MiRNA-d ja nende identifitseerimine genoomist *in silico*

Triinu Kõressaar
geenitehnoloogia II, mag

Tartu Ülikool
2005

RNA-de klassifikatsioon (Brosius and Tiedge, "Rnomenclature", RNA Biology 2004):

Klass II – transkriptsioon, translatsioon; sisaldavad tansleeritavaid avatud lugemisraame (ORF)
(klass II kuuluv RNA sisaldab peptiidsideme formeerumiseks vajalikku infot)
RNA-de klassi II kuuluvad:

- standard *mRNAd*
- arvatatavad valku-kodeerivad RNAd (*putative peptide-encoding RNA*) – *pepRNA või* (sama asi teise nimega) lühikest avatud lugemisraami sisaldavad mRNAd (*short open reading frames)* (*sORFmRNA*).

PepRNAd (MacIntosh et al., 2001) oletatavad mRNA geenid, mille ORF pikkus on lühem kui 100 aminohapet (vaadatakse 3nda positsiooni varieerumist ja pikima potensiaalse ORFi konserveerumise taset kõikides teatud liikides (lähedastes sugulastes, homoloogsetes geenides); pepRNAdel esines 3ndas positsioonis (aluse muutus oli eelistatud koodoni kolmandas positsioonis) 2 esimese positsiooniga võrreldes erinevat käitumist, kuid pikim potensiaalne ORF oli konserveerunud (valku-kodeeriv geen tahab säilitada aminohappelist järjestust, võimalikud on neutraalsed mutatsioonid koodoni kolmandas positsioonis, kuid mitte-valku kodeeriv geen püüab säilitada pigem nukleotiidset järjestust).

Klass I – transkribeeritavad, mitte-transleeritavad (untranslated RNAs, *utRNA*-d, *npcRNA* – non-peptide/protein coding RNAs), transleeritav avatud lugemisraam puudub, sisaldavad palju stop-koodoneid. Transkribeeritakse RNA polümeraas I, II või III poolt.

Rnome – kõik mitte-transleeritavad RNA-d rakus.

On ennustatud, et 97-98% transkriptsiooni produktidest on npcRNAd (95% pre-mRNAdest sisaldavad introne), ca pooled transkriptsiooniproduktid (va intronid) npcRNAd ja pooled mRNAd. npcRNA-d saavad aktiivseks peale nende post-transkriptsioonilist protsessimist. (Mattick. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. BioEssays, 2003)

<u>Klass I kuuluvad</u>:

- *rRNA*
- *lRNA* (large/long RNA) – nende pikkus võib olla mRNA-dega ekvivalentne ja nad võivad läbida mRNA-le sarnaseid protsesse (protsessing, polüadenülatsioon jt), kuid nad ei kodeeri antud kontekstis (nad võivad (ei pruugi) läbida kaht tüüpi posttranskriptsioonilisi modifikatsioone – ühe tagajärjel tekib valku kodeeriv transkript, teisel juhul npcRNA, viimasel on teatud regulatsiooni roll) valku
- *nfRNA*-d (non-functional RNA-s, mis kodeerivad/sisaldavad mingeid motiive või struktuure, kuid pole funktsionaalsed; on näidatud, et paljas selle valgu transkribeerimine võib muuta downstreamselt olevate geenide ekspressiooni, sellise nfRNA pikkus ja järjestus on ebaoluline)

- *sRNA* (small/short RNA)
    - small cytoplasmic RNA (scRNA, nt tRNA),
    - short interferring RNA (siRNA), microRNA (miRNA), teised väikesed RNAd
    - small nuclear RNA (snRNA) – leiduvad eukarüootse raku tuumas, osalevad RNA splaissingus ning telomeeride säilimises
    - small nucleolar RNA (snoRNA) – rRNA geenide keemiline modifikatsioon, RNA geenide metülatsioon
    - organellar small RNAs (guide RNA) – osalevad RNA editeerimisel (mRNA-dele uridiinijääkide lisamine)

Viimastel aastatel on tehtud avastus, et rakus transkribeeritakse proteiini kodeerivate RNA-de kõrval ka suur hulk erinevaid mitte-transleeritavaid RNA-sid

npcRNA-d jäid tihti uurimiste alt kõrvale, kuna:

