Data Driven Genetic Improvement

W. B. Langdon
Computer Science, University College London
Big Data, Legacy systems

W. B. Langdon, UCL
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.

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

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

Bit rot, broken images
RNA sequence length, i.e. number of bases (log scale)
Compare RNAfold with RNAstrand

> PDB_00865
GGUAACAAUAUGCU
AAAUGUUGUUACC

Fasta text format input to RNAfold: RNAfold

Prediction:
MCC 0.956018

RNAstrand:
three D picture

W. B. Langdon, UCL
Compare RNAfold with RNAstrand

> PDB_00865
GGUAACAAU AUGCU AAA AUGUUU ACC

Non-graphics output of RNAfold

> PDB_00865
GGUAACAAU AUGCU AAA AUGUUU ACC
(((((((((((((.....))))))))))))) (-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

Prediction

Ground Truth

Non-standard G↔A pair
Compare RNAfold & RNAstrand matrices

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB_00865.ct_rnafold</td>
<td>PDB_00865.ct_rnastrand</td>
</tr>
<tr>
<td>GGUAACAAUAUGCUAAUGUUGUUACC</td>
<td>GGUAACAAUAUGCUAAUGUUGUUACC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TN</th>
<th>TP</th>
<th>FN</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>G</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>U</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>A</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>A</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>C</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>A</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>A</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>U</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>A</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>U</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>G</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>C</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>U</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>A</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>A</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>A</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>U</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>U</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>G</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>U</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>U</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>G</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>U</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>U</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>U</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>A</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>C</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>C</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>x</td>
</tr>
</tbody>
</table>
Compare RNAfold with RNAstrand

```
PDB_00865.ct_rnafold
GGUAACAAUAUGCUAAAUGUUGUUACC
```

<table>
<thead>
<tr>
<th>Prediction</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>. TN</td>
<td>. TN 705</td>
</tr>
<tr>
<td>X TP</td>
<td>X TP 22</td>
</tr>
<tr>
<td>O FN</td>
<td>O FN 2</td>
</tr>
<tr>
<td>X FP</td>
<td>X FP 0</td>
</tr>
</tbody>
</table>

```
729 = 27^2
```

Matthews Correlation Coefficient

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

\[
MCC = 0.9560
\]

Product of true (TP\times TN) minus product errors (FP\times FN) normalised. MCC lies range -1 to +1
Training RNA sequences

⅓ data below 155 bases randomly selected for training

W. B. Langdon, UCL
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
# Table 2

31 (10 scalars + 21 arrays) RNAfold parameters which can be optimised. Data structures marked with energies which are always multiples of 10. (Mutation ensures they remain multiples of ten.) The original values of Tetraloop\_E and Triloop\_E are mostly zero\(^a\) and so mutation of Tetraloop\_E is limited to the first 15 elements and in Triloop\_E to just the first element. NBPAIRS=7 and MAXLOOP=30.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Parameter Value</th>
<th>Energy Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>noLP</td>
<td>mismatch_M_E</td>
<td>[NBPAIRS+1][5][5]</td>
</tr>
<tr>
<td>uniq_ML</td>
<td>mismatch_Ext_E</td>
<td>[NBPAIRS+1][5][5]</td>
</tr>
<tr>
<td>dangles</td>
<td>dangle_3_E</td>
<td>[NBPAIRS+1][5]</td>
</tr>
<tr>
<td>min_loop_size</td>
<td>dangle_3_E</td>
<td>[NBPAIRS+1][5]</td>
</tr>
<tr>
<td>rtype</td>
<td>mismatch_H_E</td>
<td>[NBPAIRS+1][5][5]</td>
</tr>
<tr>
<td>gquad</td>
<td>stack_E</td>
<td>[NBPAIRS+1][NBPAIRS+1]</td>
</tr>
<tr>
<td>special_hp</td>
<td>bulge_E</td>
<td>[MAXLOOP+1]</td>
</tr>
<tr>
<td>pair</td>
<td>int_1_E</td>
<td>[NBPAIRS+1][NBPAIRS+1][5][5]</td>
</tr>
<tr>
<td>noGUClosure</td>
<td>int_2_1_E</td>
<td>[NBPAIRS+1][NBPAIRS+1][5][5][5]</td>
</tr>
<tr>
<td>Terminal_AU_E</td>
<td>internal_loop_E</td>
<td>[MAXLOOP+1]</td>
</tr>
<tr>
<td>ML_intern_E</td>
<td>[NBPAIRS+1]</td>
<td>ninio_2_E</td>
</tr>
<tr>
<td>ML_closing_E</td>
<td>mismatch_1_n_E</td>
<td>[NBPAIRS+1][5][5]</td>
</tr>
<tr>
<td>ML_base</td>
<td>int_2_2_E</td>
<td>[NBPAIRS+1][NBPAIRS+1][5][5][5][5]</td>
</tr>
<tr>
<td>hairpin_E</td>
<td>[31]</td>
<td>mismatch_2_3_1_E</td>
</tr>
<tr>
<td>Tetraloop_E_E</td>
<td>[200] (15)</td>
<td>[NBPAIRS+1][5][5]</td>
</tr>
<tr>
<td>Triloop_E_E</td>
<td>[40] (1)</td>
<td>[NBPAIRS+1][5][5]</td>
</tr>
</tbody>
</table>

