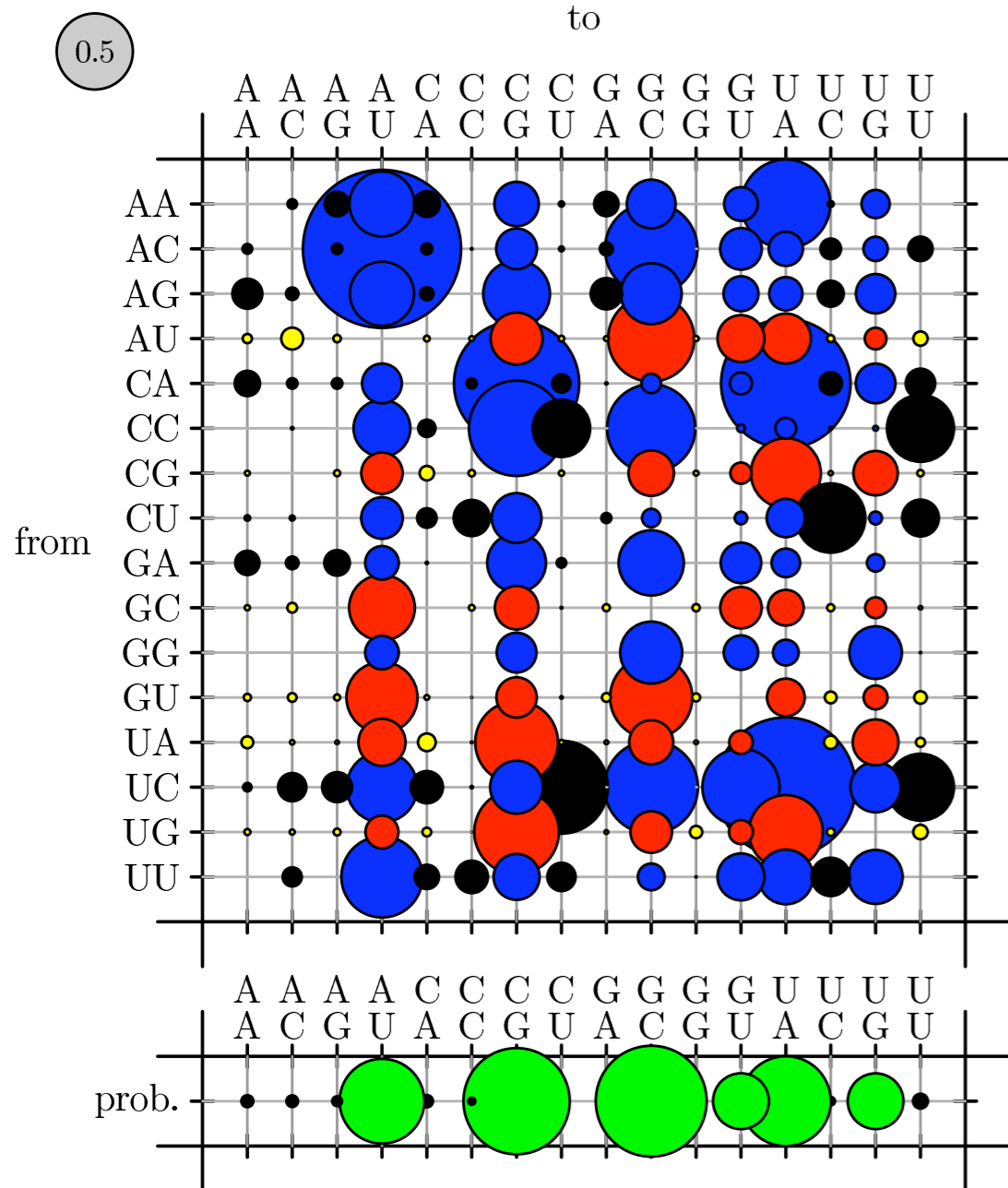
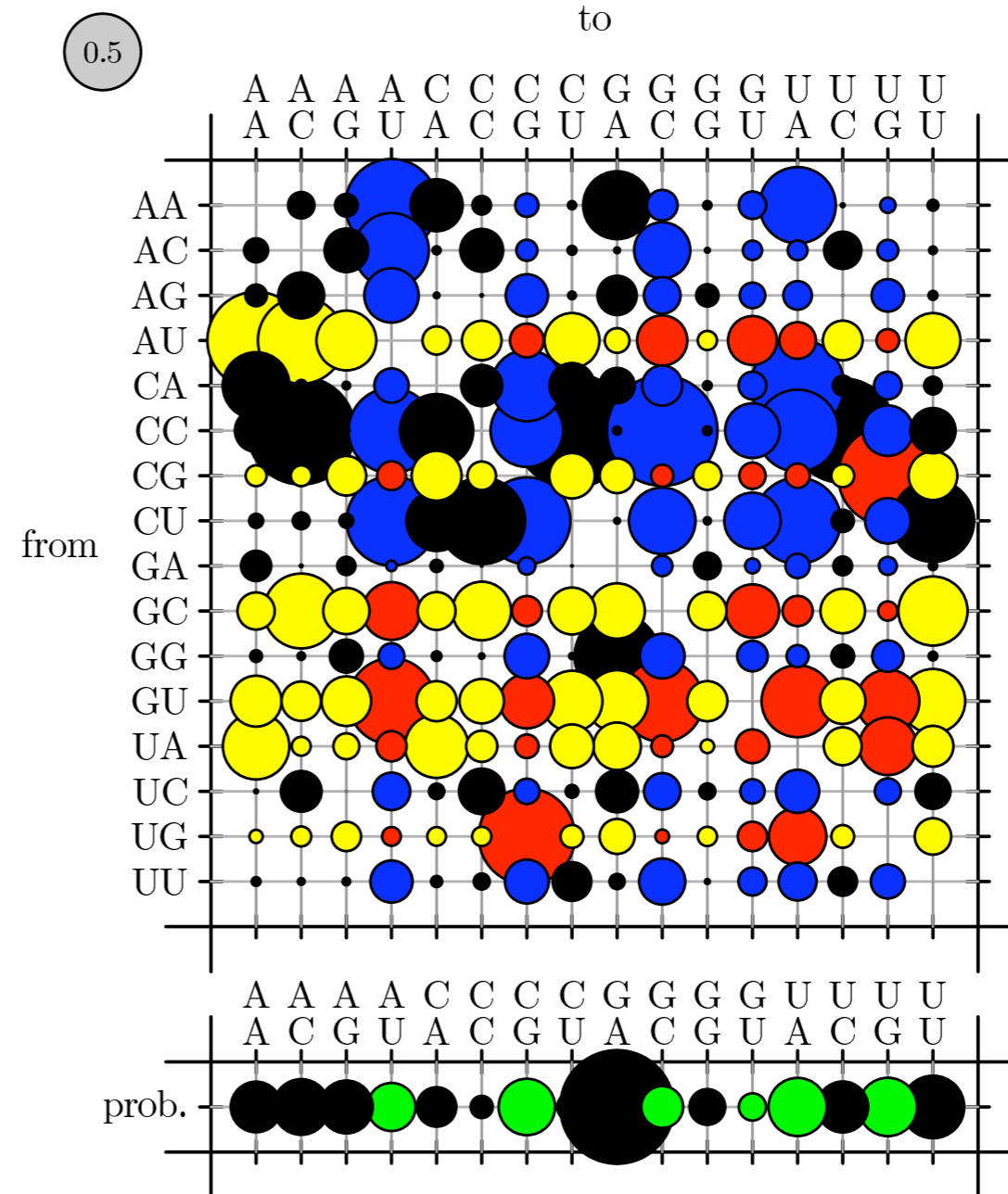


# Discovery, alignment & reconstruction of ncRNAs

Ian Holmes



**Base pairs in stems**



**Loop-closing base pairs**

A C C C  
G U  
12Y X5  
C G  
G C  
U A  
G C

# False positives

Features of real genome alignments: low-complexity DNA, repetitive sequence, microsatellite expansions, big indels...

```
UUUAAACU-----CCUAGAAGAACUAGAAGCUCUGGAAG---CUUCU-UAAGCGUUU-----UAU-----AUAUCAGAAU-AUAUUGCGCGUCAA---
AUCCAGCA-----GGAGCAGAGGCUCUGGCAG---CUUUU-GCAGCG-UU-----UAU-----AUAACGAGAAU-AUAUACGCCUCCG---
-----AAUUGCUGAACUCUGGCAGUUU--CUCGGGAUGUUUUAUUAACAGAUUAUUA-----AUCCACUC
AUCCAGCU-----CU-----GGAGCAGAGGCUCUGGCAG---CUUUU-GCAGCG-UU-----UAU-----AUAACAUGAAU-AUAUACGCAUCCG---
-----UGUGGGAGAAGCUCUGGUAGUUU--UUCA--UACGUGUACAUAUCUAUAUUAUUAUAGCUAUAUUGUAUAUAUCGAUUAUAUUAUAUA--UAUGUAUACGUGAUCCACU
GUCUAGCCGAAGAACCCACACGGA-----ACUCUGGCAG---UUUUAUCUGCG-CG-----UAU-----AUAGC--UUUAU-AUAUUGCGAUUCCG---
AUCCAGCU-----GGAGCAGAGGCUCUGGCAG---CUUUU-GCAGCG-UU-----UAU-----AUAACAAGAAU-AUAUACGCAUCCG---
-----AGUGGAAGAAGCUCUGGCAGCUU--UUUA--AGCGUUUAUAUAAGAGUUUAUA-----UAUGCGCGUCCACAC
AUCCAGCA-----GGAGCAGAAGCUCUGGAAG---CUUUU-GUAGCG-UU-----UAU-----AUAGCAAGAAU-AUAUACGCCUCCG---
```

# False positives

Features of real genome alignments: low-complexity DNA, repetitive sequence, microsatellite expansions, big indels...