- suhteliselt labiilsed
- rakus madalas koguses ekspresseeritud ja/või koe-spetsiifilised ning on seetõttu mitte-väga hästi normaliseeritud cDNA raamatukogudest enamsti välja jäänud
- pole osatud piisavalt npcRNAdele tähelepanu pöörata, pole tuldud selle peale, et neid põhjalikumalt uurida
- enamus biokeemilisi analüüse pole orienteeritud npcRNA-de detekteerimisele
- npcRNA-de funktsioon võib olla vaevu märgatav

Seoses miRNA-de ja teiste mitte-transleeritavate valkude suure hulga avastamisega, on tekkinud vajadus (võimalus) uurida npcRNA-sid süstemaatiliselt.

RNAi – RNA interference, so raku mehhanism, post-transkriptsiooniline geeni vaigistamine, mis reguleerib geeni ekspressiooni. Mehhanism seisneb selles, et kaheahelaline RNA tuntakse rakus ära, kui võõras kompleks, mistõttu kaheahelalises vormis olev RNA degradeeritakse.

siRNA – short interfering RNA (20-25nt), tekivad dsRNA protsessimise teel, siRNA-d vahendavad järjestuse-spetsiifilist RNA degradatsiooni

miRNA:

- üheahelaline ~22nt (19-25nt) pikkune RNA molekul
- kodeeritud genoomis kui fülogeneetiliselt konserveerunud juuksenõelstruktuuri (*hairpin or stem-loop structure*) omavad järjestused (on näidatud, et primaatide miRNA-d omavad suurt konserveerumist miRNA regioonides *(hairpin* struktuuride piirkonnas*),* kuid suhteliselt madalat konserveerumist miRNA regioonidega piirnevates alades)

MiRNA geenide jaotus genoomis pole juhuslik, rohkem kui pooled imetajate miRNA geenidest asuvad peremees-geenide intronites. Mikroarray analüüsid näitavad, et rohkem kui pooled miRNAdest on koekspresseeritud nende peremeesgeenidega.

On ennustatud, et ca 30% inimese genoomi geenidest on reguleeritud microRNA-de poolt.

MicroRNA andmete jaoks on realiseeritud andmebaas miRBase. Versioon 2.0 (2004, jaan) sisaldab infot seitsmest erinevast organismist pärit 506 miRNA kohta; praegune versioon 7.0 sisaldab infot 36 organismi (taimed, viirused, loomad) 2909, kusjuures inimeses 321.
http ://microrna.sanger.ac.uk

