JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 1, 2012

Pp. 1–12
DOI: 10.1089/cmb.2011.0070

Research Articles

Separating Significant Matches from Spurious Matches in DNA Sequences

HUGO DEVILLERS[1,2] and SOPHIE SCHBATH[1]

ABSTRACT

Word matches are widely used to compare genomic sequences. Complete genome alignment methods often rely on the use of matches as anchors for building their alignments, and various alignment-free approaches that characterize similarities between large sequences are based on word matches. Among matches that are retrieved from the comparison of two genomic sequences, a part of them may correspond to spurious matches (SMs), which are matches obtained by chance rather than by homologous relationships. The number of SMs depends on the minimal match length (ℓ) that has to be set in the algorithm used to retrieve them. Indeed, if ℓ is too small, a lot of matches are recovered but most of them are SMs. Conversely, if ℓ is too large, fewer matches are retrieved but many smaller significant matches are certainly ignored. To date, the choice of ℓ mostly depends on empirical threshold values rather than robust statistical methods. To overcome this problem, we propose a statistical approach based on the use of a mixture model of geometric distributions to characterize the distribution of the length of matches obtained from the comparison of two genomic sequences.

Key words: comparative genomics, match length, maximal exact matches, mixture model.

1. INTRODUCTION

For a long time, DNA sequence comparisons essentially relied on local or global alignments (Batzoglou, 2005). However, faced with the overwhelming number of completely sequenced genomes, the development of new methods able to investigate long sequences (several Mb) has exploded (Field et al., 2006; Treangen and Messeguer, 2006). Indeed, complete alignment of entire genomes is impossible with classical dynamic programming approaches such as the original Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). This is mainly due to computation time and space limitation (Miller, 2001; Ureta-Vidal et al., 2003) but also because most of the time such long sequences cannot be considered as collinear (Mantaci et al., 2008; Forêt et al., 2009).

There are two different approaches to compare long DNA sequences such as complete genomes. The first approach is the calculation of a complete alignment based on the principle of anchor alignment (Delcher et al., 1999; Devillers et al., 2010, 2011). Briefly, it is derived into two steps: (1) all the extremely

[1]INRA, UR1077, Mathématique, Informatique, et Génome, Jouy-en-Josas, France.
[2]LIRMM–CNRS, Université Montpellier 2, Montpellier Cedex 5, France.

conserved regions between the compared sequences are retrieved; and (2) these regions are sorted, and a part of them are selected to anchor sequences together. Anchored regions form the backbone that may correspond to the common ancestral sequence of the compared genomes; the rest are considered as specific regions (Chiapello et al., 2005). The second approach is more global. It consists in computing a score evaluating either a distance or the similarities between the compared sequences. These approaches are called ''alignment-free methods'' and are based on various theoretical foundations (Vinga and Almeida, 2003).

The critical step in the two above approaches is the detection of highly conserved regions between the compared genomes. There are two ways to retrieve these regions. The first one consists in performing local alignments, and the second one consists in identifying word matches (WMs). WMs correspond to exactly conserved sub-sequences shared by all the compared sequences. While local alignments have been intensively studied, few efforts have been provided regarding the use of WMs in comparative genomics.

Basically, there are two approaches to using WMs in comparative genomics. The first approach is to retrieve all the matches of a given length, the k-word matches (Forêt et al., 2009), where k is the length of the matches. k-WMs are essentially used in alignment-free methods such as for the D_2 statistic (Forêt et al., 2009). The second approach consists in retrieving the maximal exact matches (MEMs), which are WMs that cannot be extended either from their left nor from their right. MEMs are widely used as anchor points in complete genome alignment (Höhl et al., 2002) as well as in alignment-free methods (Deloger et al., 2009). It is noteworthy that, in comparative genomics, most of the studies that use WMs are based on the detection of MEMs rather than k-words. Indeed, it has been shown that k-word approaches often suffer from serious shortcomings to treat long genomic sequences (Reinert et al., 2009). However, a few recent studies have proposed k-word methods dedicated to the comparison of complete genomes (Guyon et al., 2009; Sims et al., 2009).

Computation of k-WMs and MEMs has been well studied (Lippert et al., 2005; Ohlebusch and Kurtz, 2008), and various efficient algorithms to retrieve them are available (Delcher et al., 2002; Khan et al., 2009). They only require one essential parameter that is the minimal length (ℓ) of the matches to retrieve. For the k-WM approaches, ℓ directly corresponds to k, and for the methods using MEMs, ℓ is the length of the shortest MEMs retrieved.

Selecting a suitable value for ℓ is a challenging task. Indeed, when retrieving WMs, two kinds of matches are found: (1) the significant matches that are related to the homologous relationship of the compared sequences; and (2) the spurious matches that are obtained only by chance. The shorter the retrieved WMs, the higher the number of spurious matches. Thus, the challenge is that ℓ has to be high enough to avoid spurious matches (Guyon and Guénoche, 2008) but not too high to discard significant matches. Figure 1 exemplifies this problem of providing the number and the type of the MEMs retrieved between two short sequences for different values of ℓ.

Until now, approaches aiming at choosing a reliable value for ℓ mostly relied on empirical analyses such as in Chiapello et al., (2005, 2008) and Wen et al., (2005). There are a few studies that propose statistical methods to solve this issue. For example, Reinert and Waterman (2006), inspired by the pioneering work of Karlin and Ost (1985), used a mixed Poisson approximation to analyze the length of the longest match in random sequences. They showed that, with their model, the probability for a spurious match to be longer than 22 nucleotides was very low. Guyon and Guénoche (2008) used a Poisson approximation to evaluate the expected number of MUMs (MEMs that are observed only once in the compared genomes) between

