IEEE CEC 2008, Hong Kong, 1-6 June, pages 4152–4157

Evolving GeneChip Correlation Predictors on Parallel Graphics Hardware

W. B. Langdon

***Abstract*—A GPU is used to datamine five million correlations between probes within Affymetrix HG-U133A probesets across 6685 human tissue samples from NCBI's GEO database. These concordances are used as machine learning training data for genetic programming running on a Linux PC with a RapidMind OpenGL GLSL backend. GPGPU is used to identify technological factors influencing High Density Oligonuclotide Arrays (HDONA) performance. GP suggests mismatch (PM/MM) and Adenosine/Guanine ratio influence microarray quality. Initial results hint that Watson-Crick probe self hybridisation or folding is not important. Under GPGPGPU an nVidia GeForce 8800 GTX interprets 300 million GP primitives/second (300 MGPops, approx 8 GFLOPS).**

I. Introduction

Affymetrix GeneChips, such as their HG-U133A, provide multiple measurements per gene transcript. Individual measurements are provide by short (25 base) DNA sequences (known as probes, cf. Figure 2). These sequences of DNA bases are designed to be complementary to known locations in human genes. Being complementary, the gene product (mRNA) preferentially binds to the probe. Probes are tightly placed on a glass slide in a square grid pattern. A fluorescent dye is used to quantify how much mRNA is bound to each probe.

Measurement of ultra low (pico molar) concentrations of long chain molecules, like mRNA, is noisy. Affymetrix provides various control signals, including multiple measurements to reduce noise. One controversial mechanism is adjacent to each measuring probe is a control "mismatch" probe. The MM probe is identical to the "perfect match" PM probe except its central base is anti-complementary. The intention being the MM measurement would give an extremely sensitive background reading for its PM partner. The true signal being given by subtracting the MM from the PM signal. However in many cases the MM signal is actually higher than the PM signal. This has led to mismatch probes being widely distrusted and often ignored.

While nothing is simple in Biology, to a first approximation the amount of mRNA produced by a gene should be the same no matter which part of the mRNA molecule is bound to a probe. Affymetrix groups probes into probesets. Each probeset targets a gene. Excluding controls, the HG-U133A has 22 215 probesets. For simplicity we concentrate upon the 21 765 HG-U133A probesets with exactly 11 pairs of probes. Figure 1 shows for an example probeset its 231 correlations as a "heatmap" (yellow/lighter corresponds to greater consistency between pairs of probes).

Mathematical and Biological Sciences, University of Essex,Colchester CO4 3SQ, UK; email: w1angdon@essex.ac.uk.

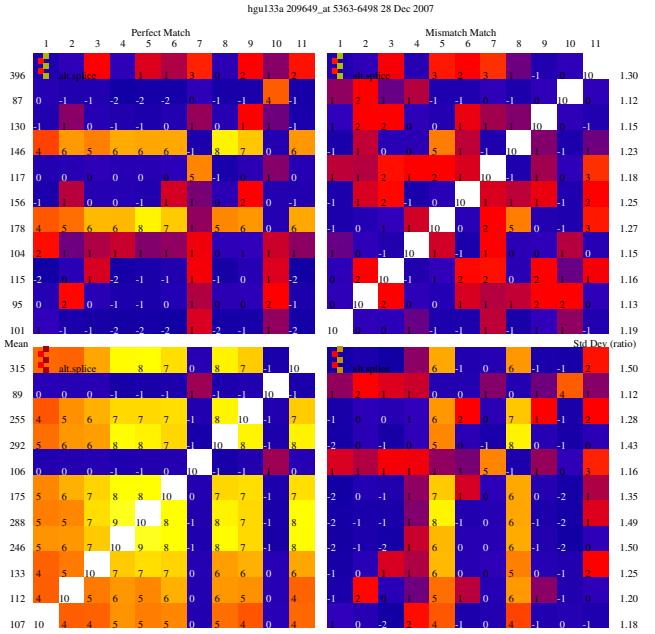

Fig. 1. Correlation coefficients between 22 probes for gene "signal transducing adaptor molecule (SH3 domain and ITAM motif) 2" STAM2. Nine of the perfect match (lower left) are correlated but probes PM_7 and PM_{10} are not. PM_7 and PM_{10} have stable low intensities (106, 89). The mismatch probes are not well correlated, either amongst themselves (top right) or with the PMs (lower right).

