Algorithm to capture local RNA sequential- and structural- motifs

Hiroshi Miyake

PhD student

THE UNIVERSITY OF TOKYO

Computational Approaches to RNA Structure and Function
2018.7.15 – 7.27 @ Benasque

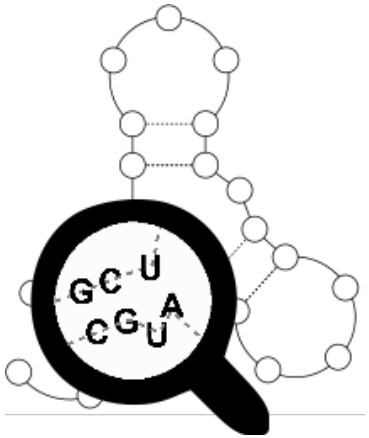
Algorithm to capture local RNA sequential- and structural- motifs

Pico de Aneto

Great landscape !

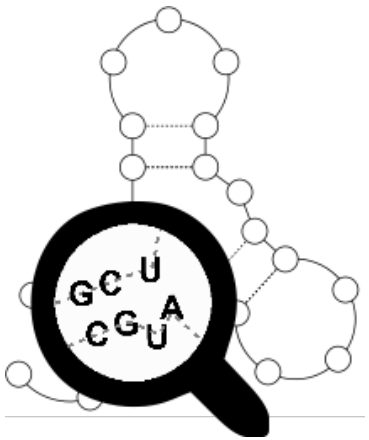
RNA motif discovery

RNA motif = combination of conserved subsequence and local structure



RNA motif discovery

RNA motif = combination of conserved subsequence and local structure



HMM

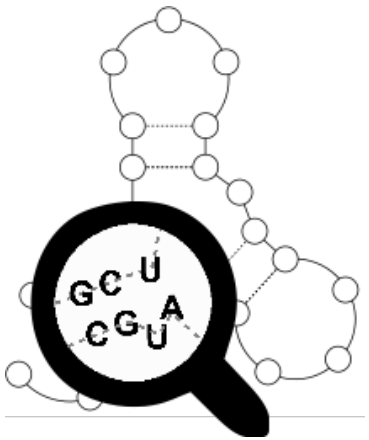
In / Del

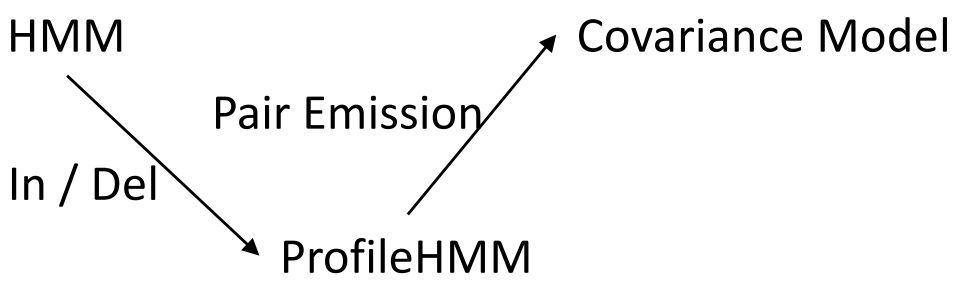
ProfileHMM

Krogh et.al.(1994). Hidden Markov models in computational biology: Applications to protein modeling. *Journal of molecular biology*, *235*(5), 1501-1531.

RNA motif discovery

RNA motif = combination of conserved subsequence and local structure

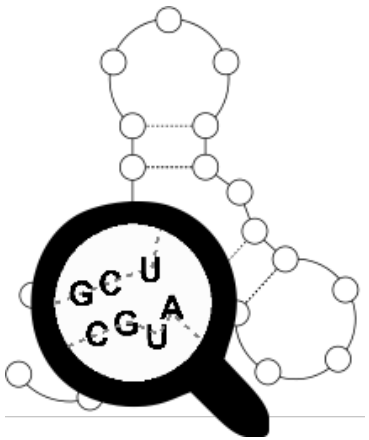




Eddy, S. R. et.al. (1994). RNA sequence analysis using covariance models. *Nucleic acids research*, *22*(11), 2079-2088.

