ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Read Classification for Next Generation Sequencing

James M. Hogan[1], Peter Holland[1], Alexander P. Holloway[1],
Robert A. Petit III[2], Timothy D. Read[2]

1- School of EECS - Faculty of Science and Engineering, QUT
GPO Box 2434, Brisbane, QLD, 4001 - Australia

2- Department of Human Genetics - Emory University School of Medicine
Whitehead Biomedical Research Building, 615 Michael Street, Suite 301
Atlanta GA 30322 - United States of America

Abstract. *Next Generation Sequencing (NGS)* has revolutionised molecular biology, allowing routine clinical sequencing. NGS data consists of short sequence reads, given context through downstream assembly and annotation, a process requiring reads consistent with the assumed species or species group. The common bacterium *Staphylococcus aureus* may cause severe and life-threatening infections in humans, with some strains exhibiting antibiotic resistance. Here we apply an SVM classifier to the important problem of distinguishing *S. aureus* sequencing projects from other pathogens, including closely related *Staphylococci*. Using a sequence *k-mer* representation, we achieve precision and recall above 95%, implicating features with important functional associations.

1 Introduction

Next Generation Sequencing (NGS) [1] has revolutionised molecular biology, increasing efficiency to the point that clinical sequencing is now routine. This paper considers sequencing related to bacterial infection, although the ideas are more broadly applicable. Normally, a lab may isolate DNA from a bacterial colony, and sequence the genome before the species is known with certainty. Clinical signs and non-molecular diagnostics may offer some insight, but the suspicion needs to be confirmed if species-specific downstream informatics are to work successfully. Broadly, the task is to take the raw output of the sequencer – the project FASTQ file – and determine whether the *whole project* best resembles a sequencing project for a single species. In the alternative, the project might be corrupted, drawn from a mix of species, or reflect a previously unknown species or strain. Confounding factors include variable quality and length of data from different instruments and the inherent genetic variability of bacterial species, which include inversion of novel DNA in the form of plasmids – molecular DNA independent of the main chromosome – and prophages – genetic material which remains a part of bacterial DNA as a result of viral infection.

It is important to distinguish this problem from the related task of species identification – as practised in metagenomics [2]. Metagenomic studies rely on a common, highly conserved target, the *16S* region; in contrast, clinical sequencing is characterised only through the read distribution, and these are not selected

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

a priori, and indeed vary significantly across technologies. Our task is to find a representation which allows ready comparison across projects. In the following sections, we develop a simple approach based on a spectrum of short *k-mers* – sequence substrings of length k – which proves appropriate for the task.

The *Staphylococcus* genus includes at least forty species, of which several are harmless. However, some species, such as *S. aureus*, may cause serious, life threatening infections. Antibiotic resistance in *Staphylococcus* species has become widespread, purportedly as a result of *horizontal gene transfer* [6]. This has created strains which are particularly difficult to treat, notably methicillin-resistant *S. aureus (MRSA)* and vancomycin-resistant *S. aureus (VRSA)*, both of which have been implicated in serious outbreaks in major hospitals.

Both *Staph* and *non-Staph* projects are made available through the project FASTQ file, the *de facto* standard for NGS sequence reads. FASTQ combines a traditional FASTA sequence format with quality scores indicating confidence in the base prediction, scores whose utility is limited by inconsistency across vendors; we ignore them for this study. The file is then a collection of generated reads, held separately without additional structure. Each individual read is treated equally, and resulting k-mer counts are combined.

A general discussion of NGS technologies lies well beyond the scope of this paper, but the field was reviewed recently by [1]. Here we note only the principal characteristics of the machines. Of most interest is the variation in the read length. Homogeneity among large sets of sequencing project data cannot be assumed, and one of the challenges is to provide a unified representation. The majority of genomes examined in this research were sequenced using the Illumina and 454 technologies, with others based on PacBio and Ion Torrent machines. Broadly, these may be grouped according to the longer reads from Ion Torrent, Roche 454 and PacBio (200-400,330,964) and the far shorter reads of the Illumina platform (75-90). While we do not address this issue in the current study, an important consequence of this variation lies in the associated depth of coverage – the number of reads overlapping a particular base. For our purposes, the coverage provided with each of these technologies is adequate to ensure some reliable calculation of the relative counts. Working, as we do, with short k-mers avoids some pitfalls which might arise when working with the longest fragments directly, and offers the promise of lower computational overhead.

The remainder of this paper is organised as follows: In Section 2.1, we briefly introduce the classifier and the data representation, before considering the data set and our initial exploration and feature selection in Section 2.2. Details of the classifier performance are provided in Section 3 and we conclude in Section 4 with some discussion and future directions for this area of research.

