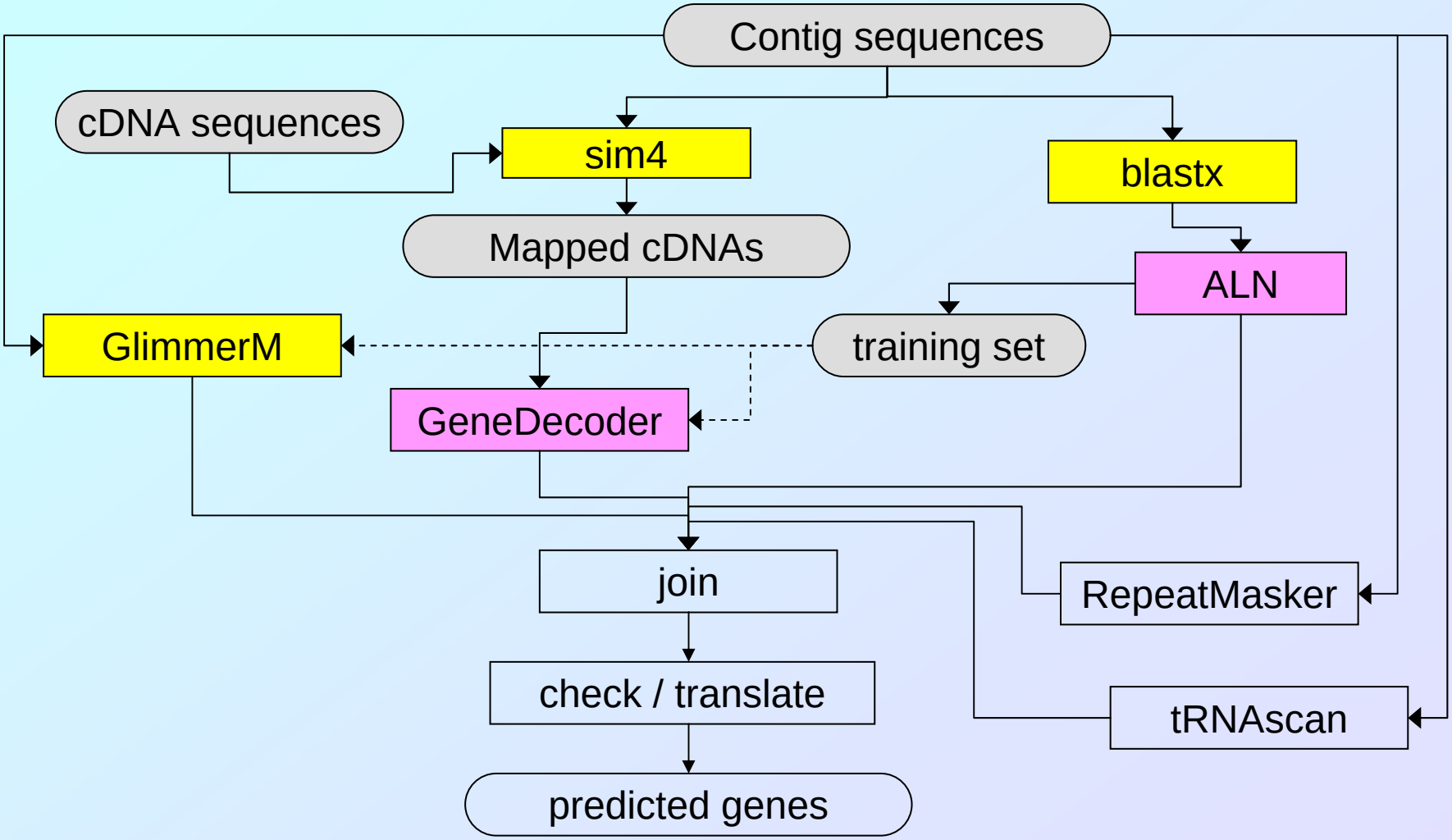
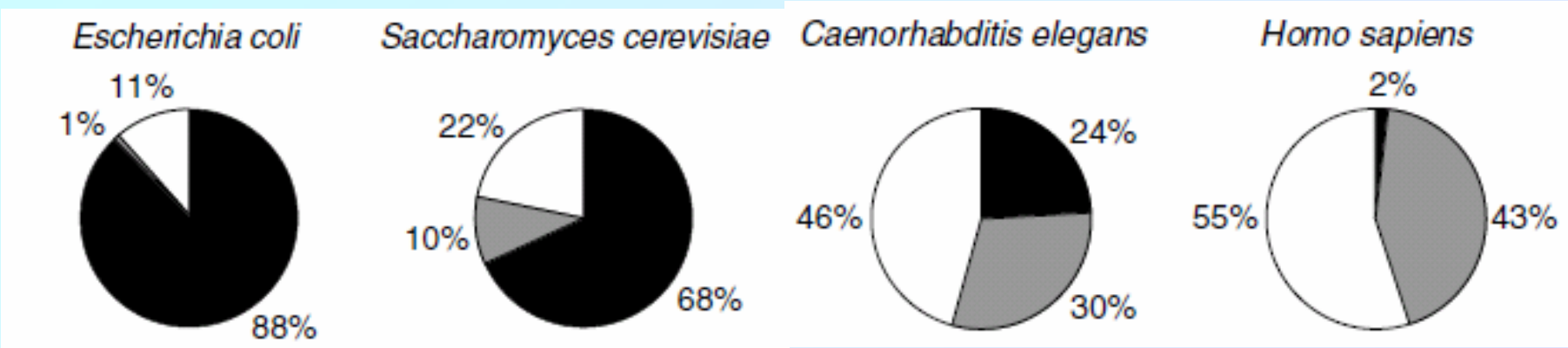
RNA配列の比較アルゴリズム

１．東京大学大学院新領域創成科学研究科
２．産業技術総合研究所生命情報科学研究センター
浅井　潔

麹菌
Aspergillus oryzae遺伝子発見パイプライン



転写産物(RNA)の大部分は非コード領域 !!



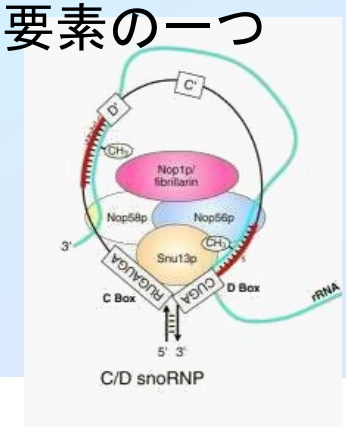
Genome ***Biology*** 2004, **5:**105

様々なタイプの機能性非コードRNAがある

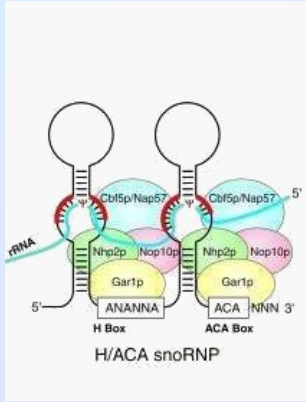
機能性RNAもゲノム上にコードされている
miRNAの発見、RNAi（RNA干渉）の開発
ncRNAは、生体内のネットワークの重要な要素の一つ

miRNA

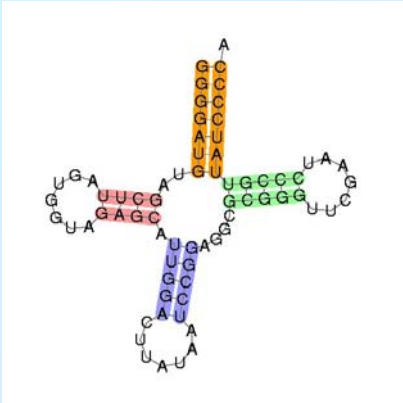
siRNA

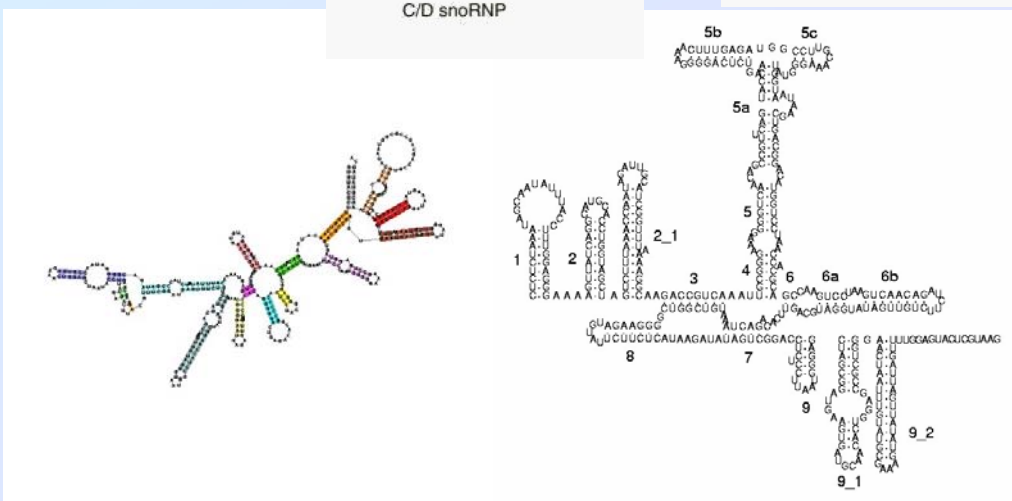


C/D snoRNP

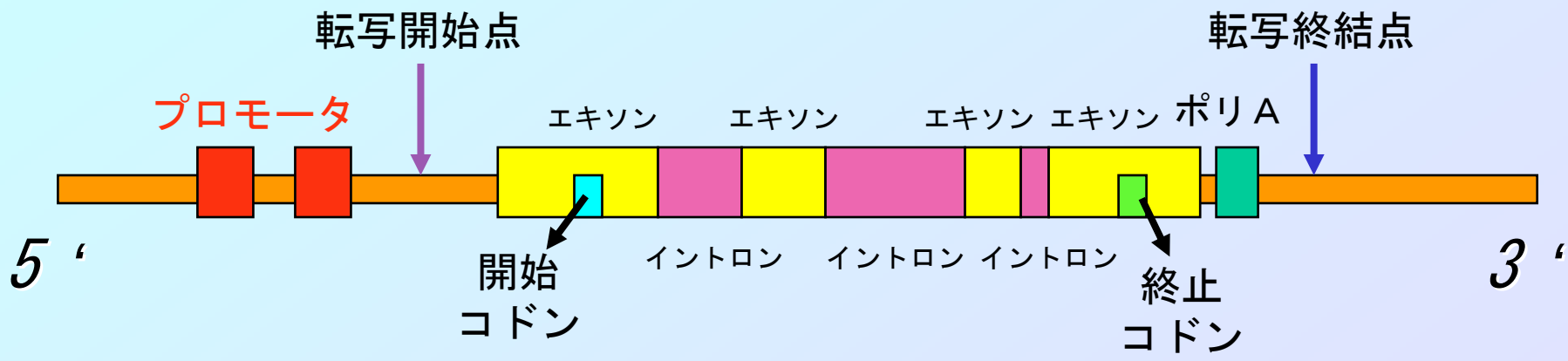


H/ACA snoRNP





遺伝子の構造(真核生物)


