Infinite String Block Matching Features for DNA Classification

Wendy Ashlock
joint work with Sierra Gillis and Daniel Ashlock

Abstract

Automatic classification of DNA can be performed in a number of ways using a variety of features. This study introduces a novel technique for generating global features for DNA classification based on block-matching with infinite strings.

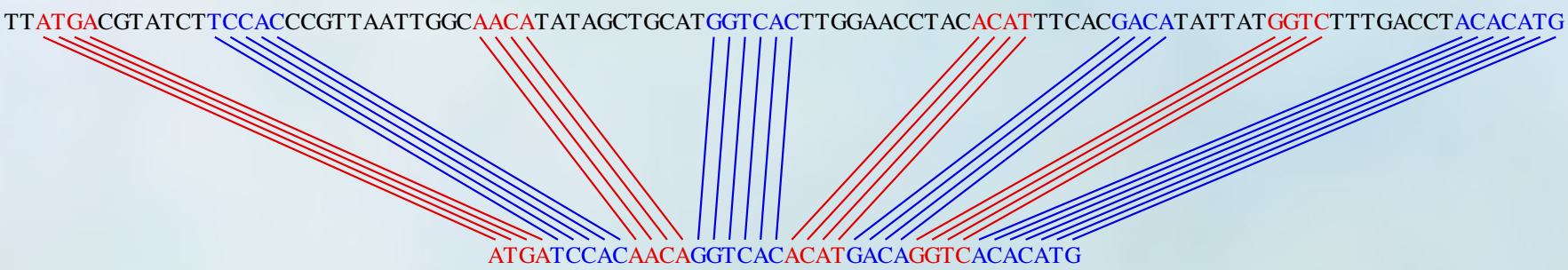
DNA features

Using machine learning to distinguish different classes of DNA is a natural application, but it suffers from a wealth of possible features.

- Features like *GC-content* or *Fourier transforms* are global and diffuse.
- Features like *motifs* are local and specific.
- *Side effect machines* tend to discover generalizations of GC-content.
- *Woven string kernels* are generalized motifs.

This presentation introduces a novel type of specific, global feature based on *block matching with infinite strings*, intended to fill a gap in the available features.

Block Matching Features



This diagram shows matching of a prefix of an infinite string (top) to a DNA sample (bottom). Matches have a minimum length, in this example 4, and the next such available block in the infinite string is matched to the sample. It the lengths of block are b_k then the matching score is:

$$R(S) = \sum_{k=1}^{n} b_k^2$$

which rewards longer blocks more than shorter ones.

Generating infinite strings with self-driving automata.

An example automata

		Transitions			
State	Response	On **C**	On **G**	On **A**	On **T**
0	AC	→7	→8	→6	→10
1	G	→5	→6	→1	→3
2	G	→0	→4	→1	→3
3	C	→1	→9	→6	→1
4	T	→10	→0	→3	→6
5	G	→1	→6	→7	→3
6	CG	→4	→11	→8	→6
7	T	→9	→6	→5	→10
8	T	→5	→7	→9	→9
9	A	→3	→1	→0	→3
10	A	→3	→2	→6	→3
11	A	→2	→4	→1	→2

Output: **ACCGTCGAGACTAGAGCCGAG...**

The automata starts in state 0 and emits "AC". Thereafter it drives itself with its own buffered output. This is the encoding for infinite strings – though only prefixes are generated in practice.

