Pattern Anal Applic (2011) 14:261–272
DOI 10.1007/s10044-011-0224-z

THEORETICAL ADVANCES

# Relational space classification for malaria diagnosis

Paolo Pintus · Maria Petrou

Received: 5 August 2009 / Accepted: 24 May 2011 / Published online: 11 June 2011

**Abstract** We present a study of sera derived from the malaria medical analysis of 189 subjects. The feature space is 18-dimensional and each serum is represented by a binary number. The subjects are divided into three different groups: no malaria, clinical malaria and asymptomatic subjects. We studied the main characteristics of the data and we selected 7 out of the 18 antigens as the most important for group discrimination. We propose a novel representation of the data in the so-called relational space, where the coded data of pairs of patients are plotted. We are able to separate the groups with 58% accuracy, about 15% points better than several conventional methods with which we compare our results.

**Keywords** Relational space · Malaria · Gene expression · Classification

## 1 Introduction

Malaria is one of the most common infectious diseases and an enormous public health problem. According to the World Health Organisation, “Half of the world’s population is at risk of malaria, and an estimated 247 million cases led to nearly 881 000 deaths in 2006” (from the World malaria report 2008 [20]). Some researchers evaluated that deaths are up to 50% higher than those reported by the World Health Organisation [15]. Snow et al. estimated that there were approximately 515 (range 300–660) million cases of *Plasmodium falciparum* malaria in 2002, while the number of deaths was between 700,000 and 2.7 million [15]. This represents at least one death every 30 s.

The vast majority of cases occur in children under the age of 5 and especially in Africa where 90% of malaria fatalities occur [2]. Despite efforts to reduce its transmission and increase treatment, there has been little change in those areas which are at risk by this disease since 1992 [10]. Indeed, if the prevalence of malaria stays on its present upwards course, the death rate could double in the next 20 years [2]. Precise statistics are unknown, because many cases occur in rural areas where people do not have access to hospitals or the means to afford health care. Consequently, the majority of cases are undocumented.

Malaria is caused by protozoan parasites of the genus *Plasmodium* [19]. Only four types of the plasmodium parasite can infect humans [19]. The most serious forms of the disease are caused by *Plasmodium falciparum* and *Plasmodium vivax*, but other related species (*Plasmodium ovale*, *Plasmodium malariae*) can also affect humans [19]. These human-pathogenic plasmodium species are usually referred to as malaria parasites and they are transmitted by female anopheles mosquitoes [19]. The parasites multiply within the red blood cells, causing symptoms that include symptoms of anaemia (light headedness, shortness of breath, tachycardia, etc.) as well as other general symptoms such as fever, chills, nausea, flu-like illness, and in severe cases, coma and death [19].

This work was done when he was visiting Imperial College as a student of the Leonardo da Vinci programme, ‘‘Bioinformatics and Bionanotechnology (BIO-NANO)’’, European project coordinated by Consorzio TUCEP, Perugia, Italy.

P. Pintus
Crim-Lab, Scuola Superiore di Studi Universitari e
Perfezionamento Sant’Anna, 56127 Pisa, Italy
e-mail: paolo.pintus@sssup.it

M. Petrou (✉)
Dept. of Electrical and Electronic Engineering,
Imperial College London, London SW7 2AZ, UK
e-mail: maria.petrou@imperial.ac.uk

No vaccine is currently available for malaria; preventative drugs must be taken continuously to reduce the risk of infection (http://www.malariavaccine.org/). These prophylactic drug treatments are often very expensive for most people living in endemic areas [20]. Malaria infections are treated by the use of antimalarial drugs, such as quinine or artemisinin derivatives, although drug resistance is increasingly common [19].

Research on malaria vaccine is mainly focused on finding differences between affected people, healthy people and people that are protected against it. A considerable role is played by new technologies such as microarray techniques and bio-informatics. Algorithms of pattern recognition and signal processing are speeding up the malaria vaccine development. Consider only that the completed genome sequences of bacteria and parasites have disclosed the presence of ~1,000–5,000 candidate proteins (and variants thereof) for each microorganism [8].

In this work, we analyse and cluster sera derived from the medical analysis of 189 subjects. The data are the result of a study using microarray technology to characterise the serum reactivity profiles of 189 children, 3–9 years old, living in Gambia. In this study, Gray et al. [8] assessed the antibody responses against 18 recombinant proteins derived from 4 leading blood-stage vaccine candidates for *Plasmodium falciparum malaria*.

The sera are divided into three different groups: no malaria (group A) including children having no evidence of infection throughout the study period; clinical malaria (group B), children having at least 1 episode of fever (temperature more than 37.5°C) and parasitemia ≥5,000 μl; and asymptomatic (group C), children with parasitemia or acquired splenomegaly but with no evidence of fever or other symptoms of clinical malaria [8].

To analyse the data and cluster them, Gray et al. [8] used the *k*-means clustering technique. The data were clustered into three clusters. Their results are reported in Table 1, where we present the number of subjects from each group that were classified in each of the three clusters created.

We note from these results that it is not possible to associate the three identified clusters with the three classes of subjects we have, without committing a very significant error. For example, cluster 2 could easily be considered to represent group A or group B. If we consider cluster 2 to represent group B, cluster 1 to represent group C and cluster 3 to represent group A, 108 out of the 189 subjects will be wrongly classified, producing a classification accuracy of 43% only.

**Table 1** Data clustering with the *k*-means technique, Gray et al. [8]

| | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| Group A | 14 | 31 | 10 |
| Group B | 22 | 43 | 16 |
| Group C | 28 | 19 | 6 |

In this paper, first, in Sect. 2, we analyse the data to work out which antigens have the most discriminatory power between the groups, and then examine what the minimum possible classification error one can achieve is. In Sect. 3, we present novel methodology to approach this ideally minimum error, by considering the context of the data expressed by their pairwise relations. Our results are presented in Sect. 4 and our conclusions in Sect. 5.

## 2 Data characteristics

Figure 1 shows the data divided into three sets: group A, group B and group C. In this figure, each matrix column represents the subjects' immunology response, while the *i*th row shows the positive ("1") or negative ("0") response of all subjects to the *i*th antibody. The three groups are composed, respectively, of 55 subjects (group A), 81 subjects (group B) and 53 subjects (group C). Each subject is represented by a binary 18-dimensional vector: a component is 1 if the child's serum has a positive response and 0 if it has negative response to the corresponding antibody.

