CMfinder - a covariance model based RNA motif finding algorithm

ZIZEN YAO, ZASHA WEINBERG and WALTER L. RUZZO

source: http://bioinformatics.oxfordjournals.org/content/22/4/445.abstract

BIOINFORMATICS ORIGINAL PAPER

Vol. 22 no. 4 2006, pages 445–452
doi:10.1093/bioinformatics/btk008

Sequence analysis

CMfinder—a covariance model based RNA motif finding algorithm

Zizhen Yao[1,*], Zasha Weinberg[1] and Walter L. Ruzzo[1,2]

[1]Department of Computer Science and Engineering and [2]Department of Genome Sciences, University of Washington, Seattle WA 98195-2350, USA

Received on June 9, 2005; revised on December 12, 2005; accepted on December 13, 2005
Advance Access publication December 15, 2005
Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: The recent discoveries of large numbers of non-coding RNAs and computational advances in genome-scale RNA search create a need for tools for automatic, high quality identification and characterization of conserved RNA motifs that can be readily used for database search. Previous tools fall short of this goal.

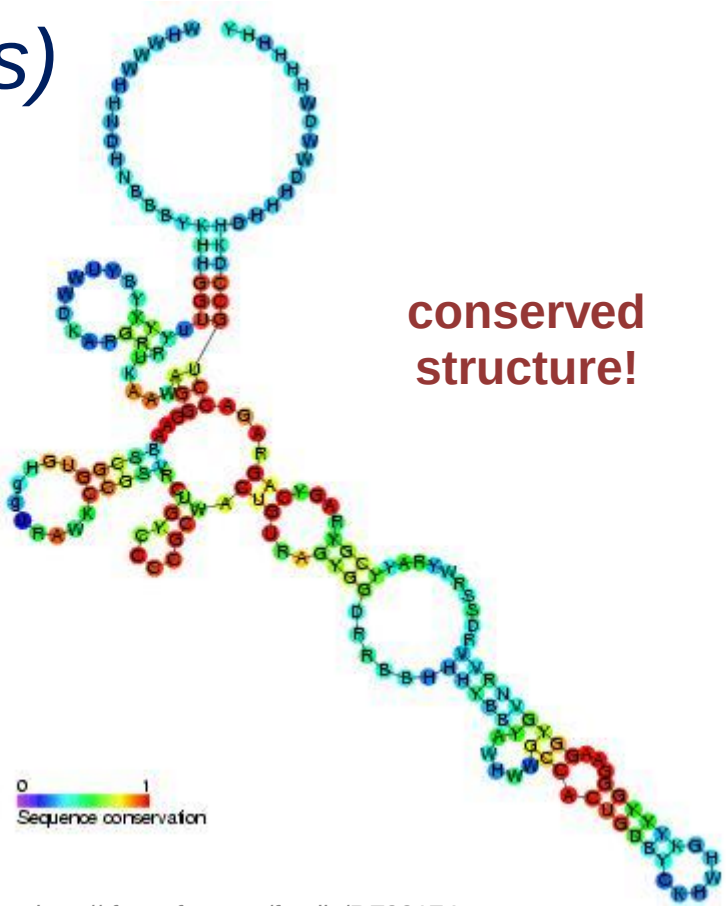
Results: CMfinder is a new tool to predict RNA motifs in unaligned sequences. It is an expectation maximization algorithm using covari-

Important computational advances, such as development of the Rfam database (Griffiths-Jones *et al.*, 2003) and fast genome-scale covariance model (CM) searches (Eddy and Durbin, 1994; Weinberg and Ruzzo, 2004a,b), aid RNA research significantly. A key problem remaining is to identify conserved secondary structure motifs among related sequences, and characterize them by models that can be used for homology search. For example, identification of such motifs in untranslated regions of orthologous

Non-coding RNAs (ncRNAs)

- functional important
- e.g. riboswitches –

  regulate own expression



conserved structure!