\(^a\) 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 subtract 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

- 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

681 short training sequences

681 new predictions
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
2.1 sec each
# Impact of 42 GI Changes

Table 3. Impact of the 42 components of the cleaned up evolved patches to 51521 int parameters 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 indication 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.

| internalLoop | +49 | \(-0.91\%\) 667 | Add 40 to internalLoop(230) ([0] and [1] are INF and so cannot be incremented) |
| MLintern | +10 | \(-3.25\%\) 437 | MLintern(0.7) were all 99. now -80 except [3] is 150 |
| MLintern 3c | +150 | \(-2.50\%\) 501 | Was 90 now 80 |
| mismatch23T 799000000 | +108 | \(-1.40\%\) 131 |  |
| dangle5 **+60 | +40 | \(-1.27\%\) 104 |  |
| int22 320868 | int22 1809280 | int22 **+2 | int22 28990 | int22 200190000 | 1054 | 0.05% | 37 |
| mismatch **,0+100 | mismatch **,1+10 mismatch **,14+100 mismatch **,1+40 mismatch **,1+400 | 96 | 0.05% | 617 |
| int11 **,1+1200 | int11 6,0+20 | 1600 | 1.22% | 1306 |
| dangle3 5 **+80 | 5 | 1.86% | 13 |
| mismatchu1d 70+110 | 150 | 1.80% | 375 |
| TerminalA U 80 | 3.04% | 759 | Was 50 now 80 |
| rtype 660 rtype 2*+1 | 3 | 3.05% | 1257 [2] 1++2 and [0] was 5 becomes 6, page 14 |
| mismatchExt **,4+80 mismatchExt **,1<40 | 200 | 3.50% | 320 | +80 is added to all elements, except 1 in 5 is set to -40 |
| stack 100060 stack 1400 | stack 2,2+20 stack *4<50 | 14 | 6.08% | 2135 [0.4] 100060000 | -50 [1.4] -140+50 |
| | | [0.6] -100+60 [0.5] +100+0 [0.5] -60-50 |
| | | [0.6] -100+60 [0.5] +100+0 [0.5] 30-50 |
| int21 2390260 | 1600 | 6.51% | 287 | 383 values that were 330 replaced by 300. |
| int22 **,1,1+70 | 1600 | 4.14% | 220 replaced by INF. And |
| int22 220+10000000 | 1225 cases (of a possible 1060) where int21 **,1,1+70 is reduced by 70 |
| bulge **+50 | 30 | 7.33% | 639 | All bulge[1,39] increased by 40, ([0] is INF and so cannot be incremented) |
| mismatchM 70+130 | 142 | 10.79% | 1227 15 cases where -70 is replaced by -130. 2 cases where -110 is replaced by -130. 30 |
| mismatch **,1+10 mismatch **,2000 mismatch **,7,0+40 mismatch **,0,0+170 mismatch **,0+90 | 142 | 10.79% | 1227 15 cases where -70 is replaced by -130. 2 cases where -110 is replaced by -130. 30 |
| mismatch **,4<100 mismatch **,1,1+130 mismatch **,1,5<40 | 180 | 16.39% | 1810 39 cases where mismatchM [*,3] is set to -130. 8 cases mismatchM [*,1,2] becomes |
| | | -80 and 135 where other values in mismatchM are reduced by 90 |

Total: 14732 of 51521 (29%) changed
Improving RNAfold parameters

- 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:
    • \( \text{cbrt} < 5 \text{ minutes}, \log_2 6 \text{ secs}, \text{invsqrt} 6 \text{ secs} \)

• 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 \( \text{cbrt}, \log_2, \text{invsqrt}, \text{reciprocal}, \text{etc.} \)
• 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 `RNAfold2018.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 `log2`
  - `invsqrt` \[ \frac{1}{\sqrt{x}} \]
  - division less division, 4\sqrt{}, etc.
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

Matthews Correlation Coefficient of Prediction, holdout RNA_STRAND

W. B. Langdon, UCL
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.

W. B. Langdon, UCL
Evolved $\frac{1}{\sqrt{x}}$

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 $<1.5 \times 10^{-37}$ or $>3.3 \times 10^{38}$
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
A personalised list of every author’s GP publications.

blog

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