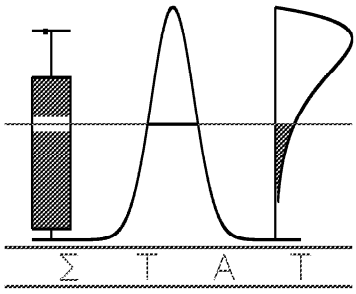
T E C H N I C A L

R E P O R T

0236

A Similarity Measure and Test Between Two DNA Sequences Based on a Mahalanobis Distance Between Word Frequencies

I. Jansen, K. Van Steen, G. Molenberghs, M. De Wit, M. Peeters



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

A Similarity Measure and Test Between Two DNA Sequences Based on a Mahalanobis Distance Between Word Frequencies.

Ivy Jansen,[1] **Kristel Van Steen,**[1] **Geert Molenberghs,**[1]
Mieke De Wit,[2] **and Monika Peeters** [2]**.**

[1] Biostatistics, Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium

[2] Tibotec-Virco, B-2800 Mechelen, Belgium.

Summary

Searching genetic databases for similarities between DNA sequences is nowadays widely used. Well-known and very popular packages are FASTA and BLAST. Nevertheless, past research has shown that they do not find all (dis)similarities, found by word-based search tools.
In this paper, we use the word-based method, introduced by Wu, Burke and Davison (1997), to develop and evaluate an equivalence test, using simulation studies. The "closeness" of DNA sequences is quantified via a Mahalanobis-type distance, accounting for variances and covariances between frequencies of n-words.
Knowing the closeness between two DNA sequences is very important in the field of accrediting new laboratories. It can also be used as a measure of consistency if multiple sequences are generated by a single lab. The introduced methodology can be extended to amino acid sequences instead of sequences of adjacent letters.

Key words: DNA Sequence Comparison, Mahalanobis Distance, Word Counts.

1 Introduction

Any organism's genetic information is coded and stored in its DNA, a linear sequence of molecules that can be seen as a necklace of simple building stones, called nucleotides and represented by the letters a (adenine), c (cytosine), t (thymine) and g (guanine). An n-word is a subsequence of n adjacent letters constructed of these four nucleotides, so there are 4^n different possible n-words. Each sequence of three bases can be thought of as a word that specifies a particular amino acid, of which there are 20 common types. Those molecules are the building blocks of proteins. Some parts of the DNA code for these proteins, some code for the time point at which a protein has to be produced, some code for how much of a protein needs to be produced.

Nowadays, quantifying the similarity between two DNA sequences is one of the cornerstones of modern molecular biology. With the development of dynamic programming theory and with the availability of high-speed computers, alignment algorithms became a popular and widely used tool in biological sequence comparisons. The prototype of a global algorithm is the classic Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). The Smith-Waterman algorithm is the best known local alignment algorithm (Smith and Waterman, 1981). Both alignment methods assign scores to insertions, deletions and replacements, and compute an alignment of two sequences that corresponds to the least costly set of such mutations, hereby maximizing the similarity between the two sequences.

Rather than comparing individual residues in two sequences, FASTA (Fast Alignment) searches for matching sequence patterns or words, or k-tuples (Wilbur and Lipman, 1983; Lipman and Pearson, 1985; Pearson and Lipman, 1988). These patterns comprise k consecutive matches in both sequences. The program attempts to build a local alignment based on

these word matches. FASTA thus compares an input DNA or protein sequence to all of the sequences in a target database and reports the best-matched sequences and local alignments of these matched sequences with the input sequence. The FASTA approach is very popular and is accepted in the biological community as being sensitive and selective. Nevertheless, FASTA does not find all (dis)similarities, found by word-based search tools (Hide, Burke and Davison, 1994; Blaisdell, 1989).

The paper of Wu, Burke and Davison (1997) characterized and compared the relative performance of a family of word-based dissimilarity measures that define a distance between two sequences by simultaneously comparing the frequencies of all n-words in the two sequences. The Mahalanobis distance, which accounts for both the variances and covariances between frequencies of n-words, turned out to give the best performance.

Van Steen et al. (2001, 2002) focused on the analysis of accrediting new labs, i.e., on testing whether DNA sequences composed in different labs are sufficiently close. In Van Steen et al. (2001), the classical Mahalanobis distance between nucleotides is used to make inferences about the similarity of the DNA sequences. An alternative approach based on generalized estimating equations and pseudo-likelihood is adopted in Van Steen et al. (2002).