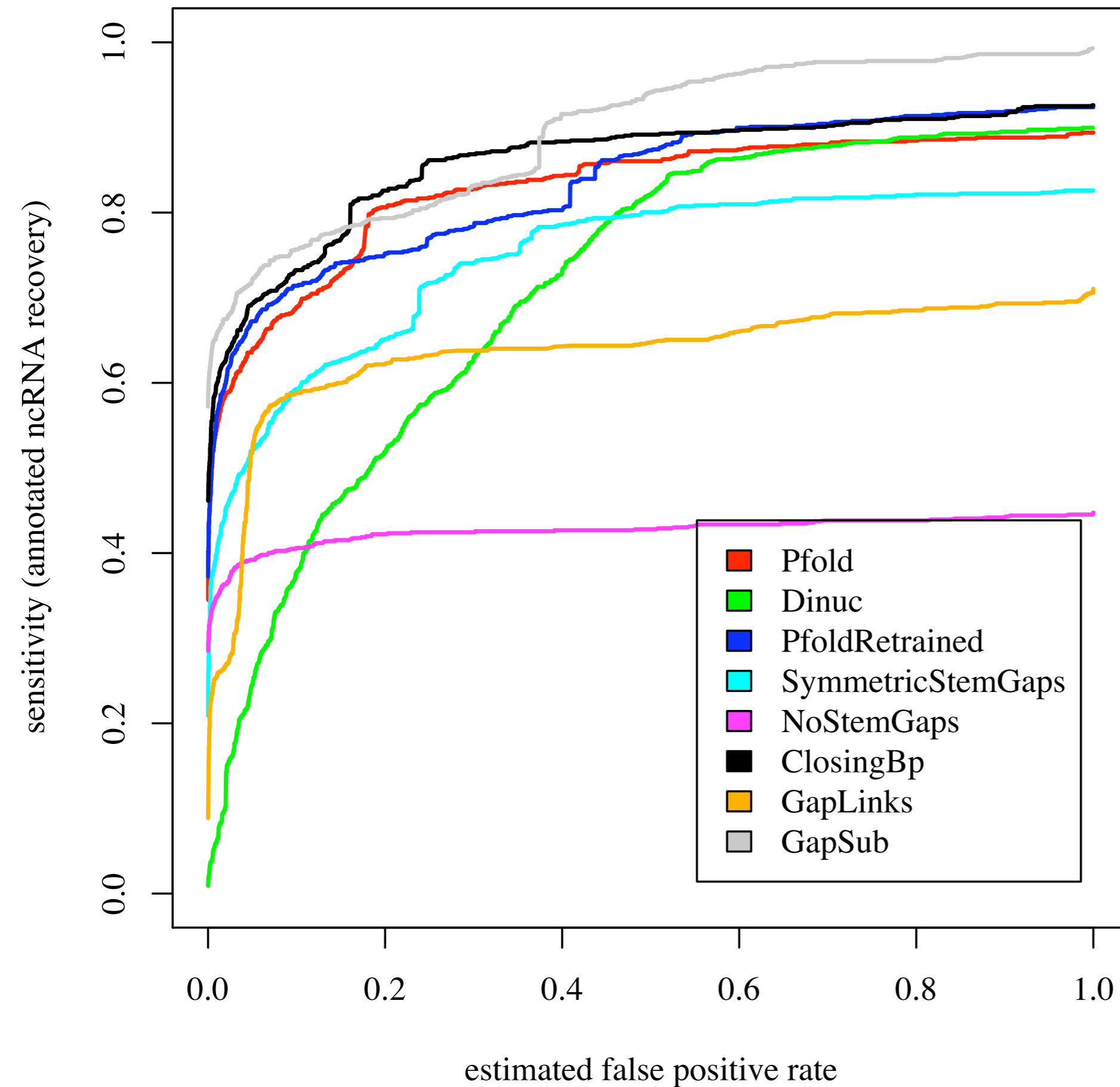
```
UUUAAACU-----CCUAGAAGAACUAGAAGCUCUGGAAG---CUUCU-UAAGCGUUU-----UAU-----AUAUCAGAAU-AUAUAGCGCGUCAA---
AUCCAGCA-----GGAGCAGAGGCUCUGGCAG---CUUUU-GCAGCG-UU-----UAU-----AUAACGAGAAU-AUAUACGCCUCCG---
-----AAUUGCUGAACUCUGGCAGUUU--CUCGGGAUGUUUUAUUAACAGAUUAUUA-----AUAUCAGAAU-AUAUACGCCUCCG---
AUCCAGCU-----CU-----GGAGCAGAGGCUCUGGCAG---CUUUU-GCAGCG-UU-----UAU-----AUAACGAGAAU-AUAUACGCCUCCG---
-----UGUGGGAGAAGCUCUGGUAGUUU--UUCA--U-----GUGUACAUAUUAUUAUUAUAGCUAUAUUAUGUAUUAUUAUCGAUUAUUAUUAUUA-----UAUAGUAUACGUGAUCCACU
GUCUAGCCGAAGAACCCACACGGA-----ACUCUGGCAG---UUUUAUCUGCG-LG-----UAU-----AUAACGAGAAU-AUAUACGCCUCCG---
AUCCAGCU-----GGAGCAGAGGCUCUGGCAG---CUUUU-GCAGCG-UU-----UAU-----AUAACGAGAAU-AUAUACGCCUCCG---
-----AGUGGAAGAAGCUCUGGCAGCUU--UUUA--AGCGUUUAUUAUAGAGUUUAUA-----UAU-----AUAACGAGAAU-AUAUACGCCUCCG---
AUCCAGCA-----GGAGCAGAAGCUCUGGAAG---CUUUU-GUAGCG-UU-----UAU-----AUAACGAGAAU-AUAUACGCCUCCG---
```











# ROC curves comparing ncRNA genefinding grammars

Simulations:  
GSIMULATOR

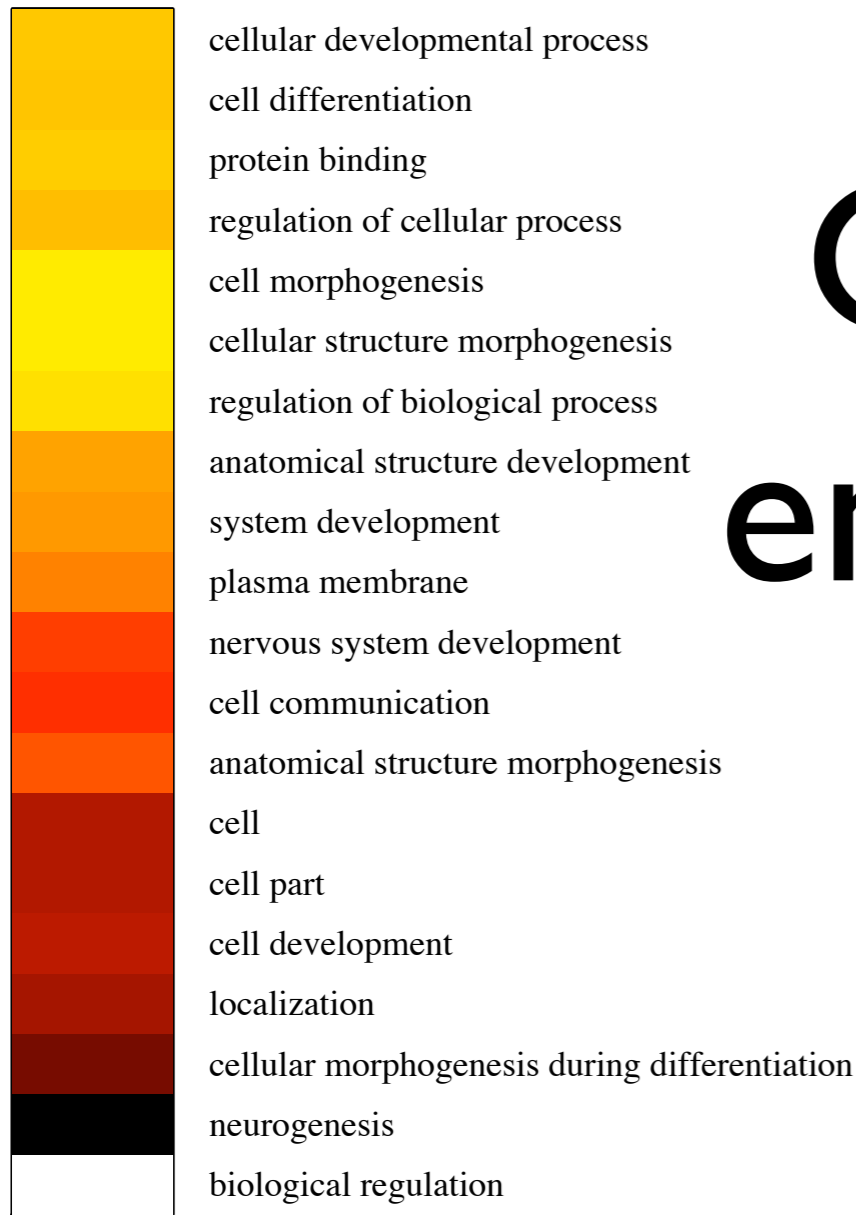
# Genome-wide scan, ClosingBp

	Predicted	Estimated false +ves
<b>Putative genes</b>	<b>1,949</b>	1%
In 3' UTRs	1,788	2%
In 5' UTRs	2,252	1%
In CDS	37,758	n/a
In pseudogenes	13	3%

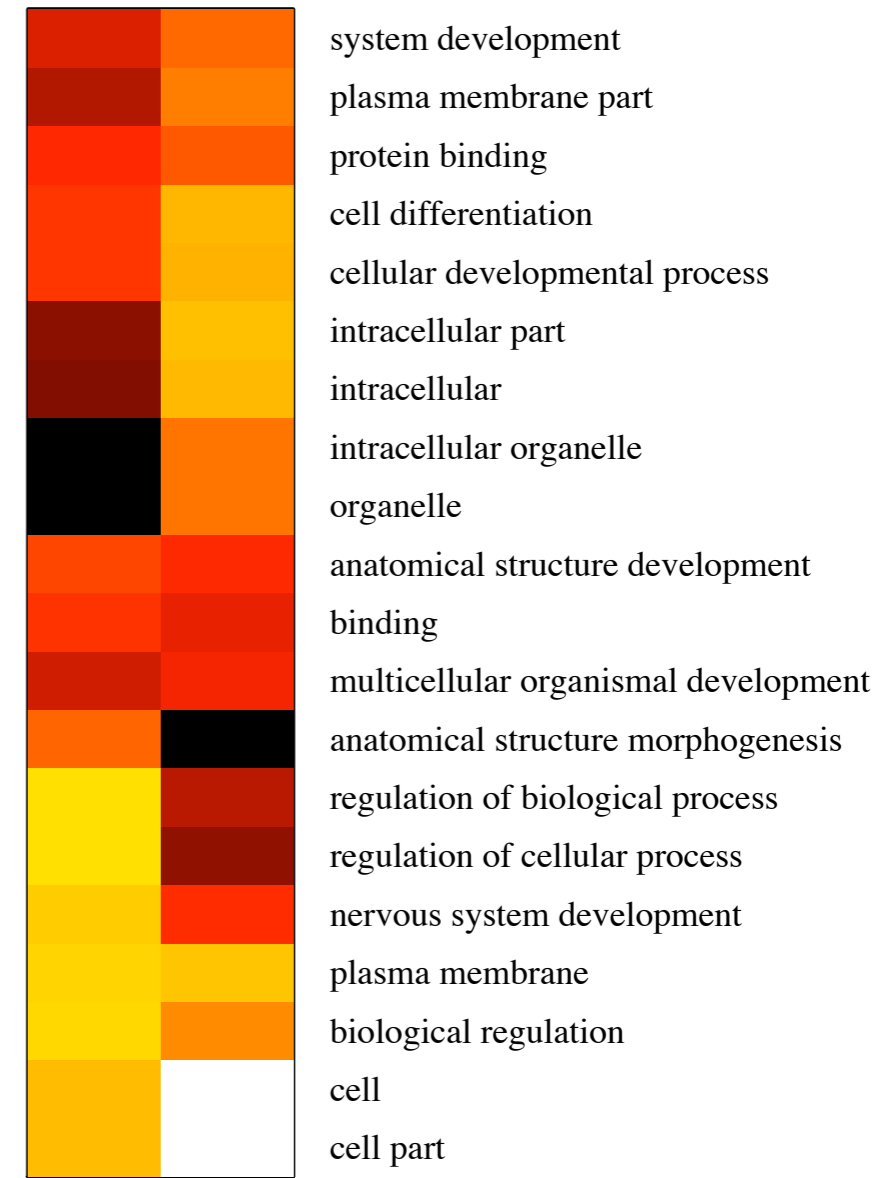
GROSS under-estimates!