Experiments performed

Two kinds of data were used,

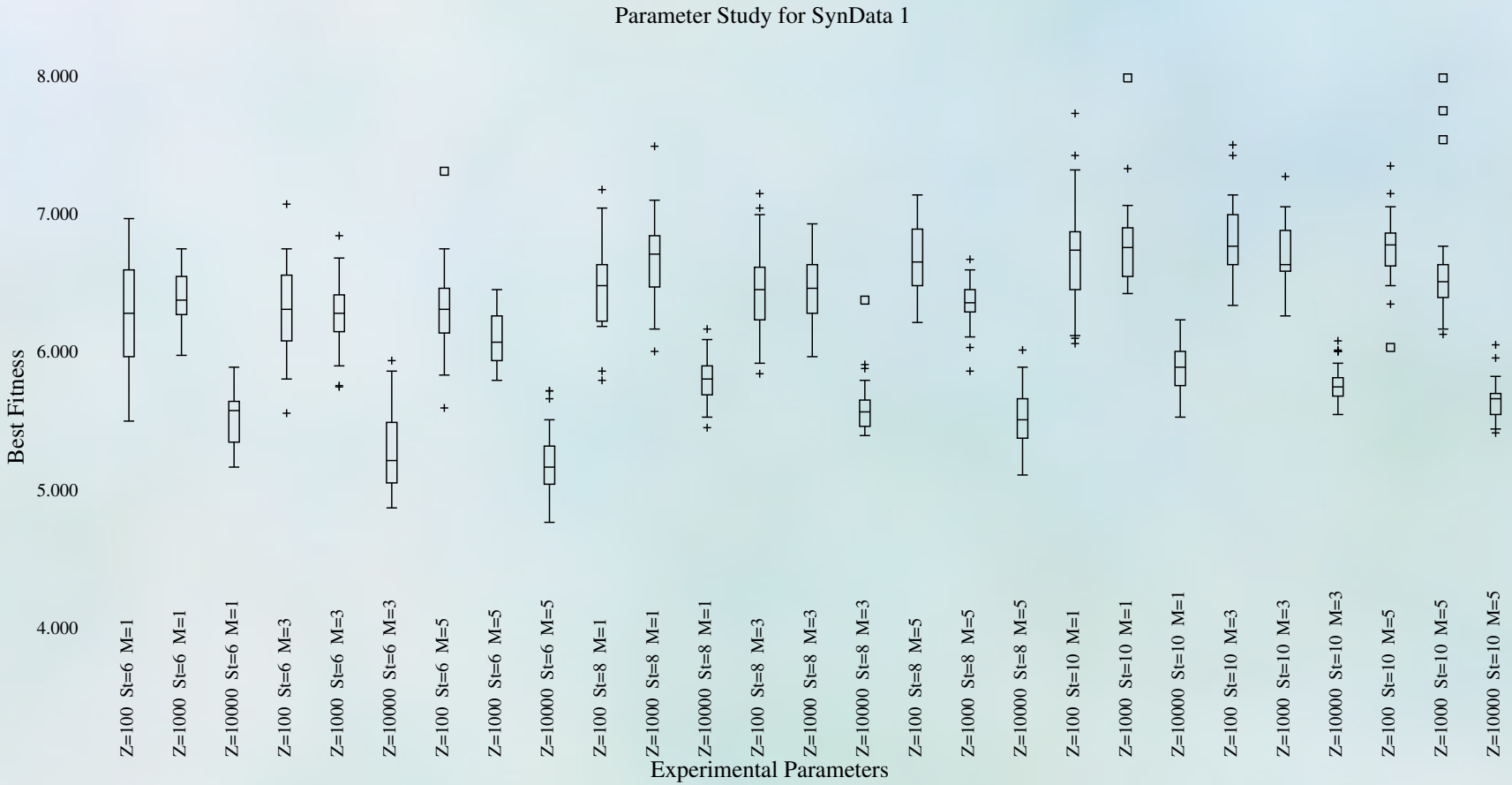
- six sets of synthetic data with different distributions of length-6 substrings
- A set of biological data selected from available human endogenous retroviruses (HERVs) and control sequences.

