BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 20 2011, pages 2782–2789
doi:10.1093/bioinformatics/btr476

Sequence analysis

Advance Access publication August 16, 2011

Identifying associations between amino acid changes and meta information in alignments

L. Spangenberg[1], F. Battke[2], M. Graña[1], K. Nieselt[2] and H. Naya[1,*]

[1]Bioinformatics Unit, Institut Pasteur Montevideo, 11400 Montevideo, Uruguay and [2]Integrative Transcriptomics, Center for Bioinformatics, University of Tübingen, Sand 14, 72026 Tübingen, Germany

Associate Editor: Martin Bishop

ABSTRACT

Motivation: We present a method that identifies associations between amino acid changes in potentially significant sites in an alignment (taking into account several amino acid properties) with phenotypic data, through the phylogenetic mixed model. The latter accounts for the dependency of the observations (organisms). It is known from previous studies that the pathogenic aspect of many organisms may be associated with a single or just few changes in amino acids, which have a strong structural and/or functional impact on the protein. Discovering these sites is a big step toward understanding pathogenicity. Our method is able to discover such sites in proteins responsible for the pathogenic character of a group of bacteria.

Results: We use our method to predict potentially significant sites in the RpoS protein from a set of 209 bacteria. Several sites with significant differences in biological relevant regions were found.

Availability: Our tool is publicly available on the CRAN network at http://cran.r-project.org/

Contact: naya@pasteur.edu.uy

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 27, 2011; revised on July 25, 2011; accepted on August 9, 2011

1 INTRODUCTION

Next-generation sequencing has dramatically increased the amount of biological data available. This technology has made sequencing cheaper and faster, while stimulating the development of several methodologies. The number of sequenced genomes has increased dramatically over the past few years, reaching 1815 sequenced genomes at the time of writing (Genome Project database at the NCBI). In bacteria, several strains are sequenced from different organisms with the number of available genomes steadily increasing. At this time, >3800 strains corresponding to >400 species are being sequenced (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, last accessed date April 25, 2011).

The multitude of genomes gives the possibility to identify orthologous groups of proteins among many closely related strains. For very similar strains, thousands of these orthologous groups may exist, implying a dimension expansion of problems involving alignments. Indeed, aligning thousands of proteins from hundreds of strains, exploring the alignments and identifying interesting sites is a daunting task. When dealing with very large alignments, the overview is easily lost. This leads to the idea of establishing a more automated way of finding significant positions in the alignments, that are sites in the protein in which amino acid changes could alter the function.

In general, multiple sequence alignments identify regions of similarity that may result from functional, structural or evolutionary relationships between the sequences and frequently a mixture of them. If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations and gaps as insertions or deletions introduced in one or both lineages since they diverged from that ancestor. Hence, behind a multiple sequence alignment of orthologs lies a phylogeny. When analyzing alignments in order to perform association studies with some meta-information, such as specific phenotypic data, one should therefore consider the phylogenetic relationship between the organisms. The method presented here identifies associations between amino acid changes in 'interesting' positions in an alignment (taking into account several amino acid properties) with some meta-information (e.g. phenotypic data). The proposed method has general applicability to other organisms, different amino acid properties, different metadata. As a motivating example, we applied it to a set of 209 bacterial strains belonging to several genera (72 genera, 117 species) with the aim of finding amino acid changes that might be correlated with the pathogenicity of the bacteria.

The pathogenicity character is mainly determined by the presence of specific genes, such as toxins, often referred to as virulence genes. However, several studies have shown that the pathogenicity character of different bacterial strains is determined by changes in amino acids causing changes in protein structure, and hence function (Conenello *et al.*, 2007; Marjuki *et al.*, 2010; Sokurenko *et al.*, 1998). Thus, the pathogenicity character can also be conferred by specific genetic variations having an effect on protein function and not solely by the presence or absence of virulence factor genes as previously assumed (Falkow, 1997). We have assembled a screening method that identifies significant sites in an alignment (which might confer the pathogenicity character to some bacteria) through the application of linear mixed models on different amino acid properties in each of those columns.

Amino acid properties can be grouped according to many different characteristics, such as size, polarity, alpha helix or beta sheet propensity. Substitutions severely changing the value of some key properties (e.g. from polar to non-polar) tend to have a stronger effect on the tertiary structure, and probably on the function of the protein.

*To whom correspondence should be addressed.

If those substitutions are associated with the label (metadata), they define an 'interesting' alignment position, which might be responsible for pathogenicity. The amino acid properties considered depend on the specific problem as will be discussed later.

Our method is divided into several steps, which we shall outline below. Additionally, we provide results from a real biological example. Finally, we discuss the flexibility of the method as well as potential applications.

2 METHODS

2.1 Overview

As explained above, pathogenicity can result from a mutation of some specific amino acids of the protein. Such a mutation could alter some local structural feature of the protein, with potential functional impact. In order to find this variation, we need to look for potentially 'interesting' columns in alignments. To those columns, given the class label (metadata) of the organisms and some relevant amino acid properties, (generalized) linear mixed models are applied to determine the effect of some relevant amino acid properties on pathogenicity. The method can be summarized as follows:

(1) Find orthologous groups of n genomes where the label of the organism (such as pathogen/non-pathogen) is known.

(2) For each orthologous group, let A be the alignment of n genomes restricted to the group.

(3) Filter out irrelevant columns, keeping columns with sufficient changes. Only consider these columns in the following steps.

(4) Apply a phylogenetic mixed model to each column of the subalignment on each relevant amino acid property using the phylogenetic relationship matrix (from the phylogenetic tree) to account for non-independent observations.

(5) Identify for each column the correlation of the label and the amino acid property, generally via the contrast of 'fixed' effects.

(6) Summarize results and compute statistical measures.

