal, A.; Vastrik, I.; and Clamp, M. 2002. The Ensembl genome database project. *Nucleic Acid Res.* 30(1):38–41.

Johnson, S.; Lin, S.; and Slack, F. 2003. The time of appearance of the *C. elegans* let-7 microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. *Dev Biol.* 259(2):364–379.

Johnston, R., and Hobert, O. 2003. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* 426(6968):845–849.

Khvorova, A.; Reynolds, A.; and Jayasena, S. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* 115(2):209–216.

Lagos-Quintana, M.; Rauhut, R.; Lendeckel, W.; and Tuschl, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* 294(5543):853–858.

Lagos-Quintana, M.; Rauhut, R.; Yalcin, A.; Meyer, J.; Lendeckel, W.; and Tuschl, T. 2002. Identification of tissue-specific microRNAs from mouse. *Curr Biol.* 12(9):735–739.

Lagos-Quintana, M.; Rauhut, R.; Meyer, J.; Borkhardt, A.; and Tuschl, T. 2003. New microRNAs from mouse and human. *RNA* 9(2):175–179.

Lai, E.; Tomancak, P.; Williams, R.; and Rubin, G. 2003. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* 4(7):R42.

Lau, N.; Lim, L.; Weinstein, E.; and Bartel, D. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294(5543):858–862.

Lee, R., and Ambros, V. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* 294(5543):862–864.

Lee, E.; Feinbaum, R.; and Ambros, V. 1993. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complentarity to lin-14. *Cell* 75:843–854.

Lee, Y.; Jeon, K.; Lee, J.; Kim, S.; and Kim, V. 2002. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.* 21(17):4663–4670.

Lee, Y.; Ahn, C.; Han, J.; Choi, H.; Kim, J.; Lee, J.; Provost, E.; Radmark, O.; Kim, S.; and Kim, V. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425(6956):415–419.

Lim, L.; Lau, N.; Weinstein, E.; Abdelhakim, A.; Yekta, S.; Rhoades, M.; Burge, C.; and Bartel, D. 2003. The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* 17(8):991–1008.

Lund, E.; Guttinger, S.; Calado, A.; Dahlberg, J.; and Kutay, U. 2004. Nuclear export of microRNA precursors. *Science* 303(5654):95–98.

Mathews, D.; Sabina, J.; Zuker, M.; and Turner, D. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288:911–940.

Ohler, U.; Yekta, S.; Lim, L.; Bartel, D.; and Burge, C. 2004. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA* 10(9):1309–1322.

Olsen, P. 1999. The lin-4 regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev Biol.* 216(2):671–680.

Pasquinelli, A.; Reinhart, B.; Slack, F.; Martindale, M.; Kuroda, M.; Maller, B.; Hayward, D.; Ball, E.; Degnan, B.; Muller, P.; Spring, J.; Srinivasan, A.; Fishman, M.; Finnerty, J.; Corbo, J.; Levine, M.; Leahy, P.; Davidson, E.; and Ruvkun, G. 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* 408(6808):86–89.

Rajewsky, N., and Socci, N. 2004. Computational identification of microRNA targets. *Dev Biol* 267(2):529–535.

Reinhart, B.; Slack, F.; Basson, M.; Pasquinelli, A.; Bettinger, J.; Rougvie, A.; Horvitz, H.; and Ruvkun, G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* 403(6772):901–906.

Rhoades, M.; Reinhart, B.; Lim, L.; Burge, C.; Bartel, B.; and Bartel, D. 2002. Prediction of plant microRNA targets. *Cell* 110(4):513–520.

Schneider, R., and Sander, C. 1996. The HSSP database of protein structure-sequence alignments. *Nucleic Acid Res.* 24(1):201–205.

Schwarz, D.; Hutvagner, G.; Du, T.; Xu, Z.; Aronin, N.; and Zamore, P. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* 115(2):199–208.

Slack, F.; Basson, M.; Liu, Z.; Ambros, V.; Horvitz, H. R.; and Ruvkun, G. 2000. The lin-41 RBCC gene acts in the *C. elegans* heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. *Mol Cell* 5(4):659–669.

Smith, T., and Waterman, M. 1981. Identification of common molecular subsequences. *Mol Biol.* 147(1):195–197.

Stark, A.; Brennecke, J.; Russell, R. B.; and Cohen, S. M. 2003. Identification of *Drosophila* MicroRNA Targets. *PLoS Biologiy* 1(3):001–113.

Wuchty, S.; Fontana, W.; Hofacker, I.; and Schuster, P. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 49(2):145–165.

Yi, R.; Qin, Y.; Macara, I.; and Cullen, B. 2003. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev.* 17(24):3011–3016.

Zeng, Y., and Cullen, B. 2003. Sequence requirements for microRNA processing and function in human cells. *RNA* 9(1):112–123.

Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31(13):3406–3415.

Lisa 1

Tabel 3.1: Kattuvustabel miRanda ja EmiRSA tulemuste vahel.

miRNA	Ühisosa	miRanda	EmiRSA
let-7	32	263	658
mir-1	20	136	992
mir-10	11	139	394
mir-100	6	114	356
mir-11	29	229	764
mir-12	29	322	594
mir-125	26	324	577
mir-133	41	635	450
mir-14	19	316	479
mir-210	120	1379	762
mir-219	9	96	670
mir-275	55	842	576
mir-280	42	292	630
mir-283	0	6	455
mir-286	58	435	658
mir-3	78	558	959
mir-303	14	284	322
mir-305	48	392	648
mir-309	58	343	1073
mir-311	55	398	623
mir-313	31	220	662
mir-316	46	917	425
mir-317	92	912	547
mir-318	49	415	809
mir-33	64	351	1212
mir-4	1	8	611
mir-5	10	101	630
mir-7	33	230	890
mir-79	0	5	647
mir-87	5	50	808
mir-9a	36	358	634