There are several known Biological reasons which might lead to probes on the same gene giving consistently unrelated readings. (Alternative splicing, alternative polyadenylation and 3'-5' degradation, come to mind [1]. See also the next section.) However these seem unable explain all the many cases of poor correlation. Can we find technological reasons?

The next section will describe the preparation of datasets containing the correlation coefficients and facts about GeneChip technology. Section III describes the genetic programming system and its operation on a graphics processing unit (GPU) [2], [3], [4], [5]. Our genetic programming (GP) uses facts about the HG-U133A probes to predict which are well correlated with gene activity and which are not. Section IV describes how well the GP does. It uses the evolved population to suggest the relative importance of various components of the GeneChip technology. Then it gives the speed of the GPU. In Section V we conclude that PM/MM and the relative numbers of As and Gs are the most important.

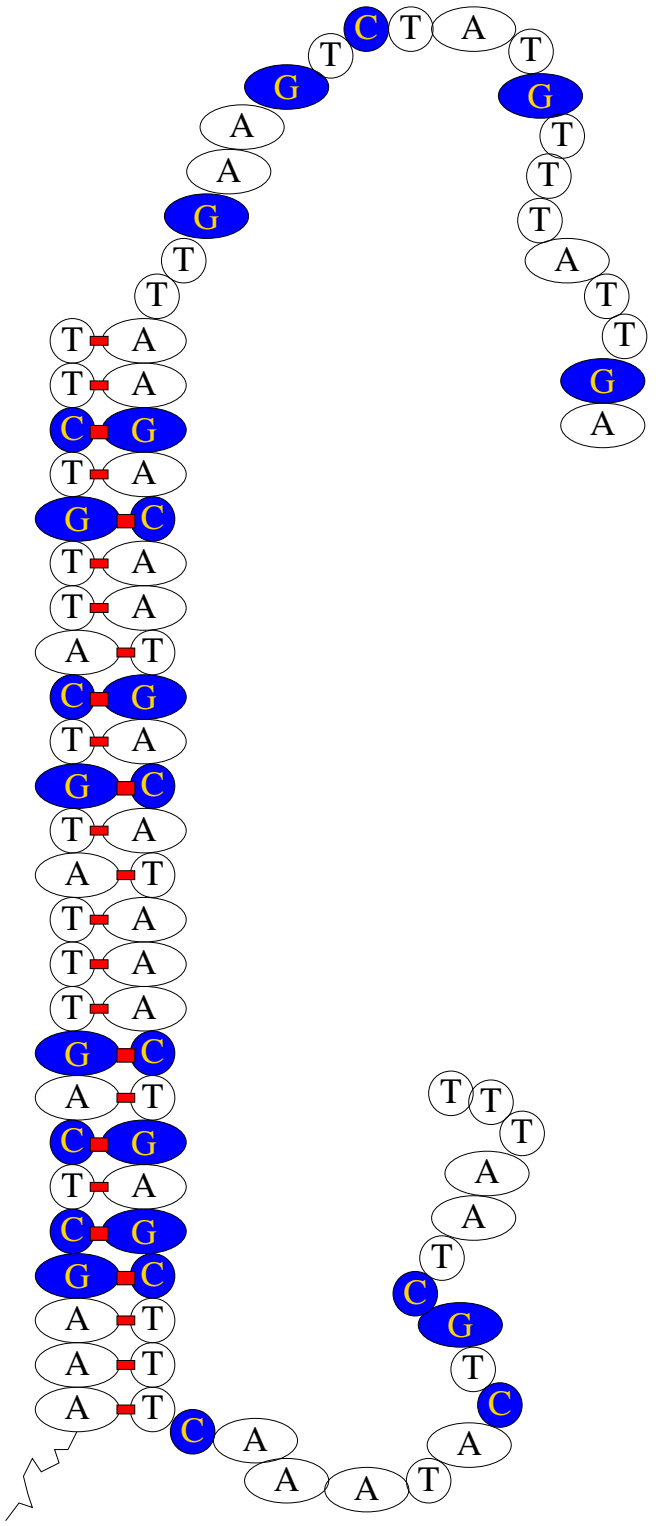

Fig. 2. Schematic of an Affymetrix probe (209649_at PM_5, left) bound with complementary target sequence (right). DNA double helix represented as straight vertical ladder. Note complementary T–A and C–G base bindings are shown by red rectangles. The 25 bases of the probe are tethered to the glass slide by a flexible linker (black lower left). Firmly bound target sequences can be detected by treatment with a florescent dye, whose location is detected with a laser and an optical microscope. The florescent intensity is approximately proportional to the amount of bound target and so gives some indication of target gene activity.