The approach was implemented in the bcool R package available at the CRAN network http://cran.r-project.org/.

2.2 Applying the phylogenetic mixed model

Orthologous groups are determined for the proteins of interest, and alignments are generated. Before proceeding with each alignment, we discard irrelevant columns from the alignment. First, the columns without variance, mainly the perfectly conserved sites, are discarded. In addition, we filter out sites with a high number of gaps. On each of the remaining columns, a linear mixed model is applied. The phylogenetic mixed model of Lynch (PMM) partitions phenotypic values into three components:

$$\vec{Y} = \mathbf{X}\vec{\beta} + \mathbf{Z}\vec{a} + \vec{e}, \tag{1}$$

where $\vec{Y}$ is the vector of observations (the dependent variable), $\vec{\beta}$ is the vector of fixed effects, $\vec{a}$ is the vector of phylogenetic heritable additive effects and $\vec{e}$ is the vector of independent and identically distributed residual errors. $\mathbf{X}$ is the incidence matrix that associates effects with observations. The number of columns of $\mathbf{X}$ are the number of fixed effect levels one wants to consider. $\mathbf{Z}$ is the matrix that associates additive effects with observations. Both of them, $\mathbf{X}$ and $\mathbf{Z}$, are matrices relating the observations $\vec{Y}$ to the regressors $\vec{\beta}$ and $\vec{a}$.

Equation (1) is applicable to very general cases, especially $\vec{\beta}$ could be a vector holding the regressors for many different fixed effects and several link functions can be used for Y, extending the theory to generalized linear mixed models. In the case of a binary labeling (such as pathogen/non-pathogen), $\mathbf{X}$ contains the labels of the organisms, thus having a dimension of $n \times 2$ (n being the number of organisms considered). Vector $\vec{X}_{i1}$ corresponds to the pathogens and it holds $x_{i1} = 1$ for pathogens, and $x_{i1} = 0$ for non-pathogens. Vector $\vec{X}_{i2}$ stands for non-pathogens and it holds $x_{i2} = 0$ for pathogens and $x_{i2} = 1$ for non-pathogens.

Each y_i is the value of the amino acid property considered in the organism i. $\mathbf{Z}$ is the matrix relating species to observations and in our case corresponds to a diagonal matrix of dimensions $n \times n$. Random effects are normally distributed with mean 0 and variance matrices $\mathbf{R}$ and $\mathbf{G}$, corresponding to residual and additive effects, respectively. In the univariate case, $\mathbf{R} = \mathbf{I_n} \times \sigma_e^2$ and $\mathbf{G} = \mathbf{A} \times \sigma_a^2$, σ_e^2 and σ_a^2 standing for residual and additive variances. The $\mathbf{A}$ matrix represents the phylogenetic relations between the n organisms. It holds evolutionary 'time' values t_{ij} representing the time that organism i shared with organism j before speciation. The a_i and e_i values stand for random organism effects and the error term for each organism, respectively. These two vectors, and the fixed effects, $\vec{\beta}$, are the ones to be estimated. $\vec{\beta}$ has dimension 2 (β_p: pathogen, β_{np}: non-pathogen) in our binary case, since we are calculating the fixed effects of the pathogenicity. Henderson and coworkers (Henderson, 1949) found an approach to efficiently estimate the fixed and random effects via BLUE (best linear unbiased estimator) and BLUP (best linear unbiased predictor), when the variance of $\vec{a}$ and $\vec{e}$ are known. However, this is not the most general scenario which, in effect, is not applicable to our specific case. Hence, estimation methodologies are applied for inferring the values of $\vec{\beta}$, $\vec{a}$ and $\vec{e}$ together with the variances. Bayesian and frequentist approaches are usually applied, each of them with their advantages and disadvantages (Blasco, 2001). In this work, a Bayesian approach similar to the one presented by Naya *et al.* (2006) is chosen, hence not just a single value for the difference between $\vec{\beta}_p$ and $\vec{\beta}_{np}$ is determined, but a posterior probability distribution. Our package makes extensive usage of the main function implemented in the MCMCglmm package (Hadfield, 2010).

As stated above, we filtered non-significant columns and we apply a PMM to each of the remaining for each amino acid property. Let us say p 'interesting' columns are found (after filtering or with Analysis of Variance (ANOVA), entropy or conditional entropy, see Section 2.5) and m amino acid properties are considered relevant for the labels of the n organisms (e. g. pathogen, non-pathogen). For each column $1 \ldots p$ and each property $1 \ldots m$ Equation (1) is solved, each combination giving a different $\vec{Y}$ vector, providing estimations of $\vec{\beta}$, $\vec{a}$ and $\vec{e}$ (Fig. 1). Instead of working with two distributions, $\vec{\beta}_p$ and $\vec{\beta}_{np}$, we are interested in the difference $\vec{\beta}_{\text{diff}} = \vec{\beta}_p - \vec{\beta}_{np}$. The processing of the columns, that is, applying the PMM to each of the p $n \times m$ matrices, generates $m \times p$ posterior distributions of the fixed effects. These are further processed for significance determination, as explained in the next section.