In this paper, we focus on comparing DNA sequences generated by two different labs, which can be seen as ratings on 1047 characteristics (loci). From this perspective, we are interested in the closeness or similarity of ratings as a measure of agreement, or as a measure of consistency if multiple sequences are generated by a single lab. Closely linked to the concepts of similarity of ratings is the idea that the level of disagreement between any two ratings might be represented by a distance measure. Measuring this disagreement by allowing insertions or deletions, will be counterintuitive. Using word-based (dis)similarity measures may therefore

be a better choice. We will focus on the Mahalanobis distance between frequencies of words, as defined in Wu, Burke and Davison (1997). These word-based methods are also applicable to amino acid sequences with minor modifications.

A major difference with the distance defined in Wu, Burke and Davison (1997), is that we use it to find the degree of similarity between sequences, and not to find matching parts in sequences. Therefore some crucial modifications are necessary.

This paper has the following organization. Section 2 provides a brief description of the data and introduces the format of the data for further reference. In Section 3, we will shortly introduce the Mahalanobis distance between frequencies of words. Section 4 focuses on the derivation of a formal test to assess whether or not a particular level of disagreement is statistically significant or has occurred by chance. It will be applied immediately to the data. Conclusions are drawn in Section 5. The latter is followed by a discussion in Section 6.

2 Data Description

Proficiency testing is a key part of a laboratory's quality control activities. It offers to laboratory customers independent evidence of the laboratory's performance. The purpose of the proficiency testing programme, as set up by Tibotec-Virco, is to enable ongoing monitoring of a laboratory's competence in the genotypic sequencing of HIV-containing samples. A battery of seven samples was selected which contained all relevant genotypic resistance profiles. These samples were first sent to four reference laboratories for replicate (5 times) sequencing. Participating laboratories sequenced the samples only once. All nucleotide sequences were summarized using IUPAC-IUB Ambiguity Codes, allowing for mixtures of nucleotides at certain positions.

Practically, each genotype sequence exists of (i) nucleotide/amino acid of HIV-1 PR (codons 1 through 99 $\times$ 3) and (ii) nucleotide/amino acid of HIV-1 RT (codons 1 through 250 $\times$ 3). This results in a total of L=1047 nucleotides. Only one of the samples (sample 002) is selected for illustrative purposes. Since one reference lab sequenced all samples only once, the pool of reference sequences comprises 16 DNA sequences.

All information within the reference labs is summarized in one consensus sequence (of length 1047) by selecting the most frequent nucleotide per locus. Ties are broken arbitrarily. Moreover, if all sequences are non-informative with respect to a particular locus, then the derived consensus sequence shows a missingness code at that locus. In addition, mixtures of nucleotides are treated as missing observations.

To assess the quality of a query (new) DNA sequence, we will calculate the distance between the query sequence and the consensus sequence, using the technique of word counts as described in the next section.

3 The Model

Wu, Burke, and Davison (1997) introduced a technique to compare a query sequence with a library sequence, based on a Mahalanobis distance between frequencies of words. In fact, their goal was to find parts of sequences that match with the query sequence, to make inferences about the characteristics of the query sequence based on known characteristics of the library sequences.

We will adapt this method to assess the quality of a new query sequence, calculating how well this query sequence maps with the consensus sequence.

We focus on the independent model of genome composition, while the four bases that make up a sequence are assumed to occur independently. The independence of bases is an approximation to the actual dependence in DNA sequences.

3.1 Covariance Between Frequencies of Words

Following the development in Wu, Burke and Davison (1997), we define a strand of DNA of length $(\ell + n - 1)$ (having ℓ n-words, where $\ell > 1$) on the alphabet $\{a, c, t, g\}$. For $\omega = (a_1, \ldots, a_n)$, an arbitrary but fixed n-word within the strand, $N_\omega = \sum_{i=1}^{\ell} X_i$ refers to the frequency of ω. Here X_i represents an indicator variable which takes on the value 1 if ω begins at position i and 0 otherwise. If the product of the probabilities of the first k and the last k letters in ω are denoted by respectively P_k and P_k^* then Gentleman and Mullin (1989) obtained

$$E(N_\omega) = \ell P_n \tag{3.1}$$

and

$$var(N_\omega) = \ell P_n(1 - \ell P_n) + P_n^2(\ell - n + 1)(\ell - n) + 2P_n \sum_{k=1}^{n-1}(\ell - k)P_k Q_{n-k}, \tag{3.2}$$

where $Q = (Q_1, \ldots, Q_n)$ is the overlap capability of ω, defined as

$$Q_i = \begin{cases} 1, & \text{if } (a_1, \ldots, a_i) = (a_{n-i+1}, \ldots, a_n), \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