source: http://rfam.xfam.org/family/RF00174

Aim

- predict RNA consensus motif

  → find unknown

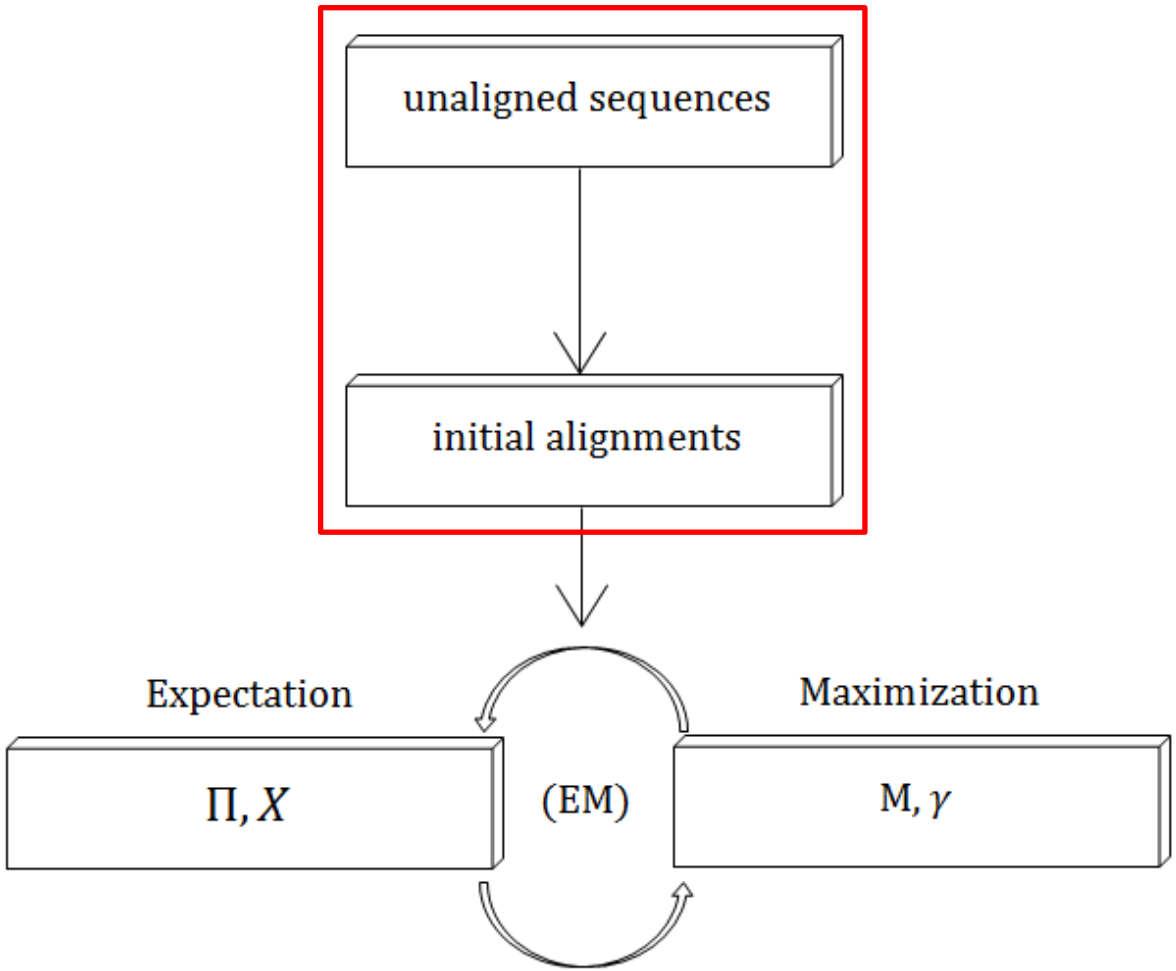
Key Idea

- simultaneous folding and aligning

Methods

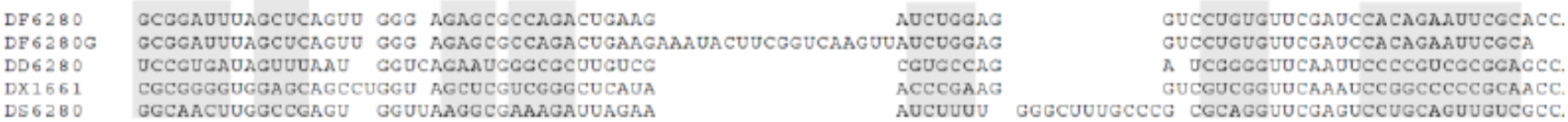
I. construction of heuristic initial alignment

