Machine learned regression for abductive DNA sequencing

David Thornley*, *member IEEE*, Maxim Zverev and Stavros Petridis, *member IEEE*
Department of Computing, Imperial College London
180 Queen's Gate, South Kensington, London SW7 2RH

Abstract

We construct machine learned regressors to predict the behaviour of DNA sequencing data from the fluorescent labelled Sanger method. These predictions are used to assess hypotheses for sequence composition through calculation of likelihood or deviation evidence from the comparison of predictions from the hypothesized sequence with target trace data. We machine learn a means for comparing the measures taken from competing hypotheses for the sequence. This is a machine learned implementation of our proposal for abductive DNA basecalling. The results of the present experiments suggest that neural nets are a more effective means for predicting peak sizes than decision tree regressors, and for assembling evidence for competing hypotheses in this context. This is despite the availability of variance estimates in our decision tree regressors.

1 Introduction

In his thesis of 1993 [5], Blanchard examined sequence dependent variations in Sanger sequencing [13] trace data [7]. He expressed the opinion that this knowledge would not be useful in basecalling. Indeed, no basecalling package in current use leverages this peak height knowledge. The leading third party basecaller, PHRED uses peak spacing to excellent effect [3] with quantifiable error rates [4], although it tracks overall trends rather than using detailed knowledge [6]. In unrelated work, Thornley counsels against dismissing interest in peak size variation, and explains that the accompanying attempt to reduce peak height variation is throwing away important information [14]. He also provides a simple functional model for peak size variation, and formulates a method of analysis which actively takes advantage of peak size variation. That approach comprises abduction of basecalls in which we hypothesize a sequence composition, and assess the peak sizes predicted from each hypothesis against the target trace data to find that which fits best [1].

*This work is a deliverable of EPSRC grant GR/S60266/01

The primitive model used in that initial work enabled validation of our suggestion that there is information encoded in the peak size behaviour in the context of a base position. We used a contrived approach which we refer to as blind spot analysis or BSA, in which we isolate the contextual information by entirely omitting the data at the basecalling position, or *pivot* as we will refer to this position hereafter. This means that the information normally used for basecalling is omitted. Thus we use only the contextual information which we have proposed is encoded in the repeatable, sequence motif correlated behaviour of DNA sequencing trace data from the Sanger method.

Our exploration of this information using machine learning tools has demonstrated that viable basecalls can be made using contextual information alone by direct classification [17, 18]. In this work we found that the dependencies examined by Lipshutz [9] during work to estimate confidence in existing basecalls relate to the information we seek to exploit.

When a classifier is given access to the data at the basecalling position – which we refer to as the *pivot* – it only uses that information in its decision, effectively ignoring the contextual data provided. This is because high quality data can generally be called using the pivotal data alone. Indeed, this pivotal data provides the only peak heights used in current basecalling methods. To enable comparison of classifier effectiveness in using context information, we excluded the pivot data to perform "blind spot analysis" in the sense introduced in [14].

We now seek to build a machine learned approach to the basecall abduction proposed in [14]. In this new work we move on from proof of principle toward establishing components for the abductive process as originally intended. The goal of the present work is to find an effective means for regressing peak sizes, and to explore comparison methods. Since we intend to use the resulting regressor in a general basecaller [1], it must use all the information available. We have found that if we supply the regression information at the pivotal position to the hypothesis comparison step, regardless of which regressor or comparison method is used, the success rate is approximately 100%. This is because

in high quality data, the pivot data is sufficient for inference of the base call. We therefore exclude the regression result at the pivot from the information made available to the comparator. We refer to this as partial blind-spot analysis (PBSA). The data at the pivot is used in the formulation of the regression, but prediction and comparison of data to form a likelihood measure is not carried out for the pivot position.

The contextual information has previously enabled us to call a base by direct classification in high quality data with a success rate of approximately 80%. This contrasts with a lower band expected through random guessing. This is not 25%, because the base composition is not a uniformly distbuted process. A classifier trained to identify a base through using only its contextual sequence (without the trace peak data) achieves a success rate of approximately 34% [17]. The trace data used in the present and previous work comes from genomic work at the Wellcome Trust Sanger Institute to sequence the human X chromosome.