2.3 Finding the really significant columns

In our case, m distributions (fixed effects) are calculated for each column. A way of summarizing the data can be seen in Figure 2. In the matrix $\mathbf{M}$, rows represent the m properties, while the p columns represent the interesting sites in the alignment. Each entry M_{ij} holds a selected summary statistic, for example the number of times (iterations) the difference of $\vec{\beta}_p$-$\vec{\beta}_{np}$ is >0 (gt0_{ij}) for property i in column j. Also, P-values from a t-test (or other tests) might be used as a significant summary statistic. Here, we focus on gt0. Since gt0 displays the proportion of the distribution being >0, not only large values of gt0 represent a large difference between $\vec{\beta}_p$ and $\vec{\beta}_{np}$, but also small values ($\vec{\beta}_{\text{diff}}$ negative values). Hence, very large or very small gt0 values are relevant, since they show a significant difference between the two labels. In order to assess the global significance of the alignment position based on the gt0 of each property, in a first step a transformation is performed: $T_{ij} = 2 \cdot (\text{gt0}_{ij} - 0.5)$, to center the values. T_{ij} values close to 1 and -1 are significant. One could simply set a significance level, say 0.9 and -0.9, and consider the entries above or below that number significant. Here, we summarize all entries corresponding to one column even further:

$$S_{T_j} = \frac{\sum_{i=1}^{m} T_{ij}^2}{m},$$

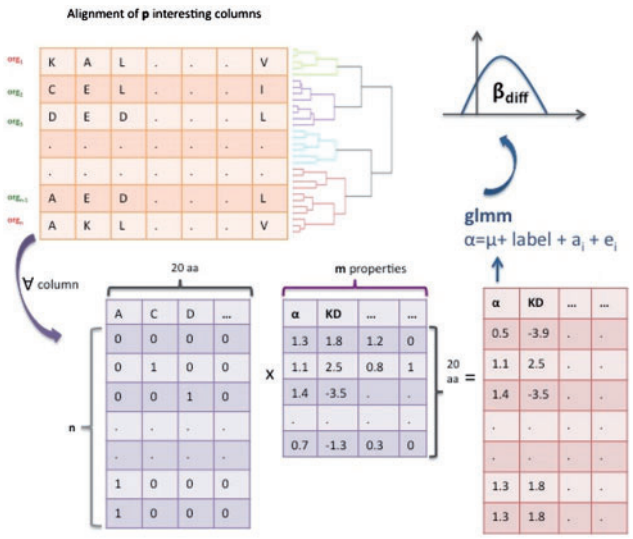

Fig. 1. Scheme of the method: n organisms are being studied and m amino acid properties are considered relevant for the determination of the organisms' labels (green and red). One preselects p 'interesting' columns. For each column, one obtains an $n \times m$ matrix representing the values of the property for each amino acid in the column (rightmost matrix). On this matrix, a phylogenetic mixed model is applied, obtaining as a result estimations for the $\vec{\beta}_{\text{diff}}$ vector. In this case, $\vec{\beta}_{\text{diff}}$ is the difference of the distributions of pathogen and non-pathogen. m such distributions are determined for each of the p columns, resulting in an $m \times p$ matrix.

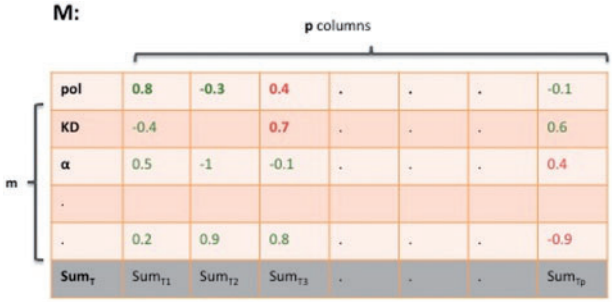

Fig. 2. For each column, a statistical value is calculated and a summary statistic is determined. S_{T_j} represents the relevance of the column j, large S_{T_j} values (close to 1) suggest a crucial alignment column.

S_{T_j} lies between 0 and 1. The larger the value, the more significant the column. A value of 1 suggests that all properties have values of 1 or -1 in that column, hence the relevance of the column is very high.

One would like to know the distribution of the score in order to be able to assess a significance value for the column. It is easy to see that a linear relationship of S_{T_j} with a standard χ^2_k distribution exists: under the null hypothesis of no differences between fixed effects ($\vec{\beta}_{\text{diff}} = 0$), gt0 has a binomial distribution, with success probability $P = 0.5$. For big k (number of iterations), $\frac{\text{gt0}-P}{\sqrt{\frac{PQ}{k}}}$ follows a standard normal distribution. It follows that

$$\frac{\text{gt0}_{ij} - P}{\sqrt{\frac{PQ}{k}}} = 2\sqrt{k} \cdot (\text{gt0}_{ij} - 0.5) = \sqrt{k} \cdot T_{ij}.$$

The sum of squares of standard normally distributed variables is χ^2 distributed with degrees of freedom equal to the number of summands. Summarizing for each column we obtain

$$\text{S'} = \sum_{j=1}^{m} \left(2\sqrt{k} \cdot (\text{gt0}_j - 0.5)\right)^2 = k \cdot \sum_{j=1}^{m} \left(2 \cdot (\text{gt0}_j - 0.5)\right)^2,$$

which is χ^2_m distributed with m degrees of freedom. Normally k corresponds to the number of iterations in each MCMC run, which is the same for each property. However, since autocorrelation is expected in each run, the real number of independent iterations becomes smaller. Hence, we also calculated here the S' score with the effective number of iterations $k_{m,j}$, so the formula becomes a little different: since k is changing with the property and site ($k_{m,j}$) it cannot be factored out from the sum.

Since S' has a known distribution, statistical significance can be assessed. Nevertheless, when testing the significance on real data, almost all columns were considered significant. That is, the null hypothesis (proportion of differences > 0) is too weak to represent the biological problem under investigation, making the consideration of the score's significance impractical. Still, S_{T_j} is intuitive and easy to understand, since it ranges from 0 to 1. Moreover, the sole fact of a column being statistically significant does not mean that it is biologically significant. Hence, the effect size has to be measured as well. In addition to the S_T score, we calculated the effect sizes for each property by dividing the median of $\vec{\beta}_{\text{diff}}$ by the range of the corresponding amino acid property. These values (for each of the five considered properties per column) were squared, then added together and finally the square root was taken. This value correspond to the l^2 norm for the effect sizes. When analyzing the correlation between the S_T score and the effect sizes one finds high correlations, as discussed below.

