Non-redundant sampling and estimating properties of RNA secondary structures

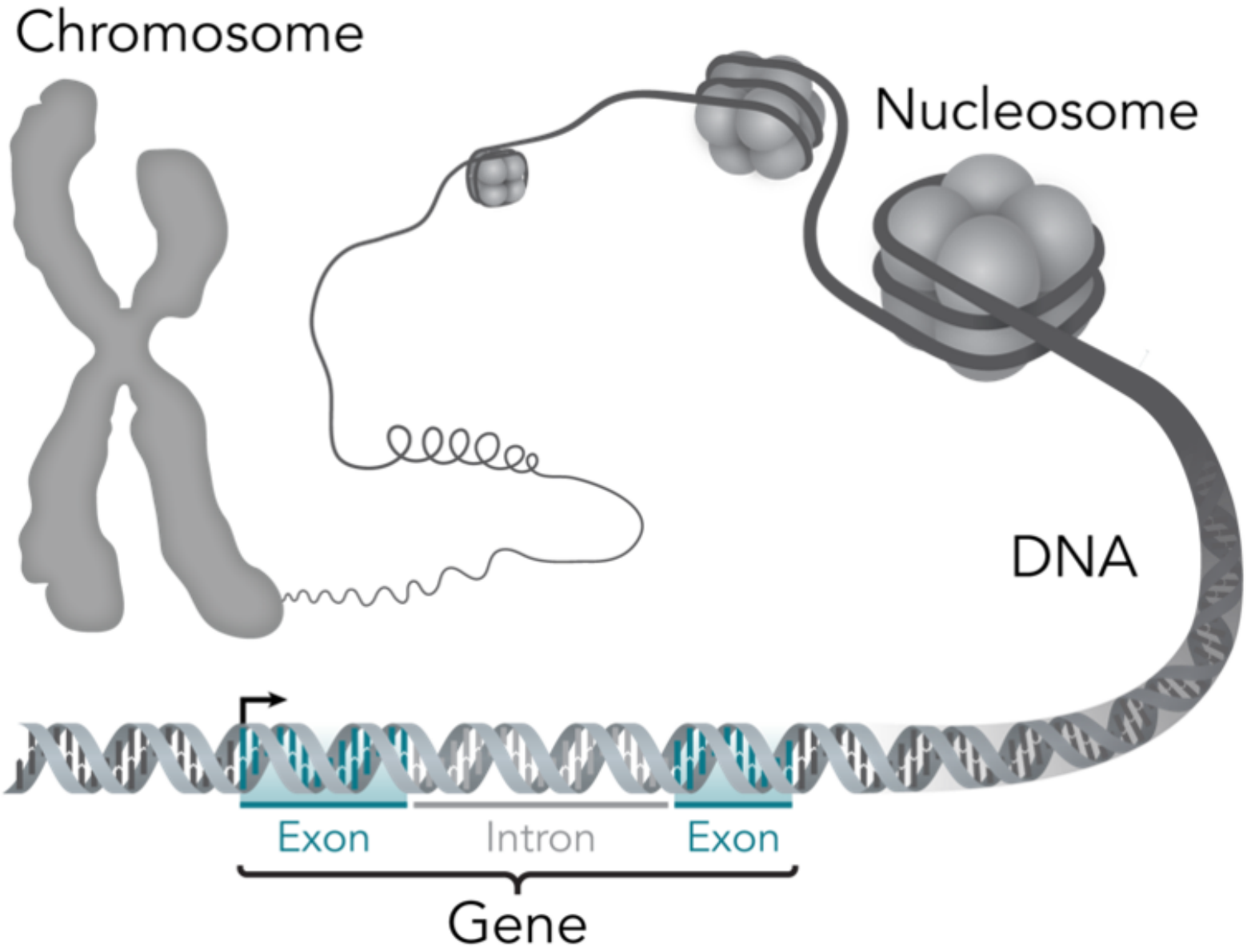
ALEA 2018

Juraj Michálik[1] Yann Ponty[1] **Christelle Rovetta**[1,2]

