Bioinformatic tools for RNAi

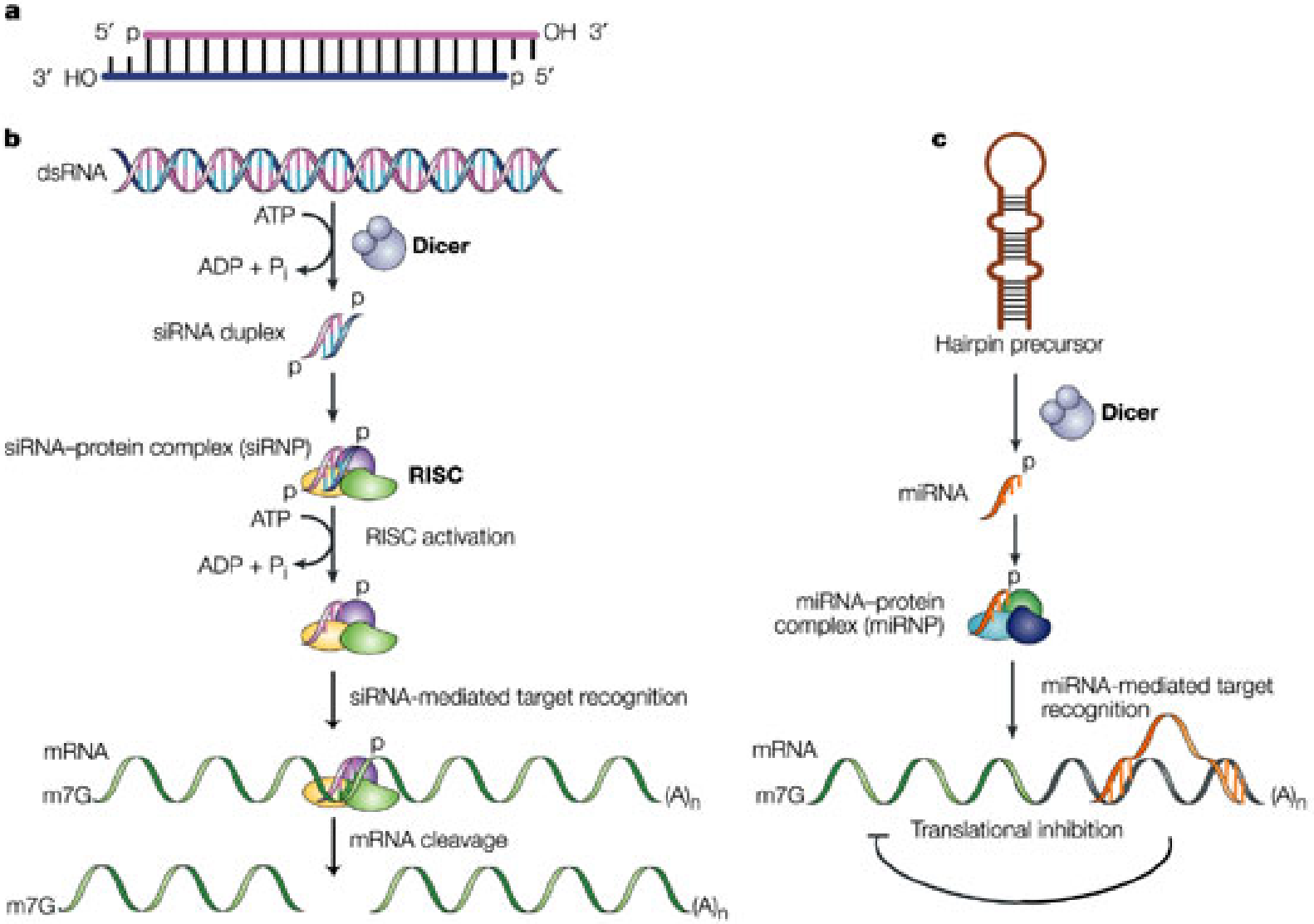
Ulrike Mückstein

Institute for Theoretical Chemistry and Structural Biology
University Vienna
http://www.tbi.univie.ac.at/~ulim/

