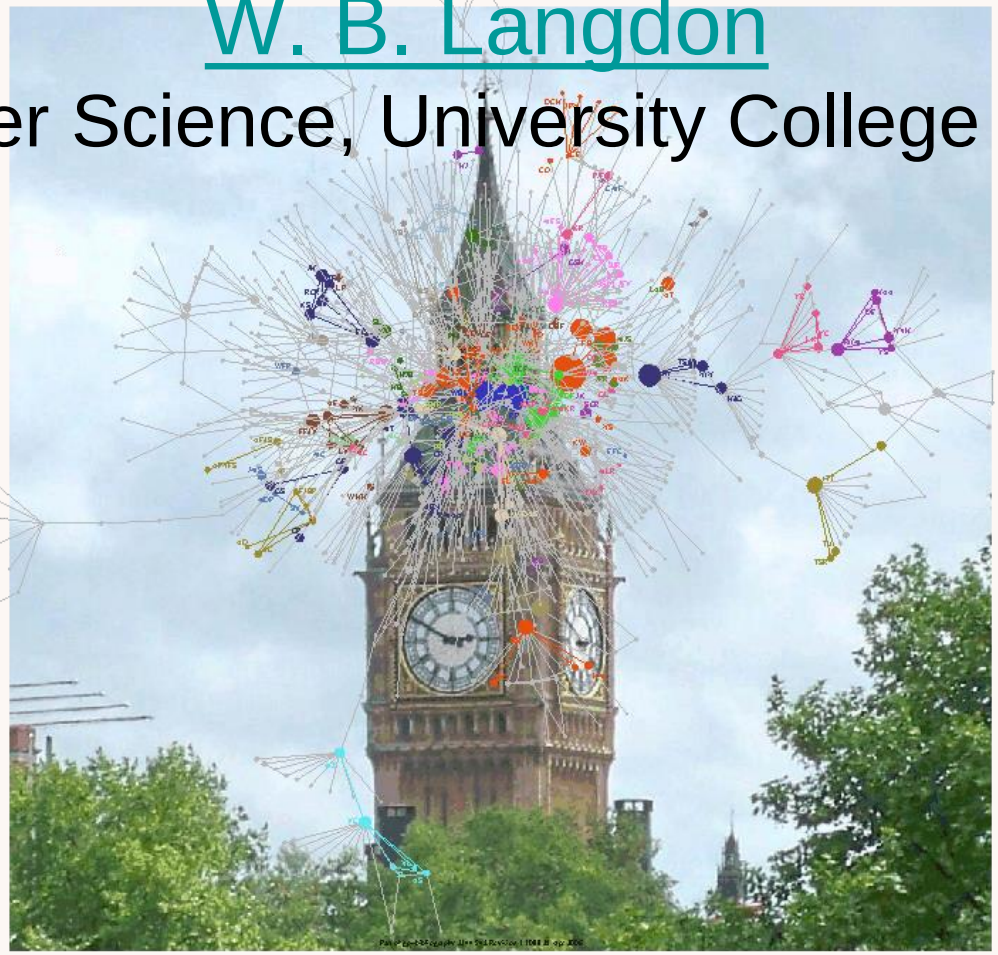
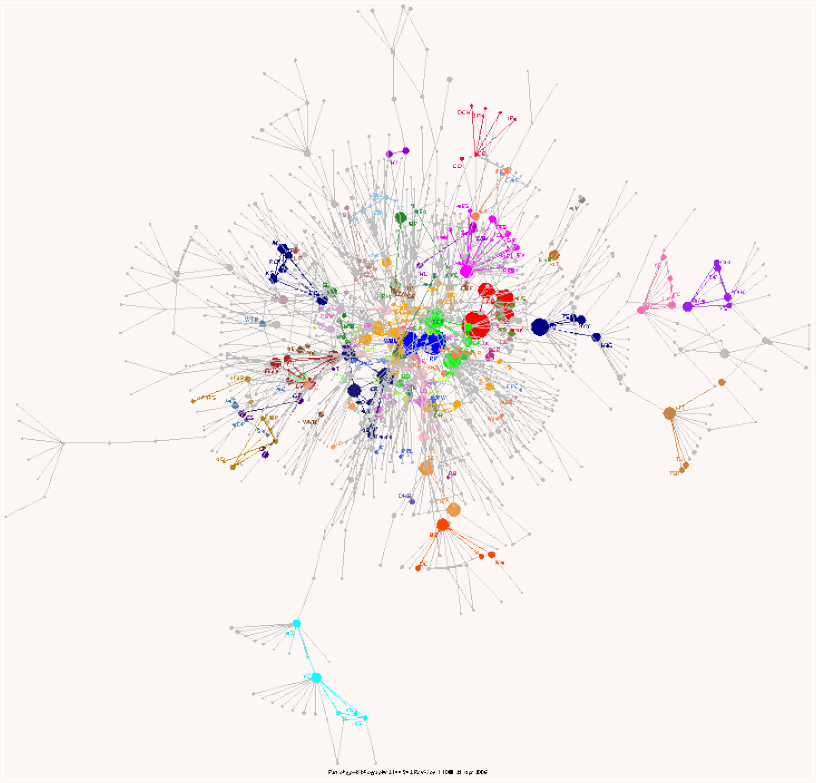
CREST Open Workshop COW61, 22nd October 2019

Data Driven Genetic Improvement

W. B. Langdon
Computer Science, University College London



Big Data, Legacy systems



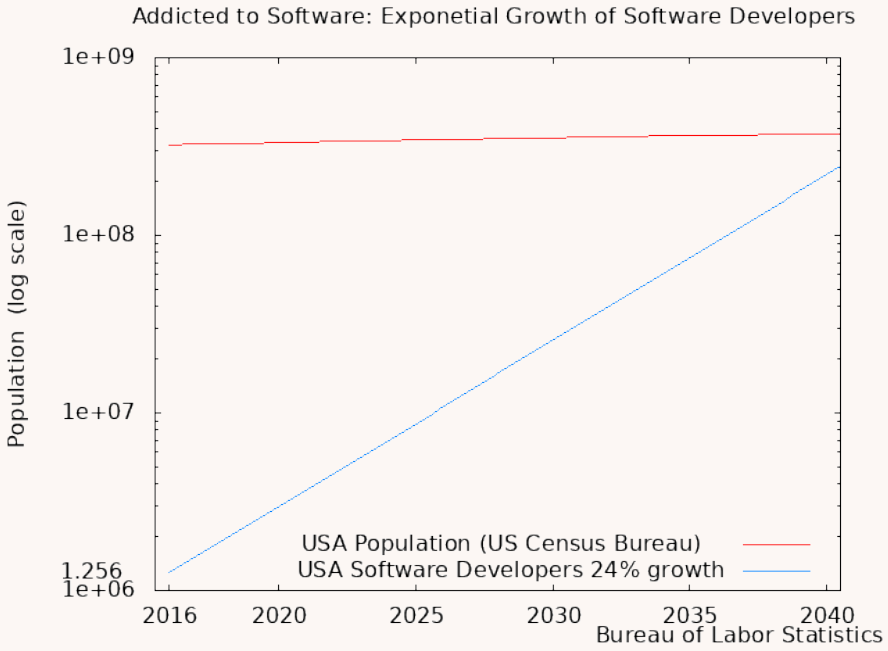
Big Data Legacy systems

Data Driven Genetic Improvement

- Need to automate software maintenance
- Search Based Software Engineering has concentrated on program source code
- FGIP: apply search to data in programs
- Better prediction of RNA structure
- What Next?
  - FGIP proposal submitted to EPSRC
- Conclusions