RNA motif discovery

RNA motif = combination of conserved subsequence and local structure

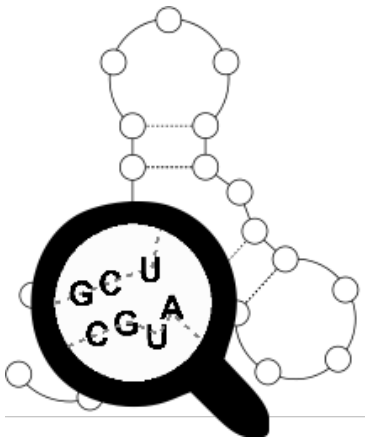


HMM → (In / Del) → ProfileHMM → (Pair Emission) → Covariance Model → (Energy model) → CMfinder

Yao, Z., Weinberg, Z., & Ruzzo, W. L. (2005). CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics*, *22*(4), 445-452.

RNA motif discovery

RNA motif = combination of conserved subsequence and local structure



HMM → ProfileHMM (In / Del)

ProfileHMM → Covariance Model (Pair Emission)

Covariance Model → CMfinder (Energy model)

Local structural motif

RNAcontext

GraphProt

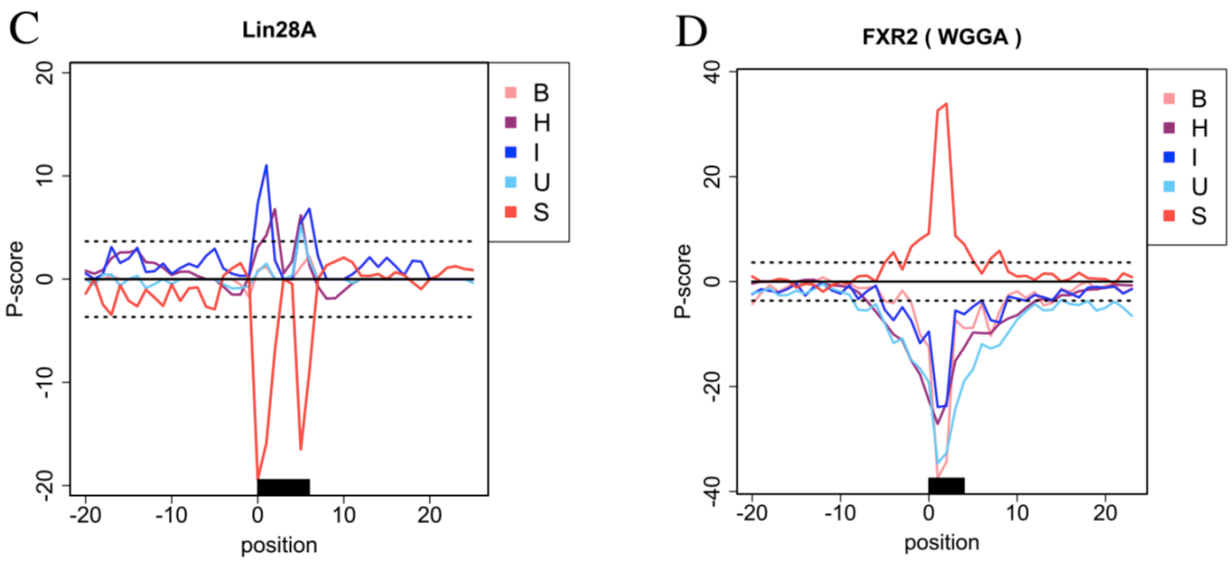
Kazan, H., Ray, D., Chan, E. T., Hughes, T. R., & Morris, Q. (2010). RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS computational biology*, *6*(7), e1000832.

Maticzka, D., Lange, S. J., Costa, F., & Backofen, R. (2014). GraphProt: modeling binding preferences of RNA-binding proteins. *Genome biology*, *15*(1), R17.