To discover whether there are proteins which permit us to distinguish better one group from the others, we show in

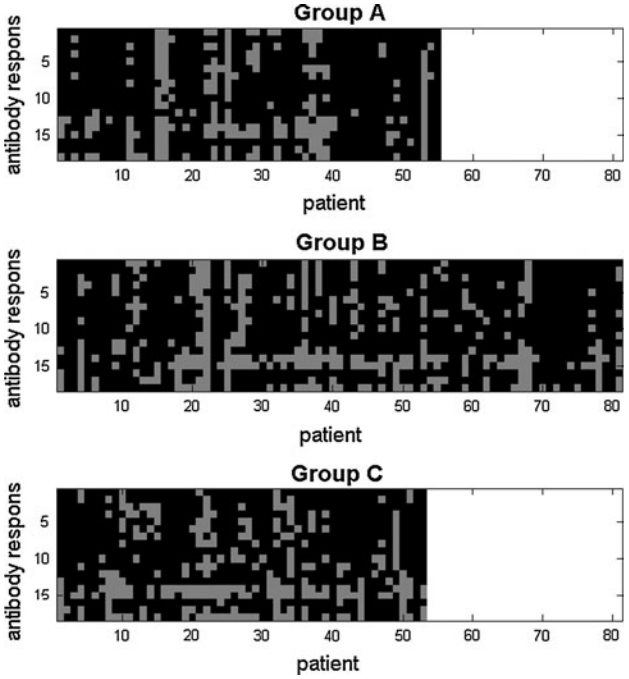

**Fig. 1** Data corresponding to groups A, B and C. A *column* represents a subject's immunology response, while *each row* is an antibody response. Each subject is represented by an 18-dimensional vector. *Black squares* represent negative responses to the antibody, while *grey squares* represent positive responses

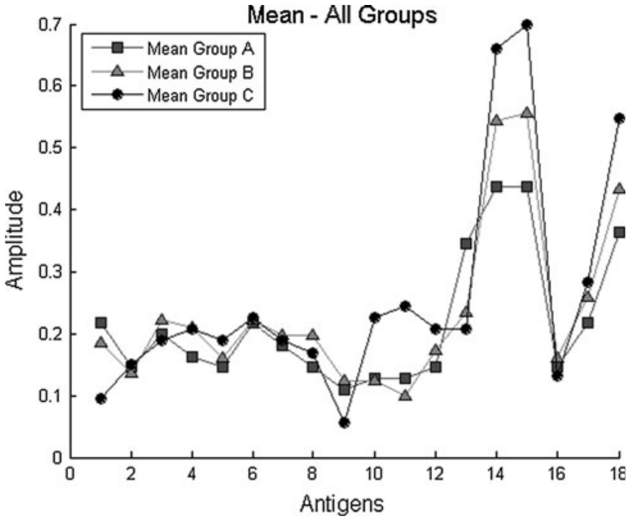

**Fig. 2** Mean values of positive responses for 18 antigens and for each group

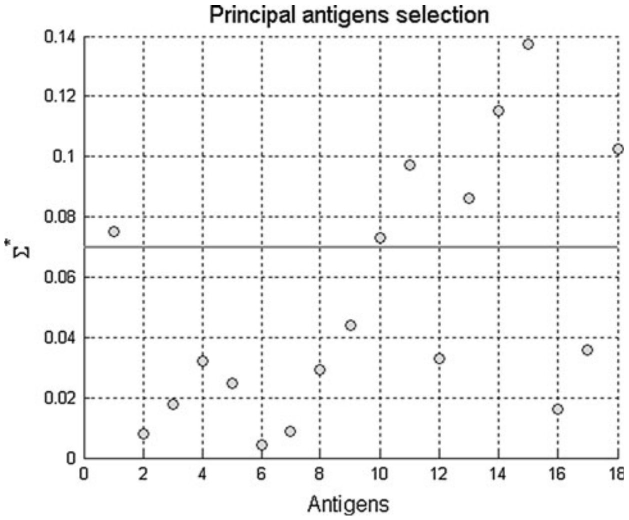

**Fig. 3** Principal antigens selection: we select the antigens that have a mean value for at least one group, which is considerably different from the mean value of the whole population

Fig. 2 the mean value for each matrix row. We observe that some antigens have the same mean value (i.e. antigens 2 or 6), while for antigens 1, 10, 11, 13, 14, 15, 18, the mean values are very different. We select the most discriminative antigens by considering the mean value $\mu^{\mathrm{A}}$, $\mu^{\mathrm{B}}$ and $\mu^{\mathrm{C}}$ for sets A, B and C and the mean $\mu$ for the whole population. We choose those antigens that have a mean value, which, for at least one set, is considerably different from the mean value of the whole population. Defining $\Sigma^*$ vector as

$$\Sigma^* = \max\{|\mu^{\mathrm{A}} - \mu|, |\mu^{\mathrm{B}} - \mu|, |\mu^{\mathrm{C}} - \mu|\}, \qquad (1)$$

we select the $i$th antigen if the $i$th component of $\Sigma^*$ is considerably large.[1] Figure 3 shows the value of $\Sigma^*$ for each antigen. There is a natural gap in the values of these antigens that allows us to divide them into two categories. However, one may also define a threshold value following Otsu's method [11], where the threshold is selected to make the two populations as compact as possible. It turns out that a threshold in the range [0.05–0.07] minimises the total intraclass spread. Using such a threshold allows us to select the most meaningful antigens. The selected antigens are those that lie above the line in Fig. 3. We call these antigens *principal antigens*.

## 2.1 Confidence intervals and significance of the principal antigens

To quantify how significant the difference in mean value is for the principal antigens selected, we considered the value of each antigen as a random variable with a binomial probability distribution.

Assuming that the outcome of a random experiment is either 0 or 1, the probability of getting $k$ 1s in $n$ trials is given by

$$\mathrm{pmf}(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad (2)$$

where $p$ is the probability of getting 1 in a single trial. Let us denote by $k_i^{\mathrm{A}}$ the total number of subjects of group A that are positive to the $i$th antigen (they have "1" in their $i$th component). In a similar way, we define $k_i^{\mathrm{B}}$ and $k_i^{\mathrm{C}}$. According to these definitions, the probability of finding a person that has "1" at the $i$th component is

$$p_i = \frac{k_i^{\mathrm{A}} + k_i^{\mathrm{B}} + k_i^{\mathrm{C}}}{N^{\mathrm{A}} + N^{\mathrm{B}} + N^{\mathrm{C}}}, \qquad (3)$$

where $N^{\mathrm{A}} = 55$, $N^{\mathrm{B}} = 81$ and $N^{\mathrm{C}} = 53$. Therefore, the probability of getting $k_i$ people positive with respect to the $i$th antigen is given by

$$\mathrm{pmf}(k_i; N, p_i) = \binom{N}{k_i} p_i^{k_i} (1-p_i)^{N-k_i}. \qquad (4)$$

where $N = 189$. We also introduce the probability of finding a person that is positive with respect to the $i$th antigen in groups A, B and C as[2]

$$p_i^{\mathrm{A}} = \frac{k_i^{\mathrm{A}}}{N^{\mathrm{A}}}, \quad p_i^{\mathrm{B}} = \frac{k_i^{\mathrm{B}}}{N^{\mathrm{B}}}, \quad p_i^{\mathrm{C}} = \frac{k_i^{\mathrm{C}}}{N^{\mathrm{C}}}. \qquad (5)$$

Now we want to check whether $p_i^{\mathrm{A}}, p_i^{\mathrm{B}}$ and $p_i^{\mathrm{C}}$ are significantly different from $p_i$. For example, if the probability to get $k_i$ positive subjects in the whole

[1] The mean values $\mu = [\mu_1\ \mu_2\ \ldots\ \mu_{18}]^{\mathrm{T}}$, $\mu^{\mathrm{A}} = [\mu_1^{\mathrm{A}}\ \mu_2^{\mathrm{A}}\ \ldots\ \mu_{18}^{\mathrm{A}}]^{\mathrm{T}}$, $\mu^B = [\mu_1^{\mathrm{B}}\ \mu_2^{\mathrm{B}}\ \ldots\ \mu_{18}^{\mathrm{B}}]^{\mathrm{T}}$ and $\mu^{\mathrm{C}} = [\mu_1^{\mathrm{C}}\ \mu_2^{\mathrm{C}}\ \ldots\ \mu_{18}^{\mathrm{C}}]^{\mathrm{T}}$ are 18-dimensional vectors where each component is the mean value of the $i$th antigen.

