Creating Regular Expressions as mRNA Motifs with GP to Predict Human Exon Splitting

W. B. Langdon, J. Rowsell, A. P. Harrison

Crest, Department of Computer Science, King's College, London, WC2R 2LS, UK
Departments of Mathematical and Biological Sciences, University of Essex, UK

Wi11iam.Langdon@kcl.ac.uk,Mark.Harman@kcl.ac.uk,Yue.Jia@kcl.ac.uk

ABSTRACT

RNAnet [3] `http://bioinformatics.essex.ac.uk/users/wlangdon/rnanet/` allows the user to calculate correlations of gene expression, both between genes and between components within genes. We investigate all of Ensembl and find all the Homo Sapiens exons for which there are sufficient robust Affymetrix HG-U133 Plus 2 GeneChip probes. Calculating correlation between mRNA probe measurements for the same exon shows many exons whose components are consistently up regulated and down regulated. However we identify other Ensembl exons where sub-regions within them are self consistent but these transcript blocks are not well correlated with other blocks in the same exon. We suggest many current Ensembl exon definitions are incomplete.

Secondly, having identified exon with substructure we use machine learning to try and identify patterns in the DNA sequence lying between blocks of high correlation which might yield biological or technological explanations. A Backus-Naur form (BNF) context-free grammar constrains strongly typed genetic programming (STGP) to evolve biological motifs in the form of regular expressions (RE) (e.g. `TCTTT`) which classify gene exons with potential alternative mRNA expression from those without. We show biological patterns can be data mined by a GP written in `gawk` and using `egrep` from NCBI's GEO database. The automatically produced DNA motifs suggest that alternative polyadenylation is not responsible. (Full version in TR-09-02 [7].)

Blocky exons can be found in `http://bioinformatics.essex.ac.uk/users/wlangdon/tr-09-02.tar.gz`

Categories and Subject Descriptors: J.3 [Life and Medical Sciences]: Biology and Genetics

General Terms: Experimentation

1. INTRODUCTION

Particularly in Man, there are multiple ways messenger RNA is transcribed from its gene. E.g. alternative splicing may lead to the removal of exons or to exons being repeated. Usually it is assumed that exons are indivisible. Usually measurements of mRNA taken at various positions in an exon are highly correlated. However Figure 1 shows this is not true for an exon in Aspartyl-tRNA synthetase. Indeed we have found several hundred other Homo sapiens Ensembl exons with pronounced structure in their correlation heatmaps.

GECCO'09, July 8–12, 2009, Montreal, Canada
ACM 978-1-60558-325-9/09/07.

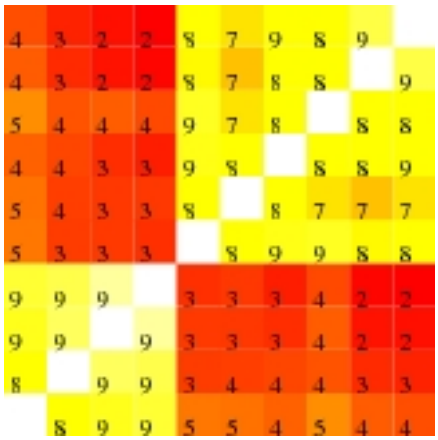

Figure 1: Correlation ($\times 10$) between different locations in an Ensembl exon across 2757 tissue samples. The first four locations (lower left) clearly fall into a different block than the others (top right).

Having identified these DNA sequences, we used strongly typed grammar based [5] GP [1, 6, 9] to evolve biologically meaningful motifs to explain them using only the mRNA sequences. A motif gets high fitness if it matches many blocky exons (i.e. positive examples) but fails to match many exons with high correlations but without blocks (negative examples).

We excluded suspect GeneChip data [8], suspect probe sequences [4] and probes which map to more than one exon [11] from the training data. The regular expression grammar in [4] was used, except `^` and `$` were omitted and the maximum Kleene closures was 7 instead of 9.

2. FITNESS OF STGP RE MOTIFS

In each generation, GP generates a unix command file which contains an `egrep -c '`*RE*`'` command for each individual in the population [2]. (*RE* is the individual's regular expression.) The command is run on a file holding the 100 sequences lying between two blocks. (Collars of 50 additional bases mean the last 50 bases of the first well correlated block and the first 50 bases of the trailing well correlated block are also included.)

`egrep -c` counts the number of probes which match the evolved motif (*RE*). The same command is also run on a file holding the 100 sequences of exactly the same length taken from exons which do not contain well separated blocks.

The fitness score of the regular expression is the difference between the number of lines in the two files which match *RE*. Expressions which either match all probes or fail to match any are penalised by subtracting 100 from their score. Each generation a new 100 positive and 100 negative sequences are used.