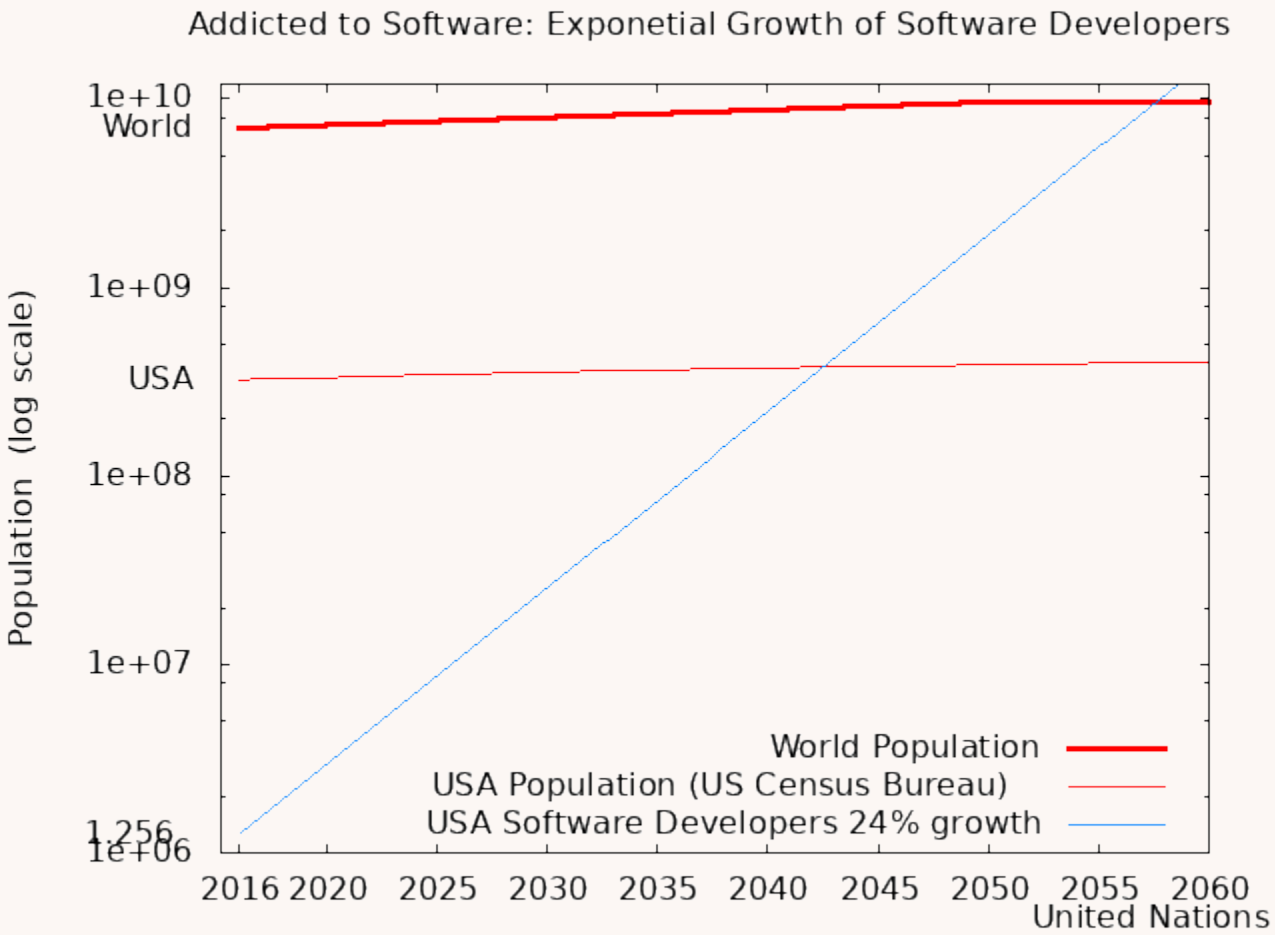
Need to Automate Software Maintenance

- Exponential increase in demand
- Cheaper/faster hardware does not help
- Cost of computing ~ cost of software
- Cannot exponential increase people in s/w



- Demand 24% per year
- 20 years most of USA are software developers
- Not possible, instead
- Need exponential improvement in software productivity
- Need to automate

Offshoring does not help



Desperate Need to Automate: SBSE so far

- Automatic code testing: eg EvoSuite
- Automatic bug fixing: eg C and Java code
- Genetic Improvement: eg faster, port code

Next

- Optimise software if objective measure
  - Automatic optimisation of program's *data*
    - more acceptable to programmers?
    - optimise *numbers*, use existing techniques
    - genetic search on *data*: eg
      - RNAfold, GNU C library
  - Others?

What is RNAfold

- RNAfold is the state of the art prediction of how RNA molecule will fold up based on its sequence of bases.

GI 33471 downloads (993826 www) since Apr 2017

- Open source program RNAfold 7100 lines of C source code.
- 51521 parameters (10 scalars+21 arrays)
- Training data ⅓ RNAstrand 4655 known structures

  (only use training sequences < 155 bases)

Training Data RNAstrand

http://www.rnasoft.ca/strand/

RNAstrand contains known RNA secondary structures.
4666 secondary structures in total.
Example screen shot for PDB_00865

www.rnasoft.ca/strand/show_results.php?molecule_ID=pdb_00865&Submit=Search+RNA+STRAND+ID

RNA STRAND v2.0 - The RNA secondary STRucture and statistical ANalysis Database

[Home | Search | Analyse | Submit structures | News | Help]

General features for molecule pdb_00865
(click to expand/contract all tables)

Format: CT — View the RNA sequence and secondary structure for molecule pdb_00865