2.4 Permutation test for significance assessment

We implemented a simple permutation test to check for significance on the obtained S_T scores. For this, we randomly permuted the amino acids at a given position (and thus, the properties) and calculated the S_T score. This approach is computationally very expensive and only practical for few sites. However, as discussed below, the permutation test displayed beneficial properties that allows its usage for all potentially significant sites.

2.5 Alternative fast determination of potentially significant columns

Since our approach is computationally expensive for large alignments, particularly with evolutionary distant species, we considered three fast alternative methods that could lead to potentially similar results: (i) the entropy of the columns, (ii) the reduction of the conditional entropy and (iii) ANOVA P-values. The well-known entropy (Shannon, 1948) is associated with the degree of uncertainty of a variable. The logic behind considering entropy as a fast approach to identify potentially 'interesting' sites is that only the columns with enough variation can reasonably explain the differences between labels. Tightly related, the conditional entropy quantifies the remaining entropy of a random variable Y given that the value of another random variable X is known. In our case, for each column the random variable X is the label (the pathogenicity) of the organism and Y is the amino acid at the selected position in the alignment. Intuitively, the conditional entropy measures the amount of uncertainty that remains, e.g. how little you know about the amino acid frequency, when the pathogenicity of the organism is known. If the remaining conditional entropy is small, the label 'determines' the property of the amino acid. When high, the amino acid frequency is independent from the label. We calculated the Conditional Entropy Reduction (*CER*), a value between 0 and 1 for each column of the alignment. High *CER* values are promising, while low ones are less significant. It is important to note that these (entropy) approaches are independent of the amino acid properties selected.

In the same sense, we implemented an ANOVA approach. In its simplest form, ANOVA provides a statistical test of whether or not the means of several

groups are all equal. For each column and each property, one may calculate the F statistics and a P-value in order to evaluate the significance of the differences of variances within and between the groups. In our example, there are only two groups, pathogenic and non-pathogenic. The P-values are *log*-transformed and the mean and median of the properties are determined. Moreover, for each column the number of significant properties (P-values $<$ cutoff) is calculated. The latter together with the mean and median give an idea of which columns have differences in amino acid properties in each group (pathogens and non-pathogens). Since several ANOVA assumptions are violated, mainly the independence of the observations, results from this method can only be considered as a fast and rough approximation to the phylogenetic mixed model (PMM) endpoints.

2.6 Example dataset

Sigma factors are a family of proteins, which are a subunit of RNA polymerases in eubacteria and they provide the catalytic core RNA polymerase with the ability of promoter sequence recognition and initiation of specific transcription (Helmann and Chamberlin, 1988). In this work, the sigma factor σ^{38} (also known as σ^{S} or RpoS) was used to test the method. From the KEGG database (Kanehisa and Goto, 2000), all the orthologs for this gene were downloaded and a group of 209 organisms for which the (see Supplementary Material). The alignment of the 209 RpoS sequences was performed with MUSCLE (Edgar, 2004). Since we are incorporating the phylogenetic inertia into the calculations with the PMM, a phylogenetic tree is needed. For this, we chose seven genes present in all considered organisms: *secA*, *secE*, *secY*, *rpoS*, *srp54*, *ftsY* and *yidC*. Their alignments were also computed with MUSCLE and the tree reconstruction was performed with Phyml (Guindon *et al.*, 2005) based on the concatenation of these alignments using default parameters.

The amino acid properties considered were the residue accessible surface area in folded protein [RAS] (Chothia, 1976), relative mutability [MUT] (Jones *et al.*, 1992), hydropathy index [KD] (Kyte and Doolittle, 1982), normalized frequency of alpha-helix [ALPHA] and beta-sheet [BETA] with weights (Levitt, 1978).

3 RESULTS

In a first step, we explored the ability of the three fast scanning methods with two aims: (i) as fast alternatives to the PMM, and (ii) as fast initial filtering of columns to further apply the PMM. We defined S_{T_j} as the gold standard. For this reason, our method was applied to the 275 columns in the alignment of RpoS that remained after removing columns that were totally conserved or that contained >5% gaps. Subsequently, we calculated the S_T score for each column. Unexpectedly, ANOVA and CER performed extremely bad. The rank correlation between S_{T_j} and CER was -0.204 ($P < 10^{-3}$) while between ANOVA (median) and S_T it was 0.221 ($P < 10^{-3}$). On the other hand, the simple entropy approach displayed a rank correlation of 0.617 ($P < 10^{-15}$). However, despite the reasonable correlation value, only 25% of the columns can be safely ignored based on the entropy value. In the light of these results, it is clear that none of the previous fast methods reasonably predict the PMM results and therefore cannot be used in the initial screening procedure. Given this situation, we continued the analysis with the 275 columns in the alignment. However, as these methods are based on completely different assumptions and even theoretical foundations, it could be interesting to keep them at hand to compare predictions.

A cutoff corresponding to the 95 percentile of S_T was chosen in order to keep the most significant columns for further analysis, which led to 14 remaining columns. Table 1 shows those significant positions and the respective scores for each property as well as the S_T score. Furthermore, we predicted the secondary structure of RpoS using the PSIPRED server (Bryson *et al.*, 2005) applied to the RpoS protein from *Escherichia coli* as model. For each of the significant columns in Table 1, we denoted whether it lies within an alpha helix or coil region. In addition, the amino acids enriched (5% difference in the general frequency) at each position in pathogens and non-pathogens are shown.