Bled, Feb. 2004

Possible strategies for locating ncRNA genes in genomic sequences

- ⋆ Sequence similarities
- ⋆ Comparative genomics
- ⋆ Transcription signals
- ⋆ Statistical analysis
- ⋆ blabla


a
5′ p
OH 3′
3′ HO
p 5′
b
dsRNA
ATP
ADP + P$_i$
Dicer
siRNA duplex
siRNA–protein complex (siRNP)
RISC
ATP
ADP + P$_i$
RISC activation
siRNA-mediated target recognition
mRNA
m7G
(A)$_n$
mRNA cleavage
m7G
(A)$_n$
c
Hairpin precursor
Dicer
miRNA
miRNA–protein complex (miRNP)
miRNA-mediated target recognition
mRNA
m7G
(A)$_n$
Translational inhibition

RNA interference

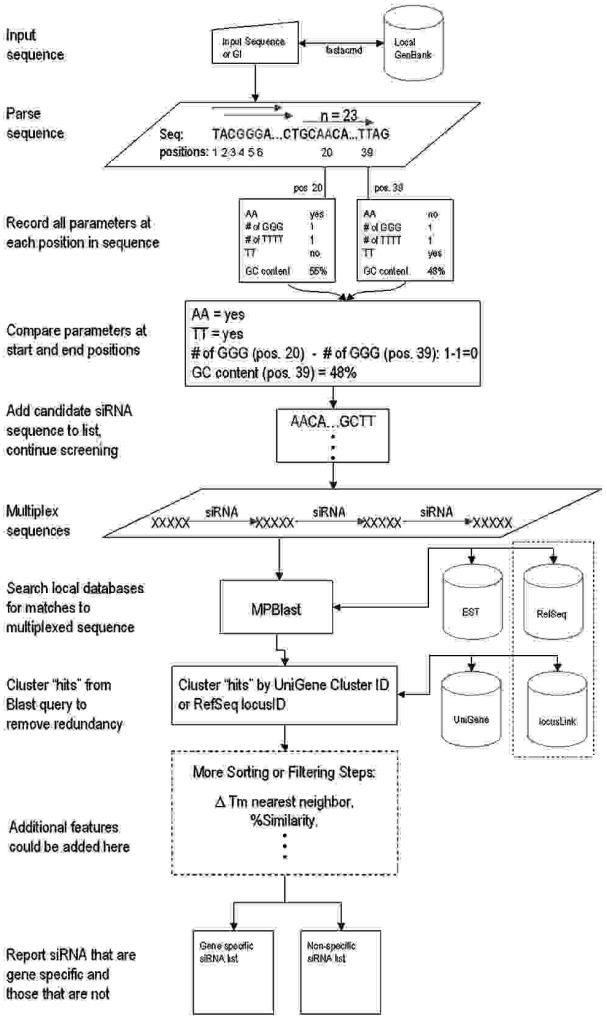
- ⋆ sequence specific, post-transcriptional gene silencing
- ⋆ induced by double-stranded (ds) RNAs
- ⋆ only a small portion of siRNAs designed for any gene is effective.

Empirical rules for siRNA design

category	features	autors
Length	duplexes, length 21 (default) to 23 nt	all
Sequence	G/G content (30% to 70%)	all
	location of mismatches	A
	no homopolymers longer than 3 bases	A, D, Q
	Remove sequences with tandem repeats	G
Overhangs	2-nt symmetric 3' overhang (TT)	T,Q
	nucleotide content of the 3' overhangs	A
Motives	NAR(N17)YNN *	T
	AA(n19)	A, D, Q
Specifity	blast-search against EST libraries	all
Thermodynamics	TM	A, G
Structure	no strong internal secondary structure	D, G

* where R is purine (A, G) and Y is pyrimidine (C, U)
A Ambion, T Tuschl, Q Qiagen, D Dharmacon, G GenScript

