A BAYESIAN APPROACH TO TOP-SCORING PAIRS CLASSIFICATION

Emre Arslan and Ulisses M. Braga-Neto

Texas A&M University
Department of Electrical and Computer Engineering
College Station, TX, USA

ABSTRACT

We extend the popular Top Scoring Pair (TSP) classification rule to a Bayesian setting, with the purpose of obtaining robust and effective classifiers for small-sample, high-dimensional data. We employ the Bradley-Terry model for rank data, and infer its parameters using a previously proposed Gibbs sampling algorithm. The parameters are then used to define a Bayesian TSP score, which is used to select the gene pairs to define the proposed Bayesian TSP classifiers. Accuracy of the proposed Bayesian classification rules is evaluated against those of the conventional TSP classifiers as well as other well-known machine learning methods, using a total of 12 gene-expression data sets. The results indicate that the Bayesian k-TSP classifier obtained the best overall average accuracy rate and the best accuracy rate over the majority of the individual data sets.

Index Terms— Gene expression classification, Top Scoring Classifier, Bayesian methods.

1. INTRODUCTION

Classification of high-dimensional gene expression data has been a topic of ongoing interest in Genomics, with applications in diagnosis and prognosis of cancer, infectious diseases, and more [1, 2, 3, 4, 5, 6, 7, 8]. When selecting a classification algorithm, a balance has to be struck between “black-box” accuracy and interpretability; many classification rules appear to produce accurate results on test data, but they produce complex decision boundaries that have little hope of interpretability and further validation by physicians and other biomedical experts. The Top-Scoring Pair (TSP) classifier was introduced by D. Geman and collaborators in [9], and further studied in [10, 11, 12]. It is a rank-based classifier that ignores the raw gene expression values and instead seeks simple rank changes across the two conditions. Its main virtue is to produce interpretable classifiers, which are yet powerful in terms of its classification accuracy rates.

In this paper, we extend the basic TSP and k-TSP classification rules to a Bayesian setting, with the purpose of obtaining robust and effective classifiers for small-sample, high-dimensional data. We employ the Bradley-Terry model for rank data [13], and infer its parameters using the Gibbs sampling algorithm proposed by Caron and Doucet in [14]. The parameters are then used to define a Bayesian TSP score, which is used to select the gene pairs to define the proposed Bayesian TSP classifiers. Accuracy of the proposed Bayesian classification rules is evaluated against those of the conventional TSP classifiers as well as other well-known machine learning methods, using a total of 12 gene-expression data sets. The results indicate that the Bayesian k-TSP classifier obtained the best overall average accuracy rate and the best accuracy rate over the majority of the individual data sets.

2. BRADLEY TERRY MODEL

We begin by introducing the Bradley-Terry model. Let \( i \) and \( j \) be a pair of individuals randomly drawn from a population of size \( M \). The Bradley-Terry model [13] stipulates that:

\[
\pi_{ij} = P(i > j \mid \lambda_i, \lambda_j) = \frac{\lambda_i}{\lambda_i + \lambda_j},
\]

where parameter \( \lambda_i > 0 \) can be interpreted as the “skill of player \( i \),” for \( i = 1, \ldots, M \). For our purposes, \( \lambda_i \) denotes the propensity of the rank of gene \( i \) being larger than the rank of gene \( j \). As expected, (1) implies that \( \pi_{ij} + \pi_{ji} = 1 \). The Bradley-Terry (BT) model can also be identified as a logistic model by a non-linear reparameterization of the parameters, \( \lambda_i = e^{\beta_i} \), in which case

\[
\pi_{ij} = \frac{1}{1 + e^{-(\beta_i - \beta_j)}} = \eta(\beta_i - \beta_j)
\]

where \( \eta(\alpha) = 1/(1 + e^{-\alpha}) \) is the inverse logit function [15]. There are numerous applications of the BT Model, such as ranking of Chess and Go players by their respective international federations, estimation of the influence of statistical journals, and more.