Filtered by max gappiness, min number of species in alignment, min number of basepairs, min number of compensatory mutations; intergenic hits (predicted “genes”) further restricted by intersection w/Affy transfrag

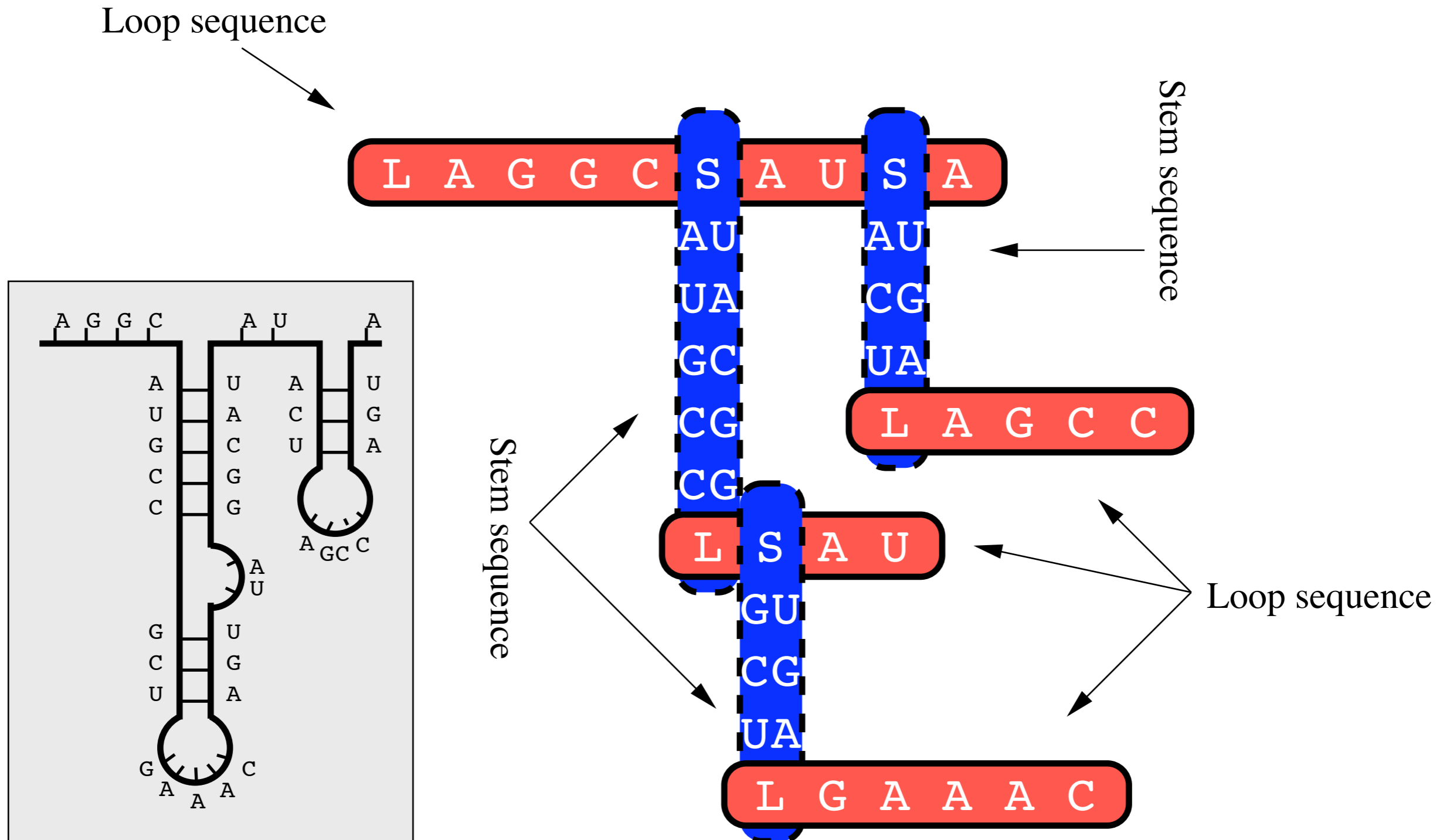
# GO term enrichment



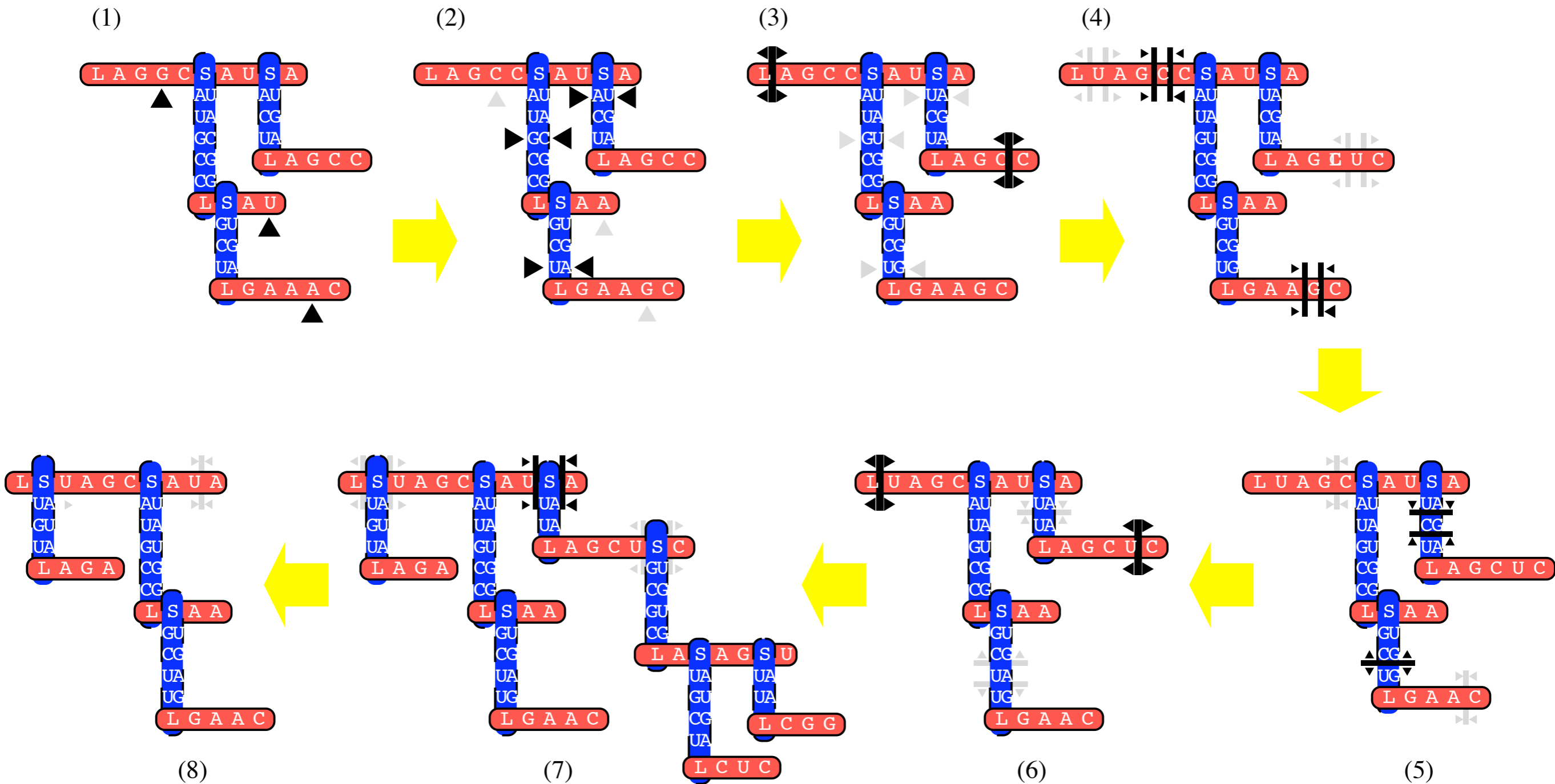
GO::TermFinder  
(Gavin Sherlock)



# Evolution of RNA structure



# Allowed mutations



Rule	lhs	→	rhs	$P(a)$	$P(d a)$
1.	$L_1$	→	$X_a Y_d L_1$	$\kappa_1 p_1(X)$	$(1 - \beta_1) \alpha_1 M_1(X, Y)$
2.			$Y_d L_1$	1	$\beta_1 p_1(Y)$
3.			$X_a L_2$	$\kappa_1 p_1(X)$	$(1 - \beta_1)(1 - \alpha_1)$
4.			$S_1 L_1$	$\kappa_1 p_1(S)$	$(1 - \beta_1) \alpha_1$
5.			$S_4 L_1$	1	$\beta_1 p_1(S)$
6.			$S_3 L_2$	$\kappa_1 p_1(S)$	$(1 - \beta_1)(1 - \alpha_1)$
7.			$\epsilon$	$1 - \kappa_1$	$1 - \beta_1$
8.	$S_1$	→	$W_a Y_d S_1 Z_d X_a$	$\kappa_2 p_2(WX)$	$(1 - \beta_2) \alpha_2 M_2(WX, YZ)$
9.			$Y_d S_1 Z_d$	1	$\beta_2 p_2(YZ)$
10.			$W_a S_2 X_a$	$\kappa_2 p_2(WX)$	$(1 - \beta_2)(1 - \alpha_2)$
11.			$L_1$	$1 - \kappa_2$	$1 - \beta_2$
12.	$L_2$	→	$X_a Y_d L_1$	$\kappa_1 p_1(X)$	$(1 - \gamma_1) \alpha_1 M_1(X, Y)$
13.			$Y_d L_1$	1	$\gamma_1 p_1(Y)$
14.			$X_a L_2$	$\kappa_1 p_1(X)$	$(1 - \gamma_1)(1 - \alpha_1)$
15.			$S_1 L_1$	$\kappa_1 p_1(S)$	$(1 - \gamma_1) \alpha_1$
16.			$S_4 L_1$	1	$\gamma_1 p_1(S)$
17.			$S_3 L_2$	$\kappa_1 p_1(S)$	$(1 - \gamma_1)(1 - \alpha_1)$
18.			$\epsilon$	$1 - \kappa_1$	$1 - \gamma_1$
19.	$S_2$	→	$W_a Y_d S_1 Z_d X_a$	$\kappa_2 p_2(WX)$	$(1 - \gamma_2) \alpha_2 M_2(WX, YZ)$
20.			$Y_d S_1 Z_d$	1	$\gamma_2 p_2(YZ)$
21.			$W_a S_2 X_a$	$\kappa_2 p_2(WX)$	$(1 - \gamma_2)(1 - \alpha_2)$
22.			$L_1$	$1 - \kappa_2$	$1 - \gamma_2$
23.	$L_3$	→	$X_a L_3$	$\kappa_1 p_1(X)$	1
24.			$S_3 L_3$	$\kappa_1 p_1(S)$	1
25.			$\epsilon$	$1 - \kappa_1$	1
26.	$S_3$	→	$W_a S_3 X_a$	$\kappa_2 p_2(WX)$	1
27.			$L_3$	$1 - \kappa_2$	1
28.	$L_4$	→	$Y_d L_4$	1	$\kappa_1 p_1(Y)$
29.			$S_4 L_4$	1	$\kappa_1 p_1(S)$
30.			$\epsilon$	1	$1 - \kappa_1$
31.	$S_4$	→	$Y_d S_4 Z_d$	1	$\kappa_2 p_2(YZ)$
32.			$L_4$	1	$1 - \kappa_2$

