Computational Biology and Chemistry 47 (2013) 221–230

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Probabilistic model based error correction in a set of various mutant sequences analyzed by next-generation sequencing

Takuyo Aita[a], Norikazu Ichihashi[a,b], Tetsuya Yomo[a,b,c,*]

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

ARTICLE INFO

Article history:
Received 30 July 2013
Received in revised form
13 September 2013
Accepted 27 September 2013

Keywords:
Base call error
Image restoration
Quality score
Quasispecies
Sequence analysis
SMRT

ABSTRACT

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 (next-generation) sequencing technologies, which enable rapid and parallel analysis of multikilobase sequences. However, the observed sequences include many errors of base call. Therefore, if next-generation 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 errors of base call in the observed sequences, and to subject the sequences to error-correction processes. To address this issue, we have developed a novel method of error correction based on the Potts model and a maximum a posteriori probability (MAP) estimate of its parameters corresponding to the "true sequences". Our method of error correction utilizes (1) the "quality scores" which are assigned to individual bases in the observed sequences and (2) the neighborhood relationship among the observed sequences mapped in sequence space. 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.

1. Introduction

Many kinds of evolutionary experiments, by using bacteria (Barrick et al., 2009), viruses (Meyer et al., 2012) or self-replicating molecular systems (Kita et al., 2008; Ichihashi et al., 2013), have been conducted over the world, and then it becomes more and more necessary to analyze an evolutionary dynamics and statistical properties of fitness landscapes by sequencing a vast number of mutants in the evolving population (Betancourt, 2009; Pitt, and Ferr-D'Amar, 2010; Steinbruck et al., 2011; Otwinowski et al., 2013). This can be realized by high-throughput deoxyribonucleic acid (DNA) sequencing (called the "next-generation" sequencing) technologies (Eid et al., 2009; Quail et al., 2012; Liu et al., 2012), which enable rapid and parallel analysis of multikilobase sequences. However, the observed sequences include many errors of base call. Therefore, if next-generation 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 errors of base call in the observed sequences, and to subject the sequences to error-correction processes.

Many studies of error correction methods for next-generation sequencing have been reported (e.g. Wijaya et al., 2009; Zhao et al., 2010; Ilie et al., 2011; Yang et al., 2011). Several groups reported error correction methods of next-generation sequencing data to estimate the genetic diversity in quasispecies of viruses, that is, the occurrence-frequency distribution of mutant sequences in quasispecies (Zagordi et al., 2010a, 2011; Prosperi et al., 2011; Prosperi and Salemi, 2012; Astrovskaya et al., 2011; Skums et al., 2012). For example, in Zagordi et al. (2010b), each of the sequenced fragments (called the "reads"), which contain point mutations and base-call errors, is classified into several clusters by using a model-based probabilistic clustering algorithm. This clustering algorithm is based on a Bayesian statistics using the "Dirichlet process mixture (DPM)". Then, the consensus sequence of each cluster represents the true (original) sequence.

We have developed a novel method of error correction based on the Potts model and a maximum a posteriori probability (MAP) estimate of its parameters corresponding to the "true sequences". Our method of error correction utilizes (1) the "quality scores" which are assigned to individual bases in the observed sequences and (2) the neighborhood relationship among the observed sequences mapped in sequence space. The quality scores are encoded by

* Corresponding author at: Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Yamadaoka 1-5, Suita, Osaka, Japan. Tel.: +81 668794151.
E-mail address: yomo@ist.osaka-u.ac.jp (T. Yomo).

1476-9271/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2013.09.006

Ascii code in FASTQ files (Cock et al., 2010). Based on the concept of sequence space (Eigen and Winkler-Oswatitsch, 1990), several mutant sequences which are similar to an arbitrary sequence X are located near the sequence X in sequence space. These mutant sequences are called the "neighborhoods" of the sequence X. Error correction of the sequence X is conducted by referring to the sequence information in the neighborhoods. For example, let us consider that a letter at the ith position on the sequence X is "A", while letters at the same position on other sequences in the neighborhoods are "T". If the quality score assigned to "A" takes a high value and those assigned to "T" take low values on average, the "A" is likely to be the true letter. If the quality score assigned to "A" takes a low value and those assigned to "T" take high values on average, the "A" is likely to be an error and should be replaced with "T". The purpose of this study is to develop a method to implement these processes automatically in a rational way. A marked difference between our method and other ones resides in that our method does not perform the clustering of reads based on their similarity, but utilizes the quality scores which are transformed to the error probabilities of base call.

Our method of error correction is analogous to a probabilistic-model-based method of image restoration developed in the field of information engineering (Bilbro et al., 1992; Pryce and Bruce, 1995). Therefore, we can obtain several useful ideas from this field. To evaluate the performance of our method, we carried out computer experiments of error correction by generating 1024 artificial DNA sequences in a simulated evolving population. The results of the computer experiments supported the effectiveness of our method. This paper reports a novel approach toward a systematic error correction in a set of various mutant sequences through next-generation sequencing.