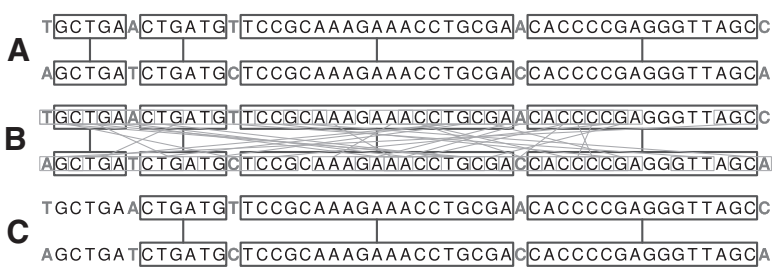

FIG. 1. Retrieving MEMs between two DNA sequences: impact of the choice of the minimal length ℓ. **(A)** The compared sequences have five point mutations (bold gray) yielding to four significant matches (black boxes). **(B)** Retrieved MEMs when $\ell = 3$; all the significant matches are found but with 28 spurious matches (gray boxes). **(C)** Retrieved MEMs when $\ell = 6$; no spurious match is retrieved, but a significant match is discarded.

random DNA sequences. They showed that the probability for a MUM to be longer than 22 nucleotides between two random sequences of about 2 Mb was almost null. By using a Bernoulli model and a Monte Carlo method, Lippert et al. (2005) showed that a threshold of 35 nucleotides was reliable to avoid spurious matches. However, most of these studies only investigated random sequences and did not test whether or not their thresholds were reliable on real DNA sequences. In addition, only a few of them consider the characteristics of the compared sequences such as their length, their nucleotide composition, or their relative divergence, although intuitively, one expects that the optimal value for ℓ should depend on these characteristics.

In this context, we developed a statistical approach based on the use of a mixture model of geometric distributions to characterize the distribution of the lengths of MEMs retrieved when comparing two genomes. In this article, after a brief presentation of the method, we show how our model can be used to find an optimal value for ℓ (ℓ_{opt}). The strengths and weaknesses of our method are then discussed through examples drawn from simulated and real genomic sequences.

2. METHOD

2.1. MEM lengths and geometric distribution

When comparing sequences, two kinds of MEMs are retrieved: the significant and the spurious MEMs. In this subsection, we show how it is possible to consider that the lengths of these two types of MEMs present a geometric distribution. Briefly, a geometric distribution can be defined as follows. Let's consider a sequential experiment in which independent Bernoulli trials with a probability of success p are repeated. We define the random variable X that corresponds to the number of repeats before the first success of a Bernoulli trial. The probability that $X = k$, i.e., k Bernoulli trials that fail followed by one success, is given by:

$$Pr(X = k) = (1 - p)^k p, \tag{1}$$

for $k \in \{0, 1, 2, \ldots\}$. These probabilities define the geometric distribution, and X has a geometric distribution with parameter p.

2.1.1. Significant MEMs. Significant MEMs correspond to the MEMs obtained from the homologous relationship of the compared sequences. If we consider two aligned homologous sequences, the length of a MEM is given by the number of identical nucleotides between two point-mutations (Fig. 1A). Let p_{mut} be the probability of a point-mutation. Here, we assume that point-mutations are independent and identically distributed (i.i.d.) in the sequences. We define the random variable L_{sign} as the length of significant MEMs; it has a geometric distribution with parameter p_{mut}. From equation (1), the probability that a significant MEM has a length k is given by:

$$Pr(L_{sign} = \ell) = (1 - p_{mut})^\ell p_{mut}. \tag{2}$$

Literally, a MEM of length ℓ corresponds to ℓ failures (probability: $(1 - p_{mut})^\ell$) followed by one success (probability: p_{mut}).

2.1.2. Spurious MEMs. Spurious MEMs randomly occur between the compared sequences. Let us consider random positions in the genomes. The length of a spurious MEM starting at these positions is given by the number of nucleotides that match before the first mismatch. Let p_{mis} be the probability of a mismatch between random positions in the compared sequences, and let the random variable L_{spur} be defined as the length of spurious MEMs; it has a geometric distribution with parameter p_{mis}. From equation (1), the probability that a spurious MEM has a length ℓ is given by:

$$Pr(L_{spur} = \ell) = (1 - p_{mis})^\ell p_{mis}. \tag{3}$$

2.2. Mixture model

To model the length L of the MEMs retrieved between two genomic sequences, we propose the following K-component geometric mixture distribution:

$$Pr(L = y) = \sum_{j=1}^{K} \pi_j f(y; p_j), \tag{4}$$

where

$$f(y; p_j) = (1 - p_j)^y p_j,$$

and $\Psi = (\pi_1, \cdots, \pi_{K-1}, p_1, \cdots, p_K)^T$, the vector containing all the unknown parameters in the mixture model. Here, π_j denotes the mixing proportion of the jth component of the model such that $\sum_{j=1}^{K} \pi_j = 1$ and p_j is the parameter of the jth component. The number of components is at least equal to two ($K = 2$), i.e., one component for the significant MEMs, equation (2), and one for the spurious MEMs, equation (3). In that case, $\Psi = (\pi_{\text{mut}}, p_{\text{mut}}, p_{\text{mis}})$. It is also possible to have $K > 2$, for example, if it is necessary to consider several different point-mutation probabilities simultaneously.

2.3. EM algorithm

An Expectation-Maximization (EM) algorithm (Dempster et al., 1977) was used to estimate the parameters Ψ of the mixture model. EM algorithms are iterative optimization methods that estimate unknown parameters in statistical models. A comprehensive and detailed presentation of EM algorithms is available in Mclachlan and Krishnan (1997). Briefly, let $y_1, \cdots, y_n$ be the observed data, i.e., the lengths of all the MEMs retrieved between two sequences. The EM algorithm we used consists in applying the following procedure:

1. **Initialization.** The parameter values $\Psi^{(0)} = (\pi_1^{(0)}, \cdots, \pi_{K-1}^{(0)}, p_1^{(0)}, \cdots, p_K^{(0)})^T$ are randomly initialized.
2. **Expectation Step.** The E-step, on the (t)th iteration, only requires the computation of the *a posteriori* probability $\gamma_{ij}^{(t)}$, which is the probability that the observation y_i belongs to the jth component of the mixture model. Here, $\gamma_{ij}^{(t)}$ is computed as follows:

$$\gamma_{ij}^{(t)} = \frac{\pi_j^{(t-1)} f\left(y_i; p_j^{(t-1)}\right)}{\sum_{k=1}^{K} \pi_k^{(t-1)} f\left(y_i; p_k^{(t-1)}\right)}. \tag{5}$$

3. **Maximization Step.** The M-step, on the (t)th iteration, consists in identifying the values of Ψ that maximize the Q-function:

$$Q(\Psi, \Psi^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{K} \gamma_{ij}^{(t)} \log\left(\pi_j f(y_i; p_j)\right),$$

knowing the value of $\gamma_{ij}^{(t)}$ computed in the E-step. In the case of a finite geometric mixture model, the maximization yields:

$$\pi_j^{(t)} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ij}^{(t)}, \tag{6}$$

and

$$p_j^{(t)} = \frac{\sum_{i=1}^{n} \gamma_{ij}^{(t)}}{\sum_{i=1}^{n} \gamma_{ij}^{(t)} y_i}. \tag{7}$$

4. **Iterations.** Steps 2 and 3 are iterated until convergence to a stationary value of the Q-function.

2.4. Selecting the number of components K

The selection of the number of components in the mixture model approaches is a critical task that requires a statistic evaluation (Mclachlan and Krishnan, 1997). This question has been intensively

investigated, and different solutions based on various statistical approaches have been proposed (Leroux, 1992; Green, 1995; Solka et al., 1998). In the domain of mixture modeling, the Bayesian Information Criterion (BIC) (Schwarz, 1978) is probably the most used statistical tool to face this task. It is defined as follow:

$$\mathrm{BIC} = -2 * \ln(\mathcal{L}) + k * \ln(n),$$

where $\mathcal{L}$ is the likelihood of the model, k is the number of parameters to estimate, and n the number of observed data (i.e., the total number of MEMs). The smaller the BIC, the better the model. Increasing the number of components K simultaneously yields the decrease of $-\ln(\mathcal{L})$ and the increase of k. In this work, the BIC criterion was used to choose the value of K.

2.5. Software tools and implementation

The EM algorithm was implemented in C and run under the statistical environment R (version 2.11.1). The source code is available upon request. We used the MUMmer tool (version 3) (Kurtz et al., 2004) to retrieve the MEMs. MUMmer is rooted on a suffix tree algorithm and is widely used to retrieve maximal matches between sequences.

3. RESULTS

The mixture model presented in the previous section can be used to identify an optimal value for ℓ to limit the number of spurious MEMs. Considering the lengths of the MEMs that are retrieved between a pair of sequences, the EM algorithm estimates values for the parameters (p_i) for the K components of the mixture model. Depending on these estimations, it is then possible to determine if a component describes the length of spurious MEMs or the length of significant MEMs. Changing the minimal length (ℓ) of the MEMs impacts the proportion of each type of MEMs. More especially, if ℓ is high enough, the proportion of spurious MEMs will be low enough to be ignored, and hence, the component corresponding to them will not be identified by our model. This latter value of ℓ could be used as a threshold to limit the amount of spurious MEMs.

All the results presented in this section were obtained from 1000 repetitions of the EM algorithm and selection of the estimation that led to the highest likelihood. First, controlled simulated sequences were investigated to illustrate the use of our method, and second, real DNA sequences were studied to show its practical interest.

3.1. Simulated random DNA sequences

3.1.1. Selecting an optimal value for ℓ:ℓ_{opt}. Let's consider a pair of 2-Mb sequences with equal proportion of each base (25%). The first sequence (S_1) was randomly generated, and the second one (S_2) was obtained by applying substitutions on S_1 with a rate of 0.01 substitution per site. Figure 2 displays the results obtained when adjusting a two-component mixture model ($K = 2$) to compare S_1 and S_2.

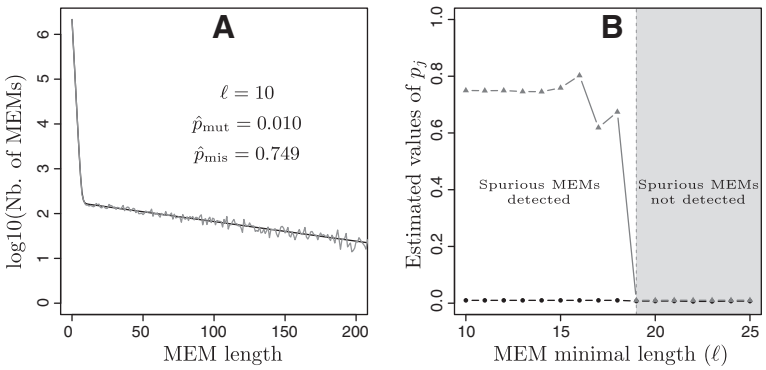

FIG. 2. Estimations obtained with the EM algorithm with two components ($K = 2$) for the comparison of S_1 and S_2. **(A)** Number of MEMs observed with respect to their length (gray curve) and number of MEMs expected according to the mixture model (black curve) with $\ell = 10$. **(B)** Estimated values for the parameters p_j with respect to the minimal length (ℓ) of MEMs.

Figure 2A provides an observed versus fitted plot of MEM length distribution when the minimal length of the retrieved MEMs is fixed at 10 nucleotides ($\ell = 10$). It clearly shows that the model fits the observed data exactly. This means that, if all the MEMs with a minimal length of 10 nucleotides are considered, it is possible to characterize the two distributions that correspond to significant MEMs and spurious MEMs. Interestingly, the estimated $\hat{p}_{\text{mut}}$ value is exactly equal to the substitution rate value chosen to obtain S_2. The parameter p_{mis}, the probability that two nucleotides mismatch, is estimated by $\hat{p}_{\text{mis}} = 0.749$. This corresponds to the probability of mismatch between random position from two sequences with equal proportion of each base ($p_{\text{mis}} = 0.75$).