# Pair SCFG

$$\alpha_n = \exp(-\mu_n t)$$

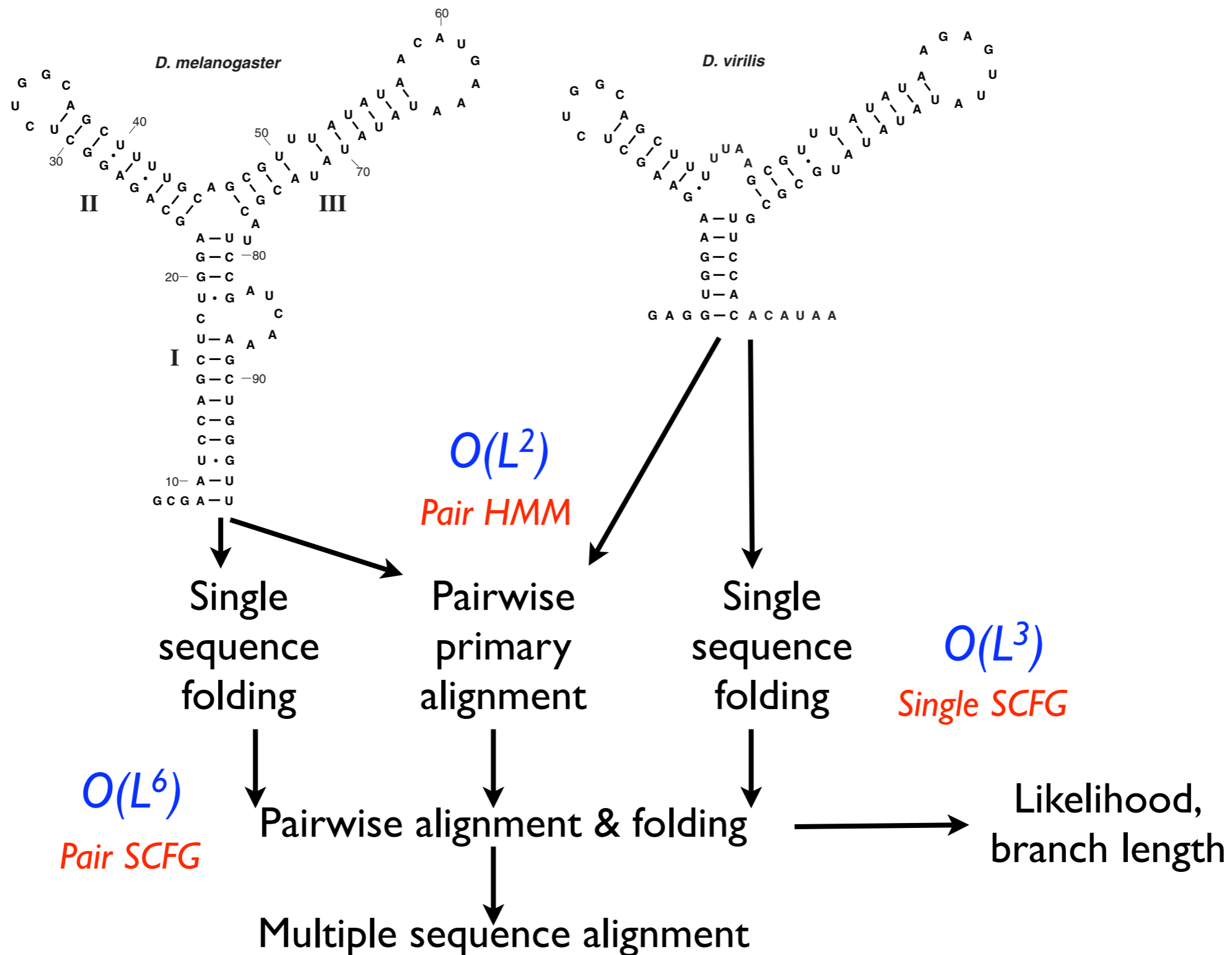
$$\beta_n = \frac{\lambda_n (1 - \exp((\lambda_n - \mu_n)t))}{\mu_n - \lambda_n \exp((\lambda_n - \mu_n)t)}$$

$$\gamma_n = 1 - \frac{\mu_n (1 - \exp((\lambda_n - \mu_n)t))}{(1 - \exp(-\mu_n t))(\mu_n - \lambda_n \exp((\lambda_n - \mu_n)t))}$$

