Uncovering a Novel Mechanism for Gene Expression Regulation with a Random Forest Classifier

Anonymous Author(s)

Affiliation
Address
@email

Abstract

*Trypanosoma brucei* is a devastating parasite which infects tens of thousands of people in sub-Saharan Africa, causing a deadly disease known as African sleeping sickness [1]. The parasite is transmitted from human to human via an insect vector known as the tsetse fly, which feeds on human blood [1]. *T. brucei* possesses a highly unusual genome. Unlike most organisms in its phylogenetic domain, this single-celled protozoan and several other closely related species lack a vital biological mechanism for regulating the expression of their genes [2]. Scientific evidence indicates that the parasite does indeed feature different levels of gene expression in its two host environments (the human bloodstream and the tsetse fly midgut) [2]. This suggests that *T. brucei* exploits a different mechanism than most other organisms for regulating the expression of its genes [2]. This study will explore the hypothesis that the parasite exploits the change in temperature across its two host environments to induce changes in RNA structure that affect gene expression. Using a random forest classifier, we demonstrate that nine variables describing change in RNA structure for *T. brucei* genes across host temperatures may be successfully utilized to classify genes according to their expression in the parasite’s two hosts. A bootstrapping procedure is implemented to verify that the classifier performs significantly better than chance.

1 Introduction

1.1 Biological background

1.1.1 Gene expression

All living cells contain deoxyribonucleic acid (DNA) [3]. DNA is a long molecular chain composed of four types of small subunits called nucleotides [3]. The function of DNA is to store information vital to cellular processes. This information is encoded by the order of the nucleotides in the DNA chain, known as the DNA ‘sequence’ [3]. In order to use the information stored in DNA, living cells make molecular ‘copies’ of certain crucial portions of the DNA sequence, known as genes [3]. These copies are made of ribonucleic acid (RNA), a compound similar to DNA but one that is more easily degraded [3]. Once synthesized, most RNA molecules are used as templates for the production of proteins, which are the molecular machines that carry out critical cellular functions. Proteins may function as enzymes (biological catalysts), structural building blocks (responsible for cell shape and motility), and signalling molecules (such as hormones), for example [3].

Making RNA copies of DNA is critical to cellular functioning [3]. When an organism is exposed to different environmental conditions (e.g. a high concentration of sugar) it may require more proteins to cope with this environmental change (e.g. more enzymes that break down sugar for energy) [3]. In order to cope, the cell cannot alter the amount of DNA it possesses, but it can respond by making
more RNA copies of certain DNA genes under certain conditions, and less under other conditions [3]. This is known as ‘differential gene expression’ [3].

1.1.2 Gene expression in *Trypanosoma brucei*

The single-celled parasite *Trypanosoma brucei* is the causative agent of African sleeping sickness, which affects an estimated 30,000 people worldwide [1]. It is transmitted from human to human by the tsetse fly, an insect found in certain sub-Saharan regions which feeds on human blood [1]. The parasite is highly adept at evading the human immune system, and the disease is therefore difficult to treat [4]. The development of effective drugs that are affordable to people living in affected regions has been hampered by the lack of economic incentive for pharmaceutical companies, and the World Health Organization has therefore classified African sleeping sickness as a Neglected Tropical Disease [1]. Currently, the sole affordable medications available are arsenic-based, cause side effects similar to arsenic poisoning, and are themselves fatal in approximately 1 in 10 cases [1]. If left untreated, the disease is fatal in 100% of cases [1].

*T. brucei* is a relatively simple parasite which can survive only in two specific environments: the human bloodstream, and the midgut of the tsetse fly [2]. These environments are sufficiently different that the parasite must differentially regulate genes in the two hosts [2]. However, past research has demonstrated that *T. brucei* features a highly unusual genome, and lacks a crucial mechanism for regulating how many RNA copies of each DNA gene are synthesized [4].

