CREST Open Workshop 25th September 2017

Faster folds, Better folds: Genetic Improvement of RNAfold

W. B. Langdon
Computer Science, University College London

WIKIPEDIA
Genetic Improvement

GI 2018, Göteborg, ICSE-2018 *proposed* workshop

Genetic Improvement of RNAfold

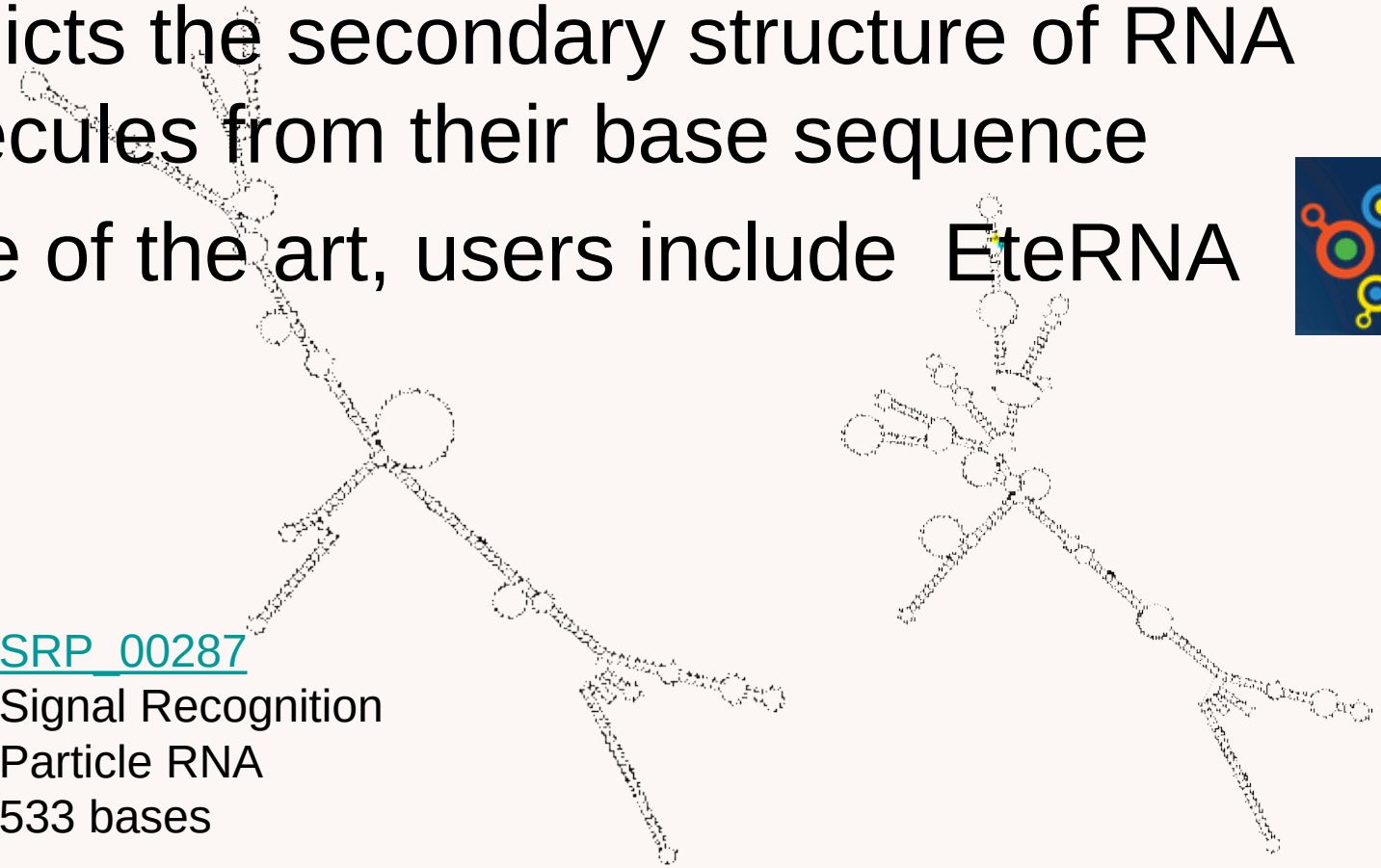
- What is RNAfold
- Grow and Graft Genetic Programming

  1. speed up, 2. functional improvement
- GGGP RNAfold
  - 31% speed up via SSE, GI 2017 workshop
  - Optimise C code, 1% better predictions
  - Optimise 50,000 parameters
    - net 20% better prediction of RNA structures
  - Next: try 512 bit hardware

What is RNAfold?

