Comparison of genomic sequences using the Hamming Distance [*]

Hildete Prisco Pinheiro
Departamento de Estatística
UNICAMP - SP, Brazil

Aluísio de Souza Pinheiro
Departamento de Estatística
UNICAMP - SP, Brazil

Pranab Kumar Sen
Department of Biostatistics
UNC-Chapel Hill

Abstract

The paper considers the problem of homogeneity among groups by comparison of genomic sequences. Some alternative procedures that attach less emphasis on the likelihood approach, and more on alternative measures that deal with similar homogeneity problems are considered here. On this approach, a one-sided hypothesis test is considered and the classical ANOVA decomposition can be directly adapted to sample measures based on the Hamming distance, without necessarily going through their second moments. Some results of U-statistics theory will be useful for the decomposition of the test statistic and to find its asymptotic distribution. An application of this test with real data is shown and the p-value of the test statistic is found via bootstrap resampling.

Keywords: Amino Acid; Asymptotic distribution; Bootstrap; Categorical Data; Genome; Hamming Distance; Nucleotide; Nonparametric; Statistical Genetics; U-statistics.

[*] MSC 2000 subject classifications: Primary-62G10; secondary-62G09, 62G20, 92D20.

This research was funded in part by Fundação de Amparo à Pesquisa do Estado de São Paulo (2002/04332-3).

1 Introduction

Most problems in computational sequence analysis (CSA) are essentially statistical. Stochastic evolutionary forces act on genomes. In genomic sequence analysis, typically, we encounter data on a large number (K) of positions or *sites*, and in each position, we have a purely qualitative (nucleotides or amino acid labels) categorical response with 4 to 20 categories depending on the DNA or protein sequence. The spatial (functional as well as stochastic) dependence (or association) patterns of these sites may not be known, nor can they be taken to be *stochastically independent*. Also, regular and nearly identical structures of the DNA solicitate statistical appraisal based on other variational properties which exhibit more statistical variation and information too.

In this high-dimensional qualitative response setup, it is difficult to incorporate standard (discrete or continuous) multivariate analysis tools, in a parametric formulation (as the number of associated parameters may be exceedingly large and the underlying model may not be that well specified or anticipated).

If we restrict ourselves to a single site, in most cases there is little statistical information. In a multiple site context, we need to consider high-dimensional qualitative categorical data models trying to preserve intersite dependence as much as possible, and then to proceed to CSA statistical appraisals.

Both parametric and nonparametric procedures for categorical data tests are available. Some attack the problem parametrically on a gene level context (Chernoff and Lander, 1995; Chernoff, 1993). We consider here procedures that attach less emphasis on the likelihood approach, and more on alternative measures that deal with homogeneity problems of nucleotide/amino acid distributions. We may refer to Pinheiro et al. (2000, 2001) for some related work.

The forementioned works, although different in their paradigms and their level of analysis (gene vs nucleotide/amino acid), have a common ground. Both situations deal with single vs multiple sources of variation. For instance, in (Chernoff and Lander, 1995), hypotheses concern the existence of genetic markers. In Pinheiro et al. (2000, 2001), one tests homogeneity of nucleotide distribution among groups. Both solutions depend heavily on asymptotic testing. Moreover, the discrete nature of each problem results in mixture-type distributions.

On Section 2, we present the problem and motivate the use of Hamming distance in its solution. Avoiding second moment considerations, a decomposition of the within and in between groups is proposed and used in a one-sided hypothesis test. On Section 3 we review some results of U-statistics theory which will be useful for the decomposition of our test statistic and to find the asymptotic distribution of the statistic presented on Section 4. Finally, on Section 5 an application of this test with real data is presented and the p-value of the test statistic is found via bootstrap resampling.

2 Some nonparametrics and the Hamming Distance as a measure of diversity

Consider a general CSA with K sites, each one having a categorical response with $C(\geq 2)$ qualitative categories, indexed as $1, \ldots, C$. For the ith sequence, let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iK})'$ be a random vector of responses where X_{ik} denotes the category outcome $c(= 1, \ldots, C)$ at site $k(= 1, \ldots, K)$. Recalling that these sites may not be stochastically independent, we need to have a measure of divergence which takes into account the inter-site stochastic dependence to a certain extent. The primary motivation for using a diversity measure stems from the fact that HIV or some other retrovirus have the ability to have higher mutation rates which can be traced with a diversity index, without going through some likelihood formulations. With that in mind, we define the *Hamming distance* between a pair (i, i') of sequences as

$$D_{ii'} = \frac{1}{K} \sum_{k=1}^{K} I(X_{ik} \neq X_{i'k}), \tag{1}$$

so that $D_{ii'}$ is the proportion of sites where $\mathbf{X}_i$ and $\mathbf{X}_{i'}$ do not match. Since the K coordinate indicator functions are not necessarily independent, this measure attempts to take into account their dependence, albeit in a symmetric manner. It is easy to see that the expected value of $D_{ii'}$ is the average (over the K positions) *Gini-Simpson* diversity indexes (Gini, 1912), which we denote by $\mathcal{H}$.

$$E(D_{ii'}) \quad = \quad \frac{1}{K} \sum_{k=1}^{K} P(X_{ik} \neq X_{i'k}) = \frac{1}{K} \sum_{k=1}^{K} P(X_{ik} = x_{ik}, X_{i'k} \neq x_{ik})$$

$$= \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{kc}(1-\pi_{kc}) = 1 - \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{kc}^2 = \mathcal{H}$$

where π_{kc} is the probability of being in category c at position k.

It is also possible to employ other measures of diversity which have nonparametric flavor; for details, we refer to Pinheiro et al. (2000).