[1] Equipe AMIB, Laboratoire LIX - [2] BioInfo Group LRI - RNALands

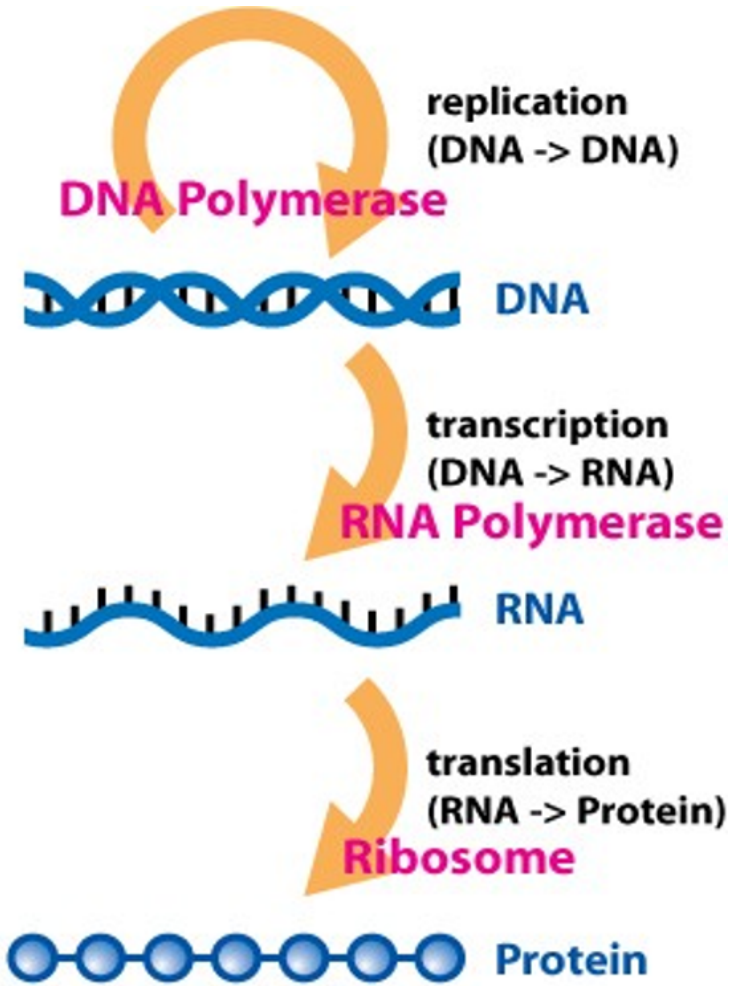
12th March 2018

Chromosome DNA Gene



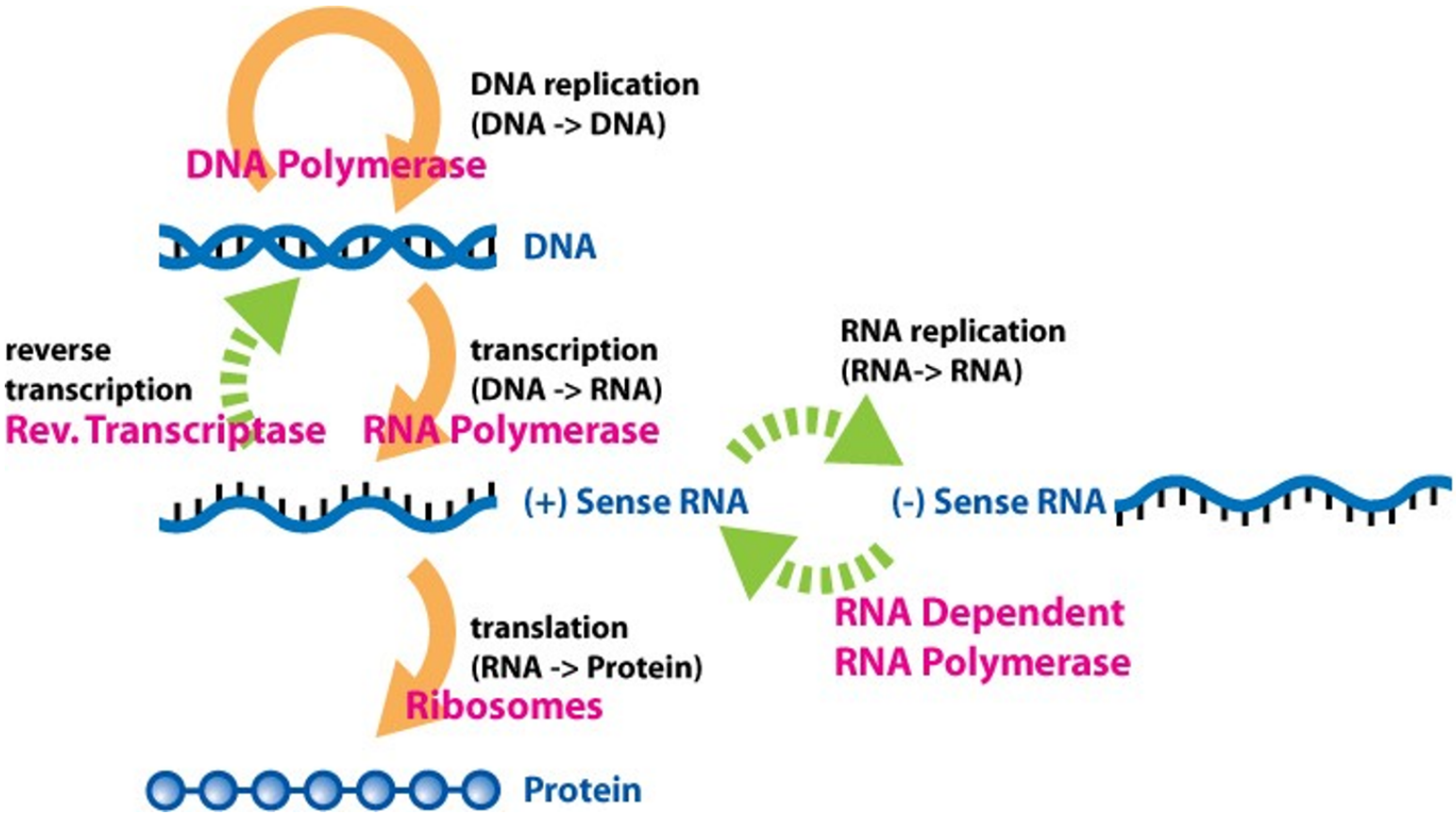
Central Dogma

- Francis Crick in 1958



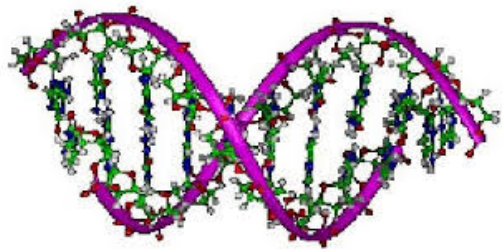
Central Dogma

- Francis Crick in 1958
- He got this one somewhat wrong...
- The RNA World?



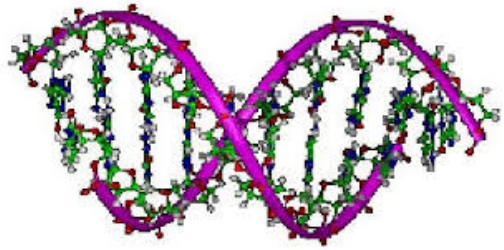
DNA and RNA

- DNA is made up of Adenine (A), Guanine (G), Cytosine (C), and Thymine (T)
  - Base pairs (Watson-Crick) : A-T, G-C

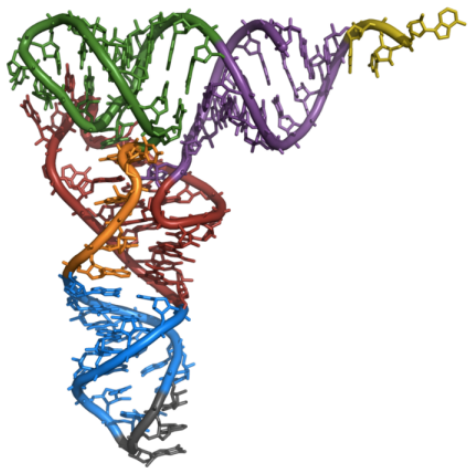


DNA and RNA

- DNA is made up of Adenine (A), Guanine (G), Cytosine (C), and Thymine (T)
  - Base pairs (Watson-Crick) : A-T, G-C



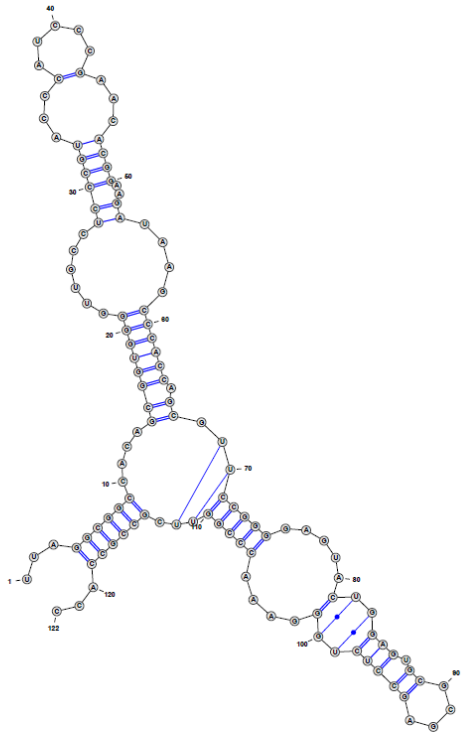
- RNA is made up of A, G, C, and Uracil (U)
  - Base pairs : A-U, G-C, G-U (Wobble)



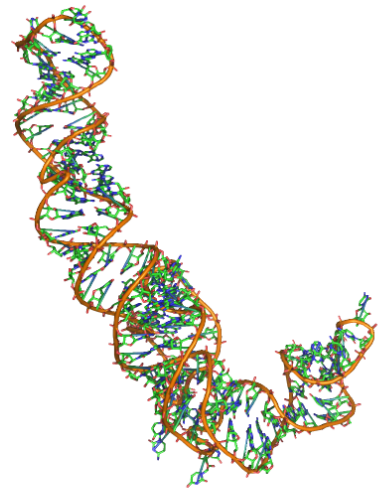
RNA representation

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```

Primary structure



Secondary structure



Tertiary structure

Source: 5s rRNA (PDB 1K73:B)

Secondary structure

- $BP = \{A-U, G-C, G-U\}$
- **Valid base pairs** for a sequence seq :

$$\mathcal{P}(\text{seq}) := \{(i,j) \mid j - i > 3 \text{ and } (\text{seq}[i], \text{seq}[j]) \in BP\}$$

- Example:

Secondary structure

Secondary structure

- $BP = \{A-U, G-C, G-U\}$
- **Valid base pairs** for a sequence seq :

$$\mathcal{P}(\text{seq}) := \{(i,j) \mid j - i > 3 \text{ and } (\text{seq}[i], \text{seq}[j]) \in BP\}$$

- Example:

Secondary structure

1. $s \subseteq \mathcal{P}(\text{seq})$

Secondary structure

- $BP = \{A - U, G - C, G - U\}$
- **Valid base pairs** for a sequence seq :

$$\mathcal{P}(\text{seq}) := \{(i, j) \mid j - i > 3 \text{ and } (\text{seq}[i], \text{seq}[j]) \in BP\}$$

- Example:

Secondary structure

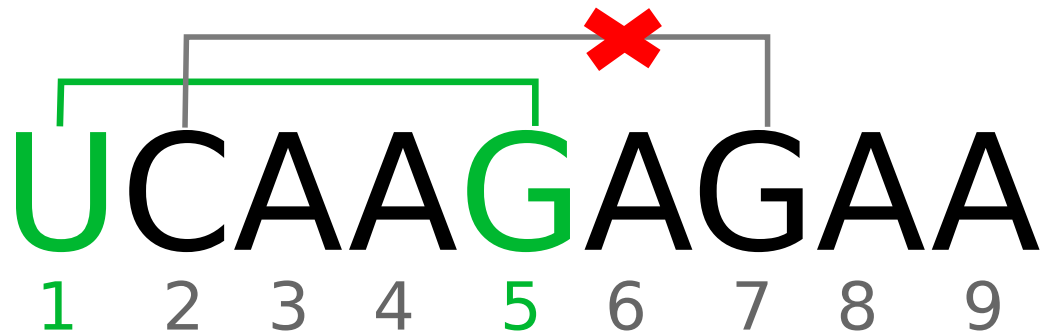
1. $s \subseteq \mathcal{P}(\text{seq})$

Secondary structure

- $BP = \{A-U, G-C, G-U\}$
- **Valid base pairs** for a sequence seq :

$$\mathcal{P}(\text{seq}) := \{(i,j) \mid j - i > 3 \text{ and } (\text{seq}[i], \text{seq}[j]) \in BP\}$$

- Example:



Secondary structure

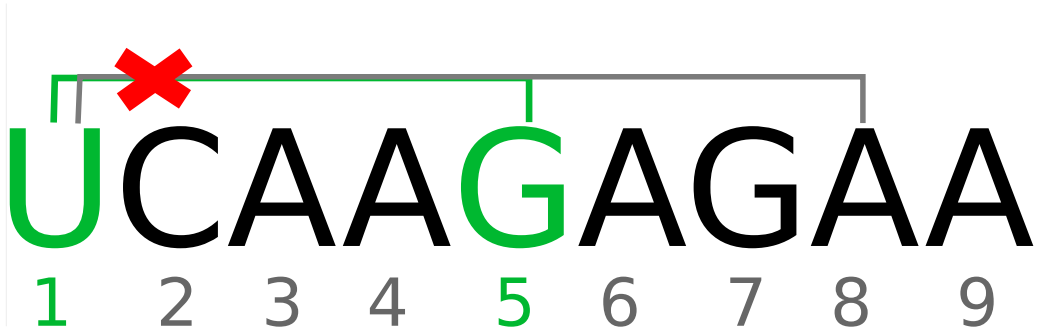
1. $s \subseteq \mathcal{P}(\text{seq})$
2. No crossing between bases pairs