[2] Note that $\mu = [\mu_1\ \mu_2\ \ldots\ \mu_{18}]^{\mathrm{T}}$ and $p = [p_1\ p_2\ \ldots\ p_{18}]^{\mathrm{T}}$ have the same values, in the same way as $\mu_i^{\mathrm{A}} = p_i^{\mathrm{A}}$, $\mu_i^{\mathrm{B}} = p_i^{\mathrm{B}}$ and $\mu_i^{\mathrm{C}} = p_i^{\mathrm{C}}$ for $i = 1, \ldots, 18$; we changed the notation here in order to underline the probabilistic meaning of these quantities.

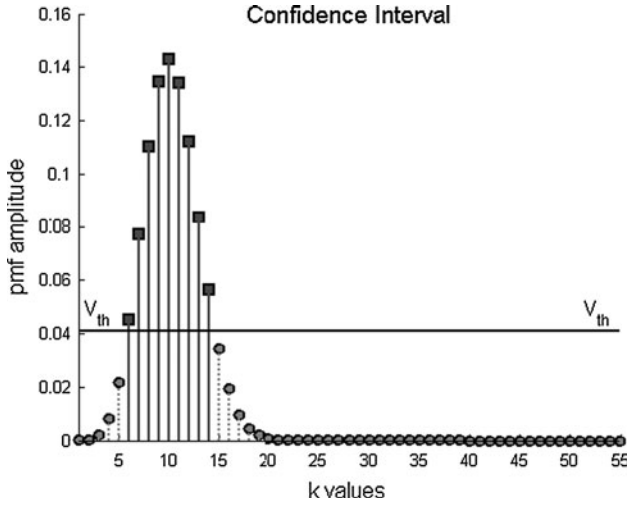

**Fig. 4** In this figure, we represent the confidence interval of 89.60% for the first antigen of group A. This means that the sum of the probability impulses for $k \in [6, 14]$ is 0.8960

population is the same as that of getting $k_i^{\mathrm{A}}$ positive subjects in set A, then

$$\mathrm{pmf}(k_i^{\mathrm{A}}; N^{\mathrm{A}}, p_i) \simeq \mathrm{pmf}(k_i^{\mathrm{A}}; N^{\mathrm{A}}, p_i^{\mathrm{A}}). \tag{6}$$

For this purpose, we defined the confidence intervals for the binomial distribution as well as for the normal distribution.[3] Considering Fig. 4 and set A, for each antigen we defined two sets:

$$\overline{\mathcal{K}}_{i,\mathrm{A}} = \{k : \mathrm{pmf}(k; N^{\mathrm{A}}, p_i) < V_{\mathrm{th}} \text{ and } k \in \{1, \ldots, N^{\mathrm{A}}\}\} \tag{7}$$

$$\mathcal{K}_{i,\mathrm{A}} = \{k : \mathrm{pmf}(k; N^{\mathrm{A}}, p_i) \geq V_{\mathrm{th}} \text{ and } k \in \{1, \ldots, N^{\mathrm{A}}\}\}. \tag{8}$$

Then

$$\overline{\mathcal{K}}_{i,\mathrm{A}} \bigcup \mathcal{K}_{i,\mathrm{A}} = \{1, \ldots, N^{\mathrm{A}}\}. \tag{9}$$

According to the above definitions, we define the confidence interval of at least $\epsilon$% for group A, with respect to the $i$th antigen, as the set $\mathcal{K}_{i,\mathrm{A}}$ where threshold $V_{\mathrm{th}}$ is chosen so that the sum of the impulses with amplitude greater than $V_{\mathrm{th}}$ is at least $\frac{\epsilon}{100}$, while the sum of the impulses of the probability distribution with amplitude less than $V_{\mathrm{th}}$ is less than $(1 - \frac{\epsilon}{100})$:

$$\sum_{k \in \mathcal{K}_{i,\mathrm{A}}} \mathrm{pmf}(k, N^{\mathrm{A}}, p_i) \geq \frac{\epsilon}{100}. \tag{10}$$

In the above formulae, we considered group A and probability $p_i$ in place of $p_i^{\mathrm{A}}$. For fixed $\epsilon$, if $k_i^{\mathrm{A}}$ belongs to $\mathcal{K}_{i,\mathrm{A}}$ for small values of $\epsilon$ (big values of $V_{\mathrm{th}}$) it means that $p_i^{\mathrm{A}}$ is close to $p_i$, while if $k_i^{\mathrm{A}}$ belongs to $\overline{\mathcal{K}}_{i,\mathrm{A}}$ for big values of $\epsilon$ (small values of $V_{\mathrm{th}}$) it means that $p_i^{\mathrm{A}}$ is significantly different from $p_i$.

**Table 2** Antigen significance

| Confid. interval | Antigen and corresponding group |
|---|---|
| 92% | {15,A} |
| 86% | {14,A} |
| 83% | {15,C} |
| 80% | {1,C} |
| 75% | {11,C} |
| 72% | {18,A} {14,C} |
| 63% | {9,C} |
| 59% | {18,C} |
| 56% | {13,C} |
| 55% | {10,C} {13,A} |

Changing the level of the confidence interval, we can analyse the importance of the antigens for the three groups. For confidence ranging from 50 to 99%, Table 2 presents the antigens that are out of the corresponding interval.

Let us focus our attention to antigen 15. The mean value for antigen 15 in group A is outside the 92% confidence interval. This means that if we consider the whole population and a subset of it composed only of group A, this subset does not represent the population for antigen 15 in 92% of the cases. According to that, we can say that the "*significance*" of antigen 15 is 92%.

From Table 2, we can see that all *seven principal antigens* have a significance value between 92 and 55%. Moreover, some antigens are significant for more than one group. However, note that there is no antigen that characterises in any significant way group B. This is as expected, because group B is the biggest group and forms a great part of the population (43%). This means that the value of $\mu$ is close to $\mu_{\mathrm{B}}$ and there are no antigens that characterise the difference between group B and the whole population. We do not include antigen 9 to the principal antigens, because the ninth value of $\Sigma^*$ is considerably lower than its first and tenth components (see Fig. 3).