View the output of the RNA Secondary Structure Analyser for molecule pdb_00865

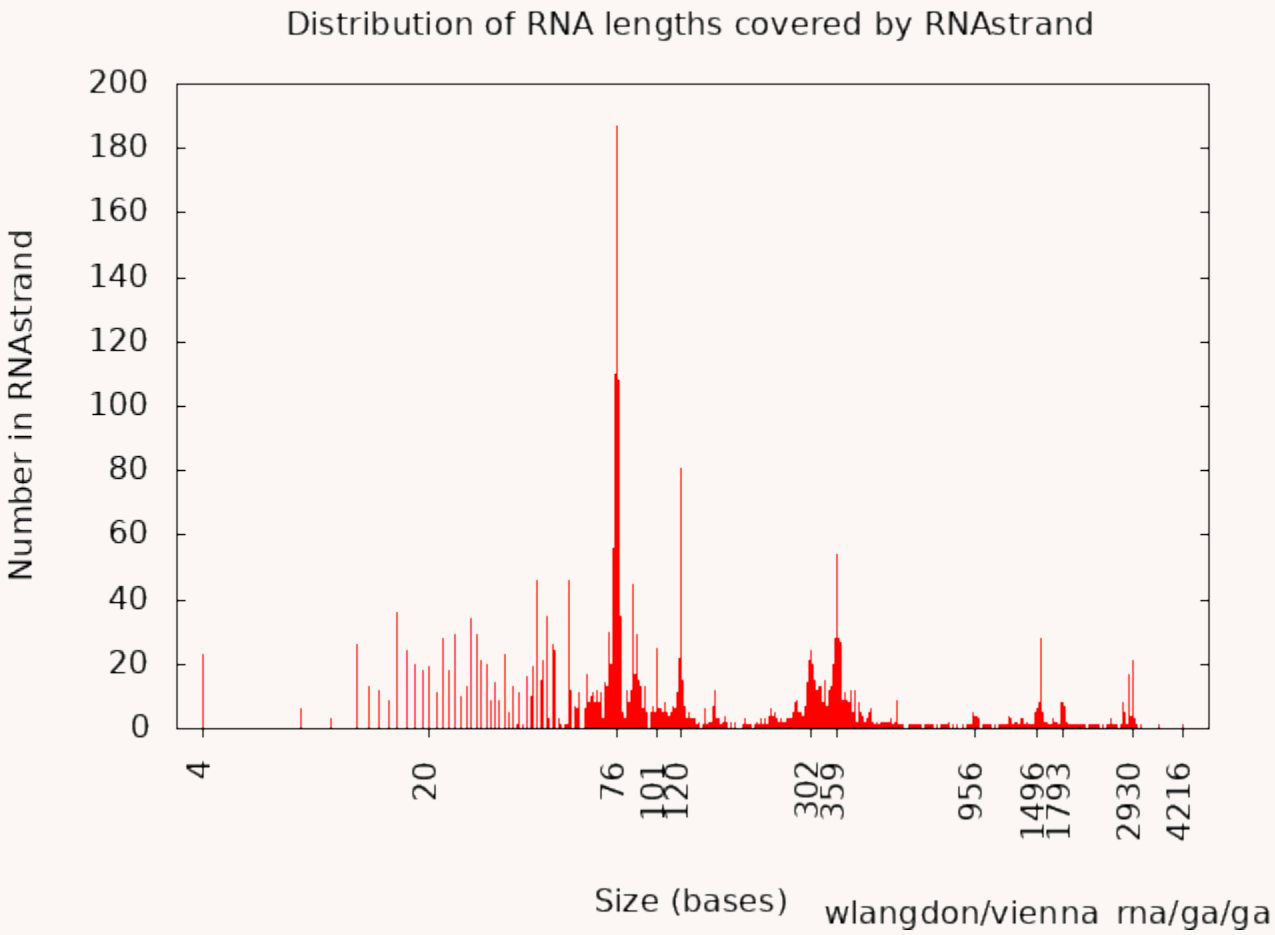
Molecule ID [?]:	PDB_00865
Molecule name [?]:	SOLUTION STRUCTURE OF THE CENTRAL REGION OF THE HUMAN GLUR- B R/G PRE-MRNA
Source [?]:	RCSB Protein Data Bank
Source ID [?]:	1YSV
Reference [?]:	R.STEFL,F.H.ALLAIN. A NOVEL RNA PENTALOOP FOLD INVOLVED IN TARGETING ADAR2.. RNA V. 11 592 2005 ASTM RNARFU UK ISSN 1355-8382
Type [?]:	Synthetic RNA
Organism [?]:	SYNTHETIC
Validated by NMR or X-Ray [?]:	Yes
Method for secondary structure determination [?]:	NMR; ran through RNAview
Number of molecules [?]:	1
Length [?]:	27
Fragments used [?]:	No
Duplicated sequence [?]:	No other molecule in the database has the same sequence
Number of domains [?]:	1
Number of unpaired bases [?]:	3
Number of paired bases [?]:	24

[?] PS figure PDF figure
Figure from original source

Structure centre of human GluR-B R/G pre-mRNA
https://www.rcsb.org/structure/1YSV

Bit rot, broken images

RNAstrand

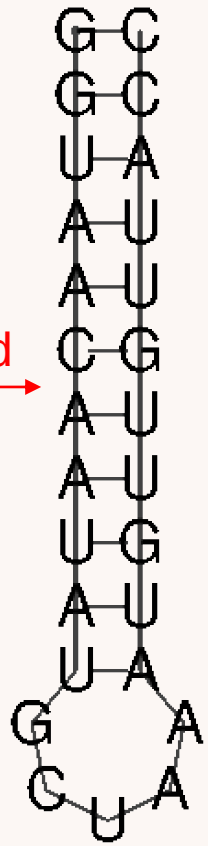


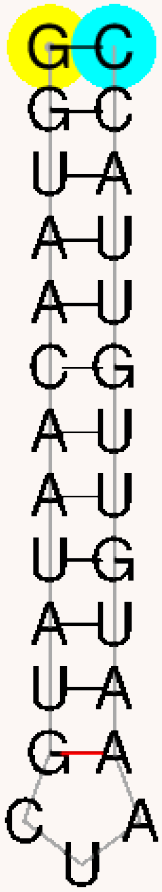
RNA sequence length, i.e. number of bases (log scale)

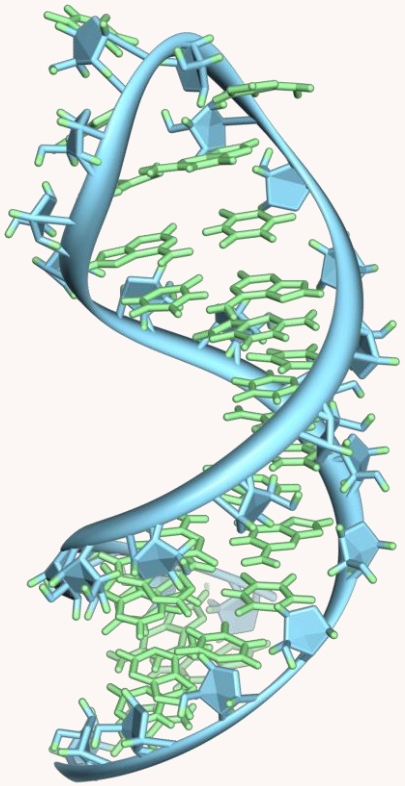
Compare RNAfold with RNAstrand

>PDB_00865
GGUAACAAUAUGCU
AAAUGUUGUUACC

RNAfold







Fasta text format
Input to RNAfold

Prediction
MCC 0.956018

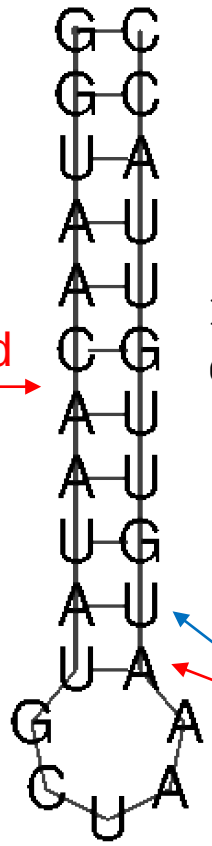
RNAstrand

three D picture

Compare RNAfold with RNAstrand

>PDB_00865
GGUAACAAUAUGCU
AAAUGUUGUUACC

RNAfold



Non-graphics output of RNAfold

```
>PDB_00865
GGUAACAAUAUGCUAAAUGUUGUUACC
((((((((((.....)))))))))) (-12.20)
```

Nested brackets, showing which base binds with another.
E.g. U↔A and A↔U

Calculated Binding energy

Fasta text format
Input to RNAfold

Prediction
MCC 0.956018

Nested brackets to connection matrix

PDB_00865.ct_rnafold

```
  GGUAACAAUAUGCUAAAUGUUGUUACC
G .........................x
G ........................x.
U .......................x..
A ......................x...
A .....................x....
C ....................x.....
A ...................x......
A ..................x.......
U .................x........
A ................x.........
U ...............x..........
G ..........................
C ..........................
U ..........................
A ..........................
A ..........................
A ..........x...............
U .........x................
G ........x.................
U .......x..................
U ......x...................
G .....x....................
U ....x.....................
U ...x......................
A ..x.......................
C .x........................
C x.........................
```