Given sample data, let \( w_{ij} \) denote the number of times \( i > j \), \( w_i \) be the number of “wins” by \( i \), and \( n_{ij} \) represent the number of comparisons between \( i \) and \( j \) over the data. These statistics can be used to estimate the model parameters \{\lambda_i\} by maximum likelihood (ML). If pairings are assumed
independent, the log-likelihood function for the BT Model is given by
\[ \ell(\lambda) = \sum_{1 \leq i \neq j \leq M} w_{ij} \log \lambda_i - w_{ij} \log(\lambda_i + \lambda_j) \]
\[ = \sum_{1 \leq i \neq j \leq M} w_{ij} \log \lambda_i - \sum_{1 \leq i < j \leq M} n_{ij} \log(\lambda_i + \lambda_j). \] (3)

The ML estimator can be found by an iterative procedure [16],
\[ \lambda_i^{(k+1)} = w_i \left( \sum_{i \neq j} \frac{n_{ij}}{\lambda_i^{(k)}} + \lambda_j^{(k)} \right)^{-1}, \quad i = 1, \ldots, M, \] (4)
which is repeated until convergence. The main drawback with the ML approach is the strong assumption that no player may win all games against another (i.e., \( n_{ij} \neq 0 \) for all \( i, j \)).

3. BAYESIAN INFERENCE FOR THE BRADLEY-TERRY MODEL

To overcome the difficulties associated with maximum-likelihood methods, Caron and Doucet [14] introduced a Bayesian approach to the inference of the parameters \( \lambda_i \) in the BT model, which we briefly describe next.

The BT Model has a “Thurstonian” interpretation: for each pair \( 1 \leq i < j \leq K \) and for each pair comparison \( k = 1, 2, \ldots, n_{ij} \), consider two independent random variables \( Y_{ki} \sim \text{Exp}(\lambda_i) \) and \( Y_{kj} \sim \text{Exp}(\lambda_j) \), where \( \text{Exp}(\lambda) \) denotes an exponential distribution with rate parameter \( \lambda \). Then a simple calculation reveals that
\[ P(Y_{ki} < Y_{kj}) = \frac{\lambda_i}{\lambda_i + \lambda_j}. \] (5)

In order to get a simpler complete log-likelihood, Caron and Doucet seek to introduce new latent variables \( Z_{ij} \) given by
\[ Z_{ij} = \min_{k=1}^{n_{ij}} \{Y_{kj}, Y_{ki}\}. \] (6)
Owing to the facts that \( U_k = \min \{Y_{kj}, Y_{ki}\} \sim \text{Exp}(\lambda_i + \lambda_j) \) and that \( U_k \) and \( U_l \) are independent for \( k \neq l \), we conclude that \( Z_{ij} \sim \text{Gamma}(n_{ij}, \lambda_i + \lambda_j) \), a Gamma distribution with shape parameter \( n_{ij} \) and rate parameter \( \lambda_i + \lambda_j \). The resulting density of the variables \( Z = \{Z_{ij}\} \) given the data \( D \) and vector of parameters \( \lambda \) is
\[ p(z \mid D, \lambda) = \prod_{1 \leq i < j \leq M \atop s.t. n_{ij} > 0} \text{Gamma}(z_{ij}; n_{ij}, \lambda_i + \lambda_j) \] (7)

and the resulting complete data log-likelihood is,
\[ \ell_c(\lambda) = \sum_{1 \leq i \neq j \leq M \atop s.t. w_{ij} > 0} w_{ij} \log \lambda_i - \sum_{1 \leq i < j \leq M \atop s.t. n_{ij} > 0} (\lambda_i + \lambda_j)z_{ij} + (n_{ij} - 1) \log z_{ij} - \log \Gamma(n_{ij}) \] (8)

where \( \Gamma \) is the Gamma function.

The prior for \( \lambda \) is assigned as in [17, 18]:
\[ p(\lambda) = \prod_{i=1}^{M} \text{Gamma}(\lambda_i; a, b), \] (9)
where \( a \) and \( b \) are hyper parameters. At this point, one needs to sample from the posterior \( p(\lambda \mid Z, D) \). Canon and Doucet suggest the following Gibbs sampling scheme to accomplish that. First, update \( Z \) from the previous value of \( \lambda \) using (6):
\[ Z_{ij}^{t+1} \mid D, \lambda^t \sim \text{Gamma}(n_{ij}, \lambda_i^t + \lambda_j^t); \] (10)