Secondary structure

- $BP = \{A-U, G-C, G-U\}$
- **Valid base pairs** for a sequence seq :

$$\mathcal{P}(\text{seq}) := \{(i,j) \mid j - i > 3 \text{ and } (\text{seq}[i], \text{seq}[j]) \in BP\}$$

- Example:



Secondary structure

1. $s \subseteq \mathcal{P}(\text{seq})$
2. No crossing between bases pairs
3. Each base is at most involved in one base pair

Boltzmann Model

- $\Omega(\text{seq})$ set of all secondary structures

$$\mathbb{P}(s) := \frac{e^{-\beta E(s)}}{Q} \quad \text{where} \quad Q = \sum_{v \in \Omega(\text{seq})} e^{-\beta E(v)}.$$

  - $E(s)$ free energy of the structure s
  - $\beta = -\frac{1}{RT}$
  - R Boltzmann constant, T absolute temperature
  - Q the partition function

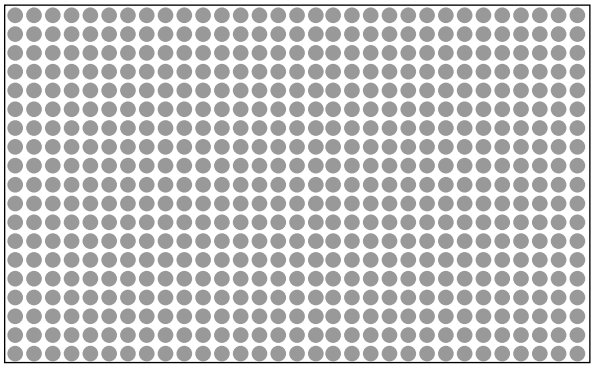
- **Dynamic programming** (DP) to compute Q in $\Theta(n^3)$

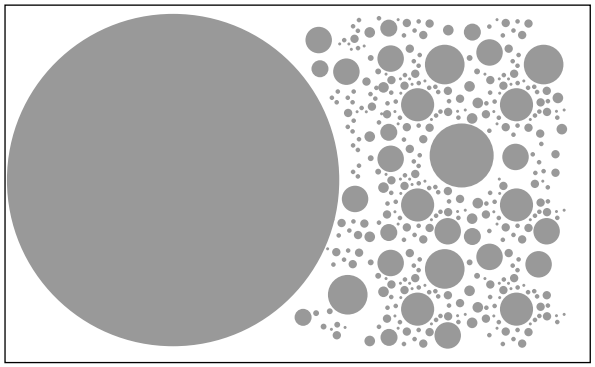
Motivations

- Sampling not a big deal using DP, but ...

Motivations

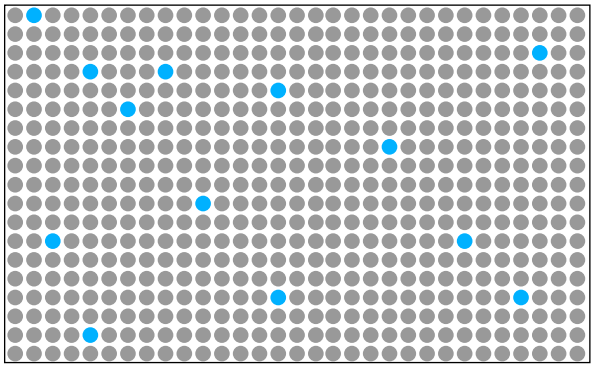
- Sampling not a big deal using DP, but ...

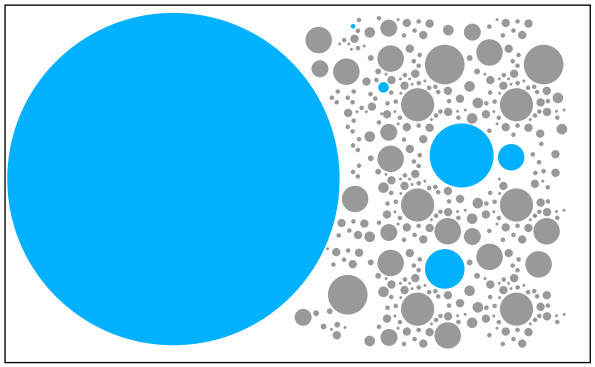




Motivations

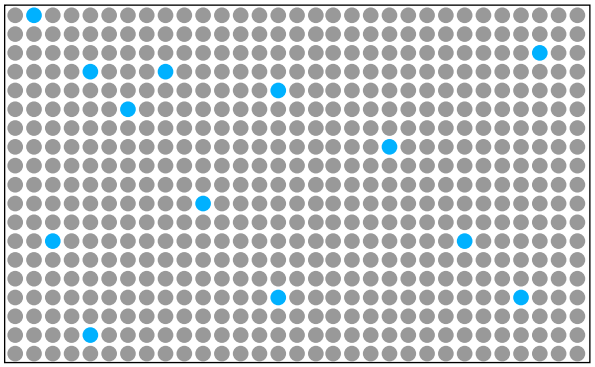
- Sampling not a big deal using DP, but ...

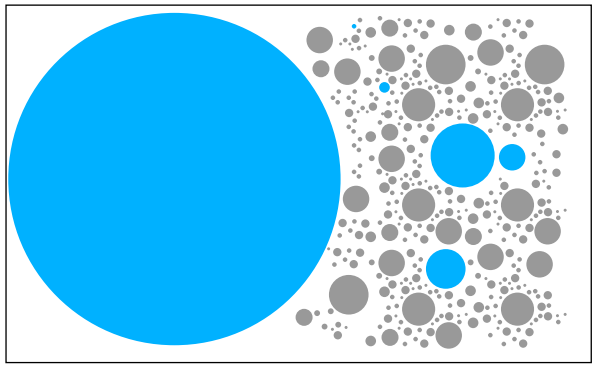




Motivations

- Sampling not a big deal using DP, but ...





- Non-redundant sampling [Lorenz, Ponty, '13]
  - $\mathcal{H}$ the set of samples already generated

$$\mathbb{P}_{\mathcal{H}}(s) = \begin{cases} \frac{e^{-\beta E(s)}}{Q - \sum_{s' \in \mathcal{H}} e^{-\beta E(s)}} & s \notin \mathcal{H} \\ 0 & \text{if otherwise.} \end{cases}$$

Recursive equation

- Construction of Ω using $\Omega[i:j] := \Omega(\text{seq}[i:j])$

$$\Omega[i:j] = \begin{cases} \emptyset & \text{if } j - i < 3 \\ \Omega[i+1:j] \bigcup_{k} \bigcup_{(i,k)\in\mathcal{P}} \{(i,k)\} \times \Omega[i+1:k-1] \times \Omega[k+1:j] & \text{otherwise.} \end{cases}$$