As in our direct classification experiments, we partition our trace data corpus into a number of folds. In these experiments, each fold comprises all local behaviour instances taken from the high quality region of 100 randomly selected trace data files. Data from a given file appears in only one fold. A local behaviour instance comprises the basecalls and trace data covering 5 base positions. To ensure consistency between experiments, and to test generality, we use a single training fold, a single regressor validation fold, a single comparator training fold, a single comparator validation fold, and 10 test folds. Each fold comprises 100 randomly selected trace files.

In this paper we describe the formulation of a regressor, which predicts a single peak height based on its context in terms of basecalls, and a comparator which integrates the information from comparisons of behaviour between prediction and target data at a number of contextual positions. We carry out two main sequences of experiments: one with decision tree regression, and another using neural network regression. In each case we experiment with simple functional comparison of evidence, and with a neural net to perform that comparison.

2 Abductive basecalling

Abduction is a method of reasoning in which the cause of an observed phenomenon is sought by evoking suitable hypotheses and comparing the predictions that can be made from each with the observation. In our case, we hypothesize the base composition of a sample, and predict trace data which would arise from each hypothesis.

Our original conceit in [14] was that we should be able to call bases without reference to the data at the calling position, because each possible different base will have a different – and hence, we hope – discriminable effect on the peak size behaviour in its context. Peak height patterns are strongly repeatable [5] , and correlate with sequence motif to exhibit distinct expectations and variance [10, 11]. In [14] we suggest that an incorrect hypothesis of base composition will predict patterns which differ from the trace data we see, and hence allow us to reject such hypotheses.

In other work we are attempting to construct a phenomenological model which will capture all scales of behaviour in trace data [15, 16]. While machine learning relationships in the data, we restrict ourselves to data configurations for which we can construct a simple learning strategy. This essentialy means capturing local influences over a small number of base positions. For the present work, we use data from 5 consecutive base positions in every case, since this has proven an effective configuration for machine learning with this data.

Skylining [2] is used to detrend the data to make it possible to compare characteristics of the signal among any points in the trace. We find that a quadratic polynomial is an effective and intuitively reasonable function for modelling the skyline [17]. We believe this to be because there is a simple decay characteristic in the data due to the activity of the Sanger reaction, and a simple rising characteristic early in the data due to the efficiency of physical transfer of the different lengths of DNA fragments between the reaction vessel and the separation medium.

3 Machine learning for abduction

We have conducted ANOVA experiments on recent trace data to see if Lipshutz results [9] still pertain. Our results were essentially identical, showing that peak heights trace are affected by surrounding sequence, with the peak height in a given position influenced most strongly by the three bases to the left and one to the right. We performed a range of experiments [17] to determine the most effective context for use in direct classification of bases. This emerged as basecalls and peak data from three positions to the right, and one position to the left. This is then clearly because those are the positions most directly affected by the choice of basecall, which was the output class.

In the present work, which performs abduction rather than direct classification, there are two forms of local influence to be taken into account when designing the input data for a machine learned entity. In regressing (predicting) a peak, we include basecalls and data from three positions to the left, and one position to the right of the base position for which we wish to predict a peak height because those are the positions which most directly influence that peak. These peak heights are used in the calculation of a likelihood measure through comparison with the target data. The ultimate goal is to classify a base, and a base affects be-

haviour at one position to the left and three to the right. Therefore, we consider three footprint shapes. The regressor footprint is (3,1),*i.e.* three bases to the left and one to the right of the pivot. The footprint of the classifier is (1,3) in the likelihoods, and therefore (4,4) in the trace data itself. That means that data from 9 base positions contribute to each basecall in the form of evidence from 4 in these experiments, and 5 in a general basecaller when we correctly incorporate the pivot evidence.

Considering the direct classifier training problem briefly, we require a number of example instances of each local behaviour to train a tree or net. In our previous work, we have found that a (3,1) footprint is most effective (a combination of accuracy and cost of training, with a strongly diminishing return at larger footprints). There are $4^5 = 1024$ (3,1) footprints, taking into account the base at the pivot, which must be covered in the training. If we were to attempt direct classification with a (4,4) footprint, we would have to consider $4^9 = 262144$ distinct base sequences, each with a requirement to cover several examples of peak size behaviour not constrained by sequence. This would lead to an infeasible training task.