It may be remarked that an optimal nonparametric estimator of $\mathcal{H}$ is the Hoeffding (1948) U-statistic (corresponding to the kernel D_{ij} of degree 2):

$$\bar{D}_n = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} D_{ij}. \tag{2}$$

This U-statistic formulation enables us to use conventional statistical tools for testing homogeneity of the $\mathcal{H}$ for different groups. In conventional way of using ANOVA procedures based on U-statistics, such tests for homogeneity of the $\mathcal{H}$ for the different groups were considered by Pinheiro et al. (2000, 2001).

Pinheiro et al. (2001) defined the average distance within a group as

$$\bar{D}_{gg} = \binom{n_g}{2}^{-1} \frac{1}{K} \sum_{1\le i<j\le n_g} \sum_{k=1}^{K} I(X_{ik}^g \neq X_{jk}^g)$$

which is a U-statistic of degree 2 (Lee, 1990). Therefore it is an unbiased estimator of the population average distance within group g, say $\mathcal{H}_{gg}$.

$$\begin{aligned}\mathcal{H}_{gg} &= E(\bar{D}_{gg}) = \frac{2}{n_g(n_g-1)} \sum_{1\le i<j\le n_g} \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{kgc}(1-\pi_{kgc}) \\ &= 1 - \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{kgc}^2\end{aligned}$$

The average distance between groups g and g' is

$$\bar{D}_{gg'} = \frac{1}{n_g n_{g'}} \sum_{i=1}^{n_g}\sum_{j=1}^{n_{g'}} \frac{1}{K}\sum_{k=1}^{K} I(X_{ik}^g \neq X_{jk}^{g'})$$

which is a two-sample U-statistic of degree (1,1) (Hoeffding, 1948; Puri and Sen, 1971; Lee, 1990) and therefore, it is an unbiased estimator of the population average distance between groups g and g', say $\mathcal{H}_{gg'}$.

$$\begin{aligned}\mathcal{H}_{gg'} &= E(\bar{D}_{gg'}) = \frac{1}{n_g n_{g'}} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g'}} \frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} \pi_{kgc}(1 - \pi_{kg'c}) \\ &= 1 - \frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} \pi_{kgc}\pi_{kg'c}\end{aligned}$$

Note that

$$\pi_{kgc}\pi_{kg'c} \leq \frac{1}{2}(\pi_{kgc}^2 + \pi_{kg'c}^2) \tag{3}$$

Therefore,

$$\begin{aligned}\mathcal{H}_{gg'} &\geq 1 - \frac{1}{2}\frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} (\pi_{kgc}^2 + \pi_{kg'c}^2) \\ &= \frac{1}{2}\left[1 - \frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} \pi_{kgc}^2 + 1 - \frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} \pi_{kg'c}^2\right] \\ &= \frac{1}{2}\left[\mathcal{H}_{gg} + \mathcal{H}_{g'g'}\right] \end{aligned} \tag{4}$$

Let $w_{gg'}$, $g, g' = 1, \ldots, G$, be a set of nonnegative weights such that $\sum_{g=1}^{G} w_{gg'} = 1$. Also, let $w_g^o = \sum_{g=1}^{G} w_{gg'}$ and $w_g^\star = w_g^o - w_{gg} = \sum_{g=1(\neq g')}^{G} w_{gg'}$, $g = 1, \ldots, G$. Thus, we have $\sum_{g=1}^{G} w_g^o = 1$ and $\sum_{g \neq g'=1}^{G} w_{gg'} = \sum_{g=1}^{G} w_g^\star = 1 - \sum_{g=1}^{G} w_{gg}$.

Then, it readily follows from (4) that

$$\frac{\sum_{g \neq g'=1}^{G} w_{gg'}\mathcal{H}_{gg'}}{\sum_{g \neq g'=1}^{G} w_{gg'}} \geq \frac{\sum_{g=1}^{G} w_g^\star \mathcal{H}_{gg}}{\sum_{g=1}^{G} w_g^\star}. \tag{5}$$

The RHS of (5) represents a "within group" measure, while the LHS is a "between group" one; both being nonnegative. This suggests that the classical ANOVA decomposition can be directly adapted to the sample measures $\bar{D}_{gg'}$, $1 \leq g, g' \leq G$, without necessarily going through their second moments (as done in Pinheiro et al, 2001). Motivated by this feature, here we pursue the ANOVA decomposition on $\bar{D}_{gg'}$, and proceed as follows.

$$\bar{D}_n^{(o)} = \binom{n}{2}^{-1}\left(\sum_{g=1}^{G}\binom{n_g}{2}\bar{D}_{gg} + \sum_{1\le g<g'\le G} n_g n_{g'}\bar{D}_{g,g'}\right) \qquad (6)$$

which is a linear combination of U-statistics. $\bar{D}_n^{(o)}$ is the overall distance or the Hamming distance for the pooled sample of the G groups as a combined set of $n = \sum_{g=1}^{G} n_g$ sequences.

Note that we can write

$$\begin{aligned}\bar{D}_n^{(o)} &= \sum_{g\neq g'}\frac{n_g n_{g'}}{n(n-1)}\bar{D}_{gg'} + \sum_{g=1}^{G}\left[\frac{n_g}{n} - \frac{n_g}{n} + \frac{n_g^2}{n(n-1)} - \frac{n_g}{n(n-1)}\right]\bar{D}_{gg} \\ &= \sum_{g=1}^{G}\frac{n_g}{n}\bar{D}_{gg} + \sum_{g\neq g'}\frac{n_g n_{g'}}{n(n-1)}\bar{D}_{gg'} - \sum_{g=1}^{G}\frac{n_g(n-n_g)}{n(n-1)}\bar{D}_{gg} \\ &= D_n(W) + D_n(B)\end{aligned}$$

