Aberystwyth University

Computer Science and Statistics (GG34)

CS396: Minor Project

Outline Project Specification

Application of Machine Learning Techniques to Next Generation Sequencing Quality Control

Author:
Sam Nicholls msn

Supervisor:
Dr. Amanda Clare afc

Version 1.0

February 7, 2014

0.1 Introduction and Background

Over the past few years advances in genetic sequencing hardware have introduced the concept of massively parallel DNA sequencing; allowing potentially billions of chemical reactions to occur simultaneously, reducing both time and cost required to perform genetic analysis[8]. However, these "next-generation" processes are complex and open to error[2], thus quality control is an essential step to assure confidence in any downstream analyses performed.

0.2 Description

During sample sequencing a large number of quality control metrics are generated to determine the quality of the reads from the sequencing hardware itself. At the Wellcome Trust Sanger Institute, the automated QC system currently relies on hard thresholds to make such quality control decisions with individual hard-coded values on particular metrics determining whether a lane has reached a level that requires a warning, or has exceeded the threshold and failed entirely. Whilst this does catch most of the very poor quality lanes, a large number of lanes are flagged for manual inspection at the warning level; a time consuming task which invites inefficiency and error.

In practise most of these manual decisions are based on inspecting a range of diagnostic plots which suggests that a machine learning classifier could potentially be trained on the combinations of quality control statistics available to make these conclusions without the need for much human intervention.

0.3 Proposed Tasks

0.3.1 Research

Due to my unfamiliarity with the problem space and the experimental nature of the end-goal; the project invites various elements of research. I foresee needing to investigate the following:

Problem Domain
: Understand terminology related to the problem and environment; what terms are used by QC team members to describe possible reasons for failure? What might cause these failure modes, how badly does this affect the readings from the sample?

Quality Control Procedures
: Detail what QC procedures are currently in place; what are the hard thresholds used to define a failure or a warning? What manual methods are applied to samples that cannot be automatically processed?

Output and Statistics
: Decipher the output files from the sequencing and current automatic QC process; what lines are relevant to quality and what can be safely discarded and ignored? What could turn out to be a useful indicator of quality? What statistics are available at QC level and how are they used by QC staff?

Information Theory
: Investigate potential applications of information theory to QC metrics.

Collect Data Sets

Collect and catalogue available data for training. Investigate relationships between attributes and possible sources of noise. Construct or locate useful tools for manipulation and management of these data sets.

0.3.2 Development

Define Problem

What exactly will the classifier attempt to learn? How will the problem be represented?

Machine Learning Classifiers

Evaluate algorithm options and select an appropriate learning strategy to implement for the final classifier.

0.3.3 Testing

Test Suite

Develop an easy to configure and deploy test suite that measures and stores the performance of the learning algorithm at the current time. It will be critical to be able to identify if changes impact the performance of the classifier and measure performance gain over time.

Continuous Integration

Cloud solutions such as Wercker or Travis may provide a simple to deploy and use system. Standalone systems like Jenkins could offer more control if required.

0.3.4 Deployment

It may be necessary to provide additional configuration options, interfaces or minor changes once the software has been deployed at the institute as part of their QC pipeline.

0.4 Deliverables

Outline Project Specification, Start of February An initial outline of the project, tasks and expected deliverables

Progress Reports Provide weekly status reports in the form of a diary or blog

Project Specification, Mid February Document the problem in greater detail, providing answers to the research questions posed previously. Consider learning goal and possibly strategies, outline the implementation

Data Sets, Mid February Collate the data sets to be used for training and validation

Initial Classifier, End of February Commit the initial classifier

Mid-Project Demo, March Perform a demonstration of the classifier's current capabilities

Test Report Complete analysis on the classifier's ability to generalise

Final Submission, May Submit the classifier itself along with the final report and documentation

0.5 Annotated Bibliography

[1] B.-S. Kang and S.-C. Park, "Integrated machine learning approaches for complementing statistical process control procedures," *Decision Support Systems*, vol. 29, no. 1, pp. 59–72, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923600000634

Summary of machine learning techniques and brief examples of their applicability to statistical process control.

[2] M. Kircher, U. Stenzel and J. Kelso, "Improved base calling for the Illumina Genome Analyzer using machine learning strategies," *Genome Biology*, vol. 10, no. 8, p. R83, 2009.

Useful introduction to relevant Illumina hardware and the errors that can occur during sequencing.

[3] D. J. MacKay, *Information Theory, Inference and Learning Algorithms*, 9th ed. Cambridge University Press Cambridge, 2010.

Discussing the possible implementation options with my supervisor, we are interested in investigating whether information theory topics such as compression and gain could provide solutions to the problem of deciding whether removing samples will affect variant calling downstream.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

A machine learning framework for Python.

[5] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2010. [Online]. Available: http://books.google.co.uk/books?id=8jZBksh-bUMC 9780136042594.

Recommended reading for last semester's Machine Learning module, providing an in-depth explanation of various machine learning algorithms that are available.

[6] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature Biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.

In-depth review of next-generation sequencing hardware, including the history of sequencing technology, a comparison of current hardware and their advantages and disadvantages, and an outline of available tools for post-processing.

[7] M. Shewhart, "Interpreting statistical process control (SPC) charts using machine learning and expert system techniques," in *Aerospace and Electronics Conference, 1992. NAECON 1992., Proceedings of the IEEE 1992 National*, vol. 3, May 1992, pp. 1001–1006.

Relevant as a case study of the application of machine learning to statistical quality control.

[8] T. Strachan and A. Read, *Human Molecular Genetics*, 4th ed. Garland Science, 2011, pp. 214–254.

A concise introduction to the processes involved in massively parallel DNA sequencing.