ELSEVIER

Journal of Microbiological Methods 26 (1996) 225–236

Journal of Microbiological Methods

# A graph-theory method to establish serological relationships within a bacterial taxon, with example from *Porphyromonas gingivalis*

Philippe Casgrain[a,*], Pierre Legendre[a], Jean-Louis Sixou[b], Christian Mouton[c]

[a]*Département de Sciences Biologiques, Université de Montréal, C.P. 6 128, succ. Centre-ville, Montréal, Québec, H3C 3J7, Canada*
[b]*Équipe de Biologie Buccale, Université de Rennes-1, 2, place Pasteur, F-35000 Rennes, France*
[c]*Groupe de Recherche en Écologie Buccale, Faculté de médecine dentaire, Université Laval, Sainte-Foy, Québec, G1K 7P4, Canada*

Received 30 August 1994; revised 22 April 1996; accepted 22 April 1996

**Abstract**

This paper develops a rationale for transforming serological data obtained by indirect immunofluorescence (IF) into a meaningful character-state matrix, and uses this matrix for numerical phylogenetic analysis. Typically, immunofluorescence data come in square asymmetrical matrices; columns correspond to strains used for adsorption and rows to strains used in the IF test. Such matrices can be decomposed into a symmetric and a skew-symmetric part. We first show that all pertinent biological information needed to reconstruct a phylogeny lies in the skew-symmetric component. Then we show how to transform the skew-symmetric matrix into a character-state tree, and how to obtain a binary character-state matrix from it. The character-state matrices obtained for different hyperimmune serum antibodies are assembled into a total character-state matrix, on which phylogenetic analysis is conducted. The data that motivated this methodological development concern *Porphyromonas gingivalis*, a major pathogen in adult periodontitis. Various proposals have been put forward in the literature, concerning the number of major serogroups found in this taxon. Six human and two animal strains of *P. gingivalis* were subjected to serotyping and to the phylogenetic analysis described above. Using a test of statistical significance recently developed to compare independently-obtained phylogenetic trees, or to compare hypotheses to trees, we show that our results best fit the hypothesis that there are three groups of serotypes, one animal and two human. Alternate hypotheses are not, or less strongly supported by our data. The algorithms developed to implement the new phylogenetic analysis method are presented in appendices.

*Keywords:* Graph theory; Immunofluorescence; Phylogenetic analysis; *Porphyromonas gingivalis*; Serology; Triple-permutation test (TPT)

## 1. Introduction

When doing serotyping studies, one is quickly submerged by the amount of collected data, since it increases as the cube of the number of strains. For instance, the complete set of homologous and heterologous adsorptions for 8 strains contains 512 data points; 12 strains contains 1728 and 30 contains 27 000! Therefore, extracting all the relevant information from this massive data set becomes a daunting task, which few scientists are likely to complete.

*Porphyromonas gingivalis* is considered a major

*Corresponding author. Department of Biology, University of Michigan, Ann Arbor, Michigan 48109-1057, USA (starting May, 1996). E-mail: casgrain@umich.edu, casgrain@ere.umontreal.ca


*PII* S0167-7012(96)00911-6

pathogen in adult periodontitis; for an up-to-date review see [1]. Among *P. gingivalis* isolates, antigenic heterogeneity was demonstrated by several serotyping studies. The proposed numbers of serogroups are two [2–5], three [6,7] or four [8,9]. Even if these studies show considerable disagreement, they did bring valuable information. One of them [5] formally recognized the basic separation of animal and human biotypes; others have recognized the distinction between strains of human origin that are virulent or non-virulent in an experimental model. Recent phenotypic [10] and genotypic [11,12] studies also recognized the human-animal dichotomy.

The purpose of the present paper is to develop a rationale for transforming indirect immunofluorescence (IF) serological data into a meaningful character-state matrix, and to show that this matrix can be used as the basis for numerical phylogenetic studies. *P. gingivalis* will be our case study.

Since the actual number of serogroups in *P. gingivalis* is still a subject of controversy, we will also show how different hypotheses found in the literature can be confronted to the results of our phylogenetic analysis, using a recently-developed statistical test for comparing additive trees (phylogenies).

## 2. Materials and methods

### 2.1. Biological protocol

Eight strains, two of animal origin and six human isolates, were selected from our collection to encompass a large biological and geographical diversity. All strains were identified to species level using conventional physiologic and biochemical criteria, and the identification was confirmed through gas-liquid chromatography analysis of the cellular fatty acids [10]. Hyperimmune serum to each strain was produced by intravenous injection of formalinized whole cells in rabbits, as described by [13] and [5]. These aliquots of the hyperimmune sera were processed through homologous (same strain as injected to the rabbit) and heterologous (some other strain) adsorption. The number of reactions performed for heterologous adsorption was determined by the number of adsorptions needed to extinguish the reactivity by homologous adsorption. Adsorption of each serum on bacteria was performed by mixing a volume of serum, diluted 1:50, with an equal volume of bacterial suspension at 109 formalin-killed cells/ml, followed by incubation for 60 min at 37°C, then 12 h at 4°C. The samples were centrifuged at 5000 × *g* for 15 min, the pelleted bacteria were discarded, and the efficiency of the adsorption was evaluated by IF.

Eight matrices corresponding to the eight rabbit antisera were thus generated. In each one, the columns correspond to the strains used for the 'adsorbed sera', while the rows represent strains tested by IF. The IF results were recoded as '0' for a negative reaction, '1' for a variable, weak reaction (inconsistently-reproducible results), and '2' or '3' for serological reactions of increasing intensity; this is a recoding of the reaction classes of [13]. Our data matrices are presented in Table 1.

In a serological study such as this one, the objective is to determine how closely the various strains are related to one another in terms of their surface antigens. The adsorption on homologous or heterologous bacteria removes the antibodies that have avidity to the bacterial strain originally used to produce the hyperimmune serum. After centrifugation, there remains only, in the supernatant, those antibodies that do not have enough avidity for the surface antigens of the adsorbing strain to bind with it. The IF test shows the intensity of the serological reaction between the residual antibodies in the adsorbed serum and the test strains.