Table 1. Properties computed for significant columns for the RpoS alignment

	Aminoacid properties								Enriched AA	
Position	RAS	MUT	KD	ALPHA	BETA	Score	2D	Reg.	in P	in NP
138	**0.005**	**−0.040**	**−0.006**	0.026	**−0.038**	0.868	H	1.2	L	I
146	−0.032	**0.033**	**0.046**	0.026	0.045	0.766	C	1.2	AV	PR
151	0.031	**−0.038**	−0.037	0.046	**−0.060**	0.822	H	1.2	QV	I
163	**0.080**	−0.028	**−0.078**	0.056	−0.044	0.802	H	1.2	AE	L
337	−0.030	0.010	**0.037**	−0.033	0.055	0.786	H	3	V	KR
341	**0.047**	0.010	**−0.057**	**−0,064**	0.034	0.768	H	3	IT	AGV
374	**−0.060**	**−0.036**	0.091	−0.045	0.053	0.889	H	3	AEY	DKRT
376	**0.026**	**0.091**	**−0.044**	**−0.079**	**0.059**	0.978	H	3	ST	L
419	**0.041**	−0.020	**−0.045**	**−0.075**	−0.010	0.747	H	3	DR	AEGK
427	**−0.012**	**0.017**	**0.019**	−0.040	0.045	0.924	C	3	V	L
429	−0.017	**0.039**	0.022	−0.028	**0.074**	0.756	C	3	T	LS
462	**−0.021**	**−0.026**	**0.028**	0.023	−0.029	0.751	C	3	E	S
465	**−0.016**	**0.045**	**0.027**	−0,055	**0.084**	0.942	H	3	TV	AL
514	0.018	**−0.064**	−0.016	0.045	**−0.066**	0.868	H	4.2	L	IV

Significant scores are shown in bold face. 2D, secondary structure (helix or coil); Reg., region in the RpoS protein (as described in the text); amino acids enriched in pathogenic (P) and non-pathogenic (NP) strains are given in the rightmost columns.

In order to assess the significance of the S_T scores, bootstrapping was performed, permuting the properties in each bootstrap. It can easily be appreciated that the distribution of bootstraps for intermediate or high scores are similar (Fig. 3).

A different approach to find 'interesting' sites come from the effect sizes. Figure 4 shows the l^2 values per position, smoothed considering the neighbor positions with a DEMA function (double exponential moving average). That is, for each position we consider the whole 'region' ($n = 3$, the adjacent neighbors at each side) and average the l^2 values weighted by a double exponential according to the distance. We permuted the positions and calculated the DEMA for each position, taking only the maximum into account before permuting positions again. Three different peaks were observed over the 97.5 percentile of the bootstrap, corresponding to sites 96–98, 374–376 and 510–514 (Fig. 4). Table 2 shows, for each of these peaks, the difference of the effect size (pathogen−non-pathogen) of each property, in order to investigate the contribution of each of (see Supplementary Material for a complete list of effect sizes).

Cumulative effect sizes were calculated in order to observe overall tendencies along the alignment. Figure 5 shows the cumulative sum of the effect size in each of the five properties along the positions in the alignment. The cumulative effect sizes were calculated based only on the significant positions. Were the distribution of the effect sizes random, meaning that no tendency in pathogens or non-pathogens would be detectable, a line oscillating around zero should be observed (similar to the dotted one).

Finally, in our example, an alignment with 209 species, 275 relevant columns and 5 properties took 8 h for running 10^5 iterations in a MacBook Pro with Intel Core i7 processor (2.66 GHz) and 8 GB of RAM. While the three approaches we assayed for quick

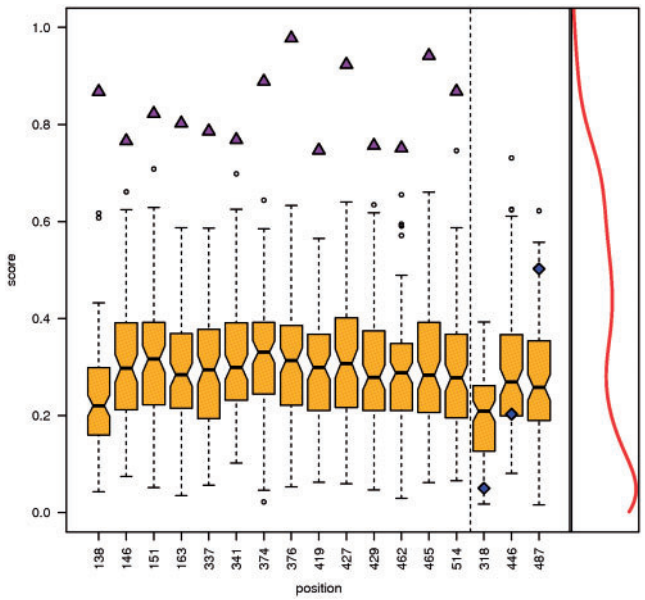

Fig. 3. The box plots represent the scores calculated from 100 bootstraps for all positions with significant scores (Fig. 2), and three arbitrarily chosen positions with insignificant scores (to the right of the dashed line). The triangles and rhomboids are the actual S_T scores. While triangles (significant scores) lie above the bootstrap distribution, rhomboids (insignificant scores) are falling into the distribution range. This is in agreement with our assessment of significance based on S_T. The empirical distribution of S_T is plotted as a rotated density plot on the right, i.e. the the *y*-axis represents the score's values while the *x*-axis represents their abundance.