転写開始点
転写終結点
プロモータ
エキソン
エキソン
エキソン
エキソン
ポリA
5‘
開始
コドン
イントロン
イントロン
イントロン
終止
コドン
3‘

タンパク質コード遺伝子の発見法

- 既知転写産物との類似性
  - DB上の既知遺伝子（DNA）で相同性検索
  - DB上の既知タンパク質で相同性検索
  - cDNAをゲノムDNA配列に「貼り付ける」
- 比較ゲノム
  - 保存領域（synteny）は「重要」
- “ab initio” 遺伝子発見
  - 統計的な情報を用いた遺伝子の確率モデル

タンパク質コード遺伝子の発見法

- 既知転写産物との類似性

  アラインメントによる配列の比較
  動的計画法（DP）
  BLAST

- 比較ゲノム

- “ab initio” 遺伝子発見
  - 統計的な情報を用いた遺伝子の確率モデル

  モデルによるゲノムDNAの「局所検索」

非コードRNA遺伝子の発見法

- 既知転写産物との類似性

アラインメントによる配列の比較
動的計画法（DP）
BLAST？？？？

- 比較ゲノム
  - きわめて重要(QRNA, RNAz)
- “ab initio” 遺伝子発見
  - 統計的な情報を用いた遺伝子の確率モデル

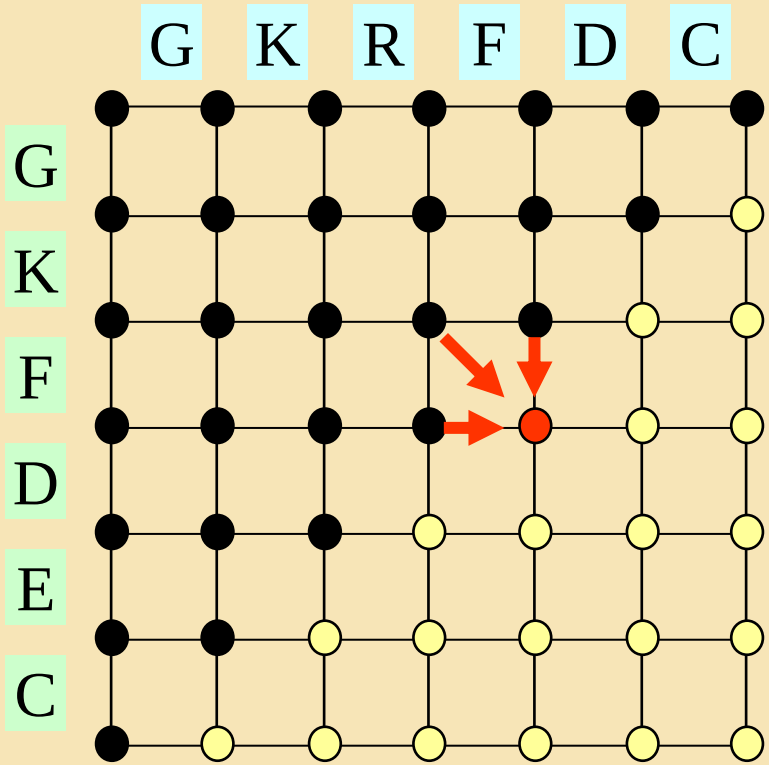
However, is it OK to define the “conserved region” by sequence similarities ?

モデルによるゲノムDNAの「局所検索」

RNA配列情報解析の課題

- ゲノム配列からの既知ncRNA発見
  - tRNAscan-SE, Snoscan, Infernal
- cDNA配列からのncRNA発見
  - クラスタリング、二次構造予測（xxxfold）
- 比較ゲノム保存領域からのncRNA発見
  - 保存領域抽出、ncRNA判定（QRNA, RNAz）
- 特定モチーフを持つncRNA発見
  - 二次構造モチーフと特定配列。標的予測

配列の比較:2次元DP

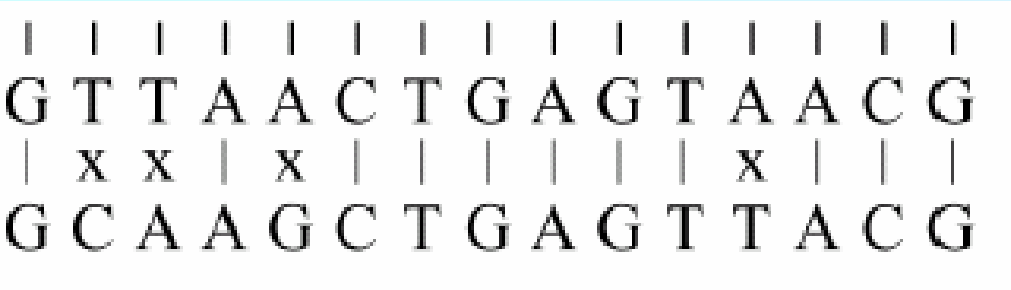


$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i, j-1) - d \\ F(i-1, j) - d \end{cases}$$

メモリ　$O(L^2)$

計算時間　$O(L^2)$

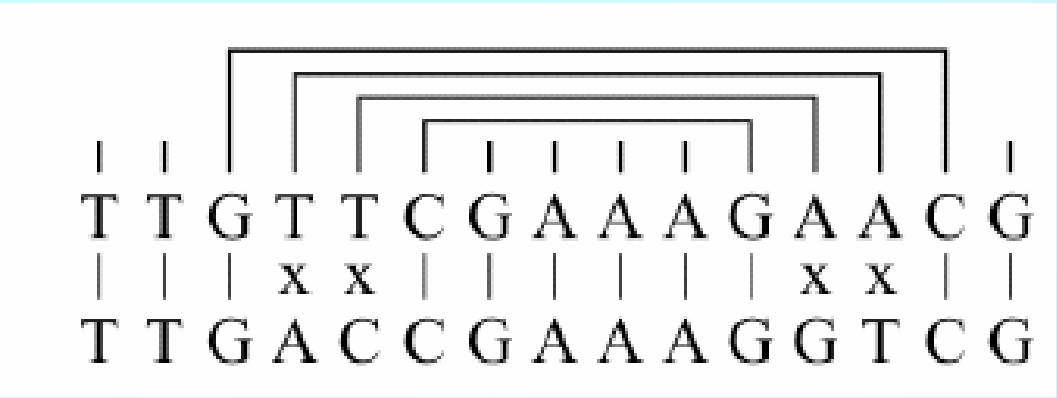
通常のアライメント法

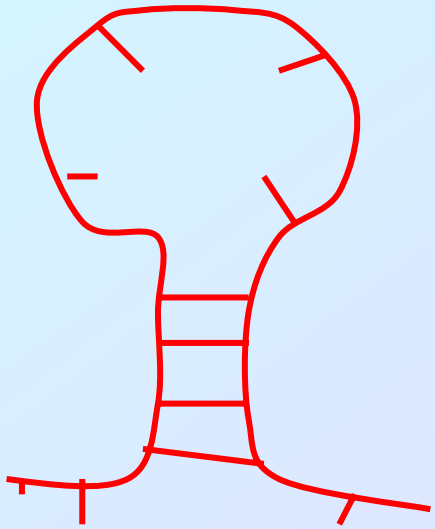


	A	C	G	T
A	1.1	-2.7	-1.2	-2.8
C	-2.0	1.4	-2.1	-1.1
G	-1.1	-2.1	1.4	-2.1
T	-2.7	-1.2	-2.6	1.1

- 塩基配列の各位置が、進化の過程で、それぞれ独立に他の塩基への置換を起こすというモデルに基づいている。
- ClustalWなど

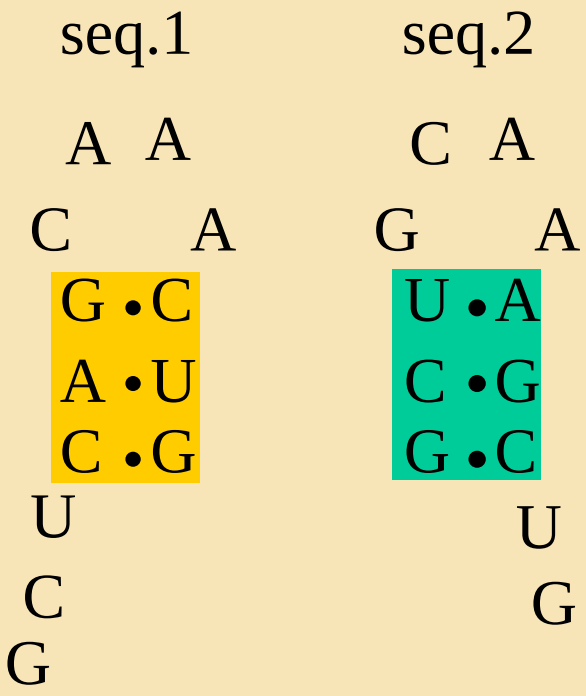
二次構造をもつRNAの進化





- 塩基対をつくる塩基ペアは、塩基対を保つような塩基置換の仕方（共置換）をする。
- →離れた2点での塩基の進化は独立でない。
- 配列の相同性が低くなってくると従来のアライメント法はうまくいかなくなってくる。