If the laboratory procedure has been carefully carried out, it is expected that a serum processed through homologous adsorption should produce a negative IF reaction on the homologous strain. Moreover, none of the other strains tested by IF should react with this adsorbed serum. Therefore, each of our matrices should have zeroes in its diagonal and in one of its columns. The numerical analysis described below does not take the diagonal into account; it is merely used as a quality control.

### 2.2. Biological information needed to construct the skew-symmetric matrix

Assume that an antigen is characteristic of the surface of a given bacterial strain. A bacterial cell of that strain will bind the corresponding antibody in

Table 1
Immunofluorescence data matrices. Matrix names refer to the hyperimmune sera (the criteria). Matrix columns correspond to the strains used for adsorption, and rows to strains in the IF test

| anti-ATCC 49417 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H |
| A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| B | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| H | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| anti-ATCC 33277 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H |
| A | 0 | 0 | 0 | 2 | 2 | 2 | 3 | 2 |
| B | 3 | 0 | 2 | 3 | 2 | 3 | 3 | 3 |
| C | 2 | 0 | 1 | 3 | 2 | 3 | 3 | 3 |
| D | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 3 |
| E | 3 | 0 | 3 | 2 | 0 | 3 | 2 | 3 |
| F | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| G | 2 | 0 | 3 | 2 | 2 | 2 | 2 | 2 |
| H | 2 | 0 | 1 | 2 | 2 | 2 | 2 | 0 |

| anti-16.1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H |
| A | 0 | 2 | 0 | 2 | 0 | 3 | 2 | 2 |
| B | 2 | 0 | 0 | 2 | 1 | 2 | 3 | 3 |
| C | 3 | 3 | 0 | 3 | 3 | 3 | 3 | 3 |
| D | 1 | 3 | 0 | 0 | 3 | 3 | 3 | 2 |
| E | 3 | 2 | 0 | 3 | 0 | 3 | 3 | 3 |
| F | 0 | 2 | 0 | 2 | 0 | 0 | 2 | 2 |
| G | 2 | 1 | 0 | 1 | 1 | 1 | 0 | 2 |
| H | 2 | 3 | 0 | 2 | 2 | 3 | 2 | 1 |

| anti-A7A1-28 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H |
| A | 0 | 3 | 2 | 0 | 2 | 2 | 3 | 3 |
| B | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 2 |
| C | 1 | 2 | 0 | 0 | 2 | 2 | 2 | 3 |
| D | 3 | 3 | 3 | 0 | 3 | 3 | 3 | 3 |
| E | 0 | 3 | 2 | 0 | 0 | 1 | 2 | 2 |
| F | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| G | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| H | 1 | 2 | 1 | 0 | 1 | 1 | 1 | 0 |

| anti-17A3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H |
| A | 1 | 3 | 0 | 1 | 1 | 1 | 3 | 2 |
| B | 2 | 0 | 2 | 2 | 0 | 2 | 2 | 2 |
| C | 2 | 1 | 0 | 3 | 8 | 3 | 2 | 2 |
| D | 0 | 0 | 8 | 0 | 2 | 8 | 0 | 2 |
| E | 3 | 2 | 2 | 2 | 0 | 2 | 3 | 3 |
| F | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| G | 3 | 2 | 2 | 2 | 0 | 2 | 8 | 2 |
| H | 2 | 2 | 2 | 2 | 2 | 3 | 0 | 0 |

| anti-W50 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H |
| A | 0 | 2 | 1 | 0 | 1 | 0 | 2 | 1 |
| B | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| C | 2 | 2 | 0 | 1 | 0 | 0 | 3 | 3 |
| D | 2 | 2 | 0 | 0 | 0 | 0 | 2 | 2 |
| E | 2 | 2 | 0 | 1 | 0 | 0 | 2 | 2 |
| F | 3 | 3 | 3 | 2 | 2 | 0 | 3 | 3 |
| G | 2 | 2 | 2 | 2 | 0 | 2 | 1 | 2 |
| H | 2 | 2 | 0 | 0 | 0 | 0 | 1 | 0 |

| anti-T22* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H |
| A | 0 | 2 | 0 | 1 | 2 | 2 | 0 | 3 |
| B | 0 | 0 | 0 | 2 | 1 | 8 | 0 | 2 |
| C | 2 | 2 | 0 | 2 | 0 | 3 | 0 | 3 |
| D | 2 | 2 | 1 | 0 | 0 | 2 | 0 | 2 |
| E | 2 | 2 | 0 | 2 | 0 | 2 | 0 | 3 |
| F | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 2 | 2 | 2 | 2 | 0 | 2 | 0 | 2 |
| H | 3 | 3 | 3 | 3 | 2 | 3 | 1 | 3 |

| anti-Wolf 1.1* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H |
| A | 0 | 1 | 0 | 1 | 2 | 2 | 0 | 0 |
| B | 0 | 0 | 0 | 2 | 0 | 2 | 0 | 0 |
| C | 3 | 3 | 0 | 3 | 2 | 3 | 2 | 0 |
| D | 1 | 1 | 0 | 0 | 1 | 2 | 0 | 0 |
| E | 2 | 0 | 0 | 2 | 0 | 2 | 0 | 0 |
| F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 2 | 1 | 2 | 2 | 0 | 2 | 0 | 0 |
| H | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 0 |

Letter codes: A, strain ATCC 49417; B, ATCC 33277; C, 16.1; D, A7A1 28; E, 17A3; F, W50; G, T22*; H, Wolf 1.1*. Numerical codes for the serological reactions: 0, no reaction; 1, faint; 2, some reaction, close to positive reaction threshold; 3, positive reaction; 8, missing data. All strains represent human *P. gingivalis* isolates, except (*) which are animal isolates: T22 from a cynomolgus monkey, and Wolf 1.1 from a wolf.

the adsorption process. Centrifugation should pellet all bacterial cells and the bound antibodies to the bottom of the test tube and therefore deplete the hyperimmune serum of these specific antibodies, provided that bacterial cells are added in sufficient number.