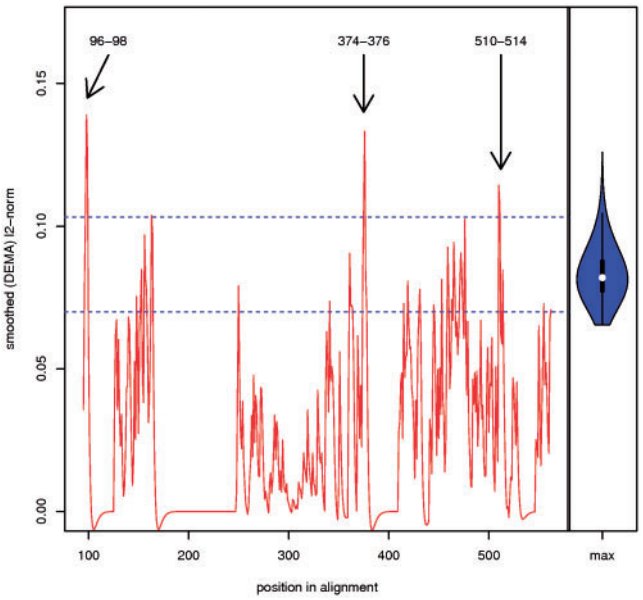

Fig. 4. Effect size per position in the alignment, smoothed using the DEMA function. Horizontal dashed lines represent the 2.5 and 97.5 percentiles of the maximums' bootstrap distribution. Peaks above the upper dashed line were considered significant sites. To the right, we show the distribution of the maximum values obtained by each bootstrap in a violin plot (basically a box plot that also shows the probability density: the white circle indicates the median, the central black box and lines extending from it indicate the interquartile range (IQR) and the 1.5-IQR, respectively; a kernel estimate of the probability density is plotted as a blue area to the left and right of the box plot). The marked peaks are all well above the 97.5 percentile and two of them even greater than the bootstrap maximum.

selection of significant columns failed to select good candidates (which precludes their usage in the filtering step), a two stages approach of the PMM can be used with straightforward results. The rank correlation between the final S_T and estimates obtained at 1000, 5000 and 10000 iterations was always >0.960 (data not shown). The number of columns to include from a first quick run to retain the final 1% to 5% top values was below the double for 5000 and 10 000 iterations. This suggest a first run with 10 000 iterations, selecting a number of top-ranking columns that double the desired final proportion of sites of interest and doing the final run only for this set.

4 DISCUSSION

In the present work, we have shown how linear mixed models may serve to pinpoint 'interesting' residues in protein alignments, associated with meta-information labels. Our method is closely connected to the PMM (Lynch, 1991), albeit cast within a Bayesian framework, as previous work has put forward (Naya *et al.*, 2006). Linear mixed models are extremely flexible, being particularly well suited to account for non-independence of observations, especially when dealing with closely related taxa. We applied the PMM, using five amino acid descriptors of our own choice. It is worth mentioning that while arbitrary, our selection is based on sensible physicochemical assumptions, under which their mutation is expected to impact a functional protein. The PMM was performed over the whole set of relevant columns. As a first step, we created a summary statistic (S_T) that allows assessing a site's significance, with all properties being jointly considered. Several studies divided the sigma factors (and the RpoS protein) into four regions with different roles (Gopal and Chatterji, 1997; Gruber and Gross, 2003; Ohnuma *et al.*, 2000; Reddy *et al.*, 1997). The highest S_T score observed in Table 1 was 0.978, which corresponds to a residue in Region 3 of RpoS' structure (position 376, 202 in the reference *E.coli* RpoS protein). This region is involved in recognition of the −10 promoter elements (subregions 2.3, 2.4 and 3). Other interesting residues in Region 3 are 465 and 427 with scores of 0.942 and 0.924, respectively. While columns 376 and 465 are predicted to lie in alpha-helix secondary structures, column 427 belongs to a coil region. Following the S_T scores, Region 3 harbors most residue changes associated with pathogenicity. It is noteworthy that leucine is among the residues enriched at these positions in non-pathogenic organisms. In pathogenic, valine is enriched for columns 337, 427 and 465. While both residues, valine and leucine, are aliphatic, they show strong differences in frequencies at alpha-helix and beta-sheet structures. Additionally, position 514 in the alignment (with score 0.868), corresponding to subregion 4.2, is very interesting. Recognition of the −35 promoter element is mediated by a helix-turn-helix unit in this subregion, and amino acids from this region may also provide a contact point for some activator proteins.

S_T scores near 1 are 'good', even though no statistical significance value is attached to them. Moreover, by ordering the scores increasingly one can rank the columns, from less to more 'interesting', analyzing only the most 'interesting' of them.

Table 2. Significant sites (peaks) with respect to effect sizes

Position	RAS	MUT	KD	ALPHA	BETA	RpoS region
96	–	–	–	0.098	–	1.1
97	–	−0.080	–	–	–	1.1
98	–	–	−0.084	–	−0.092	1.1
374	−0.060	−0.036	−0.091	–	–	3
375	−0.094	–	–	–	–	3
376	0.026	0.091	−0.044	−0.079	0.059	3
510	0.113	–	−0.130	–	−0.089	4
511	–	0.028	0.072	–	–	4
513	–	–	−0.091	–	–	4
514	–	−0.064	–	–	−0.066	4

Values correspond to the effect difference between pathogenic and non-pathogenic species. Only the values of significant properties are shown.

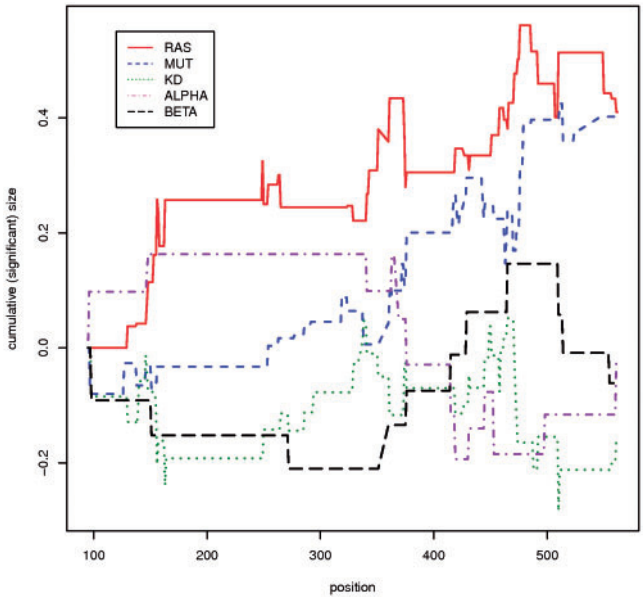

Fig. 5. Cumulative effect sizes per site in alignment. Only significant changes (according to the gt0 value) were included. Each line corresponds to one amino acid property.

