William B. Langdon and Justyna Petke and Ronny Lorenz
GGGP, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK (Email: w.langdon@cs.ucl.ac.uk)

Evolving better RNAfold structure prediction

Central region of the human glutamate receptor stem-loop pre-mRNA

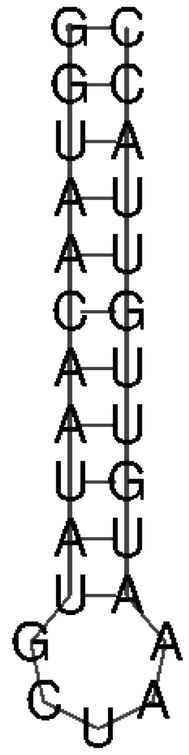


Figure 1a: RNAfold prediction (MCC 0.956018)

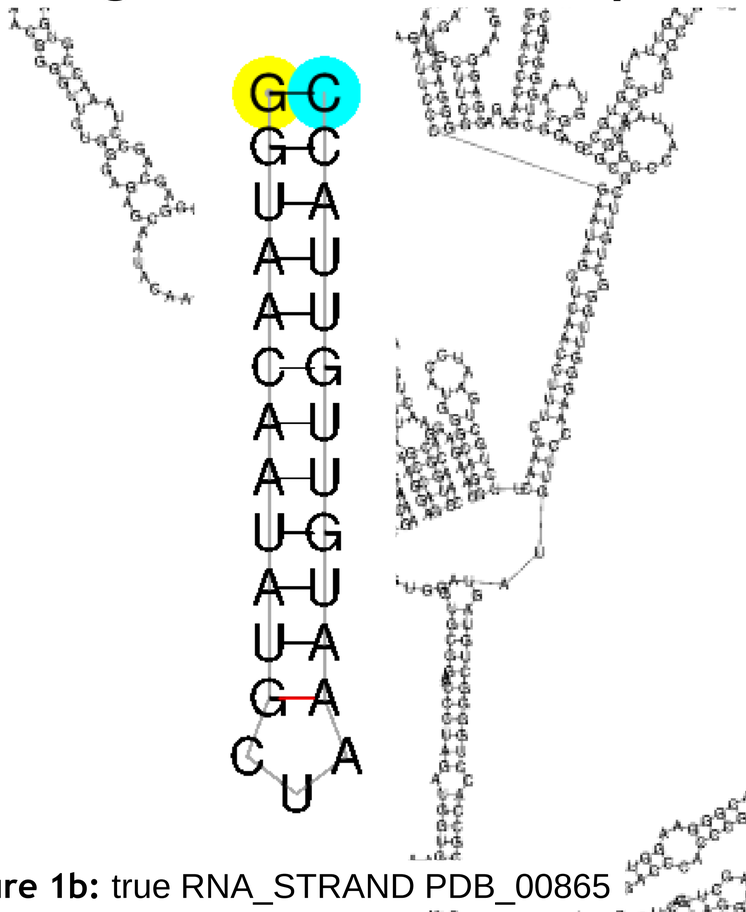


Figure 1b: true RNA_STRAND PDB_00865

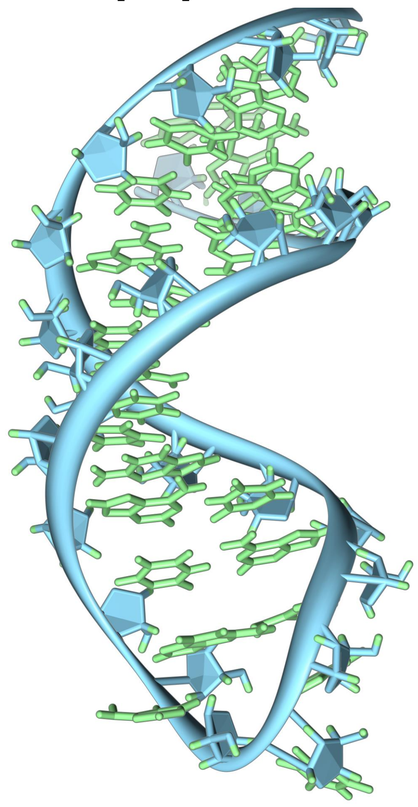


Figure 1c: True, Protein Data Bank 1YSV

1 Why

Scientific programs require parameters. RNAfold uses 50000 parameters to predict RNA structure. Parameters are subject to measurement noise. ViennaRNA's have been revised by hand. We use evolution to fit them to ground truth (RNA_STRAND).
Better than ViennaRNA 2.3.0 and indeed better than Andronescu et al., 2007, $p<10^{-54}$

2 Variable length list of parameter changes

3 Replace mutation >

```
mismatchM -60>-40
```

Replace every element in array mismatchM whose values is currently -60 with -40

4 Overwrite mutation <

```
mismatchH *,1,2>-80
```

Overwrite eight elements in array mismatch (ie mismatchH[*,1,2]) with -80

5 Increment mutation +=

```
mismatchH *,*,*+=-90
```

Add -90 (ie substract 90) from every element in array mismatch (ie mismatchH[*,*,*] 200)