II. Inputs to Explain Poor Correlation

In 2007, as part of major bioinformatics datamining exercise, all the human Affymetrix GeneChips in the USA National Center for Biotechnology Information's GEO [7] had been down loaded [6]. In particular 6685 HG-U133A, each containing half a million data points from a wide range of Human tissues and disease states were available.

Since it is impossible to accurately control the total amount of mRNA dropped on each GeneChip, it is necessary to normalise the data. We used quantile normalisation, which non-linearly rescales the data so that the mean and standard deviation (and all higher moments) are the same. We normalised each chip against a reference average chip [6].

The correlation coefficient of each pair of probes within each probeset in the HG-U133A design was calculated. (A total of 5 310 652 correlation coefficients.) To avoid undue influence of outliers: if either probe was more than three standard deviations from its average value that tissue sample was exclude from the calculation of that probe pairs correlation. Also, each GeneChip was checked for spatial defects [6] and the sample exclude if either probe was within 4 probes of a known spatial error. Even after excluding these data, all correlations used probe pairs from several thousand samples.

The placement of probes on GeneChips, like the HG-U133A, is designed to put similar probes next to each other. This is done to ease the photolithographic chemical process used in manufacture and leads to horizontal bands in average intensity. Probe variability tends to increase with average intensity (until the signal is so strong that the probe saturates, which reduces its ability to vary). Since signal variation (as opposed to noise) is required for correlation, correlation also tends to increase with average intensity leading to horizontal bands in correlation across the GeneChip. Figure 3 shows the expected bands in average correlation for each probe.

The distribution of probe-probe correlation coefficients for probes in the same probeset has two peaks: one near 0.8 and the other near 0, cf. Figure 4. As the peak at zero emphasises, there are many probes which are simply not well correlated with the other members of their probeset. Explaining why is a major research effort. The peak at 0.8 corresponds better behaved probes.

Part of the sample preparation process requires an enzyme to act upon purified mRNA starting at the 3' end. I.e. at the normal back end of mRNA. The enzyme works its way forward towards the 5' front of the mRNA. Unfortunately it has a tendency to fall off. Therefore the strength of signals tends to fall further away from the 3' end. This is a known problem and laboratories regularly check that the effect is not too extreme in each sample. Nonetheless, there is a small trend for correlation between probes in the same probeset to fall as the distance between where they measure along the mRNA gene transcript increases. (The 5' end is the negative direction in Figure 4).

To exclude genes which are either never expressed (or are constantly expressed) we selected 13 863 probesets where

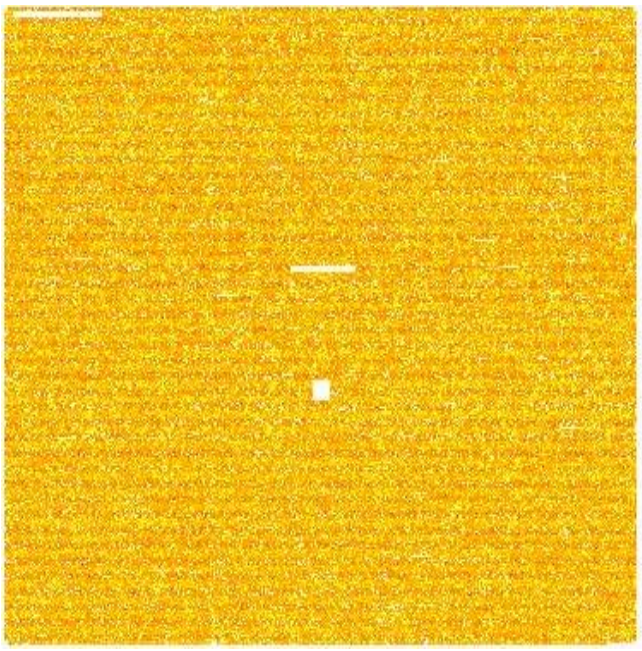

Fig. 3. Mean of 5 310 652 correlations between probes in probesets across 6685 HG-U133A GeneChips. White regions contain no probesets. Smallest -0.63 (red). Median 0.16. Max 0.97 (yellow). 0.7% 3341 probes have a mean correlation greater than 0.8 with the other probes in their probeset.