Research aim

Open question: a mathematical model which can

- Capture complex 2D structural context



- Capture covariance between two loci
- Capture a gap inside motif

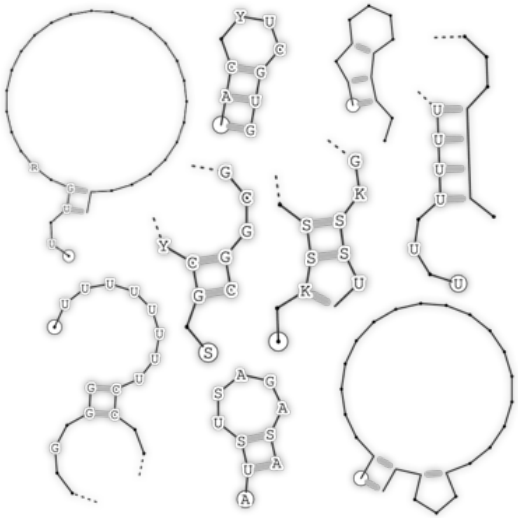
Fukunaga, T., Ozaki, H., Terai, G., Asai, K., Iwasaki, W., & Kiryu, H. (2014). CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. *Genome biology*, *15*(1), R16.

Development of RNAelem

iyak/RNAelem

input.fa

```
...GGGGGAGGAAGTGGCTAGCTCAGGGCTTCAGGG...
...ACAGACAGGGAGAGATGACTGAGTTAGATGAGA...
...CGAGGGGGCGGGCTGGGGGTGCGAGAAGGAAGC...
...TGGCAAGGAGACTAGGTCTAGGGGGACCACAGG...
...GGGCAGGCTGCATGGAAAGGGGGCGGGGCCTGG...
...CTGCAGGCGGACCCCGTGGAAGGGTTTCGCGGG...
...TAGCGGGGACTCCTCGGGAGTCTTACAGGGCGG...
...AGCTTAAGGTGCCGGAAAAGTGGAAAATTACCA...
...AAAGCAGGAAGGGAGGGTTAGCCTTGGGAAACC...
...AATCTGGGTTTGCCACGGGGGCTTACTGAGTCA...
...GGCCCCCAGTCCCACAATTGGAAGAGATTGACG...
...GTGTAGTGTCTTCAAGCTTGCTTTTTGGTGGGG...
...ATTGGGGAGCTGTCGGGGCGGCTGCCTTTGGTA...
...GCTGTTGAGGGAGTCTGGGGCTTGTGAGCTGTA...
```



SCFG for 2D structure

Discriminative model of 2D structure σ under given sequence x

$$P(\sigma \mid x) = \frac{g(\sigma, x)}{Z(x)}, \qquad Z(x) = \sum_{\sigma} g(\sigma, x)$$

$$g(\sigma, x) = e^{-\frac{1}{kT}\Delta G(\sigma, x)}$$

SCFG for 2D structure

Discriminative model of 2D structure σ under given sequence x

$$P(\sigma \mid x) = \frac{g(\sigma, x)}{Z(x)}, \qquad Z(x) = \sum_{\sigma} g(\sigma, x)$$

$$g(\sigma, x) = e^{-\frac{1}{kT}\Delta G(\sigma, x)}$$

x CAUGCUAGCUAGUCUGCGUAGCUGCGUACUAGCGCGUACGUCGGAU

		$\ln g(\sigma, x)$
$\sigma \in$	.(((((..((((((.(((.....))).))))))))))).....	-13.8
	...((((((.....)).)))).((((((.....))))))....	-13.4
	.....((((((((.((.((((....))))))))).))))))	-13.1
	.....((((((((.(.((((....))))))))).))))))	-12.8
	⋮	

SCFG for primary sequential profile

Discriminative model of alignment ψ with base- and base pair- emission, under given sequence x and pattern of interest

$$P(\psi \mid x) = \frac{h(\psi, x, \theta)}{Z\,(x, \theta)}, \qquad Z(x, \theta) = \sum_{\psi} h(\psi, x, \theta)$$

SCFG for primary sequential profile

Discriminative model of alignment ψ with base- and base pair- emission, under given sequence x and pattern of interest

$$P(\psi \mid x) = \frac{h(\psi, x, \theta)}{Z\,(x, \theta)}, \qquad Z(x, \theta) = \sum_{\psi} h(\psi, x, \theta)$$

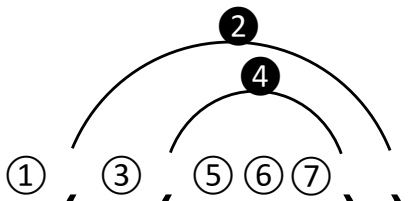
x CAUGCUAGCUAGUCUGCGUAGCUGCGUACUAGCGCGUACGUCGGAU

Pattern: .(.(...))