miRNA geenist küpse miRNA-ni

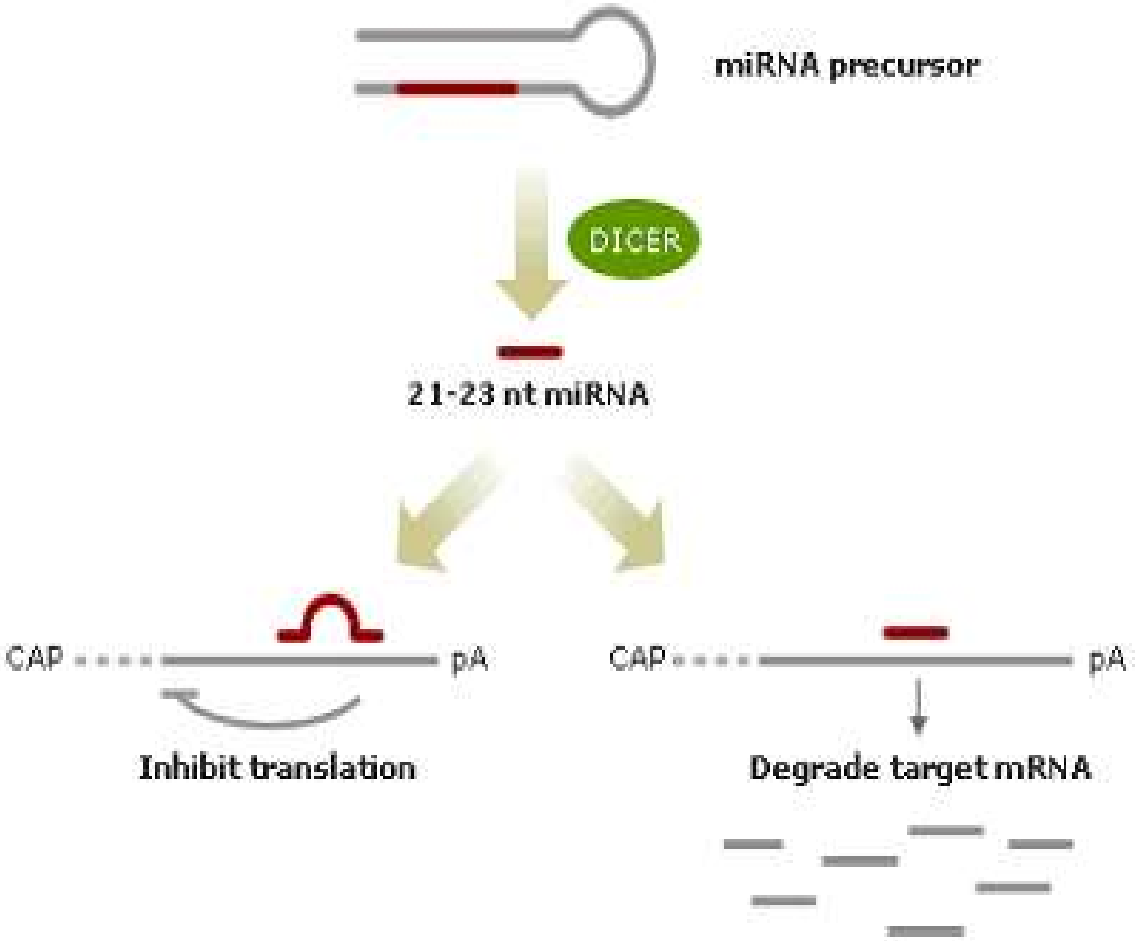
- miRNA geen transkribeeritakse polümeraas II poolt
- saadakse pikk primaarne *hairpin* struktuuri omav miRNA (pri-miRNA)
- pri-miRNA protsessitakse tuumas ~70-80 nt pikkuseks pre-miRNA-ks
- pre-miRNA transporditakse tsütoplasmasse, kus ta lõigatakse ~22 nt pikkuseks miRNA-ks
- miRNA lahti-harutamine ühe-ahelaliseks, enamasti saab üks ahel funktsionaalseks ja teine degradeeritakse

RNA polümeraas II — **Drosha** (tuuma rnaas) — Dicer (rnaas tsütoplasmas)

miRNA geen -> pri-miRNA (<500 bp) -> pre-miRNA (*hairpin-shaped*, 60-70 nt)-> küps miRNA (19-25 nt)

pre-miRNA sisaldab ~10nt *loopi* ja ~22nt kaksik-ahelat.

Arvatakse, et miRNA funktsioneerib kahe mehhanismi alusel, sõltudes miRNA ja sihtmärk mRNA komplementaarsuse tasemest – sihtmärk mRNA lagundamine (miRNA:mRNA komplementaarne või praktiliselt komplementaarne; Arvatakse, et miRNA seondub oma sihtmärgi 3' UTR regiooniga, kusjuures on tarvilik 2-7 nukleotiidi täielik komplementaarsus) ja translatsiooniline inhibitsioon.



Programmid leidmaks miRNA sihtmärkjärjestusi (*miRNA target sites*). Keeruline probleem, kuna miRNA võib oma ülesande täita seondudes mRNA-le madala komplementaarsuse tasemega. Üks miRNA võib omada mitut sihtmärk–mRNA-d. Taimedes miRNAde sihtmärkide leidmine lihtsam, kuna on näidatud taimede miRNAde suuremat komplementaarsuse taset sihtmärk mRNAdega.

Programmid leidmaks miRNA geene. miRNA algoritmid lähtuvad enamasti evolutsiooniliselt konserveerunud juuksenõel-struktuuri otsimisest (homoloogia otsingud ning võrdlev genoomika), mis võiksid kodeerida prekursor miRNA-sid (*cross-species comparision*). Keeruline ülesanne, kuna miRNA ise lühike ja miRNA geenijärjestuse primaarstruktuur vähe-konserveerunud erinevates liikides. Sekundaarstruktuur on rohkem konserveerunud liikide lõikes.

MicroScan (Ohler et al., 2003) – ennustab lähedasi homolooge nematoodidest, võetakse arvesse ka pre-miRNA sekundaarstruktuuri ning miRNA geeni *upstream* järjestuse motiive.