Recursive equation

- Construction of Ω using $\Omega[i:j] := \Omega(\text{seq}[i:j])$

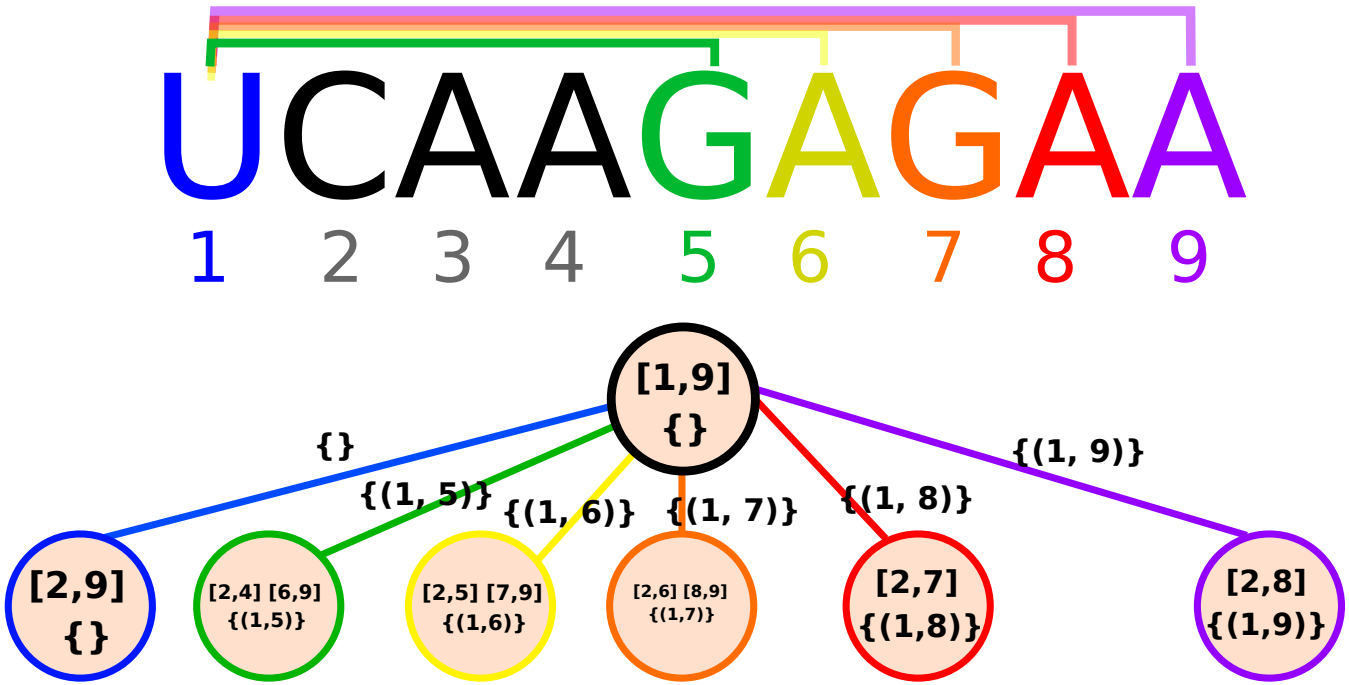
$$\Omega[i:j] = \begin{cases} \emptyset & \text{if } j - i < 3 \\ \Omega[i+1:j] \bigcup_{k} \bigcup_{(i,k)\in\mathcal{P}} \{(i,k)\} \times \Omega[i+1:k-1] \times \Omega[k+1:j] & \text{otherwise.} \end{cases}$$

Tree of secondary structures

- Recursive method, random walk in a tree

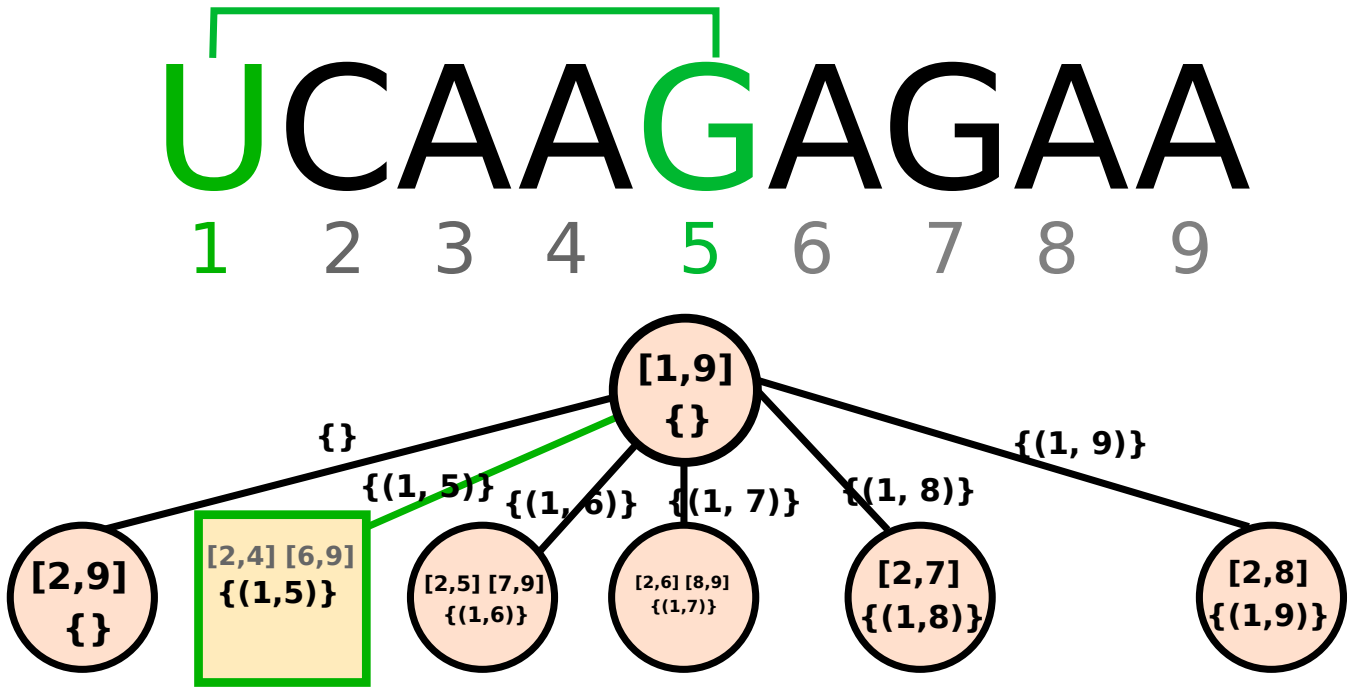
Tree of secondary structures

- Recursive method, random walk in a tree
- Tree construction



Tree of secondary structures

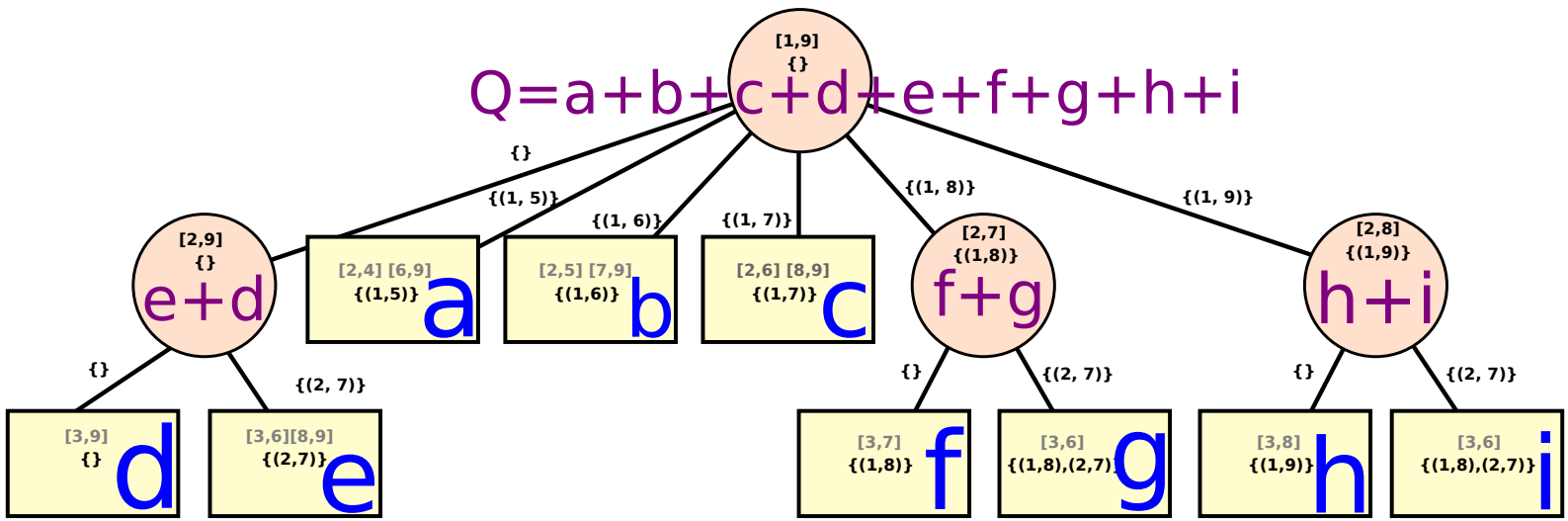
- Recursive method, random walk in a tree
- Tree construction



Tree of secondary structures

- Probability of $s \in \Omega$:

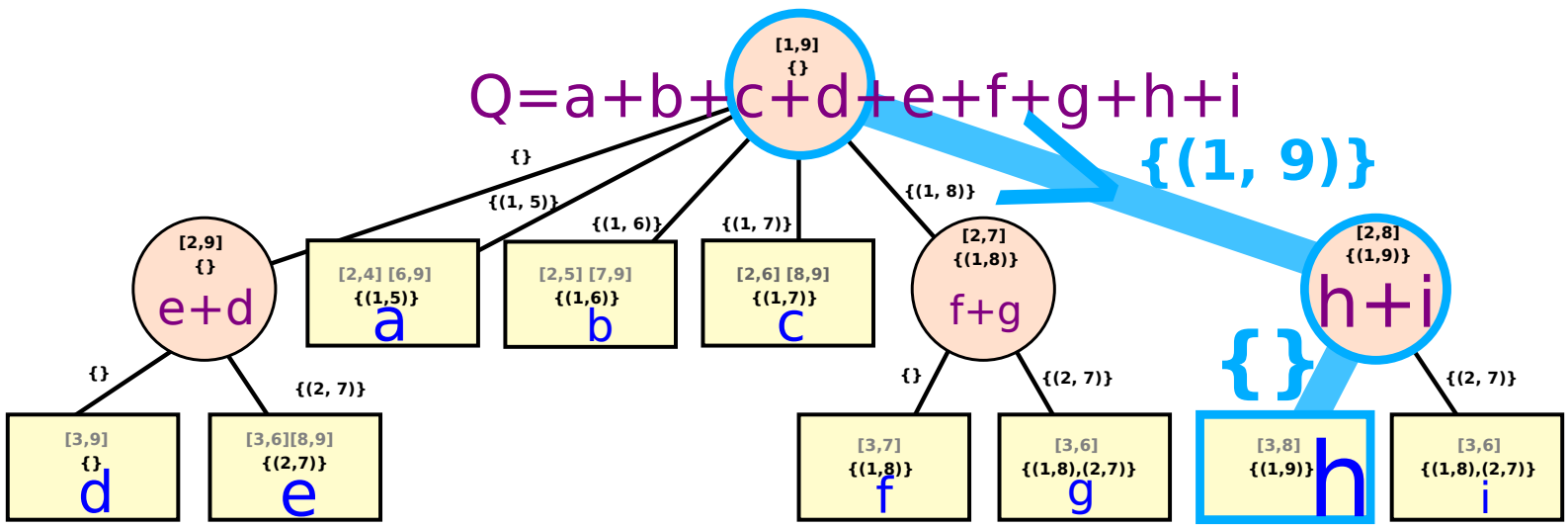
$$\mathbb{P}(s) := \frac{e^{-\beta E(s)}}{Q} \quad \text{where} \quad Q = \sum_{v \in \Omega} e^{-\beta E(v)}.$$



(Redundant) Sampling

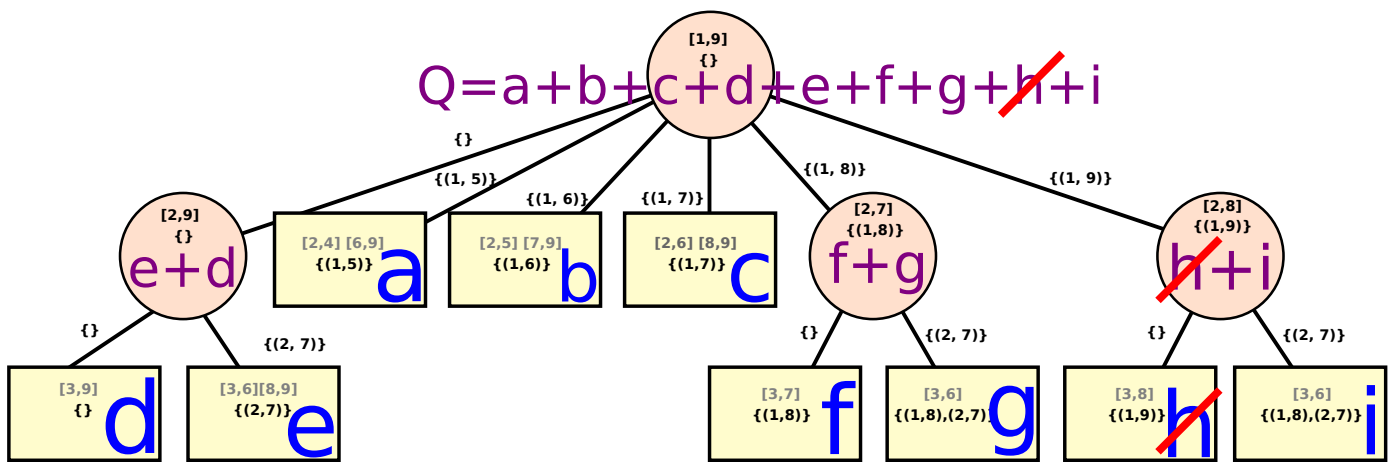
- Probability of $s = \{(1, 9)\}$:

$$p_s = \frac{h+i}{a+\ldots+i} \times \frac{h}{h+i} = \frac{h}{a+\ldots+i} = \frac{e^{-\beta E(s)}}{Q} := \mathbb{P}(s)\,.$$



Non-redundant sampling

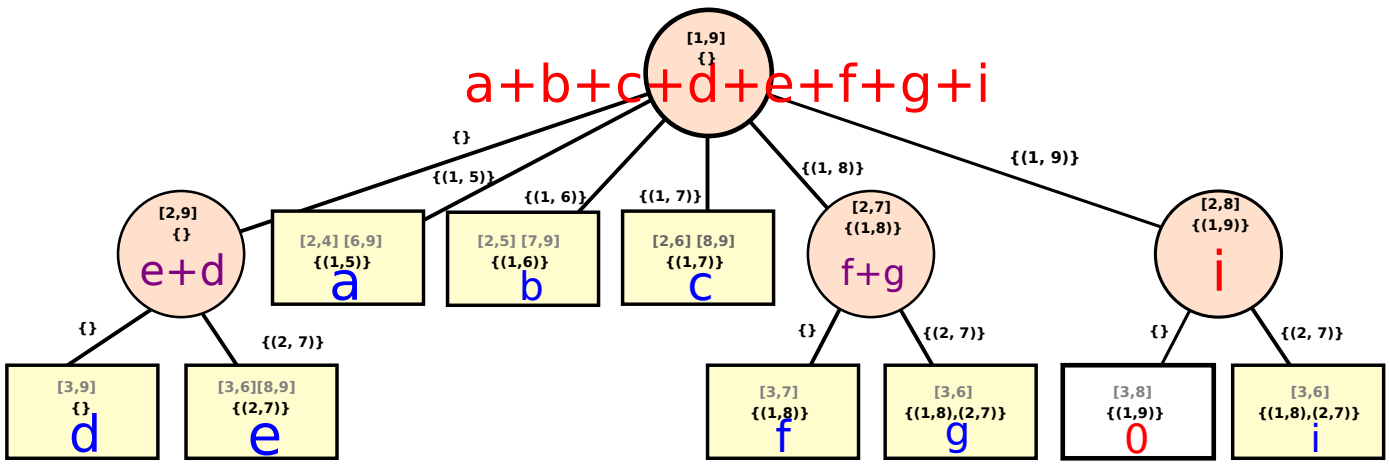
- RNA Non-redundant sampling [Michálik et al., 17]
- $\mathcal{H} = \{v_1\}$, $v_1 = \{(1, 9)\}$



Non-redundant sampling

- RNA Non-redundant sampling [Michálik et al., 17]
- $\mathcal{H} = \{v_1\}$, $v_1 = \{(1, 9)\}$
- Probability of $v_2 = \{(1, 8), (2, 7)\}$:

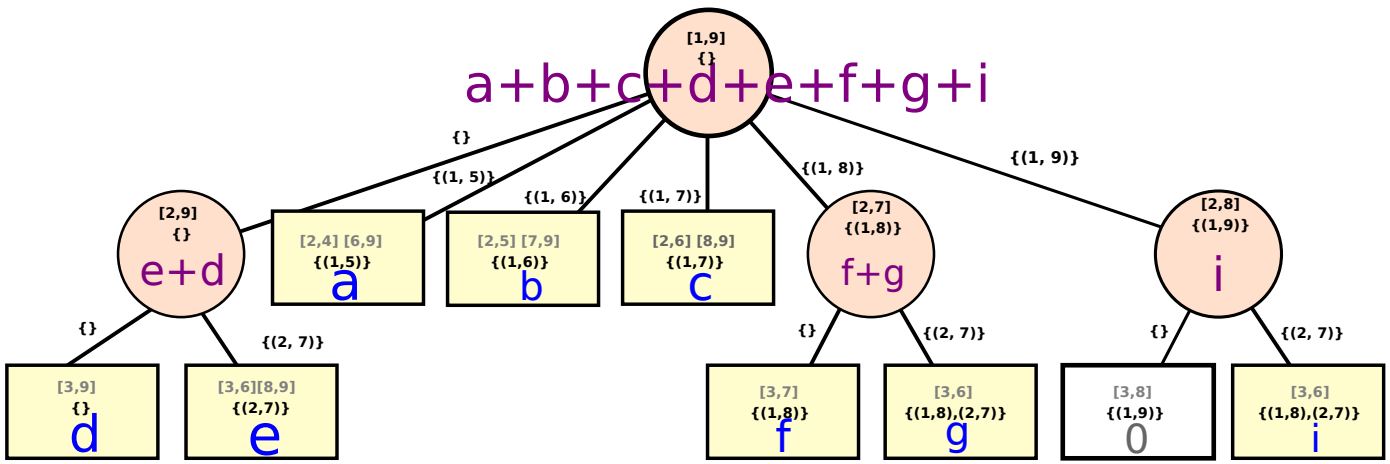
$$p_{v2} = \frac{f + g}{a + \ldots + f + g + i} \times \frac{g}{f + g} = \frac{e^{-\beta E(v_2)}}{Q - e^{-\beta E(v_1)}} := \mathbb{P}_{\mathcal{H}}(v_2)$$