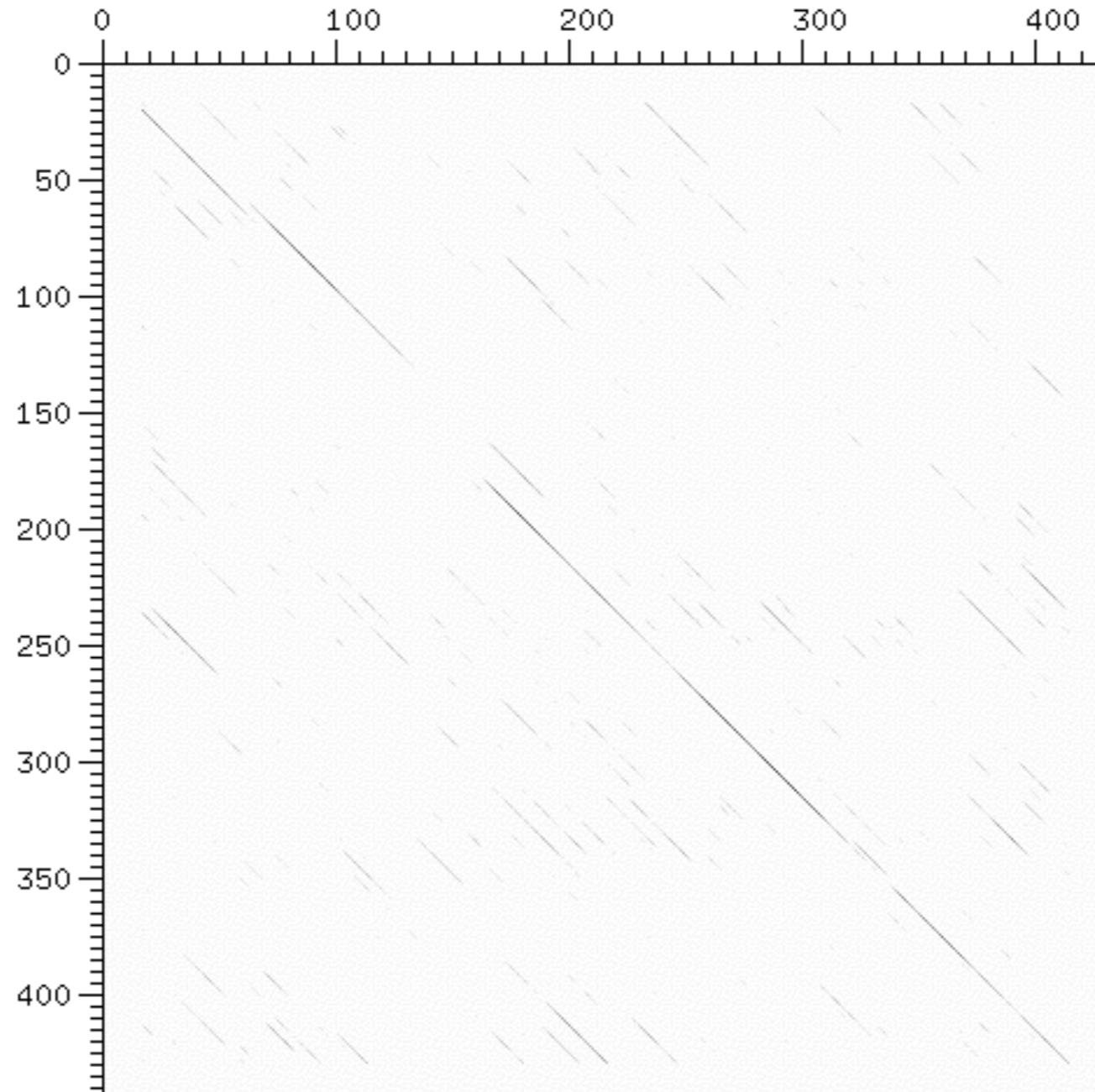
$$\kappa_n = \lambda_n / \mu_n$$

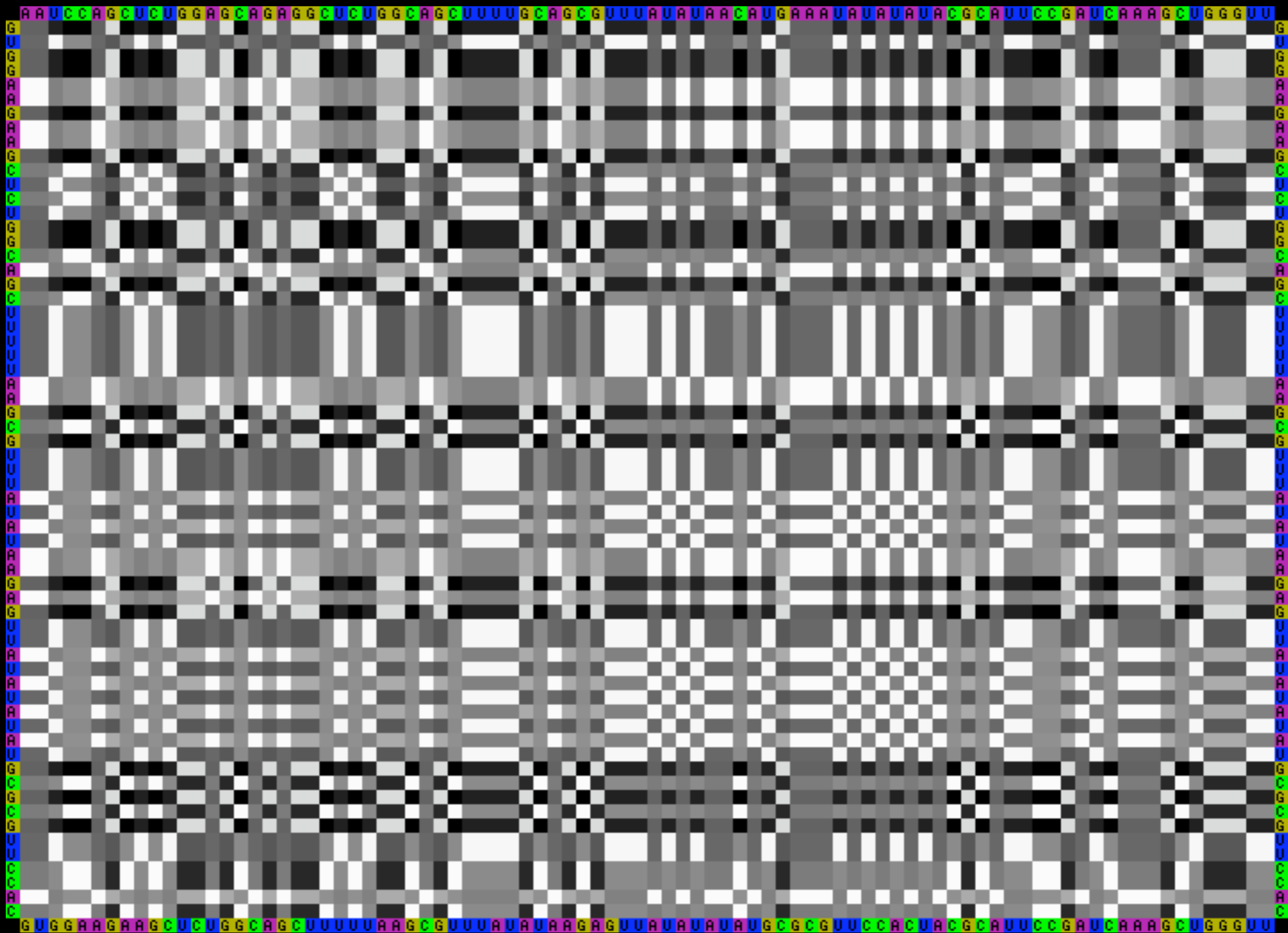
$$M_n(i, j) = \exp(\mathbf{R}_n t)_{ij}$$