MiRseeker (Lai et al., 2003) – miRNA geenide ennustamine *Drosophilast* otsides kogu genoomist homoloogseid *stem-loop* struktuure.

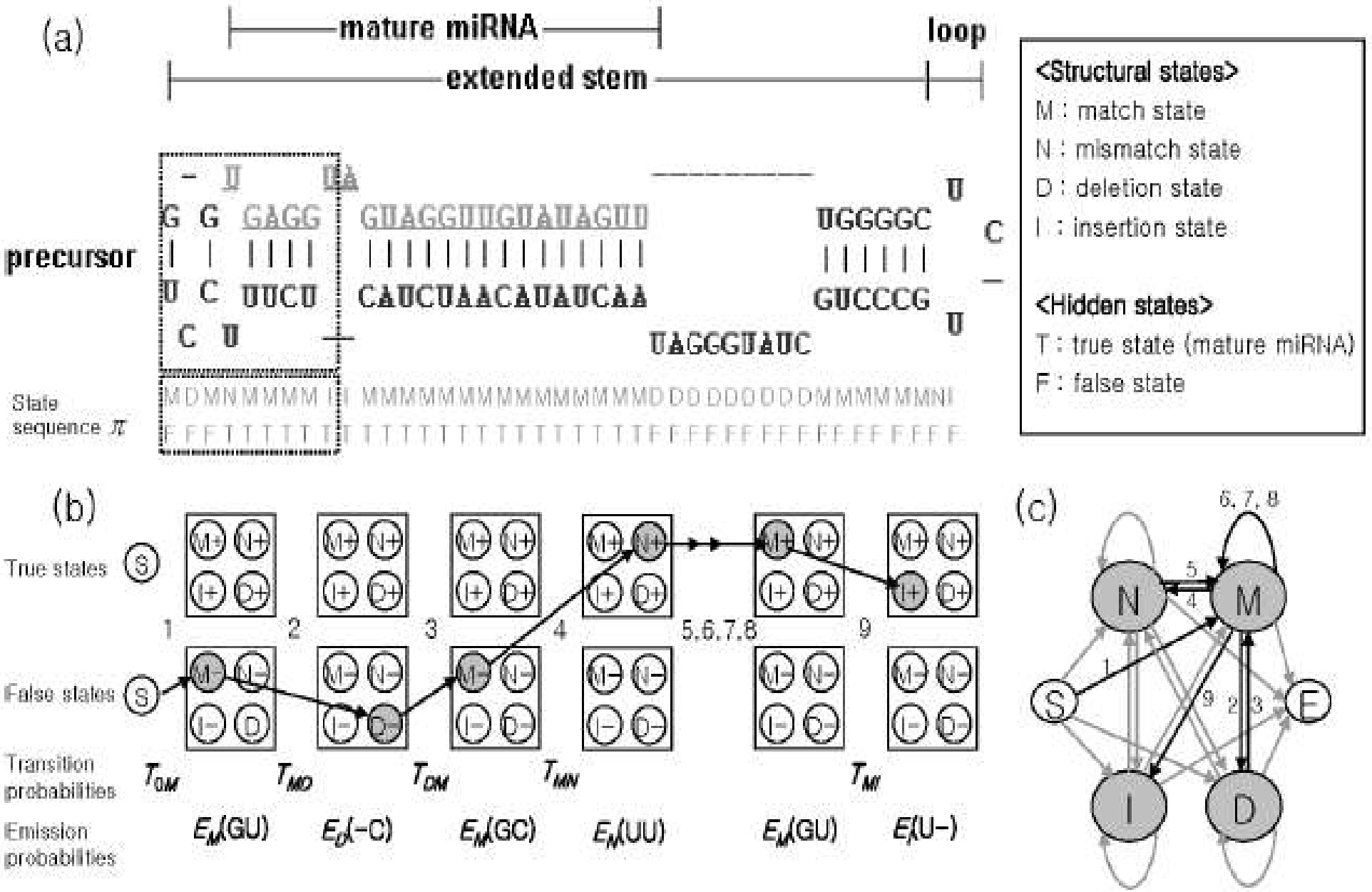
MiRfold (Billoud et al., 2005; arendatud välja *A. thaliana* taustal) – mirRNA geeni leidmiseks vajab küpse miRNA järjestust. Leitakse genoomist RNA sekundaarstruktuure ennustava programmiga (Vienna sek. str ennustav pakett) sek.str moodustavaid piirkondi küpse miRNA oletatava regiooni lähedusest. Sekundaarstruktuurile arvutatakse skoor mitte-paardunud nukelotiidide ja *bulgesid* moodustavate nukleotiidide summeerumisel. Tõeseks struktuuriks loetakse kõige väiksema skooriga struktuuri. Juhul, kui jääb alles kaks või enam ühesuguse kõige madalama skooriga struktuuri, siis arvutatakse deltaG väärtus ning võetakse kõige stabiilsem struktuur.