- Part of ViennaRNA package (170000 lines)
- RNAfold 7100 lines .c (i.e. excluding .h)
- Predicts the secondary structure of RNA molecules from their base sequence
- State of the art, users include EteRNA

SRP_00287
Signal Recognition
Particle RNA
533 bases



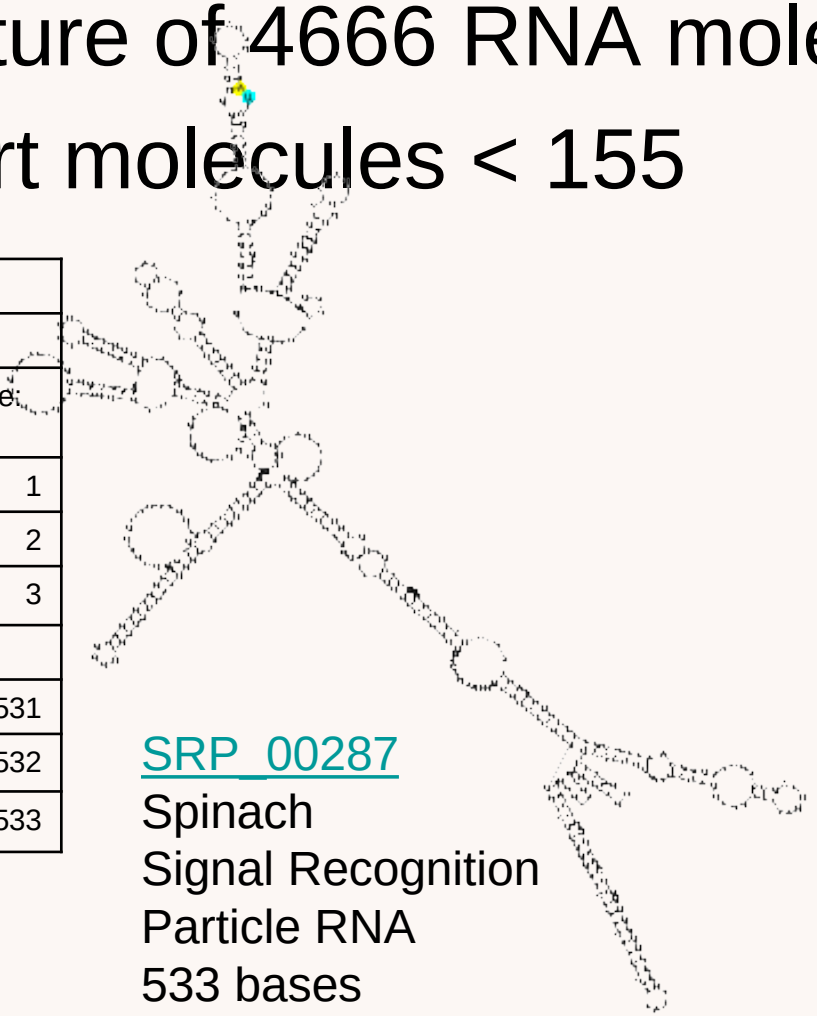
Matthews correlation coefficient MCC 0.519169

Training/Test data: RNA STRAND

Known structure of 4666 RNA molecules

Train on short molecules < 155

# File SRP_00287.ct					
# RNA SSTRAND database					
# External source: SRP Database, file name: SAC.CAS..ct, ID: SAC.CAS.					
1	A	0	2	15	1
2	G	1	3	14	2
3	G	2	4	13	3
…					
531	A	530	532	0	531
532	C	531	533	0	532
533	U	532	534	0	533



SRP_00287
Spinach
Signal Recognition
Particle RNA
533 bases

RNAfold

- Uses dynamic programming to select structure with minimum energy.
- Source code contains 31 read only scalars and arrays which hold parameters for model of interactions between RNA bases.
- Total 51745 parameters (all int)
- Use evolution GGGP to optimise 51745 parameters

Optimise 50,000 parameters in RNAfold

- Mutate read-only arrays before RNAfold runs dynamic programming
- Compare new predicted structure with correct structure from RNA STRAND
- Use ⅓ molecules for training
- Run time excessive:
  - use small molecules for training, size < 155
  - still running RNAfold 681 times (too many?)

Representation: Genotype→Phenotype

- Variable length genotype. Each gene specifies one or more changes to one scalar or array parameter.
- Apply changes in order (canonical operator removes some redundant genes, bloats anyway).
- Multiple types of mutation
- Two point (variable length) crossover

Mutate scalar or array values

- > Replace all values with another

  `int22 260>80` Replace every 260 with 80

- < Replace one or more values with another

  `mismatchI *,*,0<100` Volume replace [*,*,0] by 100

- Increment one or more values with another

  `mismatchM *,3,*+=20` Add 20 to all `mismatchM[*,3,*]`(40)

- Respect energy values (all multiples of 10 or INF) and “small values” (0…8). Cannot inc/dec INFinity.
- 20% creep mutation: change value in existing mutation.