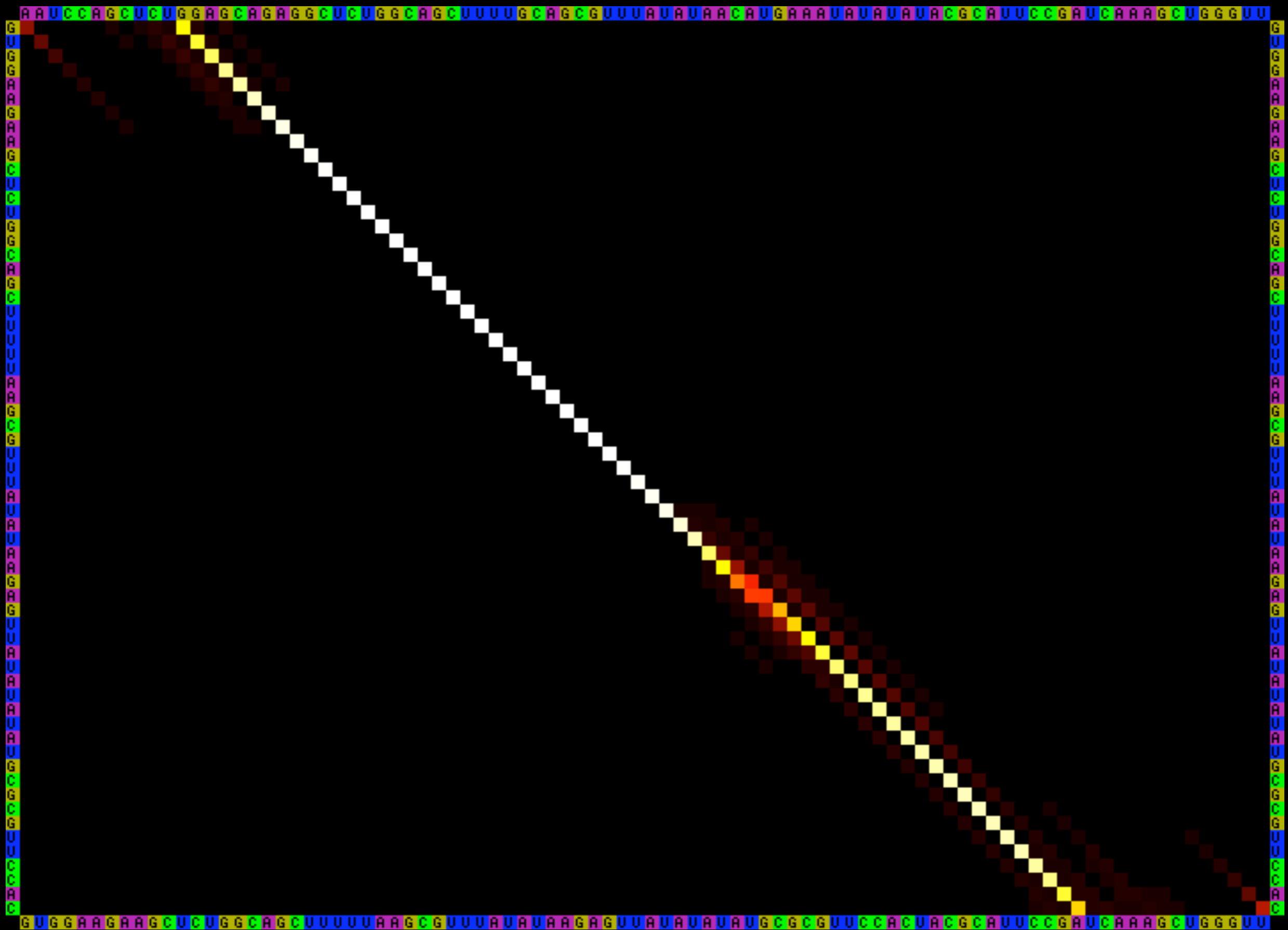
# Constrained Sankoff



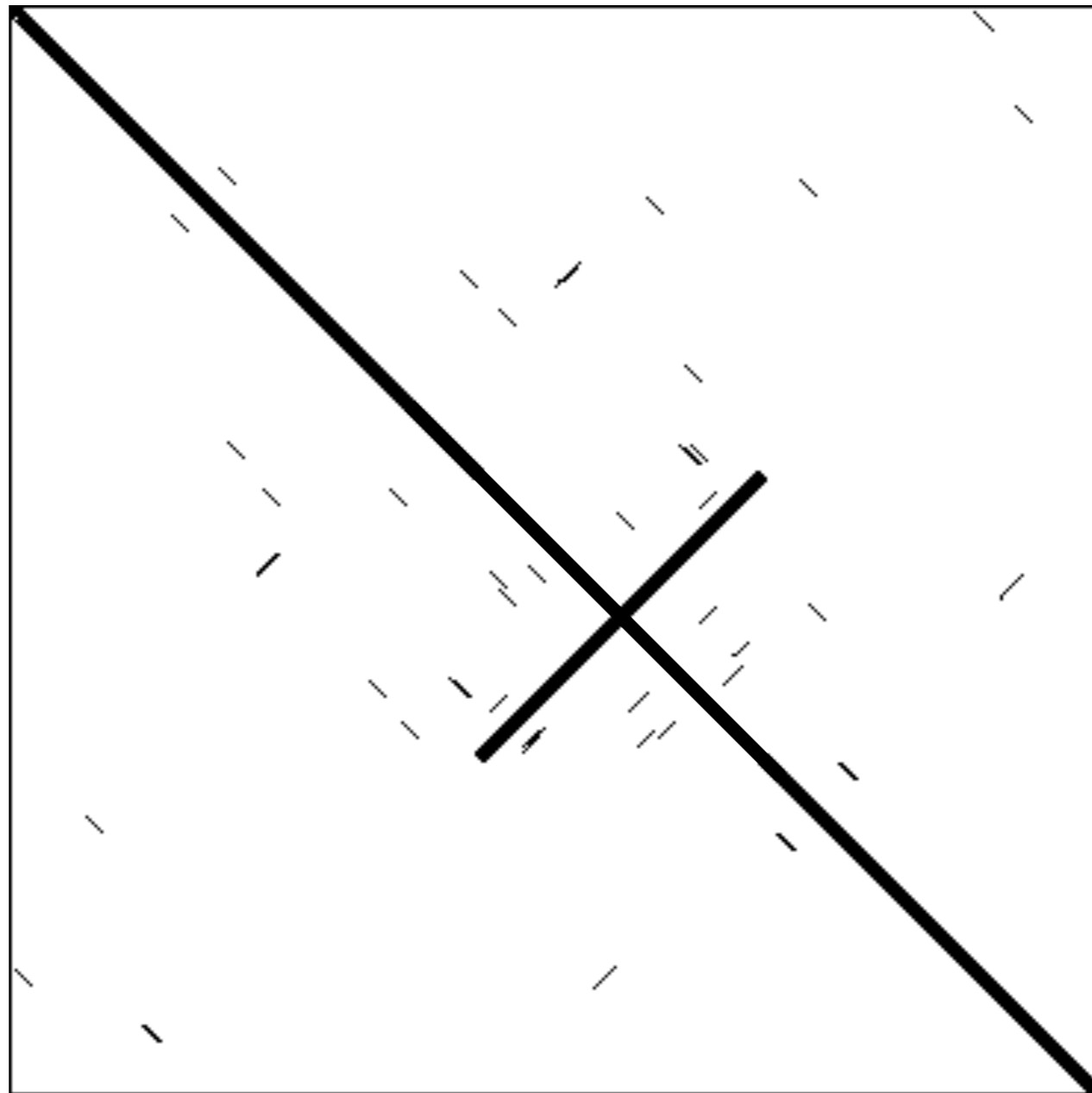
# Dotplots

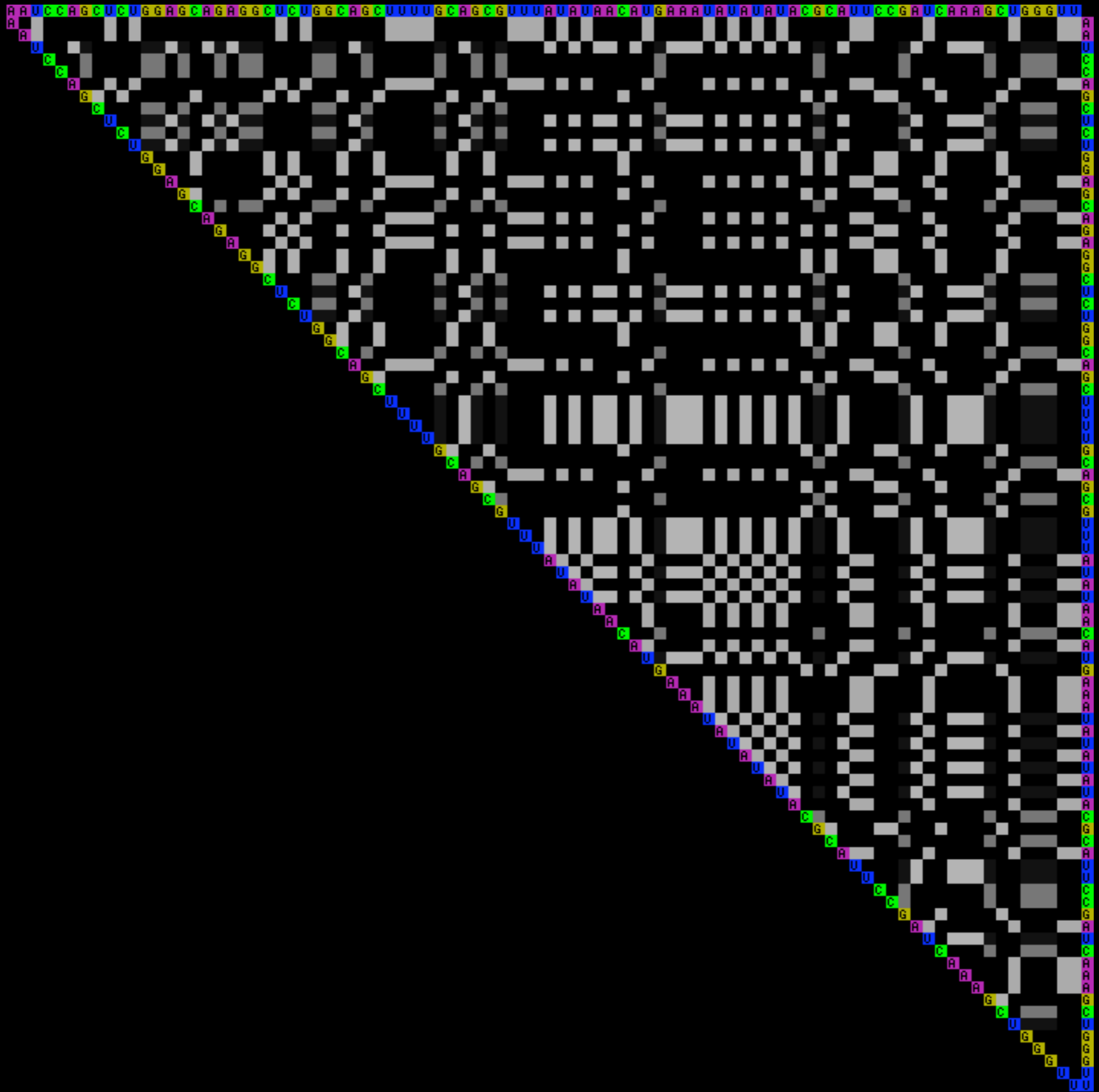




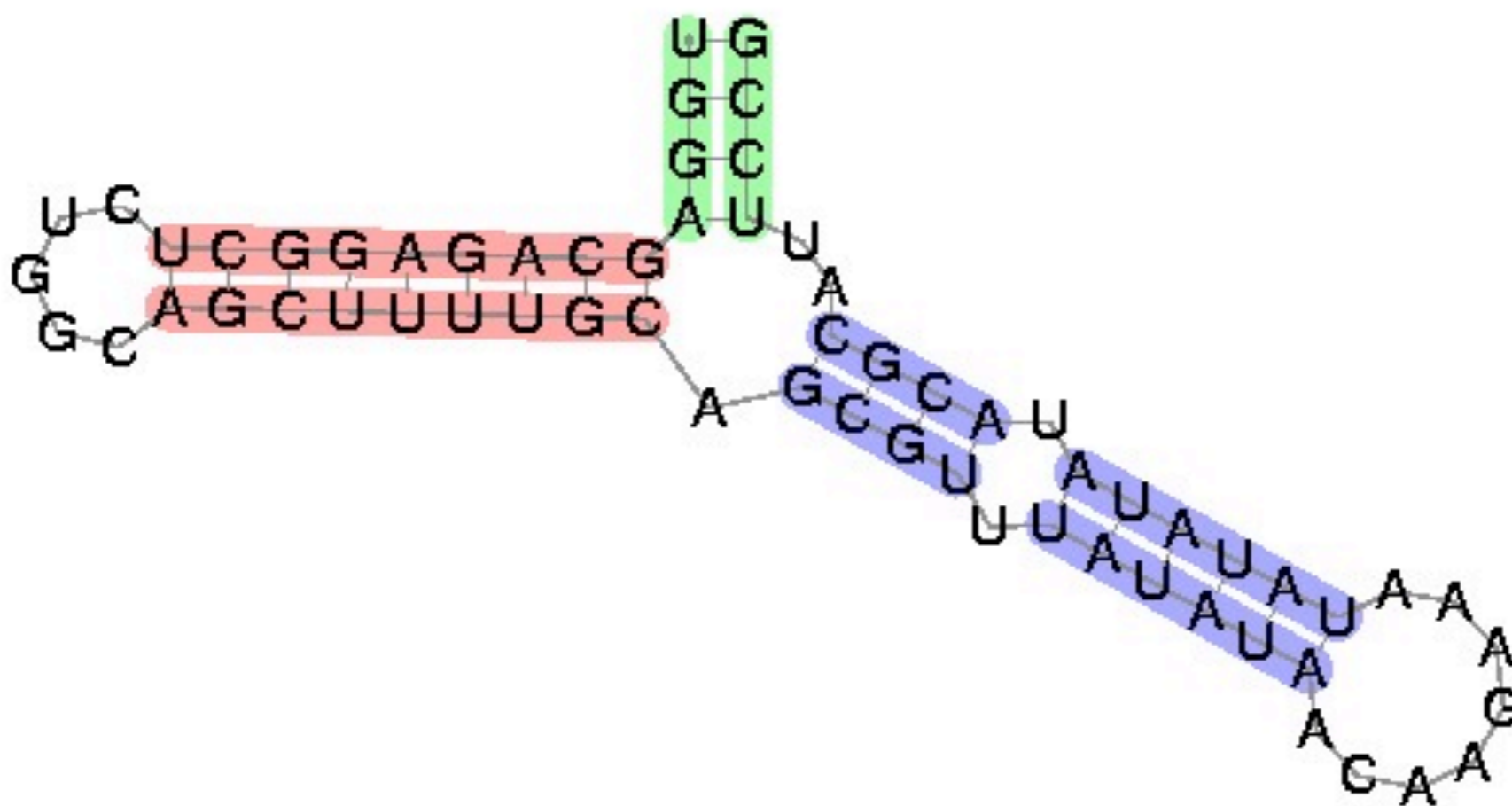