Prediction

PDB_00865.ct_rnastrand

```
  GGUAACAAUAUGCUAAAUGUUGUUACC
G .........................x
G ........................x.
U .......................x..
A ......................x...
A .....................x....
C ....................x.....
A ...................x......
A ..................x.......
U .................x........
A ................x.........
U ...............x..........
G ..............x...........
C ..........................
U ..........................
A ..........................
A ...........x..............
A ..........x...............
U .........x................
G ........x.................
U .......x..................
U ......x...................
G .....x....................
U ....x.....................
U ...x......................
A ..x.......................
C .x........................
C x.........................
```

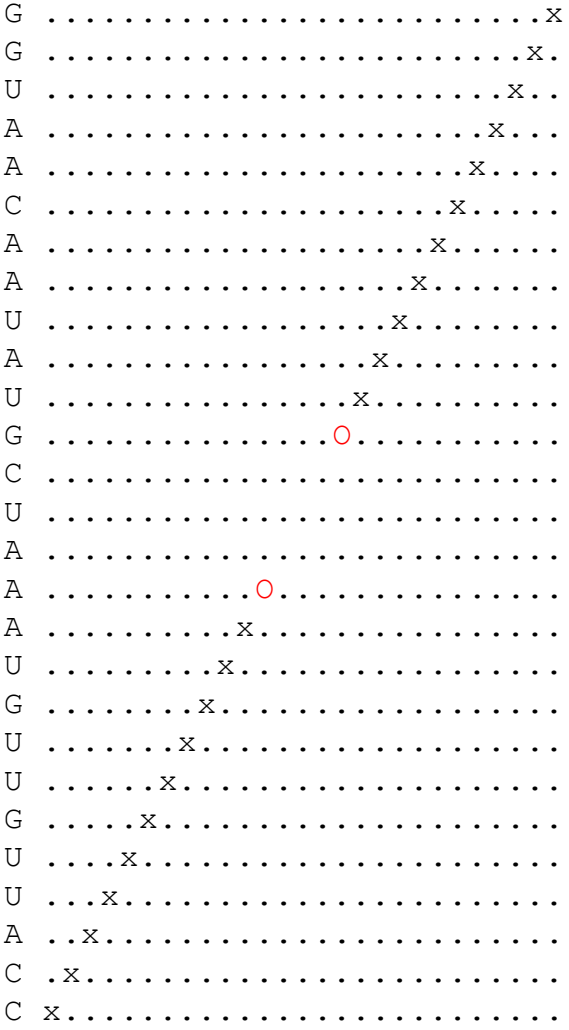
Ground Truth

Non-standard
G↔A pair

Compare RNAfold & RNAstrand matrices

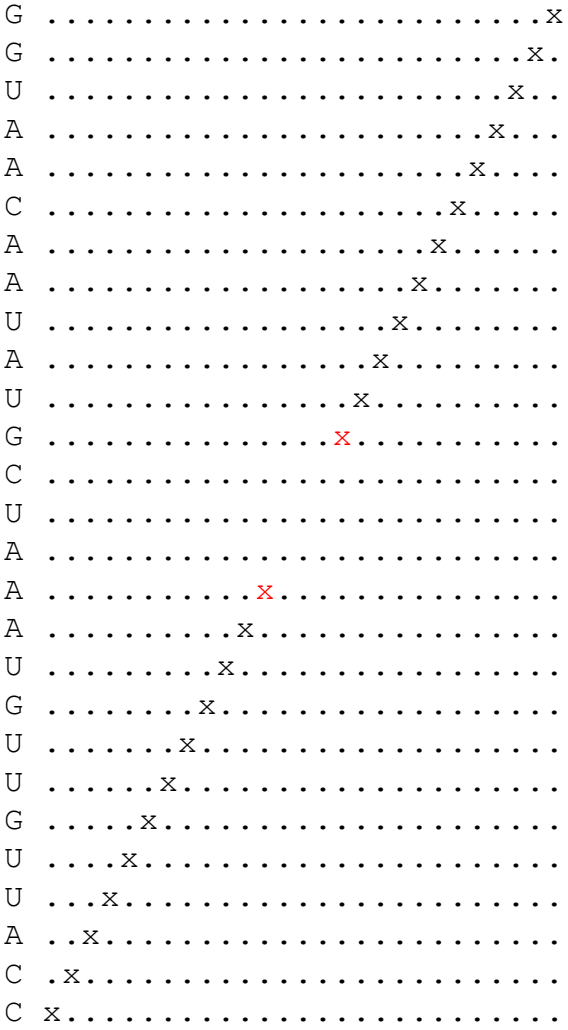
. TN
X TP
O FN
X FP

PDB_00865.ct_rnafold



Prediction

PDB_00865.ct_rnastrand

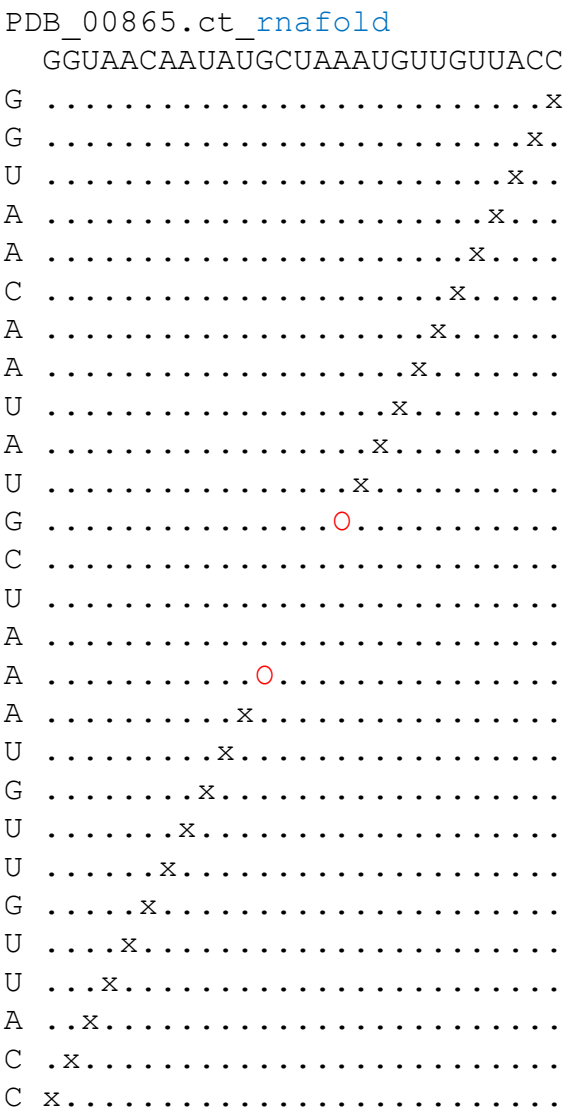


Ground Truth

Compare RNAfold with RNAstrand

. TN
X TP
O FN
X FP

PDB_00865.ct_rnafold



Prediction

.	TN	705
X	TP	22
O	FN	2
X	FP	0

729 = 27^2

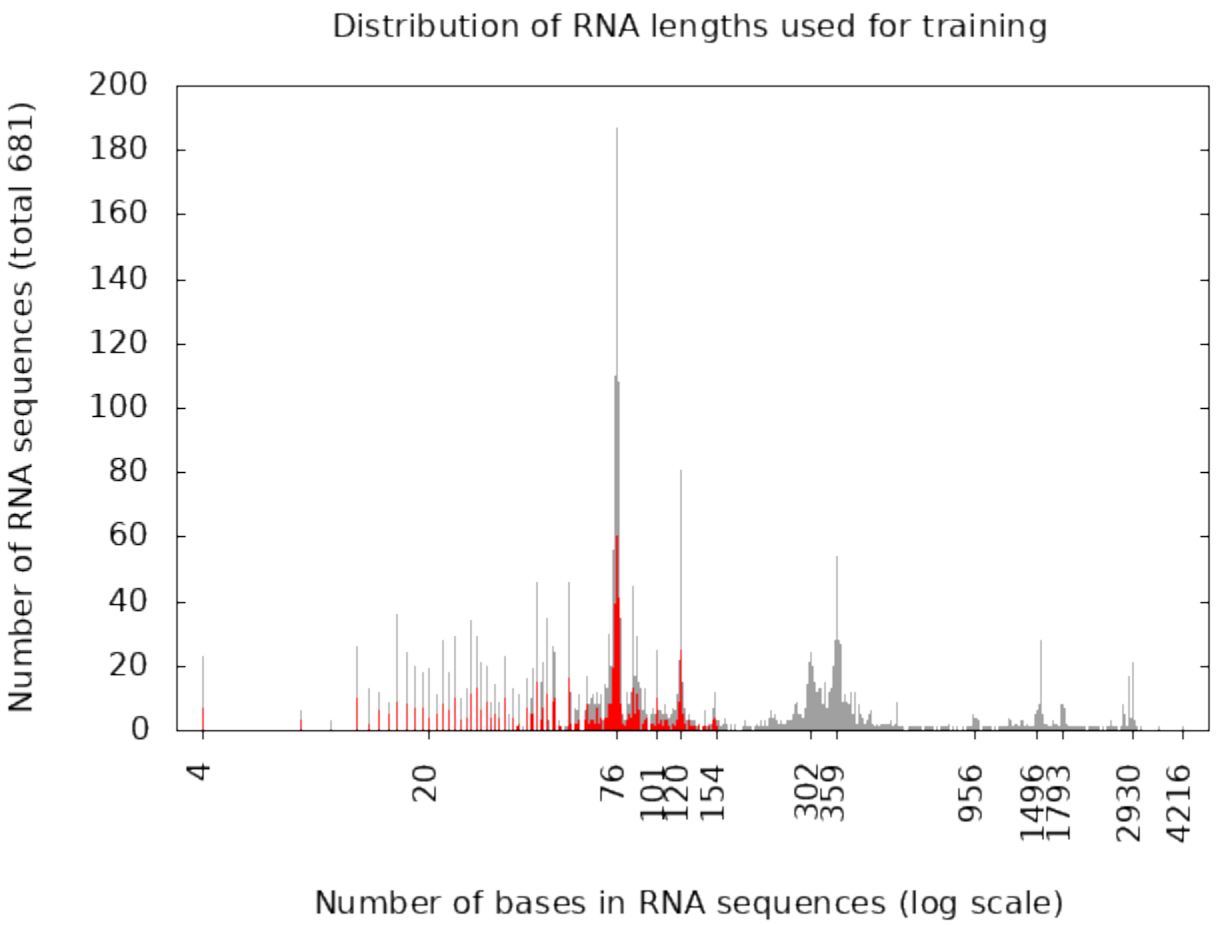
Matthews Correlation Coefficient

$$\frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}.$$

MCC = 0.9560

Product of true (TP×TN) minus product errors (FP×FN) normalised. MCC lies range -1 to +1

Training RNA sequences



⅓ data below 155 bases randomly selected for training

GI Fitness Function

- Run RNAfold with modified internal data on 681 short training RNA sequences
- Calculate Matthews Correlation Coefficient for each prediction.