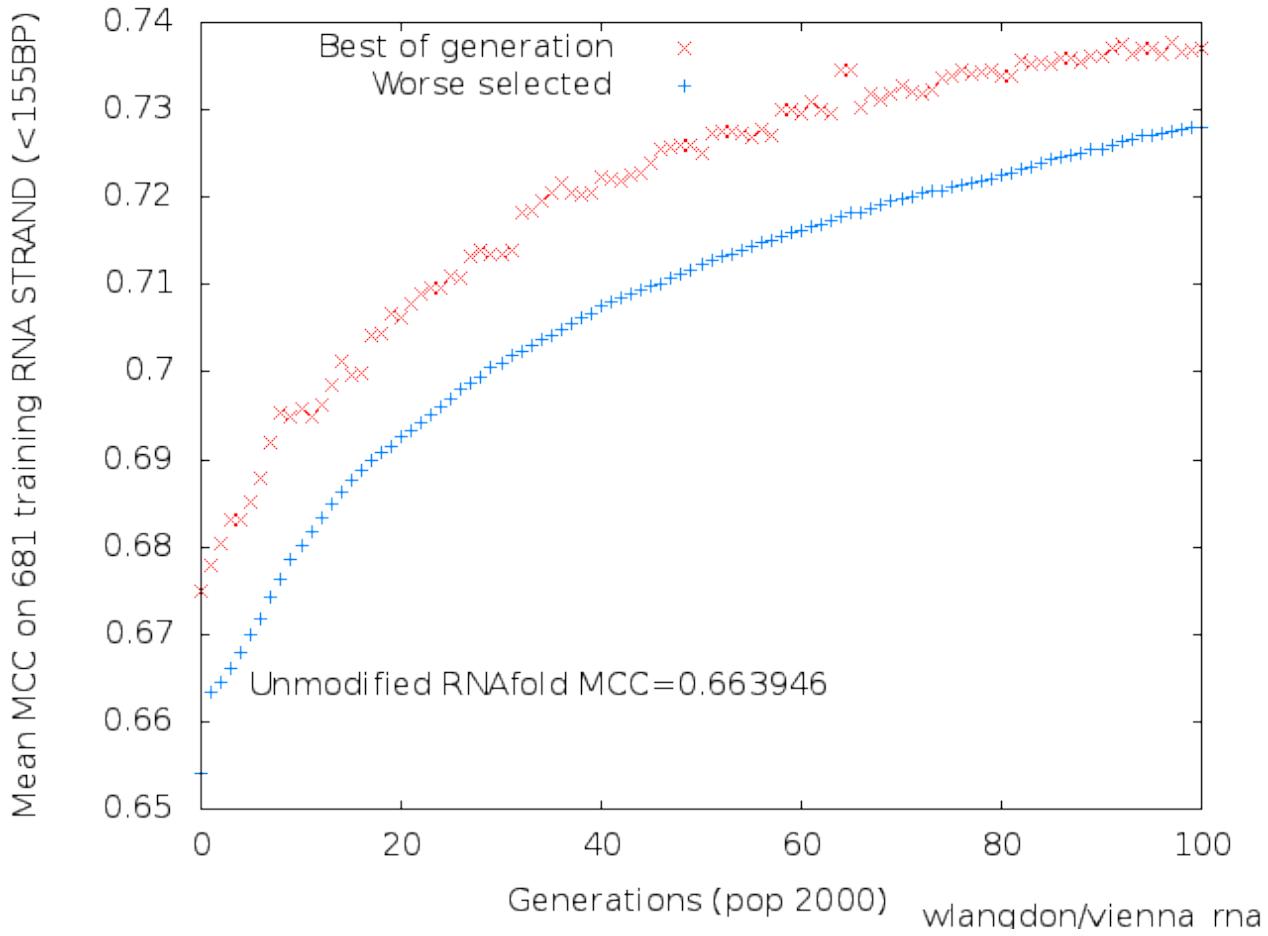
Fitness

- Run RNAfold on whole of training set of RNA molecules (len<155) from RNA STAND
- Compare each predicted structure against actual structure in RNA STRAND using Matthews Correlation Coefficient (MCC) and against unmuted prediction. Fitness is mean MCC, but
  - If no changes: cannot be parent
  - If RNAfold segfaults: cannot be parent
  - If can't mutate params: cannot be parent
- Select best half of population to be parents

Evolution

- 50% mutation, 50% crossover
- Promote search:
  - Reduce to canonical form
  - Tabu search to prevent repeated evaluation of genetically identical children
  - Anti-elitism: fitness cannot be parent more than 20 times (ie 1% popsize).
- 100 generations, population 2000

Evolution of Training Fitness



Results

- Take best of last generation (100)
  - Length 2849, MCC 0.737044
- Remove bloat by removing genes which do not help (two passes).
  - Length 42, MCC 0.737752
- Little over fitting: holdout MCC 0.730137