Next, sample the new value of \( \lambda \) from \( p(\lambda \mid Z, D) \propto p(\lambda) \ell_c(\lambda, z) \), which has a Gamma distribution with known parameters:
\[ \lambda_i^{t+1} \mid D, Z^{t+1} \sim \text{Gamma} \left( a + w_i, \frac{b + \sum_{i < j} z_{ij}^{t+1} + \sum_{i > j} z_{ji}^{t+1}}{s.t. n_{ij} > 0} \right). \] (11)

4. TOP SCORING PAIR CLASSIFIERS

Consider \( p \) genes where \( X_i \) represents the expression value of the \( i \)-th gene. The quantity of interest is \( p_{ij}(c) = P(X_i < X_j \mid C = c) \) where \( C \) represents a given class (in this paper, we will assume the two-class case, for the sake of simplicity). A TSP classifier seeks the pair which maximizes the TSP score \( \hat{\Delta}_{ij} = |\hat{p}_{ij}(0) - \hat{p}_{ij}(1)| \), where \( \hat{p}_{ij} \) is the sample estimate of \( p_{ij} \). After choosing the pair \((i^*, j^*)\) with the highest score, and assuming \( \hat{p}_{i^*, j^*}(0) > \hat{p}_{i^*, j^*}(1) \), the TSP classifier is defined as;
\[ h_{TSP} (x_{new}) = \begin{cases} C_0, & \text{if } \text{x}_{new,i^*} < \text{x}_{new,j^*}, \\ C_1, & \text{otherwise}. \end{cases} \] (12)

The classes \( C_0 \) and \( C_1 \) are flipped in the definition of the TSP classifier if \( \hat{p}_{i^*, j^*}(0) \leq \hat{p}_{i^*, j^*}(1) \).

Many variations of TSP classifier have been proposed [12, 11, 19]. Among these, the most popular is the \( k \)-TSP classifier[10]. It is a generalization of TSP; instead of using one pair as in TSP, it selects the top \( k \) pairs according to \( \Delta_{ij} \).
(usually an odd number $k$ of disjoint pairs are selected), and performs majority voting to classify the given data point:

$$h_{kTSP}(x_{new}) = \arg\max_{C \in \{C_0, C_1\}} \sum_{r=1}^{k} I(h_r(x_{new}) = C),$$

(13)

where $h_r(\cdot)$ is the TSP classifier based on pair $r$, for $r = 1, \ldots, k$.

5. BAYESIAN TOP SCORING PAIRS

In this section we describe the application of the Bradley-Terry model in the design of TSP classifiers. Let $\lambda^0$ and $\lambda^1$ be the class-specific “skill” parameters sampled from the posterior distributions for each class separately, as described in the previous section, and define

$$\pi^0_{ij} = \frac{\lambda^0_{ij}}{\lambda^0_i + \lambda^0_j} \quad \text{and} \quad \pi^1_{ij} = \frac{\lambda^1_{ij}}{\lambda^1_i + \lambda^1_j}. \quad (14)$$

The main goal is to find pairs that swap frequently between classes in terms of $\lambda^0$ and $\lambda^1$. For this purpose, we define the Bayesian TSP score

$$\Omega_{ij} = |\pi^0_{ji} - \pi^1_{ji}| = \left| \frac{\lambda^0_{ij}}{\lambda^0_i + \lambda^0_j} - \frac{\lambda^1_{ij}}{\lambda^1_i + \lambda^1_j} \right|, \quad (15)$$

and choose the best pair according to this score,

$$(i^*, j^*) = \arg\max_{(i,j) \in S} \Omega_{ij}, \quad (16)$$

where $S$ is the whole feature space. If $\pi^0_{i^*j^*} < \pi^1_{i^*j^*}$, the Bayesian Top Scoring Pair (BTSP) classifier is defined as

$$h_{BTSP}(x_{new}) = \begin{cases} C_0, & \text{if } x_{new,i^*} < x_{new,j^*}, \\ C_1, & \text{otherwise}. \end{cases} \quad (17)$$

The classes $C_0$ and $C_1$ are flipped in the definition of the BTSP classifier if $\pi^0_{i^*j^*} \leq \pi^1_{i^*j^*}$.

