

Identifikation af potentielle microRNA gener ved hjælp af komparativ genomanalyse

Per Tøfting

23. september 2008



Speciale i softwarekonstruktion
IT-Vest
Aarhus Universitet

Agenda

- Formål
- microRNA
- Strategien
- Pipelinen
- Test
- Forbedringer

Formål

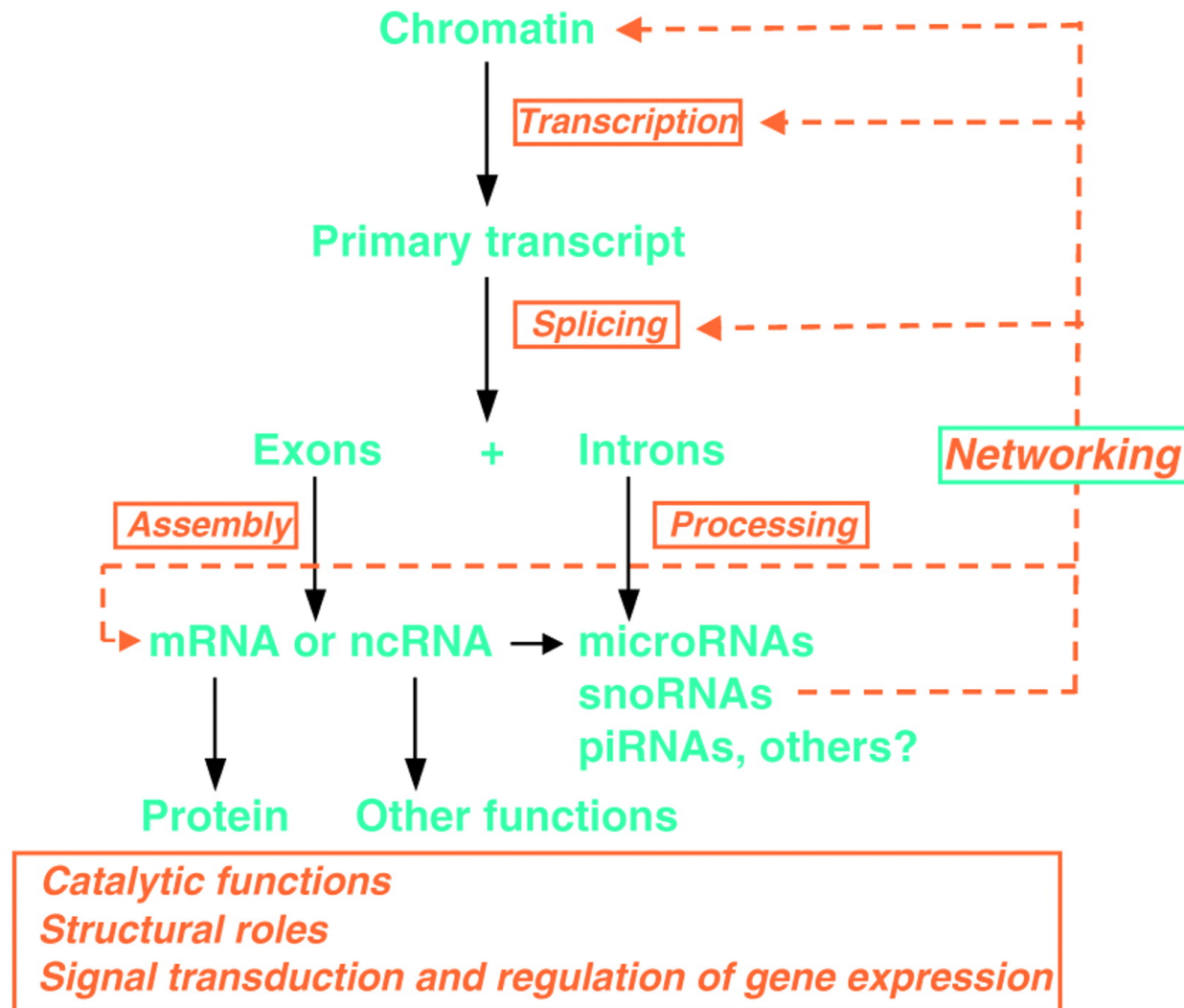
Dette speciale projekt har til formål at udvikle en pipeline, der gennem komparativ genomanalyse af tre arters genomer kan finde DNA-sekvenser, der indeholder potentielle miRNA gener.

Om de fundne DNA-sekvenser, rent faktisk koder for miRNAer, skal afgøres eksperimentelt, som er uden for dette projekts område.

Den udviklede pipeline vil bestå af egenudviklet programmer og allerede eksisterende programmer.

Hvorfor er microRNAer interessante?

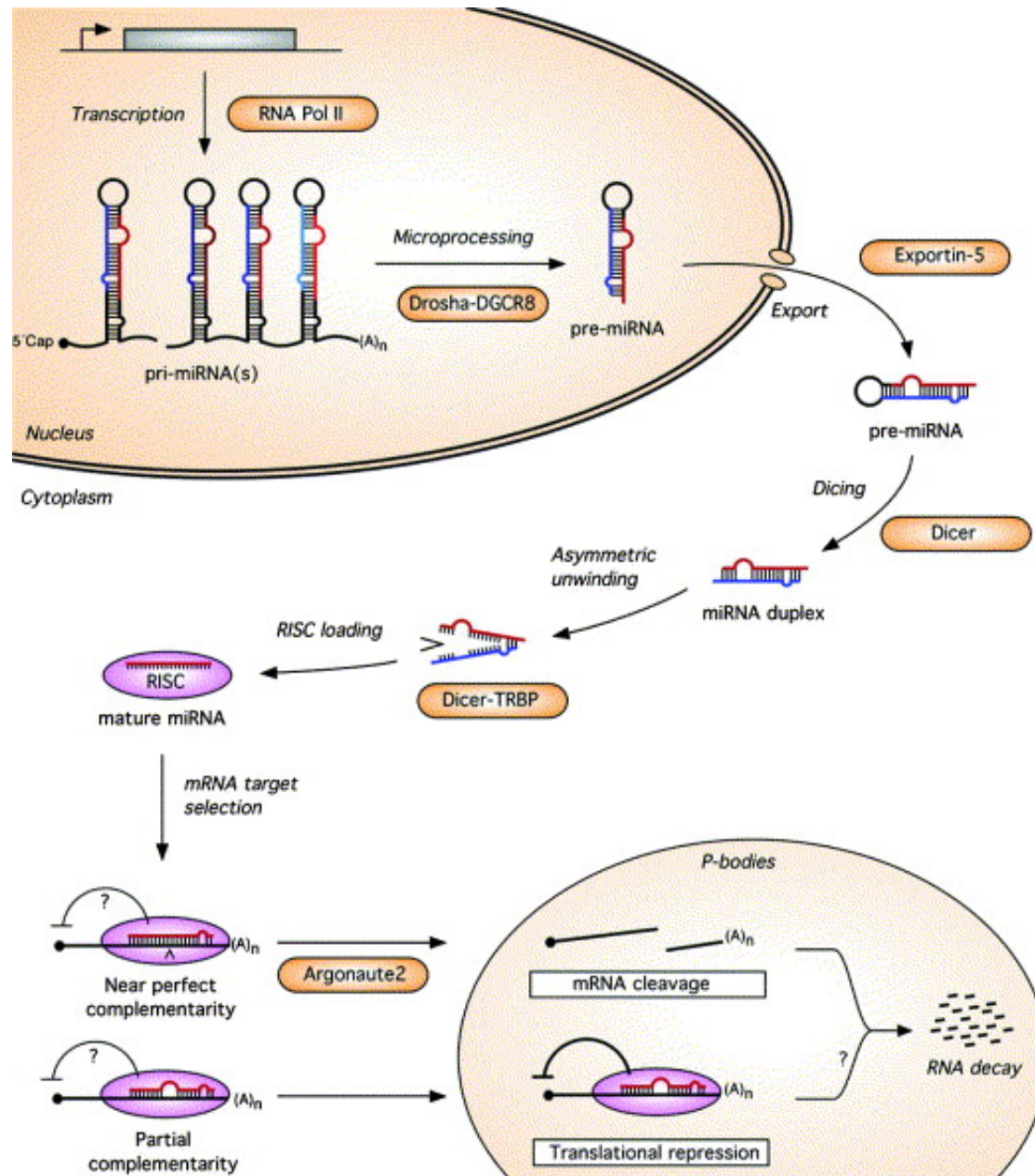
Revised definition of gene and flow of genetic information