All neural networks in the present work are composed of threshold sigmoid units. We generally use 2 layers of 30 nodes for the regression step, since these can approximate any non-linear mapping from inputs to outputs [8]). For the comparison stage in the final experiments however, we find a deeper net more effective, using three layers of 20 nodes. All nets are trained using the resilent propagation algorithm [12].

3.1 Overfitting

Overfitting occurs naturally when applying machine learning techniques to real data which includes some form of noise, random errors or inconsistencies. In this particular application we have prior knowledge (or at least strong suspicion) that the data includes modes of behaviour which cannot be entirely explained by local base sequence [15, 16]. However, we believe these behaviours are realtively simple dynamic oscillatory effects which could potentially be constrained by examining peak sizes. Therefore we provide the contextual peak sizes to the regression component as they contain both direct information about sequence composition, and may express aspects of any longer range influences in the behaviour. Our hope is that a machine learned entity can capture this in a helpful manner.

To train such a machine learned entity, we require a corpus of examples for it to summarize as a function. When we assess the error of such a neural net after training against the training corpus, there will generally be a significant residual error. Conversely, an unpruned decision tree may exactly regress its training set, because each datum follows through to a leaf.

We have ensured generality of the machine learned regressors in two ways. For the decision tree regressor, we pruned the tree to a level which minimizes the mean square error in peak size prediction against our validation data corpus. For our neural net regressor, we instead optimized its performance explicitly in the target application of abductive basecalling. We did this by testing the net at a number of training epochs, selecting the net at the number of epochs whose resulting parameterization achieves the minimum classification error of partial blind basecalls in the validation corpus through selection of minum z-score sum. Of course, this does not necessarily produce a net which most optimally predicts the peak sizes in the training corpus: instead it optimizes its output to be a value which is most effective in the specific application of selecting a hypothesis by minimum sum of absolute differences. We find 500 epochs approximately optimal. Our hypothesis comparator net's performance on a validation set increases monotonically with epochs and network size in the experimentation we have performed so far.

4 Regression and comparison experiments

Our approach to abduction of basecalls here is to posit a hypothesis which comprises a suggestion for an unknown base in the context of known surrounding sequence. For each hypothesis (there are always four in these experiments - one for each of A, C, G and T) we form predictions of the data expected at the contextual base positions if that hypothesis were true. Evidence for or against that hypothesis is then calculated for each target data peak in comparison with its corresponding prediction from the hypothesis. These evidence measures – four per hypothesis – are then weighed against each other to select the hypothesis which is most likely correct. We use very simple evidence calculations, intending that the machine learned components evoke appropriate functions.

The comparator footprint is (1,3) in the likelihoods, so for each hypothesis (differing by the base at the pivot), we use the regressor to predict the peak height at each position in that (1,3) footprint but not at the pivot. We then calculate the measure to be used in the comparison step for each prediction.

In experiment 1, this is the likelihood value calculated as the value of the normal distribution specified by the mean and variance given by the regression tree, read at the size of the observed peak. We now have four sets of four likelihoods. The likelihoods for a given hypothesis are summed to give a total for that hypothesis. The hypothesis with the largest total likelihood is taken to be the correct basecall.

In all the experiments we describe here, predictions are made for the data in a window of five base positions indexed

1 through 5. The basecall we wish to make is at position 2 in that window. Each prediction is provided by a regressor which responds to a window of data. The regression window is also 5 base positions, but the configuration is quite different. We regress the peak size of position 4 in the window of five bases - this is because our analyses in [17] demonstrate that this configuration is effective and efficient. A larger window enables marginally improved success rates in classification, but at the expense of longer training times. We therefore present this exploratory work with a five base window.

Likelihoods for four base positions (1, 3, 4 and 5 in the five base window) for each alternative hypothesis of the pivot base (A, C, G or T) are tabulated for the inference process. These likelihoods are calculated as a function of the measured peak and the prediction information from the regressor. A decision tree regressor provides a mean and variance, so we can derive a probability density from a normal distribution with this mean and variance (*probability evidence*), or either the absolute difference between mean and measured peak (*difference evidence*), or that difference divided by the variance (*deviation evidence*). A neural net regressor provides a single value prediction, so we use difference evidence.