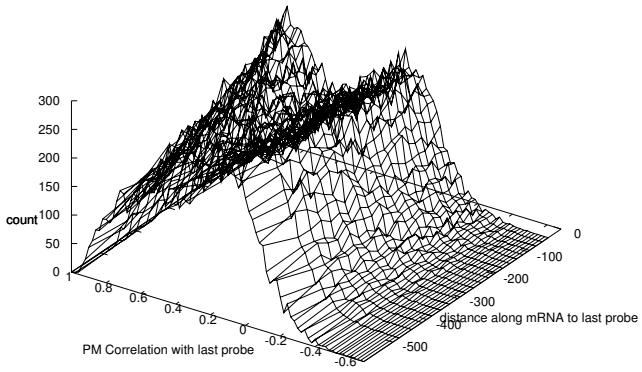

Fig. 4. Correlation between PM probe and last PM probe in each GEO HG-U133A probesets. (Approx 10000 data points per distance bin. Controls excluded.)

three or more probe pairs had correlations of 0.8 or more. These were evenly split into three to provide independent training, test and validation data.

To give each probe the best chance we looked at all 21 of its pairings with other members of its probeset and took the one for which it was most correlated. The technological data for the probe is summarised in Table I.

A. GP Training Set

As Figure 4 has shown correlation coefficients cover a wide range with many taking intermediate values. Since we are using correlation only as an indication of how well a probe is working we decided to exclude the middle values from training and instead use probe pairs that were highly correlated or were very poorly correlated.

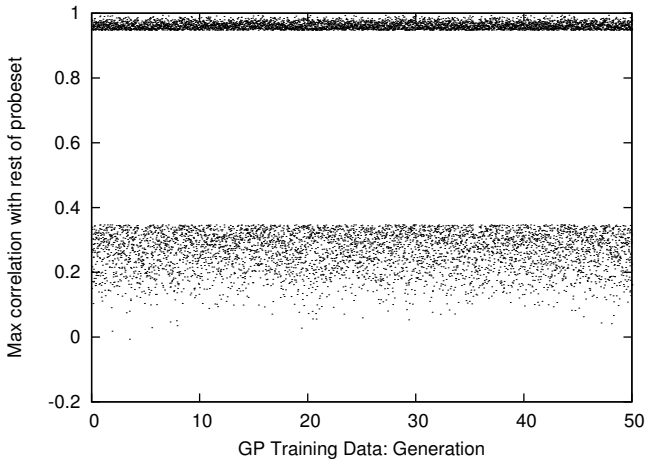

Fig. 5. GP training data changed each generation. GP trained on 100 probes well correlated with rest of their probeset (top) and 100 poorly correlated (bottom).

Of the 101 662 available training examples, the 5200 most correlated and the 5200 least correlated we chosen. Each high correlation example was paired with the corresponding low correlation example and then the pairs were put into a random order.

Each generation the GP uses two hundred randomly chosen but different probes for training. This ensure there are 100 high corelation probe and 100 low corelation probes. Cf. Figure 5.

B. Probe Folding

Affymetrix discounts the idea that poor probe performance arrises from its single stranded DNA probes binding to themselves to form a stable DNA double helix and so not being available to bind to mRNA. Nevertheless the suspicion remains that this is a possible explanation for poor probe performance. We provided the GP with the results of two of the many possible simple probe bindings (see Figures 6 and 7). For each the GP is given the fraction of the top of the probe left exposed (i.e. not part of a DNA spiral). Secondly we crudely model the thermodynamic strength of the binding by counting the number of complementary base pairing the spiral contains. Since the probes are only 25 bases long the optimal binding is readily estimated by exhaustive search.