- Selection fitness is mean MCC over 681 predictions. (Select top 50% of population)
- Ignore data mutations which make no difference on training RNA examples

51521 RNAfold parameters

Table 2. 31 (10 scalars + 21 arrays) RNAfold parameters which can be optimised. Data structures marked^{E} hold energy values which are always multiples of 10. (Mutation ensures they remain multiples of ten.) The original values of Tetraloop_E^{E} and Triloop_E^{E} are mostly zero [a] and so mutation of Tetraloop_E^{E} is limited to the first 15 elements and in Triloop_E^{E} to just the first element. NBPAIRS=7 and MAXLOOP=30.

noLP		mismatchM^{E}	[NBPAIRS+1][5][5]
uniq_ML		mismatchExt^{E}	[NBPAIRS+1][5][5]
dangles		dangle5^{E}	[NBPAIRS+1][5]
min_loop_size		dangle3^{E}	[NBPAIRS+1][5]
rtype	[8]	mismatchH^{E}	[NBPAIRS+1][5][5]
gquad		stack^{E}	[NBPAIRS+1][NBPAIRS+1]
special_hp		bulge^{E}	[MAXLOOP+1]
pair	[21][21]	int11^{E}	[NBPAIRS+1][NBPAIRS+1][5][5]
noGUclosure		int21^{E}	[NBPAIRS+1][NBPAIRS+1][5][5][5]
TerminalAU^{E}		internal_loop^{E}	[MAXLOOP+1]
MLintern^{E}	[NBPAIRS+1]	ninio[2]^{E}	
MLclosing^{E}		mismatch1nI^{E}	[NBPAIRS+1][5][5]
MLbase		int22^{E}	[NBPAIRS+1][NBPAIRS+1][5][5][5][5]
hairpin^{E}	[31]	mismatch23I^{E}	[NBPAIRS+1][5][5]
Tetraloop_E^{E}	[200] (15)	mismatchI^{E}	[NBPAIRS+1][5][5]
Triloop_E^{E}	[40] (1)		

total 51521 `int`

[a] The energy contributions for Tetraloop and Triloop are only used under special circumstances. They represent tabulated exceptions of small hairpin loops that do not follow the values provided in hairpin. They are only used when the sequences in question match the corresponding patterns stored in the character arrays Tetraloop and Triloop.

GI Representation

Variable length list of problem dependent mutations to data inside RNAfold.

