C-5-22

Probabilistic Model Based Error Correction of Various Mutant Sequences Analyzed by the Single-Molecule Real-Time Sequencing

Takuyo Aita[1]
aita-takuyo@bio.eng.osaka-u.ac.jp

Norikazu Ichihashi1[1,2]
ichihashi-norikazu@bio.eng.osaka-u.ac.jp

Tetsuya Yomo[1,2,3]
yomo@ist.osaka-u.ac.jp

[1] Exploratory Research for Advanced Technology, Japan Science and Technology Agency, Yamadaoka 1-5, Suita, Osaka, Japan
[2] Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Yamadaoka 1-5, Suita, Osaka, Japan
[3] Graduate School of Frontier Biosciences, Osaka University, Yamadaoka 1-5, Suita, Osaka, Japan

Keywords: base call error; next generation sequencing; image restoration; quality score; quasispecies; sequence analysis; SMRT

To analyze the evolutionary dynamics of a mutant population in an evolutionary experiment, it is necessary to sequence a vast number of mutants by high-throughput sequencing technologies. Particularly, we focus on the single-molecule real-time (SMRT) sequencing technology [1], which enables rapid and parallel analysis of multikilobase sequences. However, the observed sequences as "circular consensus sequences (CCS)" obtained by the SMRT sequencing include many random errors of base call. Therefore, if the SMRT sequencing is applied to analysis of a heterogeneous population of various mutant sequences, it is necessary to discriminate between true bases as point mutations and random errors of base call in the observed sequences, and to subject the sequences to error-correction processes [2]. To address this issue, we have developed a novel method of error correction based on the Bayesian theory with the Potts model and a maximum a posteriori probability (MAP) estimation. The available information for error correction is (1) "quality scores" which are assigned to individual bases in the observed sequences [3] and (2) a spatial distribution of the observed sequences in sequence space [4]. The computer experiments of error correction of artificially generated sequences supported the effectiveness of our method, showing that 50-90 % of errors were removed. Interestingly, this method is analogous to a probabilistic model based method of image restoration developed in the field of information engineering [5].

[1] Eid J et al., Real-time DNA sequencing from single polymerase molecules. *Science*,**323**, 133-138, 2009.
[2] Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N., Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J Comput Biol*.,**17**,417-428, 2010.
[3] Cock P.J., Fields C.J., Goto N, Heuer M.L. and Rice P.M., The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, **38**, 1767–1771, 2010.
[4] Eigen M, Winkler-Oswatitsch R., Statistical geometry on sequence space. *Methods Enzymol.*, **183**, 505-530, 1990.
[5] Tanaka K. and Morita T., Cluster variation method and image restoration problem. *Physics Letters* A, **203**, 122-128, 1995.