ProMiR (Nam et al., "Human microRNA prediction through a probabilistic co-learning model of sequence and structure" Nucleic Acids Res. 2005)

miRNAde leidmisel võtab üheaegselt arvesse nii miRNA-prekursori struktuuri kui ka järjestuse.

Algoritm baseerub tõenäosuslikul *co-learning* meetodil ja kasutab *paired* HMM-i

Metooodikat kasutades leitakse nii lähedased kui ka kauged homoloogid.


(a)
mature miRNA
loop
extended stem
precursor
State sequence π
<Structural states>
M : match state
N : mismatch state
D : deletion state
I : insertion state
<Hidden states>
T : true state (mature miRNA)
F : false state
(b)
True states
False states
Transition probabilities
Emission probabilities
1
2
3
4
5.6.7.8
9
T_{0M}
T_{MD}
T_{DM}
T_{MN}
T_{MI}
E_M(GU)
E_D(-C)
E_M(GC)
E_M(UU)
E_M(GU)
E_I(U-)
(c)
6, 7, 8
N
M
I
D
S
E

Kahe-ahelalisele pre-miRNA järjestusele antakse tõenäosus:
vaja teada emissiooni ja transitsiooni tõenäosust

Transitsiooni tõenäosus:

Olgu vaatlusalune seisundeid sisaldav järjestus (rada) π (pii). Eeldatakse, et antud seisundi tõenäosus sõltub vaid eelmisest seisundist. Kui π_i viitab i-ndale seisundile rajas, siis transitsiooni tõenäosus on arvutatav valemiga:

$$T_{kl} = P(\pi_i = l \mid \pi_{i-1} = k),$$

kus transitsioon on seisundist π_{i-1} =k seisundisse π_i=l.

Emissiooni tõenäosus:

Olgu x_i i-ndas seisundis emiteeritud sümbol. Seisundis k vaadeldava sümboli b emissiooni tõenäosus on:

$$E_k(b) = P(x_i = b \mid \pi_i = k)$$

Kasutades transitsiooni ja emissiooni tõenäosusi saame hinnata tõenäosuse P(x), et järjestus x on genereeritud tõenäosusliku *co-learning* mudeli poolt.

$$P(x,\pi) = T_{0\pi_1} \prod_{i=1}^{L} E_{\pi_i}(x_i) T_{\pi_i \pi_{i+1}}.$$

kus L on akna suurus
Kui ennustuses kasutada ainult ühte rada π*, siis kõrgeim tõenäosus tuleks valida

$$\pi^* = \arg\max_{\pi} P(x,\pi).$$

Kõige tõenäosuslikuma seisundeid sisaldava transitsiooni raja ja tema tõenäosuse leidmiseks kasutatakse laialt levinud Viterbi algoritmi(1973-st aastast).
Kuna algoritmi poolt tagastatavad tõenäosuslikud väärtused on väga väikesed ning vaadeldava järjestuse pikenedes tõenäosuste väärtused vähenevad eksponentsiaalselt, siis kasutatakse Viterbi tõenäosust fikseeritud pikkusega järjestuse jaoks – kasutatakse küpse miRNA pikkust kogu pre-miRNA pikkuse asemel (L=22).

Maksimaalne P(x,π) leitakse libiseva akna meetodil. Leitakse kaks maksimaalset P(x,π) väärtust – pre-miRNA 3' ja 5' otsast vaadelduna. Kui leitud maksimaalsed väärtused on kõrgemad kui *threshold* väärtus, siis võib klassifitseerida antud järjestust kui pre-miRNA.