Replace mutation > `mismatchM -60>-40`

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

Overwrite mutation < `mismatchH *,1,2<-80`

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

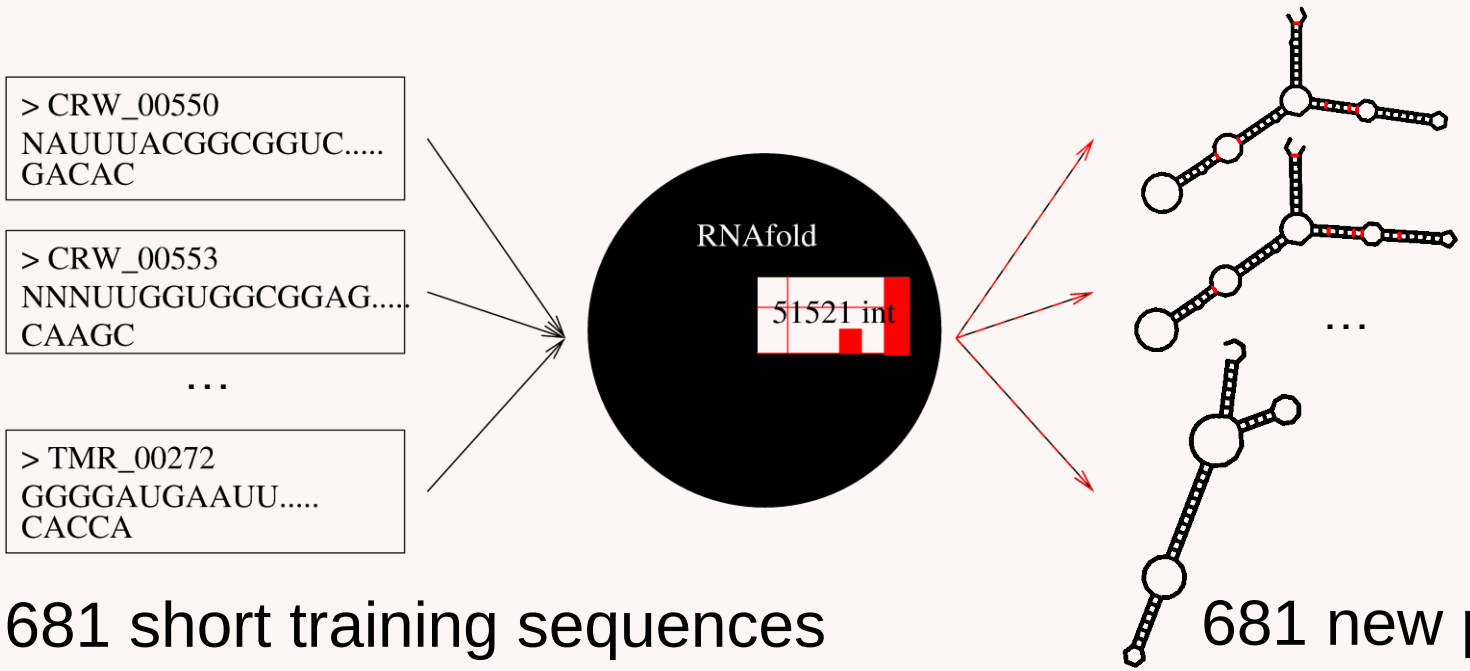
Increment mutation += `mismatchH *,*,*+=-90`

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

Creep mutation Small change (<20%) to value of existing mutations

Two point crossover

Fitness of Mutated RNAfold



681 short training sequences

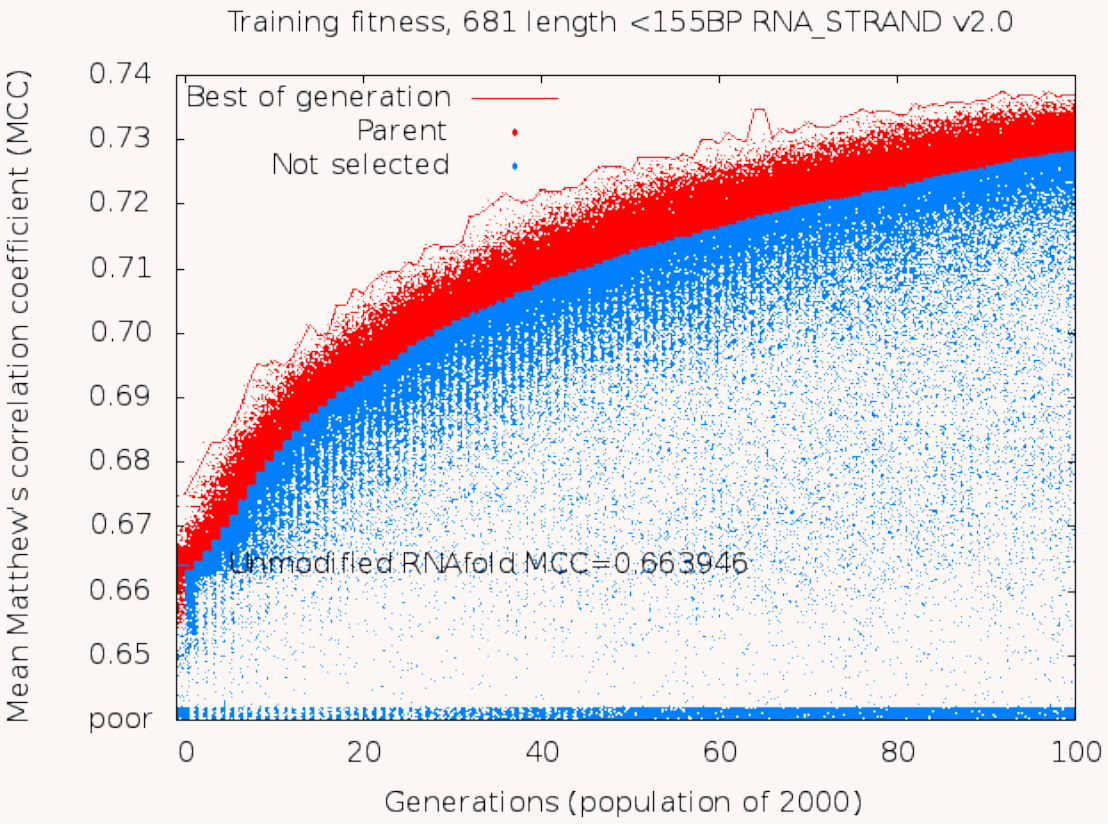
681 new predictions

- Mutate constants inside RNAfold and recompile
- Run mutated RNAfold on training RNA sequences
- Compare each new prediction with real structure
- Fitness mean Matthew's correlation coefficient on 681 training RNA molecules

GI RNAfold

- Pop 2000, 50% mutation 50% crossover
- Bloat removed. Best individual at gen 100 2849 mutations, hill climbing(2), 42 left

2000*101
fitness evals
<5 days
2.1sec each



Impact of 42 GI Changes

Table 3. Impact of the 42 components of the cleaned up evolved patches to 51 521 `int` paramters of RNAfold's dynamic programming model of RNA secondary structure. First column: components grouped by data structure (order in group is still significant). 2nd number of `int` changed. 3rd responsibility for fitness change (mutations build on each other, so isolated changes only give an indiaction of their importance). 4th again impact, this time on number of bonds changes across the whole training set. Last column describes changes with impact >2%. See also Sect. 3.3.

internal_loop *+=–40	29	–6.91%	667	Add 40 to internal_loop[2..30] ([0] and [1] are INF and so cannot be incremented)
MLintern *+=10 MLintern 3<–150	8	–3.25%	437	MLintern[0..7] were all –90, now -80 except [3] is –150
ninio[2] 80		–2.50%	501	Was 60 now 80
mismatch23I 70>10000000	108	–1.40%	131	
dangle5 *,*+=60	40	–1.27%	101	
int22 260>80 int22 180>280 int22 *,*,2,*,*,*+=10 int22 280>200 int22 200>10000000				
	10454	0.05%	37	
mismatchI *,*,0<100 mismatchI *,*,1+=–10 mismatchI 2,3,1+=–100 *,4,*+=–40				
	96	0.05%	617	
int11 *,*,*,*<200 int11 6,*,*,2+=–70	1600	1.22%	1306	
dangle3 5,*+=–80	5	1.28%	13	
mismatch1nI 70>110	125	1.89%	173	
TerminalAU 80		3.04%	759	Was 50 now 80
rtype 6<6 rtype 2+=1	2	3.05%	1257	[2] 1←2 and [6] was 5 becomes 6, page 14
mismatchExt *,*,*+=80 mismatchExt *,*,1<–40	200	3.90%	320	+80 is added to all elements, except 1 in 5 is set to –40
stack –100>60 stack –140>0 stack 2,2+=–20 stack *,4<–50				
	14	6.08%	2135	[0,4] 10000000← –50 [1,4] –140← –50 [1,7] –140←0 [2,2] –340← –360 [2,4] –150← –50 [3,5] –140←0 [4,1] –140←0 [4,4] 30← –50 [4,6] –100←60 [5,3] –140←0 [5,4] –60← –50 [6,4] –100← –50 [7,1] –140←0 [7,4] 30← –50
int21 230>260 int21 *,*,*,*,3+=–70 int21 220>10000000	1669	6.51%	287	283 values that were 230 replaced by 260. 161 values of 220 replaced by INF. And 1225 cases (of a possible 1600) where int21[*,*,*,*,3] is reduced by 70
bulge *+=40	30	7.53%	635	All bulge[1..30] increased by 40. ([0]is INF and so cannot be incremented)
mismatchM –70>–130 mismatchM *,3,*+=20 mismatchM *,1,*+=–40 mismatchM -110>–130 mismatchM *,0,*+=–170 mismatchM -60>–40	142	10.70%	1227	15 cases where –70 is replaced by –130. 2 cases where –110 is replaced by –130. 20 cases where -60 is replaced by -40. 40 cases where [*,0,*] is reduced by –170, 35 [*,1,*] by -40, and 30 [*,3,*] by –40
hairpin *<560	30	14.75%	1217	All hairpin[*] are set to 560 (Fig. 5)
mismatchH *,*,*+=–90 mismatchH *,*,3< –130 mismatchH *,1,2<–80	180	16.30%	1610	39 cases where mismatchH [*,*,3] is set to -130. 8 cases mismatchH [*,1,2] becomes -80 and 133 where other values in mismatchH are reduced by –90
Total:	14732			