where

$$D_n(W) = \sum_{g=1}^{G}\frac{n_g}{n}\bar{D}_{gg} \qquad (7)$$

and

$$\begin{aligned}D_n(B) &= \sum_{g\neq g'}\frac{n_g n_{g'}}{n(n-1)}\bar{D}_{gg'} - \sum_{g=1}^{G}\frac{n_g(n-n_g)}{n(n-1)}\bar{D}_{gg} \\ &= \frac{1}{n(n-1)}\left\{\sum_{g=1}^{G-1}\sum_{g'=g+1}^{G} n_g n_{g'}(2\bar{D}_{gg'} - \bar{D}_{gg} - \bar{D}_{g'g'})\right\} \qquad (8)\end{aligned}$$

$$E(\bar{D}_n^{(o)}) = \sum_{g=1}^{G}\frac{n_g}{n}\mathcal{H}_{gg} + \sum_{g\neq g'}\frac{n_g n_{g'}}{n(n-1)}\mathcal{H}_{gg'} - \sum_{g=1}^{G}\frac{n_g(n-n_g)}{n(n-1)}\mathcal{H}_{gg}$$

In order to test the hypothesis of homogeneity of $\mathcal{H}$ for differente groups, based on the results given in (3) and (4) we can write the hypothesis

$$\begin{aligned}&H_0 : 2\mathcal{H}_{gg'} = \mathcal{H}_{gg} + \mathcal{H}_{g'g'}, \forall g \neq g' \\ &H_1 : 2\mathcal{H}_{gg'} > \mathcal{H}_{gg} + \mathcal{H}_{g'g'} \qquad (9)\end{aligned}$$

Under the null hypothesis described in (9), it is easy to see that $E(D_n(B)) = 0$ and also

$$\begin{aligned} E_0(D_n^{(o)}) &= \sum_{g=1}^{G} \frac{n_g}{n} \mathcal{H}_{gg} + \sum_{g \neq g'} \frac{n_g n_{g'}}{n(n-1)} \left(\frac{\mathcal{H}_{gg} + \mathcal{H}_{g'g'}}{2} \right) - \sum_{g=1}^{G} \frac{n_g(n-n_g)}{n(n-1)} \mathcal{H}_{gg} \\ &= \sum_{g=1}^{G} \frac{n_g}{n} \mathcal{H}_{gg} = E(D_n(W)) \end{aligned}$$

Therefore, we can define $D_n(B)$, described in (8) as our test statistic.

3 U-statistics Theory Results

Using U-statistic theory we know that if F denotes the distribution function of X_i and U^m is a U-statistic of degree m, computed from a sample of size n, with kernel $\phi(X_1, \ldots, X_m)$ and $E(U^m) = \theta(F) = \theta$.

$$U^m \equiv U(X_1, \ldots, X_n) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \cdots < i_m \leq n} \phi(X_{i_1}, \ldots, X_{i_m}), \;\; n \geq m \tag{10}$$

where $\theta(F) = E_F\{\phi(X_1, \ldots, X_m)\} = \int \ldots \int \phi(x_1, \ldots, x_m)\, dF(x_1) \ldots dF(x_m)$

Let

$$\Psi_c(x_1, \ldots, x_c) \equiv \mathrm{E}\{\phi(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m)\} \tag{11}$$

The function Ψ_c has the following properties (Lee, 1990, p. 11):
(i) $\Psi_c(x_1, \ldots, x_c) = \mathrm{E}\{\Psi_d(x_1, \ldots, x_c, X_{c+1}, \ldots, X_d)\}$ for $1 \leq c < d \leq m$,
(ii) $\mathrm{E}\{\Psi_c(x_1, \ldots, x_c)\} = \mathrm{E}\{\phi(X_1, \ldots, X_m)\}$.

$F_n(x)$ is the empirical distribution function (d.f.)

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \epsilon(x - X_i) \qquad x \in I\!R^p, \qquad n \geq 1$$

with $\epsilon(u)$ being 1 if all p coordinates of u are nonnegative and 0 otherwise.

We may rewrite (10) as

$$U^m = n^{-[m]} \sum_{1 \leq i_1 \neq \ldots \neq i_m \leq n} \int_{I\!R^{pm}} \ldots \int \phi(x_1, \ldots, x_m) \prod_{j=1}^{m} d(\epsilon(x_j - X_{i_j})), \tag{12}$$

where $n^{-[m]} = (n^{[m]})^{-1} = \{n \dots (n-m+1)\}^{-1}$.

Writing $d(\epsilon(x_j - X_{i_j})) = dF(x_j) + d[\epsilon(x_j - X_{i_j}) - F(x_j)],\ 1 \le j \le m$, we obtain

$$U^m = \theta(F) + \sum_{h=1}^{m} \binom{m}{h} U_h^m, \qquad n \ge m \tag{13}$$

where

$$U_h^m = n^{-[h]} \sum_{1 \le i_1 \ne \cdots \ne i_h \le n} \int_{I\!R^{ph}} \cdots \int \Psi_h(x_1, \dots, x_h) \prod_{j=1}^{h} d[\epsilon(x_j - X_{i_j}) - F(x_j)],$$

for $1 \le h \le m$.

Further, if we write

$$\begin{aligned} \Psi_h^\circ(x_1, \dots, x_h) &= \Psi_h(x_1, \dots, x_h) - \sum_{j=1}^{h} \Psi_{h-1}(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_h) \\ &\quad + \cdots + (-1)^h \theta(F), \qquad \forall\ (x_1, \dots, x_h) \in I\!R^{ph}, \end{aligned}$$

for $1 \le h \le m$, we obtain

$$U_h^m = \binom{n}{h}^{-1} \sum_{1 \le i_1 < \cdots < i_h \le n} \Psi_h^\circ(X_{i_1}, \dots, X_{i_h}), \qquad 1 \le h \le m \tag{14}$$

and the U_h^m are themselves U-statistics.