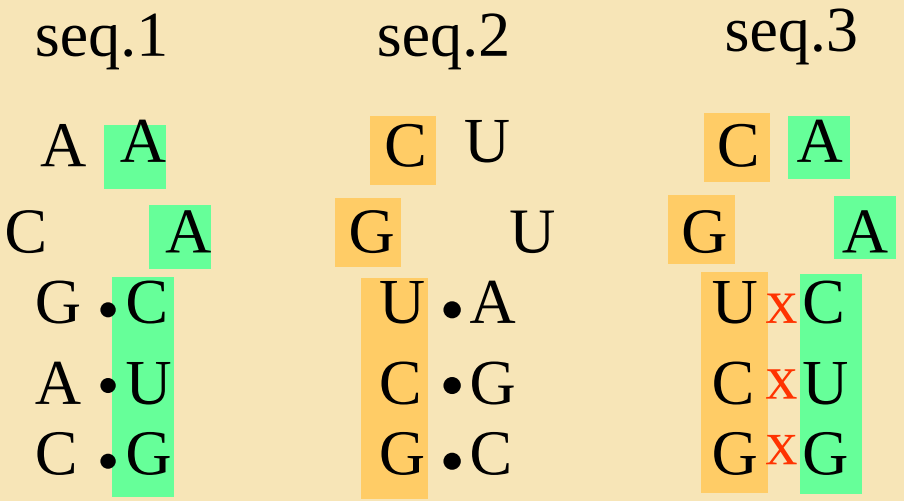
二次構造を考慮したアラインメント

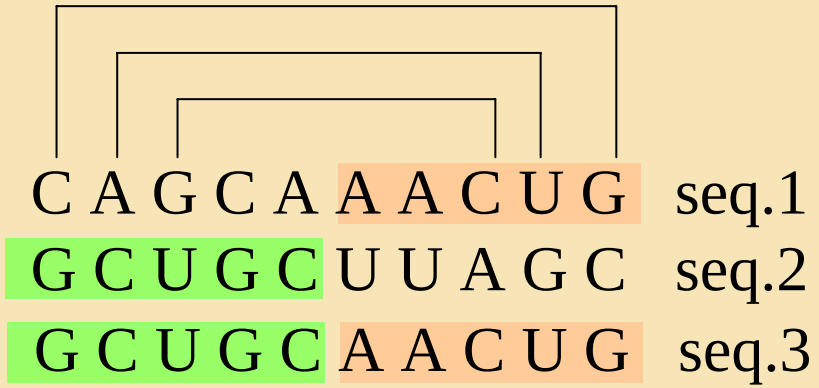


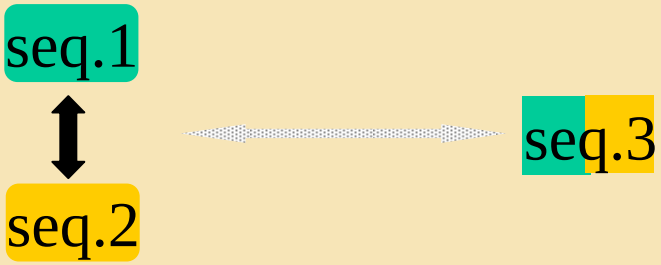
G C U C A G C A A A – C U G seq.1
G C U – – G C A A A G C U G seq.2

G C U C A G C A A A C U G seq.1
G C U G C U U A G C U G seq.2

二次構造を考慮したスコア







あらゆるRNA配列情報解析の基礎

RNA配列の比較

- 配列の文字列としての類似性
  - ステム部分の塩基対類似性
  - 非塩基対部分の配列類似性

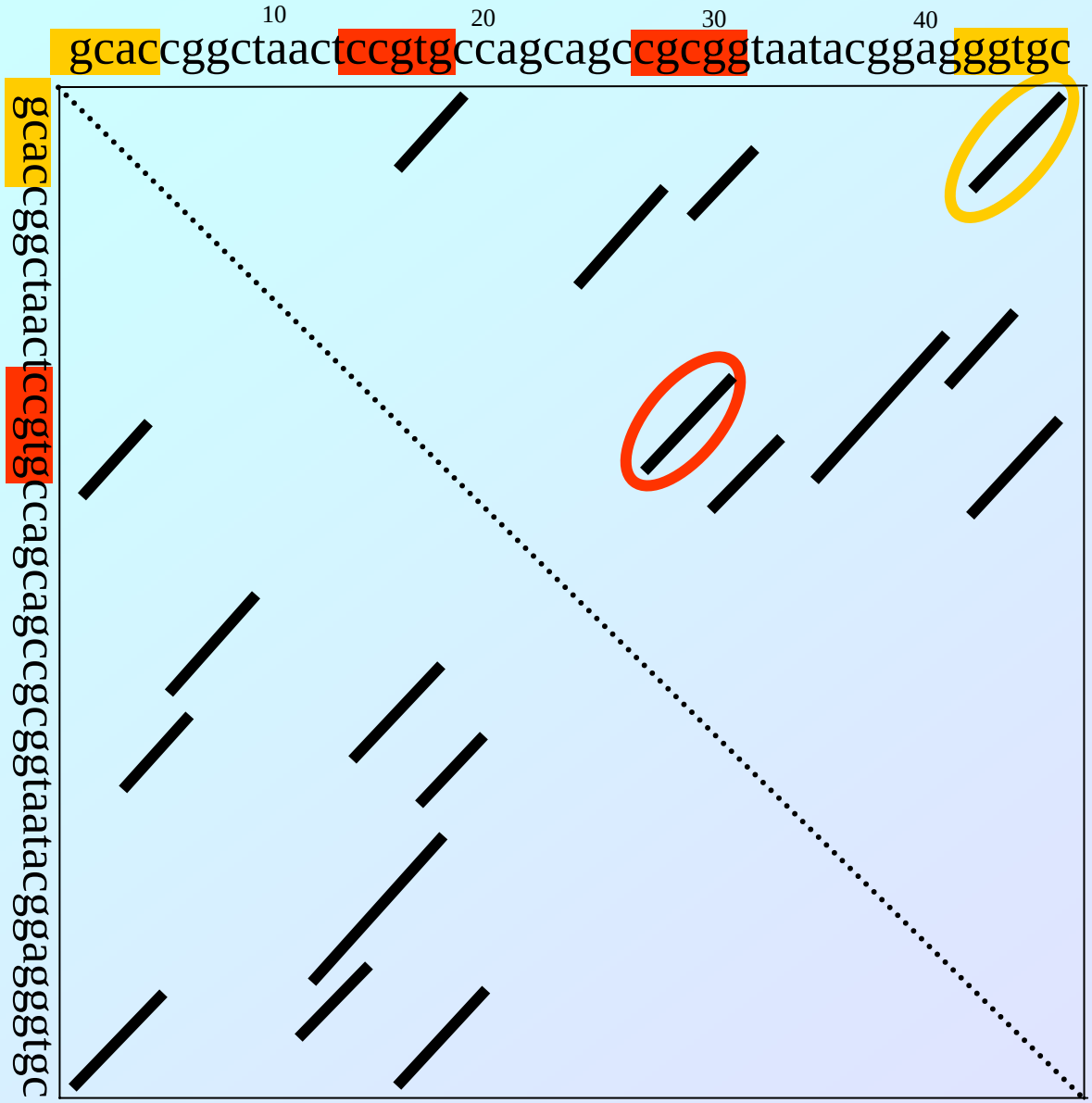
- RNAの二次構造の類似性
  - 二次構造が異なる　　→　どう比較するのか？
  - 二次構造が分からない→　どう比較するのか？

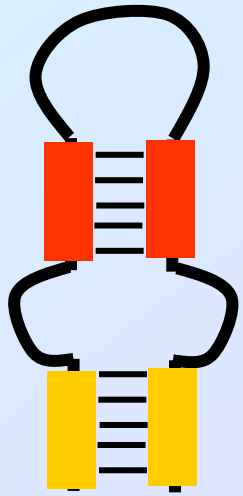
RNAの２次構造予測

- 自由エネルギー最小化（含：準最適構造）
  - Nussinov & Jacobson (“Nussinov Algorithm, 1980)
  - Mathews, Turner & Zuker (MFOLD, 2000)

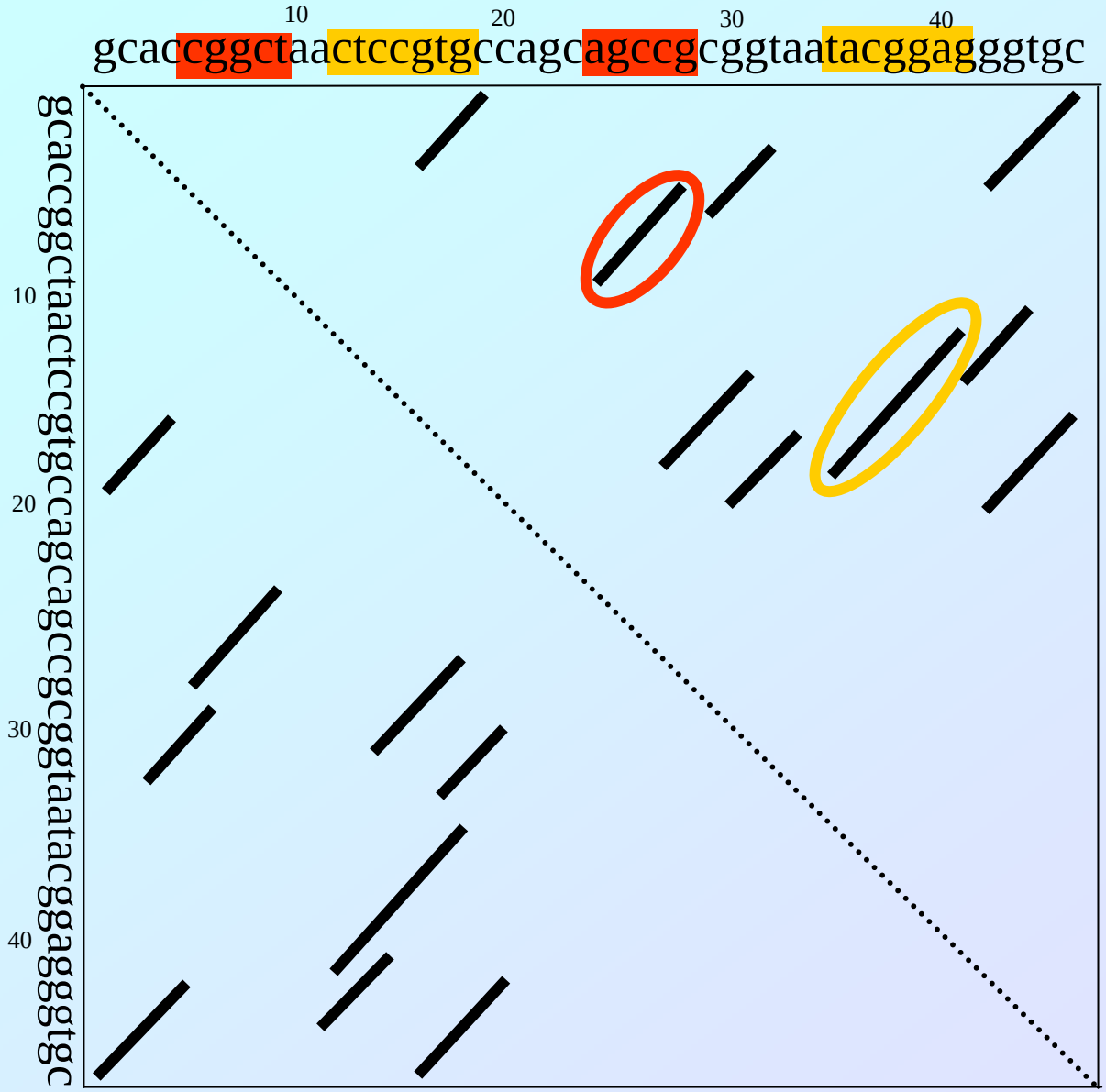
- 配列比較解析
  - 相同RNA配列群が必要、結果が初期整列に依存
  - Chiu and Kolodziejczak (1991)
  - Eddy and Durbin (CM=プロファイルSCFG, 1994)