There are two cases where one should expect a negative IF reaction. The first one occurs when a strain is reacted with the serum previously processed through homologous adsorption; the corresponding data column will be filled with zeroes.

The second case is observed when the strain used in the IF test was also used for heterologous adsorption. The matrix diagonals should also be filled with zeroes.

Four other, more informative cases can be described, considering two bacterial strains 'A' and 'B', they are summarized in Table 2.

In the first case, strain 'A' (columns of Table 1), used to create the adsorbed serum (the 'criterion'), shares a large set of antigens with the immunizing strain 'B', i.e. the two strains are closely antigenically related. The heterologous adsorption therefore leaves little antibodies for the IF-tested strain B (rows of Table 1) to react with. The expected IF reaction (Col A/Row B) is weak. Conversely, when B has little antigenic relatedness with A, there is plenty of antibodies left for the now IF-tested strain A to react with and the IF reaction (Col B/Row A) is strong. Both observations indicate that, from a serological point of view, A has more antigenic determinants than B; so one can express that relationship using either set theory: A contains B, or directed graphs: A→B.

Reversing the names of the strains (B sharing a

Table 2
Expected IF reactions for various criterion (antibody) avidities, according to strain relatedness

| Case no. | Relatedness with immunizing strain | | Expected IF reaction in matrix | |
|---|---|---|---|---|
| | A | B | Col A/Row B | Col B/Row A |
| 1 | High | Low | Weak | Strong |
| 2 | Low | High | Strong | Weak |
| 3 | High | High | Weak | Weak |
| 4 | Low | Low | Weak | Weak |

large set of antigens with the criterion and A sharing little) corresponds to the second case, B→A, which is the opposite of the first.

The third and fourth cases are obtained when A and B both share either a large or very little set of antigens with the criterion. In these two cases, it is easy to show that the IF reaction should be weak at all times, either (a) because there is very little left to react with, or (b) because the strain used in the IF test has only a distant relationship with the criterion. We code A = B when one cannot say which of the two strains has the larger antigenic repertoire.

### 2.3. The skew-symmetric matrix

The square data matrices in Table 1 have the property that corresponding cells above and below the diagonal do not necessarily contain the same value; for that reason, they are called non-symmetric. Matrix algebra tells us that any non-symmetric matrix can be expressed as the sum of two derived matrices, one symmetric and one skew-symmetric, without loss of information [14]. Consider for instance the two numbers 0 and 3, found in opposite positions (1,2) and (2,1) of matrix anti-ATCC 33277 of Table 1 and Fig. 1a. The symmetric part is obtained by averaging these two numbers: $(0 + 3)/2 = 1.5$; the skew-symmetric part is obtained by subtracting one from the other and dividing by 2: $(0 - 3)/2 = -1.5$ and $(3 - 0)/2 = +1.5$. So, when the symmetric and skew-symmetric parts are added, the result is the original matrix: $1.5 - 1.5 = 0$ for the upper original number, and $1.5 + 1.5 = 3$ for the lower one. Using letters instead of numbers, one would obtain a simple algebraic proof of the additivity of the symmetric and skew-symmetric components, so that we can write the matrix equation:

$$M = {}_{sym}M + {}_{skew}M$$

Consider matrix M to be our anti-ATCC 33277 matrix (Fig. 1a) and construct the skew-symmetric matrix ${}_{skew}$M (Fig. 1b). The upper-triangular part of ${}_{skew}$M contains all the information needed to compare strains: from Table 2, one can see that subtracting a strong IF reaction from a weak one can be expressed as a positive number, and corresponds to the case where the row-component of the matrix has greater complementarity to the criterion than the column, and vice versa with negative numbers. Alternatively, one can use either the negative numbers in the whole matrix (as in the next paragraph), or the positive numbers only.

One can represent the signs of the numbers in ${}_{skew}$M by a directed graph. If there is a negative number at the intercept of row A and column B, this can be interpreted as B→A. If that number is positive, it means that there is a negative number at the intercept of row B and column A and A→B. If it is zero, then A = B.

This reasoning shows that the most important information in evolutionary terms is found in the asymmetry of the numbers in the original biological data matrix; this information is now concentrated in the skew-symmetric matrix, which we will analyze in more detail.

### 2.4. Transformation of the skew-symmetric matrix into a path matrix (directed graph)

Matrix ${}_{skew}$M can be transformed into a path matrix ${}_{path}$M, which is a binary matrix filled with only 0s and 1s, following this simple rule: put a '1' at the intercept of row A and column B if and only if B→A, or in other words if there is a negative number in that position in ${}_{skew}$M; otherwise put a '0'. The B→A (or A→B) information is read directly from ${}_{skew}$M. The path matrix extracted from ${}_{skew}$M in Fig. 1b is shown in Fig. 1c.

This path matrix can be used to create a graphical representation of the matrix as a directed graph where each edge corresponds to a '1' in the matrix.

### 2.5. Elimination of non-independent information

One problem posed by path matrices is that they contain redundant information. Consider the ${}_{path}$M matrix in Fig. 1c: one can clearly see that B→C and C→A. Following the directed graph equivalent, it follows that B→A. Thus, the information B→A is redundant, or non-independent; it can be deduced from the data at hand. However, the B→A information is present in the ${}_{path}$M matrix. Therefore, our next task is to remove the non-independent information from ${}_{path}$M.