Evolved change

hairpin *<560
mismatchM -70>-130| *,3,*+=20| *,1,*+=-40| -110>-130| *,0,*+=-170| -60>-40
internal_loop *+=-40
MLintern *+=10| 3<-150
rtype 6<6| 2+=1
int11 *,*,*,*<200| 6,*,*,2+=-70
int21 230>260| *,*,*,*,3+=-70| 220>10000000
int22 260>80| 180>280| *,*,2,*,*,*+=10| 280>200| 200>10000000
dangle3 5,*+=-80
mismatchH *,*,*+=-90| *,*,3<-130| *,1,2<-80
mismatchExt *,*,*+=80| *,*,1<-40
TerminalAU 80
mismatch23I 70>10000000
mismatchI *,*,0<100| *,*,1+=-10| 2,3,1+=-100| *,4,*+=-40
ninio[2] 80
dangle5 *,*+=60
stack -100>60| -140>0| 2,2+=-20| *,4<-50
mismatch1nI 70>110
bulge *+=40

mismatchM many changes

rtype base A treated as C, X as K

mismatchH Rewrite array

mismatchI many changes

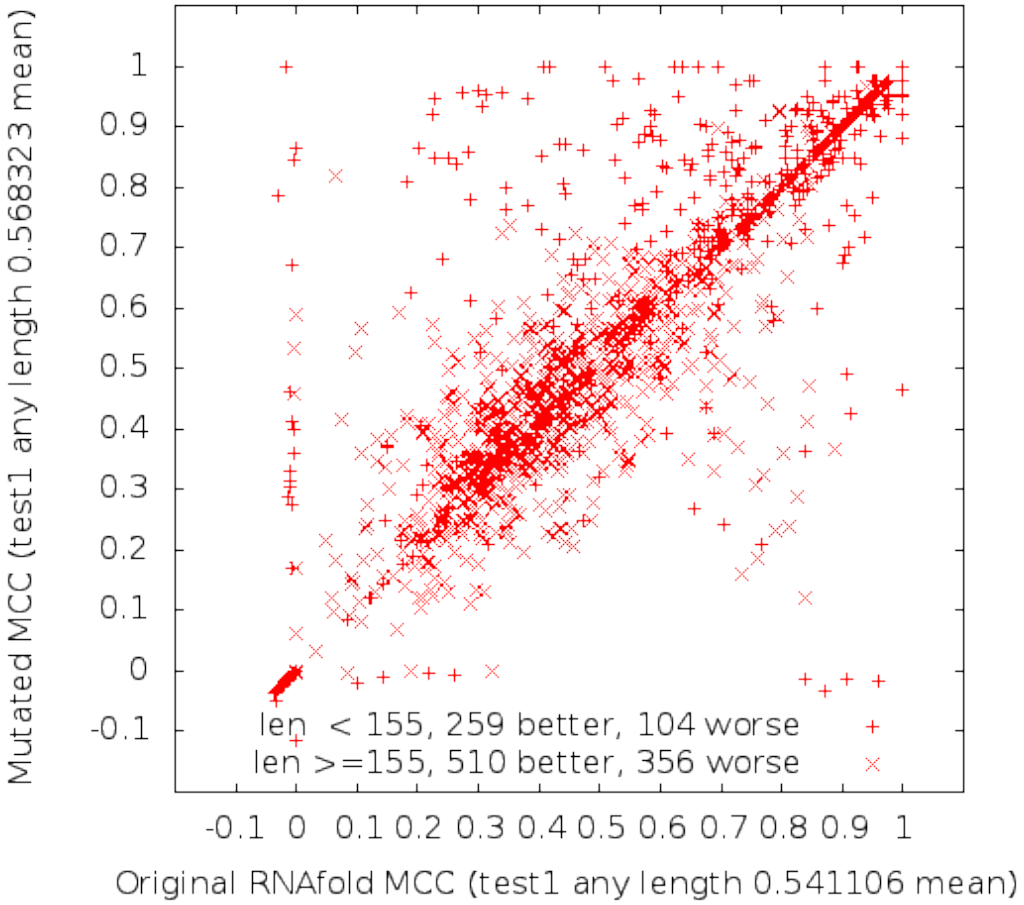
stack many changes

Impact on MCC

mismatch1nI	0.47%
mismatch23I	0.64%
int22	1.11%
dangle3	1.86%
int21	4.12%
dangle5	4.43%
bulge	5.15%
TerminalAU	6.02%
ninio[2]	7.53%
int11	10.70%
MLintern	10.72%
internal_loop	10.89%
hairpin	10.97%
mismatchExt	15.45%
stack	20.32%
mismatchI	21.12%
rtype	21.48%
mismatchM	21.62%
mismatchH	27.91%

Fraction of improvement in MCC lost if remove changes to each scalar or array.
(Measured on training data.)