2 Approach and Data Selection

2.1 Classifier and Representation

The Support Vector Machine is a linear classifier embodying the principles of Statistical Learning Theory [3] and supporting extensions to non-separable datasets

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

through the use of slack variables and penalty terms, and extensions to higher dimensions and subtle similarity relationships – including those defined over structures – through the use of Mercer kernels [4]. As is well-known, the decision surface is chosen so as to maximise the margin of separation between examples of the positive class – here $STAPH\_AUREUS$ – and examples of the negative class – NOT_STAPH_AUREUS. The formulation employed is the more traditional C-SVM. The SVM is here a binary classifier, predicting the label $y_i \in \{-1, +1\}$ for a given example pattern x_i, based on training over the labelled set $\{(x_i, y_i)|i = 1 \dots l\}$. Ultimately, the classifier is determined by solving the following dual problem:

$$\max_{\{\alpha_i\}} Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \tag{1}$$

subject to the constraints:

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \tag{2}$$

and

$$0 \leq \alpha_i \leq C, i = 1 \dots l. \tag{3}$$

Here the α_is are the Lagrange multipliers associated with the constraints, $K(\cdot, \cdot)$ is the kernel similarity function, and C is the coefficient of the primal constraint penalty term. The final form of the function is then:

$$f(x) = \sum_{i \in S} \alpha_i y_i K(x_i, x) + b, \tag{4}$$

where S is the index set of support vectors, and b is the bias term.

Our use of a k-mer spectrum follows the approach of Leslie et. al. [5], though it was not necessary to consider non-zero values of the mismatch parameter. Counts are accumulated via a sliding window over each sequence read, the results being summed and subsequently normalised. Thus, the representation presented to the model initially is that of a vector of counts associated with each project, with the dimension determined in principle as 4^k, where k is the chosen k-mer length. In the following sections, we consider the effective dimension of these datasets after applying feature selection.

2.2 Data Set and EDA

Initial explorations were based on 20 projects from the Read Lab, balanced evenly between *S.aureus* and other genomes. Other data included 5 *E. coli*, and one each of *Pseudomonas aueriginosa, Clostridium botulinum, Bordetella bronchideptica, Myobacterium colombiense* and *Neisseiria meningitidis*, a set chosen for diversity. The 6-mer count Gram matrix showed abundant evidence differentiating in-class from out-class projects. SVM trials delivered perfect classification, with little disruption from an additional ten out-class examples.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

For more serious runs, a set was assembled of the raw FASTQ project files for a total of 60 *S. aureus* genomes and a total of 70 additional genomes of other kinds. Critically, this additional set included some 15 *Staphylococcus* genomes other than *S. aureus*. These data were received in raw SRA form from NCBI, and converted to FASTA files using NCBI utilities. Tokenization of the sequences was performed for $k = 6, 8$ and 10, thereby yielding comprehensive feature sets at each level. Feature selection approaches were investigated for $k = 8$ and 10, and after a number of trials, feature sets were determined using Relief [7] as supported by Weka. Relief determines an attribute relevance score based on the difference from the feature values of its nearest neighbours, and is known to be fast even in the presence of a large number of features.

Reduced datasets were determined for $k = 8$ and 10, with the full set employed for $= 6$ due to the relatively limited scale of the vector. Based on the distribution of the rankings returned from Relief, the feature set dimension for $k = 8$ was reduced to around 5000, with $10,000$ selected for $k = 10$. While the selection at the $k = 8$ level was straightforward from inspection of the scores, a good deal of experimentation was undertaken for the 10-mer case, with classification runs for widely varying feature sets, noting both accuracy and the quality of the model, as reflected in the number of support vectors relative to the training set size. The selection of 10,000 features provided an optimal trade-off between classification cost and accuracy. Smaller feature sets resulted in significant increases in both false positive and false negative classifications, while larger feature sets presented unacceptable computational cost for commodity hardware. A full set of the organisms and top scoring k-mers used in the study is provided in the supplementary material at `http://eprints.qut.edu.au/57694/`.

3 Results

SVMs employed in this study were trained rapidly on commodity hardware using the R environment's e1071 package. A range of kernels and parameter settings were explored, but were found to offer no additional benefit over the base spectral representation and a linear, dot-product combination of the feature vectors. As noted in the previous section, the dimension of the pattern space was reduced through the use of Relief, allowing far more rapid training and testing of the models. These feature sets were employed throughout the training and testing of the models. The full set of 4096 features were employed for $k = 6$, with no zero counts encountered. Reduced sets were employed for $k = 8 : 65536 \rightarrow 4975$ and $k = 10 : 1048576 \rightarrow 10000$.

The full data set of 130 sequencing projects was employed in the study, partitioned into training and holdout sets, with the split varying as described below. Training results shown are the mean and standard deviation from classification of the test fold when 10-fold cross validation is employed on this training set. Direct training results over the 9 folds are not reported. The columns at right show the results of applying the best model obtained during the training process to the hold out set for the current split. Initial exploration was performed with a

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

model development set of 30 count vectors, with a hold-out set of the remaining 100. A wide variety of values were considered for C, with 2^-8 proving optimal and being used for subsequent runs. The tables below show increasing training set sizes in successive increments of 20[1], for $k = 6, 8$ and 10.