## 2.2 The minimum possible error

When each subject is represented by a small number of antigens (in our case 18 or 7), expressed as binary patterns, it is possible that different subjects have identical representations. If subjects with identical representations belong to the same group, this is a good characteristic of the data. If, however, they belong to different groups, then one has to realise that it is not possible to distinguish these two subjects using only the information that is available. In order to work out the minimum possible error allowed by

[3] For the normal distribution, $N(\mu, \sigma)$, about 68% of the values are within $[\mu - \sigma, \mu + \sigma]$, about 95% of the values are within $[\mu - 2\sigma, \mu + 2\sigma]$ and about 99.7% lie within $[\mu - 3\sigma, \mu + 3\sigma]$.

the information that is available, we introduce here the concept of the ideal classifier: the ideal classifier can classify the subjects by committing the minimum possible error. This becomes clear with an example. Let us assume that three subjects have identical representation when we use only 18 genes, although two of them belong to group A and one of them to group B. The ideal classifier will put all three to the same class, which will be identified with group A, so that only one out of the three subjects will be wrongly classified. A less than ideal classifier will place all three in the same class, which may be identified with group B (two of the subjects will be wrongly classified) or with group C (all three subjects will be wrongly classified). The ideal classifier, of course, is expected to classify with the minimum error subjects with identical data and with no error subjects with discriminating data.

The ideal classifier will have an 82.54% accuracy in the 18-dimensional space and a 62.96% accuracy in the 7-dimensional space. Note that these minimum possible errors may not be achievable in practice, as the non-ambiguous patterns may be very diverse and incoherent, even when the subjects belong to the same group, and thus, no real classifier will be able to group together subjects of the same group. The classifiers we shall propose in this paper will attempt to approach this minimum possible error.

## 3 Proposed methodology

Figure 5 shows schematically the essence of our idea: let us consider that each subject is represented by two features, and let us assume that we have two classes of data that have a strong overlap. If we try to classify a new pattern using these two features, any pattern that happens to fall in the overlapping area of the training patterns will be classified with the same accuracy as pure chance. Let us consider now the pairwise relationships in the training data, between patterns that belong to the same group. It is as if we replace the individual points in this feature space by a structure that connects the members of each group with each other. Let us then introduce the new pattern to the structure of each class. Obviously, the structure of each class will be perturbed when a new pattern has to be incorporated into it. We shall classify the new pattern to the class with the least perturbed structure.

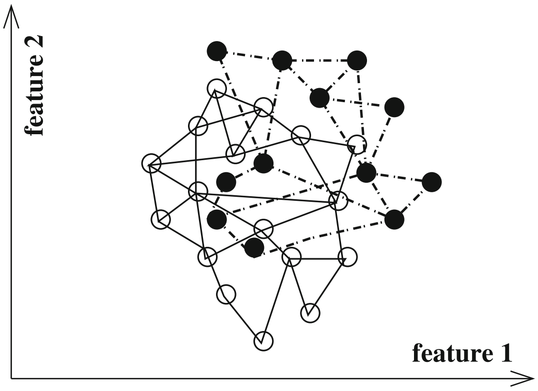

**Fig. 5** Two overlapping classes in a 2D feature space. We may be able to separate them if we consider pairwise relations between the patterns that make up each class

To use this idea in practice, we have to decide two things:

(a) how to construct features from a pattern that describes a subject and
(b) how to describe the structure with the pairwise relations.

These two issues will be discussed in the subsections that follow.

### 3.1 Feature creation: binary code

As we stated earlier, the value of each component can be "0" or "1". This characteristic suggested to us the idea of encoding the subjects with a binary number. Let us consider, e.g., one element $P_{\mathrm{A},1}$ of set A:

$P_{\mathrm{A},1} = [0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,0\,0\,1]$.

Interpreting this pattern as a binary number, we have:

$$\begin{aligned} P_{\mathrm{A},1} &= 0 \times 2^{17} + 0 \times 2^{16} + \cdots + 0 \times 2^{1} + 1 \times 2^{0} \\ &= (57)_{10} \end{aligned} \tag{11}$$

The order of "0" and "1" is not immunologically important. In fact, what is important is the positive or negative response against antigens and not the order in which we consider the antigens. Changing the antigen order we can codify the same subject in different ways. Indeed, we have as many as 18! = 6402373705728000 possible permutations.

To deal with smaller numbers, we decided to consider just the seven *principal antigens*. Table 3 lists the frequencies of occurrence of these seven antigens. The antigens are ordered in each row from the least frequent (L.F.) to the most frequent (M.F.) occurrence in each group and in all groups together.

Using this table, we may select convenient permutations that may allow us to distinguish the groups.

Let us consider, e.g., permutation {13, 1, 11, 10, 18, 14, 15}. We called it α-code. In this permutation, antigen 13 represents the most significant bit (MSB), while antigen 15 is the least significant bit (LSB). In this way, antigens {18, 14, 15} have little weight, because they do not permit us to separate the groups. The positions of antigens 13 and 1 (MSB) allow us to maximise the distance between groups

**Table 3** Frequency of seven antigens

| | L.F. | | | | | M.F. |
|---|---|---|---|---|---|---|---|
| Freq.A | 10 | 11 | 1 | 13 | 18 | 14 | 15 |
| Freq.B | 11 | 10 | 1 | 13 | 18 | 14 | 15 |
| Freq.C | 1 | 13 | 10 | 11 | 18 | 14 | 15 |
| Freq.ToT | 11 | 10 | 1 | 13 | 18 | 14 | 15 |

**Table 4** Binary code

| | Binary value | $\alpha$-code | $\beta$-code |
|---|---|---|---|
| $P_1$ | 0001110 | $(1000011)_2 = (67)_{10}$ | $(0001011)_2 = (11)_{10}$ |
| $P_2$ | 1101101 | $(1110110)_2 = (118)_{10}$ | $(0111110)_2 = (62)_{10}$ |
| $P_3$ | 1010010 | $(0101001)_2 = (41)_{10}$ | $(1010001)_2 = (81)_{10}$ |

**Table 5** Sorted binary code

| Subject | Sorted $\alpha$-vector | Subject | Sorted $\beta$-vector |
|---|---|---|---|
| $P_3^\alpha$ | $(0101001)_2 = (41)_{10}$ | $P_1^\beta$ | $(0001011)_2 = (11)_{10}$ |
| $P_1^\alpha$ | $(1000011)_2 = (67)_{10}$ | $P_2^\beta$ | $(0111110)_2 = (62)_{10}$ |
| $P_2^\alpha$ | $(1110110)_2 = (118)_{10}$ | $P_3^\beta$ | $(1010001)_2 = (81)_{10}$ |

C and {A, B}. To maximise the distance between groups B and C, we consider antigen 11 as the third significant bit. The fourth significant bit then has to be antigen 10.

Another possible choice is considering the global occurrence. Let us call $\beta$-code the antigen order given by the last row in Table 3: {11, 10, 1, 13, 18, 14, 15}.

Considering different permutations we can create mono- and multi-dimensional feature spaces where along each axis we represent a different code.

### 3.2 Pairwise structure: the relational space

The pairwise relations we consider are represented in the feature space by lines that join the corresponding points, i.e. edges that connect corresponding vertices (see Fig. 5). The relational space is a space where each edge is represented by a point, with coordinates that measure something which characterises the vertices it connects. For our data, we define this relational space as follows. We consider each group separately and encode subjects according to different codes. Let us consider the $\alpha$-code and $\beta$-code and then create two vectors for each group containing the codes of the subjects. Let us call these vectors $\alpha$-vector and $\beta$-vector. Before drawing the graph in the relational space, we sort the elements of the $\alpha$- and $\beta$-vectors in increasing order and pair the $i$th element of the sorted $\alpha$-vector with the $i$th element of the sorted $\beta$-vector. Then, we plot these new pairs in the relational space. Obviously, each point now does not represent a single person but a pair of subjects belonging to the same group.