III. Evolving Correlation Prediction

The genetic programming system is a traditional tree GP system with subtree crossover and a range of mutation operators [9], [10], [3] (cf. Table I). Whilst [11] demonstrates (albeit for evolutionary programming rather than for GP) that a GPU can implement mutation and selection, these are done by the host CPU. This means each generation the whole population and the training data are transfered to the GPU. However the run time is dominated by the time taken to interpret each GP individual, rather than the genetic operations (see Section IV-C). Therefore we anticipate only a modest improvement might be possible by implementing mutation etc. on the GPU.

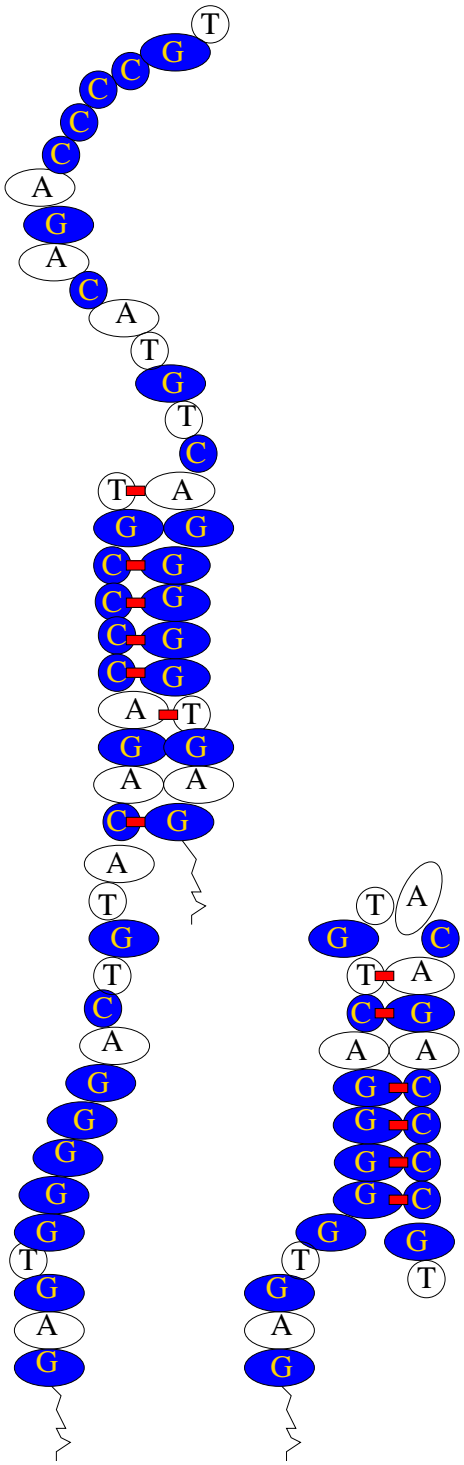

Fig. 6. Possible pairing of probe with neighbour (which will have same sequence). Watson-Crick binding occurs between ovals and circles of the same colour. Ovals represent the larger bases (A, G) and circles T and C. Blue indicates stronger binding. 25 base DNA probes 20.38 10^{-9}m (unwound) are tethered to glass slide by flexible polymers. The average distance (along glass) between probes is 2.5 10^{-9}m [8, e70]. I.e. the probes are close enough to interact with each other.

Fig. 7. (Right) DNA sequences may be self complementary forming hairpins loops. Red rectangle indicates binding between complementary bases.