As Figure 1A illustrates, most eukaryotic (complex) cells independently control the expression of individual genes. (B) In *Trypanosoma brucei*, genes are expressed together. The resulting long RNA molecules are then separated, such that an equal number of copies of each gene is produced. Through an unknown mechanism, some of these RNA copies are then degraded, permitting differential gene and protein expression.
The mechanism by which certain RNA molecules are preserved in *T. brucei* while others are degraded has long puzzled researchers [2]. In this paper, we propose a novel mechanism which may underlie differential gene expression in the two hosts of *T. brucei*. Using a dataset acquired through RNA-sequencing, we apply machine learning methods to test the hypothesis, demonstrate the robust nature of the results, and propose future experiments which may validate our findings.

### 1.2 Hypothesis: temperature-induced changes in RNA structure

RNA molecules, like DNA, are long chains composed of small nucleotide subunits. DNA consists of two opposing strands of nucleotides linked by chemical bonds, such that it remains stable and linear. RNA, on the other hand, is single-stranded. Due to this fact, molecular bonds can form between different nucleotides within an RNA chain, giving each RNA molecule a distinct structure that is dictated by its nucleotide sequence.

Each RNA molecule is composed of three distinct regions. The coding region of RNA is the one utilized in the production of protein, while the 5’ untranslated region (5’UTR) and 3’ untranslated region (3’UTR) are regulatory regions [3]. Past studies indicate that structures which form within the two regulatory regions may serve as signals to target certain RNA molecules for degradation [5].

Temperature affects the stability of bonds between RNA nucleotides [5]. Human bloodstream temperature is approximately 37°C, while the temperature of the tsetse fly is closer to the ambient temperature in sub-Saharan Africa, approximately 28°C [5]. Thus, the 5’UTR and 3’UTR regions of *T. brucei* RNA may have evolved to form different structures at these two temperatures, and the change in structure across temperatures may serve as a cue to differentially regulate the expression of genes in the two host environments. Such temperature-induced regulation of gene expression via RNA structure has not been previously reported in any organism.

### 1.3 Research aims

In order to test the above hypothesis we utilize a comprehensive dataset featuring the expression of thousands of *T. brucei* genes in the parasite’s two forms: the human bloodstream form (BF), and the insect or ‘procyclic’ form (PF). We set out to computationally predict the structures that form in the 5’UTR regulatory region of these genes at both 37°C and 28°C, and compute the magnitude of structural difference (for each gene) across these two temperatures using several distance measures.

Our aim is then to predict gene expression in the two forms of the parasite, using only those measures of structural difference across the two temperatures. If this classification method performs significantly better than chance, then it supports the hypothesis that changes in RNA structure due to differences in temperature mediate gene expression regulation across the two host environments of *Trypanosoma brucei*.

## 2 Methods and Results

### 2.1 Overview

The dataset utilized consists of 5199 genes (the samples), their class labels (highly expressed in humans, highly expressed in the tsetse fly, or not differentially expressed), and nine measures of the magnitude of change in the gene’s RNA structure across temperatures (37°C in humans versus 28°C in the tsetse fly). A random forest classifier was trained on a portion of the data, and its predictive accuracy tested on the remainder of the data. The model parameters were optimized using cross-validation, and the misclassification error rate was shown to be consistently and significantly below chance using a bootstrapping procedure. Finally, the relative importance of the nine predictors for classification was evaluated.

### 2.2 Data preparation

We utilized the structure modelling program Mfold to predict the structure of the 5’UTR segments of the RNA molecules corresponding to $N = 5199$ genes, at both 37°C and 28°C [6].
Based on their expression in the human bloodstream form (BF) of the parasite and the insect ‘procyclic’ form (PF) of the parasite, these genes were placed in one of three mutually exclusive classes:

1. Class BF: genes more than two-fold over-expressed in BF relative to PF (362 genes or 6.96% of the dataset)

2. Class PF: genes more than two-fold over-expressed in PF relative to BF (313 genes or 6.02% of the dataset)

3. Class N: genes expressed approximately equally in the parasite’s two forms (4524 genes or 87.02% of the dataset)

Next, we computed nine different measures (the predictors or explanatory variables), which aimed to capture different aspects of the magnitude of change in RNA structure across temperatures. For all measures, the change represents the value at $37^\circ$C (BF) less the value at $28^\circ$C (PF).