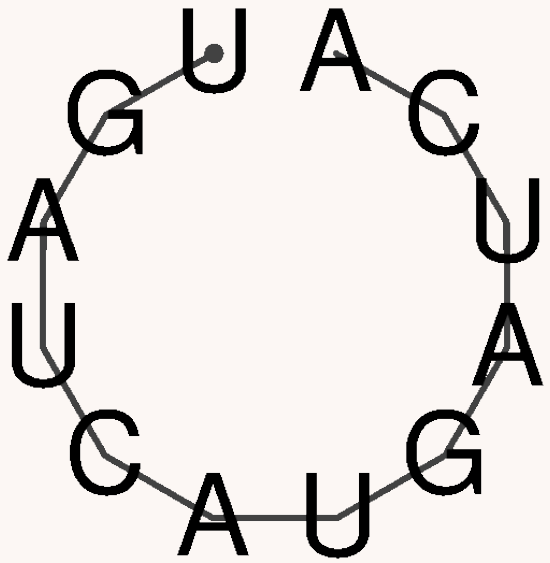
Out of Sample Performance



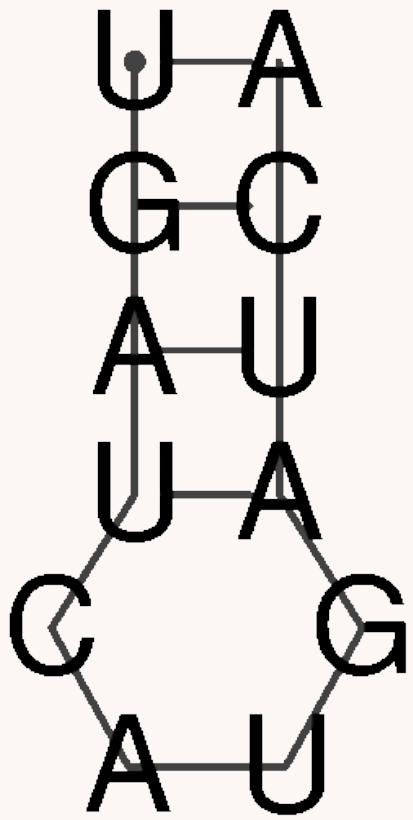
Both generalises (MCC on test set ≈ training) and extrapolates (MCC long RNA similar to training).

Total 769 better, 460 worse, holdout ⅓ RNA STRAND (1553).
Total overall out-of-sample improvement 19.897%