We observe that P_n is the probability of occurrence of the n-word ω. For later use, we also need the covariance between the frequency of two n-words within a DNA sequence. Therefore, let $\gamma = (b_1, \ldots, b_n)$ represent another n-word that is different from ω and define Y_i, N_γ, R_k, R_k^* and R_n similar to X_i, N_ω, P_k, P_k^* and P_n. As explained in Wu, Burke and Davison (1997), a key concept is the overlap capability $C(\omega\gamma) = (C_1(\omega\gamma), \ldots, C_n(\omega\gamma))$ of ω

with respect to γ:

$$C_i(\omega\gamma) = \begin{cases} 1, & \text{if } (a_{n-i+1}, \ldots, a_n) = (b_1, \ldots, b_i), \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

$C(\gamma\omega) = (C_1(\gamma\omega), \ldots, C_n(\gamma\omega))$ is defined similarly. Note that, in general, $C(\omega\gamma) \neq C(\gamma\omega)$ when $\omega \neq \gamma$. Now, using the overlap information,

$$\begin{aligned} E(X_iY_j) &= P_nP[Y_j = 1|X_i = 1](= R_nP[X_i = 1|Y_j = 1]) \\ &= \begin{cases} P_nR^*_{j-i}C_{n-(j-i)}(\omega\gamma) = R_nP_{j-i}C_{n-(j-i)}(\omega\gamma), & \text{if } 1 \leq j-i \leq n-1, \\ R_nP^*_{i-j}C_{n-(i-j)}(\gamma\omega) = P_nR_{i-j}C_{n-(i-j)}(\gamma\omega), & \text{if } 1 \leq i-j \leq n-1, \\ P_nR_n, & \text{if } |j-i| > n-1. \end{cases} \end{aligned} \tag{3.5}$$

By definition, the probability of having different words ω and γ starting at the same position is zero, so

$$\begin{aligned} cov(N_\omega, N_\gamma) &= \sum_{i,j-1}^{\ell} cov(X_i, Y_j) = (\ell - n + 1)(\ell - n)P_nR_n \\ &\quad + R_n \sum_{k=1}^{n-1}(\ell - k)P_kC_{n-k}(\omega\gamma) + P_n \sum_{k=1}^{n-1}(\ell - k)R_kC_{n-k}(\gamma\omega) - \ell^2P_nR_n. \end{aligned} \tag{3.6}$$

If the four bases occur with equal probability, the equations (3.2), (3.5), and (3.6) can be simplified by replacing both P_n and R_n by 4^{-n} and all of P_k, R_k, P^*_k, and R^*_k by 4^{-k}.

3.2 Similarity Measure Based on Distances Between Word Frequencies

In this section we assume two strands of DNA sequences Q and L (for the query and consensus sequence). Let $V_{L,n} = (N_{\omega_1}/\ell, \ldots, N_{\omega_{4^n}}/\ell)$ be the vector of relative frequencies of n-words over a segment W_L, which is a (sliding) window of length $\ell + n - 1$ (thus having ℓ n-words) from the sequence L, and where $\omega_1, \ldots, \omega_{4^n}$ refer to all possible n-words. Let $V_{Q,n}$ and W_Q be defined similarly for Q. Thus, the random vector

$$Z'_n = (z_{n1}, \ldots, z_{n4^n}) = V_{L,n} - V_{Q,n} \tag{3.7}$$

is an expression of the dissimilarity of W_L and W_Q with respect to word composition. Note that Z_n depends on (1) the word size n, (2) the common window length $\ell + n - 1$ of W_L and W_Q, and (3) the particular pair of windows (W_L, W_Q) under consideration.

The Mahalanobis distance for one particular window pair $W = (W_L, W_Q)$, denoted here by $D^2_{n,W}$, is then defined as

$$D^2_{n,W} = Z_n' \Sigma_n^{-1} Z_n, \tag{3.8}$$

with Σ_n the common population covariance matrix of $V_{Q,n}$ and $V_{L,n}$. Remark that, contrary to Wu, Burke and Davison (1997), both windows W_L and W_Q have to start at exactly the same position in the sequence, and need to slide in exactly the same way through the sequences.

Under the independence and uniform model of base composition, the diagonal and nondiagonal entries of Σ_n can be evaluated according to equations (3.2) and (3.6), respectively, while the probability of each of the four bases a, c, t, g is taken to be $1/4$.

The Mahalanobis distance is well known in multivariate statistical analysis, but not yet widely used for measuring DNA sequence dissimilarities. Nevertheless, the Mahalanobis distance is very attractive, because it takes into account not only the scaling and variance of a variable, but also the variation of other variables based on the covariances.

Contrary to Wu, Burke, and Davison (1997), we define the distance between L and Q as the *maximum* of all "window" distances, because we are interested in the degree of similarity between L and Q, and not in finding parts in both sequences that match. Thus,

$$D^2_n = \max_W D^2_{n,W}. \tag{3.9}$$

This measure still depends on (1) the word size and (2) the selected window length $\ell + n - 1$.