Figure 2B gives the estimated values of the geometric distribution parameters (p_j) with respect to the minimal length (ℓ) of MEMs. The model had two components ($K = 2$), and thus two (p_j) values were estimated. The black curve corresponds to the lowest estimated values. It does not vary with ℓ values, and it is globally equal to 0.01. This is exactly the substitution rate chosen to obtain S_2. Consequently, it is possible to associate this curve with p_{mut} (i.e., the component corresponding to significant MEMs). The gray curve can be associated with p_{mis}, the component of spurious MEMs. It displays a completely different tendency. For ℓ of 10–15, it has steady values around 0.75; then it varies and rapidly falls to 0.01 for $\ell \geq 19$. This means that for $\ell \leq 15$ the model is able to clearly detect the spurious MEM component, while for $\ell \geq 19$ the parameters of the two components are identical, $p_1 = p_2 \simeq p_{\text{mut}}$, providing us a threshold to discriminate spurious MEMs. As a consequence, in this example, we considered $\ell_{\text{opt}} = 19$ as threshold value. Note that, for ℓ equal to 16, 17, and 18, spurious MEMs are still detected, but they are not enough to accurately estimate the value of p_{mis}, implying the observed fluctuations.

3.1.2. Impacts of sequence characteristics on ℓ_{opt}. Impacts of the characteristics of the compared sequences on the value of ℓ_{opt} were investigated with simulated data. Three parameters were considered: the length, the divergence (measured in substitutions per site), and the base composition (measured in GC%). For each condition of length, divergence, and base composition, the mean of ℓ_{opt} was computed over 50 trials. For each trial, the evaluation of ℓ_{opt} was performed by using the same approach than in Figure 2B. The obtained results are summarized in Figure 3.

Figure 3A is a plot of the empirical mean of ℓ_{opt} values with respect to the divergence of the compared sequences for three different lengths at 1–4 Mb. This figure clearly shows that the length and the relative divergence of the compared sequences yield the selection of different ℓ_{opt} values. First, ℓ_{opt} decreases for substitution rates ranged between 0 and 0.1, and then reached a globally steady state for higher values of substitution rates. This means that, for very closely related sequences, the optimal ℓ value should be higher than for more distant sequences. Moreover, the longer the compared sequences, the higher the ℓ_{opt}. Figure 3B gives the variations of ℓ_{opt} with respect to the GC content of the compared sequences when considering 1-Mb and 2-Mb sequences and with substitution rates set at 0.01 and 0.1. It clearly shows that values of ℓ_{opt} are highly influenced by the GC content of the compared sequences. The higher the GC bias, the higher the ℓ_{opt}.

Selection of ℓ_{opt} is influenced by the intrinsic characteristics of the compared sequences. The overall variation observed in this analysis was 15–30 nucleotides.

In the next subsection, real DNA sequences were investigated to illustrate the applicability of our method on real data.

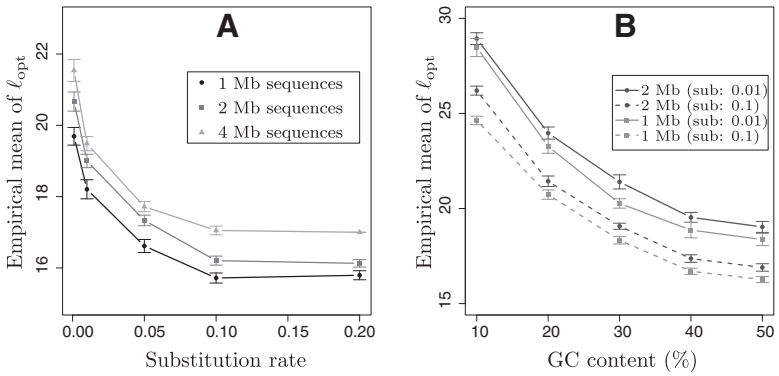

FIG. 3. Mean of ℓ_{opt} values and confidence interval (95%), computed over 50 trials, with respect to the characteristics of compared sequences on simulated data. **(A)** Impact of the length and the divergence of the compared sequences. **(B)** Impact of the base composition (GC%).

3.2. Analysis of real DNA sequences

3.2.1. Impact of the number of components K. The main difficulty when analyzing real DNA sequences is to select a suitable number of components (K) for the mixture model. Indeed, for simulated data, the substitution rate is fixed all along the sequences while real genomes are considerately more heterogeneous, including more or less conserved regions that are sometimes rearranged (i.e., non collinear). Consequently, considering more than one substitution rate is relevant to analyze real DNA sequences. Moreover, variations of K can have an impact on the selection of ℓ_{opt}. This is exemplified in Figure 4. Figure 4A provides the estimations of p_{mis}, the parameter associated with spurious MEMs, for the comparison of the genomes of two strains of *Staphylococcus aureus* (MSSA476 and JH1) when considering $K = 2$ (black curve) and $K = 3$ (gray curve). This figure clearly shows that the values of ℓ_{opt} for both models are the same: $\ell_{\text{opt}} = 18$. This is not the case in Figure 4B, which deals with two strains of *Bacillus cereus* (ATCC 10987 and E33L). In this example, the two-component model yields $\ell_{\text{opt}} = 17$, while the three-component model yields $\ell_{\text{opt}} = 18$. It is thus crucial to set K before the determination of ℓ_{opt}.

Intrinsically, the greater K, the higher the likelihood of the model. This is illustrated in Figure 5, which compares the observed data with the predictions of the mixture model when $K = 2$ and $K = 3$ for the comparison of two *Staphylococcus aureus* strains (MSSA476 and JH1). It shows that, for this example, the model $K = 3$ is significantly better than for the model $K = 2$. On the other hand, increasing the number of components also increases the complexity of the model (i.e., number of parameters to estimate). Thus, the challenge is to find a trade-off between the precision of the model (its likelihood) and its complexity. To face this task, the BIC information defined above was used.