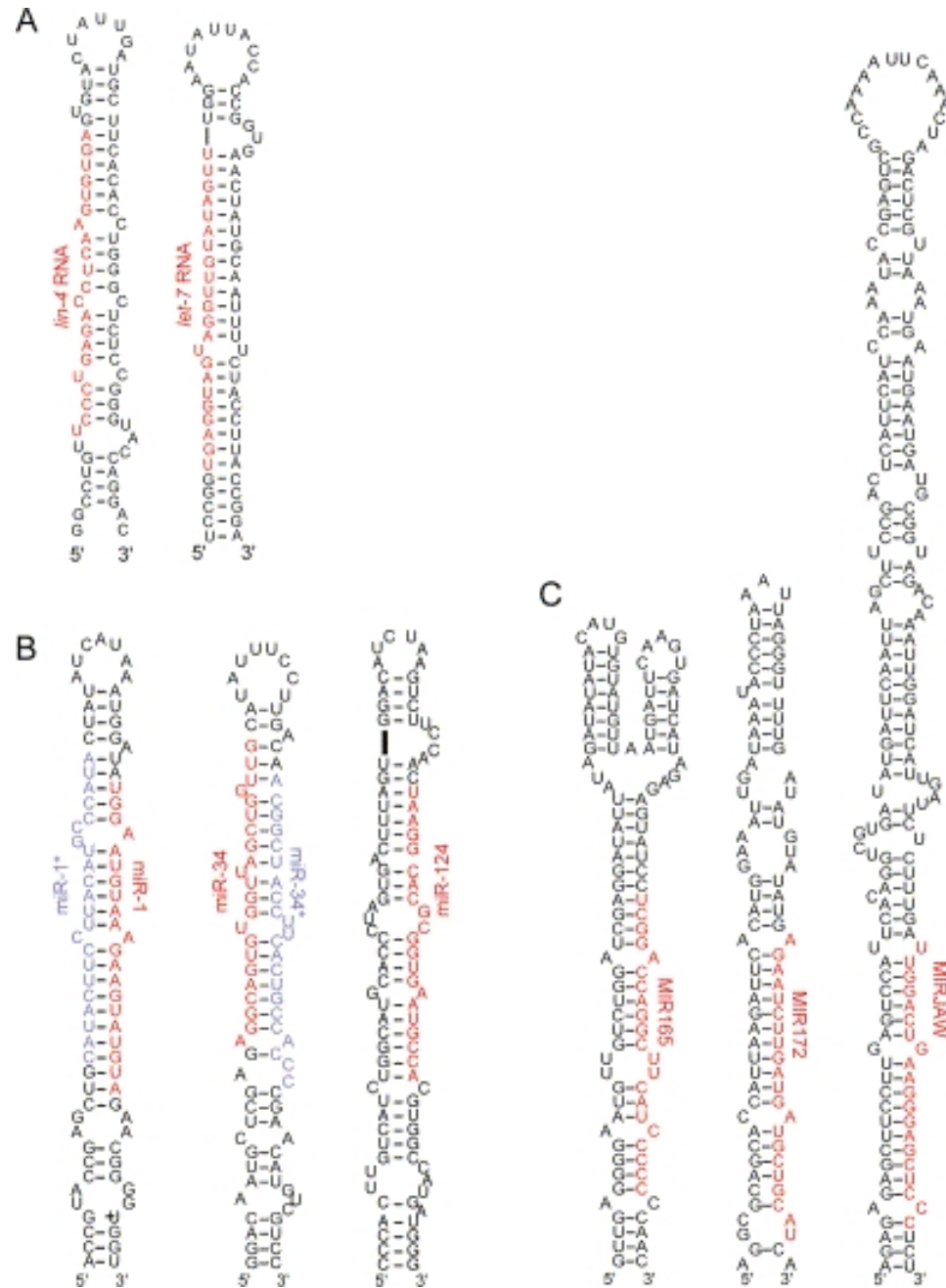
Biology's Big Bang



microRNA biogenesis



Pre-microRNAernes struktur



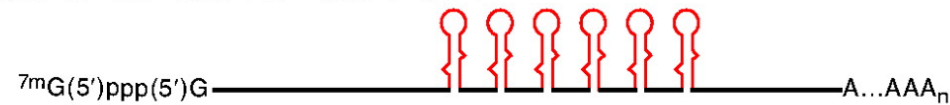
MicroRNA genes placering

A

miR-21

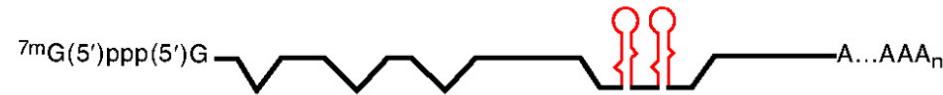


miR-17~18~19a~20~19b-1~92-1



B

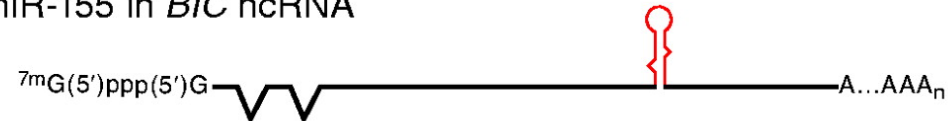
miR-15a~16-1 in *DLEU2* ncRNA intron



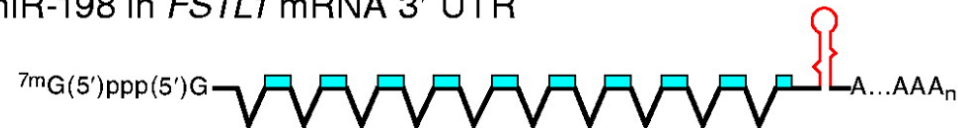
miR-106b~93~25 in *MCM7* pre-mRNA intron