Data flow


Input sequence
Input Sequence or GI
fastacmd
Local GenBank
Parse sequence
n = 23
Seq: TACGGGA...CTGCAACA...TTAG
positions: 1 2 3 4 5 6 20 39
pos. 20
pos. 39
Record all parameters at each position in sequence
AA yes
# of GGG 1
# of TTTT 1
TT no
GC content 55%
AA no
# of GGG 1
# of TTTT 1
TT yes
GC content 48%
Compare parameters at start and end positions
AA = yes
TT = yes
# of GGG (pos. 20) - # of GGG (pos. 39): 1-1=0
GC content (pos. 39) = 48%
Add candidate siRNA sequence to list, continue screening
AACA...GCTT
Multiplex sequences
XXXXX siRNA XXXXX siRNA XXXXX siRNA XXXXX
Search local databases for matches to multiplexed sequence
MPBlast
EST
RefSeq
Cluster "hits" from Blast query to remove redundancy
Cluster "hits" by UniGene Cluster ID or RefSeq locusID
UniGene
locusLink
Additional features could be added here
More Sorting or Filtering Steps:
Δ Tm nearest neighbor,
%Similarity,
Report siRNA that are gene specific and those that are not
Gene specific siRNA list
Non-specific siRNA list

Characteristics associated with siRNA functionallity

feature	autors	
low G/C content	1 (30-52%),	
bias towards low internal stability at 3'end	1,	sense strand
lack of inverted reats in siRNA	1,	
sense strand base preferences (pos.3,10,13,19)	1,	

base preferences:

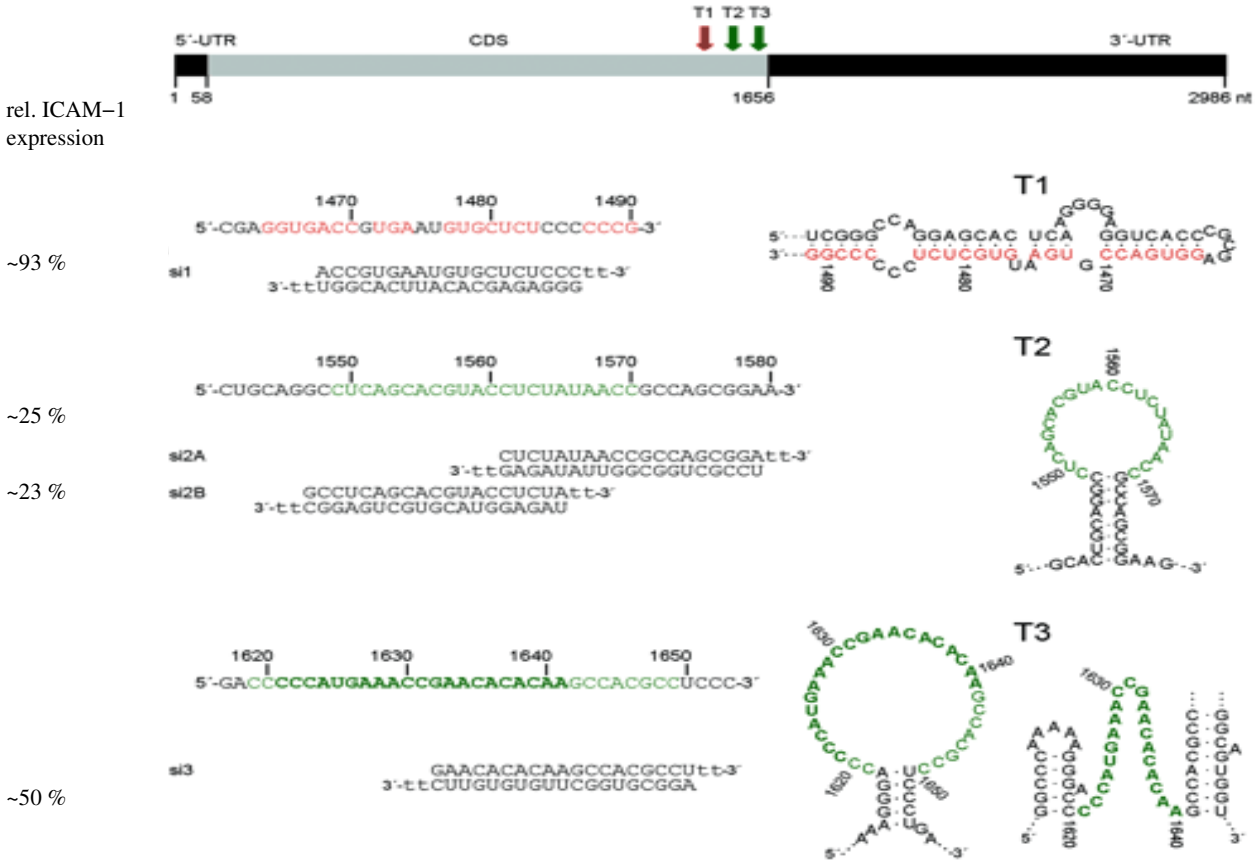
pos.3 is A:

pos.10 is U: RISC cleaves mRNA between position 10 and 11 of relative to the 5' end of the complementary target strand, most endonucleases prefer to cleave 3' of U

pos.13 is no G:

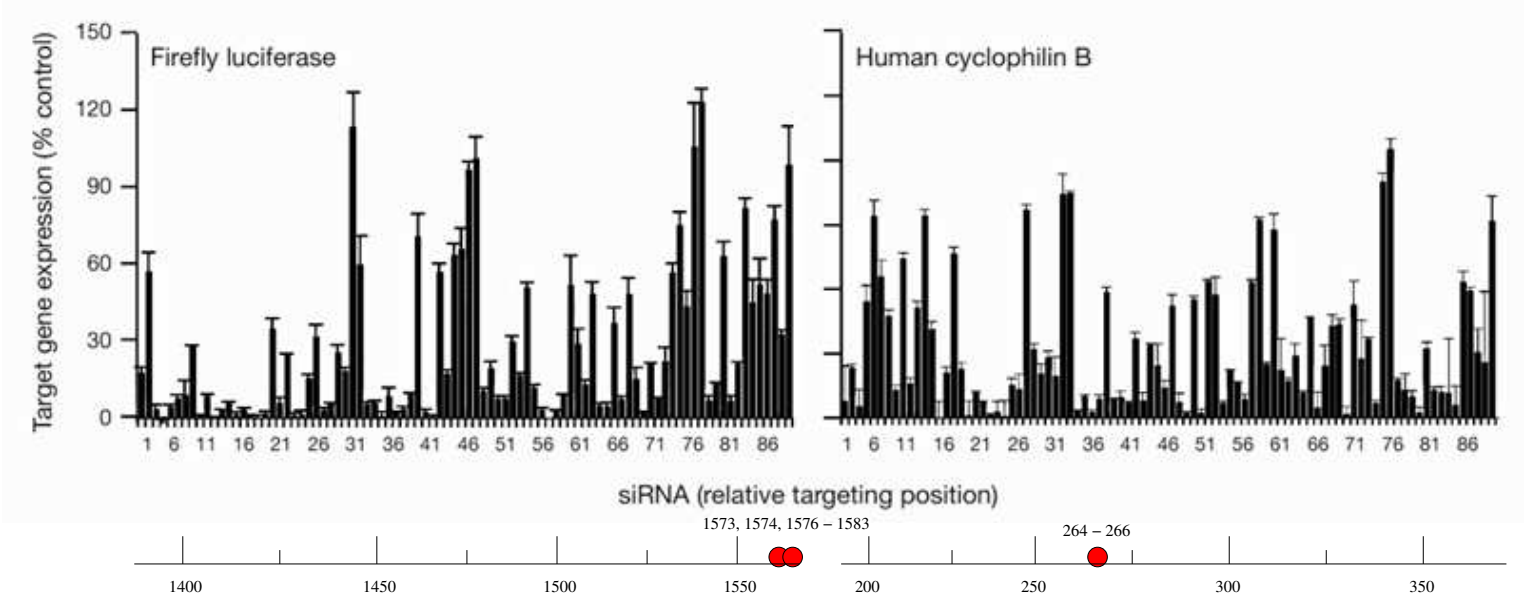
pos.19 no GC, is A:low internal stability of the sense 3' end, promote strand selection and entry into RISC. 75% of miRNA precursors contain U at pos.1 (corresponding to A in pos.19 of siRNA sense strand