II. refining alignments via covariance model-based expectation maximazation

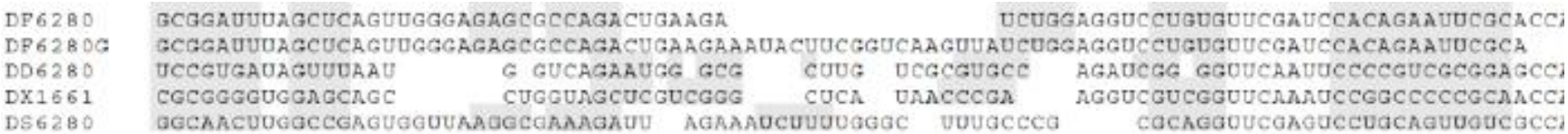

unaligned sequences
initial alignments
Expectation
Maximization
Π, X
(EM)
M, γ

I. construction of heuristic initial alignment

Trusted:

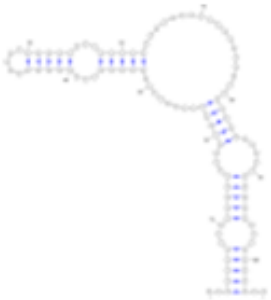


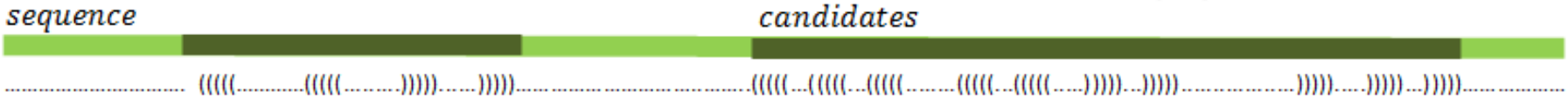
ClustalV:



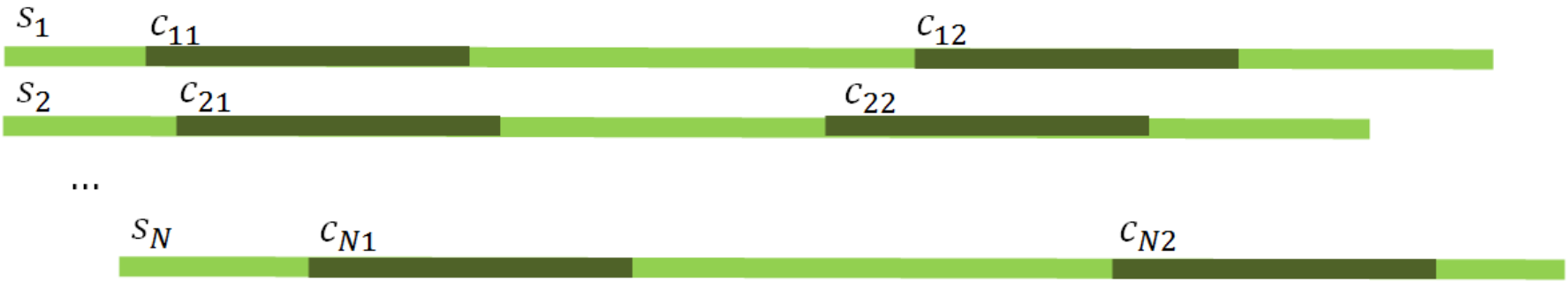
source: Paper "RNA sequence analysis using covariance model"