14732 of 51521 (29%) changed

Improving RNAfold parameters

EuroGP-2018

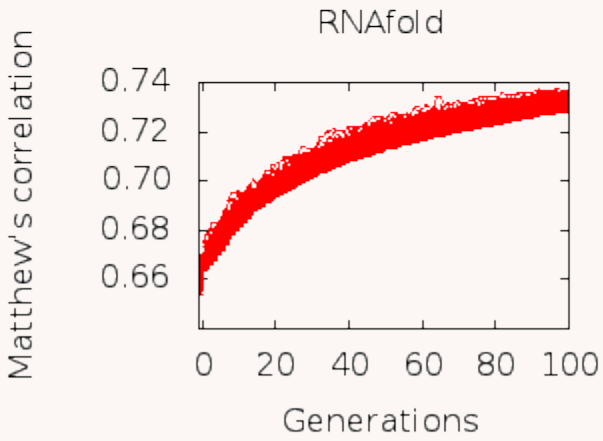
- RNAfold 7100 lines of C source code, 51521 parameters.
- Fitness correlation between prediction and true structure (Matthews Correlation, MCC).
- Post evolution tidy
- 14732 (29%) parameters changed
- Holdout set significant ($p\ 10^{-16}$) increase MCC
- Also better constrained optimisation ($p\ 10^{-15}$)
- GI parameters rna_langdon2018.par shipped with ViennaRNA since 13 Jun 2018

Automatic Software Maintenance

- In a world addicted to software, maintenance is the dominant cost of computing.
- Need to keep parameters up to date
  - New science (cf. RNAfold), new laws or regulations, new users, new user expectations
  - Change of load, new hardware (eg bigger RAM), automatic porting
  - Search can be fast:
    - cbrt < 5 minutes, $\log_2$ 6 secs, invsqrt 6 secs
- Little SBSE research
- Great scope for automation

Summary

- Problem of maintaining data in code ignored
- SBSE to optimize data
  - suitable training data
  - treat code as a black box.
- RNAfold on real data
  - No code changes
  - 50000 parameters 20% overall better prediction
- Rapidly generate cbrt, $\log_2$, invsqrt, reciprocal, etc.
- Software is not fragile



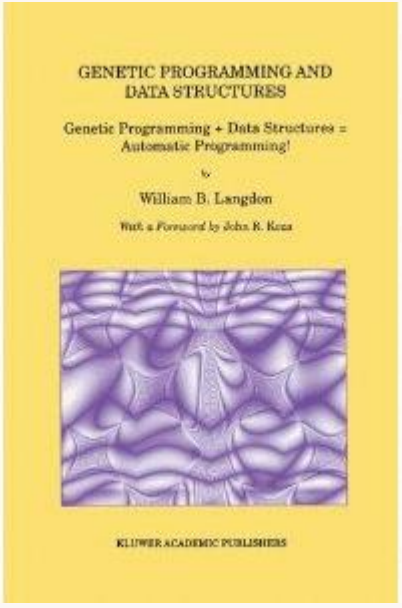
END

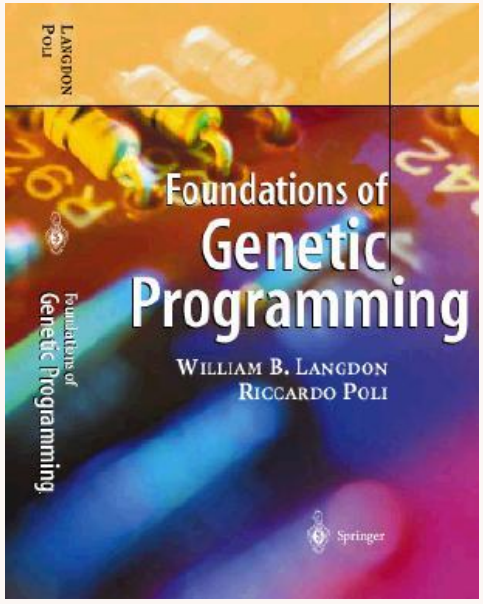
Genetic Improvement

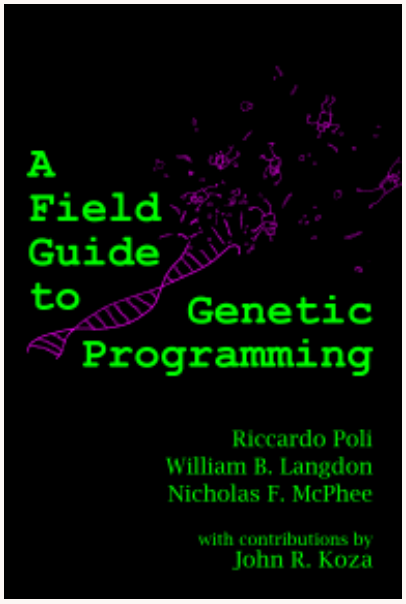
W. B. Langdon

CREST

Department of Computer Science







Genetic Improvement of RNAfold

- Speed up via Intel SSE parallel instructions GI 2017. Shipped since V2.3.5 2017-04-17
- GPU ViennaRNA Package v2.3.0cuda
- Better predictions by evolving parameters
  - On average better predictions of RNA folding.
  - Shipped since 2.4.7 2018-06-13
- AVX speedup in release 2.4.11 2018-12-17 EuroGP 2019

What has been done so far

- Mark/Fan Deep parameter tuning
- Holger Hoos et al. “constraint generation”
- RNAfold
- Converting GNU C library sqrt
  - Papers at SSBSE 2018, GECCO 2019, GI2019@GECCO
  - CREST visitor August 2019 Oliver Krauss