Target mRNA secondary structure and accessibility



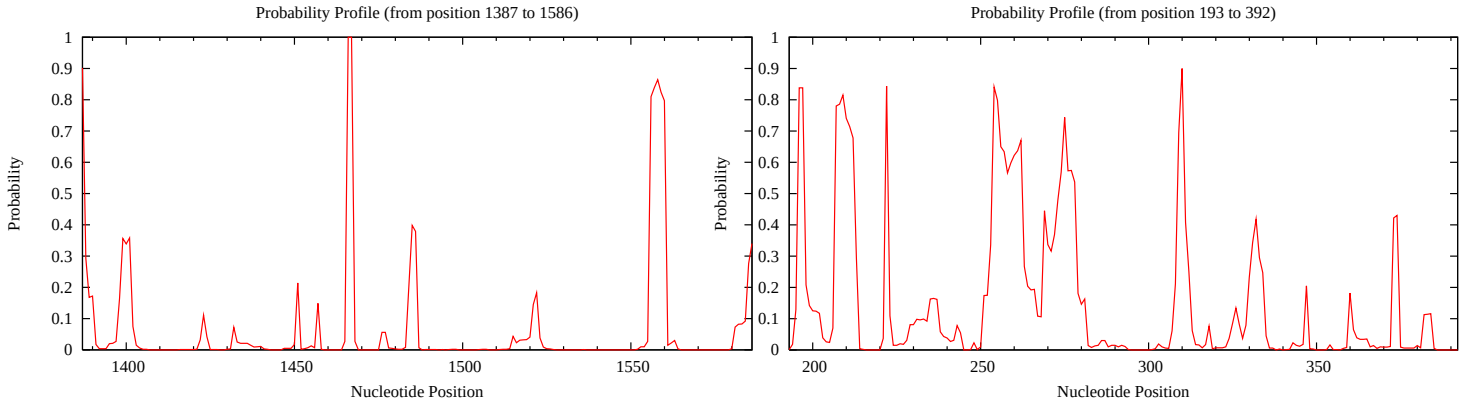
U47298.2_48_to_2185 1434–1631
absolute position: 1387 – 1583