Simple measures include: (1) the change in the fraction of nucleotides forming bonds ($\Delta$Bonds); (2) the change in the number of stem structures, which are defined as segments of four or more consecutive bonds ($\Delta$Stems); and (3) the change in the number of hairpin structures, which are segments of consecutive bonds that end in a short loop of unbound nucleotides ($\Delta$Hairpins).

Four distance metrics previously developed to quantify structural difference in RNA were also computed (for details see Moulton et al. [7]). These metrics are: (4) the tree edit distance ($\Delta$Tree); (5) the symmetric distance ($\Delta$Symmetric); (6) the Hausdorff distance ($\Delta$Hausdorff); and (7) the mountain distance ($\Delta$Mountain).

Finally, two additional measures included: (8) the difference in the free energy of the two structures ($\Delta$G); and (9) the change in the number of additional, sub-optimal structures predicted by Mfold ($\Delta$NumStructs).

### 2.3 Classification with random forests

The dataset, consisting of nine predictors and a class label for each gene, was utilized to train and test a random forest classifier using the Random Forest package in Matlab. Random forests were selected as they provide a low-bias estimate of the misclassification error rate, and can be used to determine the relative importance of the predictor variables, which aids in data interpretation.

As described above, approximately 87% of genes fell into one of the three classes (Class N). When training models with such a significant class imbalance, the classifier tends to assign almost all samples to the largest class, resulting an extremely low misclassification error rate (MCR) for that class and extremely high MCR for the other classes. Since the MCR of a test set is a weighted average of class MCRs (weighted by the size of the class), the overall error rate for the test set then appears unrealistically low when in fact the model is unable to distinguish between classes.

In order to account for class imbalance, the following procedure was adopted to construct the random forest. First the data was split into a training set (consisting of 70% of samples) and a test set (the remaining 30% of genes). The data was partitioned in this way so as to maximize the size of the training set, while still including approximately 100 genes from each of the two small classes in the test set. The number of samples in the smallest class of the training set was identified ($n_s$). Due to computational limitations, $T = 500$ trees were constructed for each run of the random forest. For each tree, an equal number of samples was drawn, with replacement, from each of the three classes in the training set. The size of the sample drawn from each class was $f_s \cdot n_s$, where $f_s$ is a parameter to be optimized. This parameter was introduced due to past research, which indicates that moderate over-sampling of the smaller classes (as opposed to simply under-sampling the largest class) improves classification with imbalanced datasets [8]. At each tree node, a number of predictor dimensions determined by the parameter $n_d$ was randomly selected. Along each dimension, all possible axis-aligned splits were examined, and the split with greatest increase in node purity (as measured by the Gini index) was selected for the given node. This was repeated until $n_{\min}$ samples remained in the leaf nodes.
2.4 Evaluating model reliability with bootstrapping

For the purpose of hypothesis testing, achieving high accuracy (e.g. through Bayesian optimization of the model parameters) is less important than demonstrating that the model performs significantly better than chance. For this reason, a bootstrapping procedure was implemented to provide an improved estimate of the misclassification error rate, as well as estimate the variance in the estimator.

Once again, the imbalanced nature of the classes presented a challenge. Since the size of the two smaller classes is approximately 6% of the original dataset, any bootstrap sample drawn from the original dataset will contain even fewer unique samples from the two small classes. The balancing procedure implemented in the random forest involves drawing samples from each class in the training set of a size similar to that of the smallest class (weighted by the parameter \( f_s \)). When combined with the bootstrapping procedure, this would result in forests trained on rather small sample sizes.

To account for this problem, a bootstrap procedure for estimating the MCR was implemented as described in Fu et al. (2005) for gene expression data \([9]\). This technique, termed bootstrap cross-validation (BCV), was demonstrated to provide superior estimation of the MCR using small simulated datasets (compared to the leave-one-out cross-validation, leave-one-out bootstrap, and bootstrap 632 estimators). For a dataset with \( N \) samples, the BCV procedure involves drawing \( b = 1, \ldots, B \) bootstrap samples of size \( N \) from the original dataset. For each bootstrap sample, cross-validation is performed on the random forest, and an error estimate is obtained. The bootstrap estimate of the MCR is then computed as the average of the \( B \) bootstrap samples. The standard deviation for the \( B \) samples is an estimate of the standard error for the misclassification rate.