Due to dependencies among frequencies of n-words, the covariance matrix Σ_n is a $4^n \times 4^n$ singular matrix so that computing the Mahalanobis distance involves finding the pseudo-inverse of Σ_n. In practice, this may be too difficult to compute for large word sizes.

4 Analysis

The analysis of the DNA sequence comparison consists of several parts. First and foremost, we have to decide which assumptions to make about word and window length, the way of sliding, etc. Later we need to perform simulations to gather knowledge about the impact of these assumptions, and to simulate a density function for the Mahalanobis distance between word frequencies, under several circumstances. The next step is to derive a formal test to assess whether or not a particular level of disagreement is statistically significant or has occurred by chance. Afterwards we will evaluate this test by calculating its power and significance. Finally we will evaluate the new lab with the developed test.

4.1 Assumptions

As in all techniques used for sequence alignment (for example, BLAST or FASTA), a number of assumptions need to be made when calculating the distance between two sequences. While in sequence alignment choices have to be made about the costs of insertions, deletions and replacements, in our setting parameters such as word and window length need to be chosen. In this paragraph we will discuss the effects of certain choices, and try to find a golden rule when using this technique in the future.

The first parameter to be fixed is the word length n. The smaller the word size, the easier differences as arising from single nucleotide polymorphisms can remain undetected. E.g.,

when $n = 1$, and we have at a certain position an a instead of a g, and 2 places further a g instead of an a, the counts for the words a and g are the same in both sequences, such that the distance in the window covering both positions will be zero, whereas the sequences are different. Hence, larger word sizes are preferred. We will stick to a word size of $n = 3$, because of the interpretation that 3 nucleotides form an amino acid, and because of computational reasons ($n > 3$ is computationally too involved).

Second, we need to specify ℓ, the number of n-words per window, or equivalent the window length $\ell + n - 1$. Here we can make similar remarks as for the word size. The longer the window size, the more n-words there are, so the larger the denominator of the proportions of occurrence of n-words, and thus the smaller those proportions. Also different nucleotides at a certain position in both sequences can be 'corrected' further on in the sequence, such that the frequencies of the words do not differ, although the sequences are not the same. So small distortions can have large consequences. In addition, the magnitude of the distances changes dramatically when changing window sizes, namely a small window leads to a higher distance measure than a large window. Bearing in mind all of these considerations, we prefer to take relatively small windows. In Section 4.4 we will formally assess the sensitivity of the number of n-words per window on the analysis results.

Third, a note on the set of window pairs over which to maximize $D^2_{n,W}$ as in (3.9). The exact way of sliding the window over the sequence depends on the user's objective. In this paper we move the window one nucleotide at a time, to make sure that every "mismatch" affects all n-words in which it appears. Another possibility is to move it three nucleotides (one amino acid) at a time, or to take disjoint windows. In this last case, a different nucleotide at the border of both windows, will only affect one n-word, and there will be less importance attached to it, than to different nucleotides in the inner part of a window.

We already mentioned before that, contrary to Wu, Burke, and Davison (1997), we define the distance between two sequences as the maximum of all "window" distances instead of the minimum, because we are interested in the degree of similarity between the sequences, and not in finding parts in both sequences that match. Another possibility for the distance measure, is to take the mean of all window distances. When choosing this mean distance the zeroes will mask high values that may appear, and we observe less difference between both sequences than when using the maximum of all window distances. That is why the maximum distance will be a better choice.

Another important issue, is how to treat incompleteness, meaning in this context positions where a mixture of nucleotides appears, or positions which are not sequenced at all. We will not take into account a word which contains a missing value, meaning that the frequencies within this window will not sum to 100%. A more formal treatment of incompleteness in this context is a topic of ongoing research.

We develop an equivalence test to decide whether or not a newly generated DNA sequence is similar or equivalent to a pool of reference DNA sequences. In doing so, we wish to account for the degree of uncertainty that is incorporated in any sampling procedure and to avoid a too strong dependence on coincidental sequences. The definition of a suitable equivalence region starts with pre-specifying the maximum number δ of loci that are allowed to differ between two DNA sequences. The choice of this number may be driven by (molecular) biological considerations. We will set $\delta = 1\%$, which is acceptable and supposed to be valid for the reference sequences too.