As in the case of TSP, the BTSP classification rule can be extended by choosing more than one pair. We choose an odd number $k$ of disjoint top scoring pairs according to $\Omega_{ij}$ and define a classifier as in (13), where $h_r(\cdot)$ this time denotes the BTSP classifier based on pair $r$, for $r = 1, \ldots, k$. We call this the $k$-BTSP classifier.

6. EXPERIMENTAL RESULTS

We used 12 genomic data sets (Table 1) to compare the proposed Bayesian TSP classification rules against the conventional TSP classifiers as well as other well-known machine learning methods. A variance of filter was employed to reduce the number of genes in all data sets to 2000. In order to simulate small sample size problem with given data, classifiers were trained on 20% of the data and tested on the remaining 80%. This procedure was repeated 50 times and the average accuracy was recorded. The Gibbs sampler was run for 1500 iterations with 300 burn-in iterations. We found that the choice of hyper parameters of the prior distribution in (9) does not change results dramatically: $b$ is a scaling parameter and the likelihood is invariant to rescaling, so changing $b$ does not have a big effect, while $a$ may be fixed or can be calculated by a Metropolis-Hasting algorithm — in both cases, accuracies did not change much.

To determine the number of pairs to be used in $k$-TSP, we used the methods described in [10] and [12], leading to two classification rules, which we called $k$-TSP1 and $k$-TSP2, respectively. The switchBox R package was used for the $k$-TSP1 analysis [24] and the ktspair R package was used for the $k$-TSP2 analysis [25]. In each case, we limit the number of pairs to at most 9. Over all experiments, we observed that $k$-TSP1 uses an average of 4.02 genes, or about 2 pairs, while $k$-TSP2 uses an average of 17.8 genes, or nearly all 9 pairs. By contrast, we used a fixed number $k = 9$ of pairs for the k-BTSP classifier. The number of genes used by the SVM-RFE classification rule was fixed at 100, while Naïve Bayes (NB) and the Decision Tree (DT) use all 2000 genes.

The accuracy results are displayed in Table 2. We observe that $k$-BTSP obtained the best overall average accuracy rate, and has the best accuracy rate over more individual data sets than all other classification rules. There doesn’t seem to be a conclusive difference between the accuracy rates of the conventional and Bayesian TSP classifiers, but there is a significant improvement of the $k$-BTSP classifier over both conventional $k$-TSP classifiers (and of course the TSP classifiers).

$k$-BTSP is robust to monotone transformations and eliminates the variations among platforms or pre-processing techniques. Therefore, the results from $k$-BTSP are more reliable and robust compared to many well-known classifiers. $k$-BTSP is also a great candidate to define clear biological connections between genes and the studies.

7. REFERENCES


[2] Margaret A Shipp, Ken N Ross, Pablo Tamayo, Andrew P Weng, Jeffery L Kutok, Ricardo CT Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S Pinkus, et al., “Diffuse large b-cell lym-
<table>
<thead>
<tr>
<th>Study</th>
<th>Number of Genes</th>
<th>Class 1 Size</th>
<th>Class 2 Size</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>2000</td>
<td>22</td>
<td>40</td>
<td>Alon et al. (1999) [20]</td>
</tr>
<tr>
<td>Leukemia1</td>
<td>7129</td>
<td>25</td>
<td>47</td>
<td>Golub et al. (1999) [1]</td>
</tr>
<tr>
<td>DLBCL</td>
<td>7129</td>
<td>58</td>
<td>19</td>
<td>Shipp et al. (2002) [2]</td>
</tr>
<tr>
<td>Lung</td>
<td>12,533</td>
<td>150</td>
<td>31</td>
<td>Gordon et al. (2002) [3]</td>
</tr>
<tr>
<td>Leukemia2</td>
<td>12,564</td>
<td>24</td>
<td>24</td>
<td>Armstrong et al. (2002) [5]</td>
</tr>
<tr>
<td>Monocytes</td>
<td>26,496</td>
<td>49</td>
<td>47</td>
<td>Maouche et al. (2008) [21]</td>
</tr>
<tr>
<td>Squamous</td>
<td>12,625</td>
<td>22</td>
<td>22</td>
<td>Kuriakose, Chen et al. (2004) [6]</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>43,931</td>
<td>37</td>
<td>31</td>
<td>Price et al. (2007) [7]</td>
</tr>
<tr>
<td>Huntington’s</td>
<td>22,283</td>
<td>14</td>
<td>17</td>
<td>Borovecki et al. (2005) [8]</td>
</tr>
<tr>
<td>CNS</td>
<td>7129</td>
<td>25</td>
<td>9</td>
<td>Pomeroy et al. (2002) [22]</td>
</tr>
<tr>
<td>Myeloma</td>
<td>12,625</td>
<td>137</td>
<td>36</td>
<td>Tian et al. (2003) [23]</td>
</tr>
</tbody>
</table>