2.5 Parameter optimization using cross-validation

Once again, the aim of this study is not to construct a model that can classify \( T. brucei \) genes with high accuracy, but rather to test the hypothesis that the predictors have some bearing on the class labels. This reason, combined with the fact that the hypothetical range of the model parameters is limited, make Bayesian optimization unnecessary. Instead, reasonable parameter values were obtained through two-fold cross-validation.

To reduce variability in the estimates, the bootstrap average error misclassification rates were computed for the training and test sets over the range of parameter values \((B = 50)\). The initial parameters were set to: \( f_s = 1 \) (no over-sampling of the smallest training class), \( n_d = 3 \) (as it is common practice to set this parameter to the square root of the number of dimensions), and \( n_{\text{min}} \) (corresponding to fully branching trees) \([10]\). The parameters were optimized sequentially \((f_s, n_d, n_{\text{min}})\), such that once optimization was completed for one parameter, this value replaced the default value for the next parameter’s optimization. The results of cross-validation are presented in Figure 2.

The parameter \( f_s \) determines the size of the sample drawn from each class for the construction of trees, as a fraction of the size of the smallest class \( n_s \). It was explored over the range of values \( \{1.0, 1.1, 1.2, 1.3, 1.4, 1.5\} \) to determine the optimal degree of over-sampling from the smallest training class. As the top-left plot in Figure 2 illustrates, the overall errors of both the training and test set decrease as \( f_s \) increases. However, a breakdown of the test misclassification by class (BF, PF, and N) reveals that the decrease in the overall test MCR is due entirely to a decrease in the MCR of the largest class (Class N). This class is drastically under-sampled, so accuracy is expected to improve as the value of \( f_s \) increases. The overall test error is a weighted average of the test class error rates, and is therefore skewed towards the error rate of the largest class. This is evident in Figure 2, as the curve of the overall test error (shown in red) closely mimics that of test class N. Since we are interested in minimizing error for all three classes equally, the unweighted mean of the three test classes was computed (shown in magenta). Unlike the overall test error, this mean is not skewed towards the largest class. The min-max principle was applied to the unweighted means of the training and test classes to determine that the optimal value of this parameter is \( f_s = 1.3 \), which corresponds to over-sampling the smallest training class by 30%.

The parameter \( n_d \) determines the number of dimensions sampled at each node during tree construction (top-right plot of Figure 2). Since the dataset features nine predictors, this parameter was explored over the range of values \( \{2, 3, 4, 5, 6, 7, 8, 9\} \). It is evident that for both the training and test sets, the misclassification error rate decreases as more parameters are sampled, but levels off at \( n_d = 7 \). This may suggest that a few of the predictors are substantially more important for clas-
sification, and that once enough dimensions are sampled such that at least one of these important predictors are highly likely to be examined, the decrease in the error rate stabilizes. Based on these results, the min-max principle suggests fixing the parameter at $n_d = 7$. However, the results also imply that examining the importance of the nine predictors is worthwhile. That analysis is described in the proceeding sections.

Finally, the parameter $n_{\text{min}}$ represents the minimum number of samples in the leaf nodes (bottom-right plot of Figure 2). It was explored over the range $\{1, 2, 3, 4, 5, 6\}$. As this parameter increases, the misclassification errors for the training set, test set, and mean of the class test sets all increase. This parameter was therefore fixed at $n_{\text{min}} = 1$, corresponding to fully branching trees.