TABLE I
GP PARAMETERS FOR GENECHIP CORRELATION PREDICTION

Function set:	ADD SUB MUL DIV MIN MAX operating on floats
Terminal set:	
	$\text{probe\_index}_{\text{self}}$, MM_{self} (0/1), $\text{probe\_index}_{\text{other}}$, MM_{other} (0/1), Position of two probes on HG-U133A GeneChip (X_{self}, Y_{self}, X_{other}, Y_{other}).
	Distance along mRNA transcript, as defined by Affymetrix, from last probe in probe set and distance from other probe.
	The same two distances expressed as a fraction of the length of mRNA spanned by the probeset (Loc_{self}(ratio) and i-o(ratio)).
	Number of Adenosine (A), Thymine (T), Guanine (G) and Cytosine (C) bases in the probe both (as integers and as fractions of 25).
	The twenty five bases in the probe (coded as A,T,G,C = $1/\pi$, $-1/\pi$, $e^{-3/4}$, $-e^{-3/4}$).
	Fraction of probe exposed assuming it was bound to neighbour and number of complementary pairs in the binding, cf. Figure 6.
	The same ratio and count assuming probe binds to itself via a single hairpin, cf. Figure 7.
	1001 Constants -5, -4.99, -4.98,... 4.98 ,4.99, 5
Fitness:	$\sum^{200}$ |best correlation−prediction| To avoid problems with calculations (e.g. divide by zero) producing infinity, the absolute prediction error calculated for each of the 200 fitness cases was limited to at most 10^{10}.
Selection:	tournament size 4 in overlapping fine grained 21×21 demes [10], non elitist, Population size 128×128 = 16384
Initial pop:	ramped half-and-half 1:3 (50% of terminals are constants)
Parameters:	50% subtree crossover. 50% mutation (point 22.5%, constants 22.5%, subtree 5%). Max tree size 63, Max tree depth 8.
Termination:	50 generations

On the GPU each of the 16 384 GP individuals is interpreted on 200 training examples. (As mentioned in Section II-A, every generation two hundred new examples are used.)

Since the GPU provides SIMD parallel operation [12] the GPU interpreter is stack based and uses reverse polish notation (RPN/postfix) rather than the usual Lisp prefix tree structure. To avoid data conversion between the CPU and GPU the GP genetic operations have been modified to use the linearised RPN representation. Linearised RPN gives a compact and very fast implementation. Details of the GPGPU implementation are given in [3], whilst [4] provides and example of its use in Bioinformatics. C++ code is available via FTP `ftp://cs.ucl.ac.uk/genetic/gp-code/gpu_gp_1.tar.gz`

IV. RESULTS

In the first run GP evolved a predictor (see Figure 8) which on the last generation's training data is on average 0.16 from the actual correlation. To convert the evolved continuous regression problem into a binary classify we use our previous threshold of 0.8 (cf. Section II) to divide good from poor probes. Table II contains a confusion matrix which compares the actual maximum correlation of the probes with the other members of their probeset with the evolved prediction on the whole of the training set (including the 91 462 middling values which GP never saw). Unlike in many machine learning applications, there is no evidence of over fitting. Indeed the corresponding results for the test set (right of Table II) are not significantly different (χ^2, 3 dof).

TABLE II
PERFORMANCE OF EVOLVED PREDICTOR

	Whole training set		Test set	
Prediction:	poor	good	poor	good
poor ($<$0.8)	32 009	15 082	32 112	15 097
good ($\geq$0.8)	23 551	31 020	23 463	30 990

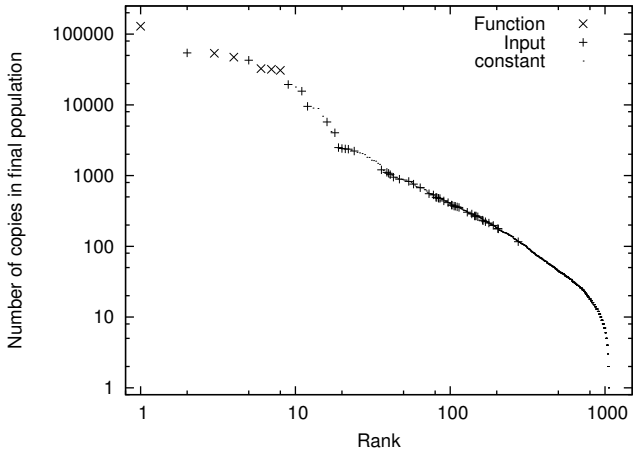

Fig. 9. End of run abundance of GP primitives in the evolved population

A. Evolved Predictor

The predictor found by genetic programming is given in Figure 8. Essentially it consists of four sub formulae and returns the maximum of them. $\mathrm{MM}_{\mathrm{self}}$ plays a dominate role. For perfect match probes $\mathrm{MM}_{\mathrm{self}} = 0$ and the predictor returns 0.97. For mismatch probes $\mathrm{MM}_{\mathrm{self}} = 1$ and usually GP predicts a correlation below 0.8 (i.e. a poor probe) unless the probe contains more Guanine than Adenosine bases. Typically if there are more than twice as many Guanine as Adenosine then GP predicts the mismatch probe will have a high correlations (with at least one other probe in its probeset). The other technological inputs have little impact of the prediction.