Training Set Size	Hold-out Set Size	Mean Accuracy	Std Deviation	Hold-out Accuracy	Precision	Recall
30	100	0.90	0.10	0.79	0.72	0.96
50	80	0.92	0.10	0.80	0.70	0.97
70	60	0.87	0.16	0.75	0.63	1.0
90	40	0.89	0.11	0.78	0.63	1.0

Table 1: Result summary for $k = 6$.

Training Set Size	Hold-out Set Size	Mean Accuracy	Std Deviation	Hold-out Accuracy	Precision	Recall
30	100	0.93	0.06	0.79	0.71	0.98
50	80	0.98	0.06	0.79	0.69	0.94
70	60	0.94	0.07	0.82	0.71	0.96
90	40	0.91	0.12	0.78	0.63	1.0

Table 2: Result summary for $k = 8$.

Training Set Size	Hold-out Set Size	Mean Accuracy	Std Deviation	Hold-out Accuracy	Precision	Recall
40	100	0.96	0.06	0.86	0.76	0.96
50	80	1.0	0.00	0.81	0.71	0.97
70	60	0.96	0.07	0.97	0.96	0.96
90	40	0.97	0.08	0.95	0.94	0.94

Table 3: Result summary for $k = 10$.

4 Conclusions

In this work we have demonstrated successful SVM classification of clinical sequencing projects based on a k-mer spectrum representation. The results offer strong evidence that larger values of k, coupled with a training set above some critical size to ensure broad coverage of the competing genomes, are necessary to ensure good performance for this problem. Certainly the results of Table 3 bear this out, with superior performance to those for $k = 6$ and 8, and a compelling

[1]Note that a training set size of 40 was the smallest used for the 10-mer study, the results at size 30 being poor.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

transition in performance for a training set size of 70 and above. These observations are confirmed when one examines the count of support vectors relative to set size, a ratio here of around 0.2 compared to almost 0.5 for the weaker runs for $k = 6, 8$. In particular we note the success in achieving not merely high recall (0.96) but equally high precision, a critical factor in this problem and one making the earlier runs unusable in practice. Additional work is being undertaken to examine higher values of k, but work from other studies suggests a trade-off between k-mer length and precision in problems of this nature. This was true of our earlier work in promoter prediction [8], and in work on similarity based on D2 and other word based distances [9]. Higher values of k, needless to say, may also come with a substantial computational cost.

Examination of the top-scoring k-mers is revealing from a biological standpoint, revealing a sharply higher GC nucleotide content (55%) than the average for the genome (34%), possibly representing selection against the prevailing $GC \rightarrow AT$ mutation bias in bacterial genomes. There were more than 1645 matches of the 10-mer features to the *S. aureus* N315 genome, 918 mapping to annotated features (mostly genes). More than 300 10-mer matches were at one locus – the sdrCDE operon, which encodes a fibrinogen binding complex – fibrinogen being a protein which assists blood clotting – and is thus an interesting diagnostic target. Additional studies of this nature may elucidate a set of such characteristic sequence fragments, of a size useful for more general analysis. Our success in distinguishing even between *S. aureus* and other *Staphylococci* suggests that fine grained distinctions are possible, and augur well for our ongoing studies over a much larger set of genomes.

References

[1] M.L. Metzker, Sequencing technologies – the next generation, *Nature Rev. Genetics*, 11:31-46, (2010).

[2] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, P. Hugenholtz, A Bioinformatician's Guide to Metagenomics, *Microbiol Mol Biol Rev*, 72(4):557-578 (2008).

[3] V. Vapnik, *The Nature of Statistical Learning Theory, 2nd Edition*, Springer Verlag, Heidelberg, 1999.

[4] C.J. Burges, A Tutorial on Support Vector Machines, *Data Mining and Knowledge Discovery*, 2:121-167, (1998).

[5] C. Leslie, E. Eskin, W.S. Noble, The spectrum kernel: a string kernel for SVM protein classification, In: *Proceedings of the Pacific Symposium on Biocomputing*, 564-575, (2002).

[6] Holden, M. T. G., Feil, E. J., ... Parkhill, J., Complete genomes of two clinical Staphylococcus aureus strains: Evidence for the rapid evolution of virulence and drug resistance, *Proceedings of the National Academy of Sciences*, 101:9786-9791, (2004).

[7] K. Kira and L. Rendell, A practical approach to feature selection, In *Proceedings of the ninth international workshop on Machine Learning*, 249-256, Morgan Kauffman, (1992).

[8] Gordon JJ, Towsey MW, Hogan JM, Mathews SA, Timms P., Improved prediction of bacterial transcription start sites, *Bioinformatics*, 22(2):142-148 (2006).

[9] S. Forêt, M.R. Kantorovitz and C.J. Burden, Asymptotic behaviour and optimal word size for exact and approximate word matches between random sequences, *BMC Bioinformatics*, 7(Suppl 5):S21, (2006).