An evolutionary algorithm was used to evolve self-driving automata to maximize the log-average block matching score. For an infinite string generator $\mathcal{S}$:

$$Fit(\mathcal{S}) = Ln\left(1 + \frac{1}{q}\sum_{k=1}^{q} R(S_k)\right)$$

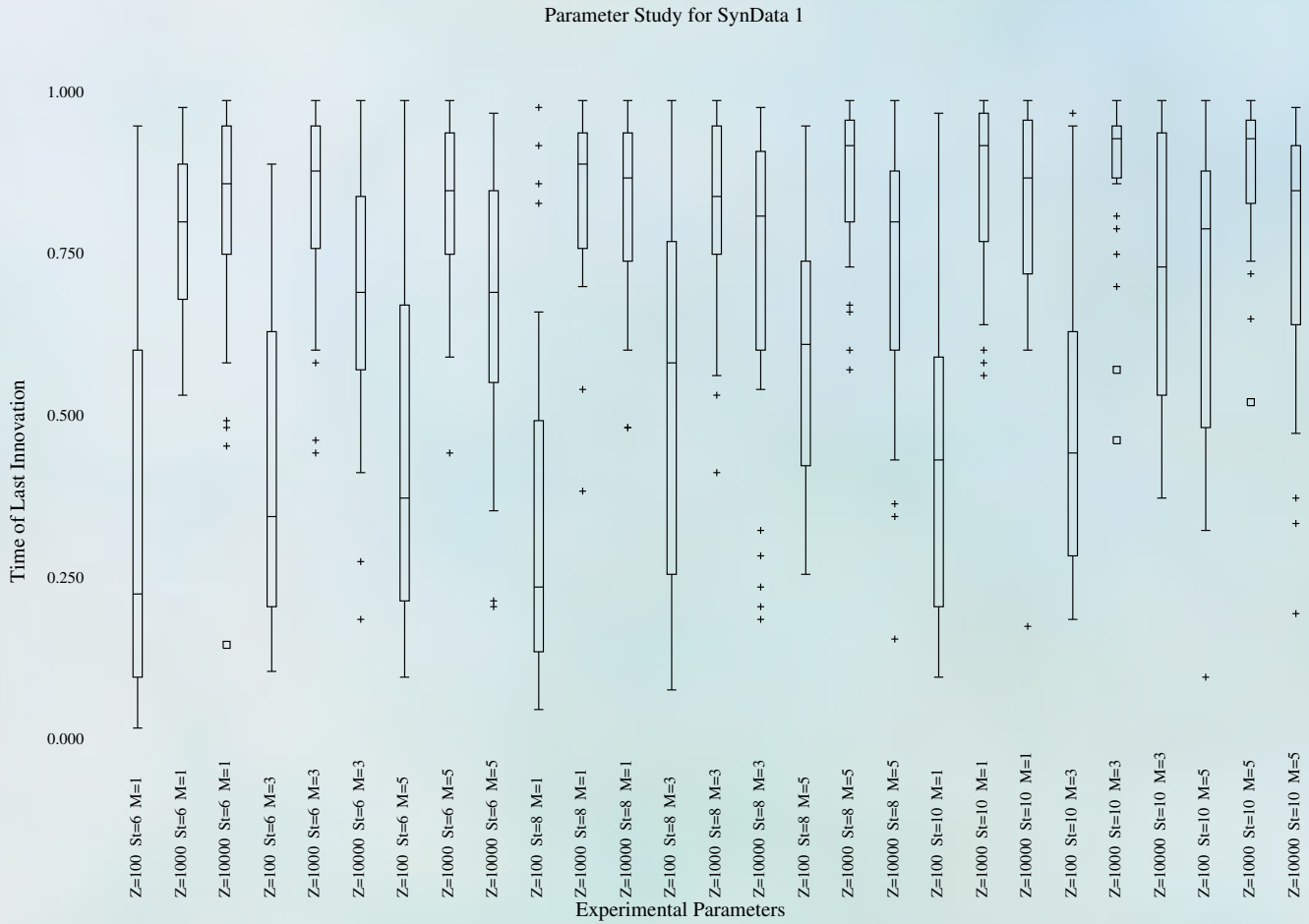
This fitness score is the value of the block matching feature on a set of data.

Parameter setting



Box plots for the final best fitness over 27 parameter setting experiments consisting of 30 replicates each with synthetic data set 2. Z is population size, S is states in the automata, M is the maximum number of mutations.