We propose a fully machine-learned approach to the abduction process. This comprises three stages:

1) predict peak sizes for each hypothesis

2) calculate likelihood measures for each data peak

3) compare the information from each hypothesis and output the best

Our first experiments uses a regression tree for step 1, a range of simple measures for step 2 (we will investigate learning a function for this in further work), and a simple combination function for selection in step 3. When we use a neural net, we briefly assess its use in a similar sense to the regression trees, then learn an optimal comparison function.

If we had perfect estimates of the distributions of peak heights expected for a context fully constraining systematic effects, then summation of likelihoods derived from that distribution would be effective. Our ongoing machine learning work both informs our pursuit of experiments to ascertain the true behaviour of the system (*e.g.* [15, 16]), but also provides functional units for the basecalling process itself which is the motivation for our efforts [1]. We are therefore pursuing a holistic approach to the research, seeking the true behaviours from the perspectives of proposal of models and testing against data, and exploratory application of machine learning tools to the data.

5 Decision tree regression

In our first experiment, we use a regression trees for predicting peak heights. An advantage of using a regression tree is that we can estimate statistics of the fit along with predictions. The variance is measured for each leaf of the regression tree during the training and gives us some information about the distribution of peak sizes around the prediction.

We find that number of nodes in an unpruned tree trained on 100 SCF files (a standard trace file format) is often as large as 14000. To circumvent memory limitations of Matlab's in-built tree pruning approach, we implemented a simplified pruning regime pruning to a constant level across the tree using a mean square error performance measure against the validation set. The level with minimum mean error is used. The classification results are laid out in table 1 and discussed below.

test data set	prod.	sum	dev.	diff.
1	55.63	56.92	64.37	65.45
2	54.39	56.50	64.00	65.16
3	55.94	57.55	64.89	65.95
4	54.74	56.25	64.02	65.46
5	55.35	57.28	64.71	65.02
6	55.42	56.73	64.14	65.00
mean	55.25	56.87	64.36	65.34
std. dev.	0.576	0.485	0.373	0.360

Table 1. Classification rates of decision tree regressor on test sets

Using probability evidence and selecting the hypothesis with the maximum product of evidence across its contributing context, the correct classification rate is **55.25%** with a standard deviation of 0.576%. Selecting the hypothesis with the highest *sum* of probability evidence leads to a *higher* rate of **56.87%**, standard deviation 0.485%. This suggests that in the stochastic view of the system, the behaviours are not independent. This seems reasonable, since the behaviours are in close physical proximity on the DNA.

Using deviation evidence and selecting the hypothesis with the lowest total gives a rate of **64.36%** and standard deviation 0.373%. Using *difference* evidence in the same manner gives a *higher rate* of **65.34%**, standard deviation 0.360%. This suggests that either our estimate of variance is poor, or the assumption of a normal distribution is inappropriate. Indeed, we believe that the behaviours we need to model are largely deterministic, and the distributions arise from dimensions in the system which are not fully constrained by data from a small window.

To explore the validity of the assumption of equal weights applied to information at each peak position, we trained a neural network to learn the most appropriate functional mapping from the likelihoods to the most likely base at the pivot in stage 3 of the abduction.

We rejected the notion of judging a hypothesis' likelihood individually - as is essentially the case with summation of likelihoods or deviations - since this removes the ability of the neural net to formulate a response outside our presumptions about the necessary inference mechanism.

The input vector to the network (2 layers of 30 nodes) comprises all the measures produced by the regressor in step 1 for each data peak for each hypothesis. Target vectors consisted of the unary representation of the correct base (*i.e.* [1 0 0 0] for base A, [0 1 0 0] for C).