Let us consider a simple example. Consider three subjects $P_1$, $P_2$ and $P_3$ who belong to group A with data shown in Table 4. The sorted $\alpha$-vector of group A is (41, 67, 118) and the $\beta$-vector is (11, 62, 81) (see Table 5). In the relational space, we plot points (41, 11), (67, 62) and (118, 81), which correspond to subject pairs: ($P_3^\alpha$, $P_1^\beta$), ($P_1^\alpha$,$P_2^\beta$) and ($P_2^\alpha$, $P_3^\beta$), respectively. Let us consider the meaning of pair (41, 11). It corresponds to patients ($P_3$, $P_1$). Patient $P_3$ is one who does not have the antigen that most discriminates between groups C and {A, B} activated. (That is why this patient's $\alpha$-code value is minimal.) Patient $P_1$ does not have the antigen that most discriminates between groups C and B also activated. So, these patients have something in common, and we may justify their pairing on this basis. Consider also the meaning of pair (118, 81). It corresponds to patients ($P_2$, $P_3$). Patient $P_2$ has activated the antigens that discriminate between groups C and {A, B}, while patient $P_3$ has activated the antigen that discriminates between groups C and B. Again, these patients have something in common, and it is reasonable to pair them.

### 3.3 Classifier

In the relational space, which we created in the previous section, every point represents a pair of subjects that are known to belong to the same group. Assume now that the

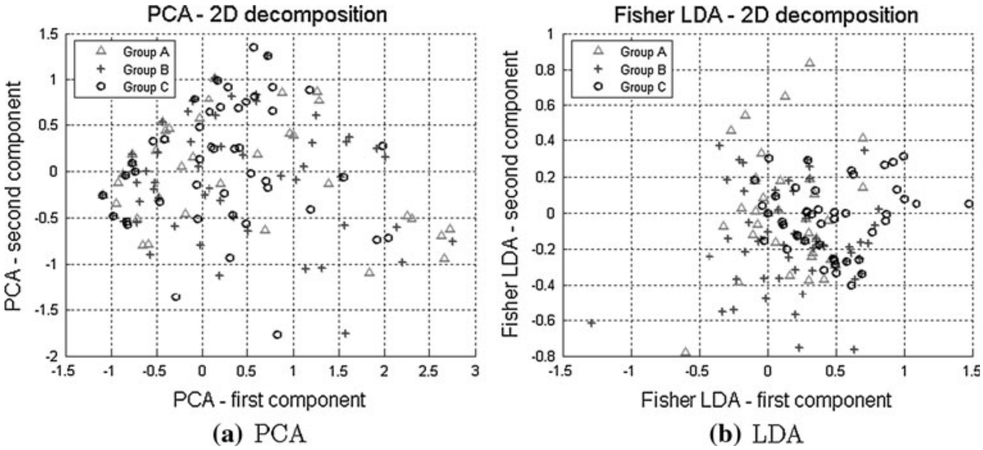

**Fig. 6** **a** PCA: considering two principal components. **b** LDA: the three groups are depicted in a 2D linear space that maximally separates them

data of a new subject arrive. We wish to classify this subject to one of the three groups. We propose the following algorithm.

We assign the test pattern to each group in turn, and re-classify the points of this group in the relational space to assess which group shows most improvement by the incorporation of the test pattern. We assign the test pattern to the group that shows the most improvement. To use this approach, we have to specify:

(a) the classifier used to classify the points in the relational space;
(b) what is meant by "showed most improvement".

As a classifier we use the *k* nearest neighbour classifier (*k*-nn). In the experimental section, we shall show that this classifier produced better results than the support vector machine (SVM), the naive Bayes and the C4.5 tree classifier.

To assess which class assignment to the test pattern improves the coherence of the clusters in the relational space most, we use the following procedure. We use the leave-one-out method to classify all points in the relational space of the group we disturb. First we consider the group as it is, before we introduce the test pattern. We consider one point of this group (representing a pair of subjects) at a time and classify it using all remaining points. This way we work out what fraction of the group would be classified correctly when we have the training data only. Then we introduce the new pattern in this group, create the pairs of patients and classify them using all training patterns in the relational space, and, thus, identify the fraction of the points that are correctly classified now. We assign the test pattern to the group for which we observed the largest improvement in classification when we incorporated it into the group. In other words, we assign the test pattern to the group that becomes most coherent in its relational space representation when we incorporate the new pattern into it.

## 4 Experimental results

First, we have to check whether the results produced by straightforward *k*-means reported in Table 1 could be improved if we used a more sophisticated algorithm, like, e.g., a SVM. Then we examine whether the results improve if we use various combinations of features constructed using the binary code. Finally, we shall show that the best results are obtained by the proposed method that uses the relational feature space.

### 4.1 Classification of the data in the original space

Let us consider first the method of principal component analysis (PCA) and linear discriminant analysis (LDA). These techniques are usually employed to reduce multidimensional data sets to a lower dimensionality feature space where the classes may be better separated. In Fig. 6a, we project the 18D feature space into a 2D plane where the first and second principal components are considered, while in Fig. 6b we depict the LDA projection into a 2D space, where the classes are assumed to be maximally separable [16, 18]. We see that the groups are overlapped and it is hard to distinguish them clearly.

We present in Table 6 the classification results using SVMs [3, 4, 7, 9] and *k*-nn [1, 5, 17] classifiers in the 18-dimensional and 7-dimensional feature space using the leave-one-out method. Since our data are "0" or "1" we did not need to scale them.[4] We tried $k = 1, 2, 3, 4, 5$ and we report here only the results with $k = 3$ because these turned out to be the best. In Table 6, $k$ is the number of nearest neighbours, $C$ is the Tikhonov constant of the SVM problem and $p$ is the parameter of the SVM with the RBF kernel. All parameters in this table were chosen according to the recommendations in [6]. The authors suggest to

[4] Also note that for binary data, the Euclidean metric, the Hamming distance and the L1 norm coincide.

**Table 6** *k*-nn and SVMs: classification results

| Space | Classifier | Parameter | Accuracy (%) |
|---|---|---|---|
| 18 antigens | *k*-nn | $k = 3$ | 37.57 |
| 7 principal antigens | *k*-nn | $k = 3$ | 46.03 |
| 18 antigens | SVM (linear kernel) | $C = 2^{-2}$ | 43.92 |
| 7 principal antigens | SVM (linear kernel) | $C = 2^{-1}, 2^{10}$ | 46.03 |
| 18 antigens | SVM (RBF kernel) | $C = 2^{-1}, p = 2^{2}$ | 48.15 |
| 7 principal antigens | SVM (RBF kernel) | $C = 2^{10}, p = 2^{-7}$ | 49.74 |

consider a "grid-search" on $C$ and $p$ using cross-validation, and to use exponentially growing sequences of $C$ and $p$. That is what we did and the combination of parameter values with the best cross-validation accuracy was picked.