Table 1. Gene expression data sets used in the numerical experiment.

<table>
<thead>
<tr>
<th>Study</th>
<th>TSP</th>
<th>$k$-TSP$_1$</th>
<th>$k$-TSP$_2$</th>
<th>BTSP</th>
<th>$k$-BTSP</th>
<th>DT</th>
<th>NB</th>
<th>SVM-RFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>0.7112</td>
<td>0.7326</td>
<td>0.7392</td>
<td>0.7304</td>
<td>0.8132</td>
<td>0.6504</td>
<td>0.6344</td>
<td>0.7268</td>
</tr>
<tr>
<td>Leukemia$_1$</td>
<td>0.8789</td>
<td>0.8893</td>
<td>0.9072</td>
<td>0.8720</td>
<td>0.9168</td>
<td>0.8091</td>
<td>0.855</td>
<td>0.8896</td>
</tr>
<tr>
<td>DLBCL</td>
<td>0.779</td>
<td>0.8216</td>
<td>0.8412</td>
<td>0.7777</td>
<td>0.8761</td>
<td>0.7408</td>
<td>0.7795</td>
<td>0.877</td>
</tr>
<tr>
<td>Lung</td>
<td>0.9529</td>
<td>0.9609</td>
<td>0.968</td>
<td>0.9246</td>
<td>0.9773</td>
<td>0.9284</td>
<td>0.9678</td>
<td>0.9734</td>
</tr>
<tr>
<td>Breast</td>
<td>0.9356</td>
<td>0.9412</td>
<td>0.9465</td>
<td>0.9322</td>
<td>0.9672</td>
<td>0.8726</td>
<td>0.8918</td>
<td>0.9134</td>
</tr>
<tr>
<td>Leukemia$_2$</td>
<td>0.8936</td>
<td>0.9178</td>
<td>0.9331</td>
<td>0.8878</td>
<td>0.9573</td>
<td>0.8436</td>
<td>0.8826</td>
<td>0.9486</td>
</tr>
<tr>
<td>Monocytes</td>
<td>0.981</td>
<td>0.9838</td>
<td>0.989</td>
<td>0.9896</td>
<td>0.9901</td>
<td>0.8866</td>
<td>0.9884</td>
<td>0.9888</td>
</tr>
<tr>
<td>Squamous</td>
<td>0.796</td>
<td>0.8062</td>
<td>0.824</td>
<td>0.7668</td>
<td>0.8388</td>
<td>0.77</td>
<td>0.6871</td>
<td>0.8708</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>0.8258</td>
<td>0.8233</td>
<td>0.847</td>
<td>0.8289</td>
<td>0.8640</td>
<td>0.7564</td>
<td>0.8133</td>
<td>0.852</td>
</tr>
<tr>
<td>Huntington’s</td>
<td>0.76</td>
<td>0.804</td>
<td>0.7728</td>
<td>0.6248</td>
<td>0.816</td>
<td>0.6824</td>
<td>0.6008</td>
<td>0.7728</td>
</tr>
<tr>
<td>CNS</td>
<td>0.7188</td>
<td>0.7222</td>
<td>0.7274</td>
<td>0.6540</td>
<td>0.7540</td>
<td>0.6044</td>
<td>0.7103</td>
<td>0.7418</td>
</tr>
<tr>
<td>Myeloma</td>
<td>0.7009</td>
<td>0.7228</td>
<td>0.7349</td>
<td>0.6755</td>
<td>0.7472</td>
<td>0.7053</td>
<td>0.7865</td>
<td>0.7689</td>
</tr>
<tr>
<td>Average</td>
<td>0.8278</td>
<td>0.8438</td>
<td>0.8525</td>
<td>0.8068</td>
<td>0.8765</td>
<td>0.7708</td>
<td>0.7997</td>
<td>0.8603</td>
</tr>
</tbody>
</table>

Table 2. Accuracy rates obtained in the numerical experiment.