4.2 Simulations

Because deriving analytic distributional properties for this Mahalanobis distance is overly complex, we will draw conclusions based on simulations. Therefore, we will generate 1000 new sequences, disagreeing with the consensus sequence for at most $\delta = 1\%$ of the loci. For all of those sequences, the distance is calculated, and a distribution based on kernel density estimation with Gaussian kernel is built. Also 1000 new sequences are generated which disagree for at least $\delta = 1\%$ of the loci with the consensus sequence, and a distribution is drawn for their distance to the consensus sequence. In Figures 1 and 2, the result of these simulations are shown. The solid line is the density under the null hypothesis, for sequences which differ at most $\delta = 1\%$ with the consensus sequence, the dashed line the density under the alternative hypothesis, for sequences which differ at least $\delta = 1\%$ with the consensus sequence.

The following conclusions can be drawn from Figures 1 and 2. For window lengths equal to 4 and 6, we observe quite some overlap between the density under the null hypothesis and the alternative hypothesis. For window lengths from 9 to 30 there is a clear distinction between the smooth densities under the null and alternative hypothesis. Once the window length becomes larger, the density under the null hypothesis is very peaked, while the density under the alternative hypothesis is spread out over a wide range of values. We also notice that the distance measure decreases with increasing window length, which implies that differences in distance measures will become less clear for larger windows.

4.3 Test Protocol

The next step is to derive a formal test to assess whether or not a particular level of disagreement is statistically significant or has occurred by chance.

Since we do not have any easily derived distributional properties of the Mahalanobis distance (3.9) between frequencies of words, we will have to rely on density estimation techniques and/or simulations.

Setting up the equivalence test, we considered it to be one-sided because our aim is one of agreement. So once the test value of D_n^2 is sufficiently large, evidence is found towards dissimilar DNA sequences.

We can approach the problem of retrieving a critical point in at least two ways. Either we can determine the critical point on the border between the acceptance and the rejection regions, based on the continuous density function, created from the simulation results, or we can construct a confidence interval for this critical point, based on the exact values obtained from the simulated distance measures.

The first method uses the quantiles of the density function, which are quite easy to obtain. The second method uses approximate Monte Carlo $100(1-\gamma)\%$ confidence limits for the $100(1-\alpha)\%$ critical value, which are the observed quantiles in positions ℓ and u with:

$$\begin{aligned} \ell &= \left\lceil S(1-\alpha) - \Phi^{-1}\left(1-\frac{\gamma}{2}\right)\sqrt{S(1-\alpha)\alpha} \right\rceil, \\ u &= \left\lceil S(1-\alpha) + \Phi^{-1}\left(1-\frac{\gamma}{2}\right)\sqrt{S(1-\alpha)\alpha} \right\rceil, \end{aligned} \tag{4.1}$$

where S is the number of sequences generated under the null hypothesis (Nettleton and Doerge, 2000).

When this interval contains the test value, more simulations need to be done to decrease the length of the interval, and to refine the test. When using the critical point based on the created density, this problem of undecidedness will not occur. Table 1 shows the Monte Carlo confidence limits (1), where $S = 1000$, $\alpha = 0.05$ and $\gamma = 0.01$, and the 95% critical point from the estimated density function (2), for several window lengths.

Table 1: 95% Critical point.

window	4	6	9	10	12	15	30
(1)	[45.17,45.17]	[10.72,10.72]	[1.98,2.07]	[1.33,1.47]	[0.73,0.82]	[0.38,0.43]	[0.0696,0.073]
(2)	47.21	10.45	2.08	1.43	0.79	0.41	0.070

window	50	100	500	1000	1047
(1)	[0.023,0.024]	[0.0054,0.0059]	[0.00027,0.00033]	[5.12E-05,6.65E-05]	[4.45E-05,5.96E-05]
(2)	0.023	0.0059	0.00029	6.60E-05	5.11E-05

As we can see, in most cases the continuously determined critical point (2) lies in the 95% Monte Carlo confidence interval (1) for the critical point. Both techniques will therefore, fortunately, lead to similar decision rules.

4.4 Significance and Power of the Test

Before using this test, it is of interest to investigate whether the test is powerful enough to detect departures from equivalence. In Table 2 the power is displayed for several window lengths, using the upper boundary of the Monte Carlo interval for the critical point (1), and the critical point from the estimated density function (2).

Table 2: Power of the test.

window	4	6	9	10	12	15	30	50	100	500	1000	1047
(1)	0.917	0.930	0.982	0.977	0.983	0.985	0.987	0.986	0.985	0.984	0.988	0.976
(2)	0.733	0.943	0.978	0.981	0.985	0.984	0.985	0.985	0.982	0.982	0.986	0.982

We can conclude that the power of the test is sufficiently high for all window lengths. Only for window length 4, the continuously calculated power is quite low.