### 4.2 Classification of the data using their binary codes

Besides the $\alpha$-code and $\beta$-code, we considered some other antigen permutations. Let us define the $\gamma$-code = {1, 10, 11, 13, 18, 14, 15} and the $\delta$-code = {10, 11, 1, 13, 18, 14, 15}. In the $\gamma$-code, we use antigen 1 as the MSB to help us separate group C from groups A and B, while in the $\delta$-code we consider antigens 10 and 11 as the most significant ones to separate group A from group B. Table 7 lists all the codes we considered.

**Table 7** Binary codes

| Code name | Antigens order |
|---|---|
| $\alpha$ | {13, 1, 11, 10, 18, 14, 15} |
| $\beta$ | {11, 10, 1, 13, 18, 14, 15} |
| $\gamma$ | {1, 10, 11, 13, 18, 14, 15} |
| $\delta$ | {10, 11, 1, 13, 18, 14, 15} |

Using these binary codes, we created new feature spaces and analysed them with the *k*-nn classifier with $k = 3$, testing the method with the leave-one-out cross-validation. We report the results of the various feature spaces we considered in Table 8.

**Table 8** Binary code spaces: the classifier used was 3-nn

| Space | Dimension | Accuracy (%) |
|---|---|---|
| $\alpha$ | 1 | 46.56 |
| $\beta$ | 1 | 47.62 |
| $\gamma$ | 1 | 46.03 |
| $\delta$ | 1 | 46.56 |
| $(\alpha, \beta)$ | 2 | 49.21 |
| $(\gamma, \delta)$ | 2 | 48.68 |
| $(\gamma, \beta)$ | 2 | 49.74 |
| $(\alpha, \beta, \gamma)$ | 3 | 49.74 |
| $(\delta, \alpha, \beta)$ | 3 | 50.26 |
| $(\gamma, \delta, \alpha)$ | 3 | 49.74 |
| $(\beta, \gamma, \delta)$ | 3 | 49.74 |

In order to get an idea which class is confused with which, we give in Table 9 the confusion matrices of the classifications we obtained in three of these feature spaces.

### 4.3 Classification of the data in the relational space

Figure 7 shows the data plotted in the relational space of the $\alpha$- and $\beta$-codes. Let us call this space $(\alpha^r, \beta^r)$. In this figure, we can distinguish easily all groups: they are overlapped just in the corner near zero where we have several subjects with all or most genes not expressed. In order to appreciate the importance of sorting the patients before pairing them, we show in Fig. 8 the relational space obtained by randomly pairing the patients and by pairing each patient with himself.

Table 10 reports the results of the *k*-nn classifier with the leave-one-out cross-validation method. The reported results correspond to the $k = 3$ classifier that produced the best results compared with classifiers with $k = 1, 2, 4, 5$. The confusion matrices of the best results, obtained in the $(\alpha^r, \beta^r)$ and the $(\delta^r, \alpha^r, \beta^r)$ spaces,[5] are shown in Table 9.

The $(\alpha^r, \beta^r)$ space was also analysed by replacing the *k*-nn classifier with an SVM with different kernels. Table 11 reports these results.

We also used the naive Bayes classifier in the relational space, obtained from [12]. The naive Bayes classifier estimates a class-conditional distribution using a histogram of feature values and assuming feature independence. The number of histogram bins may be treated as a free parameter. The algorithm counts the number of training examples for each of the classes in each of the bins, and classifies the object to the class that gives maximum posterior probability. The prior probabilities in our case were set up according to:

[5] The 3D relational space $(\delta^r, \alpha^r, \beta^r)$ was obtained in the same way as the $(\alpha^r, \beta^r)$ space, considering triplets of patients of the same group.

**Table 9** Confusion matrices for feature and relational spaces: the classifier used was 3-nn

| | Group A | Group B | Group C |
|---|---|---|---|
| $\beta$-code feature space | | | |
| True group A | 39 | 14 | 20 |
| True group B | 40 | 38 | 5 |
| True group C | 24 | 16 | 13 |
| ($\gamma$-code, $\beta$-code) feature space | | | |
| True group A | 39 | 14 | 20 |
| True group B | 36 | 39 | 6 |
| True group C | 22 | 15 | 16 |
| ($\delta$-code, $\alpha$-code, $\beta$-code) feature space | | | |
| True group A | 41 | 12 | 2 |
| True group B | 37 | 39 | 5 |
| True group C | 22 | 16 | 15 |
| ($\alpha^r$, $\beta^r$) relational space | | | |
| True group A | 40 | 8 | 7 |
| True group B | 29 | 46 | 6 |
| Tue group C | 16 | 13 | 24 |
| ($\delta^r$, $\alpha^r$, $\beta^r$) relational space | | | |
| True group A | 40 | 8 | 7 |
| True group B | 29 | 44 | 8 |
| True group C | 16 | 13 | 24 |

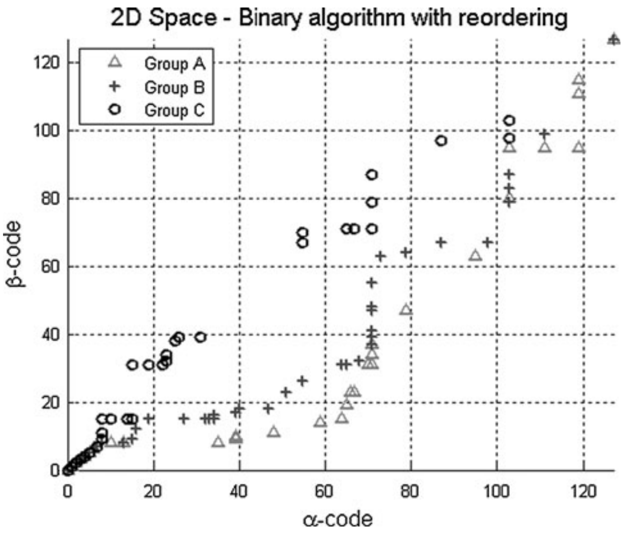

**Fig. 7** Relational space obtained with sorted α-code and sorted $\beta$-code of pairs of subjects

$$p_A = 55/189, \quad p_B = 81/189, \quad p_C = 53/189.$$

Figure 9 reports the results of this classifier for different number of bins. In all cases, they are worse than those of the $k$-nn classifier.

Finally, we explored the C4.5 decision tree as a classifier in the original 18-dimensional and 7-dimensional spaces, and in the relational space [13, 14]. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits the set of samples into subsets enriched in one class or the other. Its criterion is the normalised information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalised information gain is chosen to make the decision. In Table 12, we report the results of our analysis, where we have considered different values for the parameters "confidence-factor" and "trials". We used a grid-search, considering account pairs of ("confidence-factor", "trials"), with "confidence-factor" taking values 0.1, 0.2, …, 1.0, and "trials" being an integer between 1 and 10. For the other parameters, default values were used. The accuracies were computed with the leave-one-out technique. The best accuracy identified was lower than that of the 3-nn classifier.

The summary of the best results obtained by each classifier is shown in Table 13. When we assign the test pattern to each group in turn for the classification, we modify the placement of the points in the relational space. While the SVM, C4.5 and naive Bayes classifier are more vulnerable to this modification, the 3-nn method is more robust and produces higher accuracy.