NDB_00028

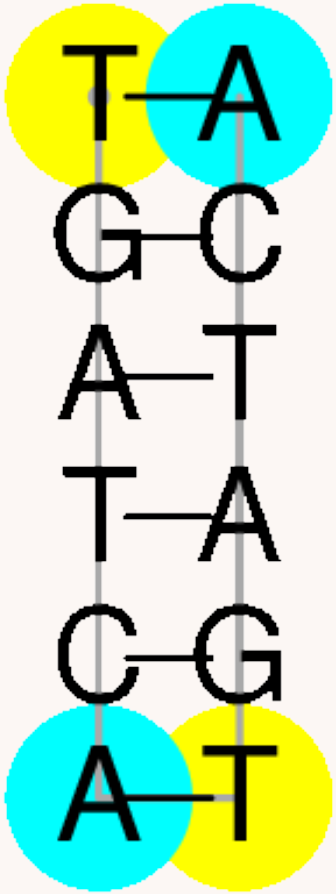


Original, MCC = 0



Mutant, MCC 0.803219

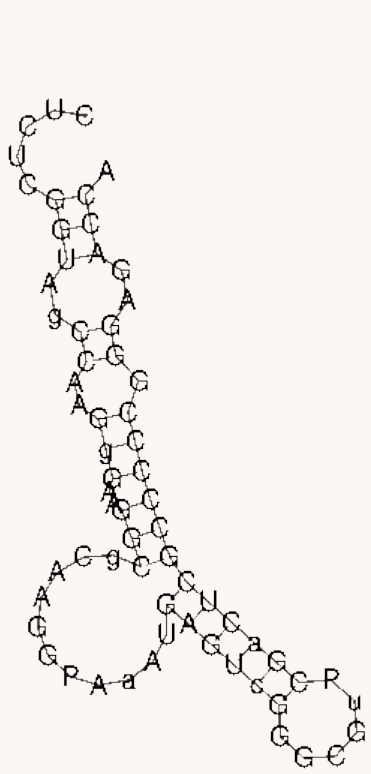
Symmetric

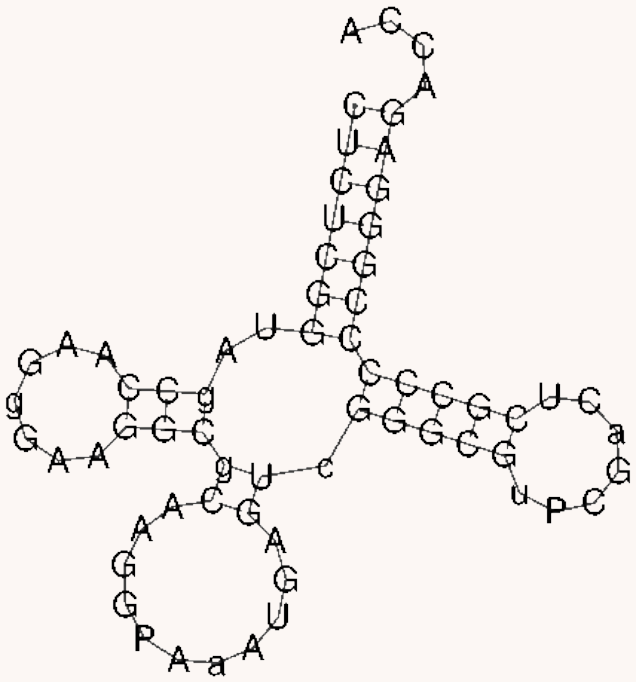


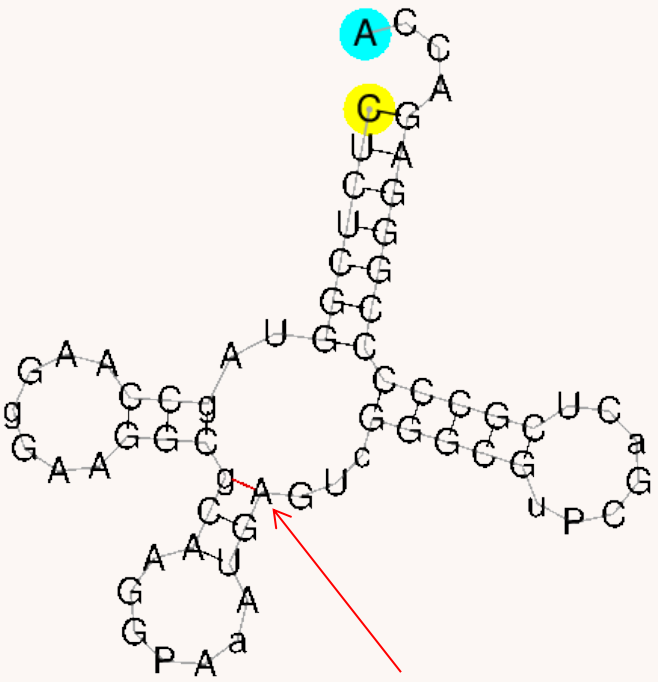
True

PDB 01001

yeast enzyme (in protein manufacture)







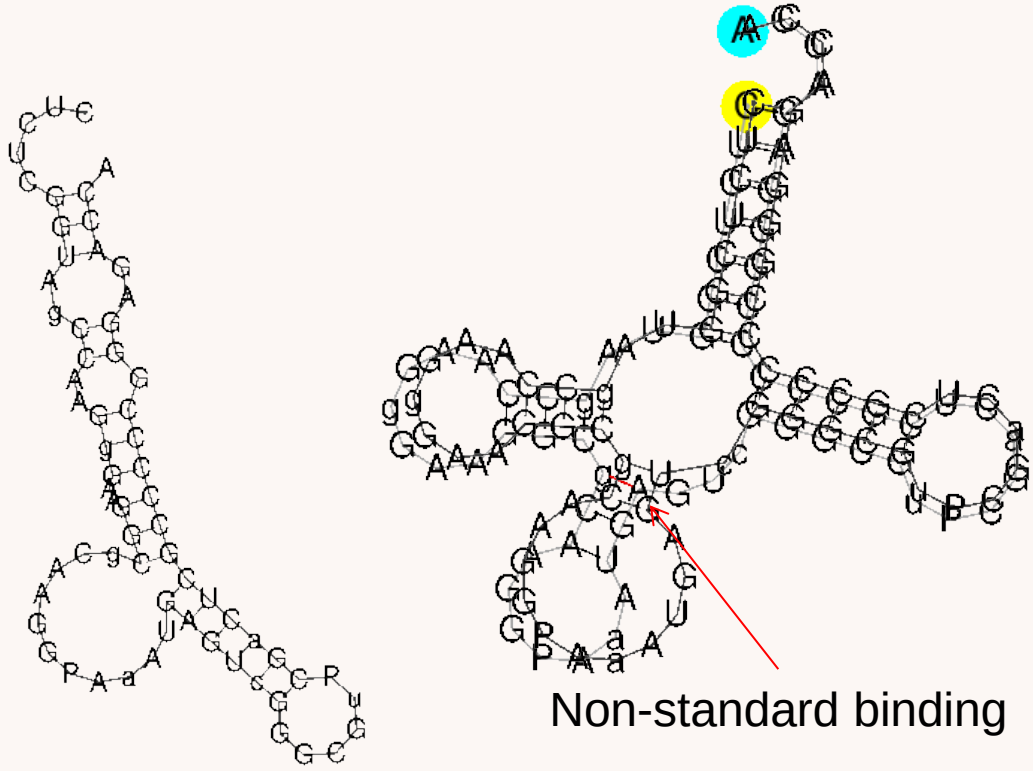
Original, MCC -0.008222

Mutant, MCC 0.856324

True

PDB 01001

yeast enzyme (in protein manufacture)