Non-redundant sampling

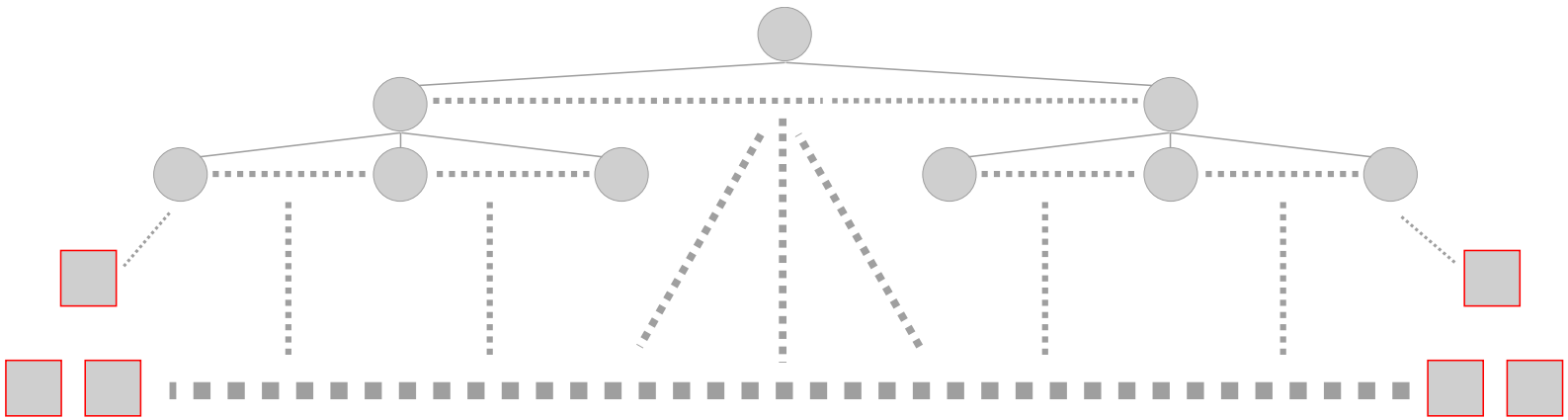
- RNA Non-redundant sampling [Michálik et al., 17]
- $\mathcal{H} = \{v_1\}$, $v_1 = \{(1, 9)\}$
- Probability of $v_2 = \{(1, 8), (2, 7)\}$:

$$p_{v2} = \frac{f + g}{a + \ldots + f + g + i} \times \frac{g}{f + g} = \frac{e^{-\beta E(v_2)}}{Q - e^{-\beta E(v_1)}} := \mathbb{P}_{\mathcal{H}}(v_2)$$



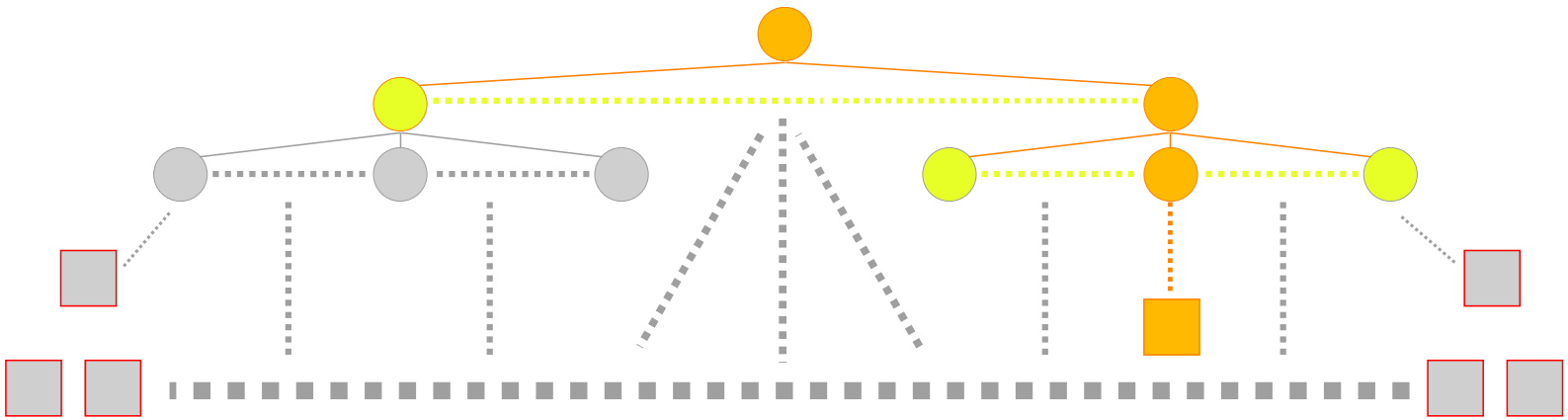
Non-redundant (and redundant) Sampling

- $|\Omega|$ huge
- Tree construction as the random generation proceeds:



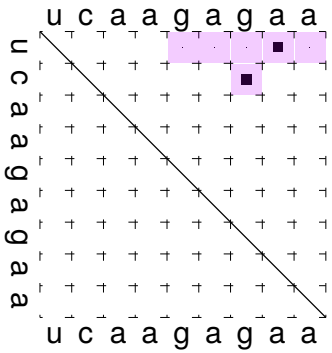
Non-redundant (and redundant) Sampling

- $|\Omega|$ huge
- Tree construction as the random generation proceeds:



Classical approach

- Feature function $F : \Omega \to \mathbb{Y}$
  - Example: probability of the base pair (i, j) ?



(`ViennaRNA`)

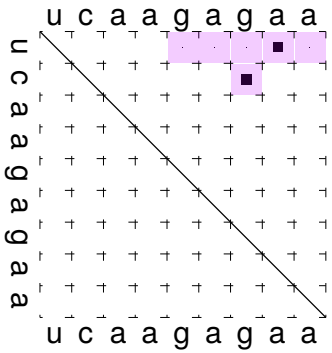
$$F_{i,j}(s) = \begin{cases} 1 & \text{if } (i,j) \in s \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}\left(F_{i,j}\right) = \mathbb{P}\left((i,j)\right)$$

- **GOAL**: estimate $\mathbb{E}\left(F(X)\right) := \sum_{s \in \Omega} \mathbb{P}\left(s\right) F(s)$ ($|\Omega|$ huge)

Classical approach

- Feature function $F : \Omega \to \mathbb{Y}$
  - Example: probability of the base pair (i, j) ?



(`ViennaRNA`)

$$F_{i,j}(s) = \begin{cases} 1 & \text{if } (i,j) \in s \\ 0 & \text{otherwise.} \end{cases}$$

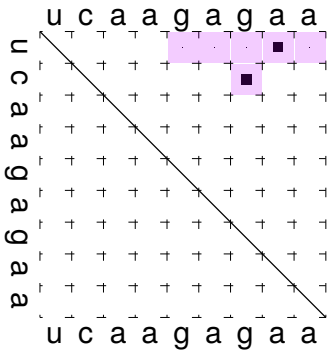
$$\mathbb{E}\left(F_{i,j}\right) = \mathbb{P}\left((i,j)\right)$$

- **GOAL**: estimate $\mathbb{E}\left(F(X)\right) := \sum_{s \in \Omega} \mathbb{P}(s) F(s)$ ($|\Omega|$ huge)
- **Empirical mean** : $(s_1, s_2, \ldots, s_m) \in \Omega^m$

$$\widehat{F}(\mathbf{s}) = \frac{1}{m} \sum_{k=1}^{m} F(s_k)$$

Classical approach

- Feature function $F : \Omega \to \mathbb{Y}$
  - Example: probability of the base pair (i, j) ?