A U-statistic of degree m can be decomposed as

$$U^m = \theta(F) + \frac{m}{n} \sum_{i=1}^{n} [\Psi_1(X_i) - \theta(F)] + O_p(n^{-1}) \tag{15}$$

The decomposition for a two-sample U-statistic of degree (m_1, m_2) can be developed similarly to the one-sample U-statistic as

$$\begin{aligned} U^{(m_1,m_2)} &= \theta(\mathbf{F}) + \frac{m_1}{n_1} \sum_{i=1}^{n_1} [\Psi_{10}(X_i) - \theta(\mathbf{F})] + \frac{m_2}{n_2} \sum_{i=1}^{n_2} [\Psi_{01}(Y_i) - \theta(\mathbf{F})] \\ &+ O_p(n_0^{-1}) \end{aligned} \tag{16}$$

where $n_0 = min(n_1, n_2)$.

Note that in our case

$$\begin{aligned}
\Psi_1(\mathbf{x}_i^g) &= 1 - \sum_{c=1}^{C} \pi_{cgk}^2 = \mathcal{H}_{gg} \\
\Psi_2(\mathbf{x}_i^g, \mathbf{x}_j^g) &= I(X_{ik}^g \neq X_{jk}^g) \\
\Psi_{10}(\mathbf{x}_i^g) &= 1 - \sum_{c=1}^{C} \pi_{cgk}\pi_{cg'k} = \mathcal{H}_{gg'} \\
\Psi_{01}(\mathbf{x}_i^{g'}) &= 1 - \sum_{c=1}^{C} \pi_{cg'k}\pi_{cgk} = \mathcal{H}_{gg'} \\
\Psi_{11}(\mathbf{x}_i^{g'}, \mathbf{x}_i^g) &= I(X_{ik}^g \neq X_{jk}^{g'})
\end{aligned}$$

Since $\bar{D}_{gg'}$ is a two-sample U-statistic of degree (1,1) and $\bar{D}_{gg}$ and $\bar{D}_{g'g'}$ are U-statistics of degree 1 we can decompose them as

$$\begin{aligned}
\bar{D}_{gg'} &= \mathcal{H}_{gg'} + \frac{1}{n_g}\sum_{i=1}^{n_g}\sum_{k=1}^{K}\frac{1}{K}\left\{[1 - P_{g'}(x_{ik}^g)] - \mathcal{H}_{gg'}\right\} \\
&+ \frac{1}{n_{g'}}\sum_{j=1}^{n_{g'}}\sum_{k=1}^{K}\frac{1}{K}\left\{[1 - P_g(x_{jk}^{g'})] - \mathcal{H}_{gg'}\right\} \\
&+ \frac{1}{n_g n_{g'}}\sum_{i=1}^{n_g}\sum_{j=1}^{n_{g'}}\sum_{k=1}^{K}\frac{1}{K}\Big\{I(X_{ik}^g \neq X_{jk}^{g'}) - [1 - P_{g'}(x_{ik}^g)] \\
&- [1 - P_g(x_{jk}^{g'})] + \mathcal{H}_{gg'}\Big\}
\end{aligned}$$

$$\begin{aligned}
\bar{D}_{gg} &= \mathcal{H}_{gg} + \frac{2}{n_g}\sum_{i=1}^{n_g}\sum_{k=1}^{K}\frac{1}{K}\left\{[1 - P_g(x_{ik}^g)] - \mathcal{H}_{gg}\right\} \\
&+ \frac{1}{n_g(n_g-1)}\sum_{1\le i<j\le n_g}\sum_{k=1}^{K}\frac{1}{K}\Big\{I(X_{ik}^g \neq X_{jk}^g) - [1 - P_g(x_{ik}^g)] \\
&- [1 - P_g(x_{jk}^g)] + \mathcal{H}_{gg}\Big\}
\end{aligned}$$

Therefore,

$$D_n(B) = \frac{1}{n(n-1)}\left\{\sum_{g=1}^{G-1}\sum_{g'=g+1}^{G}\sum_{k=1}^{K}\frac{1}{K}n_g n_{g'}\Big[2\mathcal{H}_{gg'} - \mathcal{H}_{gg} - \mathcal{H}_{g'g'}\right.$$

$$
\begin{aligned}
&+ \quad \frac{2}{n_g}\sum_{i=1}^{n_g}\left(P_g(x^g_{ik}) - P_{g'}(x^g_{ik}) - \mathcal{H}_{gg'} + \mathcal{H}_{gg}\right)\\
&+ \quad \frac{2}{n_{g'}}\sum_{i=1}^{n_{g'}}\Big(P_{g'}(x^{g'}_{jk}) - P_g(x^{g'}_{jk}) - \mathcal{H}_{gg'} + \mathcal{H}_{g'g'}\Big)\\
&+ \quad \frac{2}{n_g n_{g'}}\sum_{i=1}^{n_g}\sum_{j=1}^{n_{g'}}\Big(\boldsymbol{I}(X^g_{ik} \neq X^{g'}_{jk}) - (1 - P_{g'}(x^g_{ik}))\\
&- \quad (1 - P_g(x^{g'}_{jk})) + \mathcal{H}_{gg'}\Big)\\
&- \quad \frac{2}{n_g(n_g-1)}\sum_{1\le i<j\le n_g}\Big(\boldsymbol{I}(X^g_{ik} \neq X^g_{jk}) - (1 - P_g(x^g_{ik}))\\
&- \quad (1 - P_g(x^g_{jk})) + \mathcal{H}_{gg}\Big)\\
&- \quad \frac{2}{n_{g'}(n_{g'}-1)}\sum_{1\le i<j\le n_{g'}}\Big(\boldsymbol{I}(X^{g'}_{ik} \neq X^{g'}_{jk}) - (1 - P_{g'}(x^{g'}_{ik}))\\
&- \quad (1 - P_{g'}(x^{g'}_{jk})) + \mathcal{H}_{g'g'}\Big)\Big]\Big\}
\end{aligned}
$$