This is accomplished through a two-step algo-

(a) M =

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| A | 0 | 0 | 0 | 2 | 2 | 2 | 3 | 2 |
| B | 3 | 0 | 2 | 3 | 2 | 3 | 3 | 3 |
| C | 2 | 0 | 1 | 3 | 2 | 3 | 3 | 3 |
| D | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 3 |
| E | 3 | 0 | 3 | 2 | 0 | 3 | 2 | 3 |
| F | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| G | 2 | 0 | 3 | 2 | 2 | 2 | 2 | 2 |
| H | 2 | 0 | 1 | 2 | 2 | 2 | 2 | 0 |

(b) $_{skew}$M =

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.0 | −1.5 | −1.0 | 1.0 | −0.5 | 1.0 | 0.5 | 0.0 |
| 1.5 | 0.0 | 1.0 | 1.5 | 1.0 | 1.5 | 1.5 | 1.5 |
| 1.0 | −1.0 | 0.0 | 1.5 | −0.5 | 1.5 | 0.0 | 0.0 |
| −1.0 | −1.5 | −1.5 | 0.0 | 0.0 | 1.0 | 0.0 | 0.5 |
| 0.5 | −1.0 | 0.5 | 0.0 | 0.0 | 1.5 | 0.0 | 0.5 |
| −1.0 | −1.5 | −1.5 | −1.0 | −1.5 | 0.0 | 0.0 | 0.0 |
| −0.5 | −1.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | −1.5 | 0.0 | −0.5 | −0.5 | 0.0 | 0.0 | 0.0 |

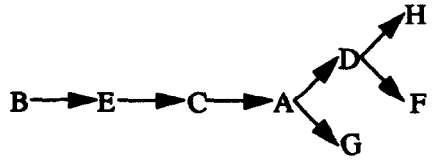
(c) $_{path}$M =

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |

(d) $_{path}$M =

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

(e)




(f)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| A | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| B | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| C | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| D | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| E | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| F | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| H | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Fig. 1. (a) Example of an immunofluorescence matrix and (b) the associated skew-symmetric matrix. (c) Path matrix derived from the skew-symmetric matrix. (d) Path matrix trimmed of redundant information, and (e) associated directed graph. (f) Character-state matrix, as read from the directed graph. There are as many rows as there are strains, and as many columns as there are edges (arrows) in the graph (e).

rithm, detailed in Appendix A. Briefly, step 1 creates a 'fat' path matrix copy of $_{path}$M, containing all the possibly redundant information it can find: if B→C and C→A then B→A; add this to the fat path matrix if it is not already there. Step 2 compares this copy to $_{path}$M to remove the non-independent information, including the one that was present before the algorithm started. What is left is a path matrix trimmed of all redundant information. It is shown in Fig. 1d, along with its associated directed graph (Fig. 1e); each '1' in the trimmed matrix $_{path}$M translates into an arrow of the graph. The arrow C→A, for instance, is drawn to mean that C has a larger antigenic repertoire than A.

Why is it important to remove the redundant information? Biologically, this is easy to understand. If B has more antigenic determinants than C, and C has more than A, then of course B has more than A; the antigenic order is transitive. Also, from a statistical point of view, if one keeps the B→A relationship (the '1' in row B, column A) in the path matrix, one adds an extra edge to the directed graph that one is trying to construct, to serve as a summary of all the relationships described by the path matrix. Since, in the following section, we will use the directed graph edges as character states (one strain either 'has' or 'does not have' the character, which is a certain level of antigenic relatedness, depending on whether it is located upstream or downstream from the directed edge), we should be careful not to over-represent some relationships with extra edges which would only add meaningless characters. These extra characters would serve no purpose since they are encapsulated in the other, non-redundant edges of the directed graph. Therefore, they should be removed from the character matrix so that their presence does not impede the subsequent numerical analysis.

### *2.6. Translating arrows into a binary character state matrix*

Each edge of the directed graph, or 'arrow', can be construed as a character. Any strain located upstream from the arrow possesses that character, since it has more antigenic determinants than those downstream. Presumably, every upstream strain has either the character itself, like the strain from which the arrow originates, or a superset of that character (more determinants, including those defined by the character). One can therefore build a binary character-state matrix that has as many rows as there are strains and as many columns as there are arrows (Fig. 1f). For each arrow (column), a strain either 'has' (state 1, upstream from the arrow) or 'does not have' (state 0) the corresponding character.

One such matrix is built for every criterion. The final matrix that will be used for phylogenetic analysis is simply a column-wise concatenation of all these matrices (Fig. 2).

To avoid the potential problem of manually drawing a directed graph and reading on it which strain is upstream or downstream from each edge, which would possibly generate coding errors, a recursive procedure has been developed (Appendix B. This algorithm reads a path matrix and achieves that goal effortlessly.

### *2.7. Phylogenetic analysis of the character-state matrix*

Such a matrix is typical of those used in cladistic analysis [15]. We used program *Mix* from the *PHYLIP* package [16] to create a cladogram representing the phylogenetic relationships among bacterial strains. The Camin-Sokal rooting criterion [17], available in that program, was used to root the tree. This criterion specifies that once state '1' of a character has been attained, it is highly unlikely that the character would revert to state '0'. Therefore, it is legitimate to speculate that a hypothetical ancestor to all Operational Taxonomic Units (OTUs), in our case *P. gingivalis* strains, should have state '0' for all characters. This effectively roots the phylogenetic tree and one no longer looks at an undirected evolutionary network, but at a directed tree. The *Mix* program implements a parsimony analysis method and tries to find the tree requiring the smaller number of mutations from state 0 to state 1 to reach the character states effectively observed in the OTUs. This method was chosen here because it seems a sensible one in the present case. Note, however, that the coding procedure for IF data described in the previous paragraphs is in no way linked to that decision.