(`ViennaRNA`)

$$F_{i,j}(s) = \begin{cases} 1 & \text{if } (i,j) \in s \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}\left(F_{i,j}\right) = \mathbb{P}\left((i,j)\right)$$

- **GOAL**: estimate $\mathbb{E}\left(F(X)\right) := \sum_{s \in \Omega} \mathbb{P}(s) F(s)$ ($|\Omega|$ huge)
- **Empirical mean** : $(s_1, s_2, \ldots, s_m) \in \Omega^m$

$$\widehat{F}(\mathbf{s}) = \frac{1}{m} \sum_{k=1}^{m} F(s_k)$$

- m' distinct structures:

$$\widehat{F}(\mathbf{s}) = \frac{1}{m} \sum_{k=1}^{m} F(s_k) = \sum_{n=1}^{m'} \frac{a_n}{m} F(s_n).$$

Non-redundant estimator

- Estimate: $\mathbb{E}(F(X)) := \sum_{s \in \Omega} \mathbb{P}(s) F(s)$
  - Non-redundant samples : $(v_1, v_2, \ldots, v_m) \in \Omega^m$
  - Exploit knowledge of $\mathbb{P}(v_k)$ $\left(\mathbb{P}(s) = \frac{e^{-\beta E(s)}}{Q}\right)$

Non-redundant estimator

- Estimate: $\mathbb{E}(F(X)) := \sum_{s\in\Omega} \mathbb{P}(s) F(s)$
  - Non-redundant samples : $(v_1, v_2, \ldots, v_m) \in \Omega^m$
  - Exploit knowledge of $\mathbb{P}(v_k)$ $\left(\mathbb{P}(s) = \frac{e^{-\beta E(s)}}{Q}\right)$
- **Estimator**:

$$\widetilde{F}(\mathbf{v}) = \frac{1}{m}\left(\sum_{i=1}^{m} F(v_i)\left(1 - \sum_{v\in\mathcal{H}_{i-1}} \mathbb{P}(v) + (m-i)\mathbb{P}(v_i)\right)\right)$$

  - $\mathcal{H}_0 = \emptyset$
  - $\forall 1 \leq i \leq m, \mathcal{H}_i := \{v_1, \ldots, v_i\}$

Estimators

- Empirical mean: $\widehat{F}(\mathbf{s}) = \frac{1}{m} \sum_{k=1}^{m} F(s_k)$
- Non-redundant:

$$\widetilde{F}(\mathbf{v}) = \frac{1}{m} \left(\sum_{i=1}^{m} F(v_i) \left(1 - \sum_{v \in \mathcal{H}_{i-1}} \mathbb{P}(v) + (m-i)\mathbb{P}(v_i) \right) \right)$$

Properties

The non-redundant estimator $\widetilde{F}$ has the following properties

1. It is unbiased, *i.e* $\mathbb{E}\left(\widetilde{F}(\tilde{\mathbf{X}})\right) = \mathbb{E}(F(X))$.
2. It is more efficient than $\widehat{F}$ for the same number of samples, *i.e* $\mathbb{V}\left(\widetilde{F}(\tilde{\mathbf{X}})\right) \leq \mathbb{V}\left(\widehat{F}(\mathbf{X})\right)$.

Base pairs

- $(i,j) \in \mathcal{P}$, $F_{i,j} : \Omega \to \mathbb{N}$

$$F_{i,j}(s) = \begin{cases} 1 & \text{if } (i,j) \in s \\ 0 & \text{otherwise.} \end{cases}$$

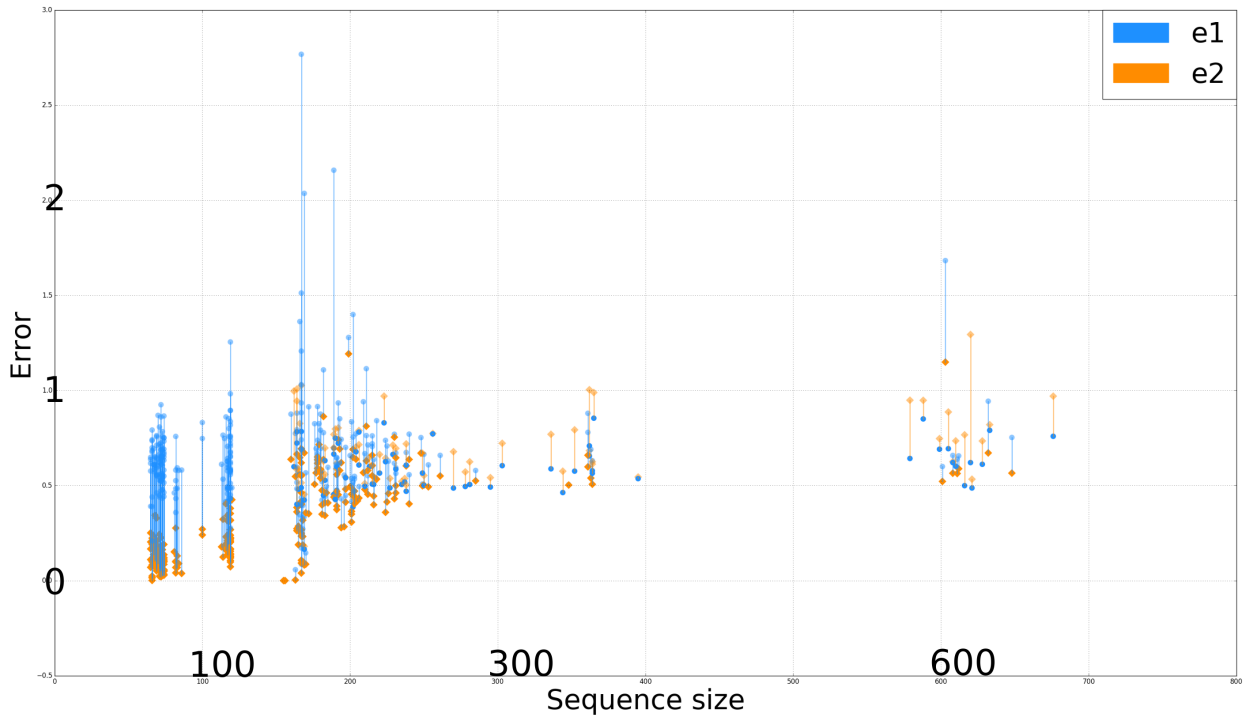
- For all $(i,j) \in \mathcal{P}$, $\mathbb{E}\left(F_{i,j}(X)\right)$ computed using DP
- Error of estimators:
  - $e_1 = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \frac{\widehat{F_{i,j}}(\mathbf{s}) - \mathbb{E}(F_{i,j}(X))}{\mathbb{E}(F_{i,j}(X))}$
  - $e_2 = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \frac{\widetilde{F_{i,j}}(\mathbf{v}) - \mathbb{E}(F_{i,j}(X))}{\mathbb{E}(F_{i,j}(X))}$

Results

- Dataset: *RF*00001, *RF*00005, *R*00061, *RF*00174, *RF*01071 and *RF*01731 (RFAM families)
  - Total 365 sequences
  - Number of samples: 10000 ($|\mathbf{s}| = |\mathbf{v}| = 10000$)

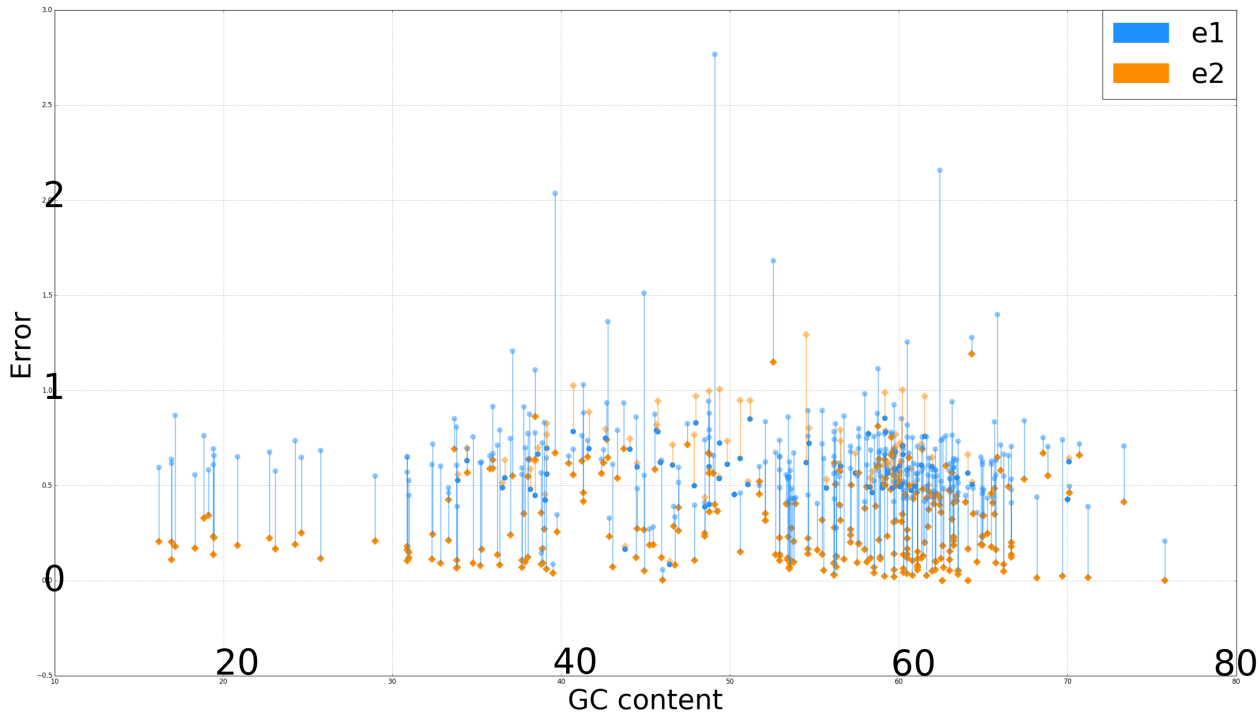
Results

- Dataset: *RF*00001, *RF*00005, *R*00061, *RF*00174, *RF*01071 and *RF*01731 (RFAM families)
  - Total 365 sequences
  - Number of samples: 10000 ($|\mathbf{s}| = |\mathbf{v}| = 10000$)
- $\widehat{F_{i,j}}(\mathbf{s})$ better than $\widetilde{F_{i,j}}(\mathbf{v})$ in 83% of cases !



