Using expected sequence features to improve basecalling accuracy of amplicon pyrosequencing data

Amplicon pyrosequencing targets a known genetic region and thus inherently produces reads highly anticipated to have certain features, such as conserved nucleotide sequence, and in the case of protein coding DNA, an open reading frame. Pyrosequencing errors, consisting mainly of nucleotide insertions and deletions, are on the other hand likely to disrupt open reading frames. Such an inverse relationship between errors and expectation based on prior knowledge can be used advantageously to guide the process known as basecalling, i.e. the inference of nucleotide sequence from raw sequencing data. The new basecalling method described here, named Multipass, implements a probabilistic framework for working with the raw flowgrams obtained by pyrosequencing. For each sequence variant Multipass calculates the likelihood and nucleotide sequence of several most likely sequences given the flowgram data. This probabilistic approach enables integration of basecalling into a larger model where other parameters can be incorporated, such as the likelihood for observing a full-length open reading frame at the targeted region. We apply the method to 454 amplicon pyrosequencing data obtained from a malaria virulence gene family, where Multipass generates 20 % more error-free sequences than current state of the art methods, and provides sequence characteristics that allow generation of a set of high confidence error-free sequences. This novel method can be used to increase accuracy of existing and future amplicon sequencing data, particularly where extensive prior knowledge is available about the obtained sequences, for example in analysis of the immunoglobulin VDJ region where Multipass can be combined with a model for the known recombining germline genes. Multipass is available for Roche 454 data at http://www.cbs.dtu.dk/services/MultiPass-1.0 , and the concept can potentially be implemented for other sequencing technologies as well.

General information
State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis, Metagenomics, Molecular Evolution, New York University
Authors: Rask, T. S. (Intern), Petersen, B. (Intern), Chen, D. S. (Ekstern), Day, K. P. (Ekstern), Pedersen, A. G. (Intern)
Number of pages: 10
Publication date: 2016
Main Research Area: Technical/natural sciences

Publication information
Journal: B M C Bioinformatics
Volume: 17
Issue number: 1
Article number: 176
ISSN (Print): 1471-2105
Ratings:
BFI (2017): BFI-level 1
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 1
Scopus rating (2016): SJR 1.467 SNIP 0.946 CiteScore 2.54
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 1.656 SNIP 1.077 CiteScore 2.77
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 1.836 SNIP 1.202 CiteScore 2.91
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 1.932 SNIP 1.335 CiteScore 3.38
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 1.857 SNIP 1.155 CiteScore 3.24
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 1.655 SNIP 1.215 CiteScore 3.34
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
Using expected sequence features to improve basecalling accuracy of amplicon pyrosequencing data.pdf

Bibliographical note
© 2016 Rask et al. Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Source: FindIt
Source-ID: 2303737824
Publication: Research - peer-review › Journal article – Annual report year: 2016