Non-standard binding

Original, MCC -0.008222

Mutant, MCC 0.856324

True

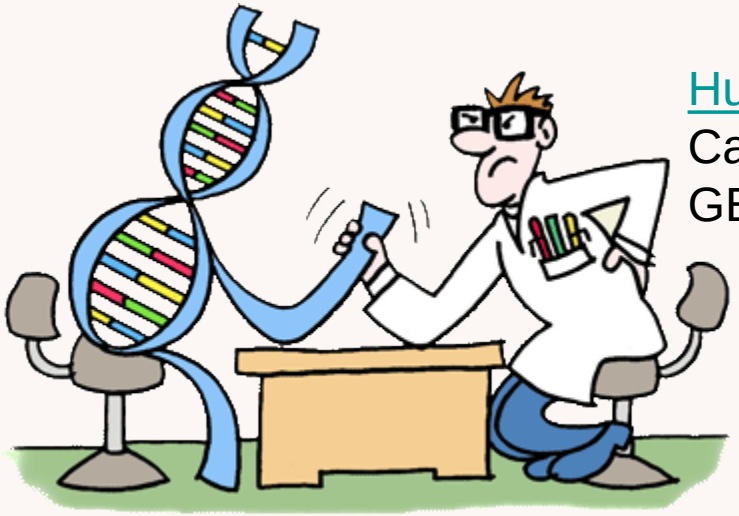
Summary

- GGGP applied to state-of-the-art RNA prediction tool on real data
- GGGP (SSE instructions) 31.9% speedup
  - Manual changes incorporated into official releases of ViennaRNA, 2190 downloads (14 April – 4 July). Used by EteRNA project.
- Better predictions
  - GGGP (code) so far modest improvement
  - GGGP 50000 parameters, cf deep parameters
    - 20% overall improved predictions

WIKIPEDIA
Genetic Improvement

ICSE

GI 2018, Göteborg, ICSE-2018 *proposed* workshop



Humies: Human-Competitive
Cash prizes
GECCO-2018

http://www.epsrc.ac.uk/ EPSRC

END

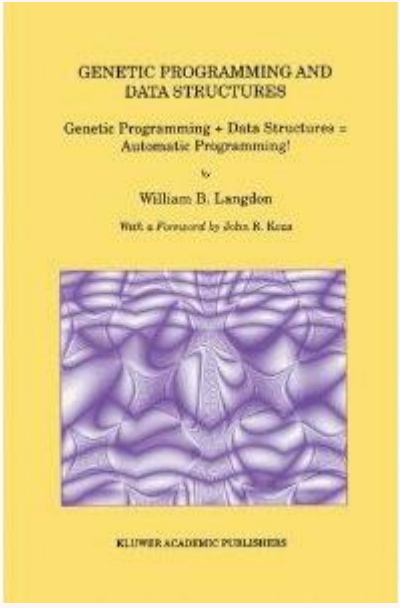
http://www.cs.ucl.ac.uk/staff/W.Langdon/ http://www.epsrc.ac.uk/ EPSRC