![Figure 2: Cross-validation for optimizing random forest parameters. Top-left: optimization for $f_s$ the coefficient affecting the degree of over-sampling from the smallest class in the training set of each tree. Top-right: optimization for $n_d$ the number of predictor dimensions sampled at each tree node. Bottom-right: optimization for $n_{\text{min}}$ the minimum number of samples present in the leaf nodes of each decision tree.](image)

2.6 Evaluating feature importance

Feature importance was evaluated by computing the total decrease in node impurity achieved by splitting on a given predictor variable, averaged over all trees in a forest and all bootstrapped samples. This computationally intensive procedure was undertaken using $B = 50$ bootstraps and $T = 500$ trees, with two sets of parameters. In both cases, the parameters were set such that $f_s = 1.3$ and $n_{\text{min}} = 1$, as suggested by the optimization procedure. In order to determine whether the results of optimizing parameter $n_d$ were due to the fact that a few predictor variables were substantially more important for classification, importance evaluation was carried out using both $n_d = 3$ (the default value) and $n_d = 7$ (the apparent optimal value).

The results of this analysis are presented in Table 1. It is clear that $\Delta G$, the difference in free energy of the RNA structure across the two host temperatures, is the most important feature for classification. When the number of dimensions sampled at each node was increased from $n_d = 3$ to $n_d = 7$, this variable appears to have been selected for splitting more often, and the total decrease in node impurity attributed to this variable increased dramatically. The free energy of a molecule represents the amount of work that it can perform [6]. The fact that the difference in the amount of work the RNA can perform across the temperatures is important for predicting expression, suggests
that a difference in the strength of bonds within the RNA across hosts may be exploited by *T. brucei* to regulate gene expression.

Two other variables appear to have moderate importance in the classification task. These include ∆NumStructs and ∆Bonds. The variable ∆NumStructs represents the difference in the number of alternative structures generated by the structure prediction program Mfold, across the two temperatures. It may be interpreted as a measure of the difference in structural flexibility across the two temperatures. Thus, a gene forming a single stable structure at one host temperature, and a diverse ensemble of structures at the other host temperature, may serve as a cue for gene expression regulation in *T. brucei*. This is consistent with the interpretation for ∆G above. The variable ∆Bonds, the difference in the number of nucleotides forming bonds across temperatures, may be interpreted (as with ∆G) to indicate that a change in the strength of bonds within the molecular structure across temperatures may be exploited for gene expression regulation.

All variables in the dataset appear somewhat useful for classification, though they measured changes in structure at different levels. The four metrics—∆Mountain, ∆Symmetric, ∆Hausdorff, and ∆Tree—all measure large-scale changes in structural configuration, and appear to have a similar level of importance for classification. The variables ∆Hairpins and ∆Stems, on the other hand, are simplistic measures representing change in the number of small-scale structural features. These appear to be the least useful for classification.

Overall, the relative importance of the nine measures for classification are consistent with common-sense interpretation, and appear to suggest that as *T. brucei* transitions from its insect host to its human host, the corresponding change in temperature affects the stability of RNA molecules within the cell in a manner that affects gene expression.

Training the model only on the top four predictor variables (with all other variables removed) was attempted, but resulted in a modest increase in misclassification error (from 0.29 to 0.31 for the test set error, using $B = 50$ bootstraps and $T = 500$ trees). In the proceeding section, we present a final bootstrap estimate for the misclassification error rates using all variables and the optimized parameter values. An estimate of the variance for the error rate is also provided.

### Table 1: Importance of the nine predictor variables for classification.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$n_d = 3$</th>
<th>$n_d = 7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>∆G</td>
<td>97.43</td>
<td>201.12</td>
</tr>
<tr>
<td>∆NumStructs</td>
<td>50.15</td>
<td>72.85</td>
</tr>
<tr>
<td>∆Bonds</td>
<td>54.51</td>
<td>60.37</td>
</tr>
<tr>
<td>∆Hausdorff</td>
<td>45.65</td>
<td>47.17</td>
</tr>
<tr>
<td>∆Mountain</td>
<td>44.60</td>
<td>39.94</td>
</tr>
<tr>
<td>∆Symmetric</td>
<td>42.97</td>
<td>40.86</td>
</tr>
<tr>
<td>∆Tree</td>
<td>38.67</td>
<td>32.64</td>
</tr>
<tr>
<td>∆Stems</td>
<td>24.42</td>
<td>26.11</td>
</tr>
<tr>
<td>∆Hairpins</td>
<td>20.17</td>
<td>20.57</td>
</tr>
</tbody>
</table>