[8] F Borovecki, L Lovrecic, Jessica Zhou, Hyun Jeong,
Florian Then, HD Rosas, SM Hersch, P Hogarth, Beren-gere Bouzou, RV Jensen, et al., “Genome-wide ex-
pression profiling of human blood reveals biomarkers
for huntington’s disease,” Proceedings of the National
Academy of Sciences of the United States of America,
vol. 102, no. 31, pp. 11023–11028, 2005.
[9] Donald Geman, Christian d’Avignon, Daniel Q Naiman,
and Raimond L Winslow, “Classifying gene expression
profiles from pairwise mrna comparisons,” Statistical
applications in genetics and molecular biology, vol. 3,
[10] Aik Choon Tan, Daniel Q Naiman, Lei Xu, Raimond L
Winslow, and Donald Geman, “Simple decision rules
for classifying human cancers from gene expression
profiles,” Bioinformatics, vol. 21, no. 20, pp. 3896–
3904, 2005.
Cope, Giovanni Parmigiani, Daniel Naiman, and Don-
ald Geman, “The ordering of expression among a few
genes can provide simple cancer biomarkers and signal
brca1 mutations,” BMC bioinformatics, vol. 10, no. 1,
[12] Bahman Afsari, Ulisses M Braga-Neto, Donald Geman,
et al., “Rank discriminants for predicting phenotypes
from rna expression,” The Annals of Applied Statistics,
of incomplete block designs: I. the method of paired
comparisons,” Biometrika, vol. 39, no. 3/4, pp. 324–
345, 1952.
[14] Francois Caron and Arnaud Doucet, “Efficient bayesian
inference for generalized bradley–terry models,” Journal
of Computational and Graphical Statistics, vol. 21,
[15] Peter McCullagh and John A Nelder, Generalized linear
[16] David R Hunter, “Mm algorithms for generalized
[17] Isobel Claire Gormley and Thomas Brendan Murphy,
“Exploring voting blocs within the irish electorate: A
mixture modeling approach,” Journal of the Ameri-
can Statistical Association, vol. 103, no. 483, pp. 1014–
1027, 2008.
for plackett-luce ranking models,” in proceedings of the
26th annual international conference on machine learn-
[19] Haiyan Wang, Hongyan Zhang, Zhijun Dai, Ming-shun
Chen, and Zheming Yuan, “Tsg: a new algorithm for
6, no. 1, pp. 1, 2013.
[20] Uri Alon, Naama Barkai, Daniel A Notterman, Kurt
Gish, Suzanne Ybarra, Daniel Mack, and Arnold J
Levine, “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tis-
uues probed by oligonucleotide arrays,” Proceedings of
the National Academy of Sciences, vol. 96, no. 12, pp.
[21] Seraya Maouche, Odette Poirier, Tiphaine Godefroy,
Robert Olaso, Ivo Gut, Jean-Phillipe Collet, Gilles Mont-
talescot, and François Cambien, “Performance compar-
ison of two microarray platforms to assess differential
gene expression in human monocyte and macrophage
[22] Scott L Pomeroy, Pablo Tamayo, Michelle Gaasenbeek,
Lisa M Sturla, Michael Angelo, Margaret E McLaugh-
lin, John YH Kim, Liliana C Goumnerova, Peter M
Black, Ching Lau, et al., “Prediction of central nerv-
ous system embryonal tumour outcome based on gene
2002.
[23] Ermeng Tian, Fenghuang Zhan, Ronald Walker, Erik
Rasmussen, Yupo Ma, Bart Barlogie, and John D Shaughnessy Jr, “The role of the wnt-signaling antag-
onist dkk1 in the development of osteolytic lesions in
multiple myeloma,” New England Journal of Medicine,
[24] Bahman Afsari, Elana J Fertig, Donald Geman, and
Luigi Marchionni, “switchbox: an r package for k–top
scoring pairs classifier development,” Bioinformatics,