And under H_0,

$$
\begin{aligned}
D_n(B) \quad &= \quad \frac{1}{n(n-1)}\Big\{\sum_{g=1}^{G-1}\sum_{g'=g+1}^{G}\sum_{k=1}^{K}\frac{1}{K}n_g n_{g'}\Bigg[\frac{2}{n_g n_{g'}}\sum_{i=1}^{n_g}\sum_{j=1}^{n_{g'}}\Big(\boldsymbol{I}(X^g_{ik} \neq X^{g'}_{jk})\\
&- \quad (1 - P_{g'}(x^g_{ik})) - (1 - P_g(x^{g'}_{jk}))\Big)\\
&- \quad \frac{2}{n_g(n_g-1)}\sum_{1\le i<j\le n_g}\Big(\boldsymbol{I}(X^g_{ik} \neq X^g_{jk}) - (1 - P_g(x^g_{ik})) - (1 - P_g(x^g_{jk}))\Big)\\
&- \quad \frac{2}{n_{g'}(n_{g'}-1)}\sum_{1\le i<j\le n_{g'}}\Big(\boldsymbol{I}(X^{g'}_{ik} \neq X^{g'}_{jk}) - (1 - P_{g'}(x^{g'}_{ik}))\\
&- \quad (1 - P_{g'}(x^{g'}_{jk}))\Big)\Big]\Big\}
\end{aligned}
$$

4 Decomposition of U-statistics and Test Statistic

Using the decompositions of U-statistics given in (12) we can write

$$
\begin{aligned}
\bar{D}_{gg'} &= \int\int \phi(\mathbf{x},\mathbf{y})dF_{n_g}(\mathbf{x})dF_{n_{g'}}(\mathbf{y})\\
&= \int\int \phi(\mathbf{x},\mathbf{y})d[F_g(\mathbf{x}) + (F_{n_g}(\mathbf{x}) - F_g(\mathbf{x}))]d[F_{g'}(\mathbf{y}) + (F_{n_{g'}}(\mathbf{y}) - F_{g'}(\mathbf{y}))]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathcal{H}_{gg'} + \int [1 - P_g(\mathbf{y})] d[F_{n_{g'}}(\mathbf{y}) - F_{g'}(\mathbf{y})] \\
&+ \int [1 - P_{g'}(\mathbf{x})] d[F_{n_g}(\mathbf{x}) - F_g(\mathbf{x}))] \\
&+ \int \int \phi(\mathbf{x}, \mathbf{y}) d[F_{n_g}(\mathbf{x}) - F_g(\mathbf{x}))] d[F_{n_{g'}}(\mathbf{y}) - F_{g'}(\mathbf{y})]
\end{aligned}
$$

and similarly,

$$
\begin{aligned}
\bar{D}_{gg} &= \mathcal{H}_{gg} + \int [1 - P_g(\mathbf{y})] d[F_{n_g}(\mathbf{y}) - F_g(\mathbf{y})] \\
&+ \int [1 - P_g(\mathbf{x})] d[F_{n_g}(\mathbf{x}) - F_g(\mathbf{x})] \\
&+ \int \int \phi(\mathbf{x}, \mathbf{y}) d[F_{n_g}(\mathbf{x}) - F_g(\mathbf{x}))] d[F_{n_g}(\mathbf{y}) - F_g(\mathbf{y})]
\end{aligned}
$$

Now, let $T_n = 2\bar{D}_{gg'} - \bar{D}_{gg} - \bar{D}_{g'g'}$. Then, under H_0,

$$
\begin{aligned}
T_n &= \int \int \phi(\mathbf{x}, \mathbf{y}) d[F_{n_{g'}}(\mathbf{y}) - F_{n_g}(\mathbf{y})] d[F_{n_g}(\mathbf{x}) - F_{n_{g'}}(\mathbf{x})] \\
&- \int \int \phi(\mathbf{x}, \mathbf{y}) d[F_{n_{g'}}(\mathbf{y}) - F_{n_g}(\mathbf{y})] d[F_g(\mathbf{x}) - F_{g'}(\mathbf{x})] \\
&+ \int \int \phi(\mathbf{x}, \mathbf{y}) d[F_{n_g}(\mathbf{x}) - F_g(\mathbf{x}))] d[F_g(\mathbf{y}) - F_{g'}(\mathbf{y})] \\
&- \int \int \phi(\mathbf{x}, \mathbf{y}) d[F_{n_{g'}}(\mathbf{y}) - F_{g'}(\mathbf{y}))] d[F_g(\mathbf{x}) - F_{g'}(\mathbf{x})]
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\int \int \phi(\mathbf{x}, \mathbf{y}) d[F_{n_{g'}}(\mathbf{y}) - F_{n_g}(\mathbf{y})] d[F_{n_g}(\mathbf{x}) - F_{n_{g'}}(\mathbf{x})] = \\
&= \frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} \sum_{d=1}^{D} \left(\frac{n_{kg'c}}{n_{g'}} - \frac{n_{kgc}}{n_g} \right) \left(\frac{n_{kgd}}{n_g} - \frac{n_{kg'd}}{n_{g'}} \right) \\
&- \frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} \left(\frac{n_{kg'c}}{n_{g'}} - \frac{n_{kgc}}{n_g} \right) \left(\frac{n_{kgc}}{n_g} - \frac{n_{kg'c}}{n_{g'}} \right) \\
&= \frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} \left(\frac{n_{kg'c}}{n_{g'}} - \frac{n_{kgc}}{n_g} \right)^2
\end{aligned}
$$