		$h(\psi, x, \theta)$
$\psi \in$	*******.(.(...))**********	0.6
	**********.(.(...))**********	0.5
	**********.(.(...))**********	0.5
	**********.(.(...))*****	0.4
	⋮	

SCFG for primary sequential profile

Pattern: `.(.(...))`



	A	C	G	U
①				
③	Base emission			
⑤				
⑥				
⑦				
*	background			

	CG	GC	AU	UA	GU	UG
❷	Base pair emission					
❹						

x CAUGCUAGCUAGUCUGCGUAGCUGCGUACUAGCGCGUACGUCGGAU

Pattern: `.(.(...))`

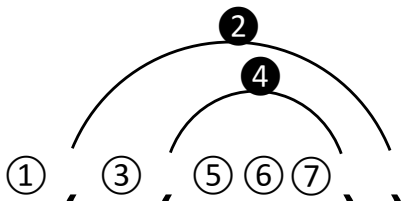
$h(\psi, x, \theta)$

$\psi \in$

```
*******.(.(...))*****************************  0.6
*******************.(.(...))******************  0.5
**********.(.(...))*****************************  0.5
*****************************************.(.(...))*****  0.4
                          ⋮
```

SCFG for primary sequential profile

Pattern: `.(.(...))`



	A	C	G	U
①				
③	Base emission			
⑤				
⑥				
⑦				
*	background			

	CG	GC	AU	UA	GU	UG
❷	Base pair emission					
❹						

Parameter θ

x CAUGCUAGCUAGUCUGCGUAGCUGCGUACUAGCGCGUACGUCGGAU

Pattern: `.(.(...))`

$h(\psi, x, \theta)$

$\psi \in$

```
*******.(.(...))*****************************  0.6
*******************.(.(...))******************  0.5
**********.(.(...))*****************************  0.5
*****************************************.(.(...))*****  0.4
```

⋮

SCFG for primary sequential profile

Pattern: `.(.(.*..))`

gap

SCFG for primary sequential profile

Pattern: `.(.(.*..))`

x `CAUGCUAGCUAGUCUGCGUAGCUGCGUACUAGCGCGUACGUCGGAU`

Pattern: `.(.(.*..))`

		$h(\psi, x, \theta)$
$\psi \in$	`*******.(.(.**********..))**********************`	0.6
	`**********************.(.(.***..))****************`	0.5
	`**********.(.(.*****..))**************************`	0.5
	`*****************************************.(.(.**..))***`	0.4
	⋮	

Combined model

Joint probability of ψ and σ, under given sequence x and pattern of interest

$$P(\psi, \sigma \mid x) = \frac{f(\psi, \sigma, x, \theta, \lambda)}{Z(x, \theta, \lambda)}, \qquad Z(x, \theta, \lambda) = \sum_{\psi} \sum_{\sigma} f(\psi, \sigma, x, \theta, \lambda)$$

$$f(\psi, \sigma, x, \theta, \lambda) = \mathbb{1}(\psi \circ \sigma)\, g(\sigma, x)^{\boldsymbol{\lambda}} h(\psi, x, \theta)$$

$\mathbb{1}(\psi \circ \sigma)$ is 1 if paired loci are consistent among ψ and σ, and 0 otherwise.

$\boldsymbol{\lambda}$ is a scaler, which can be also interpreted as “stability”

Combined model

Joint probability of ψ and σ, under given sequence x and pattern of interest

$$P(\psi, \sigma \mid x) = \frac{f(\psi, \sigma, x, \theta, \lambda)}{Z(x, \theta, \lambda)}, \qquad Z(x, \theta, \lambda) = \sum_{\psi} \sum_{\sigma} f(\psi, \sigma, x, \theta, \lambda)$$

$$f(\psi, \sigma, x, \theta, \lambda) = \mathbb{1}(\psi \circ \sigma)\, g(\sigma, x)^{\boldsymbol{\lambda}} h(\psi, x, \theta)$$

$\mathbb{1}(\psi \circ \sigma)$ is 1 if paired loci are consistent among ψ and σ, and 0 otherwise.