3.2.2. Analysis of complete bacterial genomes. Comparison of complete bacterial genomes at the nucleotide scale is very often limited to closely related genomes (Chiapello et al., 2005). This is why most of the existing approaches deal with intra-species comparisons. In this context, 53 pairs of bacterial genomes from the same species or from very close species were selected to be analyzed with our method. These bacterial genomes are listed in Table 1.

Pairs of genomes were selected in order to have a representative range of lengths and divergence. To evaluate the distance between each pair, we used the MUMi index that was especially designed to measure distances between closely related complete genome sequences (Deloger et al., 2009). Briefly, MUMi evaluates the ratio of cumulative length of maximal unique matches (MUMs) to the total genome length. It varies between 0 and 1: the lower the MUMi, the closer the genomes. For each pair, three models were considered: $K = 2$, $K = 3$, and $K = 4$. The best one, according to the BIC, was selected and used to evaluate the value of ℓ_{opt}. All the results of this analysis are given in Table 1. The pairs of genomes are sorted according to their divergence (MUMi).

The models with $K = 2$, $K = 3$, and $K = 4$ were selected 6, 30, and 17 times, respectively. A slight relation between the selected model, and the divergence of the genomes is observed; more distant sequences seem to require fewer components than closer ones. Intuitively, this can be explained by the fact that close genomes share more MEMs than distant ones, and thus more components are necessary to model the distribution of their length.

The selected values of ℓ_{opt} are 17–23. More than 50% of the selected values of ℓ_{opt} are equal to 18, and approximately 25% of the others are equal to 21. Note that, for 70% of these pairs of genomes, the value

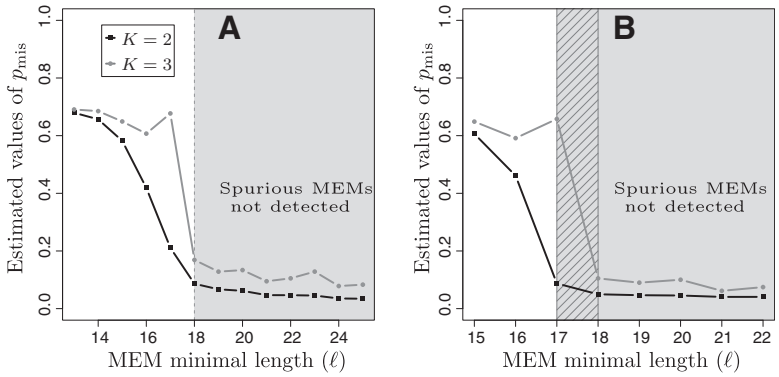

FIG. 4. Impact of the number of components (K) on the selection of ℓ_{opt}. Estimated p_{mis} with respect to the value of ℓ when $K = 2$ and $K = 3$. **(A)** Comparison of two *Staphylococcus aureus* strains (MSSA476 and JH1): similar estimations. **(B)** Comparison of two *Bacillus cereus* strains (ATCC 10987 and E33L): dissimilar estimations.

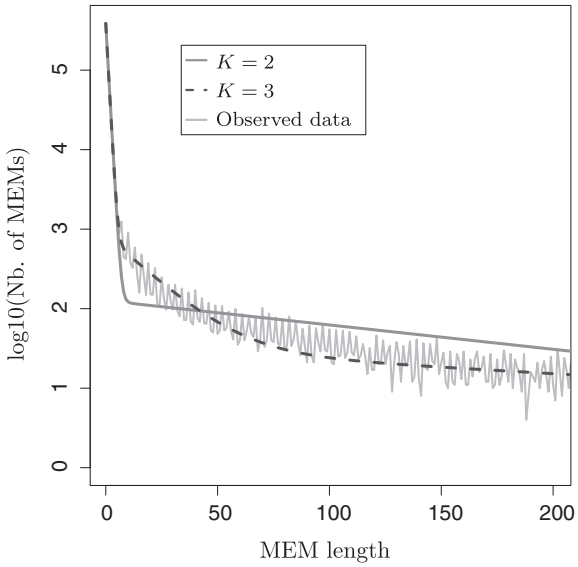

FIG. 5. Impact of the number of components (K) on the model predictions. Comparison of two *Staphylococcus aureus* strains (MSSA476 and JH1): observed data versus fitted values when $K = 2$ and $K = 3$.

of ℓ_{opt} for the best model (selected from BIC) was also found in at least one of the two other models (Table 1). These predictions were confronted with the characteristics of the compared genomes. Figure 6 summarizes these analyses when considering the length and the GC content of the sequences. Interestingly, the highest values of ℓ_{opt} (22 and 23 nucleotides) were obtained for genomes with high GC bias, and the smallest value of ℓ_{opt} (17 nucleotides) was obtained for genomes with a low GC bias. Moreover, it is noteworthy that in this dataset the length of the genomes is correlated with the GC content; small genomes have a high GC bias, and long genomes have a low GC bias. Consequently, the relationship between ℓ_{opt} and the length of the compared sequences is balanced by the GC bias. Lastly, no relation was identified between the selection of ℓ_{opt} and the MUMi index. This study confirms that the relationship between ℓ_{opt} and the characteristics of the genomes is complex and involves various factors.

4. DISCUSSION

The method presented in this article is is based on the principle that the distribution of the length of spurious and significant MEMs between two sequences can be distinguished by using a mixture model of geometric distributions. First, our model identifies the components that characterize the two types of matches. Second, it allows us to establish a threshold for the MEM minimal length (ℓ) values to discriminate between spurious and significant MEMs.

Our model was developed to analyze MEMs retrieved from two sequences. However, it is possible to treat data obtained from more than two sequences. Indeed, the inputs of our model are only the lengths of the retrieved MEMs, and thus, they can come from the comparison of two sequences or more. The only limitation is when considering more than two sequences, the biological significance of the estimated parameters is less intuitive, especially if the compared sequences are very distant.