### 4.4 A systematic way to create codes

Due to the large number of antigen possible permutations, one may device several codes and relational spaces. We estimated that the CPU time required to explore all possible permutations of 7 antigens is about 12 years! In this section, we propose the following algorithm that allows one to generate good codings and relational spaces.

1. Consider the two most important antigens.
2. Analyse all possible combinations.
3. Select the relational space that gives the highest accuracy. If two different relational spaces have the same accuracy, explore all possible solutions.
4. Add to the best relational space one antigen at every possible position with respect to the existing antigens, but without changing the relative order of the other antigens.
5. Go to Step 3.

The algorithm finishes when the accuracy starts to decrease. The accuracy is computed using our classifier and leave-one-out testing.

The results of this algorithm are reported in Table 14.

Obviously, this algorithm does not allow us to find the best combination, but permits us to construct reasonably good relational spaces in finite computing time.

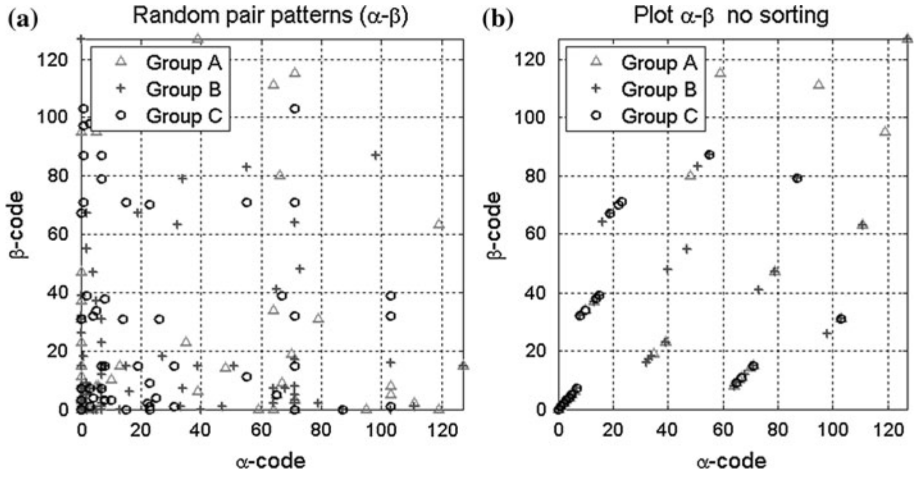

**Fig. 8** **a** Random pairing of patients. **b** Pairing each person with himself

**Table 10** Relational spaces: the classifier used was 3-nn

| Space | Dimension | Accuracy (%) |
|---|---|---|
| ($\alpha^r$, $\beta^r$) | 2 | 58.20 |
| ($\gamma^r$, $\delta^r$) | 2 | 41.80 |
| ($\gamma^r$, $\beta^r$) | 2 | 51.85 |
| ($\alpha^r$, $\beta^r$, $\gamma^r$) | 3 | 52.38 |
| ($\delta^r$, $\alpha^r$, $\beta^r$) | 3 | 57.14 |
| ($\gamma^r$, $\delta^r$, $\alpha^r$) | 3 | 51.85 |
| ($\beta^r$, $\gamma^r$, $\delta^r$) | 3 | 42.33 |

**Table 11** Classification results in the ($\alpha^r$, $\beta^r$) relational space

| Space | Classifier | Parameter | Accuracy (%) |
|---|---|---|---|
| ($\alpha^r$, $\beta^r$) | *k*-nn | $k = 3$ | 58.20 |
| ($\alpha^r$, $\beta^r$) | SVM (RBF kernel) | $C = 2^2$, $p = 2^{2.85}$ | 57.67 |
| ($\alpha^r$, $\beta^r$) | SVM (lin. kernel) | $C = 2^0$ | 49.74 |

### 4.5 Computational complexity

Let us call $\mathcal{C}^{\text{train}}$ the computational complexity for training a classifier and $\mathcal{C}^{\text{class}}$ its computational complexity for classifying an unknown pattern. $\mathcal{C}^{\text{train}}$ will be a function of the training patterns we have, $\mathcal{N}_{\text{train}}$, and both will be functions of the dimensionality of the space we use *d*. When a new pattern has to be classified, we combine it with the training patterns we have and classify the pairs we create. If we have $\mathcal{N}_{\text{class}}$ classes, we do that by assigning the new pattern to each class in turn. The overall complexity of the algorithm then is:

$$\mathcal{C}^{\text{train}}(\mathcal{N}_{\text{train}}, d) + \mathcal{N}_{\text{class}}(\mathcal{N}_{\text{train}} + 1)\mathcal{C}^{\text{class}}(d) \quad (12)$$

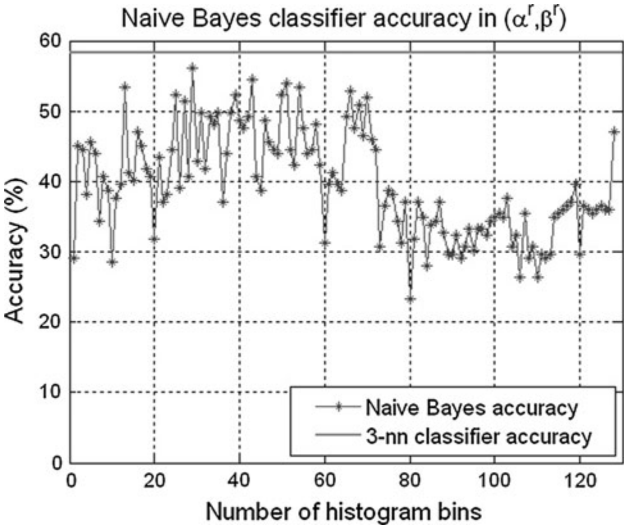

**Fig. 9** Accuracies using the naive Bayes classifier in the relational space ($\alpha^r$, $\beta^r$). The *straight line* at the top of the graph is the value of the accuracy obtained with 3-nn classifier

## 5 Discussion and conclusions

In this paper, we introduced the concept of the relational space, where instead of representing individual patterns, we represent pairs of patterns. In this space, a subject is classified not simply according to the data that describe him/her, but according to the relationship these data have with the data of other subjects. Working in the relational space allowed us to improve the classification accuracy of the malaria subjects by about 15% above the accuracy obtained by conventional methods.