### *2.8. The Triple-Permutation Test (TPT)*

There is a way to test the goodness-of-fit between a hypothesis and an actual cladogram. This is the Triple-Permutation Test (TPT), developed by [18]. The hypotheses to be tested in the present study are the proposals of the various authors, mentioned in Section 1, as to the appropriate division of *P. gingivalis* isolates into serogroups. TPT evaluates the null hypothesis that the additive trees (cladograms) under comparison — here, an actual parsimony analysis result on the one hand, and a tree depicting a hypothesis on the other — are no more similar than random additive trees with random topology, randomized labels, and randomized branch lengths. The

goodness-of-fit index used in the present study is the squared correlation coefficient ($r^2$), and the associated permutational probability, computed between the matrices representing these two trees, is obtained by the permutational method detailed in [19]. In each test, the tree representing the theoretical model was kept fixed, while the parsimony analysis tree was permuted. TPT was computed using program *Permute! 3.0* written by P.C. and available from the corresponding author.

## 3. Results

The directed graphs obtained through the procedure described above are presented in matrix form in Fig. 2; we have shown above that these matrices are equivalent to series of directed graphs. Strain names refer to the eight directed graphs that were extracted from the eight antiserum matrices. The phylogeny derived from the character-state coding table (Fig. 2), using the *Mix* parsimony program, is shown in Fig. 3.

This tree reveals three clades, one animal and two human. The animal clade is 8 mutations away from the root — the animal strains have 8 mutations in common — and thus the animal clade appears more homogeneous than the two human clades, which are respectively distant by 3 and 6 mutations from the root. The 3 mutations common to W50 and A7A1-28 are due exclusively to the anti-W50 antiserum.

Comparison of these results to a two-serogroup (human, animal) division using the TPT method yielded an $r^2$ of 0.0252 ($P = 0.456$), while comparison to a three-serogroup (human-A 'virulent', human-B 'non-virulent', animal) [6] model produced an $r^2$ of 0.2340 ($P = 0.0051$). We could not compare our results to the four-group model of [8,9], since the fourth group was comprised of strains not present in our study. Of the three remaining groups, however, [8] and [1] tended to place ATCC 33277 into the W50-A7A1-28 cluster. In that case, comparison of that three-group classification to our phylogeny produced an $r^2$ of 0.200 ($P = 0.008$).

## 4. Discussion

The phylogenetic tree in Fig. 3 shows that the *P. gingivalis* taxon is heterogeneous. It confirms the previously proposed separation between animal and human isolates [5]. It reveals that group human-A consists of two strains (W50 and A7A1-28) that have been associated with induction of pathogenic reactions in animal models [20–22], the so-called 'virulent' strains. We cannot state that group human-B represents 'non-virulent' strains, since ATCC 49417 is a known virulent strain (strain RB22D1 in [22]). Whether this illustrates the heterogeneity of the *P. gingivalis* taxon or a flaw in the method is not known at the moment.

Since our phylogenetic tree was rooted by a

ATCC 49417
ATCC 33277

| | | 16.1 | A7A1-28 | 17 A3 | W50 | T22* | Wolf 1.1* |
|---|---|---|---|---|---|---|---|
| A | 0 0010111 | 000010000 | 0101000 | 010101000 | 010000000000 | 010001000 | 000000010000 |
| B | 0 1111111 | 000000000 | 0000000 | 000101000 | 000000000000 | 000000000 | 000000001000 |
| C | 0 1010111 | 111111111 | 0000100 | 110101000 | 001000010100 | 110001000 | 101010111110 |
| D | 0 0000101 | 010001000 | 1111111 | 000000000 | 110000000010 | 000000100 | 000000000100 |
| E | 0 1110111 | 100010101 | 0100000 | 111111111 | 000100001001 | 001010110 | 000000000010 |
| F | 0 0000000 | 000000000 | 0000000 | 000000000 | 111111111111 | 000000000 | 000000000000 |
| G | 0 0000000 | 000000000 | 0000000 | 000010101 | 000000000000 | 111111101 | 010001010100 |
| H | 0 0000000 | 000000100 | 0000000 | 000000100 | 000000000000 | 001010100 | 111111111111 |

Fig. 2. Directed-path matrices, that form together the character-state matrix used in the cladistic analysis program to reconstruct the phylogeny. Strain names above the sections refer to the directed graphs that were extracted from the eight antiserum matrices (Table 1). All strains represent human *P. gingivalis* isolates, except (*) which are animal isolates.

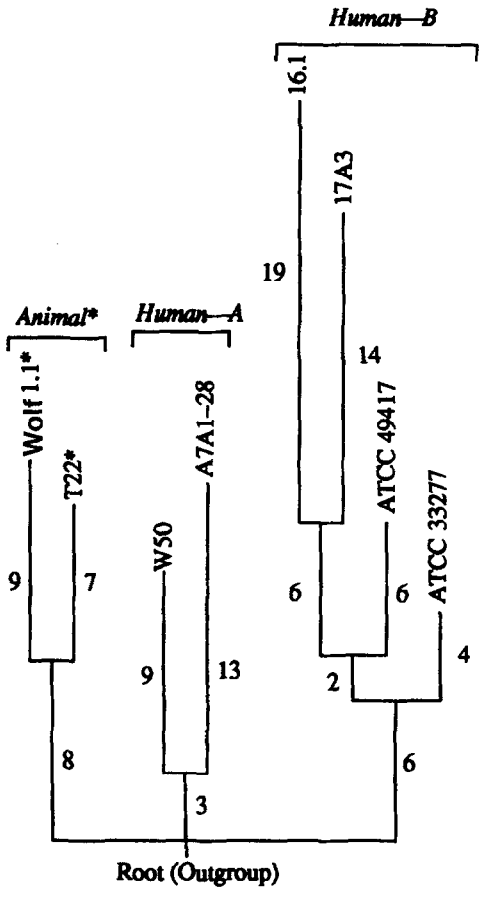



Fig. 3. Additive tree representing the relationships among the eight bacterial strains. Total tree length: 106 mutations. Numbers on the branches indicate branch lengths, in numbers of mutations; branch lengths are drawn to be proportional to these numbers of mutations. All strains represent human *P. gingivalis* isolates, except (*) which are animal isolates.

hypothetical ancestor with state '0' for all characters, our results indicate that the three groups are equally distant from the common hypothetical ancestor, meaning that the two human groups are no more related to one another than they are to the animal group.

