Finding Correlations in Functionally Equivalent Proteins by Integrating Automated and Visual Data Exploration

Daniel A. Keim Daniela Oelke
University of Konstanz
{ keim, oelke}@inf.uni-konstanz.de

Royal Truman
Mannheim
royaltruman@yahoo.com

Klaus Neuhaus
Technical University of Munich
neuhaus@wzw.tum.de

Abstract

The analysis of alignments of functionally equivalent proteins can reveal regularities such as correlated positions or residue patterns which are important to ensure a specific fold and various cellular functions. Many approaches are found in the literature which try to identify correlated positions to predict the residues that are close to each other in the three-dimensional folded structure. However, the quality of the predictions remains disappointing. One of the problems is that the statistical correlation measures that were used cannot do justice to the underlying complex biological and physicochemical realities.
In this paper we evaluate the biological requirements for a correlation measure and explain why a completely automatic approach is unlikely to succeed. We then propose a novel and flexible criteria for correlation of residue positions in protein sequences, which can be optimized for different requirements. To apply this definition we developed the tool VisAlign that combines an automatic calculation of correlations with an interactive visualization. This allows the user to visually explore alternative alignments and thereby conveniently test various hypothesis and to detect regularities in the aligned sequences.

1. Introduction

In recent years the amount of protein sequence data has grown explosively, due mainly to the availability of automated sequencing machines. The need for powerful techniques to analyze the data is greater than ever, as whole new genomes are now rapidly becoming available. Much interesting information is hidden in the databases and possibly could answer many questions of scientific and medicinal interest. One of those questions is what restrictions are protein sequences subjected to. More and more functionally equivalent variants of protein sequences are known. What do they have in common? What residue patterns would be acceptable to ensure a specific fold and a certain function? We can compare known functionally equivalent proteins by aligning their sequences. This could reveal certain regularities such as the need for an amino acid with a specific physicochemical property and correlations between positions. As a general rule, acceptable spatial and electronic features of a protein can be satisfied by alternative ensembles of amino acid residues. Identifying all these by only visual inspection of regularities in aligned sequences is unrealistic. Therefore, automatic methods are needed to support the search. However, automatic methods usually require detailed background knowledge that is not yet existent. Furthermore, generally valid principles might not be universally applicable to all proteins. In our approach we try to combine the advantages of automatic calculations with an interactive visualization. The alignment is visualized as usual but the inspection is supported by an underlying algorithm that calculates correlations between the selected columns of the alignment and all the other columns. All the relevant parameters can interactively be changed by the researcher and immediate feedback is provided. This enables the user to easily test different hypothesis as well as to perform context-sensitive studies.

2. Problem Statement

Proteins are chains of amino acids. Twenty different amino acids are coded for by most genes, although the genetic code of some organisms can occasionally also code for additional amino acids. Amino acids differ from each other in their physicochemical properties. Figure 1 illustrates these properties. Not every combination of amino acids forms a functional protein. On the other hand, there is not only a single valid sequence for a specific function, since alternative variants can be found in nature. It seems that proteins are usually fairly robust to amino acid substitutions. In order for a protein enzyme to be functional a restricted combination of amino acids that compose an active site must be present. These key sites must be held in place by a suitable scaffold provided by other portions of the protein. This is only possible if a correct, three dimensional