Threshold väärtus saadakse *receiver operator characteristic* (ROC) kaare analüüsist. ROC kaare analüüs (meetod hindamaks diagnostiliste testide toimimist) viidi läbi allpool toodud andmesettidega. *Thresholdi* valimisel arvestati spetsiifilisuse ja sensitiivsuse vahelise suhtega – *thresholdiks* valiti (P=0.033) punkt, kus sensitiivsus on 73% ja spetsiifilisus 96%.

Andmehulgad - positiivsed: 136 eelnevalt teadaolevat inimese pre-miRNAd

- negatiivsed: teatud parameetrite alusel 1000 juhuslikult inimese kromosoomidest valitud *stem-loop* struktuuri (viimased ennustati kasutades Vienna RNA tarkvara paketti).

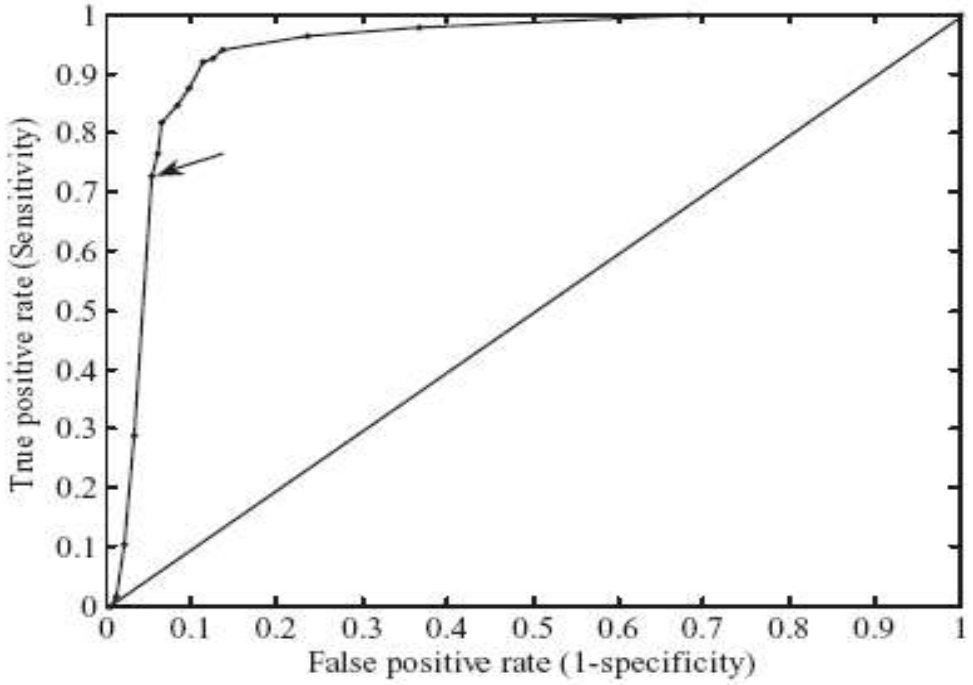


Figure 2. The ROC curve, which is defined as a plot of test sensitivity as the *y*-coordinate, versus the false positive rate (FPR; 1 − specificity) as the *x*-coordinate. The area under the ROC curve is 0.936 by non-parametric estimation. The arrow indicates the point of threshold, where $P = 0.033$, specificity is 96% and sensitivity is 73%.

Pre-miRNA valideerimine

* minimaalne vabaenergia väärtus (MFE) ja Monte Carlo simulatsioon. Vaadatakse pre-miRNA sekundaarstruktuuri termodünaamilist stabiilsust ning sekundaarstruktuuri statistilist tähtsust. Van de Peer grupp on näidanud, et miRNAde vabaenergia väärtused on võrreldes teiste ncRNAdega (statistiliselt usaldusväärselt) madalamad, samuti näidati, et ainult sekundaarstruktuuride stabiilsust ja konserveerumist kasutada miRNAde ennustamiseks pole piisav. (kasutasid Van de Peer grupi poolt välja töötatud programmi *randfold*, mis testib miRNA MFE väärtuse statistilist tähtsust baseerudes P-väärtustele (<0.05))

* kordusjärjestused. Publitseeritud miRNA järjestustest pole leitud inimese kordusjärjestuse motiive. Pre-miRNA kandidaadid, mis sisaldavad inimese kordusjärjestuse motiive tuleks elimineerida.