Analogously,

$$
\begin{aligned}
&\int \int \phi(\mathbf{x}, \mathbf{y}) d[F_{n_{g'}}(\mathbf{y}) - F_{n_g}(\mathbf{y})] d[F_g(\mathbf{x}) - F_{g'}(\mathbf{x})] = \\
&= -\frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} \left(\frac{n_{kg'c}}{n_{g'}} - \frac{n_{kgc}}{n_g} \right) (\pi_{kgc} - \pi_{kg'c}) ;
\end{aligned}
$$

$$
\begin{aligned}
&\int\int \phi(\mathbf{x},\mathbf{y})d[F_{n_g}(\mathbf{x})-F_g(\mathbf{x}))]d[F_g(\mathbf{y})-F_{g'}(\mathbf{y})] = \\
= \quad & \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\frac{n_{kgc}}{n_g}(\pi_{kg'c}-\pi_{kgc})-\mathcal{H}_{gg}+\mathcal{H}_{gg'}
\end{aligned}
$$

and

$$
\begin{aligned}
&\int\int \phi(\mathbf{x},\mathbf{y})d[F_{n_{g'}}(\mathbf{y})-F_{g'}(\mathbf{y}))]d[F_g(\mathbf{x})-F_{g'}(\mathbf{x})] = \\
= \quad & \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\frac{n_{kg'c}}{n_{g'}}(\pi_{kg'c}-\pi_{kgc})-\mathcal{H}_{gg'}+\mathcal{H}_{g'g'}
\end{aligned}
$$

Also, note that under H_0,

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}(\pi_{kg'c}-\pi_{kgc})^2 = 2\mathcal{H}_{gg'}-\mathcal{H}_{gg}-\mathcal{H}_{g'g'} = 0 \Rightarrow \\
\Rightarrow \quad & \pi_{kg'c}-\pi_{kgc} = 0 \Rightarrow \pi_{kg'c} = \pi_{kgc} \quad \forall\, c=1,\dots C \qquad (17)
\end{aligned}
$$

Therefore, under H_0,

$$
\begin{aligned}
T_n \quad = \quad & \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\left(\frac{n_{kg'c}}{n_{g'}}-\pi_{kg'c}\right)^2+\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\left(\frac{n_{kgc}}{n_g}-\pi_{kgc}\right)^2 \\
- \quad & 2\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\left(\frac{n_{kg'c}}{n_{g'}}-\pi_{kg'c}\right)\left(\frac{n_{kgc}}{n_g}-\pi_{kgc}\right)
\end{aligned}
$$

By (8), we have

$$
\begin{aligned}
D_n(B) \quad = \quad & \frac{1}{n(n-1)}\frac{1}{K}\sum_{k=1}^{K}\sum_{g=1}^{G}(n-n_g)n_g\sum_{c=1}^{C}\left(\frac{n_{kgc}}{n_g}-\pi_{kgc}\right)^2 \\
- \quad & \frac{1}{n(n-1)}\frac{1}{K}\sum_{k=1}^{K}\sum_{g\neq g'}n_gn_{g'}\sum_{c=1}^{C}\left(\frac{n_{kg'c}}{n_{g'}}-\pi_{kg'c}\right)\left(\frac{n_{kgc}}{n_g}-\pi_{kgc}\right) \\
= \quad & \frac{1}{n(n-1)}\frac{1}{K}\sum_{k=1}^{K}\left\{\sum_{g=1}^{G}(n-n_g)\mathbf{Z}'_{kg}\mathbf{Z}_{kg}-\sum_{g\neq g'}n_g^{1/2}n_{g'}^{1/2}\mathbf{Z}'_{kg'}\mathbf{Z}_{kg}\right\}
\end{aligned}
$$

where $\mathbf{Z}_{kg}=\sqrt{n_g}\left[\left(\frac{n_{kg1}}{n_g}-\pi_{kg1}\right),\dots,\left(\frac{n_{kgC}}{n_g}-\pi_{kgC}\right)\right]'$ and
$\mathbf{Z}_{kg'}=\sqrt{n_{g'}}\left[\left(\frac{n_{kg'1}}{n_{g'}}-\pi_{kg'1}\right),\dots,\left(\frac{n_{kg'C}}{n_{g'}}-\pi_{kg'C}\right)\right]'$

It is easy to see that as $n_g \to \infty$,

$$\mathbf{Z}_{kg} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{\Sigma}_{kg})$$

where $\mathbf{\Sigma}_{kg} = Diag(\boldsymbol{\pi}_{kg}) - \boldsymbol{\pi}_{kg}\boldsymbol{\pi}'_{kg}$ is a $C \times C$ matrix and
$\boldsymbol{\pi}_{kg} = (\pi_{kg1}, \ldots, \pi_{kgC})'$.

Also, since $\mathbf{Z}_{kg}$ is asymptotically $N(\mathbf{0}, \mathbf{\Sigma}_{kg})$,

$$\mathbf{Z}'_{kg}\mathbf{Z}_{kg} \approx \sum_{i=1}^{C} \lambda_i (\chi_1^2)_i, \tag{18}$$

where λ_i are the characteristic roots of $\mathbf{\Sigma}_{kg}$. In other words, the asymptotic distribution of $\mathbf{Z}'_{kg}\mathbf{Z}_{kg}$ is a linear combination of independent χ^2 random variables with 1 degree of freedom.