2. Method

2.1. A mathematical model of DNA sequencing for an evolving mutant population

We consider an evolving population of various mutant DNA sequences in an evolutionary experiment. We assume that mutational events in evolution are only base substitutions, and then the sequence lengths for all the mutants are fixed with a constant ν. Let i be the position number along each sequence ($i = 1, 2, \ldots, \nu$), and let $\lambda \equiv 4$ be the number of available letters (=bases). Arbitrary M ($= 10^3$–10^4) samples are picked from among the mutant population randomly, and each of them is referred with the serial number m or n ($m, n = 1, 2, \ldots, M$). The sequence information of each sample is provided by next-generation sequencing. For mathematical formulation, three types of "sequences" are conceptually or really assigned to each of the M samples (Fig. 1):

True sequence: the true sequence of a sample m is represented by $x_1(m)x_2(m)\ldots x_\nu(m)$, where $x_i(m)$ is the "true letter" at the ith position for the sample m. (We never know the true sequence.)
Observed sequence: this is obtained by analyzing the true sequence through next-generation sequencing (each observed sequence is called the "read"). The observed sequence of a sample m is represented by $y_1(m)y_2(m)\ldots y_\nu(m)$, where $y_i(m)$ is the "observed letter" at the ith position for the sample m.
Predicted sequence: this is obtained by subjecting the observed sequence to error-correction processes. The predicted sequence is the most probable candidate of the true sequence. The predicted sequence of a sample m is represented by $\widehat{x}_1(m)\widehat{x}_2(m)\ldots\widehat{x}_\nu(m)$, where $\widehat{x}_i(m)$ is the "predicted letter" at the ith position for the sample m.

These three types take the same sequence length ν, according to the below assumption.

Mathematically, the sequence information of the true sequences is transformed to the observed sequences with many (apparent) substitutions caused by base-call errors. We assume that, base-call errors do not cause (apparent) insertions and deletions for simplicity. [1] A "quality score" $Q_i(m)$ is assigned to the observed letter $y_i(m)$ for every position i and every sample m. The quality score $Q_i(m)$ is transformed to the error probability $p_i(m)$ that a base call at the ith position for a sample m results in failure. We assume that $p_i(m)$ is given by an unique function of $Q_i(m)$:

$$p_i(m) = f(Q_i(m)) \qquad \text{for all } m \text{ and } i. \tag{1}$$

Namely, a base call results in

$$y_i(m) \begin{cases} = x_i(m), & \text{with probability of } 1 - p_i(m) \\ \neq x_i(m), & \text{with probability of } p_i(m). \end{cases}$$

To determine the predicted sequences for the M samples through error correction, the available information is (1) a set of the observed sequences for the M samples and (2) a set of the quality scores for all letters in the observed sequences: $\{y_i(m), Q_i(m) | i = 1, 2, \ldots, \nu; m = 1, 2, \ldots M\}$.

2.2. Definition of the neighborhoods of each sample in sequence space

Our method of error correction is based on the spatial distribution of a mutant population in sequence space (Fig. 1b). We denote a pair of samples m and n by (m, n). Let $d_y(m, n)$ be the Hamming distance between the observed sequences for the samples m and n:

$$d_y(m, n) \equiv \nu - \sum_{i=1}^{\nu} \delta(y_i(m), y_i(n)), \tag{2}$$

where $\delta(*, *)$ represents the Kronecker's delta. Here, we define the "neighborhoods" of each sample in sequence space, based on the frequency distribution of the Hamming distance $d_y(m, l)$ from a certain sample m to arbitrary samples l (Fig. 1c). Let $\mathrm{E}[d]_m$ and $\mathrm{E}[d^2]_m$ be the following averages of $d_y(m, l)^b$:

$$\mathrm{E}[d^b]_m = \frac{\sum_{l \neq m} d_y(m, l)^b}{M - 1} \qquad (b = 1, 2), \tag{3}$$

where $\sum_{l \neq m}$ means the sum over all the samples except the sample m. The standard deviation of $d_y(m, l)$ is given by

$$\mathrm{SD}[d]_m \equiv \sqrt{\mathrm{E}[d^2]_m - \mathrm{E}[d]_m{}^2}. \tag{4}$$

Then, we introduce a critical Hamming distance for a sample m as follows:

$$\widehat{d}_m \equiv \mathrm{E}[d]_m - 3 \times \mathrm{SD}[d]_m. \tag{5}$$

We define that a sample n (or m) is the "neighborhood" of the sample m (or n), if

$$d_y(m, n) \leq \max\{\widehat{d}_m, \widehat{d}_n\} \tag{6}$$

(Fig. 1c). A set of all the neighborhoods of the sample m, denoted by

$$S_{\mathrm{nei}}(m) \equiv \{n | n \text{ is a neighborhood of } m\}, \tag{7}$$

is utilized in error correction of the observed sequences (Fig. 1b).

[1] This assumption does not hold in real cases. Our method can be applied to DNA sequences in a protein coding region, because insertions and deletions in these sequences cause lethal proteins through the frame shift in translation. This is the limitation of the current method. See Section 4.