In this study, two kinds of data were analyzed: simulated DNA sequences and real bacterial genomes. Results obtained with simulated data were particularly interesting. The model was able to retrieve very accurately the parameters of each component. We were also able to identify a relation between the optimal value of ℓ (ℓ_{opt}) and some characteristics of the sequences. Applications to real DNA sequence were also successful. The applicability of the method is the same as for simulated data. The main difficulty remains in the choice of a suitable number of components. We solve this problem by using the BIC. From the analysis of 53 bacterial genomes, we confirmed that characteristics of the compared sequences impact on the determination of ℓ_{opt}.

For sake of simplicity, in our model we assumed that point-mutations are i.i.d. in genomes. Although this is true in simulated sequences, it is obviously not the case in most real DNA sequences. However, we

Table 1. Analysis of 53 Intra-Species Bacterial Genomes

Species (strains)	*Length*	*MUMi*	*K*	ℓ_{opt}	*GC*
Yersinia pestis (KIM/CO92)	4600755/4653728	0.017	4	17*	47.6
Yersinia pestis (91001/CO92)	4595065/4653728	0.023	4	18	47.6
Yersinia pestis (KIM/91001)	4600755/4595065	0.025	4	18	47.6
Staphylococcus aureus (Newman/USA300)	2878897/2872769	0.049	4	19	32.8
Staphylococcus aureus (Newman/COL)	2878897/2809422	0.050	4	19	32.9
Francisella tularensis (LVS/OSU18)	1895994/1895727	0.052	3	23	32.2
Francisella tularensis (LVS/FTA)	1895994/1890909	0.055	3	22	32.2
Escherichia coli (UT189/APEC_O1)	5065741/5082025	0.073	4	21	50.6
Streptococcus pyogenes (MGAS6180/MGAS10270)	1897573/1928252	0.102	4	20	38.4
Staphylococcus aureus (N315/MSSA476)	2814816/2799802	0.105	4	18*	32.8
Staphylococcus aureus (MSSA476/JH1)	2799802/2906507	0.107	4	19	32.9
Streptococcus pyogenes (SF370/MGAS10270)	1852441/1928252	0.138	3	20	38.5
Staphylococcus aureus (RF122/COL)	2742531/2809422	0.159	4	18*	32.8
Haemophilus influenzae (86-028NP /PittGG)	1913428/1887192	0.209	3	21*	38.1
Bacillus thuringiensis (97-27 / A1Hakam)	5237682/5257091	0.210	4	19	35.4
Haemophilus influenzae (86-028NP /PittEE)	1913428/1813033	0.211	3	21	38.1
Haemophilus influenzae (PittEE/ATCC51907)	1813033/1830138	0.216	3	21	38.1
Bacillus cereus/ B. thuringiensis (E33L/A1Hakam)	5300915/5257091	0.228	4	18*	35.4
Escherichia coli (K12/O157:H7)	4639675/5498450	0.248	3	18*	50.7
Shigella flexneri/ E. coli (301/O157:H7)	4607203/5498450	0.308	3	19	50.7
Campylobacter jejuni (NCTC11168/ATCCBAA)	1641481/1845106	0.321	3	21*	30.6
Campylobacter jejuni (ATCCBAA/NCTC11828)	1845106/1628115	0.331	3	21*	30.6
Shigella dysenteriae/E. coli (Sd97/UT189)	4369232/5065741	0.371	3	18*	50.9
Bacillus cereus (ATCC10987/E33L)	5224283/5300915	0.394	4	18*	35.5
Prochlorococcus marinus (AS9601/MIT9301)	1669886/1641879	0.409	3	18*	31.3
Pseudomonas aeruginosa (UCBPP-PA14/PA7)	6537648/6588339	0.415	4	21*	66.4
Salmonella enterica (ATCC9150/RSK2980)	4585229/4600800	0.478	3	18*	51.8
Salmonella enterica (RSK2980/ATCCBAA-1250)	4600800/4858887	0.490	3	18*	51.7
Salmonella enterica (CT18/RSK2980)	4809037/4600800	0.496	3	21*	51.7
Prochlorococcus marinus (AS9601/MIT9215)	1669886/1738790	0.528	3	18*	31.2
Bacillus cereus/ B. anthracis (ATCC14579/Ames a.)	5411809/5227419	0.539	3	18*	35.3
Prochlorococcus marinus (MIT9301/MIT9215)	1641879/1738790	0.543	3	18*	31.2
Bacillus cereus (ATCC14579/ATCC10987)	5411809/5224283	0.555	3	18*	35.4
Prochlorococcus marinus (MIT9312/AS9601)	1709204/1669886	0.598	3	18*	31.3
Prochlorococcus marinus (MIT9312/MIT9215)	1709204/1738790	0.637	3	18*	31.2
Bacillus cereus (ATCC14579/KBAB4)	5411809/5262775	0.644	3	18*	35.4
Bacillus cereus (ATCC_10987/KBAB4)	5224283/5262775	0.647	3	18	35.6
Pseudomonas syringae (1448A/B728a)	5928787/6093698	0.666	3	18*	58.6
Rhodopseudomonas palustris (HaA2/BisB5)	5331656/4892717	0.716	3	21*	65.4
Lactococcus lactis (IL1403/MG1363)	2365589/2529478	0.723	4	18*	35.5
Lactococcus lactis (IL1403/SK11)	2365589/2438589	0.738	3	18*	35.6
Pseudomonas syringae (DC3000/1448A)	6397126/5928787	0.740	4	18*	58.2
Pseudomonas syringae (DC3000/B728a)	6397126/6093698	0.753	4	18*	58.8
Rhodopseudomonas palustris (BisB18/BisA53)	5513844/5505494	0.815	3	21*	64.7
Rhodopseudomonas palustris (ATCCBAA-98/BisB5)	5459213/4892717	0.817	3	21*	64.9
Rhodopseudomonas palustris (BisB5/BisA53)	4892717/5505494	0.851	3	21*	64.6
Wolbachia pipientis (wMel/TRS)	1267782/1080084	0.853	2	18*	34.7
Buchnero aphidicola (Sg/APS)	641454/640681	0.885	2	19	25.8
Prochlorococcus marinus (AS9601/MIT9515)	1669886/1704176	0.920	3	19*	31.1
Buchmera aphidicola (Sg/Cc)	641454/416380	0.932	2	22*	22.7
Candidatus Blochmannia (floridanus/BPEN)	705557/791654	0.955	2	18	28.5
Prochlorococcus marinus (CCMP1375/NATL2A)	1751080/1842890	0.982	2	18*	35.8
Prochlorococcus marinus (CCMP1378/NATL2A)	1657990/1842899	0.982	2	19	33.0

*Selected value of ℓ_{opt} that was retrieved a least once with an other value of K.