### Table 2: Estimate of misclassification rates and standard errors.

<table>
<thead>
<tr>
<th></th>
<th>MCR</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>0.1805</td>
<td>0.0157</td>
</tr>
<tr>
<td>Training class BF</td>
<td>0.0051</td>
<td>0.0048</td>
</tr>
<tr>
<td>Training class PF</td>
<td>0.0066</td>
<td>0.0055</td>
</tr>
<tr>
<td>Training class N</td>
<td>0.2066</td>
<td>0.0177</td>
</tr>
<tr>
<td>Test set</td>
<td>0.2960</td>
<td>0.0179</td>
</tr>
<tr>
<td>Test class BF</td>
<td>0.3599</td>
<td>0.0550</td>
</tr>
<tr>
<td>Test class PF</td>
<td>0.4134</td>
<td>0.0585</td>
</tr>
<tr>
<td>Test class N</td>
<td>0.2828</td>
<td>0.0206</td>
</tr>
</tbody>
</table>

#### 2.7 Estimate of misclassification rate and standard error

The final estimate of the misclassification rate and the variance for this estimator are presented in Table 2. Using the parameter values determined through cross-validation ($f_s = 1.3$, $n_d = 7$, and $n_{\min} = 1$) the model was evaluated based on $B = 500$ bootstrap forests, each with $T = 500$ trees.

By chance alone, the expected error when classifying samples into three classes is $\frac{2}{3} \approx 0.6667$. A 95% confidence interval for the results may be obtained by adding and subtracting $2 \cdot SE$ (twice the estimated standard error) from the estimated misclassification error. The results above indicate that, for all classes in the test set, the model performed significantly better than chance, as the upper limit of the confidence interval falls well below 0.6667. This supports the hypothesis that the predictor variables have some bearing on the class labels. The implications of these findings are discussed below.
3 Conclusions and Future Directions

Trypanosoma brucei is a devastating parasite with unusual genetic characteristics [4]. As the causative agent of a Neglected Tropical Disease with few affordable treatments, gaining a better understanding of the biology of this parasite is critical to improving human welfare in affected regions [1]. The results of this analysis suggest that the parasite may utilize a novel mechanism for gene expression regulation, exploiting the difference in temperature across its two host environments (the human bloodstream and the midgut of the tsetse fly) as a means of inducing changes in RNA structure that affect gene expression.

Identifying biological mechanisms that are unique to a disease-causing agent and not required by the host organism is an important step in the development of treatment [11]. For example, a class of anti-malarial drugs was developed to target the synthesis of dihydropteroate synthetase (DHPS), which is required by the malaria parasite for survival, but is not found in humans [11].

By harnessing the predictive power of machine learning tools, we demonstrated that variables capturing change in the RNA structure of genes across host temperatures may be utilized to predict gene expression, with a success rate well above chance. However, it should be emphasized that the predictor variables could not perfectly account for gene expression, and that many genes could not be classified correctly. This is to be expected in a biological system, where many complex mechanisms are at play and where any given mechanism may apply to only a subset of genes.

Validating these findings through biological experiment is the natural next step. One approach to validation may be selecting several genes that exhibit differential expression and are successfully classified by the random forest model, and subjecting their 5’UTR region to mutations that alter RNA structure. For example, the structure found in a gene over-expressed in the human host may be swapped with one found in a gene over-expressed in the insect host, and vice versa. T. brucei cells carrying these mutations may then be grown at the human host temperature (37°C) and at the insect temperature (28°C), and gene expression may be measured to determine whether the expected switch in over-expression is observed.

The possibility that the same mechanism for gene expression regulation evolved in other parasites which transition between warm-blooded hosts and insect vectors should be examined. For example, single-celled parasites of the genus Leishmania cause Leishmaniasis, an endemic neglected disease affecting approximately 12 million people in South America and southern Asia [12]. The genus Leishmania is closely related to Trypanosoma, possess a similarly unusual genome, and also cycles between an insect vector and a warm-blooded human host. Performing a complementary computational analysis for this genus of parasites may therefore prove worthwhile.

References