➔ Küpse miRNA regiooni ennustamiseks kasutavad valede seisundite transitsiooni tõenäosust.

➔ Funktsionaalse ahela ennustamiseks kasutavad mõlemale ahelale arvutatud P(x, π) väärtust. Lisaks vaatavad absoluutset ja suhtelist pre-miRNA 5' otsa (5bp) sisemist stabiilsust (kahe-ahelalise miRNA lahti harutamist initsialiseerib helikaas miRNA stabiilsemast otsast, miRNP kompleks seondub dmiRNAs 5'ahelaga, viimane saabki funktsionaalseks ahelaks)

Eksperimentaalne verifitseerimine

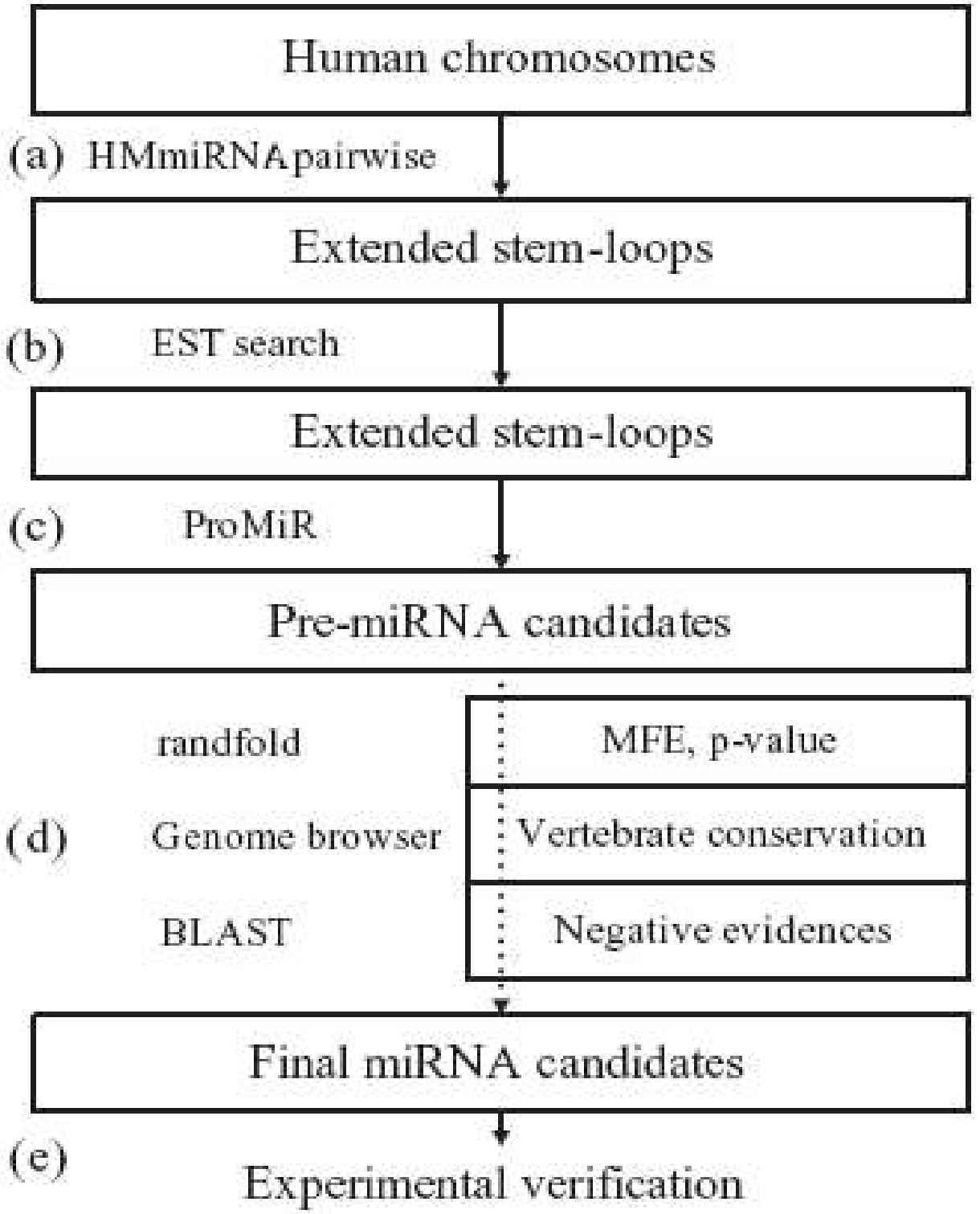
HeLa rakke inkubeeriti Drosha mRNAga komplementaarsete siRNAdega (rakkudest eemaldati Drosha, viimane lõikab kiiresti pri-miRNA pre-miRNAks, RNA interferntsi teel). Seejärel eemaldati rakkudest RNA (peaks sisaldama pri-miRNAd) ja sünteesiti vastav cDNA. Disainiti praimerid pri-miRNAde paljundamiseks ning vaadati, kas tekkis soovitud produkti.