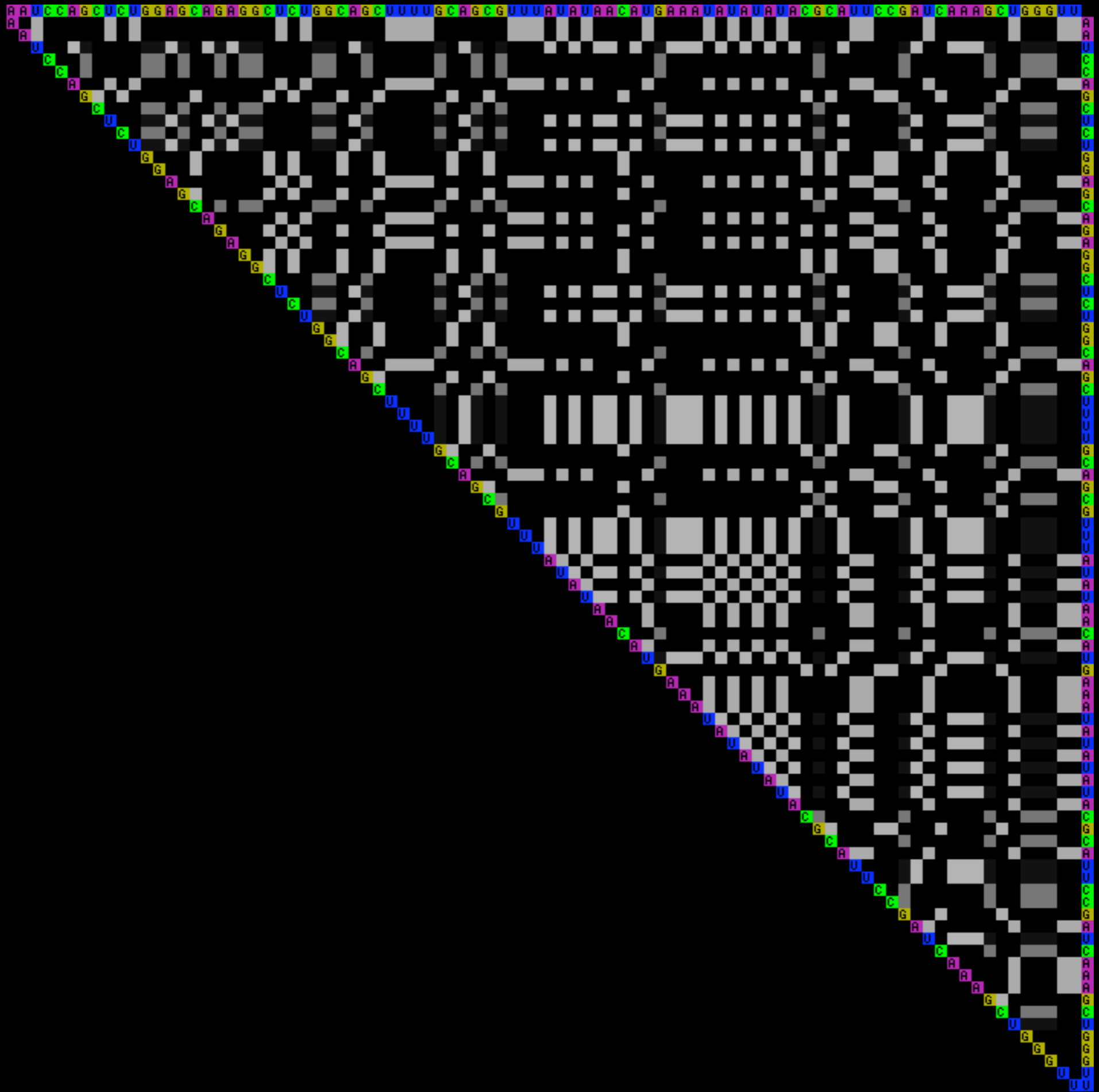
# Dotplot vs. self: inverted repeat





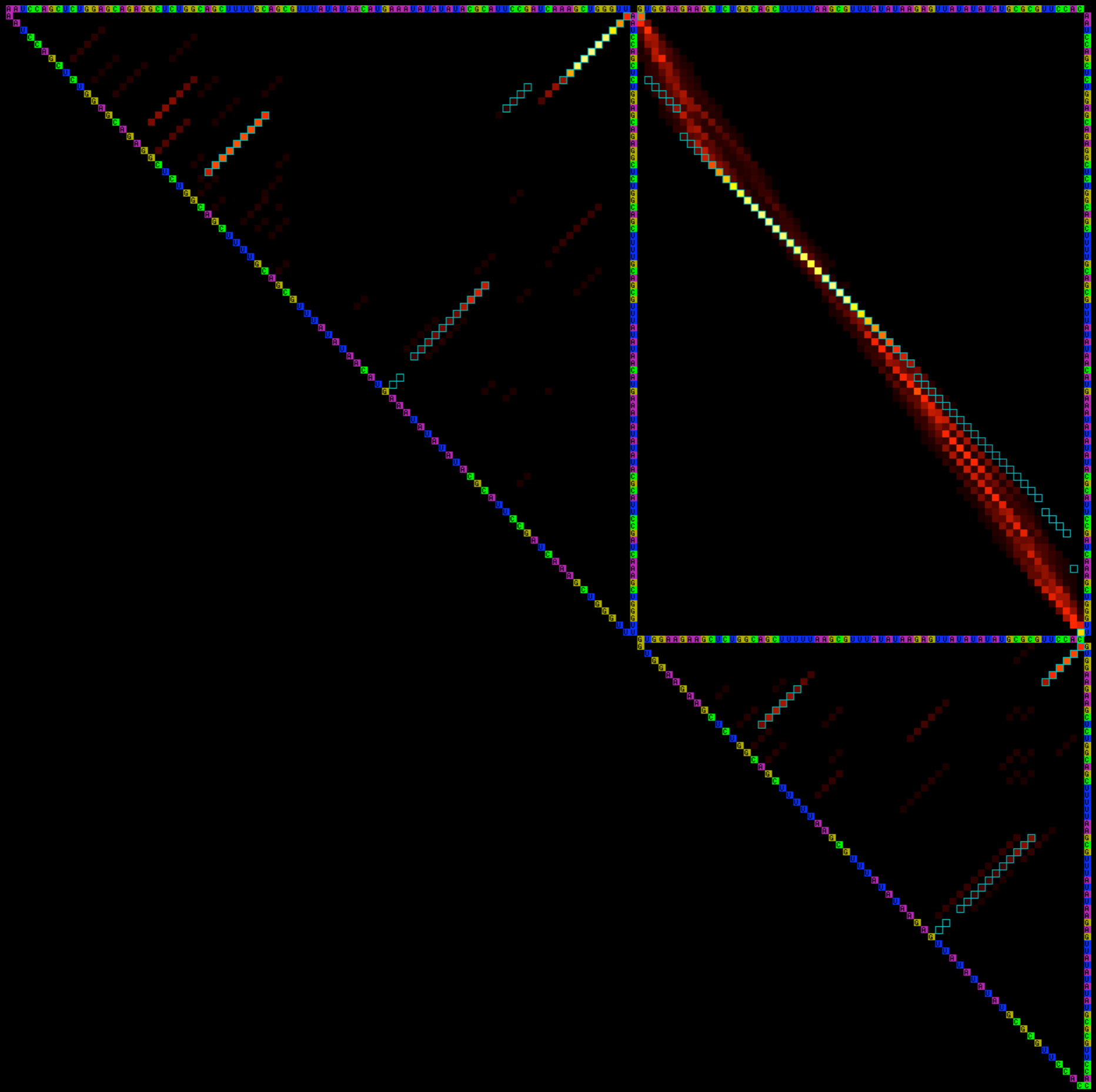
# nanos TCE

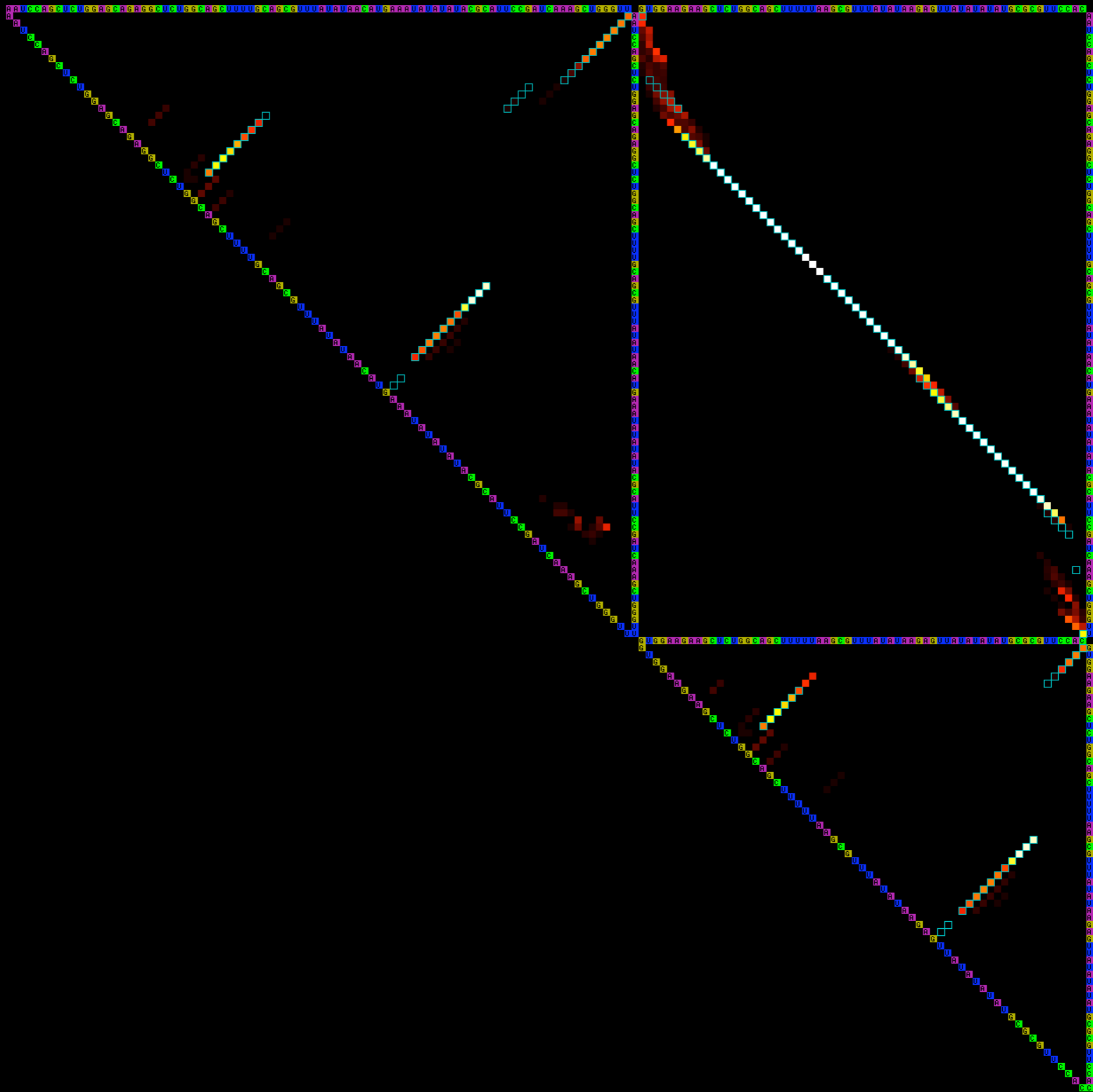




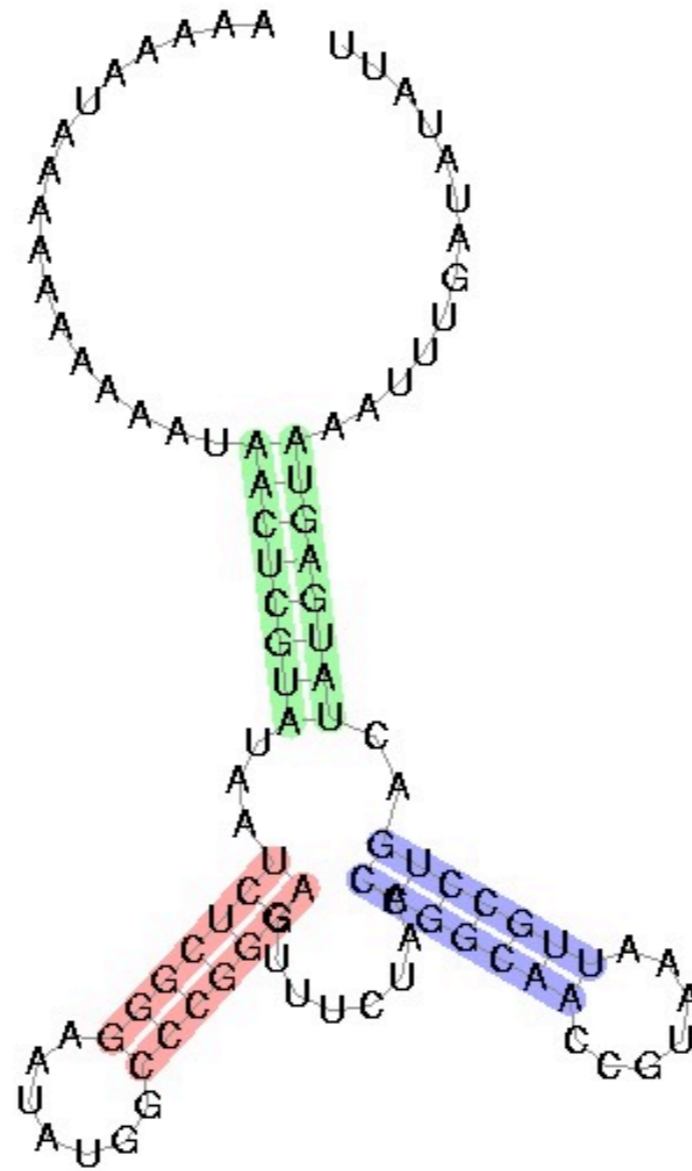