$\boldsymbol{\lambda}$ is a scaler, which can be also interpreted as "stability"

$$g(\sigma, x)^{\boldsymbol{\lambda}} = \exp\left(-\frac{\boldsymbol{\lambda}}{kT} \Delta G(\sigma, x)\right)$$

Objective function and parameter fitting

Probability of motif existence

$$
\begin{aligned}
P(\, z = 0 \mid x; \theta, \lambda \,) &= \sum_{\sigma} P(\psi_0, \sigma \mid x; \theta, \lambda) \\
&= \sum_{\sigma} \frac{f(\psi_0, \sigma, x, \theta, \lambda)}{Z(x, \theta, \lambda)} \\
&= \sum_{\sigma} \frac{f(\psi_0, \sigma, x, \theta, \lambda)}{\sum_{\psi} \sum_{\sigma'} f(\psi, \sigma', x, \theta, \lambda)}
\end{aligned}
$$

$$
P(\, z = 1 \mid x; \theta, \lambda \,) = 1 - P(\, z = 0 \mid x; \theta, \lambda \,)
$$

ψ_0: all bases are emitted by background state

$z \in \{0,1\}$: motif existence in a sequence

Objective function and parameter fitting

Probability of motif existence

$$P(z=0 \mid x;\theta,\lambda) = \sum_{\sigma} P(\psi_0,\sigma \mid x;\theta,\lambda)$$
$$= \sum_{\sigma} \frac{f(\psi_0,\sigma,x,\theta,\lambda)}{Z(x,\theta,\lambda)}$$
$$= \sum_{\sigma} \frac{f(\psi_0,\sigma,x,\theta,\lambda)}{\sum_{\psi}\sum_{\sigma'} f(\psi,\sigma',x,\theta,\lambda)}$$

$$P(z=1 \mid x;\theta,\lambda) = 1 - P(z=0 \mid x;\theta,\lambda)$$

ψ_0: all bases are emitted by background state

$z \in \{0,1\}$: motif existence in a sequence

Log likelihood over
Positive / negative
sequences

$$\mathcal{L}(\theta,\lambda) = \sum_{x^+} \ln P(z=1 \mid x^+;\theta,\lambda) + \sum_{x^-} \ln P(z=0 \mid x^-;\theta,\lambda)$$
$$= \mathcal{L}^+ + \mathcal{L}^-$$

Objective function and parameter fitting

Task:

$$\arg_{\theta,\lambda} \max \mathcal{L}(\theta, \lambda)$$

We can calculate exact derivation $\left(\frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \theta}, \frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \lambda}\right)$ by nested inside-outside algorithm (ref. Sankoff's algorithm)

$$\frac{\partial L(\theta, \lambda)}{\partial \theta_{ij}} = \sum_{x^+} \overbrace{E\left[\mathcal{N}_{ij}(\psi)\right] \ldots}^{\text{Expected value calculation DP over 2 SCFGs traversal !!}}$$

$$\frac{\partial L(\theta, \lambda)}{\partial \lambda} = \sum_{x^-} E\left[\mathcal{G}(\sigma)\right] \ldots$$

Zuker, M., & Sankoff, D. (1984). RNA secondary structures and their prediction. *Bulletin of mathematical biology*, *46*(4), 591-621.

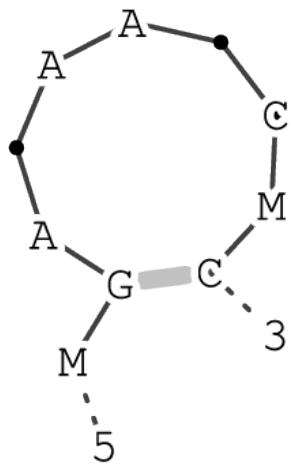
Select the best pattern after optimization

1. Enumerate 2D structural patterns
2. Optimize parameter θ and λ for each pattern
3. Model selection (k-fold cross validation) to select the best pattern

Select the best pattern after optimization

1. Enumerate 2D structural patterns
2. Optimize parameter θ and λ for each pattern
3. Model selection (k-fold cross validation) to select the best pattern

Output example



Conclusion

- We formulated a combined motif model of
$$\text{primary sequence} \times \text{secondary structure}$$
- Traversing all the RNA 2D structural space enabled more precise prediction of the local structural motif
- The new model can assess the “local stability” of the 2D structure at binding region
- Several validated structural motifs were reproduced

Acknowledgements

- The University Of Tokyo
  - Prof. Hisanori Kiryu
  - Prof. Kiyoshi Asai
  - Lab members
- AI Research Center (AIST)
  - Dr. Risa Kawaguchi

科研費 KAKENHI

GSDM