B. Relative Importance of parts of GeneChip Technology

As expected, over much of the range of values the frequence of GP primitives evolves to follow a Zipf like law [13]. This give rise to an almost straight line with a gradient near -1, when frequency is plotted against rank on log-log scales, cf. Figure 9. Of particular interest are those inputs which occur frequently, since this suggests that they can help predict if a probe works well or not.

Table III shows important factors include: 1) whether the probe is a perfect match or mismatch $\mathrm{MM}_{\mathrm{self}}$, 2) the number of A T G and C's in the probe. These reinforce the message drawn from the best individual in the final population in the previous section. That is, the most important factor in differentiating a working probe from one with low corelation with the other members of its probeset is whether or not it is a perfect match or mismatch probe. Second is the fraction of of the four bases. As Figure 9 shows there is a gap between these and the other inputs (highlighted by horizontal line in Table III.) This suggests perhaps the other inputs available to GP are of little importance.

TABLE III
POPULAR HG-U133A PROBE CORRELATION PREDICTION INPUTS

Rank	Name	Count
2	$\mathrm{MM}_{\mathrm{self}}$	54147
5	C(frac)	42710
9	G(frac)	19393
11	A	15601
12	G	9533
16	A(frac)	5725
18	T(frac)	4038
19	Seq22	2488
20	i-o(ratio)	2419
21	Seq19	2383
22	Seq18	2358
24	Seq16	2220

Plotted as $+$ in Figure 9.

Of the 8 locations inputs (be it X,Y, sequence in probeset, or location along mRNA transcript) only the relative distance between the two probes along the transcript ($\mathrm{i\text{-}o}_{\mathrm{ratio}}$) appears in the top 25. The hairpin and neighbour probe binding inputs calculated from the probe's DNA sequence (see Section II-B) appear well down the list (40 onwards, after many constants). The middle base (Seq13, which is the only difference between PM and MM probes) is even further down the list at rank 73. This is surprisingly low, since Seq13 is known to be important in the comparison of PM vs. MM probes [14].

C. RapidMind C++ Performance

On average fitness evaluation took the GPU 13.58 Sec. (Total run time 15.94 Sec.) The average program size was 25.56. Since the 51 populations each contained 16 384 programs, on average the GPU interpreted 314 million GP primitives per second.

Without detailed examination of the RapidMind GPU compiler it is difficult to estimate how many floating point operations are required to interpret each GP primitive. Since we are using defaults for all the RapidMind parameters, the GPU compiler optimises. Assuming the compiler removes common expressions, we estimate approximately 24 FLOPs are needed for each GP function or leaf. This suggests the GPU is delivering very roughly in the region of 8 GFLOP.

V. CONCLUSIONS

Using the affy bioconductor R statistical package we can calculate correlations across thousands of publicly available GeneChips. Even after excluding outliers and spatial flaws in the data, the five million correlations between probes in the same probeset, which should be measuring the same gene, show wide variation. Genetic programming running on a state of the art graphic processing unit automatically evolved a biologically feasible predictor of probe quality. Analysis of the GP's population lends support for Affymetrix' claim that poor probe performance is not due to probe's simple Watson-Crick self-hybridising. Other forms of probe-probe, probe-target [15] or target-target might be considered in future.