Also, the actual significance level can be computed, based on 1000 newly generated sequences under the null hypothesis. When using the continuous critical point (see Table 3 (2)), the actual significance level slightly under- or overestimates the significance level set in the beginning of the test ($\alpha = 5\%$), depending on the window length. Using the lower and upper confidence limits for the critical point (see Table 3 (1) $\ell\ell$ and (1) $u\ell$), they respectively overestimate and underestimate the used significance level in most cases, as it should be.

Table 3: Significance of the test.

window	4	6	9	10	12	15	30	50	100	500	1000	1047
(1) $\ell\ell$	0.000	0.093	0.121	0.054	0.068	0.066	0.077	0.059	0.069	0.039	0.051	0.052
(1) $u\ell$	0.000	0.093	0.032	0.026	0.028	0.033	0.046	0.039	0.023	0.016	0.023	0.016
(2)	0.000	0.093	0.032	0.035	0.040	0.041	0.069	0.059	0.023	0.023	0.032	0.016

A different tool to evaluate this test, is to calculate the distance between the separate sequences of the reference labs and the consensus sequence, and to check if this distance can be accepted under the null hypothesis of equivalence. Table 4 shows these distances.

There are several remarks to be made based on this table. First, the results for window length 12 are questionable. Eight out of sixteen reference sequences have a distance to the consensus sequence of more than 1%. This would mean that the consensus sequence is not a good summary sequence for the reference sequences. Therefore we will treat the results for window length 12 with caution. Second, the distance measures for the single sequence of reference lab 4 are quite large. Taking a closer look at both sequences (consensus sequence and reference sequence from lab 4), it can be seen that they differ at 21 out of 1047 places, which is 2%. We would have to reject the null hypothesis of equivalence (maximum difference

Table 4: Maximum distance between consensus sequence and reference sequences.

reference lab	window length										
	4	6	9	10	12	15	30	50	100	500	1000	1047
	17.50	5.54	1.47	1.15	1.01	0.40	0.097	0.033	0.0093	0.00066	0.00016	0.00015
	17.50	5.54	1.23	0.85	1.01	0.37	0.097	0.033	0.0093	0.00047	0.00011	0.00010
1	17.50	3.23	0.76	0.52	1.01	0.29	0.050	0.024	0.0059	0.00027	6.55E-05	5.97E-05
	17.50	3.25	1.47	1.15	0.74	0.40	0.073	0.024	0.0056	0.00046	9.79E-05	8.93E-05
	17.50	5.54	1.23	0.85	1.01	0.37	0.097	0.033	0.0093	0.00047	0.00011	0.00010
	31.83	6.80	1.55	1.10	0.97	0.47	0.063	0.019	0.0032	0.00027	5.04E-05	4.59E-05
	31.83	6.80	1.55	1.10	0.97	0.47	0.063	0.019	0.0032	0.00027	5.04E-05	4.59E-05
2	17.50	3.20	0.73	0.52	0.30	0.16	0.048	0.016	0.0036	0.00013	3.24E-05	2.95E-05
	17.50	3.20	0.73	0.52	1.01	0.29	0.048	0.018	0.0059	0.00014	3.31E-05	3.01E-05
	17.50	3.20	0.73	0.52	0.30	0.16	0.048	0.016	0.0055	0.00020	4.86E-05	4.44E-05
	45.17	10.72	1.98	1.33	0.68	0.39	0.046	0.015	0.0030	7.46E-05	1.79E-05	1.68E-05
	45.17	10.72	1.98	1.33	0.68	0.39	0.046	0.015	0.0030	7.46E-05	1.79E-05	1.68E-05
3	45.17	10.72	1.98	1.33	0.68	0.39	0.046	0.015	0.0030	7.46E-05	1.79E-05	1.68E-05
	45.17	10.72	1.98	1.33	0.68	0.39	0.046	0.015	0.0030	7.46E-05	1.79E-05	1.68E-05
	45.17	10.72	1.98	1.33	0.68	0.39	0.046	0.015	0.0030	7.46E-05	1.79E-05	1.68E-05
4	45.17	6.72	2.42	1.65	0.90	0.43	0.110	0.040	0.0092	0.00059	0.00018	0.00015

of $\delta = 1\%$). Third, large window lengths (> 15) perform worse than the smaller window lengths. This is probably due to them capturing more differences in one window, than do smaller windows, and therefore the maximum distance measure will be larger. Finally, we observe that the results for the 5 sequences of reference lab 3 are identical. This is not surprising since the 5 sequences for lab 3 are identical.

4.5 Evaluating the New Laboratory

We only have one sequence of the new lab, and we will calculate the Mahalanobis distance between this sequence and the consensus sequence of the reference labs. The results are shown in Table 5.