Kui antud miRNA geeni kanditaat tõesti ekspresseerib miRNA-d, siis eeldatakse, et PCR-i produkt akumuleerub, kuna Drosha eemaldati rakust.

Pri-miRNA kvantifitseerimiseks kasutati reaal-aja PCR-i

mi-RNAde otsimine inimese 16., 17., 18. ja 19. kromosoomist

- Iga kromosoom skaneeriti 100 bp aknaga 90 bp ülekattega
- Fragmentide pöörd-komplementaarsetele järjestustele ennustati RNA sekundaarstruktuurid RNAfold programmiga Vienna RNA sekundaarstruktuuri ennustavast pakettist (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi).
- Saadud *stem-loop* struktuuridest jäeti edasiste uuringute alla järjestused, mis vastasid teatud nõuetele (struktuuri moodustava järjestuse pikkus 64-90, *stemi* pikkus rohkem kui 22 nt, *bulge* suurus alla 15 nt, *loop'i* suurus 3-20 nt, vabaenergia väärtus alla -25 kcal).


Human chromosomes
(a) HMmiRNApairwise
Extended stem-loops
(b) EST search
Extended stem-loops
(c) ProMiR
Pre-miRNA candidates
randfold
MFE, p-value
(d) Genome browser
Vertebrate conservation
BLAST
Negative evidences
Final miRNA candidates
(e)
Experimental verification

Tabel 1. Inimese miRNA geenide otsimise tulemised

Chr	Extracted stem–loop structures	Expressed stem–loops	pre-miRNA candidates	Detected known miRNA
16	65539	8153	253	2
17	68458	9367	274	11
18	34853	3135	83	4
19	62229	7765	207	7

Eksperimentaalselt uuriti 23 ennustatud miRNA kandidaati.

Tulemused: 9 oletatavat miRNA-d leidsid kinnitust
7 kandidaadi PCR-i produkti ei suudetud detekteerida
7 kandidaati ei akumuleerunud

(7 produkti mitte-detekteerimine PCRi eksperimendis võib olla tingitud kandidaatide madala sisalduse tõttu HeLa rakkudes – tuleks uurida rakke teistest kudedest ning erinevates arengufaasides)

Küpse miRNA regioon ennustati 72% täpsusega
Funktsionaalse miRNA ahel ennustati 75% täpsusega

Seega ProMir võib ennustada miRNA geene vähemalt 40 % täpsusega.

Kirjandus

Nam et al., Human micrRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic acid Resarch, 2005 Jun 24;33(11):3570-81. Print 2005.

Billoud et al., Identification of new small non-coding RNAs from tobacco and Arabidopsis. Biochimie. 2005 Sep-Oct;87(9-10): 905-10.

Costa. Non-coding RNAs: New players in eukaryotic biology. Gene, 357-2, 2005.

Griffiths-Jones S. The microRNA Registry. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D109-11.

Berezikov et al., „Phylogenetic Shadowing and Computational Identification of Human microRNA Genes.“ Cell, 2005

Wienholds. MicroRNA function in animal development. FEBS Lett. 2005

Eckstein. „Small non-coding RNAs as magic bullets“ TRENDS in Biochemical Sciences, 30-8, 2005

Mattick. Challengin the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. Bioassay 25, 2003

Ohler et al., Patterns of flanking sequence conservation and a characteristic upstream motif for microRNS gene identification. RNA, 10, 1309-1322

Lai et al., Computational identification of Drosophila microRNA genes. Genome Biol., 4, R42.