a. candidate selection





b. candidate comparison and alignment



N: $total\ number\ of\ sequences:$

m: $number\ of\ candidates\ in\ each\ sequence:$

$S = (s_i)_{1 \le i \le N}$: $input\ sequences:$

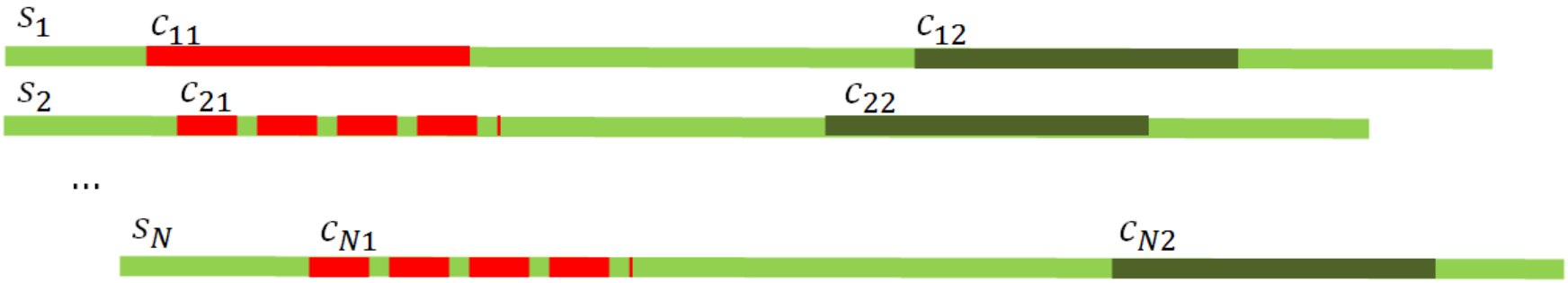
$C_i = (c_{ij})_{1 \le j \le m}$: $candidate\ set\ of\ sequence\ s_i:$

b. candidate comparison and alignment

- compare predicted secondary structure
- use modified tree-edit algorithm

→ sensitive in **sequence** and **structure**

b. candidate comparison and alignment

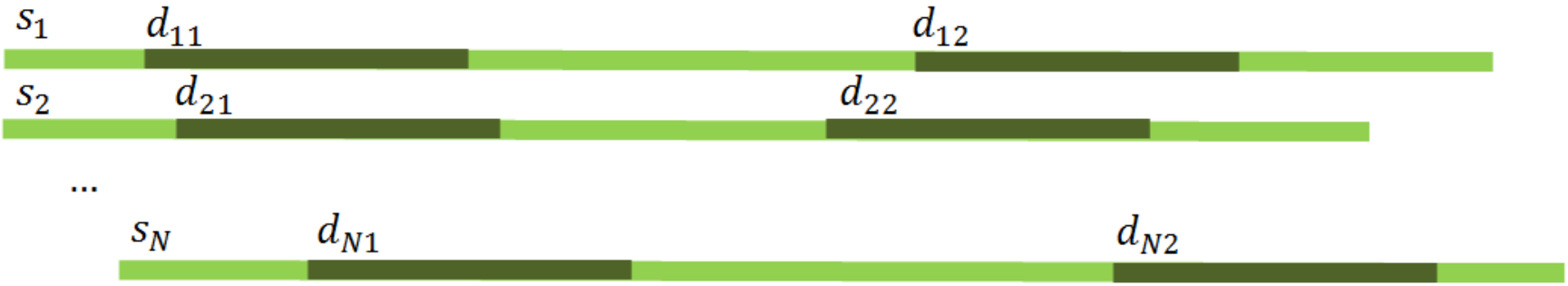


$$d_{ij} = \sum_{k \neq i} dist(c_{ij}, c_{kl}),\ l = \ argmin_{l'}\, dist(c_{ij}, c_{kl'})$$

e.g.:

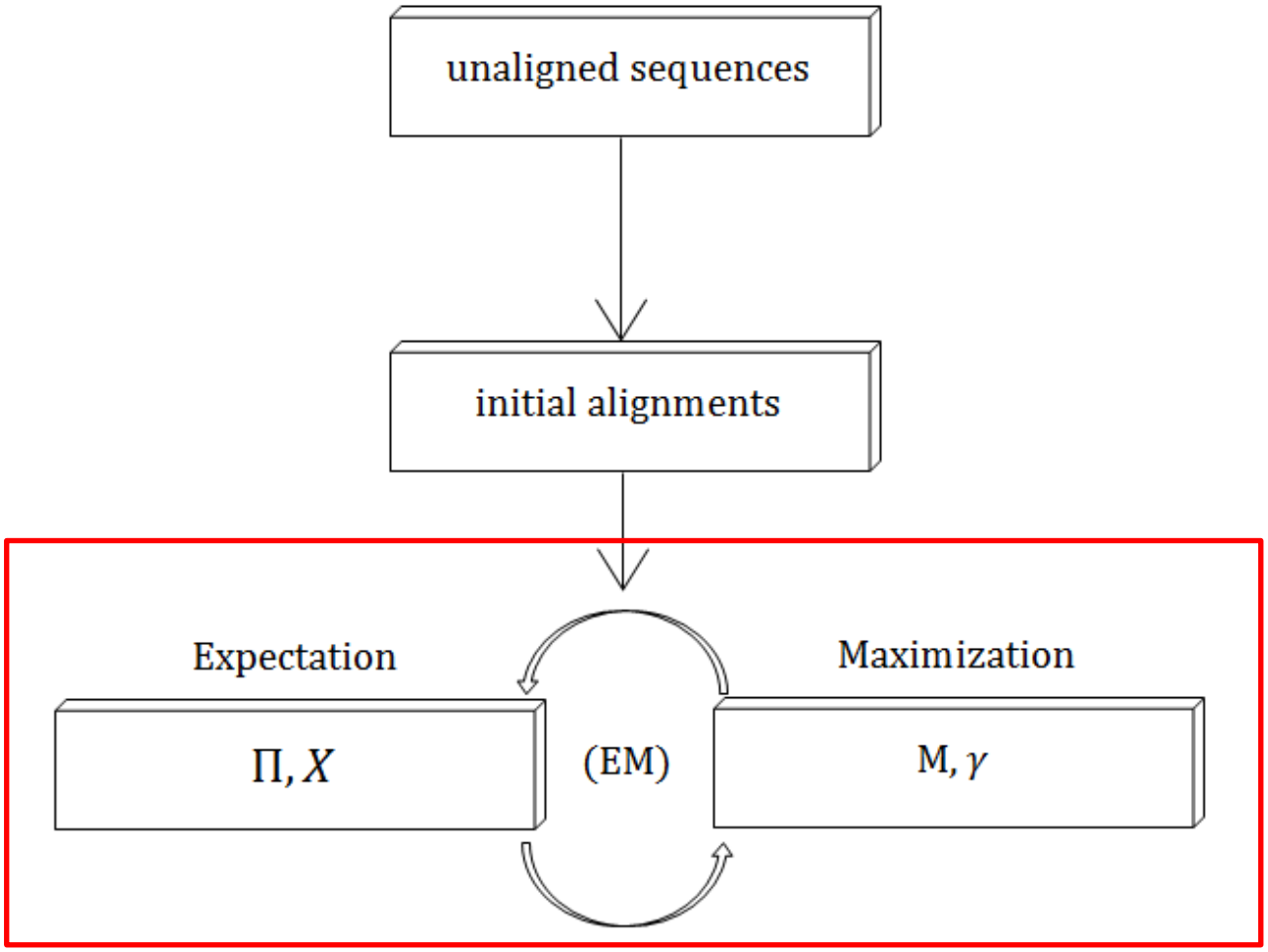
$$d_{11} = \sum_{k \neq i} dist(c_{11}, c_{kl})$$

b. candidate comparison and alignment



- "consensus candidate" $\rightarrow$ minimal d
- alignment to best match in each sequence