More parameter setting



Box plots for the *time of last innovation* for the same experiments. The time of last innovation is the fraction of the way through a run that the last increase in maximum fitness occurred.

Results: single feature classification with the block match number.

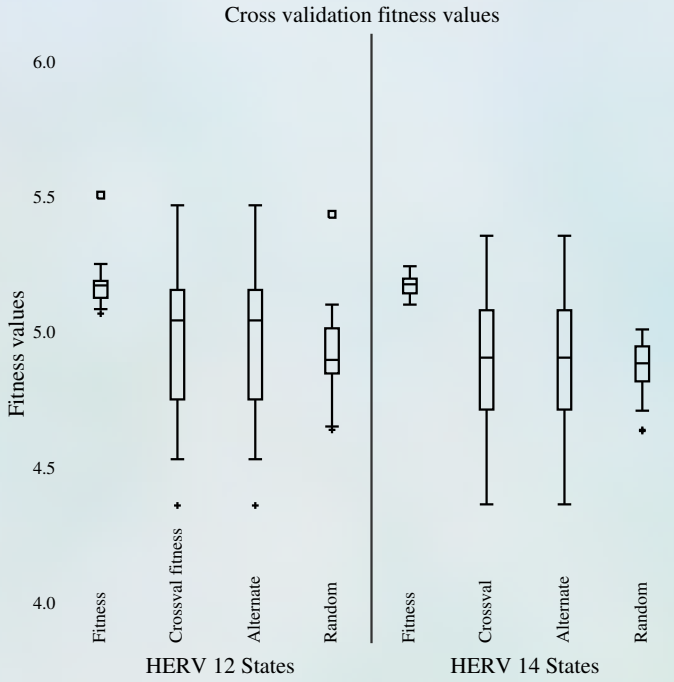
Sets of 30 additional runs were performed on all six synthetic data sets, using the parameters chosen during the parameter setting study.

Trained on data set	Evaluated on data set 1	2	3	4	5	6
1	**6.82±0.09**	3.66± 0.14	4.22± 0.15	3.44± 0.15	3.84± 0.11	3.71± 0.13
2	3.72± 0.15	**7.00±0.12**	4.02± 0.20	3.91± 0.14	3.47± 0.17	3.39± 0.12
3	3.94± 0.14	3.77± 0.16	**6.72±0.13**	3.40± 0.16	3.87± 0.16	3.63± 0.18
4	3.77± 0.13	4.01± 0.16	3.65± 0.19	**6.90±0.10**	3.33± 0.17	3.47± 0.17
5	3.66± 0.13	3.16± 0.15	4.17± 0.15	3.39± 0.19	**6.89±0.14**	4.71± 0.16
6	3.66± 0.14	3.15± 0.13	3.82± 0.16	3.49± 0.16	4.93± 0.14	**6.87±0.08**

These are the results for block match feature on individual strings from a reserved cross-validation set using the most fit infinite string generators available. The single feature classification performance is perfect on the data tested. This suggests that the block-matching fitness are an excellent type of feature.

Results: HERV Data.

The biological results were not as promising, but six of the thirty runs produces machines with good cross-validation behavior.



	Fitness:			
Run	Training	CrossVal	Alternate	Random
3	5.78	5.79	5.09	5.58
4	5.67	5.83	4.60	5.64
5	5.86	5.81	4.38	5.64
15	5.84	5.84	4.85	5.58
17	5.79	5.81	4.68	5.69
29	5.85	5.87	4.79	5.64

Scores for six runs with best cross-validation behavior.

These results suggest more work is needed to achieve the same sort of single feature classification observed on the synthetic data, but that the block match feature is still worth testing in a feature-selection context.

Next?

- Test block-match features in groups with other features with a feature selection algorithm.
- Document that block-match features are detecting a different type of signal from other types of features.
- Experiment with the minimum acceptable block length parameter.
- Experiment with more types of biological data.
- Check the impact of using a different infinite string generator.

Questions?

Many Thanks

Thanks to the *Natural Science and Engineering Research Council of Canada* and the *University of Guelph* for support of this work.