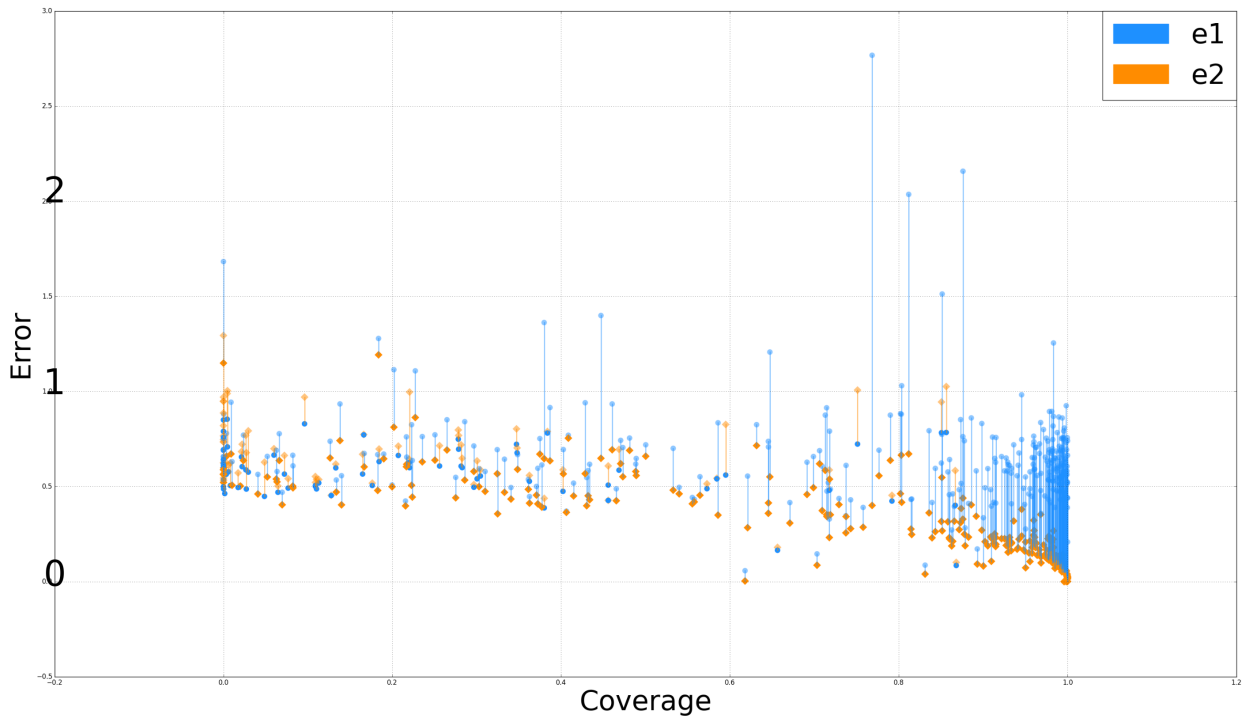
Results

- Dataset: *RF*00001, *RF*00005, *R*00061, *RF*00174, *RF*01071 and *RF*01731 (RFAM families)
  - Total 365 sequences
  - Number of samples: 10000 ($|\mathbf{s}| = |\mathbf{v}| = 10000$)
- $\widehat{F_{i,j}}(\mathbf{s})$ better than $\widetilde{F_{i,j}}(\mathbf{v})$ in 83% of cases !



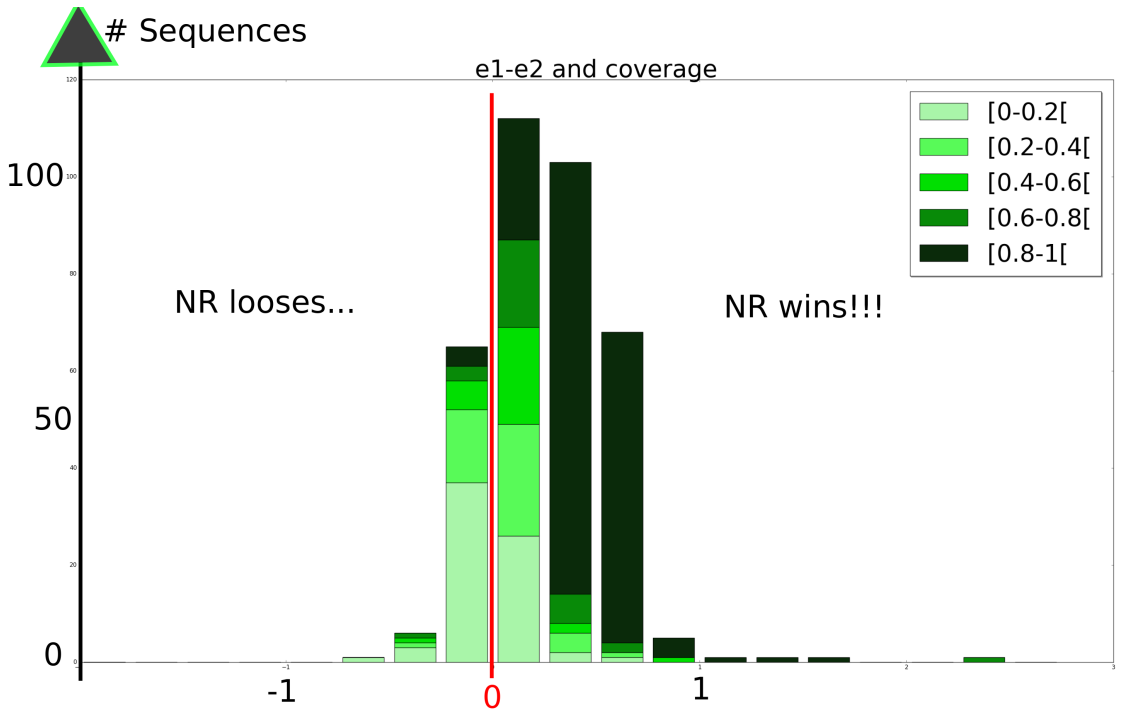
Results

- Dataset: *RF*00001, *RF*00005, *R*00061, *RF*00174, *RF*01071 and *RF*01731 (RFAM families)
  - Total 365 sequences
  - Number of samples: 10000 ($|\mathbf{s}| = |\mathbf{v}| = 10000$)
- $\widehat{F_{i,j}}(\mathbf{s})$ better than $\widetilde{F_{i,j}}(\mathbf{v})$ in 83% of cases !
- **Coverage**: $Cov(\mathcal{H}) = \frac{\sum_{v \in \mathcal{H}} e^{-\beta E(v)}}{Q}$ $(0 \leq Cov(\mathcal{H}) \leq 1)$



Results

- Dataset: *RF*00001, *RF*00005, *R*00061, *RF*00174, *RF*01071 and *RF*01731 (RFAM families)
  - Total 365 sequences
  - Number of samples: 10000 ($|\mathbf{s}| = |\mathbf{v}| = 10000$)
- $\widehat{F_{i,j}}(\mathbf{s})$ better than $\widetilde{F_{i,j}}(\mathbf{v})$ in 83% of cases !
- **Coverage**: $Cov(\mathcal{H}) = \frac{\sum_{v \in \mathcal{H}} e^{-\beta E(v)}}{Q}$ $(0 \leq Cov(\mathcal{H}) \leq 1)$



Conclusion

- Estimator for non redundant samples
- Quality of the estimator depends on coverage
- Implementation NR sampling Juraj Michálik (`ViennaRNA`)
  - NR sampling a little bit slower than "classical" one for the same number of samples (25%)
  - Most of cases, best coverage for the same execution time
- Applications in collaboration with our Austrian partners in the **RNALands** project

Conclusion

- Estimator for non redundant samples
- Quality of the estimator depends on coverage

- Implementation NR sampling Juraj Michálik (`ViennaRNA`)
  - NR sampling a little bit slower than "classical" one for the same number of samples (25%)
  - Most of cases, best coverage for the same execution time
- Applications in collaboration with our Austrian partners in the **RNALands** project
- Merci !