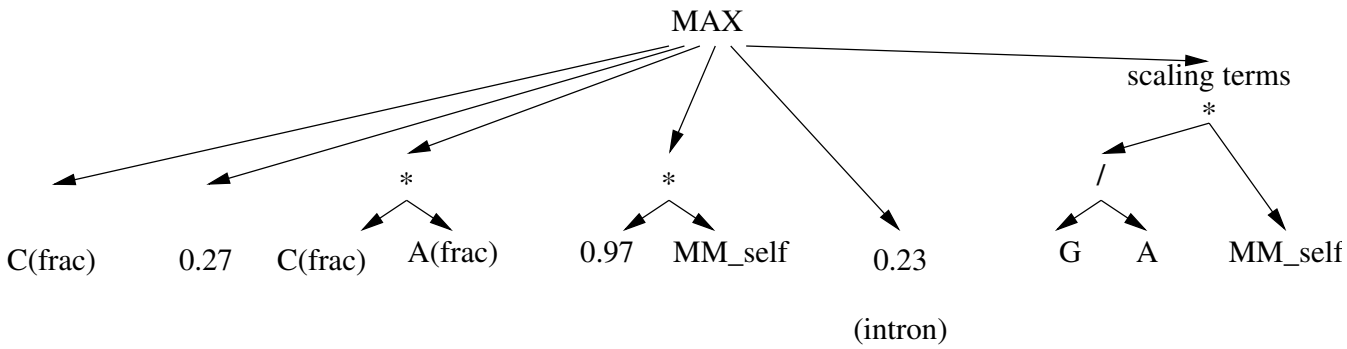

Fig. 8. Simplification of evolved HG-U133A probe correlation predictor. The evolved program contains 27 GP primitives. A subtree of 11 primitives always returns 0.23 (intron) A further branch of 8 primitives is relatively insensitive and mostly has the effect of scaling the ratio of Guanine/Adenosine to be similar to the range of other numbers in the formula.

ACKNOWLEDGEMENT

I would like to thank Tim Czyrnyj.

REFERENCES

[1] W. B. Langdon, R. da Silva Camargo, and A. P. Harrison, "Spatial defects in 5896 HG-U133A genechips," in *Critical Assesment of Microarray Data*, J. Dopazo, Ed., Valencia, 13-14 December 2007.

[2] S. Harding and W. Banzhaf, "Fast Genetic Programming on GPUs," in *Proceedings of the 10th European Conference on Genetic Programming*, ser. Lecture Notes in Computer Science, M. Ebner, M. O'Neill, A. Ekárt, L. Vanneschi, and A. I. Esparcia-Alcázar, Eds., vol. 4445. Valencia, Spain: Springer, 11 - 13 Apr. 2007, pp. 90–101.

[3] W. B. Langdon and W. Banzhaf, "A SIMD interpreter for genetic programming on GPU graphics cards," in *EuroGP*, ser. LNCS, vol. 4971. Naples: Springer, 26-28 Mar. 2008, pp. 73–85.

[4] W. B. Langdon and A. P. Harrison, "GP on SPMD parallel graphics hardware for mega bioinformatics data mining," *Soft Computing*, special Issue.

[5] J. D. Owens, David, N. Govindaraju, M. Harris, J. Kruger, A. E. Lefohn, and T. J. Purcell, "A survey of general-purpose computation on graphics hardware," *Computer Graphics Forum*, vol. 26, no. 1, pp. 80–113, March 2007.

[6] W. B. Langdon, G. J. G. Upton, R. da Silva Camargo, and A. P. Harrison, "A survey of spatial defects in Homo Sapiens Affymetrix genechips," 2007, submitted.

[7] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar, "NCBI GEO: mining tens of millions of expression profiles–database and tools update," *Nucleic Acids Research*, vol. 35, no. Database issue, January 2007.

[8] G. A. Held, G. Grinstein, and Y. Tu, "Relationship between gene expression and observed intensities in DNA microarrays–a modeling study," *Nucleic Acids Research*, vol. 34, no. 9, p. e70, 2006.

[9] R. Poli, W. B. Langdon, and N. F. McPhee, *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] W. B. Langdon, *Genetic Programming and Data Structures*. 1998.

[11] K.-L. Fok, T.-T. Wong, and M.-L. Wong, "Evolutionary computing on consumer graphics hardware," *IEEE Intelligent Systems*, vol. 22, no. 2, pp. 69–78, Mar.-Apr. 2007.

[12] H. Juille and J. B. Pollack, "Massively parallel genetic programming," in *Advances in Genetic Programming 2*, P. J. Angeline and K. E. Kinnear, Jr., Eds. Cambridge, MA, USA: MIT Press, 1996, ch. 17, pp. 339–358.

[13] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge 42, MA, USA: Addison-Wesley Press Inc., 1949.

[14] F. Naef and M. O. Magnasco, "Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 68, no. 1, p. 011906, 2003.

[15] C. Wu, H. Zhao, K. Baggerly, R. Carta, and L. Zhang, "Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays," *Bioinformatics*, vol. 23, no. 19, pp. 2566–2572, 2007.