Using the Triple Permutation Test of [18], we compared the patristic distance matrix (associated with the additive tree of Fig. 3) to a theoretical distance matrix representing a perfect trichotomy comprising one animal and two human groups. This trichotomy was deduced from [2]: 2 human groups, virulent and non-virulent; and [5]: an animal group distinct from the human strains. We found that there is a very low probability (0.0051, or 0.51%) that the observed tree of Fig. 3 does not match this perfectly trichotomous tree. (The variant model where strain ATCC 33277 is put in the human-virulent group agrees just as well with our derived phylogeny. To test this model properly, however, one should obtain serological data for the fourth *Porphyromonas* group, run the data in our analysis, and compare the experimental tree to the 4-group hypothesis). When we applied the same comparison to a simple animal-human dichotomy, as in [5], the probability rose to 0.456 (or 45.6%), indicating that the data do not support this hypothesis. Therefore, the biological reasoning now has statistical support: there are at least two human serogroups, as outlined in [2]. Whether the human-B cluster can be segregated into two or more different groups remains to be seen. The data presented here is not conclusive on that matter, although a marked differentiation is observed between (ATCC 49417-ATCC 33277) and (17A3-16.1).

The Camin-Sokal parsimony method used to reconstruct our phylogenetic tree is one that has been widely used by phylogeneticists [15,17,23]. We used it here because we felt it was appropriate to the problem at hand, but needless to say, the asymmetric matrix analysis described in Fig. 1, which leads to the character-state matrix of Fig. 2, is independent of the choice of a particular phylogenetic-tree reconstruction method. It would be possible, for instance, to develop another reconstruction method that would account for the intensity of the directed graph's relationships, instead of the simple dominance data that we used (A→B or A←B).

We will now try to ascertain the serological interest of the proposed reconstruction method. Its simplicity is appealing, but that in itself is no guarantee of success. The following facts indicate, however, that the method can be considered robust for this type of analysis. First, it was built specifically for studying the serological characteristics of bacteria using antigen-antibody reaction. Both the directed graphs and the skew-symmetric matrices have been shown to retain all the pertinent phylogenetic information contained in the original data matrices. Second, the decomposition of the graphs

into a series of binary vectors (characters) also stems from a biological observation: a bacteria either possesses, or not, a particular antigenic structure, allowing it to bind with an antibody. Third, this sequence of characters is ideal for cladistic reconstruction methods; for a complete review, see [15]. Synthesizing the biological information into binary characters allows one to easily construct a phylogenetic binary tree, much in the same way as one can build a binary identification key.

This method is not based on empirical observations nor on the skill of the observer, as were the classification methods used before the advent of numerical taxonomy. Instead, we used the biologist's knowledge to build a mathematical tool that mimics the biological process of antigenic repertoire comparison, increasing its understandability and demonstrability, unlike other methods used before [3–8].

## 5. Conclusion

The results obtained using the new graph-theory method, proposed in this paper to establish serological relationships among bacterial strains, support the existence of at least two serogroups among the human *Porphyromonas gingivalis* strains, and stress the differentiation of the animal group. This trichotomy is inconsistent with the dichotomy of animal and human strains proposed by [5], but this could be due to the number of strains used in each study. The method was developed with specific biological objectives in mind, but it appears that the need for analysis of asymmetrical matrices is much greater than for bacterial classification alone. Therefore, this new method should appeal to the growing number of researchers who produce asymmetrical matrices but have so far been unable to analyze them adequately.

## Acknowledgments

We wish to thank Deirdre Ni Eidhin for her assistance. This investigation was supported in part by MRC Grant MA-8761

## Appendix A

### The algorithms

#### List of declarations

```
const {Constants used in the program}
MaxSpecies = 10; {The maximum number of OTUs that can be read by the program}
MaxCharacters = 200; {For the output PHYLIP matrix}
type {Definition of vectors and matrices}
BinVector = array[1..MaxSpecies] of 0..1;
BinMatrix = array[1..MaxSpecies] of BinVector;
BinRectMatrix = array[1..MaxSpecies, 1..MaxCharacters] of 0..1;
var
pathMatrix:BinMatrix; {Data matrix as read from the skew-symmetric matrix}
fatPathMatrix:BinMatrix; {Initially empty, see Appendix A}
characterStateMatrix:BinRectMatrix; {The output matrix, used for phylogenetics}
numTax, numCharacters:Integer; {Actual number of OTUs and characters in the output}
```

#### Simplifying the path matrix

This procedure is an iterative, two-stage 'trimming' process that will eliminate 'shortcuts' in the directed graph from one label to the next. This simplification is necessary in order not to give too much weight to one character (antigenic relatedness) over the others.

Example — Suppose one has the following information:

A→B, A→C, A→E, B→C, C→D, D→E

Its matrix representation is:

| | *A* | *B* | *C* | *D* | *E* |
|---|---|---|---|---|---|
| *A* | 0 | 0 | 0 | 0 | 0 |
| *B* | 1 | 0 | 0 | 0 | 0 |
| *C* | 1 | 1 | 0 | 0 | 0 |
| *D* | 0 | 0 | 1 | 0 | 0 |
| *E* | 1 | 0 | 0 | 1 | 0 |

However, you will agree with me that the minimal path, or directed graph, needed to represent such information is in fact the following:

A→B→C→D→E

Therefore, there is too much weight given to some relationship. Witness the simpler matrix needed to represent the above path:

| | *A* | *B* | *C* | *D* | *E* |
|---|---|---|---|---|---|
| *A* | 0 | 0 | 0 | 0 | 0 |
| *B* | 1 | 0 | 0 | 0 | 0 |
| *C* | 0 | 1 | 0 | 0 | 0 |
| *D* | 0 | 0 | 1 | 0 | 0 |
| *E* | 0 | 0 | 0 | 1 | 0 |

There are two less paths: A→E and A→C. They are not needed, because they are already included in the graph if one 'follows the chain'. It is such a 'trimming' that this procedure attempts.