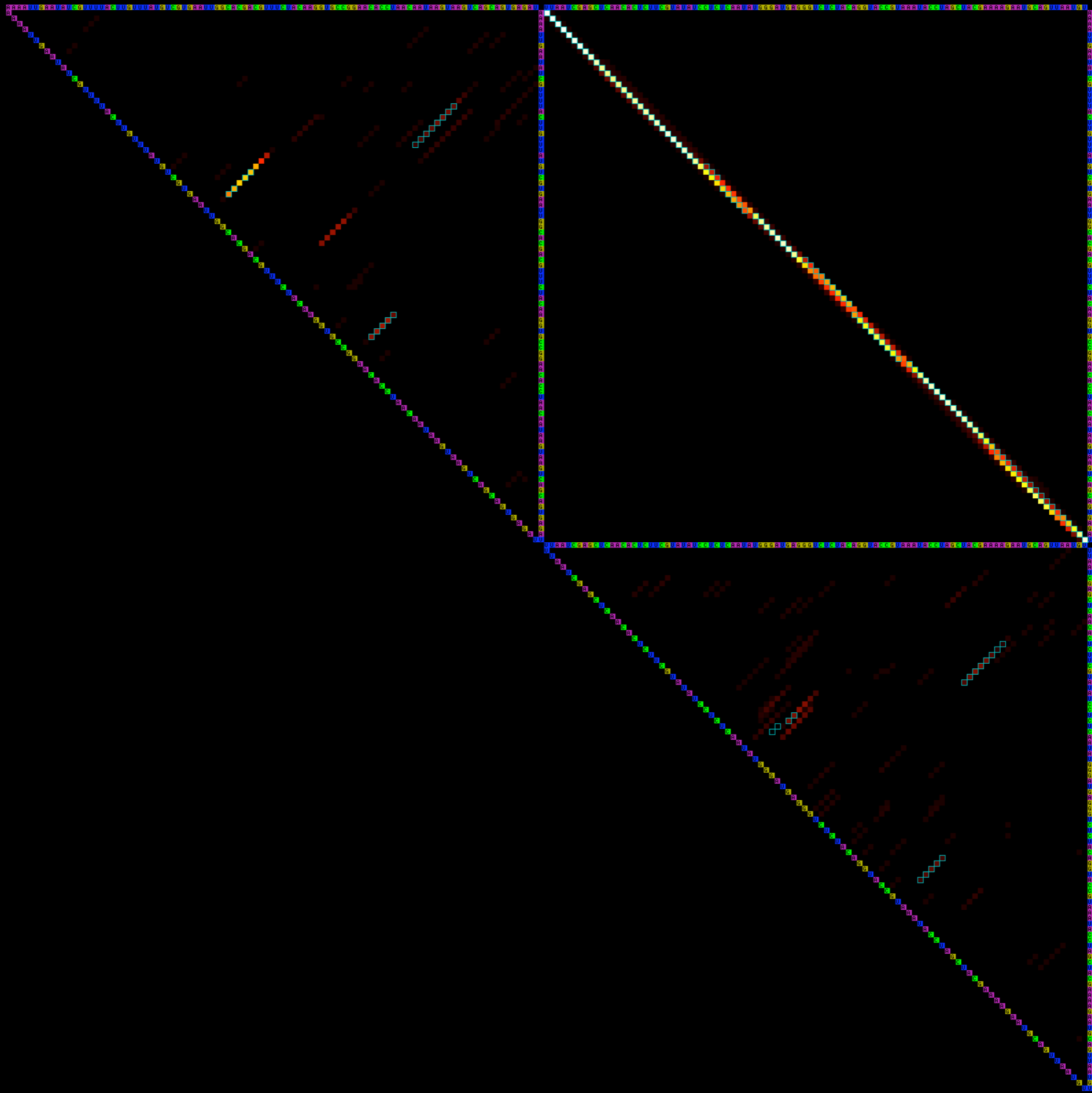


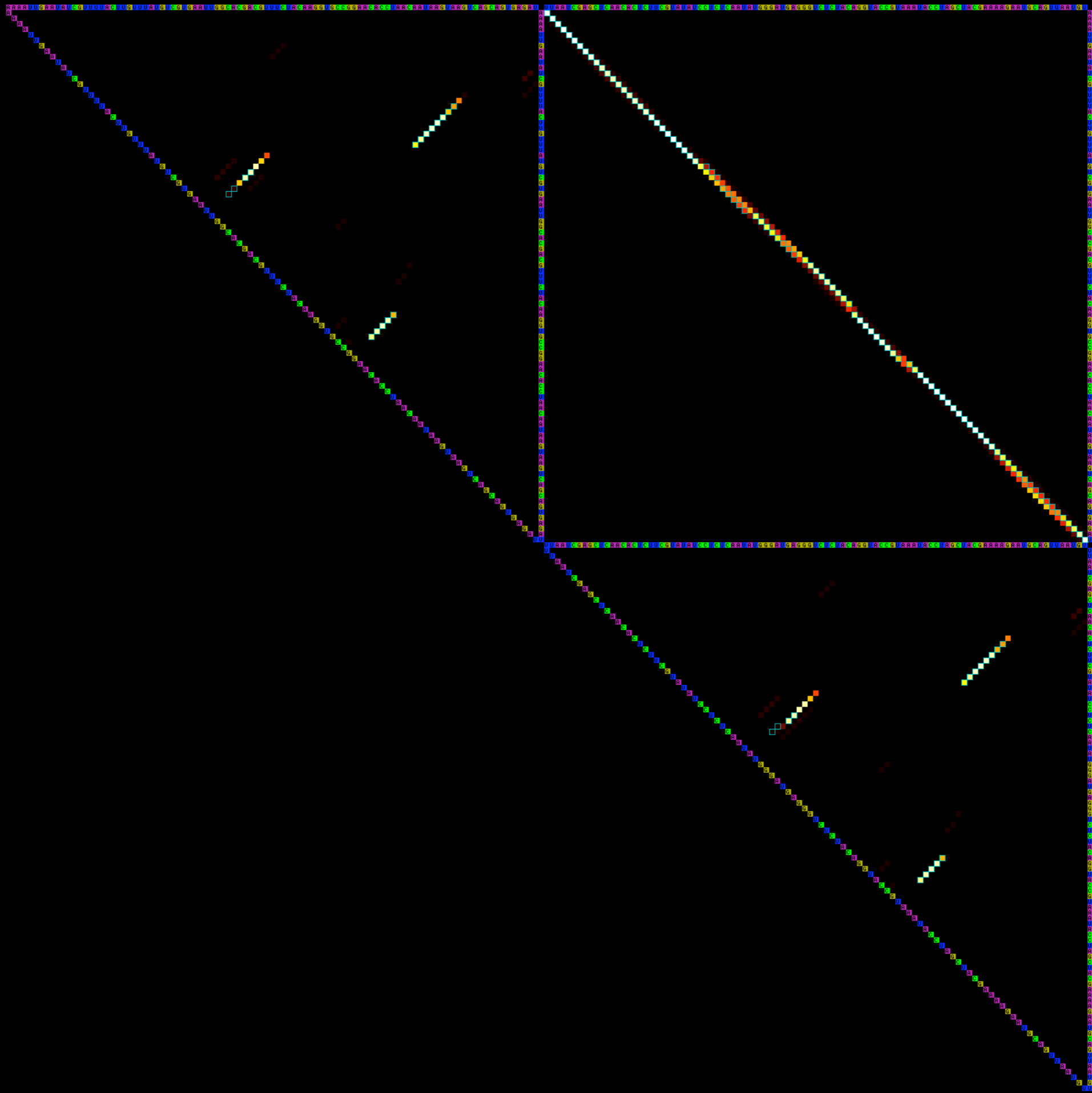




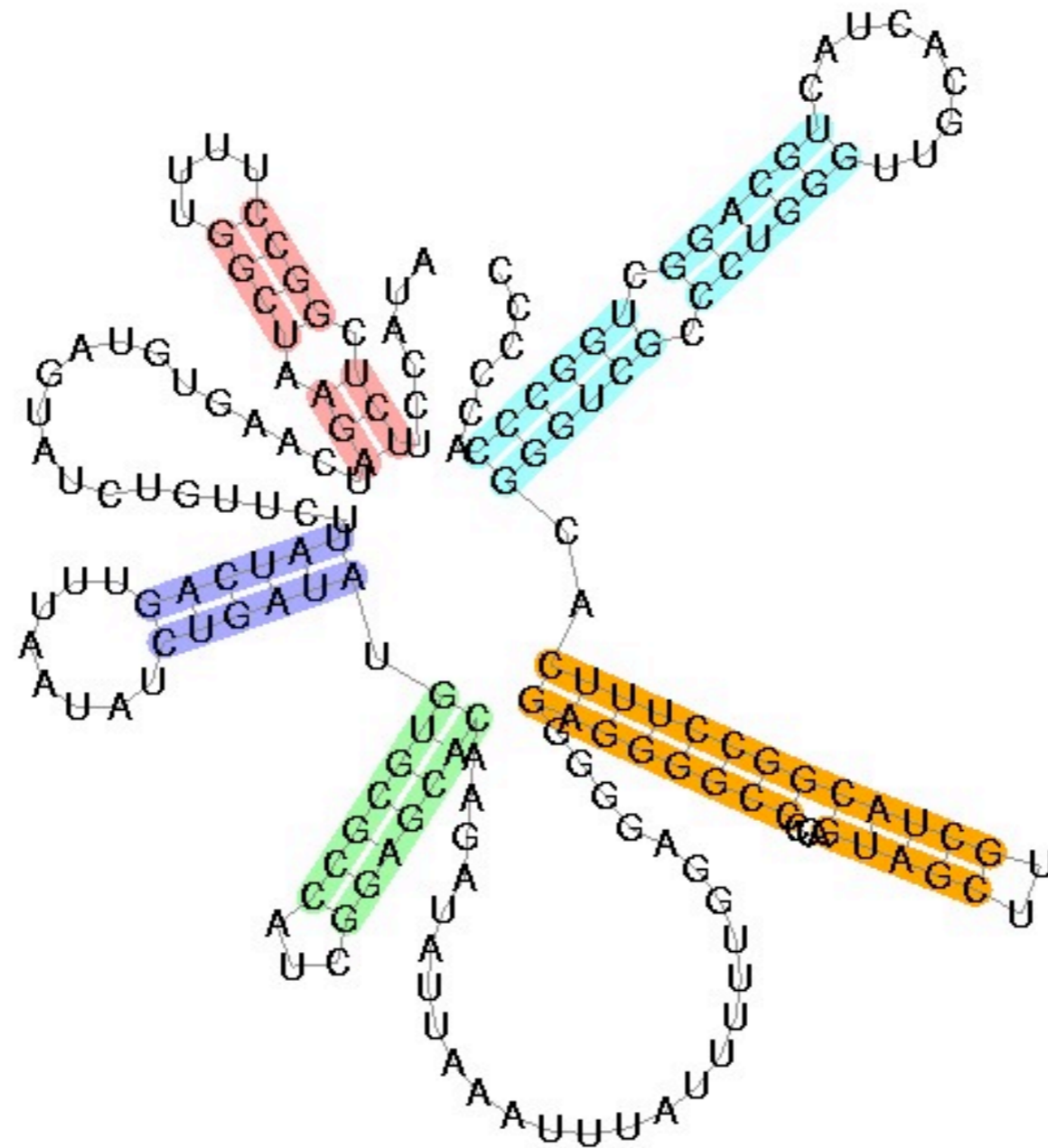
# Purine riboswitch







# U2 spliceosomal RNA



# U2 spliceosomal RNA