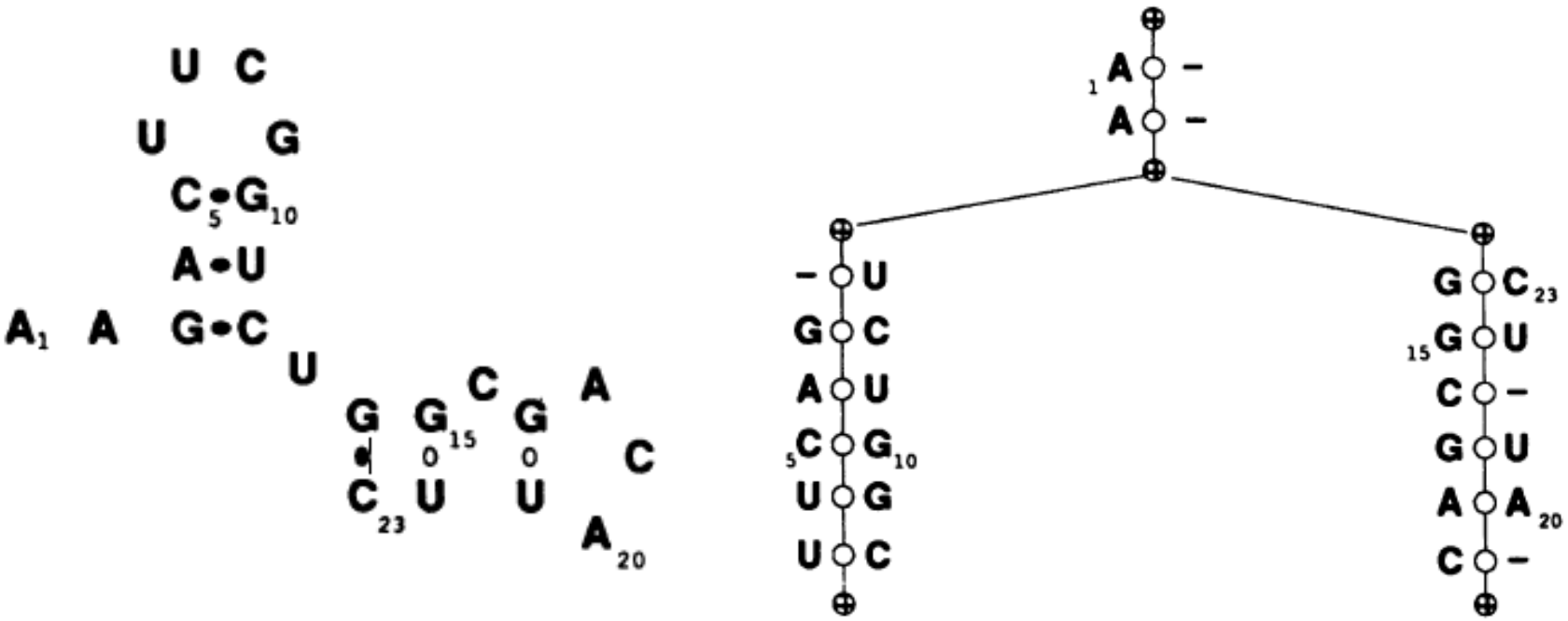
  if $< threshold$


unaligned sequences
initial alignments
Expectation
Maximization
Π, X
(EM)
M, γ

II. Refining alignments via CM based EM

- iterative expectation-maximazation-algorithm (EM) optimizes:
  - covariance model for motif description
  - bayesian framework (mutual information and folding energy) for secondary structure prediction