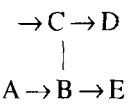
```
procedure SimplifyPathMatrix;
var
i, j, k, nbChanges:Integer;
pathVector:BinVector;
begin
{Step 1.
Take the Path Matrix and simplify it, i.e. eliminate redundant paths. But first, keep a copy of the Path Matrix!}
fatPathMatrix:=pathMatrix;
{This matrix will be 'stuffed' with all the ternary relationships (e.g. if A→B and B→C then A→C) so that the simplifying procedure will not need a recursive part, that left some 'orphans'.}
repeat
{One will repeat the 'stuffing' procedure until there are no longer any changes in the fat path matrix, that is, until one can no longer add a ternary relationship.}
nbChanges := 0;
for i := 1 to numTax do
for j := 1 to numTax do
if pathMatrix[i,j] = 0 then {There is no defined path... yet}
begin
{1.1) Copy the current line in a path vector}
pathVector := pathMatrix[i];
{1.2) Read the path vector. If one of its elements contains a '1', that is, a path, look at this element's row in the current column to find if there is an alternative path. If there is, add it in the current column.}
for k := 1 to numTax do
if pathVector[k] = 1 then
if pathMatrix[k,j] = 1 then
if fatPathMatrix[i,j] = 0 then
begin
fatPathMatrix[i,j] := 1;
nbChanges := nbChanges + 1;
end; {if fatPathMatrix[i,j] = 0 then}
end; {if pathMatrix[i,j] = 0 then}
until nbChanges = 0;
{Step 2.
Here is the path matrix simplification procedure.
One compares the actual paths (pathMatrix) to the complete paths (fatPathMatrix) to determine if there exist alternative paths that are shorter than the one currently studied. If there are, delete them (setting the value to '0' in the pathMatrix) because they are redundant. Note: one does not have to choose which path is to be deleted; it is always the shorter one because the graph is directed.}
for i := 1 to numTax do
for j := 1 to numTax do
if fatPathMatrix[i,j] = 1 then
begin
{2.1) Copy the current row in a path vector}
pathVector := fatPathMatrix[i];
{excluding the current column (the one that has a '1' into it).}
pathVector[j] := 0;
{Why exclude it? Because one knows that the label on the current column is pointing towards the label on the current row. One does not want to know what points towards the column label, but towards the row label.}
{2.2) Read the path vector. If one of its elements contains a '1', that is, a path, look at this element's row in the current column to find if there exists a longer path. If there is, remove the path in the current column (the simplified path)}
for k := 1 to numTax do
if pathVector[k] = 1 then
if fatPathMatrix[k,j] = 1 then
pathMatrix[i,j] := 0;
{Actually, one could break this loop and go a little faster, but since it would overcomplicate the code I will not do it.}
end; {if fatPathMatrix[i,j] = 1}
end; {procedure SimplifyPathMatrix}
```

## Appendix B

### Transforming a directed graph into a character-state matrix

This procedure 'reads' a path matrix to create a 'character state' matrix: each 'arrow' in the path state matrix is viewed as a character. Every label that (eventually) points to one arrow is said to 'possess' the character. Every other label does not.

Example: take the following path (directed graph):




Consider the arrow from A to B: only A is 'upstream' of the arrow, therefore only A possesses the character represented by this arrow. Likewise, the arrow from C to D represents a character possessed by A, B and C but not D (which is 'downstream') nor E (which is completely out of this path anyway).

The matrix representation of such a directed graph is, of course:

| | *A* | *B* | *C* | *D* | *E* |
|---|---|---|---|---|---|
| *A* | 0 | 0 | 0 | 0 | 0 |
| *B* | 1 | 0 | 0 | 0 | 0 |
| *C* | 0 | 1 | 0 | 0 | 0 |
| *D* | 0 | 0 | 1 | 0 | 0 |
| *E* | 0 | 1 | 0 | 0 | 0 |

This is the matrix that shall be analyzed by the 'CreateCharacterMatrix' procedure.

In layman's terms, this procedure reads the path matrix and when it finds an arrow (a '1'), it follows the path 'upstream', creating a list of every label it encounters during that process. When this is done, this list is the binary vector that tells who has the '1' state of the character and who has not.

```
procedure CreateCharacterMatrix;
var
i, j, k:Integer;
whoHasCurrentCharacter:BinVector;
begin
numCharacters := 0; {Number of characters in character state matrix}
{Initialize character state matrix: no label has state '1' of any character.}
for i := 1 to numTax do
for j := 1 to MaxCharacters do
characterStateMatrix[i,j] := 0;
{Read the path matrix until a '1' (arrow, path...) is encountered.}
for i := 1 to numTax do
for j := 1 to numTax do
if pathMatrix[i,j] = 1 then
begin
{When this happens, add one more character to the character matrix...}
numCharacters := numCharacters + 1;
if numCharacters > MaxCharacters then {error}
numCharacters := MaxCharacters;
{Initialize the 'found' vector, meaning the vector that will tell which label 'possesses' the character and which one does not. Initially, of course, no one possesses the character...}
for k := 1 to numTax do
whoHasCurrentCharacter[k] := 0;
{...no one but the current column number label, to which the arrow points to}
whoHasCurrentCharacter[j] := 1;
{Then call the recursive WhoPoints procedure to fill that binary search vector with the possession information.}
WhoPoints(pathMatrix[j], whoHasCurrentCharacter);
{Note the 'j' and not 'i': one has to find what points on the row that is currently pointing, and not what points on the row one is now analyzing.}
{Fill the character state matrix with the info from the binary search vector}
for k := 1 to numTax do
characterStateMatrix[k, numCharacters] := whoHasCurrentCharacter[k];
end; {if pathMatrix[i,j] = 1 then}
end; {procedure CreateCharacterMatrix}
```