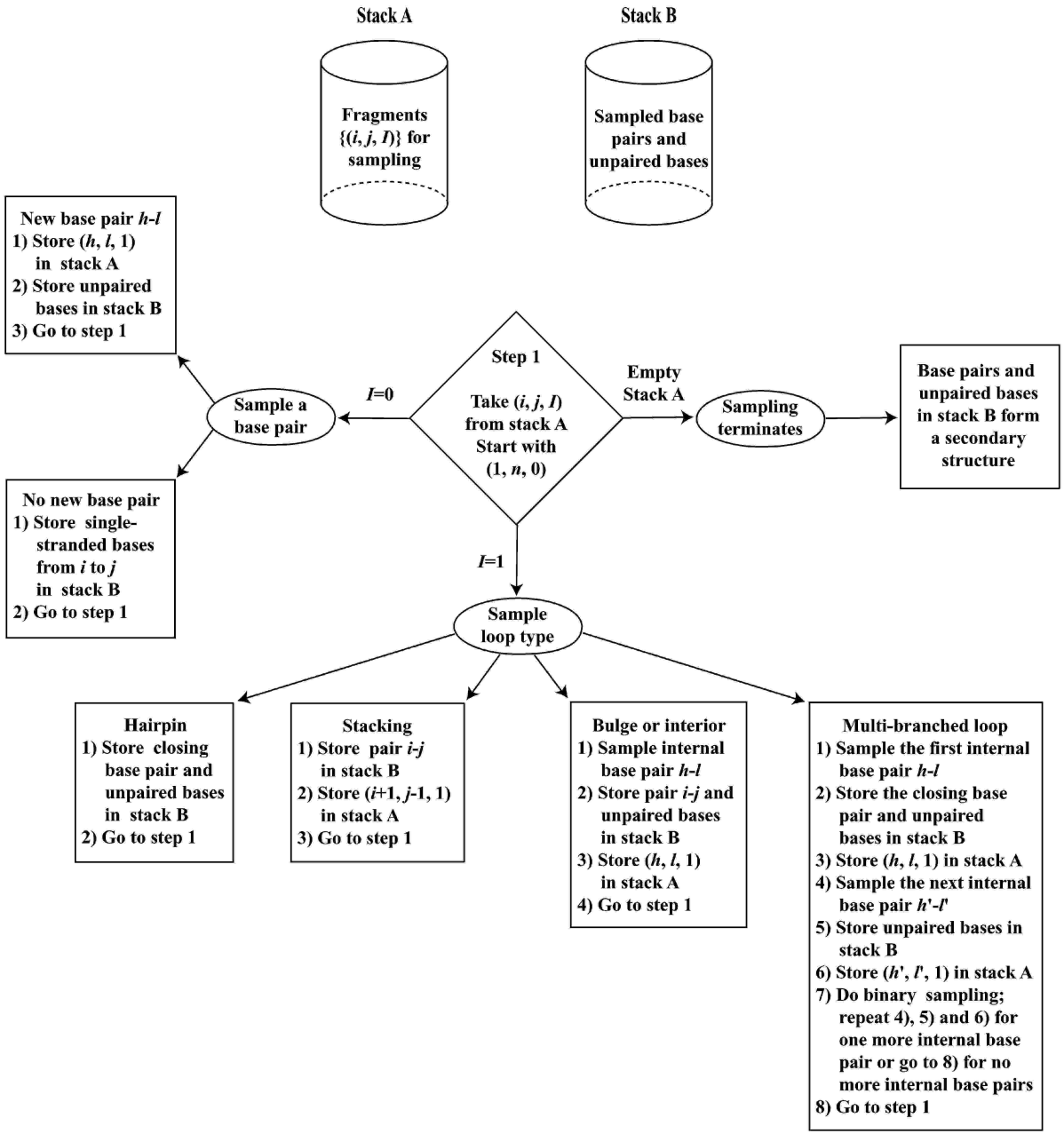
M60857 193–390

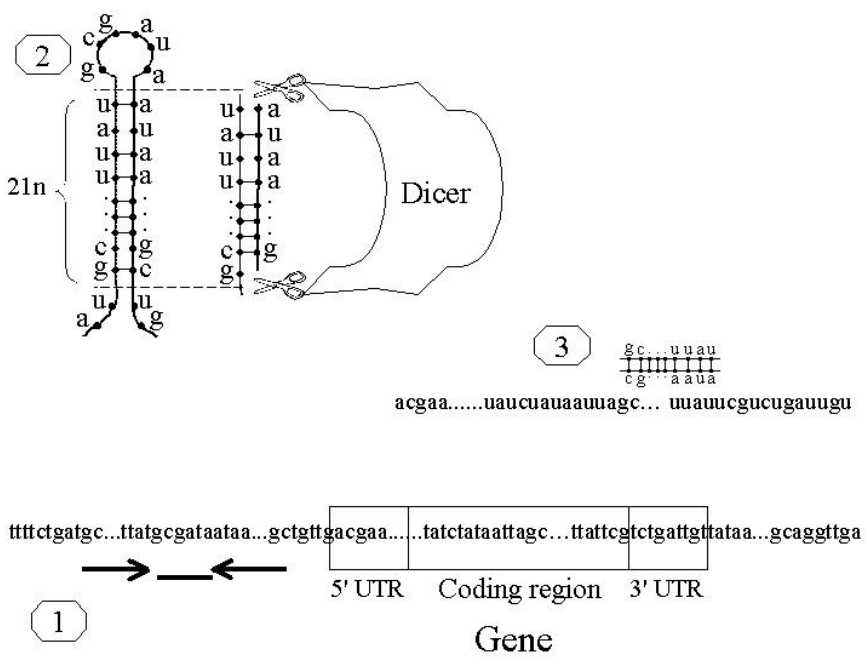


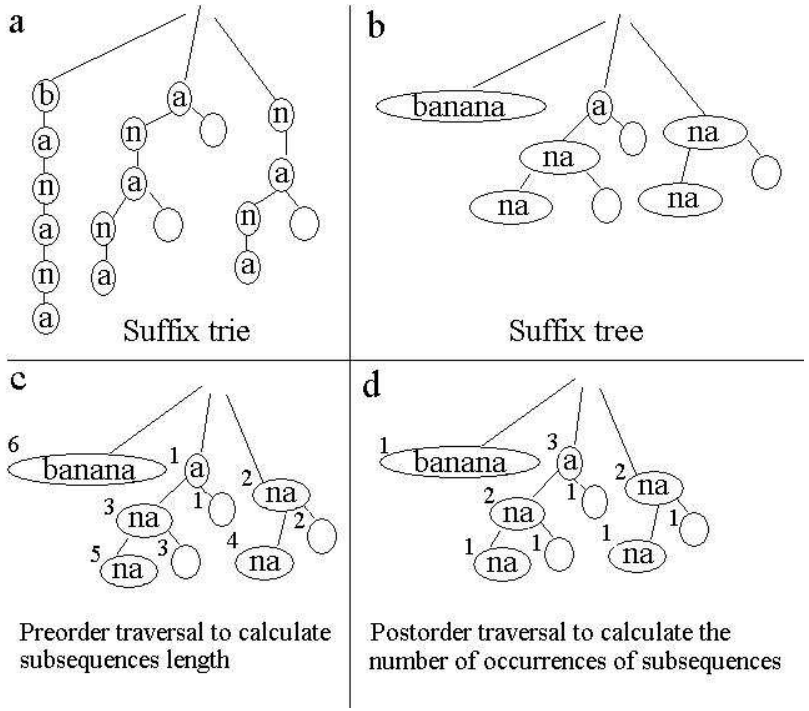
Copyright © 2003 Wadsworth Bioinformatics Center

Copyright © 2003 Wadsworth Bioinformatics Center




Stack A
Stack B
Fragments {(i, j, I)} for sampling
Sampled base pairs and unpaired bases
New base pair h-l
1) Store (h, l, 1) in stack A
2) Store unpaired bases in stack B
3) Go to step 1
Step 1
Take (i, j, I) from stack A
Start with (1, n, 0)
I=0
Empty Stack A
Sample a base pair
Sampling terminates
Base pairs and unpaired bases in stack B form a secondary structure
No new base pair
1) Store single-stranded bases from i to j in stack B
2) Go to step 1
I=1
Sample loop type
Hairpin
1) Store closing base pair and unpaired bases in stack B
2) Go to step 1
Stacking
1) Store pair i-j in stack B
2) Store (i+1, j-1, 1) in stack A
3) Go to step 1
Bulge or interior
1) Sample internal base pair h-l
2) Store pair i-j and unpaired bases in stack B
3) Store (h, l, 1) in stack A
4) Go to step 1
Multi-branched loop
1) Sample the first internal base pair h-l
2) Store the closing base pair and unpaired bases in stack B
3) Store (h, l, 1) in stack A
4) Sample the next internal base pair h'-l'
5) Store unpaired bases in stack B
6) Store (h', l', 1) in stack A
7) Do binary sampling; repeat 4), 5) and 6) for one more internal base pair or go to 8) for no more internal base pairs
8) Go to step 1


2
21n
Dicer
3
acgaa.....uaucuauaauuagc… uuauucgucugauugu
ttttctgatgc...ttatgcgataataa...gctgttgacgaa.....tatctataattagc…ttattcgtctgattgttataa...gcaggttga
5' UTR
Coding region
3' UTR
1
Gene


a
Suffix trie
b
banana
Suffix tree
c
Preorder traversal to calculate subsequences length
d
Postorder traversal to calculate the number of occurrences of subsequences