Further, we can write $\mathbf{Z}_k = (\mathbf{Z}'_{k1}, \mathbf{Z}'_{k2}, \ldots, \mathbf{Z}'_{kG})'$ as a $GC \times 1$ vector and $\mathbf{Z} = (\mathbf{Z}'_1, \mathbf{Z}_2, \ldots, \mathbf{Z}'_K)' = (\mathbf{Z}'_{11}, \ldots, \mathbf{Z}'_{1G}, \mathbf{Z}'_{21}, \ldots, \mathbf{Z}'_{2G}, \ldots, \mathbf{Z}'_{K1}, \ldots, \mathbf{Z}'_{KG})$ as a $CGK \times 1$ vector. Since the positions and groups are assumed to be independent,

$$\mathbf{Z}_k \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{\Sigma}_k) \quad \text{and} \quad \mathbf{Z} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma}_k = \mathbf{\Sigma}_{kg} \otimes \mathbf{I}_G$ and $\mathbf{\Sigma} = \mathbf{\Sigma}_{kg} \otimes \mathbf{I}_{KG}$ are block diagonal matrices of dimensions $CG \times CG$ and $CGK \times CGK$, respectively.

Alternatively, we can write

$$\begin{aligned}(n-1)D_n(B) &= \frac{1}{K}\sum_{k=1}^{K}\left\{\sum_{g=1}^{G}\left(1 - \frac{n_g}{n}\right)\mathbf{Z}'_{kg}\mathbf{Z}_{kg} - \sum_{g \neq g'} \frac{\sqrt{n_g}\sqrt{n_{g'}}}{n}\mathbf{Z}'_{kg'}\mathbf{Z}_{kg}\right\} \\ &= \frac{1}{K}\sum_{k=1}^{K}\mathbf{Z}'_k\mathbf{A}_n\mathbf{Z}_k.\end{aligned}$$

Note that $\mathbf{A}_n \to \mathbf{A} = [\mathbf{A}_{ij}]$, where

$$\mathbf{A}_{ij} = \begin{cases} (1 - p_i)\,\mathbf{I}_C, & i = j \\ -\sqrt{p_i p_j}\;\mathbf{I}_C, & i \neq j \end{cases}$$

and $n_i/n \to p_i$, $i = 1, 2, \ldots, G$.

Then,

$$(n-1)D_n(B) \longrightarrow \frac{1}{K}\sum_{k=1}^{K} \mathbf{Z}_k' \mathbf{A} \mathbf{Z}_k \approx \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{CG} \lambda_{ki}\left(\chi_1^2\right)_{ki}$$

where λ_{ki} are the characteristic roots of $\mathbf{A\Sigma}_k$. In other words, $\mathbf{Z}_k'\mathbf{A}\mathbf{Z}_k$ has as asymptotic distribution a linear combination of chi-square random variables with one degree of freedom.

In cases where the asymptotic results are not applicable, we can generate the empirical distribution of the test statistic $D_n(B)$ by resampling techniques such as the bootstrap. The empirical distribution can be generated under the null hypothesis of homogeneity among groups and then a p-value can be computed based on this distribution.

5 Application

The data set consists of mitochondrial sequences on the nucleotide level from several organisms. All the sequences can be downloaded from the address http://www.mitomap.org/MITOMAP/euk_mitos.html. The groups of sequences were defined according to taxonomic characteristics of the organisms. The proposed nonparametric test performance is illustrated on a comparison between *Homo sapiens* and a group of other primates, which include Gorillas, Chimpanzees, Orangutan and Gibbon specimens.

The original data (called **Case 1**) is composed of 87 mitochondrial sequences from the Catarrhini infraorder lineage, of which 58 are from the species *Homo sapiens* (called **Group 1**) and 29 sequences of other primates (called **Group 2**). From these 29, 28 are from the Hominidae family and one from the Hylobatidae family. The sequences were aligned using **BLAST2** (see *www.ncbi.nlm.nih.gov*) to a total of 438 positions located at the NADH dehydrogenase subunit 5 (ND5) gene. The null hypothesis of interest is that there is homogeneity between group 1 and group 2, i.e., $H_0 : 2\mathcal{H}_{12} = \mathcal{H}_{11} + \mathcal{H}_{22}$. The test statistic $D_n(B)$ is computed for the data set and its p-value is found via bootstrap resampling. We use two resample sizes, 2000 and 10000 for which little difference was found.

As a measure of scale, $D_n(B)$ is also computed on an artificially designed data set as follows. The original 58 sequences from group 1 were regrouped randomly into two subgroups of 29 sequences each. That procedure was performed for each resample size. They are respectively called **Case 2** and **Case 3** for 2000 and 10000 bootstraps.

Table 1 shows the observed values of $\bar{D}_n^{(o)}$, $D_n(B)$ and $D_n(W)$ for cases 1,2 and 3. Bootstrap p-values for Case 1 are **zero** for either 2000 or 10000 repetitions. Bootstrap p-values for Cases 2 and 3 are, respectively, 0.5615 and 0.6863.

Table 1: Observed values of the test statistics

Comparisons	$\bar{D}_n^{(o)}$	$D_n(B)$	$D_n(W)$
Case 1	0.0874	0.0348	0.0527
Case 2	0.0266	-1.183×10^{-5}	0.0267
Case 3	0.0266	-2.308×10^{-5}	0.0267

On Table 2 we show the quartiles of the empirical distributions. One can notice that $D_n(B)$ values are extremely small compared to the overall distances $\bar{D}_n^{(o)}$ in Case 1. Moreover, for all cases their bootstrap distributions are concentrated on the negative values which can be interpreted as purely random observations of very small numbers.