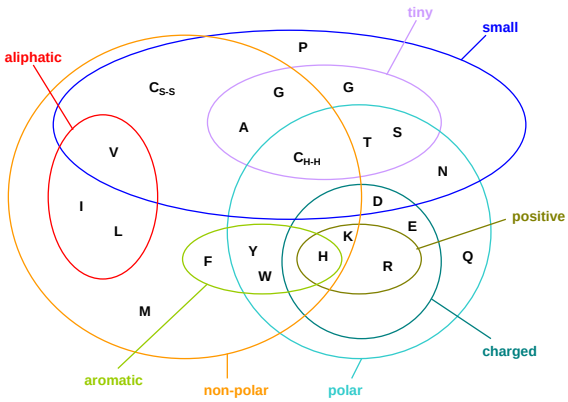


Figure 1: Physicochemical properties of the 20 amino acids (redrawn from [16])

```
21327007   NKSVYPMMSVFKVHQALALCNGFDN--KGISLDTLVNINRDKLDPKTWSPMLKDYSGPV---ISLTVRDL
56707801   ASFYFPICSTYKFLVVGAILKQSM--TDNKLLNQKIKISKN--QIVEYSPITRRHIN-----QIMTVKQL
61099446   ADQPFPMASTSKVAIAAVYLAGVDAG--------------------KWSLSSEWRLPRPG-GKYVPAKTL
16263522   GEQRFSLQSVMKVVVAAAVMQAVD---DRRIALGDRLTIRRGDLSVNIQPIADIVAERGS--FETSIGDL
23100527   ETKQMRSASIIKLFILASAYHLKE---KGKISLTDQIKLS----SNDFVQGSG-VISYLSDVKFLTYQQL
51892073   ARDPYLPASTFKLPVALCVLEAID---AGEMAWNTLVTYT----EEDYEPVGAGGFAQAAFGSRWTVRNL
79038994   GATSFPQGSLRRIWLGAVLLEAVD---QGELSLDQRVPLQ------TRARGPERHEQ---------VSAL
29376069   GQHEFYTASTIKVPLTMLVADTVAS---GQKKWTDLIPYNAEEDYEEGTGIIAYNIQP-----EYPLKTL
50954004   DHVVMPTASIGKVLLLVEVAARLQ----SGGLSSLALLDRAPQDAVGDSGIWQHLQVP-----ALPVADL
5738831    CDEPVVIASIFKVLLVLEFARQVA---AGQLDPRARVRVTAGDRLGGWGTAGCADDV------ELSLRDL
15806983   PDGVFPLASTYKQAVLWALLREFD---AGRISPNERFDVTPQNQSLGDYPYDGSNVR----------EL
62514959   VDQPFPAASLIKLGIAAFVKEKAAD---DPSQLERQVTLP---ESVGGAGILRFMSP-----QAWRVRDL
```

Figure 2: *Extract of an Alignment.* Identical or amino acids that are similar in their physicochemical properties are placed at the same location whenever possible.

fold occurs. Without the proper fold the protein sequence has no function [15].

In the three-dimensional folded state amino acids which are distant in the original linear polymer may now be close together. This fact may provide a working hypothesis for some discovered regularities among the aligned sequences which involve amino acids which are very far apart in the primary structure.

In order to dissect different proteins with the same function, their sequences are usually aligned. In a sequence alignment two or more sequences are arranged such that identical or similar amino acids are placed at the same location whenever possible. Adjustments can result in gaps within parts of the sequences. These are usually denoted by dashes. Figure 2 shows such an alignment. Some similarities between the sequences are obvious. For example there are some columns that are invariant. Those amino acids seem to be compulsory at this position for these proteins. Secondly, there are several columns in which more than one amino acid is found, but the number is much less than all 20 possibilities. An explanation might be that alternative amino acids which share a property such as similar size or

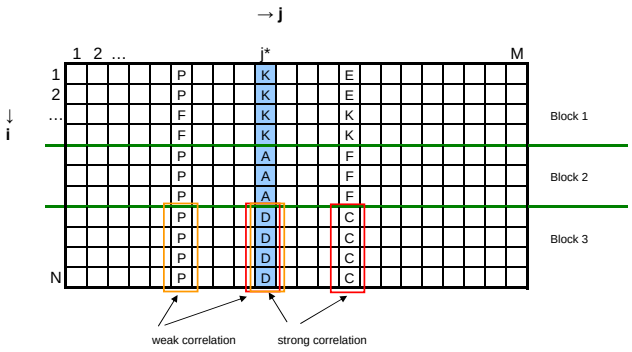


Figure 3: *Illustration of the used concepts.* j* denotes the selected basis column. A block is made up of all the sequences that have an amino acid of the same equivalence class value in the basis column(s). Correlations are searched for separately in each block, but the strength of a correlations depends on the amino acids of the whole column.