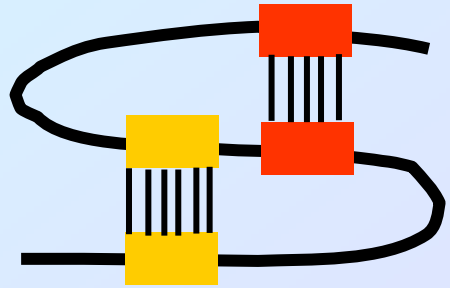
Secondary Structure Prediction of RNA



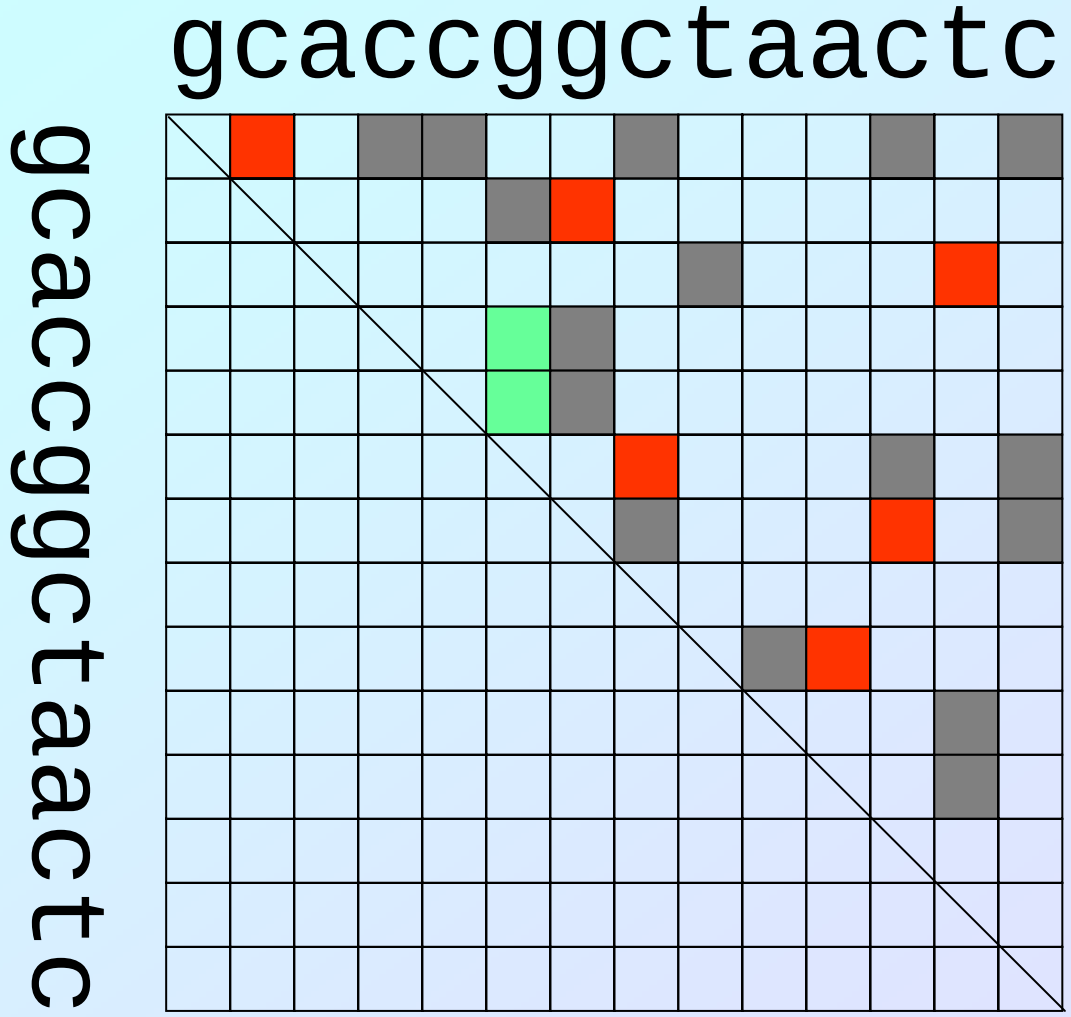


Secondary Structure Prediction of RNA



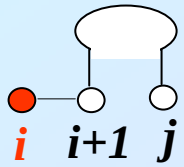


Secondary Structure Prediction of RNA

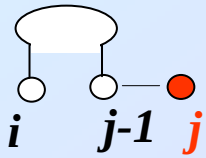


Nussinov アルゴリズム

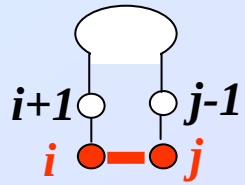
- $M(i,j)$: maximum # of base pairs in $(x_i, \ldots x_j)$
  - $M(i,j) = M(i+1, j)$



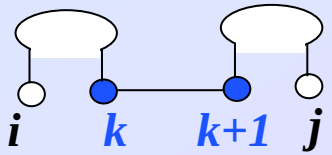
  - $M(i,j) = M(i, j-1)$



  - $M(i,j) = M(i+1, j-1) + 1$



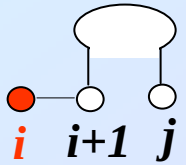
  - $M(i,j) = \max_k (M(i,k)+M(k+1,j))$



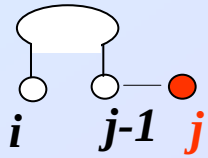
確率文脈自由文法（SCFG）による RNA二次構造予測

Σ = {a, c, g, u}, (x,y = a|c|g|u)

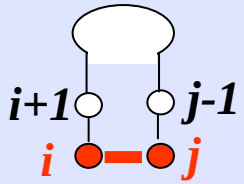
S = xS $M(i,j) = M(i+1, j)$



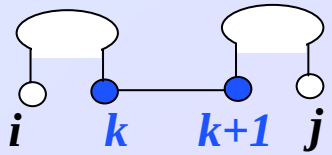
S = Sx $M(i,j) = M(i, j-1)$



S = xSy $M(i,j) = M(i+1, j-1) + 1$



S = SS $M(i,j) = \max_k (M(i,k)+M(k+1,j))$



SCFGのCYK アルゴリズム

$\gamma(i,j)$: maxmum probability of
(xi, … xj) is parsed by S

S = xS

S = Sx

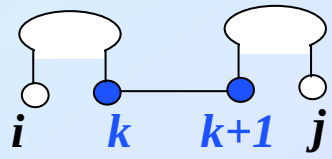
S = xSy

S = SS

$$\gamma(i,j) = \max \begin{cases} \gamma(i+1,j) + \log p(x_i S); \\ \gamma(i,j-1) + \log p(S x_j); \\ \gamma(i+1,j-1) + \log p(x_i S x_j); \\ \max_{i<k<j}\{\gamma(i,k) + \gamma(k+1,j) + \log p(SS)\} \end{cases}$$

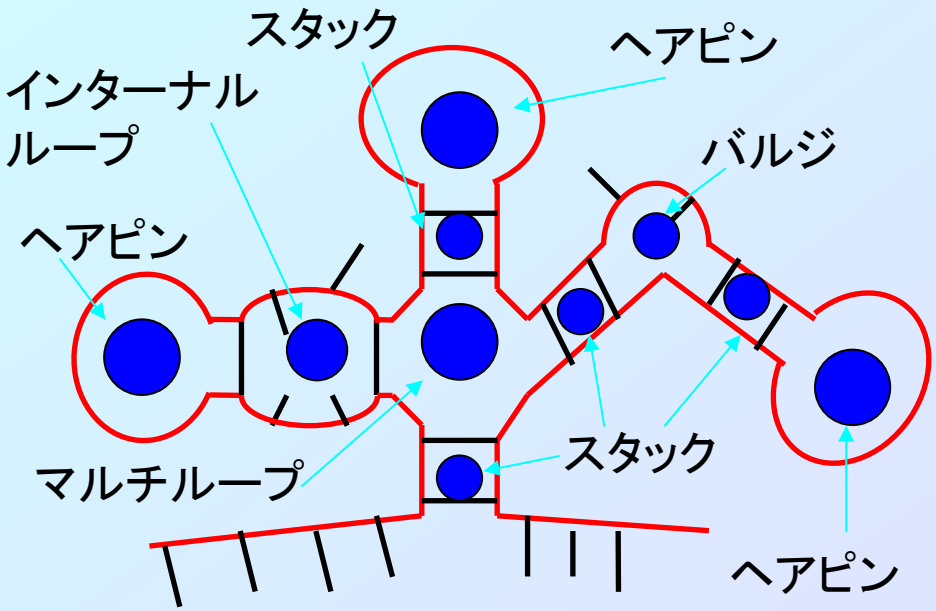
RNA二次構造予測の計算複雑度

- $M(i,j) = \max_k (M(i,k)+M(k+1,j))$



  - メモリ: $M(i,j)$ for all position $i,j = \boldsymbol{L^2}$
  - 計算時間: Iterate for all $i,j,k = \boldsymbol{L^3}$

Turnerのエネルギーパラメーター

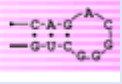


RNA配列比較の課題

- ２本配列の比較
  - 共通二次構造の推定とアラインメント
  - 局所アラインメント（検索）
- 配列群のマルチプルアラインメント
- 配列群の共通二次構造予測
- 二次構造付配列への構造アラインメント
  - 局所アラインメント（検索）
- 二次構造への構造アラインメント
  - 局所アラインメント（二次構造モチーフ検索）
- 配列クラスタリング
- 特定RNAのモデル
  - 局所アラインメント（特定RNA発見）
- 比較ゲノム保存領域からの機能性RNA発見

SCARNA

Marlet

PHMMTS 

PSTAG 

共通2次構造予測

- Sankoff (1985)
  - 2(N)本のRNAの折畳み。$O(L^{3N})$ time, $O(L^{2N})$space
- Gorodkin, Heyer & Stormo (FOLDALIGN, 1997)
  - 2本局所整列、分岐構造なし、塩基対の最大化、$O(L^4)$ time
  - 多重整列: トーナメント法＋ Greedy アルゴリズム
- Mathews & Turner (Dynalign, 2002)
  - 自由エネルギー最小化＋配列比較解析
  - ステム間距離をM以下 $O(M^3 L^3)$ time
- Perriquet, Touzet & Dauchet (CARNAC, 2003)
  - 局所配列類似性 ＋ エネルギー ＋ 塩基対共変