Either we can compare these maximum distances with the critical points in Table 1 (1), or we can calculate a p-value based on the simulation results in the same Table 1 (2). When

Table 5: Maximum distance between consensus sequence and new sequence.

window	4	6	9	10	12	15	30	50	100	500	1000	1047
max	30.5	5.54	1.23	0.85	0.98	0.38	0.10	0.033	0.0096	0.00047	0.00016	0.00013

comparing the distances in Table 5 with the critical points, we have to reject the hypothesis of equivalence for window length 12 and all window lengths larger than 30. The p-values are given in Table 6. Note that, also here, large window lengths give a highly significant result.

Table 6: P-value of the new sequence.

window	4	6	9	10	12	15	30	50	100	500	1000	1047
discrete	0.896	0.516	0.442	0.439	0.010	0.076	0.000	0.001	0.001	0.003	0.001	0.001
continuous	0.695	0.583	0.386	0.386	0.011	0.088	0.001	0.002	0.001	0.006	0.001	0.001

4.6 Conclusions

Taking into account all the performed analysis, we can conclude that there is no single best window length to perform the test for accrediting a new labo. Therefore, a good advice seems to be to choose more than one window length, but not too large, and evaluate all results simultaneously. When doing so for our new lab, and only testing for window lengths less than 15 (as was suggested when comparing the sequences of the reference labs with the consensus sequence), we can conclude that the new lab performs well.

5 Discussion

In this paper we have developed a one-sided equivalence test, based on the comparison of frequency distributions of nucleotide n-words and a Mahalanobis-type distance measure.

More intuitive as a means for measuring disagreement between DNA sequences (e.g., as com-

pared to automated alignment engines heavily relying on the cost of insertions or deletions), the test seems to perform adequately in terms of power.

However, the number and feasibility of underlying assumptions also determine the quality and usefulness of the test. To this end, we note that a consensus sequence was constructed, based on the sequences of the reference labs. We acknowledge that the use of a consensus sequence as a representative sequence is comparable to the use of summary measures in statistical analyses, and may therefore fail to make use of possibly valuable information. However, it appears to be a useful means to simulate the distribution of the Mahalanobis distance under the null hypothesis.

The distance measure selected in comparing a new sequence with a pool of reference sequences, is important as well. We have chosen for the Mahalanobis distance measure, introduced in Wu, Burke and Davison (1997). This distance measure accounts for the covariance between n-words, such that not all "mismatches" have the same impact. It takes into account the overlap capability between n-words.

When using this Mahalanobis distance, several parameters need to be chosen a priori, namely n, the word length, ℓ, the number of n-words per window, δ, the acceptable difference between consensus sequence and new sequence under equivalence, etc. There is no perfect choice. However, the size of the words mostly depends on computational possibilities, because matrices of dimension $4^n \times 4^n$ need to be inverted. The more powerful computers available, the larger n can be taken. Based on our arguments in paragraph 4.1 we recommend $n = 3$. Ideally, several window lengths are selected and all results are evaluated simultaneously. Too large window lengths should be avoided (say, smaller than $10n$), such that mistakes do not balance out in a window. With regard to the allowed disagreement, we can only say that

this is subject to requirements of the authorities deciding when a new lab may be accredited or not.

Future work will involve a sensitivity analysis of different percentages of missingness in the pool of reference sequences or the new sequence. Several ways of treating this missingness will be examined. Also the effect of random missingness or region specific missingness on the test results will be subject to further investigation.

Acknowledgements

Research supported by a PAI program P5/24 of the Belgian Federal Government (Federal Office for Scientific, Technical, and Cultural Affairs).

References

Blaisdell, B.E. (1989) Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences. *Journal of Molecular Evolution*, **29**, 526 – 537.

Churchill, G.A. and Doerge, R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963 972.

Fleiss, J.L., Cohen, J. and Everitt, B.S. (1969) Large-sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72**, 323 – 327.

Gentleman, J.F. and Mullin, R.C. (1989) The distribution of the frequency of occurence of nucleotide subsequences, based on their overlap capability. *Biometrics*, **45**, 35 - 52.

Hide, W., Burke, J. and Davison, D. (1994) Biological evaluation of d^2, an algorithm for high performance sequence comparison. *Journal of Computational Biology*, **1**, 199 - 215.

Lipman, D.J. and Pearson, W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435 – 1441.

Needleman, S.B. and Wunsch, Ch.D. (1970) A general method applicable to search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443 – 453.

Nettleton, D. and Doerge, R.W. (2000) Accounting for variability in the use of permutation testing to detect quantitative trait loci. *Biometrics*, **56**, 52 – 58.

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, **85**, 2444 – 2448.

Smith, J.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195 – 197.