presence of an aromatic ring might be acceptable. However, experimental studies have shown that all possible combinations of seemingly acceptable variants won't work [14]. Although several alternative amino acids might be found in different aligned positions, not every combination of these are actually biologically functional [4, 17]. This suggests that the positions are not all independent of each other. As an example, various amino acids may be observed at two positions. Suppose that investigation shows they all happen to be small and very small amino acids. Combinations involving only small or only very small may indeed be inacceptable. Only the "small + very small" works, although the location of each may well be unimportant. The mutual effects compensate. Clearly, learning about the construction rules for a specific protein means learning about the underlying correlation logic.

The starting point of our analysis is an alignment of functionally equivalent protein sequences.

Definition 1 (Alignment Matrix)
An alignment matrix M is a matrix that consists of elements D_{ij} *so that*
$\forall_{i=1\ldots N}\forall_{j=1\ldots M}$:
$D_{ij} \in \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y,-\}$.
(see Figure 3 for illustration)

To take into account the similarity of amino acids, equivalence classes $E(D_{ij})$ can be defined.
e.g.:
$E_p(D_{ij}) \in \{polar,\ non\ polar\}$
$E_h(D_{ij}) \in \{hydrophobic,\ hydrophilic\}$
$E_c(D_{ij}) \in \{charged,\ non\text{-}charged\}$
$E_b(D_{ij}) \in \{basic,\ acidic,\ neutral\}$

$E_a(D_{ij}) \in \{$*presence, absence of aromatic ring*$\}$
$E_r(D_{ij}) \in \{$*presence, absence of reactive carboxylic acid*$\}$
Other equivalence classes and combinations of the above mentioned are possible.

Correlations are calculated between one or multiple basis columns and the remaining columns. So n! different combinations of n basis columns can exist. One may have to distinguish between whole columns that correlate and only some members of the columns. For example, the sequences of all non-viral ubiquitins were examined recently [18]. An interesting correlation was observed between residue positions 19, 24 and 57. At the first position, 19, one observed almost only the amino acids 'P' and 'S'; at the second location almost only 'E' and 'D'; and at position, 57, almost exclusively amino acids 'S' and 'A'. One of us reported [18] that partial columns correlated perfectly to reveal a biologically interesting principle. There are three classes of ubiquitins. Animals display the three-residue pattern 'PES', plants 'SDA' and fungi 'SDS'. The classification rule was so reliable, that an exception in the case of two parasite worms was immediately apparent. Further investigation revealed that the plant-like ubiquitin manufactured in this animal was injected into the target plant and modified its metabolic behavior. Research questions were then stimulated, such as whether these parasites manufacture both versions, and if so, how segregation of two such similar proteins could be accomplished to prevent interference. We propose that partial column analysis could reveal either phylogenetic relationships, or alternative protein design patterns.

The matrix is partitioned by the basis columns. In the following a single partition is denoted as a block.

Definition 2 (Block)
A block with respect to a basis column $j^* \in \{1 \ldots M\}$ *and an equivalence class value* $e \in E(D_{ij^*})$ *is defined as follows:*
$Block_{j^*}(e) = \{i \in \{1 \ldots N\} | E(D_{ij^*}) = e\}$
The set of all blocks with respect to basis column j^* *is:*
$Blocks_{j^*} = \{Block_{j^*}(e) | \exists i : e \in E(D_{ij^*})\}$
The definition can be generalized for multiple basis columns.

Correlations can now be considered as correlated columns of blocks. Intuitively a column in a block that is strongly correlated to the basis column would be one that contains only amino acids of the same equivalence class value. However, this does not fully determine the strength of the correlation. If the same column in all the other blocks also contains many amino acids of this equivalence class value one would rather consider the correlation as weak (see fig. 3). So the strength of a correlation depends on the amino acids of the whole column.