2本のRNA配列からの共通二次構造の計算

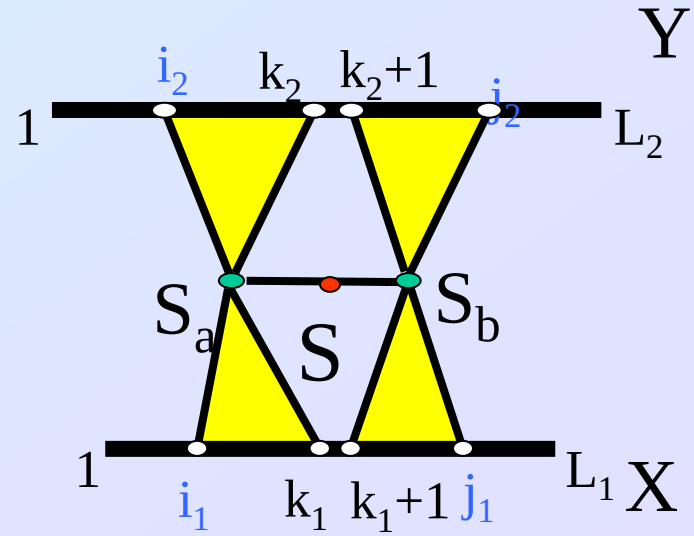
$\gamma(i_1,j_1,i_2,j_2)$: 4 次元のDP行列

(# of non-terminal is constant)

$$\gamma(i_1,j_1,i_2,j_2) = \max_{k_1,k_2} (\gamma(i_1,k_1,i_2,k_2) + \gamma(k_1+1,j_1,k_2+1,j_2))$$

メモリ : O($\mathbf{L^4}$)

計算時間 : O($\mathbf{L^6}$)



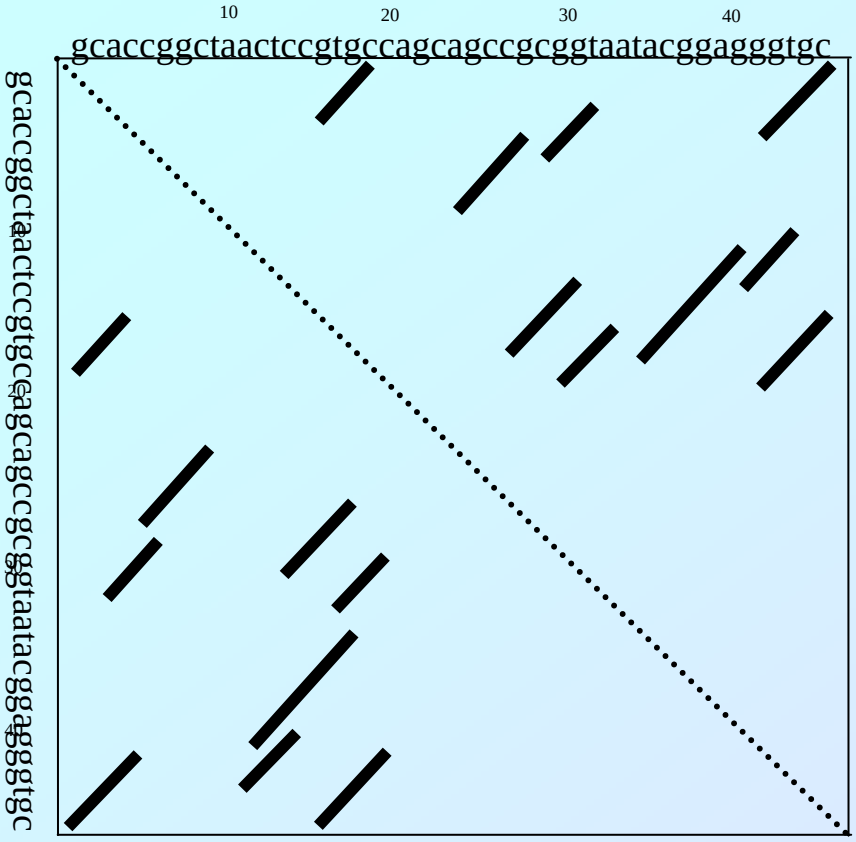
長さの6乗の計算時間とは？

- 配列長が2倍になると64倍の計算時間
- 配列長が10倍になると100万倍！！
- 1分の100万倍は約2年
- Sankoffアルゴリズムでは、150塩基の2配列の共通二次構造の計算に10分以上かかります、、、

Stem Candidate Aligner for RNA

- For each nucleotide sequences,
  - Extract stem candidates of a fixed length
  - Decompose each stem candidates into two parts:
    5’ and 3’ stem component (left SC and right SC)
  - Sort all SCs as a sequence by their positions
- Pairwise alignment DP of SCSs
- Remove inconsistent stem matches
- Construct backbone of common secondary structure
- Alignment of remaining nucleotide

Stem Candidates

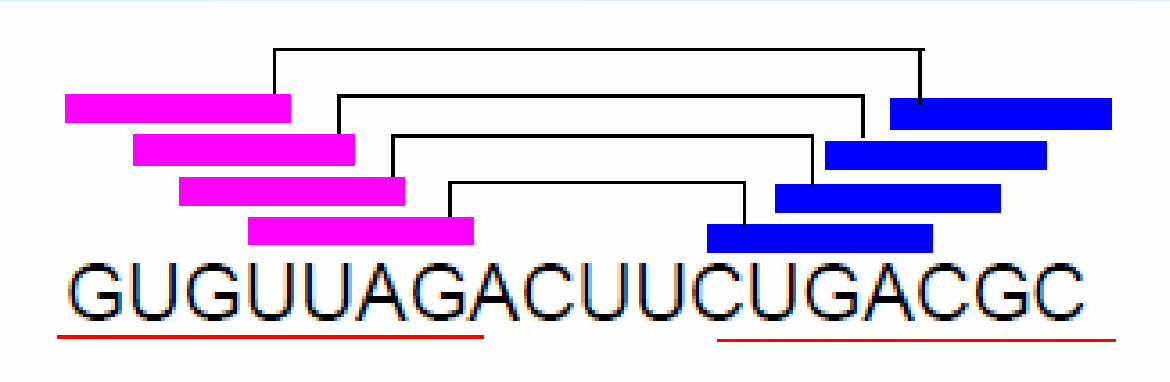


Base-pairing probabilities are calculated by McCaskill's algorithm.

Base pairs of low probabilities are not used

Stem fragments of a fixed length

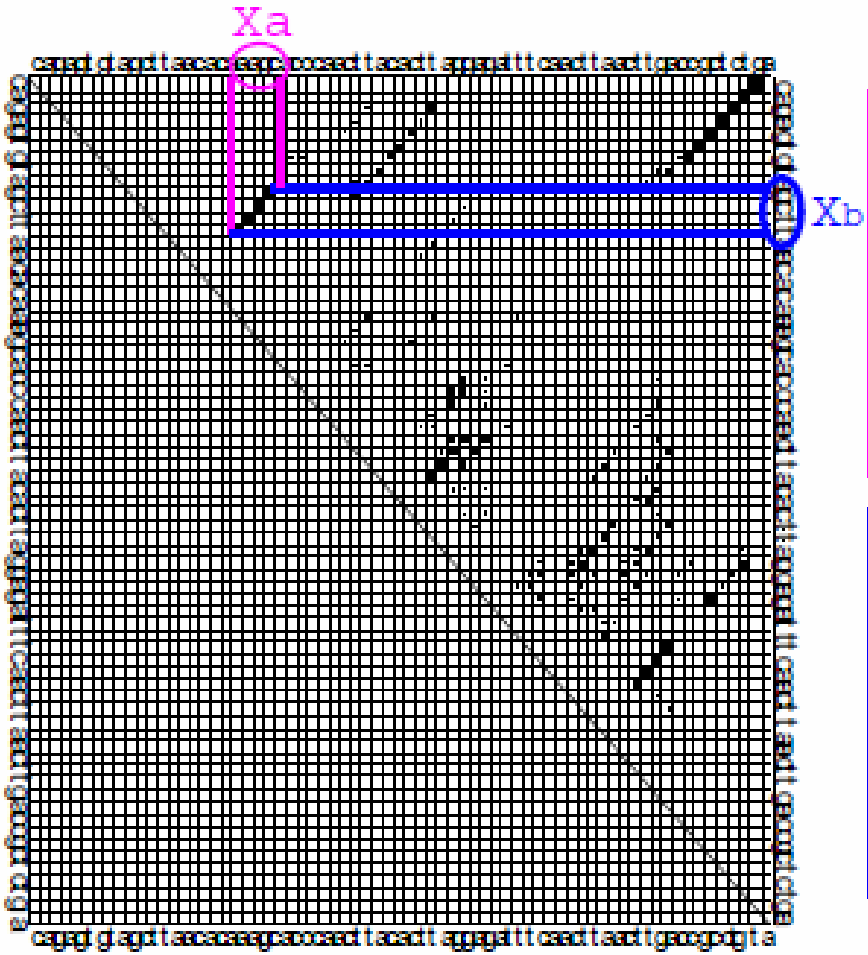
Stem candidates longer than the fixed length is treated as continuous stem fragments



SCARNA
Stem Candidate Aligner for RNA

- For each nucleotide sequences,
  - Extract stem candidates of a fixed length
  - Decompose each stem candidates into two parts:
    5’ and 3’ stem component (left SC and right SC)
  - Sort all SCs as a sequence by their positions
- Pairwise alignment DP of SCSs
- Remove inconsistent stem matches
- Construct backbone of common secondary structure
- Alignment of remaining nucleotide