Van Steen, K., Molenberghs, G., De Wit, M. and Peeters, M. (2002) Comparing DNA sequences using generalized estimating equations and pseudo-likelihood. *Biometrics*, **submitted**

Van Steen, K., Thijs, H., Molenberghs, G., De Wit, M. and Peeters, M. (2001) An equivalence test for comparing DNA sequences. *Journal of Statistical Modeling*, **submitted**

Wilbur, W.J. and Lipman, D.J. (1983) Rapid Similarity Searches of Nucleic Acid and Protein Data Banks. *Proceedings of the National Academy of Sciences*, **80**, 726 – 730.

Wu, T-J., Burke, J.P. and Davison, D.B. (1997) A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics*, **53**, 1431 – 1439.

Figure 1: Densities under the null and alternative hypothesis for small window lengths.

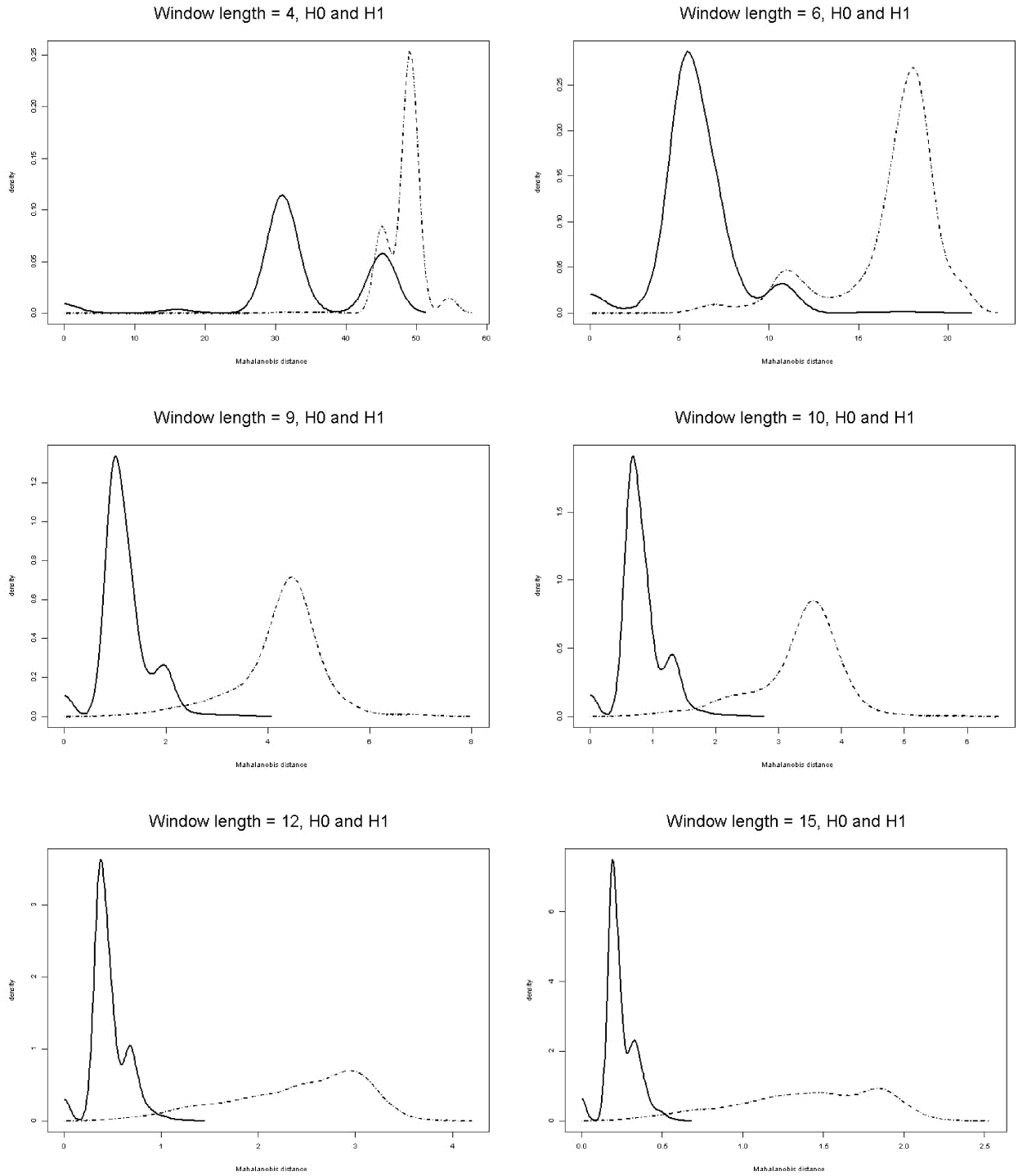


Figure 2: Densities under the null and alternative hypothesis for large window lengths.