Finding the optimal alignment of sequences is not trivial. If no structural data is available that could guide the analysis, then various algorithms can be used which rely on various substitution matrices which weight how similar the amino acids are to each other (e.g. [3]). Different matrices can be used, based on physicochemical characteristics of amino acids, putative phylogenetic relationships averaged over different proteins, characteristics of the genetic code, and so on. The resulting alignments can differ considerably according to the matrix used. The difficulty is that this approach reflects best guesses based on statistical assumptions, and may well be irrelevant when applied to various portions of the similar proteins now being aligned. An incorrect shift of the columns of some proteins can camouflage an important correlation.

Deriving patterns from aligned sequences is only meaningful if these share within their overall sequences some feature in common. This can be fulfilled by judicious selection of the sequences. In addition, the number of sequences in a block must be statistically significant to avoid spurious correlations. To a large extent subjective judgement may be necessary. The researcher might not know in advance how many amino acids contribute to various structural and signalling patterns, especially for those which remain to be discovered, and the amount of sequence data currently available might only permit tentative hypothesis. Ideally the dataset would not be limited to closely related strains or species of organisms.

3. Related Work

A lot of the existing approaches to automatically detect correlated positions are motivated by *ab initio* structure prediction, e.g. [13, 6, 11, 19]. It is assumed that the amino acids of correlated positions may often be near each other in the final folded three-dimensional structure. This would mean that correlations could give important clues with respect to the folding of the protein. However, so far the accuracy of the reported calculations is rather disappointing (e.g. [13]: 20-68% (one-time 100%), [11]: 37-68%, [19]: 14.4-38.76%.). According to the authors the enormous variations in the accuracy of the prediction of near residues are due to the different alignments and the special characteristics of the protein family. This implies that the applied correlation measure was not universally valid. To calculate the correlations statistical measures were used mainly, such as the Pearson Correlation Coefficient or the Chi-Square-Test. One of the disadvantages of these methods is that they do not take into account that sometimes amino acids may be substituted because of similar physicochemical properties. The problem is that at the current state of biological research permissible substitution cannot in general be predicted *a priori*. This would require much specific context knowledge up front. Furthermore, some of the statistical approaches can only find correlations based on all members at a column position. But the alignment might also contain

correlations which are restricted to *some* of the proteins (or organisms). For example, different environmental demands, such as extreme temperature habitats, may demand for only those organisms an especially stable folded protein. This could be attained by additional S-S bridges or a particularly tight hydrophobic core. Unfortunately, several other reasons for the disappointing quality of predictions are imaginable, such as poor alignment of the sequences, unsuitable statistical assumptions and missing consideration of other context dependencies.

Another interesting work is [14]. It was motivated by the search for the rules that lead to proteins with a specific fold. To calculate correlations, so called statistical coupling analysis (SCA) was used (see [1] for further details). The results of the study were used to build artificial proteins. In the experiment 25% of the proteins were able to adopt a fold when the calculated correlations were used to build a protein. By comparison 67% of the natural sequences could fold under the test conditions. Interestingly, none folded properly, when the only restriction was that one of the amino acids that were found in nature at this position had to be used. This demonstrates the existence of correlations but also poses the question as to why so many proteins were not able to adopt a fold.

Since correlations are nothing else than rules of the form "if ... in column x then ... in column y" instead of statistical calculations, then Association Mining might be another helpful technique [2]. To overcome the problem that this technique is usually independent of the order of items (in this case of the order of amino acids along the chain) we started with a preprocessing step in which the corresponding column index was added to every amino acid in the alignment. Afterwards the standard Association Mining algorithm was executed. Unfortunately this leads to huge numbers of rules which cannot be visually examined by an expert anymore. A major drawback is that this and other methods do not know in advance how many columns are relevant to explain some feature. A collection of 50 columns may show strong correlations, but unknown to the mathematical treatment is which of these belong to what underlying factor, or even whether they all are needed for one single protein feature.