Decomposition of stem fragments into stem components



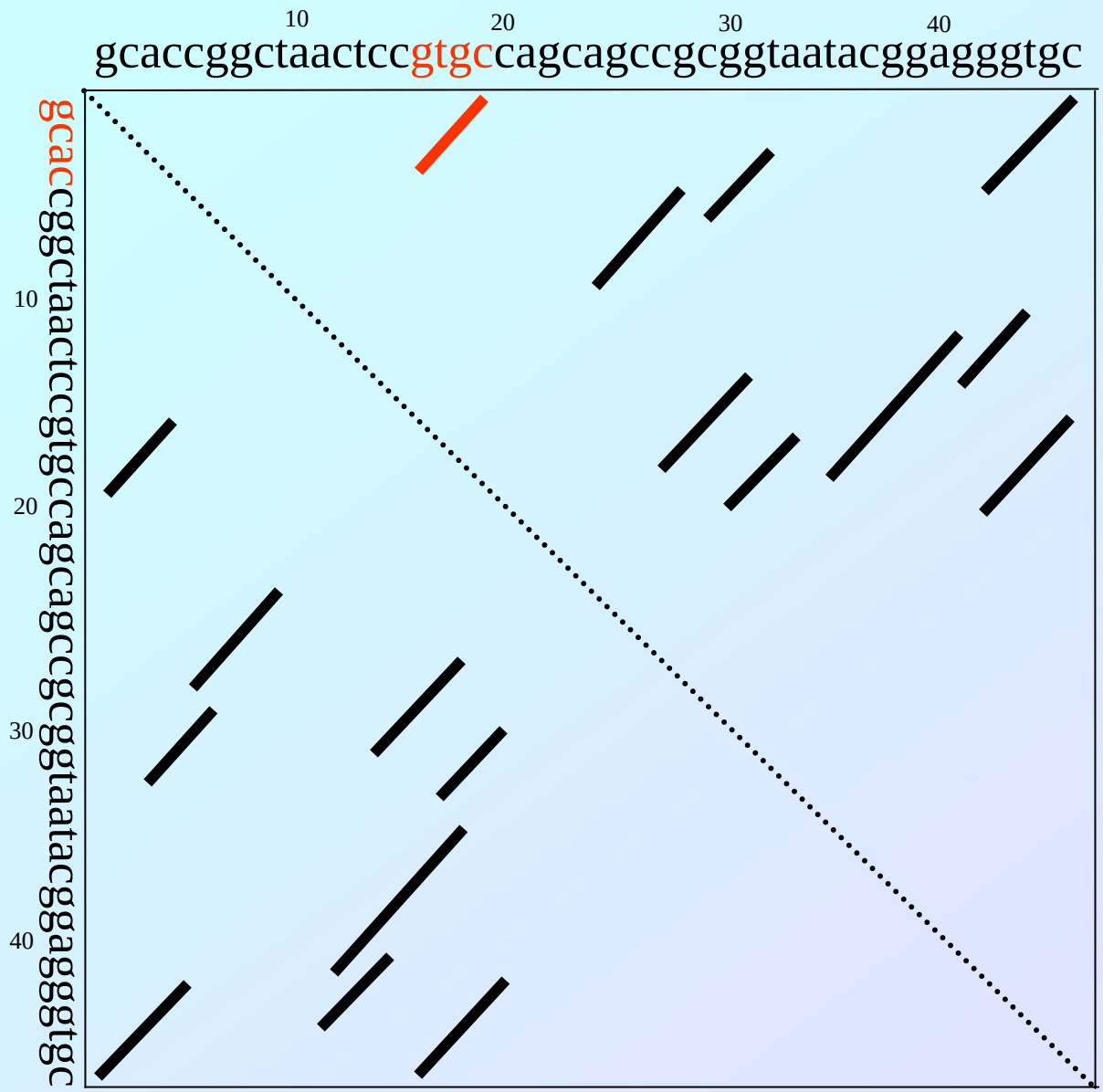
5'Stem Component (X_a)
(a)20 position
(b)aagc sequence
(c)-6 distance
(d)gctt complementary
(e)3.84 confidence score
(f)-0.9 stacking energy

3'Stem Component (X_b)
(a)10
(b)gctt
(c)6
(d)aagc
(e)3.84
(f)-0.9

Stem Candidate Aligner for RNA

- For each nucleotide sequences,
  - Extract stem candidates of a fixed length
  - Decompose each stem candidates into two parts:
    5’ and 3’ stem component (left SC and right SC)
  - Sort all SCs as a sequence by their positions
- Pairwise alignment DP of SCSs
- Remove inconsistent stem matches
- Construct backbone of common secondary structure
- Alignment of remaining nucleotide

Stem Component Sequence (SCS)

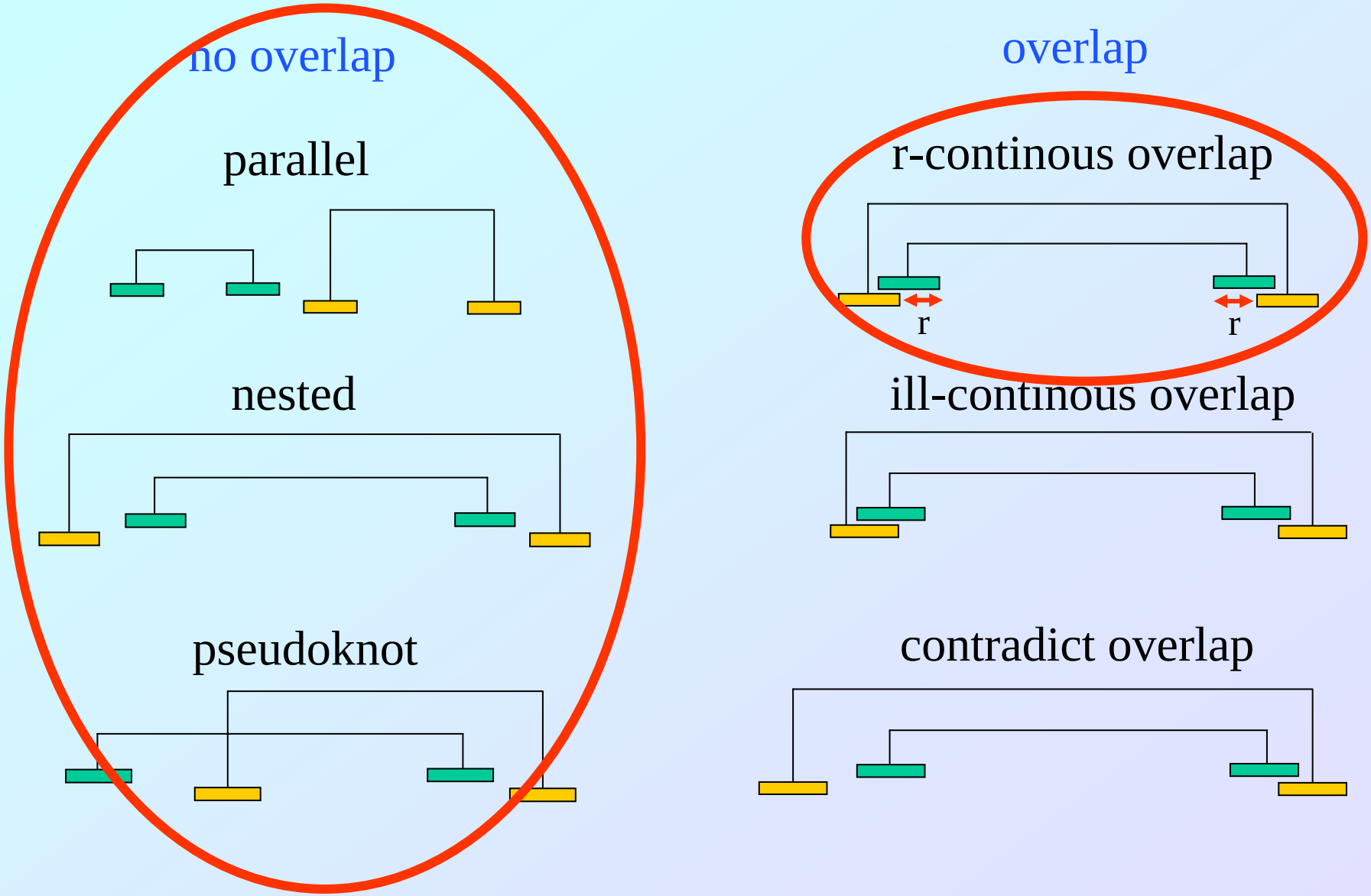


1 gcac	11	gtgc 16
1 gcac	38	gtgc 43
2 cacc	36	ggtg 42
3 accg	22	cggt 29
5 cggc	16	gccg 25
6 ggct	14	agcc 24
11 actc	26	gggt 41
12 ctcc	22	ggag 38
13 tccg	20	cgga 37
14 ccgt	10	gcgg 28
14 ccgt	18	acgg 36
15 cgtg	8	cgcg 27
15 cgtg	16	tacg 35
16 gtgc	– 11	gcac 1
16 gtgc	23	gtgc 43
17 tgcc	9	ggta 30
17 tgcc	– 21	ggtg 42
24 agcc	– 14	ggct 6
25 gccg	– 16	cggc 5
27 cgcg	– 8	cgtg 15
28 gcgg	– 10	ccgt 14
29 cggt	– 22	accg 3
30 ggta	– 9	tgcc 17
35 tacg	– 16	cgtg 15
36 acgg	– 18	ccgt 14
37 cgga	– 20	tccg 13
38 ggag	– 22	ctcc 12
41 gggt	– 26	actc 11
42 ggtg	– 36	cacc 2
42 ggtg	– 21	tgcc 17
43 gtgc	– 38	gcac 1
43 gtgc	– 23	gtgc 16

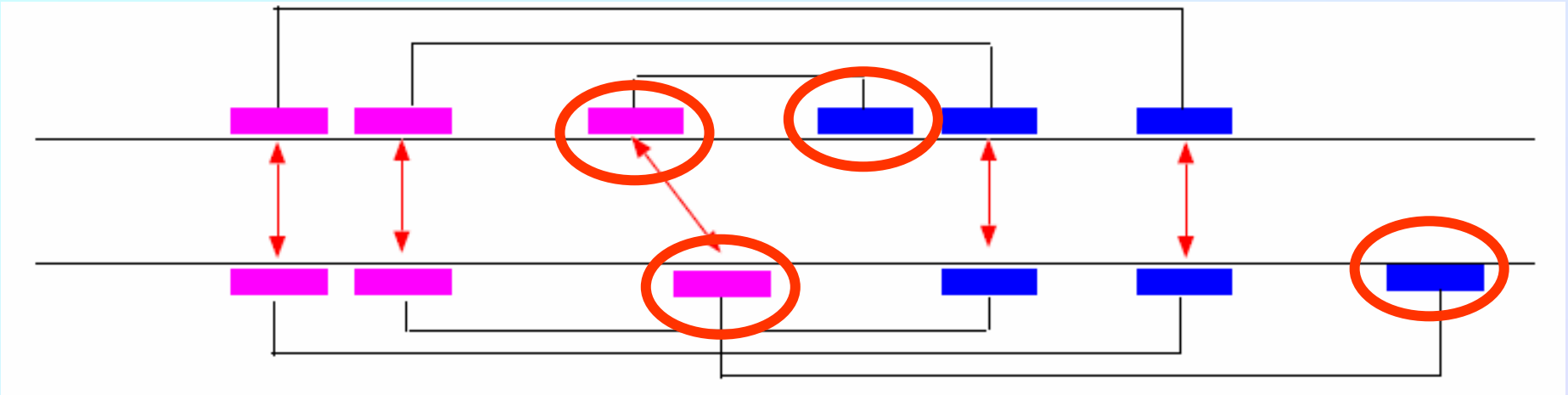
SCARNA
Stem Candidate Aligner for RNA

- For each nucleotide sequences,
  - Extract stem candidates of a fixed length
  - Decompose each stem candidates into two parts:
    5’ and 3’ stem component (left SC and right SC)
  - Sort all SCs as a sequence by their positions
- Pairwise alignment DP of SCSs
- Remove inconsistent stem matches
- Construct backbone of common secondary structure
- Alignment of remaining nucleotide