For each pair, the following is provided: the species name, the strain names and their corresponding lengths (nuc.), a measurement of their relative divergence with the MUMi index, the number of components K selected according to the BIC value, the selected value of ℓ_{opt}, and the average GC content (%).

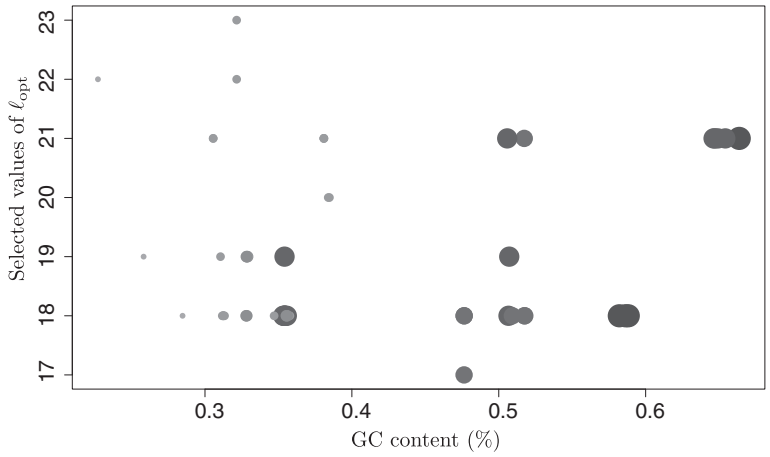

FIG. 6. Selected values of ℓ_{opt} for the 53 pairs of bacterial genomes with respect to the GC content (abscissa) and the length of the genomes (point size and grey scale).

showed (Fig. 5) that our model is able to capture the overall distribution of MEM length between two genomes, despite this assumption.

The lack of effect of the MUMi index on the values of ℓ_{opt} can have different explanations. First, measuring the distances between complete genomes is not as simple as for short DNA sequences. Most of the existing approaches deal with the evaluation of the cumulative length of common words between sequences such as the D_2 statistic (Forêt et al., 2009) or the MUMi index (Deloger et al., 2009) or, with a more flexible definition, the determination of the conserved subsequences between genomes, called the ''backbone'' (Chiapello et al., 2005). This means that distant genomes are not sequences that share homologous regions with a low percentage of identity, but rather they are sequences that share few highly conserved subsequences. This can explain why, when considering very distant genomes (Table 1), no impact was observed on the values of $\ell_{opt,}$ while when increasing the substitution rate in simulated data, ℓ_{opt} significantly decreases. Another possible explanation comes from the following observation: the greater the distance between the compared genomes, the smaller the conserved sequences and thus, the fewer the MEMs. Consequently, when considering very distant genomes, the number of significant MEMs may not be sufficient to accurately estimate the parameter p_{mut}.

For about 75% of the real sequences, estimated ℓ_{opt} values were 18–21. These results are very close to the empirical values used in former studies such as in the complete bacterial genome alignment procedure proposed in Chiapello et al., (2005) ($\ell_{opt} = 20$) or the definition of the MUMi index in Deloger et al., (2009) ($\ell_{opt} = 19$). Our results are also comparable to those obtained in some preliminary statistical studies such as in Guyon and Guénoche (2008) ($\ell_{opt} = 22$). However, the advantage of our approach is to provide a value of ℓ_{opt} adapted to each pair of genomes studied.

Note that our model provides directly estimations of the probabilities that MEMs belong to spurious or significant components, according to their length (see equation 5). Consequently, in addition to the determination of the threshold (ℓ_{opt}), these probabilities can be directly used to weigh MEMs in comparison purposes (e.g., to improve anchor selection in complete genome alignments).

The method presented here can be of interest in several domains. It helps to choose suitable anchor lengths in anchor-based alignments of complete genomes and to calibrate statistical tools to compare large sequences. Other secondary applications can be considered. For example, indexing methods are based on the use of word matches. The size and the precision of an index is directly linked to the length of WMs: increasing the length of WMs increases the precision of the index, but it also increases the index size. Our method could probably provide a good trade-off to balance between size and precision of an index.

Beyond the determination of an optimal value of MEM minimal length, our method is also able to estimate very accurately one or several substitution rates between two sequences. There only exist a few recent methods able to do such estimations on genomic sequences, but to our knowledge, none of them are able to identify several substitution rates for the same pair of genomes. In addition, mixture proportions of our model can provide fruitful information concerning the structure of the compared sequences. Thus, for example, by considering the lengths of the sequences and the mixing proportion of spurious matches (π_{mis}) over the mixing proportion of significant matches (π_{mut}), it is possible to estimate the proportion of

sequences that are specific in each compared genome, providing an estimation of the backbone coverage (Chiapello et al., 2005). Moreover, if several substitution rates are considered ($K > 2$), mixing proportions (π_j) allow us to estimate the different proportions of sequences that are governed by the different estimated substitution rates (data not shown).

Lastly, in this study, we only considered exact matches because they are used over whelmingly in comparative genomics. However, some recent developments use non-exact matches or spaced seeds (Choi et al., 2004), that consist in allowing one or several mismatches. These kind of approaches also suffer from a lack of expertise concerning the length of the seeds and the number of mismatches to consider. Consequently, adapting our method to spaced seeds should provide benefits to the development of such approaches.