The combination of a tree regressor for difference evidence with a neural net comparator led to a significantly higher success rate than simple evidence use. Perhaps surprisingly, this rate was comparable with the classification rate achieved using the neural net regressors with simple difference evidence summation as described below. Informally, this may suggest that the comparator neural net is remodelling its inputs against some notion of an optimal use of decision tree predictions.

test data set	sum	net
1	79.32	84.85
2	78.55	85.96
3	79.12	85.58
4	78.71	85.40
5	79.42	85.79
6	79.51	86.64
7	78.57	84.94
8	79.54	85.63
9	79.47	85.67
10	79.55	85.20
mean	79.18	85.57
std. dev.	0.410	0.521

Table 2. Neural net regressor through summation or net comparison

6 Artificial neural network regression

We explore the use of artificial neural nets as a regressor in step 1 to predict the peak heights. The inputs are as for the regression tree, *i.e.* peak heights of 3 bases to the left and 1 base to the right and 5 bases. Recall that the base at the pivot provided to the regressor is part of each hypothesis, not the known correct base.

Choice of the topology and size of a neural network is always a non-trivial problem, and it does not have any elegant solution. In this work we have used a two-layer feed-forward neural network, since a network with 2 layers can approximate any non-linear mapping from inputs to outputs [8]). We use 30 nodes in the hidden layer. The classification results are laid out in table 2 and discussed below.

We find that 500 epochs produce the best results on the regressor validation data set. Using the Artificial Neural Network as a regressor and summing z-scores, the PBSA success rate is **79.18%** with a standard deviation of 0.410%. This is very similar to the success rate of bagged neural net classifiers in our earlier work[17].

After experimentation with the best topology for the comparison neural network by assessing performance of a candidate network on the validation set, we have settled for a three layer network (*i.e.* network with 2 hidden layers) with 20 nodes in each layer and trained it for 2000 epochs. The resulting accuracy of the Partial Blind Spot Analysis using this approach is approximately **85.57%** with a standard deviation of 0.521%.

As we increased the number of nodes in the comparator networks, and increased the number of training epochs, the performance on the validation set did not fall, suggesting that overfitting has not yet occurred. Thus, the comparator network topology remains an open issue. We consider it likely that experiments to learn a likelihood calculation function per se may be fruitful, and this could reduce the training burden on the comparator in a similar sense to the way our subdivision of the classification problem into abduction processes makes more efficient use of data.

6.1 Other footprint sizes

Guided by findings of Thornley and Petridis in [17], we have chosen to focus on a regressor for peak height estimation based on the (3,1) footprint, *i.e.* by considering three bases to the left and one base to the right. We also briefly experimented with a (2,2) footprint.

The success rate of PBSA using a tree regressing for deviation evidence on a (2,2) footprint is approximately 56%. The result is consistent with results by Thornley and Petridis in [17], where it is suggested that 3 bases to the left and 1 base to the right have the greatest influence on the peak height at a given position and hence we expected the PBSA carried out with a (3,1) footprint to be more successful than PBSA carried out with (2,2) footprint. The difference is however small, as it was in [17] suggesting that the second base position after the basecall has an influence on its peak size behaviour.

We also tested the PBSA classification rate of a tree regressing for probability evidence. The resulting success rate of 52% shows a similar drop in performance between deviation and probability evidence as in trees using a (3,1) footprint. We conclude that the optimal 5 base regression footprint is (3,1), but it may be worthwhile investigating a (3,2) footprint in further work. In direct classification, this yielded a small advantage.

7 Discussion and conclusions

We have demonstrated that equal weighted sums of likelihood or deviation evidence of predicted peak heights for contextual comparison can produce excellent results using isolated contextual evidence, as evidenced by the classification rate of **79.18%** achieved with a relatively simple regressor. By machine learning an optimal use of this information we achieve a rate of **85.57%**. This classification rate is favourably comparable to our results using bagged neural net direct classifiers [17]. This adds weight to our suggestion that we may sensibly compare the likelihood of competing hypotheses for base sequence in terms of trace peak behaviour. We can now proceed with investigation directed at generalizing the approach to the lower quality data using regressors of similar form to those produced here.

We expected the provision of an estimate of the variance associated with a prediction to render a tree classifier more effective for use in probabilistic hypothesis assessment. The superior performance of the neural net based regressor in both even weighted summation and a machine learned comparison function was therefore surprising. We can suggest that the comparison net is effectively forming an internal model of the variance expected for the predictions. Similarly, the fact that the success rate of a comparison neural nets trained on the output of a tree regressor using any of the forms of evidence we have experimented is approximately equal suggests that the information is being remodelled to an extent. This may also suggest some notion of maximum effectiveness of regression trees applied to static, purely local information in the basecall abduction process.