In our example, the χ^2 values calculated were almost all significant, even using the correction for the effective number of samples in MCMC runs and correcting for multiple testing (data not shown). This problem, probably associated with the very relaxed null hypothesis tested, makes this statistical approach impractical. However, techniques such as bootstrapping can be used as a proxy for the significance score. When we performed 100 bootstraps for the sites with highest S_T values, the distributions of values obtained in the bootstrapping process were well below the real scores (Fig. 3). On the other hand, when we performed the bootstrap for columns with low to intermediate scores the real values fell inside the bootstrap distribution, pointing to the expected non-significance. Furthermore, for medium to very high S_T values the bootstrap distributions were strikingly similar, which in principle allows a significant reduction in the number of columns to consider for the bootstrap process.

Statistical significance of the observed fixed effects is fundamental to avoid spurious correlations. However, the real biological significance of the observed changes is probably more associated to the effect size, that is, the relative difference in each of the considered properties. While correlation between S_T and effect sizes were reasonably high and significant for all considered properties (RAS: 0.651, MUT: 0.610, KD: 0.710, ALPHA: 0.631 and BETA: 0.641), an alternative view could emerge if candidate selection is based on effect sizes. For this reason, we calculated the l^2-norm for each column in the alignment using the (standardized) effect sizes. Three peaks were significant as shown in Figure 4, corresponding to positions 96–98, 374–376 and 510–514. The change in RAS in region 374–376 is noteworthy, and the change in all properties for column 376 is of particular interest (Table 2). Residues at position 510 in the alignment showed an increase in RAS together with a decrease of hydropathy in pathogenic organisms. While the effect sizes of known reported cases are bigger than that observed in our case (e.g. Marjuki *et al.*, 2010), it is worth mentioning that we are detecting weaker signals present in a large number of organisms spread across very distant phylogenetic groups, hence smaller effect sizes are, in fact, expected. The RAS and mutability are consistently larger in pathogenic organisms as compared with their non-pathogenic counterpart. Furthermore, given the observed pattern we analyzed the mean RAS per organism with a linear mixed model and found a significant difference between pathogens and non-pathogens (mean difference 0.364, 95% confidence interval 0.161–0.568; data not shown). Additionally, for pathogens high mutability may constitute a selective advantage, e.g. to cope with a changing host environment. The behavior of other properties is more erratic, with increases and decreases alternating across the protein sequence.

Several points in our approach are worthy of discussion. First, the amino acid properties are usually limited to 20 possible values. In general, they are measured in a continuous scale, but the limited set of values (few amino acids are in each site in the alignment) suggests that a categorical representation is an interesting alternative. However, even with discrete data, the simple linear mixed model performs almost equivalent to more complex models based on generalized mixed model theory, with the enormous advantage of the simplicity for interpretation (Peñagaricano *et al.*, 2011). Second, while we focused our analysis on the 'fixed effects', essentially the difference in the properties between pathogens and non-pathogens, the 'random effects' and particularly 'additive effects' carry very important information regarding tendencies along the phylogenetic structure. Third, phylogenetic heritability (h^2) is a key component in the PMM. Heritability, usually defined as the proportion of phenotypic variance attributable to 'additive effects', indicates the importance of the phylogenetic information with regard to the estimation process. Values of h^2 near 1 indicate that characteristics are highly dependent on the phylogenetic structure, while values near 0 indicate independence of the phylogeny. In our case, the first quartile of h^2 was >0.75 for all properties, pointing toward the general importance of taking into account phylogenies in the linear model. Fourth, the computational performance of our method is adequate for medium size problems and a two-stage approach of the PMM can be used to reduce the computational burden. Moreover, our approach is trivially parallelizable, both in properties

and columns, giving a broad set of possibilities through several R packages (doMPI, Rmpi, multicore) or even manually splitting the work. Fifth, our approach is statistical. Faber and coworkers showed that a change in a single amino acid in a glycoprotein in the Rabies Virus, the mutation of asparagine at position 194 to lysine, enhances virulence and virus spread (Faber *et al.*, 2005). However, such an example would not be suited for our method, because there are essentially only two observations and then few possibilities to apply statistics.

Sixth, while we selected five properties based on our experience, there are an enormous number of different properties to combine. However, a large number of different properties are highly correlated (Kawashima *et al.*, 2008), clustering in only six groups: α and turn propensities (A), β propensity (B), composition (C), hydrophobicity (H), physicochemical (P) and other properties (O) (Tomii and Kanehisa, 1996). Based on this clustering, several choices will render similar results, identifying the same sites.

Another interesting alternative is to consider pathogenicity as the dependent variable, while considering all sites' properties as predictors. This poses a typical problem of dimensionality, as there are usually more regressors than observations. However, some shrinkage techniques such as the Bayesian LASSO (de los Campos *et al.*, 2009; Park and Casella, 2008) can accommodate the dimensionality issue, producing stronger shrinkage of regression coefficients that are close to zero and less shrinkage of those with large absolute values. Major caveats with this approach are the potential overlooking of very similar sites and the difficulty in results interpretation as the Bayesian LASSO is well suited for dependent variable prediction, while our aim is identification of relevant sites. However, the possibilities this approach raises deserve further work.