ACKNOWLEDGMENTS

We thank Dr. M. El Karoui (Harvard Medical School) for her valuable comments on this work. This work was supported by the Agence Nationale de la Recherche (project CoCoGen; grant BLAN071_185484).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Batzoglou, S. 2005. The many faces of sequence alignment. *Brief. Bioinform.* 6, 6–22.

Chiapello, H., Bourgait, I., Sourivong, F., et al. 2005. Systematic determination of the mosaic structure of bacterial genomes: species backbone *versus* strain-specific loops. *BMC Bioinform.* 6, 171–180.

Chiapello, H., Gendrault-Jacquemard, A., Caron, C., et al. 2008. MOSAIC: an online database dedicated to the comparative genomics of bacterial strains at the intra-species level. *BMC Bioinform.* 9, 498–506.

Choi, K.P., Zeng, F.F., and Zhang, L.X. 2004. Good spaced seeds for homology search. *Bioinformatics* 20, 1053–1059.

Delcher, A.L., Kasif, S., Fleischmann, R.D., et al. 1999. Alignment of whole genomes. *Nucleic Acids Res.* 27, 2369–2376.

Delcher, A.L., Phillippy, A., Carlton, J., et al. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* 30, 2478–2483.

Deloger, M., El Karoui, M., and Petit, M.A. 2009. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. *J. Bacteriol.* 191, 91–99.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* 39, 1–38.

Devillers, H., Chiapello, H., Schbath, S., et al. 2010. Assessing the robustness of complete bacterial genome segmentations. *Lect. Notes Comput. Sci.* 6398, 173–187.

Devillers, H., Chiapello, H., Schbath, S., et al. 2011. Robustness assessment of whole bacterial genome segmentations. *J. Comput. Biol.* 18, 1155–1165.

Field, D., Wilson, G., and Van Der Gast, G. 2006. How do we compare hundreds of bacterial genomes? *Curr. Opin. Microbiol.* 9, 499–504.

Forêt, S., Wilson, S.R., and Burden, C.J. 2009. Empirical distribution of *k*-word matches in biological sequences. *Pattern Recogn.* 42, 539–548.

Forêt, S., Wilson, S.R., and Burden, C.J. 2009. Characterizing the D_2 statistic: word matches in biological sequences. *Stat. Appl. Genet. Mol. B* 8, 43.

Green, P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711–732.

Guyon, F., and Guénoche, A. 2008. Comparing bacterial genomes from linear orders of patterns. *Discrete Appl. Math.* 156, 1251–1262.

Guyon, F., Brochier-Armanet, C., and Guénoche, A. 2009. Comparison of alignment free string distances for complete genome phylogeny. *Adv. Data Anal. Classif.* 3, 95–108.

Höhl, M., Kurtz, S., and Ohlebusch, E. 2002. Efficient multiple genome alignment. *Bioinformatics* 18, 18.

Karlin, S., and Ost, F. 1985. Maximal segmental match length among random sequences from a finite alphabet, 225–243. *In* Cam, L.M.L., and Olshen, R.A., eds. *Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer.* Association for Computing Machinery, New York.

Khan, Z., Bloom, J.S., Kruglyak, L., et al. 2009. A practical algorithm for finding maximal exact matches in large sequence datasets using sparse suffix arrays. *Bioinformatics* 25, 1609–1616.

Kurtz, S., Phillipy, A., Delcher, A.L., et al. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 5, 5.

Leroux, B.G. 1992. Consistent estimation of a mixing distribution. *Ann. Stat.* 20, 1350–1360.

Lippert, R.A., Zhao, X., Florea, L., et al. 2005. Finding anchors for the genomic sequence comparison. *J. Comput. Biol.* 12, 762–776.

Mantaci, S., Restivo, A., Rosone, G., et al. 2008. A new combinatorial approach to sequence comparison. *Theor. Comput. Syst.* 42, 411–429.

Mclachlan, G.J., and Krishnan, T. 1997. *The EM Algorithm and Extensions*. Wiley, New York.

Miller, W. 2001. Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics* 17, 391–397.

Needleman, S.B., and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.

Ohlebusch, E., and Kurtz, S. 2008. Space efficient computation of rare maximal exact matches between multiple sequences. *J. Comput. Biol.* 15, 357–377.

Reinert, G., and Waterman, M.S. 2006. On the length of the longest exact position match in a random sequence. *IEEE ACM Trans. Comput. Biol.* 4, 1–4.

Reinert, G., Chew, D., Sun, F., et al. 2009. Alignment-free sequence comparison (I): Statistics and Power. *J. Comput. Biol.* 16, 1615–1634.

Schwarz, G.E. 1978. Estimating the dimension of a model. *Ann. Stat.* 6, 461–464.

Sims, G.E., Jun, S.R., Wu, G.A., et al. 2009. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proc. Natl. Acad. Sci. USA* 106, 17077–17082.

Solka, J.L., Wegman, E.J., Priebe, C.E., et al. 1998. Mixture structure analysis using the Akaike criterion and the bootstrap. *Stat. Comput.* 8, 177–188.

Treangen, T.J., and Messeguer, X. 2006. M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. *BMC Bioinform.* 7, 433–447.

Ureta-Vidal, A., Ettwiller, L., and Birney E. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* 4, 251–262.

Vinga, S., and Almeida, J. 2003. Alignment-free sequence comparison—a review. *Bioinformatics* 19, 513–523.

Wen, X., Guo, X.Y., and Fan, L.J. 2005. Maximal sequence length of exact match between members from a gene family during early evolution. *J. Zhejiang Univ. Sci.* 6B, 470–476.

Address correspondence to:
Dr. Hugo Devillers
Mathématique
Informatique, et Génome
INRA, UR1077
Domaine de Vilvert
F-78352, Jouy-en-Josas, France

E-mail: hugo.devillers@jouy.inra.fr