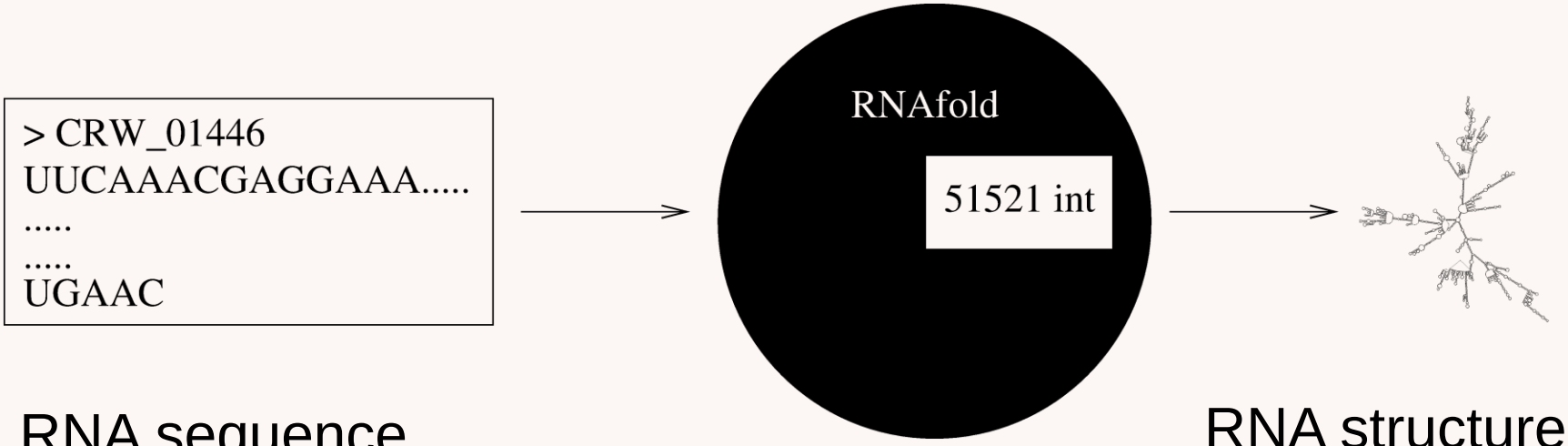
Fluid Genetic Improvement Programming

- New type of Genetic Improvement
- Update fluid embedded literals i.e. data
  1. New functionality
  2. Better non-functionality (e.g. faster)?
- Why
  1. FGIP is a new way to do GI, tackle data driven code
  2. Minimal code changes may be more acceptable?

Maintaining Embedded Constants

- EuroGP 2018
  - RNAfold 7000 lines of code 50000 numbers
  - On average better predictions of RNA folding.
  - Shipped since 2.4.7 rna_langdon2018.par
- CMA-ES evolves data in a GNU C library sqrt to give new functionality with double precision accuracy. sqrt converted to
  - cube root, cbrt
  - square root converted to $\log_2$
  - invsqrt $\frac{1}{\sqrt{x}}$
  - division less division, $\sqrt[4]{}$, etc.

RNAfold

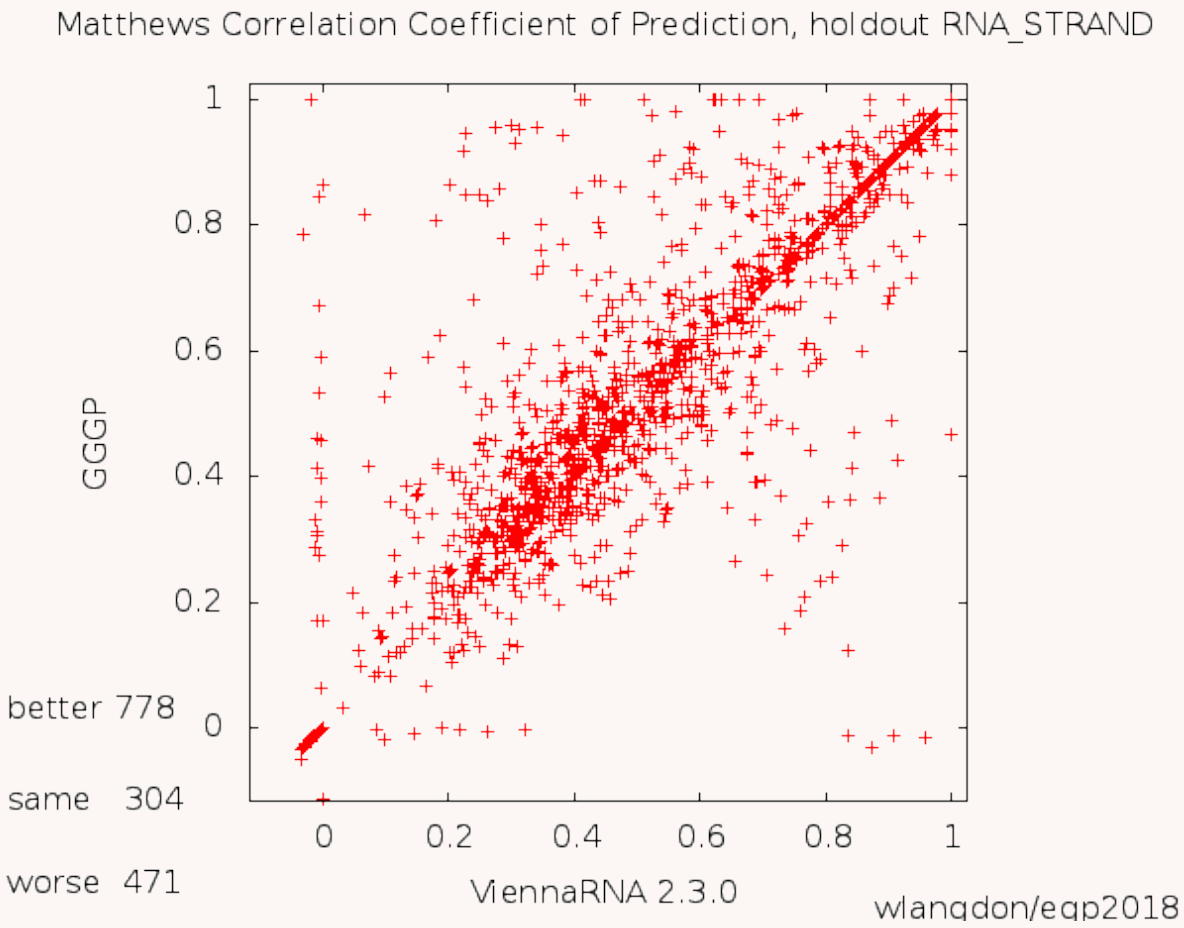


RNA sequence

RNA structure

RNAfold reads RNA molecules base sequence.
Outputs prediction of how molecule will fold up.
Internally RNAfold uses 51521 parameters.

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



Six impossible things before breakfast

- To have impact do something considered impossible.
- If you believe software is fragile you will not only be wrong but shut out the possibility of mutating it into something better.
- Genetic Improvement has repeatedly shown mutation need not be disastrous and can lead to great things.

Evolved $\frac{1}{\sqrt{x}}$ [GI@GECCO 2019]

Evolved cbrt tested many thousands of times

- Always within DBL_EPSILON
- Almost always gives best possible double

Compared to Quake (single precision approximation)

- Quake seldom gives exact answer
- Quake can be 0.17% wrong (0.43/256)
- Quake does not trap negative numbers, sometimes fails, sometimes just wrong
- Quake odd behaviour $<1.5\ 10^{-37}$ or $>3.3\ 10^{38}$

The Genetic Programming Bibliography

New home at UCL http://gpbib.cs.ucl.ac.uk

13401 references, 12000 authors

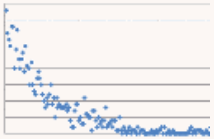
Make sure it has all of your papers!
E.g. email W.Langdon@cs.ucl.ac.uk or use | Add to It | web link

Downloads

Downloads by day

A personalised list of every author's GP publications.

blog



Your papers

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