Besides the automatic approaches several tools exist which display an alignment and allow user-defined colors to be assigned to the amino acids (see e.g. [10, 8]). The advantage of such a visualization is that context-dependency can be considered and that the experience and background knowledge of the user can be used. However, a significant number of sequences have to be used to get reliable results and an effective search for correlations by merely looking at various alignments is nearly impossible.

All the previous efforts were either pure automatic approaches or static visualizations. In our project we try to combine both methods. Instead of requiring reliable knowledge about the nature of the correlations we provide an environment in which different hypothesis can be tested. The results of automatic calculations can be intelligently visualized and are therefore convenient to analyze. The relevant parameters can easily be varied and immediate visual feedback is provided. This allows the user to search for correlations by interactively exploring various alignments and to draw on expert knowledge of patterns, both structural and informational, which he is familiar with.

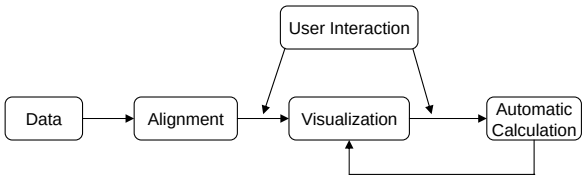


Figure 4: *Pipeline of the Exploration Process.* The aligned sequences are visualized and the correlations are calculated for the selected basis columns. After each change the visualization is updated to display the results. The user can influence the process at several stages.

4. Integrating automated and visual data exploration

The exploration process can be visualized as a pipeline (see fig. 4). As input to the process protein sequences have to be provided that all share at least one biological function. Those sequences have to be aligned. This can either be done with one of the many alignment algorithms that exist or by an expert. However, a high quality alignment is essential to get reliable results. This alignment can be visualized. Afterwards, the user can select basis columns for which the correlations are to be calculated. This triggers an update of the visualization to display the result. The user can influence the process at several stages. First of all basis columns have to be chosen for which the correlations are calculated. Secondly, the user can vary the definition of correlation. After each change the calculation of the correlations is redone and the visualization is updated. Furthermore the user can filter the results and adapt the visualization according to personal preferences and the exploration task. The following sections will explain the last two steps of the pipeline and the interaction with the user in detail.

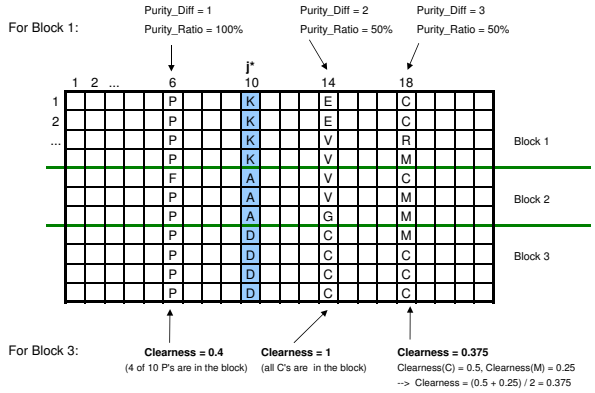


Figure 5: *Illustration of the parameters Purity_Diff, Purity_Ratio, and Clearness.* Whereas the Purity parameters represent the impurity of the correlated column, the Clearness value measures the strength of the correlation.

4.1. Automatic calculation of correlation

To be able to search for correlations a definition is needed. The definition must take into account the requirements stated in section 2. Because sometimes only part of the column is correlated we separately search in every block for correlations between the basis column and each other column. The simplest definition would be to take a column of a block as correlated to the basis column if all the amino acids of the column belong to the same equivalence class value. Thereby column 6, block 1 and columns 6 and 14, block 3 in figure 5 would be correlated to the basis column 10. But there may be reasons why we would like to tolerate some impurity in our correlated columns, such as doubt as to whether our alignment is optimal. Therefore we introduced two parameters, *Purity_Diff* and *Purity_Ratio* that allow to define how much impurity can be tolerated in a column.