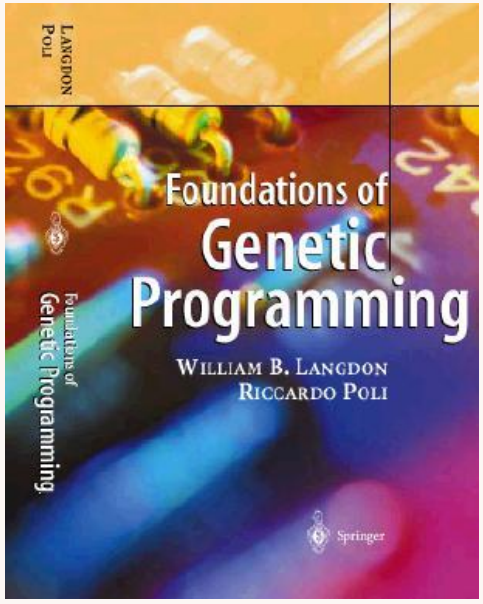
Genetic Improvement

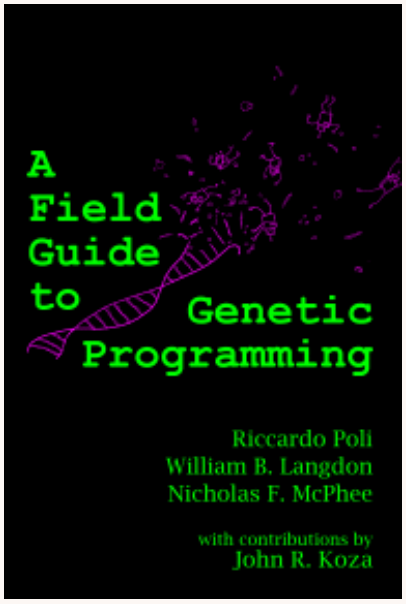
W. B. Langdon

CREST

Department of Computer Science

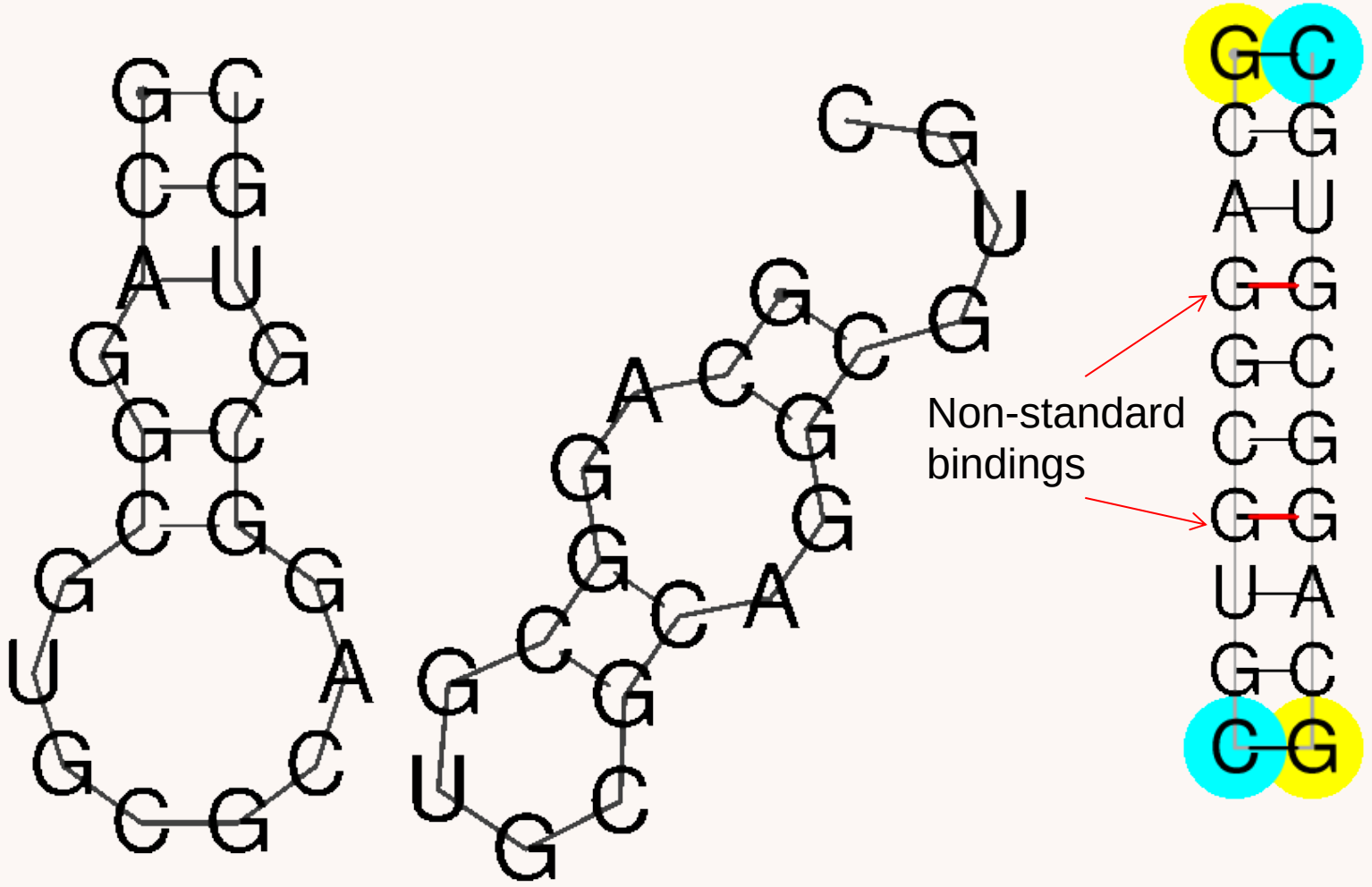






Worst training: PDB_00055

Synthetic RNA



Original MCC 0.697486

Mutant MCC -0.034565

True

The Genetic Programming Bibliography

http://www.cs.bham.ac.uk/~wbl/biblio/

11727 references, 10000 authors

Make sure it has all of your papers!
E.g. email W.Langdon@cs.ucl.ac.uk or use | Add to It | web link

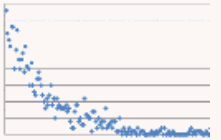

Co-authorships

RSS Support available through the Collection of CS Bibliographies.

Co-authorship community.
Downloads


Downloads by day

A personalised list of every author's GP publications.


Your papers

blog

Search the GP Bibliography at
http://liinwww.ira.uka.de/bibliography/Ai/genetic.programming.html