{The following recursive procedure is used to find which label(s) 'point to', or are along a path to, a certain label. Simply put, when it finds that there is such an arrow, it calls itself to find out where the arrow came from. Its first argument is a path vector (a row of the path matrix), and its second argument, passed as 'var' because its value could be modified,

contains the information gathered so far regarding which labels were encountered during the path search.}

```
procedure WhoPoints (rowVect:BinVector; var returnVect:BinVector);
var
m:Integer;
recursVect:BinVector;
begin
for m := 1 to numTax do
if rowVect[m] = 1 then {There was an arrow from some label in column 'm'}
begin
{The label in column 'm' 'points to' the character that interests us: set it to '1'}
returnVect[m] := 1;
{Take the path vector in this label's row}
recursVect := pathMatrix[m];
{and call the procedure again to find other arrows}
WhoPoints(recursVect, returnVect);
end; {if rowVect[m] = 1 then}
end; {procedure WhoPoints}
```

## References


[1] Van Winkelhoff, A.J., Van Steenbergen, T.J.M. and De Graaff, J. (1993) Occurrence and association with disease. In: Biology of the species *Porphyromonas gingivalis* (Eds. H.N. Shah, D. Mayrand, and R.J. Genco). CRC Press, Boca Raton, Florida, USA, pp. 33–42.

[2] Fisher, J.G., Zambon, J.J., Chen, P. and Genco, R.J. (1986) *Bacteroides gingivalis* serogroups and correlation with virulence. J. Dental Res. 65, 816, abstract no. 817.

[3] Gmür, R. (1988) Applicability of monoclonal antibodies to quantitatively monitor subgingival plaque for specific bacteria. Oral Microbiol. Immunol. 3, 187–191.

[4] Fujiwara, T., Ogawa, T., Sobue, S. and Hamada, S. (1990) Chemical, immunobiological and antigenic characterizations of lipopolysaccharides from *Bacteroides gingivalis* strains. J. Gen. Microbiol. 136, 319–326.

[5] Parent, R., Mouton, C., Lamonde, L. and Bouchard, D. (1986) Human and animal serotypes of *Bacteroides gingivalis* defined by crossed immunoelectrophoresis. Infect. Immun. 51, 909–918.

[6] Fisher, J.G., Zambon, J.J. and Genco, R.J. (1987) Identification of serogroup-specific antigens among *Bacteroides gingivalis* components. J. Dental Res. 66, 222, abstract no. 927.

[7] Umemoto, T., Ishikawa, E., Watanabe, K., Hamada, M. and Iika, M. (1989) Serological classification of *Bacteroides gingivalis* and serogroup distribution in periodontitis patients. Dentistry Jpn. 26, 27–30.

[8] Nagata, A., Man-Yoshi, T., Sato, M. and Nakamura, R. (1991) Serological activities of *Porphyromonas (Bacteroides) gingivalis* and correlation with enzyme activity. J. Periodont. Res. 67, 1075–1080.

[9] Van Winkelhoff, A.J., Appelmelk, B.J., Kippuw, N. and De Graaff, J. (1993) K-antigens in *Porphyromonas gingivalis* are associated with virulence. Oral Microbiol. Immunol. 8, 259–265.

[10] Fournier, D. and Mouton, C. (1992) The animal reservoir is excluded for transmission to humans of *Bacteroides (Porphyromonas) gingivalis*. J. Dental Res. 71, 216, abstract no. 288.

[11] Loos, B.G., Mayrand, D., Genco, R.J. and Dickinson, D.P. (1990) Genetic heterogeneity of *Porphyromonas (Bacteroides) gingivalis* by genomic DNA fingerprinting. J. Dental Res. 69, 1488–1493.

[12] Ménard, C. and Mouton, C. (1993) Randomly amplified polymorphic DNA analysis confirms the biotyping scheme of *Porphyromonas gingivalis*. Res. Microbiol. 144, 445–455.

[13] Mouton, C., Hammond, P.G., Slots, J., Reed, M.J. and Genco, R.J. (1981) Identification of *Bacteroides gingivalis* by fluorescent antibody staining. Ann. Microbiol. 132B, 69–83.

[14] Ayres, F. Jr. (1962) Theory and problems of matrices. McGraw-Hill, New York, 218 p.

[15] Felsenstein, J. (1982) Numerical methods for inferring evolutionary trees. Q. Rev. Biol. 57, 379–404.

[16] Felsenstein, J. (1989) PHYLIP — PHYLogeny Inference Package (version 3.2) (version 3.5 currently used). Cladistics 5, 164–166.

[17] Camin, J.H. and Sokal, R.R. (1965) A method for deducing branching sequences in phylogeny. Evolution 19, 311–326.

[18] Lapointe, F.-J. and Legendre, P. (1992) A statistical framework to test the consensus among additive trees (cladograms). Syst. Biol. 41, 158–171.

[19] Legendre, P., Lapointe, F.-J. and Casgrain, P. (1994) Modeling brain evolution from behavior: a permutational regression approach. Evolution 48, 1487–1499.

[20] Neiders, M.E., Chen, P.B., Suido, H., Reynolds, H.S., Zambon, J.J., Schlossman, M. and Genco, R.J. (1989) Heterogeneity of virulence among strains of *Bacteroides gingivalis*. J. Periodont. Res. 24, 192–198.

[21] van Steenbergen, T.J.M., Delamarre, F.G.A., Namavar, F. and de Graaff, J. (1987) Differences in virulence within the species *Bacteroides gingivalis*. Antonie van Leeuwenhoek 53, 233–244.

[22] Grenier, D. and Mayrand, D. (1987) Selected characteristics of pathogenic and nonpathogenic strains of *Bacteroides gingivalis*. J. Clin. Microbiol. 25, 738–740.

[23] Sneath, P.H.A. and Sokal, R.R. (1973) Numerical Taxonomy. W.H. Freeman, San Francisco, 247 p.