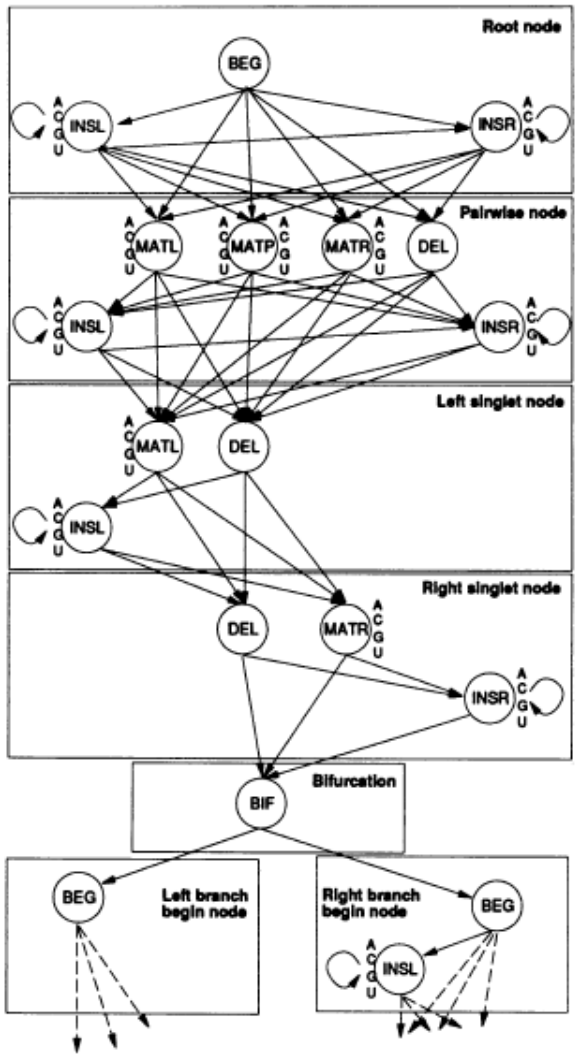
Covariance model



source: Paper "RNA sequence analysis using covariance model"

Covariance model

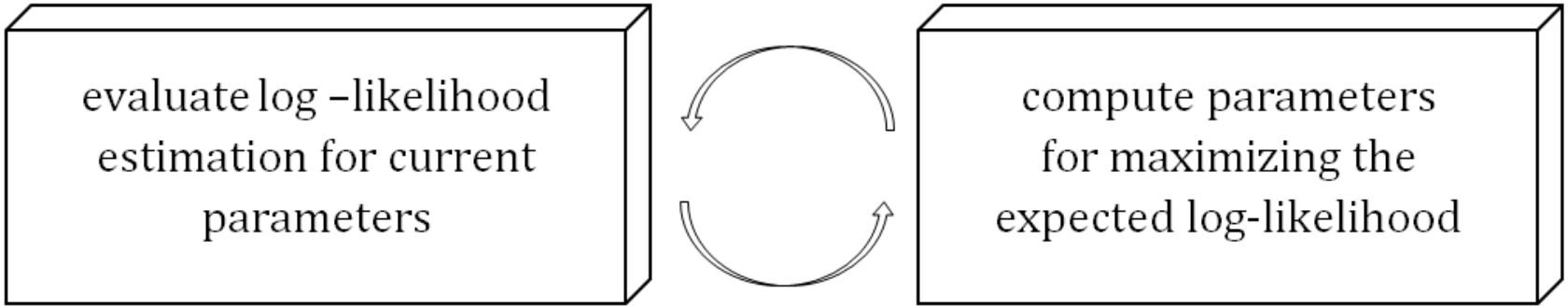
- states
  - MATP
  - MATL, MATR
  - INSL, INSR
  - DEL
- symbol emission probablilities
- state transition probablities



source: Paper "RNA sequence analysis using covariance model"

Expectation-Maximazation

- iterative methode
- Goal: find log likelihood



II. refining alignments via CM based EM

M:	motif CM
B:	background distribution
$\Gamma = (M, B, \gamma)$:	finite mixture model, γ mixture probability that sequence contains a motif
$\Pi_i = (\pi_{ij})$:	alignment of candidates C_i with M

II. refining alignments via CM based EM

$X_i = (x_{ij})$: occurence of motif in C_i
($x_{ij} = 1$ if c_{ij} is a motif instance)

$D = (L_1, L_{2,} \ldots, L_l)$: sequence alignment

$\sigma = (\alpha, \beta)$: consensus secondary structure for D.
α: indices of single stranded colums,
β: pairs of indices of base paired colums

X_i, Π_i

b. Expectation-step

- update π_{ij}: Viterbi algorithm
- update x_{ij}:

$$P(x_{ij} = 1) = \frac{\lambda \cdot \frac{P(c_{ij}|M)}{P(c_{ij}|B)}}{1 - \gamma + \sum_{k=1}^{m} \lambda \cdot \frac{P(c_{ik}|M)}{P(c_{ik}|B)}}, \qquad \lambda = \frac{\gamma}{m}$$