Artificial neural networks have achieved excellent success rates for both direct base classification and abductive calling. The abuction approach is necessary for generalization to poor data, since it is not possible to train a direct classifier to use contextual information when the pivot data is provided. We wish to use all the information available, and we have demonstrated the ability to train an effective, complete regressor. The evidence from diferent locations in the context may be ascribed approximately equal weight, and we may be able to derive a more accurate estimate of their relative importance from analysis of the comparator net. We aim to discover the relative weights to be applied to pivot and context information in low quality data through processing a larger genomic sequencing project with an excellent alignment and high quality consensus sequence provided to us by the Wellcome Trust Sanger Institute.

References

[1] International Patent Application WO96/20286 July 4, 1996, European Patent EP0799320 Mar. 7 2001 and US Patent 6,090,550, Jul. 18, 2000.

[2] L. Andrade and E. S. Manolakos. Skyline Normalization of DNA Chromatograms by Regression. *Workshop On Genomic Signal Processing and Statistics (GENSIPS)*, pages CP2–7:1–4, 2002.

[3] E. B, H. L, W. MC, and G. P. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, 8(3):175–185, March 1998.

[4] E. B and G. P. Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8(3):186–194, March 1998.

[5] A.-P. Blanchard. *Sequence Specific Effects on the Incorporation of Dideoxynucleotides by a Modified T7 Polymerase*. PhD thesis, California Institute of Technology, 1993.

[6] J. M. Bowling, K. L. Bruner, J. L. Cmarik, and C. Tibbetts. Neighbouring nucleotide interactions during DNA sequencing gel electrophoresis. *Nucleic Acids Research*, 19:3089–3097, 1991.

[7] C. Connel, S. Fung, C. Heiner, J. Bridgham, V. Chakerian, E. Heron, B. Jones, S. Menchen, W. Mordan, M. Raff, M. Recknor, L. Smith, J. Springer, S. Woo, and M. Hunkapiller. Automated DNA Sequence Analysis. *BioTechniques*, 5:342–348, 1987.

[8] G.Cybenko. Approximation by superposition of a Sigmoidal Function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989.

[9] R. J. Lipschutz, F. Taverner, K. Hennesy, G. Hartzell, and R. Davis. DNA sequence confidence estimation. *Genomics*, 19:417–424, 1994.

[10] L. T. Parker, Q. Deng, H. Zakeri, D. A. Carlson, D. A. Nickerson, and P.-Y. Kwok. Peak height variations in automated sequencing of PCR products using Taq dye-terminator chemistry. *BioTechniques*, 19:116–121, 1995.

[11] L. T. Parker, H. Zakeri, Q. Deng, S. Spurgeon, P.-Y. Kwok, and D. A. Nickerson. AmpliTaq DNA Polymerase, FS Dye-Terminator Sequencing: Analysis of Peak Height Patterns. 1996.

[12] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. *Proceedings of the IEEE International Conference on Neural Networks*, 1993.

[13] F. Sanger, S. Nicklen, and A. Coulson. DNA sequencing with chain terminator inhibitors. *Proc. Natl. Acad. Sci.*, 74:5463–5467, 1977.

[14] D. Thornley. *Analysis of Trace Data from Fluoresence Based Sanger Sequencing*. PhD thesis, Imperial College London, September 1997.

[15] D. Thornley. Modelling along the DNA template in the Sanger method: inhibition through competition and form. In *Process Algebra and Stochastically Timed Activities 2006*, June 2006.

[16] D. Thornley. Trace modelling for abduction basecalling. Technical Report 7, July 2006.

[17] D. Thornley and S. Petridis. Machine Learning in Basecalling - Decoding trace peak behaviour. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology 2006*, September 2006.

[18] D. Thornley and S. Petridis. Decoding Trace Peak Behaviour - A Neuro-Fuzzy Approach. In *Proceedings of the IEEE International Conference on Fuzzy Systems*, July 2007.