6 Creep mutation

Small (≤5,50) change to value of 20% existing mutations

7 Two point crossover

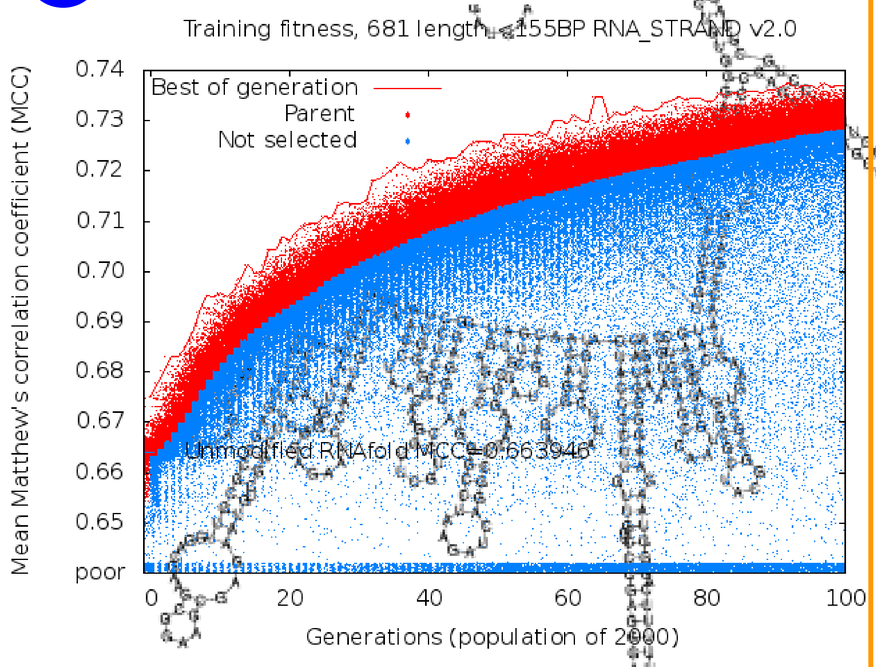
8 Fitness Function

Interpret list of mutations and apply them to RNAfold's 51521 int energy parameters. Run updated RNAfold on training RNA sequences (1/3rd RNA_STRAND 2.0 with <155 bases). It predicts structure for each of these 681 RNA, for each calculate Matthew's correlation coefficient (MCC) with true structure. At least one change is necessary to be selected.

$$\mathrm{MCC} = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

TN = true negatives, ie n(n-1)/2-(TP+FP+FN)

9 Evolution of fitness

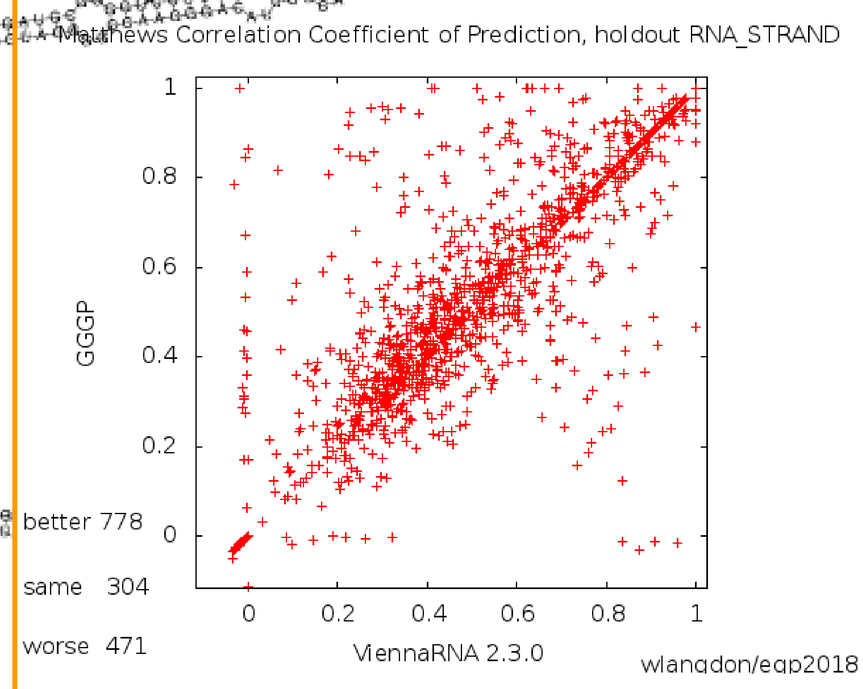


10 Post Evolution Tidy: Bloat removed

Best individual in population at generation 100 contained 2849 mutations, hill climbing (2passes) leaves just 42.

Background predicted structure of RNA from Spinach 23S Ribosome CRW_01456

11 Results $p<10^{-17}$ on holdout



12 Summary

Genetic programming is routinely used to generate from scratch small models of physical systems (e.g. Eureqa) but GP can automatically update constants within sizeable programs which have taken years to develop where keeping up with the latest empirical data is liable to drag many months behind scientific knowledge.

13 Next

CUDA RNAfold RN/18/02

Reference: W.B. Langdon, J.Petke and R.Lorenz. Evolving better RNAfold structure prediction. In M.Castelli, L.Sekanina and M.Zhang editors, EuroGP 2018, LNCS 10781, Parma, Italy, 2018. Springer Verlag.
Evolved parameters rna_langdon2018.par