$\boldsymbol{M}$, $\boldsymbol{\gamma}$

b. Maximization-step

- update $\gamma = \frac{1}{N} \cdot \sum_{i=1}^{N} \sum_{j=1}^{m} x_{ij}$
- update M:
  - i. change the structure of M
  - ii. infer emission and transition probabilities

ii. infer emission and transition probabilities

- *emission:* bayesian posterior estimate with Dirchlet prior
- *transition:* equivalent to find

$$\hat{\sigma} = argmax_{\sigma}\, P(D, \sigma)$$

ii. infer emission and transition probabilities

- maximum likelihood for:

$$\sum_{(i,j)\in\beta} K_{ij};\ K_{ij} = I_{ij} + \log\frac{p_{ij}}{q_i q_j}$$

p_{ij}: *prior based paired*

q_i: *prior single standed*

based on a thermodynamic model

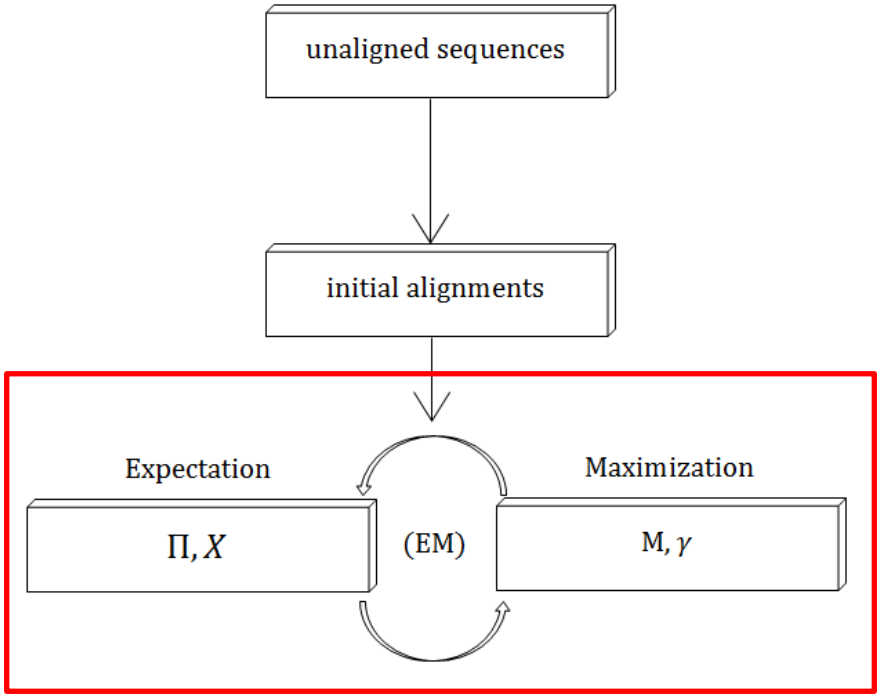
II. refining alignments via CM based EM

- iterative between

  E- and M-step

- adjusting

  candidates



Run time

- *per iteration:* $O(N \cdot L^3 \cdot |M|)$

$\|M\|$:	number of states in CM
L:	maximum sequence length
N:	total number of sequences

- <15 iteration, 1-60 minutes

Example of use

University of Washington
Computer Science & Engineering

News: CMfinder software download is available now.

You need help adjusting the parameters of CMfinder? Please read the manual. Questions?send mail

You can run CMfinder using 2 sets of parameters at one time

First configuration:

Number of stem-loops	1	
Number of motifs	3	<10
Minimum length of motif	30	>15
Maximum length of motif	100	<150
Number of Candidates	40	<100

Second configuration:

Number of stem-loops	2	
Number of motifs	3	< 10
Minimum length of motif	30	> 15
Maximum length of motif	100	< 150
Number of Candidates	40	< 100

Paste Sequences (*) in FastA Format
Please limit your dataset to

Advantages

- ✓ high average specifity (81%) and sensitivity (77%) in term of base paires
- ✓ works with low and high sequence similarity
- ✓ update and combining CM

Advantages

- ✓ robust with flanking sequences
- ✓ faster than other comparable applications *(2005)*
- ? directly useable for homology search

limitation and simplification

- × possible for < 60 sequences
- × viterbi approximation instead of running inside-outside algorithm in E-step
- × no support of pseudo nodes

Summary

- comparative methode
- prediction and localisation of consensus

  RNA motifs
- simultaneous folding and aligning

Summary

- initial alignment based on sequence and structur
- EM-algorithm use CM
- Bayesian framework for structure prediction (folding energy, sequence covariation)

Thank you for your attention.