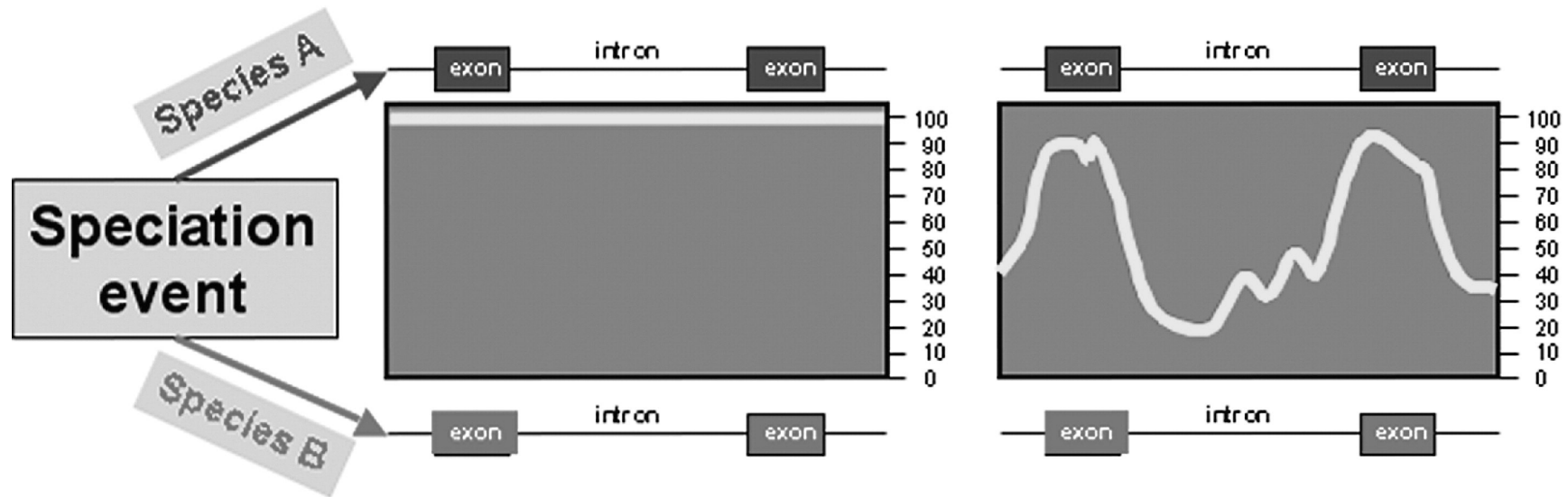
miR-155 in *BIC* ncRNA



miR-198 in *FSTL1* mRNA 3' UTR



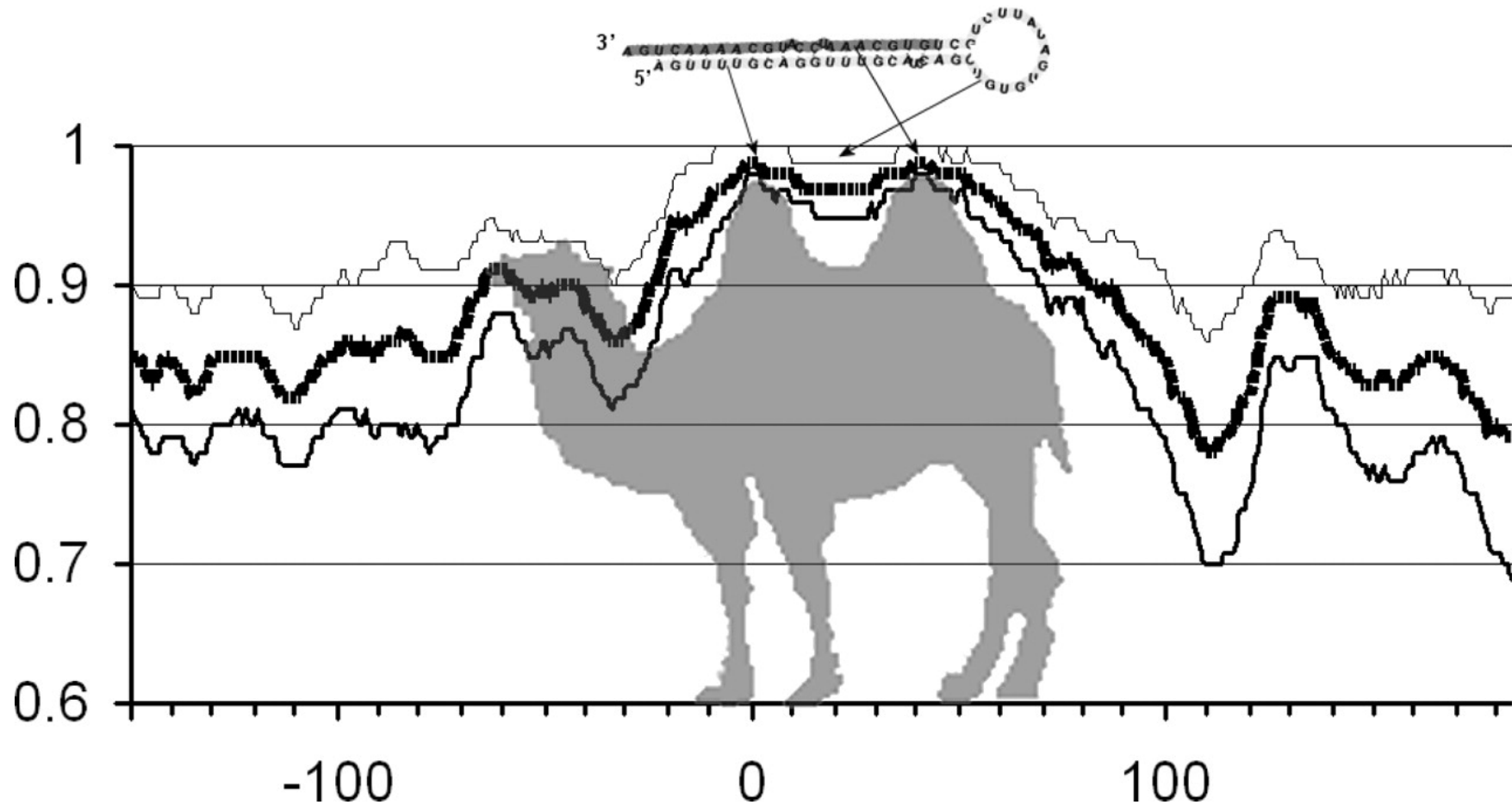
Strategien: Komparativ genomanalyse



Evolution:

- Mutationer
- Selektion

microRNAer og komparativ genomanalyse



Strategien til identifikation af microRNA gener

Strategien til computerbaseret identifikation af potentielle microRNA gener tager udgangspunkt i følgende karakteristika, for allerede kendte miRNAer:

- microRNAer er konserverede mellem beslægtede arter
- Pre-miRNAer danner stabile stem-loop strukturer
- Andre sekvens og strukturelle microRNA karakteristika

microRNAer er konserverede mellem beslægtede arter

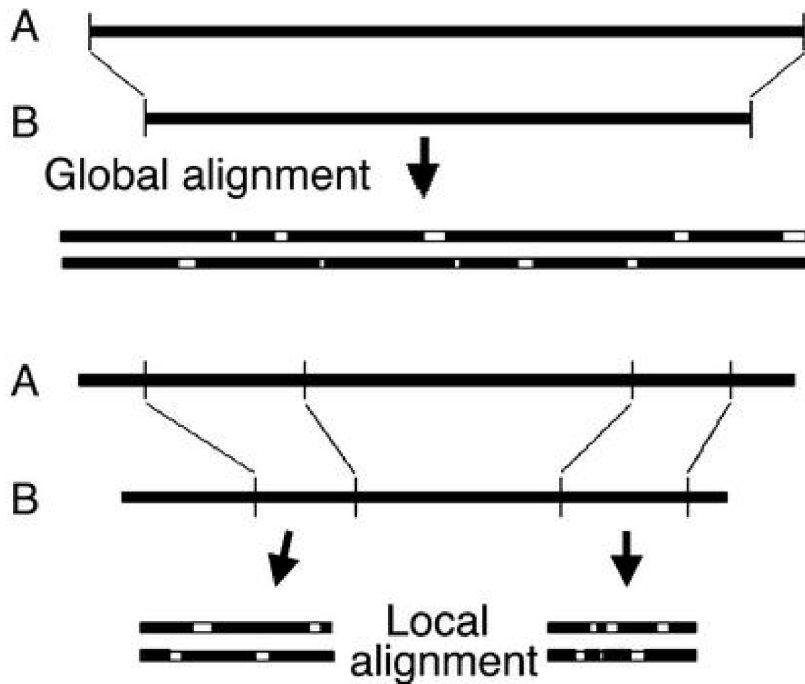
2 strategier:

- Finde DNA-sekvenser, der er homologe til kendte microRNA gener
- Finde DNA-sekvenser med potentielle microRNA gener, der ikke allerede kendes fra andre arter

Strategien bliver:

- **Identifikation af konserverede DNA-sekvenser**

Identifikation af konserverede DNA-sekvenser



Parvis alignment:

Dynamisk programmerings algoritmer:

- Global: Needleman-Wunsch algoritme
- Lokal: Smith-Waterman algoritme

Heuristiske algoritmer:

- FASTA
- BLAST
- PatternHunter
- BLAT
- BLASTZ
- BLATZ

Inputdatasættet

Hvilke arter?

- Dyr
- Planter
- Virusser

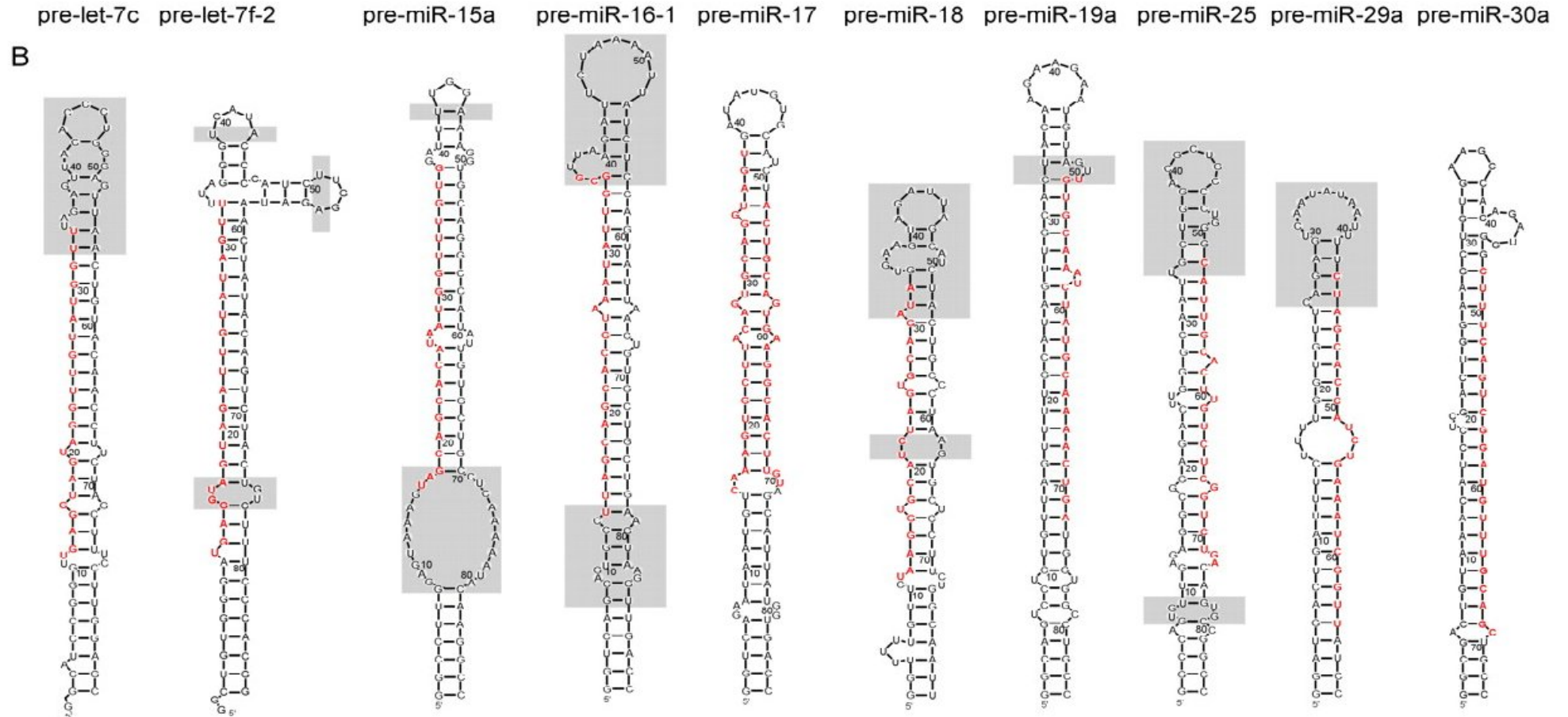
Hvor mange arter?

- 2 - 3 - ?

Hvilke DNA-sekvenser?

- Hele genomer
- Omkring kendte microRNA gener
- Repeat maskerede
- Exon maskerede

Pre-miRNAer danner stabile stem-loop strukturer



Valg af RNA foldnings program

RNA foldnings programmer

Mfold	1 sekvens	nearest-neighbor MFE
RNAfold	1 sekvens	nearest-neighbor MFE
Pfold	alignment	stokastisk kontekst-fri grammatik
RNAalifold	alignment	nearest-neighbor MFE + sekvens kovariation

RNA klassificerings program

RNAz	alignment	z-score + struktur konserverings indeks + SVM
-------------	-----------	---

Strategien Bliver:

- Identifikation af ncRNA sekvenser

Multiple alignment

Protein sekvenser

- ClustalW

Lange DNA-sekvenser

- MAVID
- Multi-LAGAN

Identifikation af potentielle microRNA gener

Triplet-SVM-classifier	1 sekvens	klassificer 1/-1	
RNAmicro	alignment	$0,50 \leq \text{score} \leq 1$	RNAz

Strategien bliver:

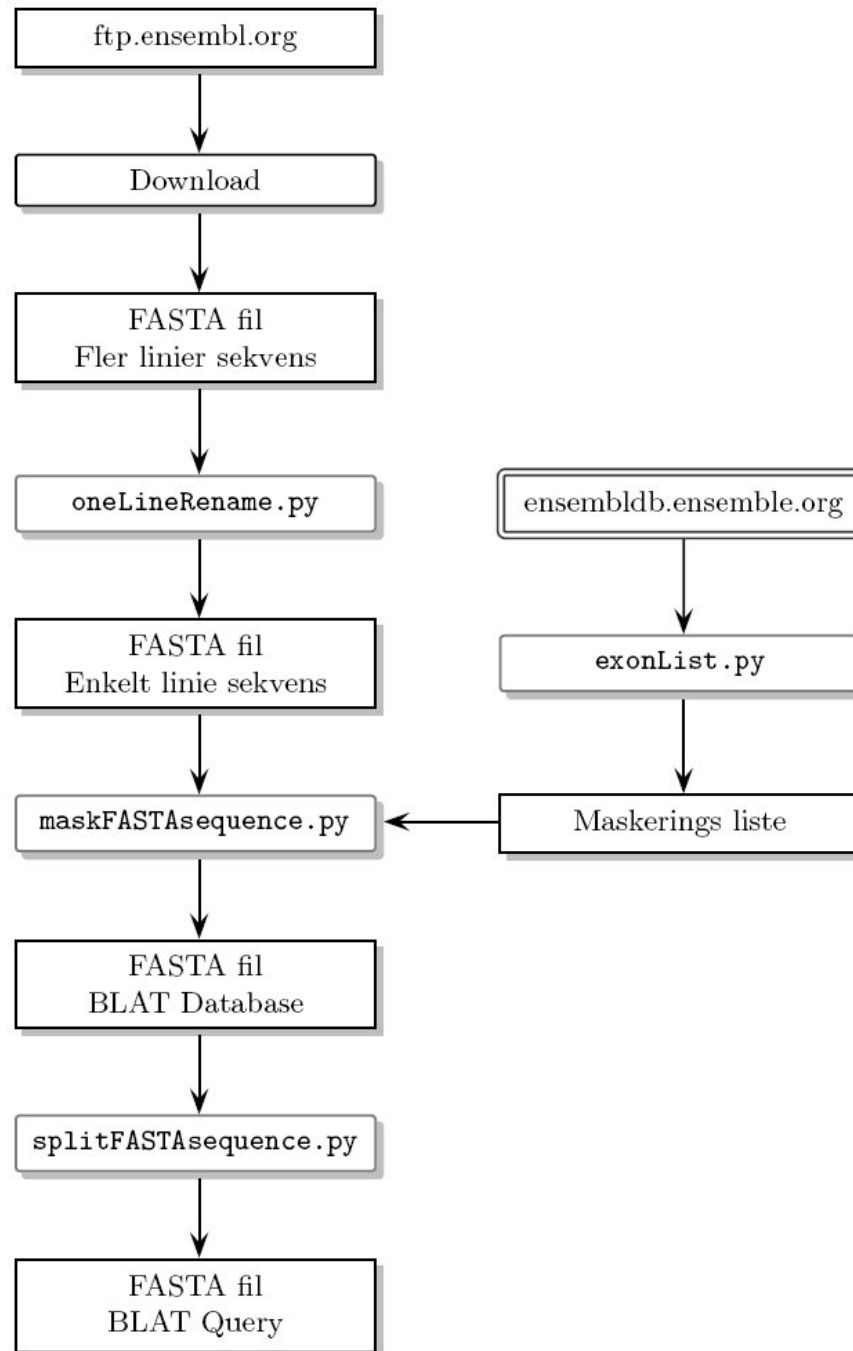
- Identifikation af potentielle microRNA gener

Pipelinen

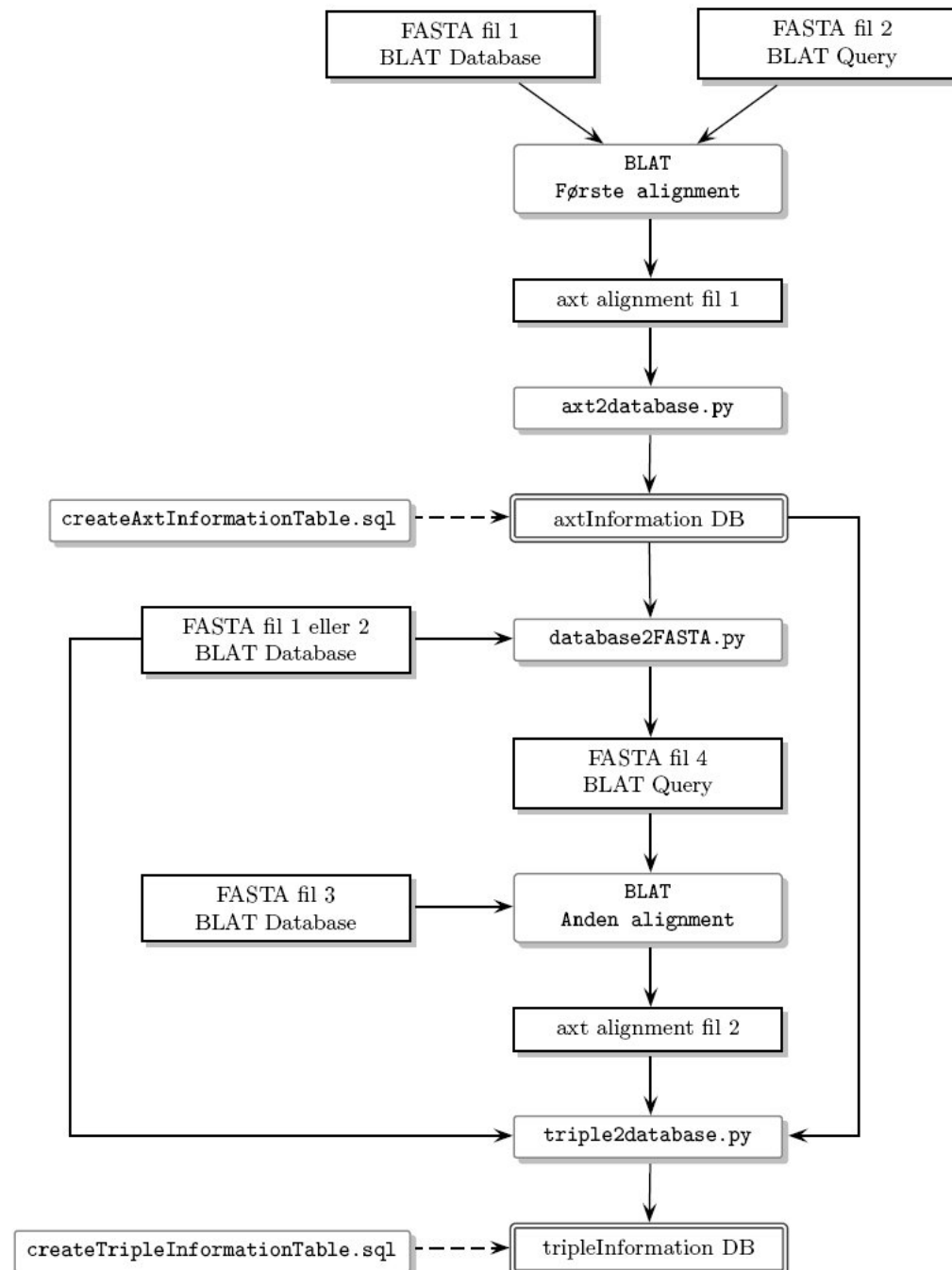
Pipelinen består af følgende tre overordnede dele:

- Forbehandling
- Konserverede DNA-sekvenser
- Identifikation

Forbehandling

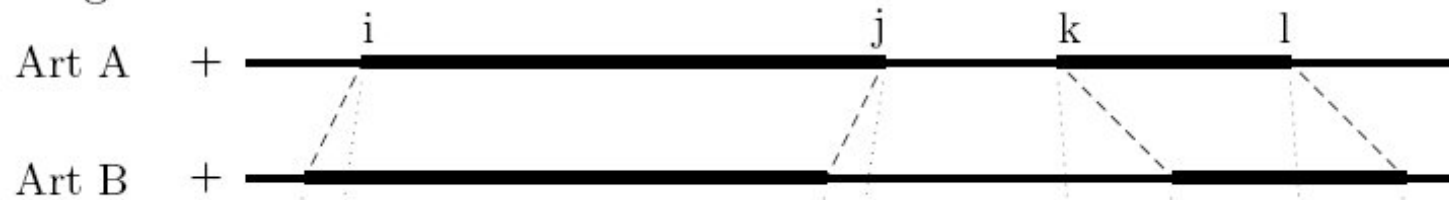


Konserverede DNA-sekvenser

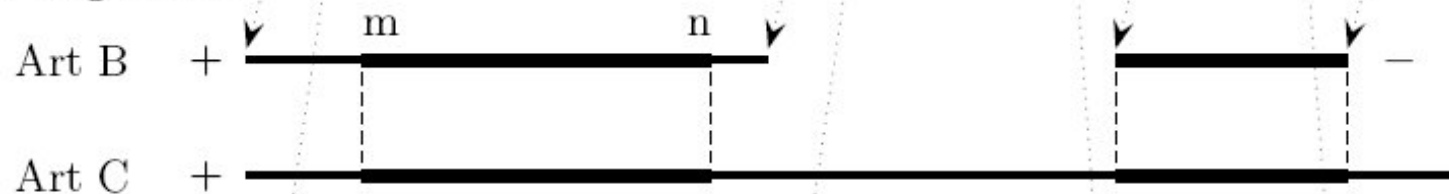


Alignmets samles og justeres

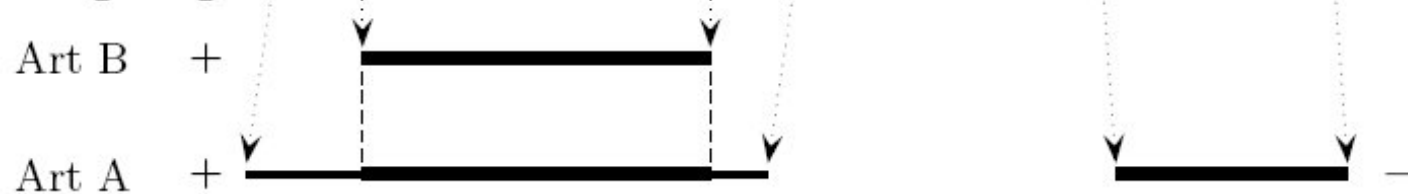
1. alignment



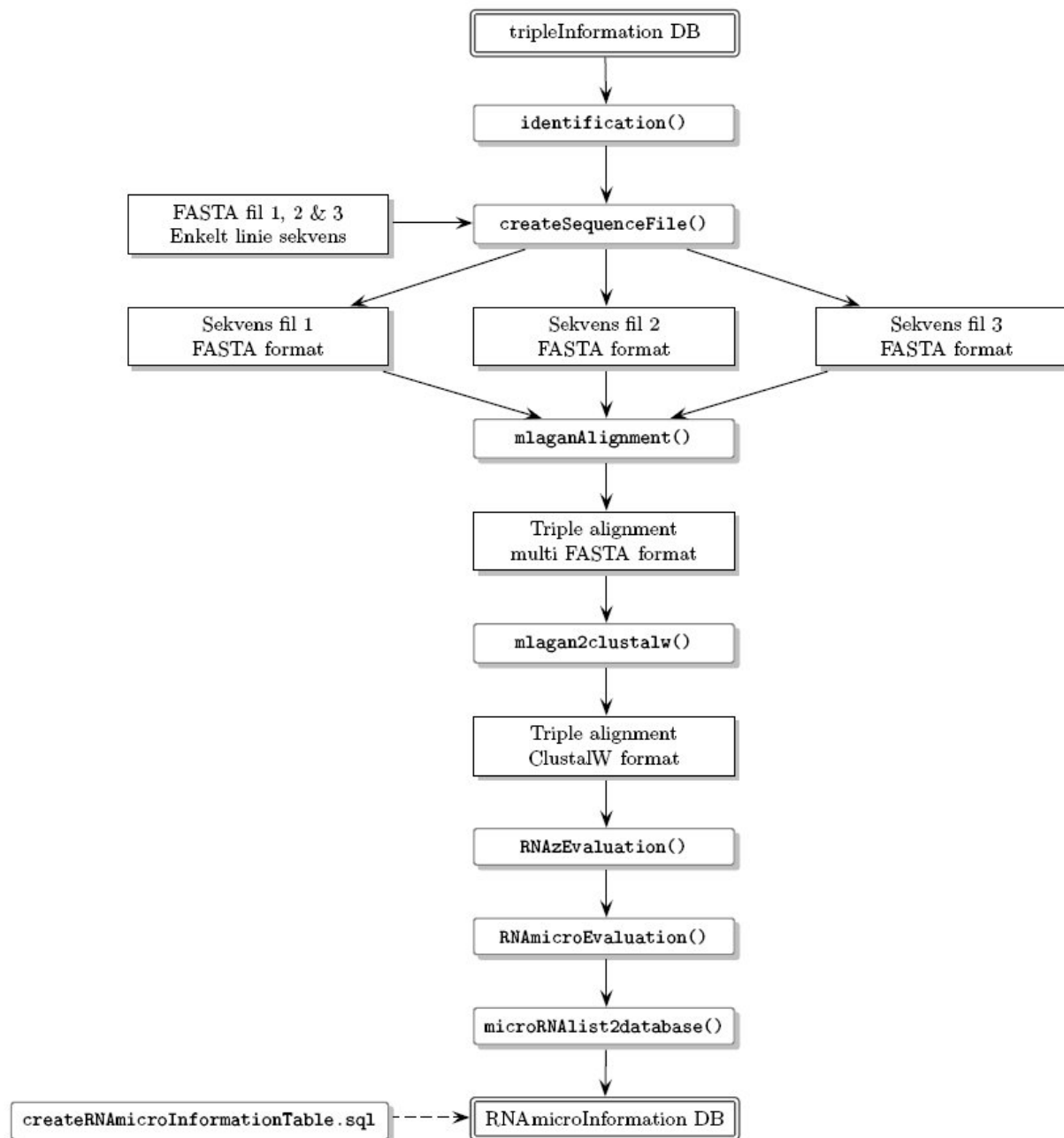
2. alignment



Korrigering



Identifikation



Test – Valg af arter

MiRBase release 9.0

Metazoa

+-- Vertebrata

+-- Mammalia