Our approach helps to efficiently identify relevant columns in an alignment that might be associated with the label of the organisms considered, based on relevant amino acid properties. It could be argued that columns in an alignment coevolve, and are not independent from each other, so one should consider this when searching for significant positions. Even though we implemented the R package considering only a univariate model (columns being independent), it is easily extendable to bivariate (two columns) models. Coevolution might be incorporated that way. Furthermore, multivariate models are straightforward to implement under this framework. Complete regions might be considered via a 'repeated measure model', though at the cost of increased computing time. Even though we applied this method to examples with two organism labels, it is extendable to multiclass labels, and not limited to the case of pathogen/non-pathogen. The ability of the proposed method to cope with different scenarios lies in its flexibility.

5 CONCLUSION

We present a method based on PMM and amino acid characteristics that helps finding 'interesting' columns in a protein sequence alignment, which might be responsible for a structural change in a protein, possibly affecting its function. We presented an example of this method through an alignment of bacterial RpoS sequences, using five amino acid properties in order to distinguish which columns could be associated with the pathogenic character of some strains. Our method is extremely flexible and can be applied to all kinds of labels, univariate or multivariate, and even taking account of several effects together. Although we mainly explored the fixed effects, a lot of additional possibilities can be exploited from random effects. In the present work, we presented one variant of all these possibilities and implemented it as an R package.

ACKNOWLEDGEMENT

Part of this work was possible thanks to travelship assistance from the Bioinformatics Master program (PEDECIBA-Uruguay). We are indebted to Alexander Herbig, Héctor Romero, Natalia Rego, Gustavo de los Campos and three anonymous reviewers for helpful discussions on the manuscript.

Funding: ANII (Agencia Nacional de Investigación e Innovación), Uruguay (to L.S.).

Conflict of Interest: none declared.

REFERENCES

Blasco,A. (2001) The Bayesian controversy in animal breeding. *J. Animal Sci.*, **79**, 2023–2046.

Bryson,K. *et al.* (2005) Protein structure prediction servers at University College London. *Nucleic Acids Res.,* **33** , W36–W38.

Chothia,C. (1976) The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.*, **105**, 1–14.

Conenello,G.M. *et al.* (2007) A single mutation in the PB1-F2 of H5N1 (HK/97) and 1918 influenza A viruses contributes to increased virulence. *PLoS Pathog.*, **3**, 1414–1421.

de los Campos,G. *et al.* (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, **182**, 375–385.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Faber,M. *et al.* (2005) A single amino acid change in Rabies virus glycoprotein increases virus spread and enhances virus pathogenicity. *J. Virol.*, **79**, 14141–14148.

Falkow,S. (1997) What is a pathogen? Developing a definition of a pathogen requires looking closely at the many complicated relationships that exist among organisms. *ASM News*, **63**, 359–365.

Gopal,V. and Chatterji,D. (1997) Mutations in the 1.1 subdomain of Escherichia coli sigma factor sigma70 and disruption of its overall structure. *Eur. J. Biochem.*, **244**, 613–618.

Gruber,T.M. and Gross,C.A. (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.*, **57**, 441–466.

Guindon,S. *et al.* (2005) PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.*, **33**, 557–559.

Hadfield,J. (2010) MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R Package. *J. Stat. Softw.*, **33**, 1–22.

Helmann,J.D. and Chamberlin,M.J. (1988) Structure and function of bacterial sigma factors. *Annu. Rev. Biochem.*, **57**, 839–872.

Henderson,C.R. (1949) Estimation of changes in herd environment. *J. Dairy Sci.*, **32**, 706.

Jones,D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences.*Comput. Appl. Biosci.*, **8**, 275–282.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kawashima,S. *et al.* (2008) AAindex: amino acid index database, progress report. *Nucleic Acids Res.*, **36**, D202–D205.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Levitt,M. (1978) Conformational preferences of amino acids in globular proteins. *Biochemistry*, **17**, 4277–4285.

Lynch,M. (1991) Methods for the analysis of comparative data in evolutionary biology. *Evolution*, **45**, 1065–1080.

Marjuki,H. *et al.* (2010) Three amino acid changes in PB1-F2 of highly pathogenic H5N1 avian influenza virus affect pathogenicity in mallard ducks. *Arch. Virol.*, **155**, 925–934.

Naya,H. *et al.* (2006) Inferring parameters shaping amino acid usage in prokaryotic genomes via Bayesian MCMC methods. *Mol. Biol. Evol.*, **23**, 203–211.

Ohnuma,M. *et al.* (2000) A carboxy-Terminal 16-amino-acid region of σ^{38} of *Escherichia coli* is important for transcription under high-salt conditions and sigma activities in vivo.*J. Bacteriol.*, **182**, 4628–4631.

Park,T. and Casella,G. (2008) The Bayesian LASSO. *J. Am. Stat. Assoc.*, **103**, 681–686.
Peñagaricano,F. *et al.* (2011) Assessment of Poisson, Probit and linear models for genetic analysis of presence and number of black spots in Corriedale sheep. *J. Anim. Breed Genet.*, **128**, 105–113.
Reddy,B.V. *et al.* (1997) Recognition of promoter DNA by subdomain 4.2 of Escherichia coli σ 70: a knowledge based model of -35 hexamer interaction with 4.2 helix-turn-helix motif. *J. Biomol. Struct. Dyn.*, **14**, 407–419.
Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423 and 623–656.
Sokurenko,E.V. *et al.* (1998) Pathogenic adaptation of Escherichia coli by natural variation of the FimH adhesin. *Proc. Natl Acad. Sci. USA*, **95**, 8922–8926.
Tomii,K. and Kanehisa,M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, **9**, 27–36.