We also introduced in this paper the concept of principal antigens, defined as those that have a statistically significant specificity for a group of subjects as well as the concept of the ideal classifier, used to indicate what the

**Table 12** C4.5 results

| C4.5 tree parameter and possible values | 18-dim. space | 7-dim. space | Relational space |
|---|---|---|---|
| Confidence-factor [0, …, 1] | 0.9 | 1 | 0.3 |
| Trials [1, …, *N*] | 7 | 10 | 1 |
| Gain-ratio (yes/no) | Yes | Yes | Yes |
| Increment [0, …, *N*] | 0 | 0 | 0 |
| Min-objects [1, …, *N*] | 2 | 2 | 2 |
| Probability-thresholds (yes/no) | Yes | Yes | Yes |
| Seed [0, …, *N*] | No seeding | No seeding | No seeding |
| Subset (yes/no) | Yes | Yes | Yes |
| Verbosity [0, …, 5] | 0 | 0 | 0 |
| Window-size [0, …, *N*] | 0 | 0 | 0 |
| Accuracy (%) | 32.33 | 38.26 | 52.91 |

**Table 13** Relational space results using different classifiers

| Relational space | Classifier | Classifier parameters | Best accuracy (%) |
|---|---|---|---|
| $(\alpha^r, \beta^r)$ | *k*-nn | $k = 3$ | 58.20 |
| $(\alpha^r, \beta^r)$ | SVM (RBF kernel) | $C = 2^2$, $p = 2^{2.85}$ | 57.67 |
| $(\alpha^r, \beta^r)$ | SVM (lin. kernel) | $C = 2^0$ | 49.74 |
| $(\alpha^r, \beta^r)$ | Naive Bayes class. | $Nbins = 29$ | 56.08 |
| $(\alpha^r, \beta^r)$ | C4.5 | Conf. fact. = 0.3, trials = 1 | 52.91 |

**Table 14** Relational space accuracies obtained by the proposed algorithm

| $\alpha$ | $\beta$ | Accuracy (%) | Acc(exhaustive) | Ideal classifier |
|---|---|---|---|---|
| 14, 15 | 14, 15 | 29.10 | 29.10 | 43.91 |
| 15, 14 | 15, 14 | 29.10 | 29.10 | 43.91 |
| 14, 15, 18 | 18, 14, 15 | 36.51 | 36.51 | 46.56 |
| 14, 11, 15, 18 | 18, 11, 14, 15 | 45.50 | 46.03 | 52.91 |
| 14, 11, 15, 18 | 18, 14, 11, 15 | 45.50 | 46.03 | 52.91 |
| 13, 14, 11, 15, 18 | 18, 11, 14, 15, 13 | 51.85 | 60.85 | 55.03 |
| 14, 11, 15, 18, 13 | 13, 18, 11, 14, 15 | 51.85 | 60.85 | 55.03 |
| 14, 11, 15, 1, 18, 13 | 13, 18, 11, 14, 15, 1 | 53.44 | – | 57.67 |
| 14, 11, 15, 1, 18, 13, 10 | 13, 18, 11, 14, 15, 1, 10 | 59.26 | – | 62.96 |
| 14, 11, 15, 1, 18, 13, 10, 9 | 13, 18, 11, 14, 15, 1, 10, 9 | 58.20 | – | 64.55 |

In the penultimate column, we show the best accuracy obtained by a code identified by exhaustive search. A dash means that exhaustive search was prohibitively slow. In the final column, we give the accuracy of the ideal classifier for the set of antigens used. Note that the accuracy in the penultimate column may refer to a different antigen combination from the one listed on the left, and that is why it may be higher than that of the ideal classifier, which refers to the same antigen combination as identified by the algorithm

possible minimum error is with which the available data could be classified. Because of a large number of subjects characterised by the same pattern of nil expression of the selected antigens, this maximum classification accuracy was found to be 63% for the case of using the 7 principal antigens. The classification accuracy we obtained using the relational space was 58%, i.e. it approached this ideal accuracy much closer than any of the other methods considered (which had accuracies of the order of 40%). Although the ideal accuracy, when using all 18 antigens, is as high as 83%, this accuracy is not really attainable, as the extra genes that allow the differentiation when we are talking in idealistic terms do not actually have any specificity for any particular group of subjects. Such accuracies are obviously not high enough for a real practical application. However, the methodology improves significantly the original raw data classification accuracy. It is possible that the improvement in classification accuracy at levels appropriate for a practical system will come from the inclusion of further, more discriminatory raw measurements. In any case, our results demonstrate that the use of a relational space may improve on the classification results of individual feature spaces.

The methodology we presented in constructing the relational space is just an example. Alternative ways of defining relationships between individuals may be devised, which may prove to lead to better classifiers. However, the concepts we introduced here, namely that of relational space, binary coding for feature construction and principal antigen are generic and they may be used for analysing other data sets and developing different classification schemes.

**Acknowledgments** The authors are grateful to Julian Gray, Giorgio Mazzoleni and Andrea Crisanti of the Department of Biological Sciences of Imperial College London, for supplying the data. Paolo Pintus is grateful to Fabio Roli and Luca Citi for useful advise. This work was also partly supported by the EPSRC portfolio grant ''Integrated Electronics''.

## References

1. Alippi C, Roveri M (2007) Reducing computational complexity in *k*-NN based adaptive classifiers. In: Proceedings of the IEEE international conference on computational intelligence for measurement systems and applications (CIMSA 2007), Ostuni, Italy, 27–29 June 2007
2. Breman J (2001) The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. Am J Trop Med Hygiene 64(1–2 Suppl):1–11
3. Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2(2):121–167
4. Boser BE, Guyon I, Vapnik V (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory. ACM Press, pp 144–152
5. Chakrabarti A, Chazelle B, Gum B, Lvov A (1999) A lower bound on the complexity of approximate nearest-neighbor searching on the hamming cube. In: Proceedings of the 31st annual ACM symposium on theory of computing (STOC99), pp 305–311
6. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm
7. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
8. Gray JC, Corran PH, Mangia E, Gaunt MW, Li Q, Tetteh KKA, Polley SD, Conway DJ, Holder AA, Bacarese-Hamilton T, Riley EM, Crisanti A (2007) Profiling the antibody immune response against blood stage malaria vaccine candidates. Proteomics Protein Markers Clin Chem 53(7):1244–1253
9. Gunn SR (1998) Support vector machines for classification and regression. Technical Report. Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, University of Southampton
10. Hay S, Guerra C, Tatem A, Noor A, Snow R (2004) The global distribution and population at risk of malaria: past, present, and future. Lancet Infect Dis 4(6):327–336
11. Otsu N (1979) A threshold selection method from grey level histograms. IEEE Trans Syst Man Cybern 9:62–66
12. PRTools, version 4.0.14, 4 Mar 2005, http://prtools.org/
13. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Francisco
14. Quinlan JR. C4.5 Release 8. http://www.rulequest.com/Personal/. Accessed 8 June 2011
15. Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI (2005) The global distribution of clinical episodes of *Plasmodium falciparum* malaria. Nature 434(7030):214–217
16. Theodoridis S, Koutroumbas K (2009) Pattern recognition. Academic Press, London
17. Vaidya PM (1989) An O($n$ log $n$) algorithm for the all-nearest-neighbors. Problem Discrete Comput Geom 4:101–115
18. Welling M. Fisher linear discriminant analysis. Department of Computer Science, University of Toronto, Canada. http://www.ics.uci.edu/~welling/classnotes/papers_class/Fisher-LDA.pdf. Accessed 8 June 2011
19. WHO (2010) Guidelines for the treatment of malaria. ISBN: 9789241547925. http://www.who.int/malaria/publications/atoz/9789241547925/en/index.html. Accessed 8 June 2011
20. WHO (2008) World malaria report. ISBN: 9789241563697. WHO reference number: WHO/HTM/GMP/2008.1. http://www.who.int/malaria/publications/atoz/9789241563697/en/index.html. Accessed 8 June 2011