+-- Primates

+-- Hominidae

+-- Homo sapiens (474) Menneske

+-- Rodentia

+-- Mus musculus (373) Mus

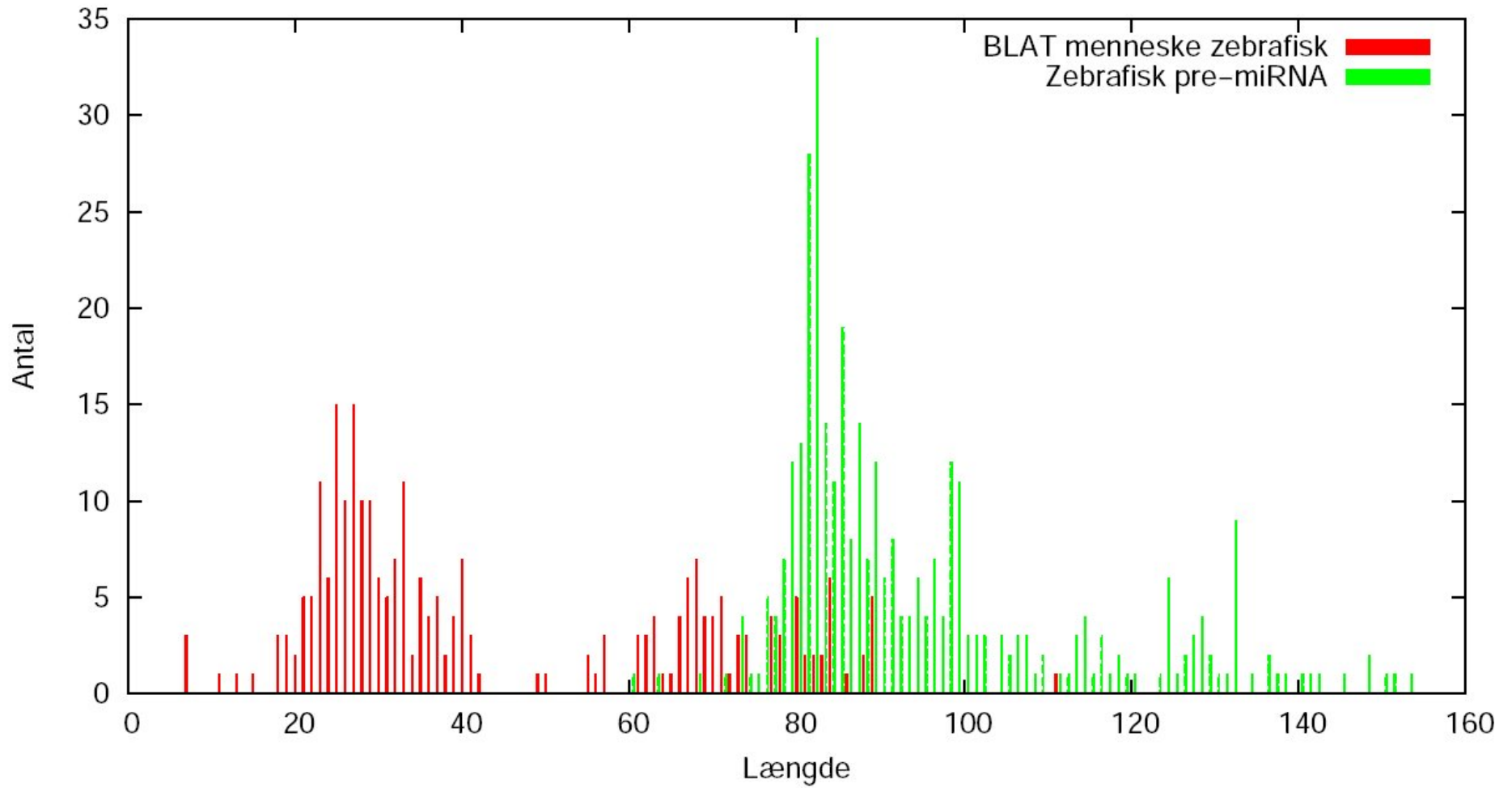
+-- Rattus norvegicus (234) Rotte

+-- Pisces

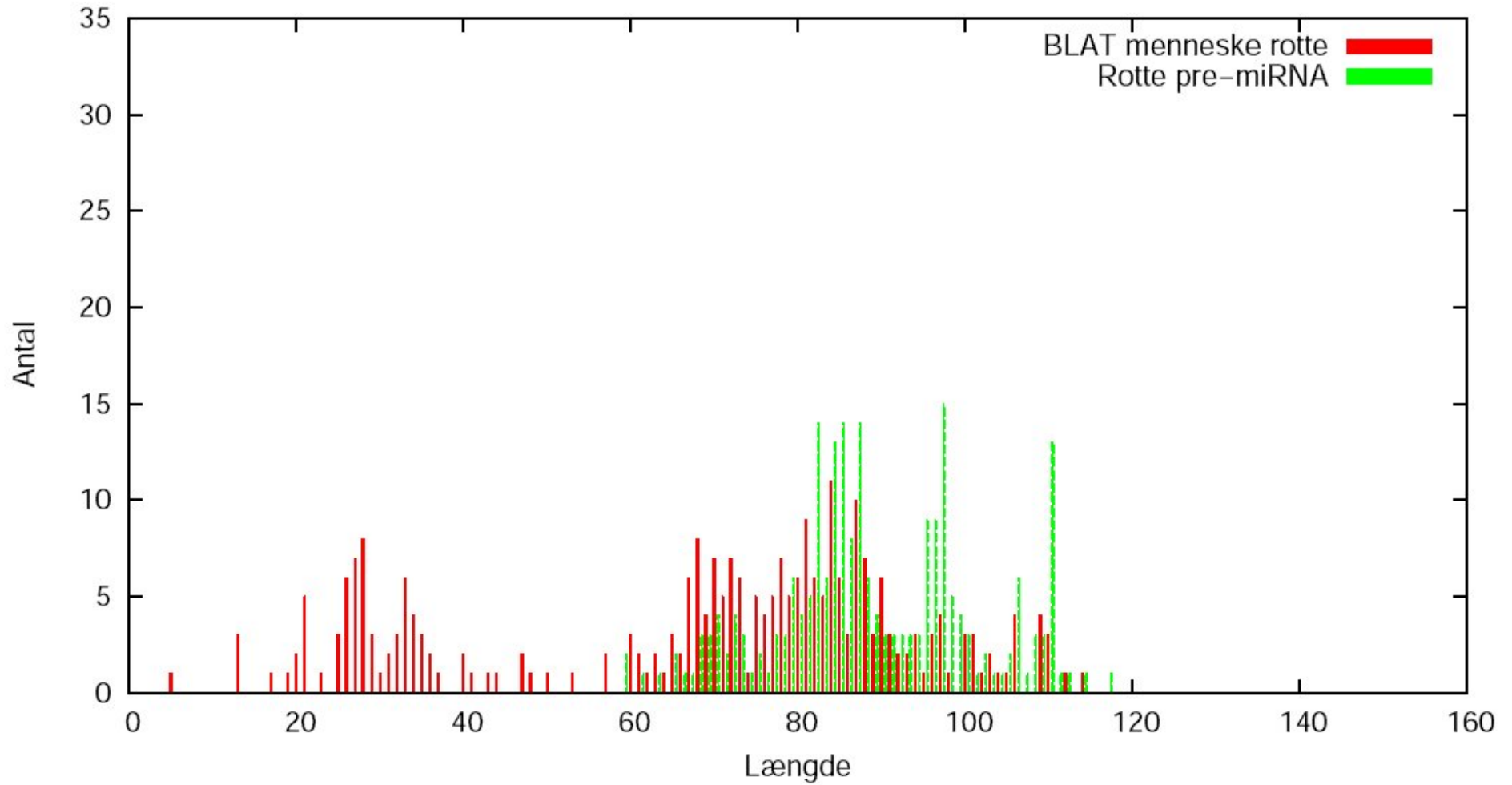
+-- Danio rerio (337) Zebrafisk

+-- Fugu rubripes (131) Fugu

Test – Valg af arter



Test – Valg af arter



Test - Valg af kromosomer

Der er mindst to ting, der skal tages højde for, når der vælges kromosomer:

- Antallet af allerede kendte miRNA gener på et kromosom
- Beslægtethed mellem miRNA generne på kromosomerne der sammenlignes

En tredje ting der kan tages med i valg af kromosomer:

- Længden af kromosomerne

Art	Kromosom	Område	Antal miRNAer
Menneske	hsa 14	99.000.000-104.000.000	46/49
Mus	mmu 12	109.000.000-114.000.000	47/50
Rotte	rno 6	132.000.000 137.000.000	34/35

Test af pipelinen

Det er ikke alle miRNA gener på de tre valgte kromosomer, der kan findes gennem pipelinen:

- De der ligger udenfor testmaterialets område
- De der er blevet maskeret væk
- De der ikke har homologe gener i alle tre arter i testsekvenserne

Test af pipelinen

BLAT -minIdentity=70 -out=axt, 2 pertekt match og seed-længde 11 nt.

MLAGAN -tree "(hsa (mmu rno))"

RNAz vindues størrelse 100 nt. og trin 10 nt.

RNAmicro vindues størrelse 75 nt., 100 nt. og 115 nt. (ma + ok), score $\geq 0,99$

BLATZ -bestScoreOnly -out=axt, seed-længde 11 nt.

Effektiviteten af pipelinen

miRNA	hsa 14	mmu 12	rno 6	Samlet
Alle i test	65,2	66,0	64,7	65,4
Kun umaskerede	81,1	79,5	84,6	81,4
Kun homologe	88,2	94,0	91,2	91,2

Nye miRNAer og miRBase release 11.0

Navn	Kromosom	Start	Slut	Streng	miRBase
hsa-mir-300	hsa 14	100577426	100577554	+/-	11.0
hsa-mir-543	hsa 14	100568044	100568186	+/-	11.0
hsa-mir-665	hsa 14	100411097	100411223	+/-	11.0
rno-mir-380	rno 6	134390959	134391089	+	11.0
rno-mir-410	rno 6	134425051	134425173	+	11.0
rno-mir-411	rno 6	134389314	134389443	+/-	11.0
rno-mir-495	rno 6	134399186	134399329	+/-	11.0
rno-mir-496	rno 6	134420333	134420469	+/-	
rno-mir-665	rno 6	134177271	134177399	+/-	
rno-mir-668	rno 6	134416200	134416313	+	
rno-mir-758	rno 6	134391982	134392081	+	11.0
rno-mir-770	rno 6	134141329	134141471	+/-	11.0

Potentielle miRNA gener eller falske

Navn	Kromosom	Start	Slut	Pipeline test		2. RNAmicro test	
				Streng	Score	Streng	Score
p-4-5	hsa 14	99935758	99935888	+/-	1,00/1,00	+/-	1,00/1,00
	mmu 12	109364008	109364147	+/-	1,00/1,00	+/-	1,00/1,00
	rno 6	133260002	133260139	+/-	1,00/1,00	+/-	1,00/1,00
p-7	hsa 14	100114002	100114094	+	1,00		
	mmu 12	109507158	109507256	+	1,00		
	rno 6	133290583	133290682	-	1,00		
p-21	hsa 14	100401265	100401378	-	1,00	-	1,00
	mmu 12	110024384	110024485	-	1,00	-	1,00
	rno 6	134153984	134154081	-	1,00	-	1,00
p-30-31	hsa 14	100421476	100421589	+/-	0,99/0,99		
	mmu 12	110043801	110043913	+/-	0,99/0,99		
	rno 6	134187664	134187778	+/-	0,99/0,99		
p-32	hsa 14	100436683	100436795	+	1,00		
	mmu 12	110056548	110056658	+	1,00		
	rno 6	134200618	134200729	+	1,00		
p-53	hsa 14	100566116	100566240	+	1,00	+	1,00
	mmu 12	110163485	110163608	+	1,00	+	1,00
	rno 6	134395190	134395313	+	1,00	+	1,00
p-122	hsa 14	100944624	100944738	-	1,00		
	mmu 12	110570781	110570895	-	1,00		
	rno 6	134802594	134802702	-	1,00		
p-123-124	hsa 14	100994123	100994245	+/-	1,00/1,00	+/-	1,00/0,97
	mmu 12	110635914	110636031	+/-	1,00/1,00	+/-	1,00/0,97
	rno 6	134900361	134900478	+/-	1,00/1,00	+/-	1,00/0,97
p-129	hsa 14	101043449	101043572	+	0,99		
	mmu 12	110680854	110680976	+	0,99		
	rno 6	134945251	134945373	+	0,99		
p-132	hsa 14	101676550	101676679	+	1,00	+	1,00
	mmu 12	111186358	111186479	+	1,00	+	1,00
	rno 6	135460822	135460943	+	1,00	+	1,00
p-134	hsa 14	102009751	102009874	+	1,00	+	0,65
	mmu 12	111409605	111409728	+	1,00	+	0,65
	rno 6	135703058	135703181	+	1,00	+	0,65
p-135	hsa 14	102129672	102129760	+	1,00	+	1,00
	mmu 12	111488018	111488106	+	1,00	+	1,00
	rno 6	135784030	135784118	+	1,00	+	1,00
p-141-142	hsa 14	103070209	103070302	+/-	1,00/1,00		
	mmu 12	112130009	112130103	+/-	1,00/1,00		
	rno 6	136464491	136464585	+/-	1,00/1,00		
p-151-152	hsa 14	103705958	103706032	+/-	1,00/0,99		
	mmu 12	112618709	112618774	+/-	1,00/0,99		
	rno 6	136997892	136997957	+/-	1,00/0,99		

Forbedringer til pipelinen

Der hvor de tre sekvenser, der er blevet fundet ved de to BLAT alignments skal samles og placeres i en databasetabel. Sekvensen for den art, der ikke er med i den anden alignment, skal nogle gange forkortes. Den måde det gøres nu med BLATz bevirker, at der ikke altid placeres et resultat i databasetabellen. I stedet for BLATz, skal der udvikles et global alignment program, der anvender dynamisk programmering, og hvor det ikke koster noget at placere gaps i enderne af den korteste sekvens.

Parametrene som BLATz, RNAz og RNAmicro anvender, er hardkodet ind i pipeline programmerne, de burde gives som parametre til pipeline programmerne.

Mere nøjagtig RNAz klassifikation og kun delalignments, der klassificeres positiv af både RNAz og RNAmicro accepteres.

De programmer, der anvender en lokal installation af miRBase databasen, bør ændres, så det bliver lettere at skifte mellem forskellige lokale installationer af miRBase.