Table 2: Bootstrap quartiles of the empirical distributions

Cases (B)	$D_n(B)$ $(\times 10^{-6})$			$\bar{D}_n^{(o)}$		Bootstrap
	Q_1	Q_2	Q_3	Q_1	Q_3	p-value
Case 1 (2000)	-488.8	-180.0	310.0	0.0746	0.0969	$< 1/2000$
Case 1 (10000)	-493.2	-197.5	295.4	0.0745	0.0962	$< 1/10000$
Case 2 (2000)	-27.37	-7.152	27.87	0.0029	0.0490	0.5615
Case 3 (10000)	-28.21	-9.471	23.28	0.0030	0.0491	0.6863

By looking at these results, one concludes that the test detects differences between the *Homo sapiens* and other primates on the molecular level. Also, the null hypotheses of homogeneity between the two pseudo-groups of *Homo sapiens* cannot be rejected, since both p-values are greater than 0.5.

On the necessary resample size, some comments are in hand. For Case 1 its group difference is so strong that both p-values are **zero** (actually, $< 1/B$). For Cases 2 and 3 p-values are both very big (0.5615 for $B = 2000$ and 0.6863 for $B = 10000$) and they support the arbitrary choice of grouping. Moreover, a larger p-value for $B = 10000$ strengths the null hypothesis. Looking at Figure 1, one can also see that there is not much difference between the empirical distributions generated by 2000 or 10000 bootstraps for Case 1. Those results show no need of large resampling for such an example.

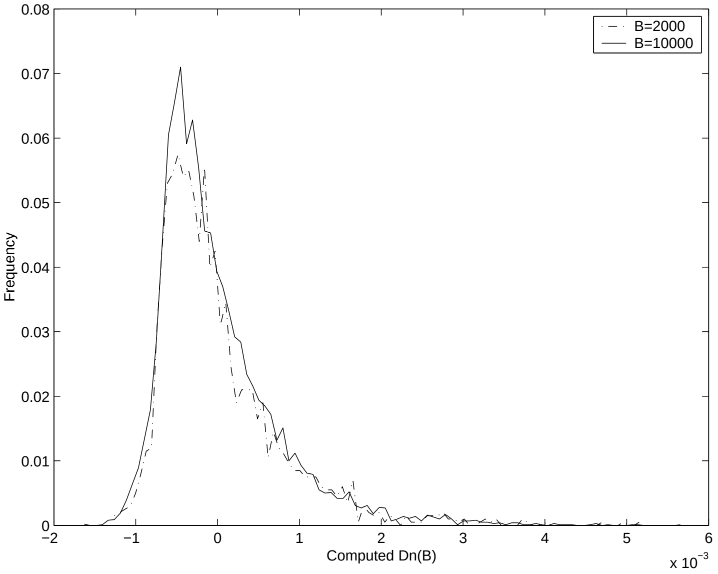

Figure 1: Empirical distributions under the null hypothesis of homogeneity between *Homo sapiens* and other primates.

6 References

Agresti, A. (1990). *Categorical Data Analysis.* Wiley, New York.

Andferson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Bahadur, R.R. (1961). A representation of the joint distribution of responses to n dichotomous items. In: H. Solomon, Ed., *Studies in Item Analysis and Prediction*, Stanford Univ. Press, California, 158-176.

Besag, J. (1974). Spatial interaction and the statistical analysis of life systems (with discussion). *J. Roy. Statist. Soc. B 48*, 192-236.

Chatterjee, S. K. and Sen, P. K. (1964). Nonparametric tests for the bivariate two-sample location problem. *Calcutta Statist. Assoc. Bull. 13*, 18-58.

Chernoff, H. and Lander, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomial is a single binomial. *Journal od Statistical Planning and Inference 43*, 19-40.

Chernoff, H. (1993). Kullback-Leibler information for ordering genes using sperm typing and radiation hybrid mapping. In: K. Matusita, M.L. Puri and T. Hayakawa, Ed., *Statistical Sciences and Data Analysis*, VSP, Zeist, The Netherlands, 1-12.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models for Proteins and Nucleic Acids.* Cambridge Univ. Press. UK.

Ewens, W. J. and Grant, G. R. (2001). *Statistical Methods in Bioinformatics: An Introduction*, Springer, New York.

Gini, C. W. (1912). Variabilita e mutabilita. *Studi Economico-Giuridici della R. Universita di Cagliari 3(2)*, 3-159.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist. 19*, 293 - 325.

Lange, K. (1997). *Mathematical and Statistical Methods in Genetic Analysis.* Springer, New York.

Lee, A. J. (1990). *U-Statisitcs - Theory and Practice.* Marcel Dekker, Inc.

Liang, K., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analysis for categorical data (with discussion). *J. Roy. Statist. Soc. B 54*, 3-40.

Light, R.H. and Margolin, B.H. (1971). An analysis of variance for categorical data. *J. Amer. Statist. Assoc. 66*, 534-544.

Pinheiro, H., Seillier-Moiseiwitsch, Sen, P.K. and Eron, J. (2000). Genomic sequence analysis and quasi-multivariate CATANOVA. In: P.K. Sen and C.R. Rao, Eds., *Handbook of Statistics, Volume 18 : Bioenvironmental and Public Health Statistics*, Elsevier, Amsterdam, 713-746.

Pinheiro, H., Seillier-Moiseiwitsch, and Sen, P.K. (2001). Analysis of variance for Hamming distances applied to unbalanced designs. Research Report No.30/01. Instituto de Matemática, Estatística e Computação Científica. Universidade Estadual de Campinas, SP, Brazil.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis.* Wiley, New York.

Sen, P. K. (1999). Utility-oriented Simpson-type indexes and inequality measures. *Calcutta Statist. Assoc. Bull. 49*, 1-22.

Sen, P. K. (2001). *Excursions in Biostochastics: Biometry to Biostatistics to Bioinformatics.* Lect. Notes, Academia Sinica Inst. Statist. Sci., Taipei, ROC.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications.* Chapmann-Hall, UK.

Simpson, E. H. (1949). The measurement of diversity. *Nature 163*, 688.

Waterman, M. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes.* Chapman-Hall, UK.