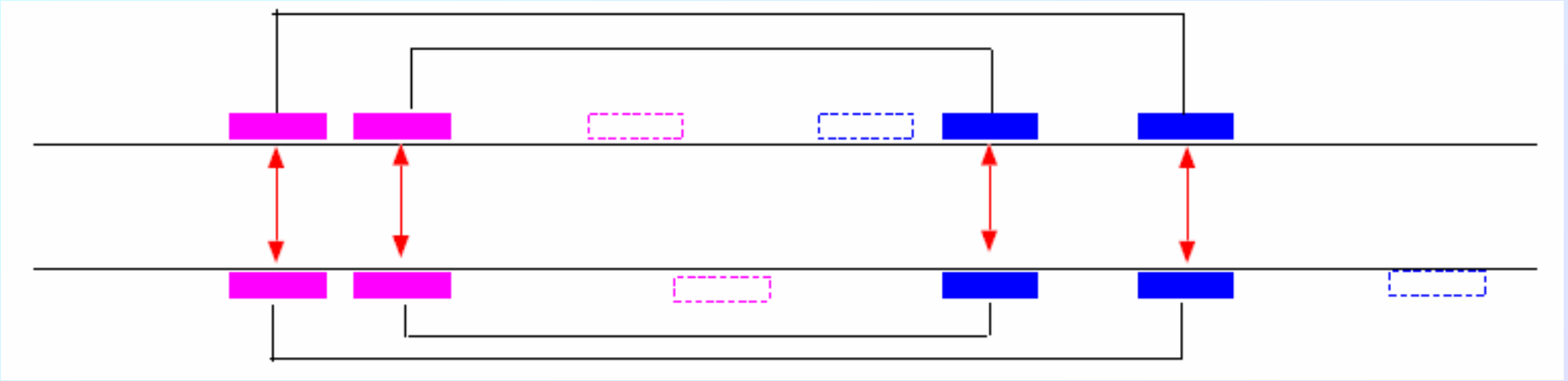
Relations of two stem fragments



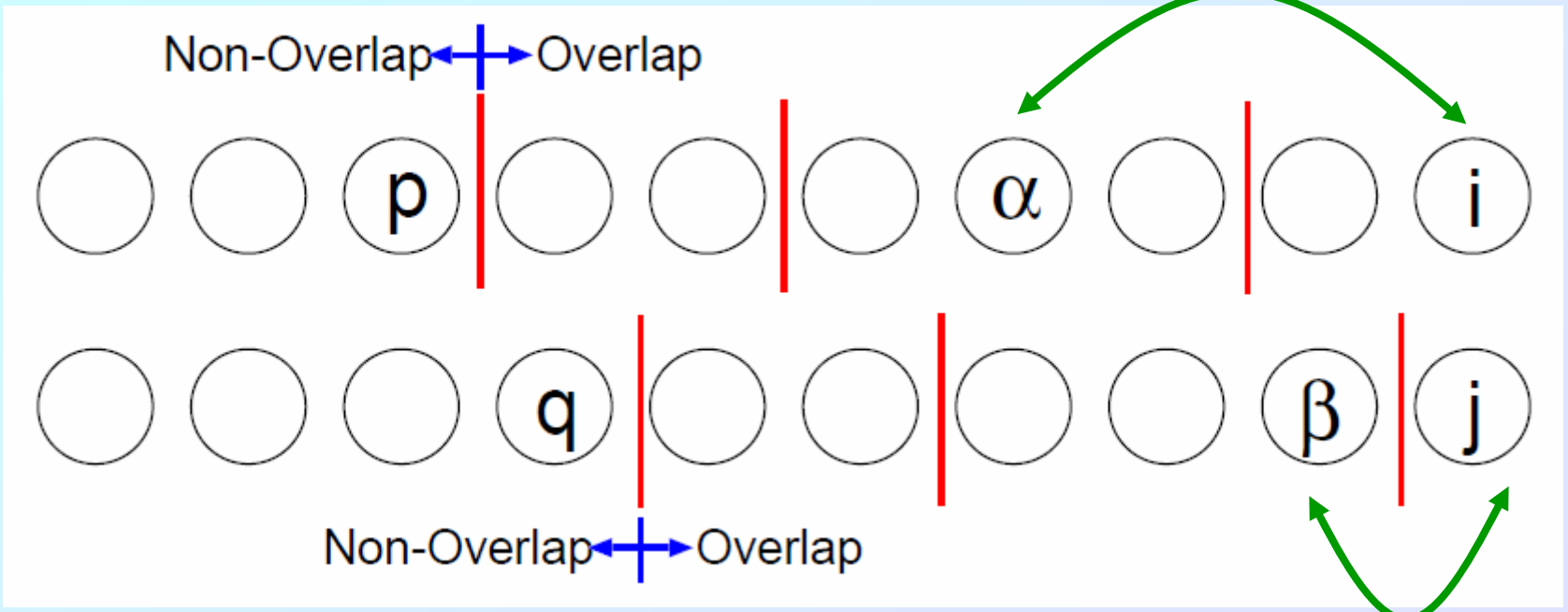
Left-right consistency of the SC match in SCS alignment



