**Introduction**

Next generation sequencing has dramatically increased the biological data available. In bacteria, several strains are sequenced from different organisms with the number of available genomes steadily increasing. The amount of genomes allows identifying 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. Aligning thousands of proteins from hundreds of strains, exploring the alignments and identifying interesting sites is a daunting task. This leads to the idea of establishing a more automated way of finding significant positions in alignments, which are sites in the protein in which amino acid changes could alter the function. Multiple sequence alignments (MSA) identify regions of similarity that may result from functional, structural, or evolutionary relationships between the sequences. Hence, behind a MSA of orthologs lies a phylogeny. When analyzing alignments in order to perform association studies with meta information one should consider the phylogenetic relationship between the organisms. Our method identifies associations between amino acid changes in “interesting” positions in an alignment (taking into account several amino acid properties) with some meta information.

As an example, we applied the method to 209 bacterial strains 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. However, studies have shown that the pathogenicity of different bacterial strains can also be determined by changes in amino acids causing changes in protein structure, and hence function. Thus, the pathogenicity character can also be conferred by specific genetic variations having an effect on protein function and not solely by the presence of virulence factors. 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.

**Methods**

Pathogenicity can result from a mutation of some specific amino acids of a protein. Such a mutation could alter some local structural feature of the protein. In order to find this variation, we need to look for “interesting” columns in alignments. To those columns, given the class label 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 (PMM; Lynch 1991) 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 on CRAN. 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 and \( m \) amino acid properties are considered.
Figure 1: $n$ organisms are considered 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. On this matrix a PMM is applied, obtaining as a result estimations for the $\beta_{diff}$ vector (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.

**Finding the really significant columns**

In our case, $m$ distributions (fixed effects) are calculated for each column (e.g. $p$ columns). For each property and each column a summary statistic can be determined, for example the number of times the difference of $\beta_{diff}$ is greater than 0 ($gt0_{ij}$) for property $i$ in column $j$. Since $gt0$ displays the proportion of the distribution being greater than 0, both, large and small values of $gt0$ represent a large difference between pathogens and non-pathogens. In order to assess the global significance of the alignment position based on the $gt0$ of each property, a transformation is performed: $T_{ij} = 2 \cdot (gt0_{ij} - 0.5)$, to center the values. $T_{ij}$ values close to 1 and $-1$ are significant. We summarize all entries corresponding to one column even further:

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

$S_{Tj}$ lies between 0 and 1. The larger the value, the more significant the column. Moreover, we performed permutation tests for significance assessments.

**Results**

$\sigma^{38}$ (RpoS) sequences of 209 organisms were downloaded and aligned. 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 more than 5% gaps. Subsequently, we calculated the $S_T$ score for each column. 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. In addition, the amino acids enriched (5% difference in the general frequency) at each position in pathogens and non-pathogens were determined. The positions found were lying in biologically relevant sites of the protein, involved in the $-10$ and $-35$ promoter recognition.