3. RESULTS

The best of generation 36 in the first GP run, with a population of 10 000 was the regular expression `TCTTT|TCGT|GG+TCAA|TTTGCCA` It has about the same performance on all the training data as it did on the two hundred exons actually used in generation 36. Whilst its performance falls on the 138 sequences used for validation it is still predictive. Most of the power of the evolved expression comes from its first term (`TCTTT`). In contrast the polyA deterministic regular expression `AATAAA.*T`$\not$`A(T|C)`$\not$`AT(G|T)`$\not$`A` [10] ($\not$A means any base except `A`) does not differentiate between mRNA sequences with blocks and those in exons without sub-blocks. This suggests that alternative polyadenylation is not responsible for the observed blocks of correlations.

4. BLOCKS DUE TO EXON-INTRON-EXON?

If the blocky big Ensembl exons are in fact two exons then there should then be an intron between the two exons. One signature of introns is a "polypyrimidine tract", which is a sequence rich in Cs and Ts. In other work we trained the GP using all the exons, i.e. including very long gaps between blocks. The evolved motifs contained runs of Cs and runs of Ts. This is reasonable if the gap is indeed due to a polypyrimidine tract associated with an undetected intron.

5. CONCLUSIONS

Hundreds of thousands of correlation coefficients for thousands of exons across thousands of people support many of the current definitions of exons. However in about a thousand cases the calculation of large scale correlations of Affymetrix GeneChip data suggests something unexpected: sub-structure within what were previously thought of as "atomic" exons.

We have used a grammar based strongly typed GP to automatically design a motif of the type biologists are familiar with. If biologists are to try an interpret our results it is important to present them in a "biologist friendly" form, rather than as a decision tree or a set of support vectors. The evolved motif uses only DNA sequence data and yet it has some ability to predict sub block structures within existing exons.

None of the four components of the evolved regular expression `TCTTT|TCGT|GG+TCAA|TTTGCCA` (and hence none of the strings it matches) themselves match either the initial indicator of polyadenylation (known as PAS) nor the "U rich" or "GU rich" signal after the cleavage site (known as DSE) [10]. If either PAS or DSE were responsible for the block structure we are seeing in Ensembl exons, we would expect both GP and our hand created regular expression should detect them. Our sequences contain more polyA sequences than would be expected by chance but the polyA regular expression does not differentiate between the sequences between correlated blocks and the negative examples. I.e. the PAS/DES motifs do not explain the observed correlations.

The fact that sometimes strong blocks of mRNA measurements for the same Ensembl exon are both not well correlated and are separated by thousands of bases along the transcript suggests that they are indeed separate exons and so Ensembl should not have grouped them together as one.

Alternative splicing and alternative polyadenylation are relatively new discoveries. The regulation of both is not well understood. It is reasonable to suggest that current bioinformatic databases may not be complete and that discoveries remain to be made. As increasingly large quantities of data from multiple disparate sources are available algorithmic tools like correlation will be more widely used. Evolutionary computation is increasingly being used in bioinformatics to aid our understanding of new aspects of biology.

6. REFERENCES

[1] Langdon, W. B. *Genetic Programming and Data Structures*. Kluwer, 1998.

[2] Langdon, W. B. Evolving GeneChip correlation predictors on parallel graphics hardware. In *2008 IEEE World Congress on Computational Intelligence* (Hong Kong, 1-6 June 2008), J. Wang, Ed., IEEE Computational Intelligence Society, IEEE Press, pp. 4152–4157.

[3] Langdon, W. B. A map of human gene expression. Tech. Rep. CES-486, University of Essex, Colchester, CO4 3SQ, UK, July 2008.

[4] Langdon, W. B., and Harrison, A. P. Evolving DNA motifs to predict GeneChip probe performance. *Algorithms in Molecular Biology*.

[5] Langdon, W. B., McKay, R. I., and Spector, L. Genetic programming. In *Handbook of Metaheuristics*, J.-Y. Potvin and M. Gendreau, Eds., second ed. Springer, ch. 7.

[6] Langdon, W. B., and Poli, R. *Foundations of Genetic Programming*. Springer-Verlag, 2002.

[7] Langdon, W. B., Rowsell, J., and Harrison, A. P. Creating regular expressions as mRNA motifs with GP to predict human exon splitting. Tech. Rep. TR-09-02, Department of Computer Science, Crest Centre, King's College, London, Strand, London, WC2R 2LS, UK, 19 Mar. 2009.

[8] Langdon, W. B., Upton, G. J. G., da Silva Camargo, R., and Harrison, A. P. A survey of spatial defects in Homo Sapiens Affymetrix GeneChips. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2009). In press.

[9] Poli, R., Langdon, W. B., and McPhee, N. F. *A field guide to genetic programming*. Published via `http://lulu.com` and freely available at `http://www.gp-field-guide.org.uk`, 2008. (With contributions by J. R. Koza).

[10] Retelska, D., *et al.* Similarities and differences of polyadenylation signals in human and fly. *BMC Genomics 7*, 1 (2006), 176.

[11] Sanchez-Graillet, O., Rowsell, J., Langdon, W. B., Stalteri, M. A., Arteaga Salas, J. M., Upton, G. J., and Harrison, A. P. Widespread existence of uncorrelated probe intensities from within the same probeset on Affymetrix GeneChips. *Journal of Integrative Bioinformatics 5*, 2 (2008), 98.