Left-right consistency of the SC match in SCS alignment

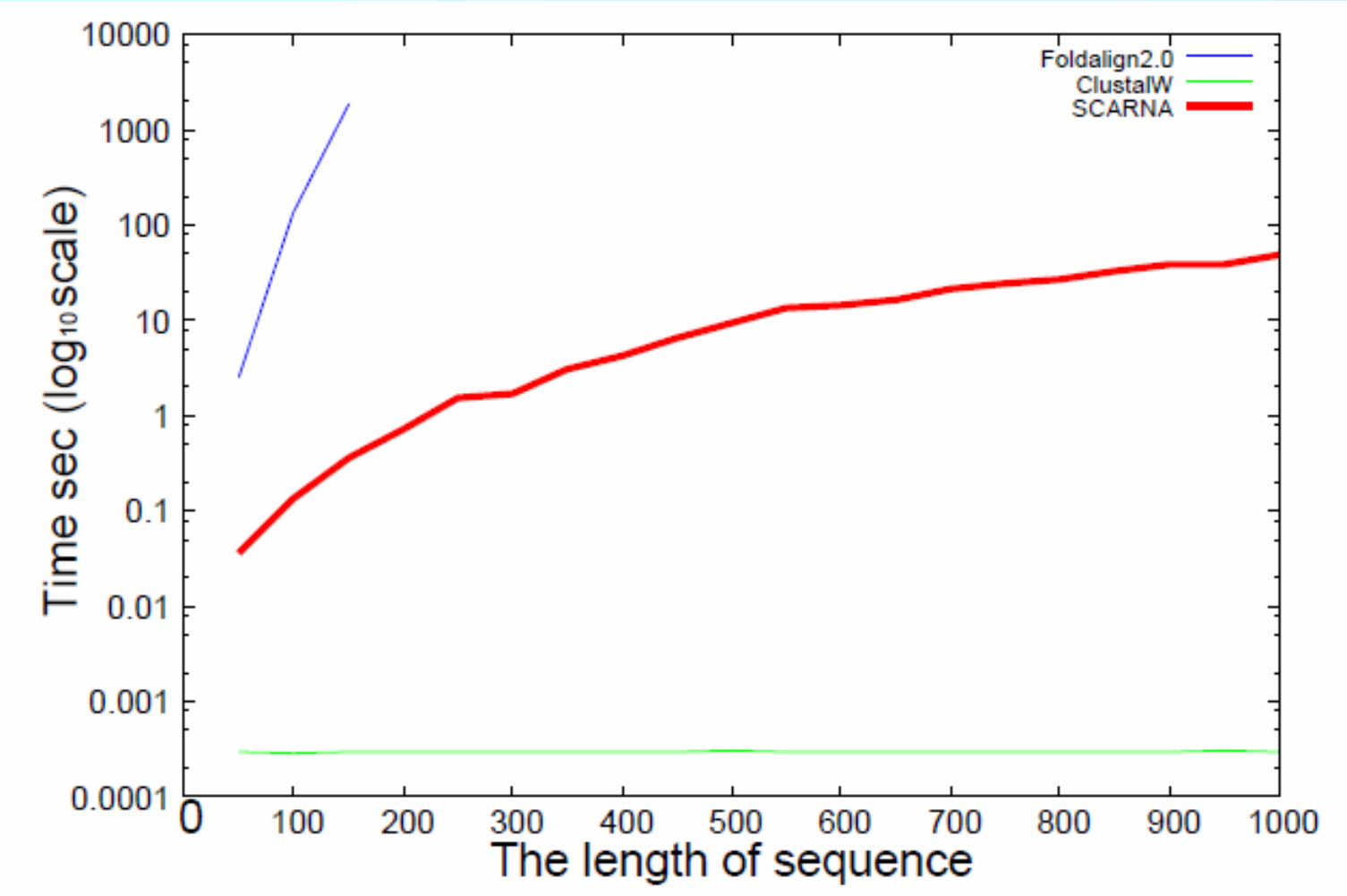


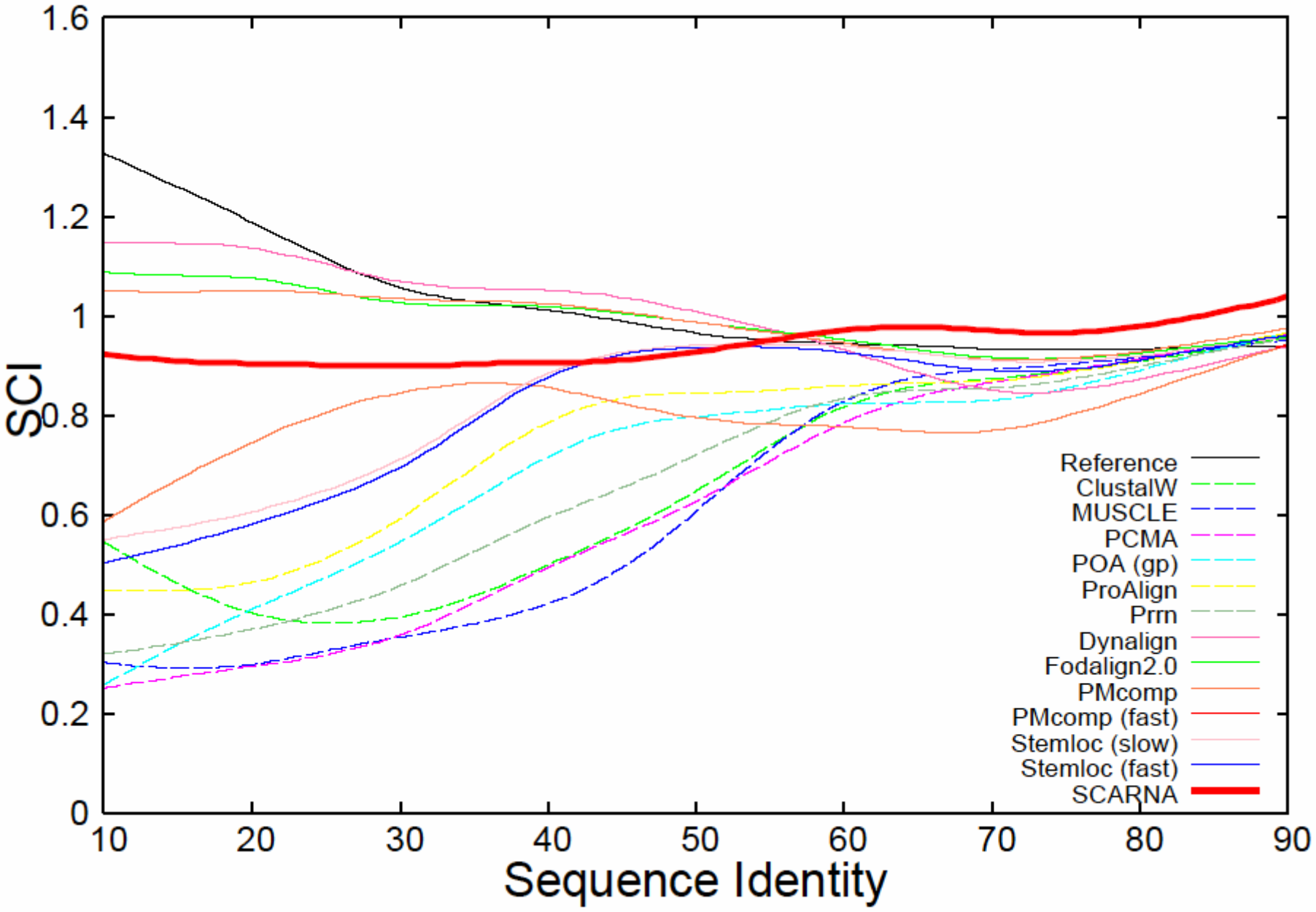
Dependencies in DP of SCSs



= SCs with same positions and different complementary partners

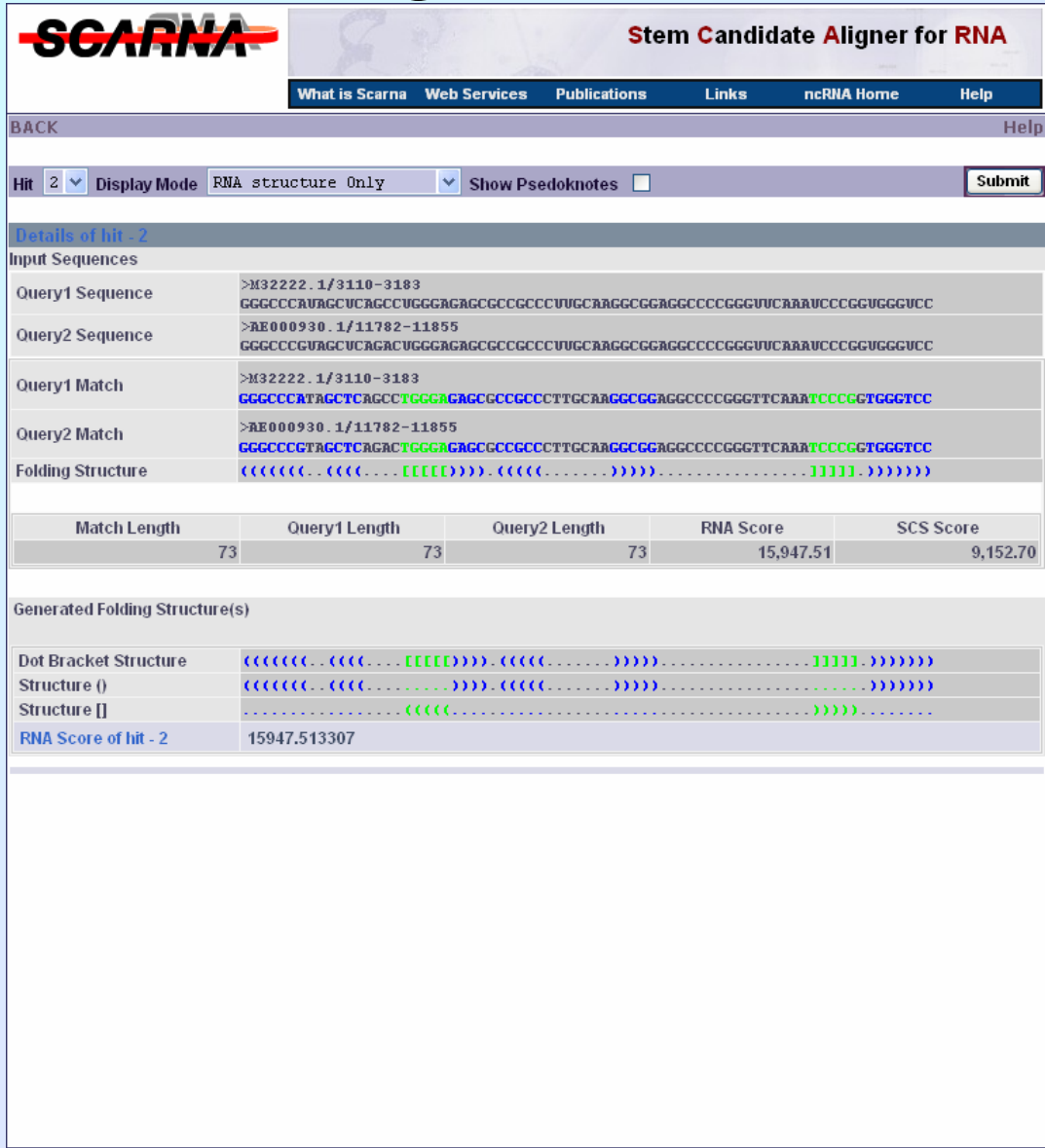
Computational Time




SCI
Sequence Identity
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
90
Reference
ClustalW
MUSCLE
PCMA
POA (gp)
ProAlign
Prrn
Dynalign
Fodalign2.0
PMcomp
PMcomp (fast)
Stemloc (slow)
Stemloc (fast)
SCARNA

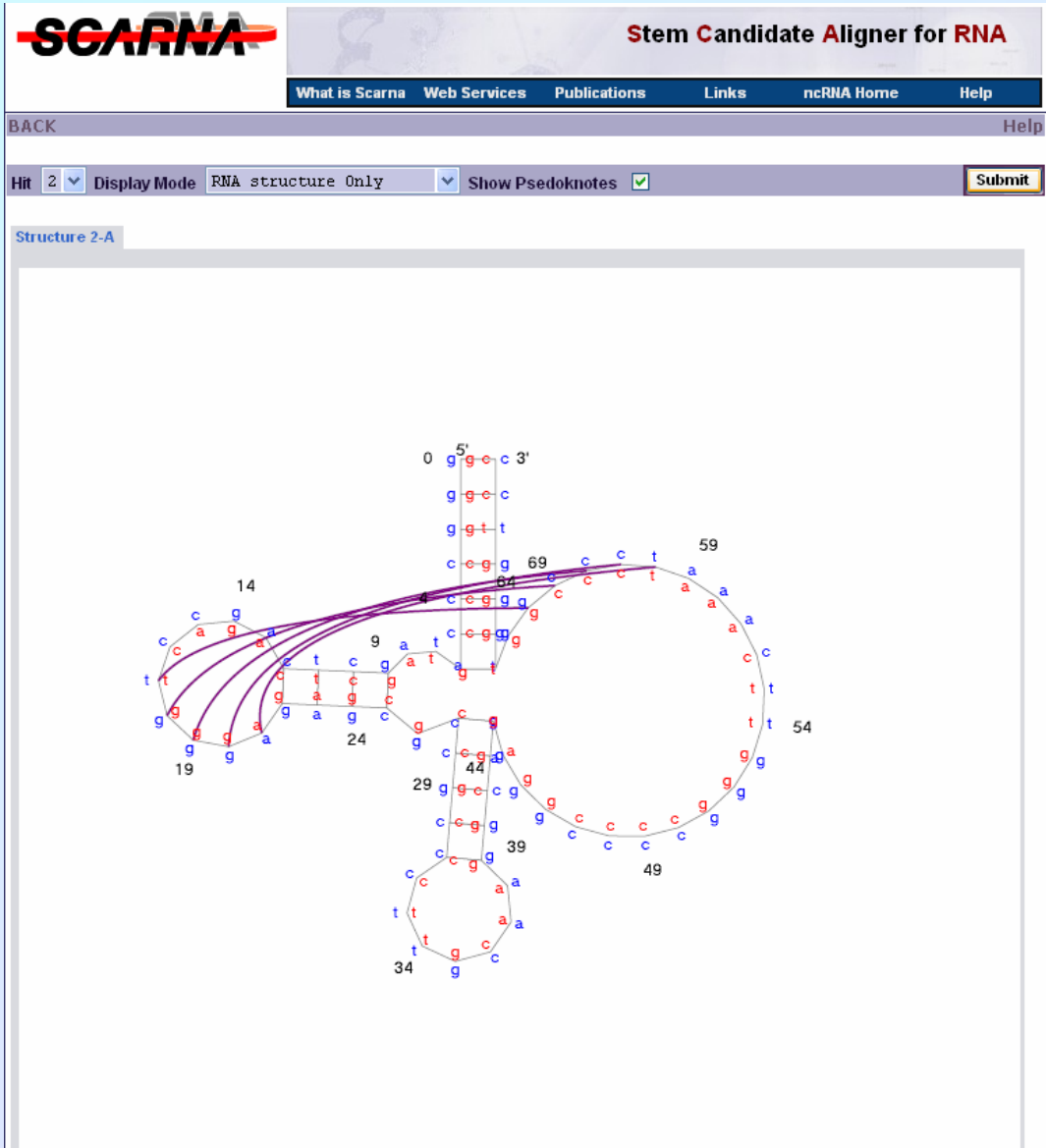
Scarna web server

http://www.scarna.org



Scarna web server

http://www.scarna.org



Acknowledgments

- “Functional RNA Project” of METI
- Grant-in-Aid for Scientific Research on Priority Area “Comparative Genomics”
- AIST, CBRC, BIRC
- JBIC
- The University of Tokyo