Whereas *Purity_Diff* counts the number of different equivalence class values, *Purity_Ratio* measures the rate of the most frequent equivalence class value in this column of the block.

Definition 3 (Purity_Diff)
The Purity_Diff of column j' with respect to equivalence class value e^* of basis column j^* is defined as:
$Purity\_Diff_{j^*,j'}(e^*) = \|\cap_{k \in Block_{j^*}(e^*)} E(D_{kj'})\|$ [1] [2]

[1] $\|\ \|$ denotes the absolute value

[2] Note that this is a set and not a multiset which means that multiple entries of the same value are reduced to one entry

Definition 4 (Frequency of equivalence class value in column of block)
The Frequency $\text{freq}(e^*, e')$ *of equivalence class value e' in column j' with respect to equivalence class value e^* of basis column j^* is defined as:*
$freq_{j^*,j'}(e^*, e') = \|\{i \in Block_{j^*}(e^*) | E(D_{ij'}) = e'\}\|$

Definition 5 (Frequency of most frequent equivalence class value)
The Frequency max_freq *of the most frequent equivalence class value m in column j' with respect to equivalence class value e^* of basis column j^* is defined as:*
$max\_freq_{j^*,j'}(e^*) = \{freq_{j^*,j'}(e^*, m) | \exists m : \forall k \neq m,\ m,k \in E(D_{ij'}),\ i \in Block_{j^*}(e^*),\ freq_{j^*,j'}(e^*, m) \geq freq_{j^*,j'}(e^*, k)\}$

Definition 6 (Purity_Ratio)
The Purity_Ratio of column j' with respect to equivalence class value e^* of basis column j^* is defined as:
$Purity\_Ratio_{j^*,j'}(e^*) = \frac{max\_freq_{j^*,j'}(e^*)}{|Block_{j^*}(e^*)|}$

Figure 5 illustrates the two parameters. In block 1 column 14 contains two different amino acids. Let us assume that all amino acids have separate equivalence class values. The *Purity_Diff* for column 14, block 1 would be 2. Because both equivalence class values occur twice, the *Purity_Ratio* would be 50% (2/4). In column 18 of block 1 we have three different amino acids, so *Purity_Diff* would be 3. However, *Purity_Ratio* is still 50% since the most frequent amino acid is *C* and occurs twice and in total we have four amino acids in the block. Combining the two parameters allows even finer tuning. For example setting the maximum value of *Purity_Diff* to 2 and the minimum value of *Purity_Ratio* to 50% would mean that column 14, block 1 would still be taken as correlated to the basis column but not column 18, block 1.

However, the purity values do not capture the strength of a correlation. In figure 5 both column 6 and column 14 in block 3 have a *Purity_Diff* value of 1 and a *Purity_Ratio* value of 100%. But of course column 14 is stronger correlated to column 10 than column 6 since in column 6 almost all the amino acids outside of block 3 are the same than the ones in block 3. It seems rather that in column 6 a *P* would always be the best choice no matter what amino acid we have in column 10. To reflect this we introduced a new parameter, the *Clearness*. The *Clearness* of a column measures the frequency of an equivalence class value in the block in relation to the frequency of this equivalence class value in the whole column. However, this is not a sufficient definition, since we can have more than one equivalence class value in a column. Therefore we calculate the *Clearness* separately for every equivalence class value of the block and take the average value as the *Clearness* for the whole block.

Definition 7 (Frequency of equivalence class value in whole column)
The Frequency $freq(e')$ *of equivalence class value* e' *in column* j' *is defined as:*
$freq_{j'}(e') = \|\{i \in \{1 \ldots N\} | E(D_{ij'}) = e'\}\|$

Definition 8 (Clearness)
The Clearness of column j' *with respect to equivalence class value* e^* *of basis column* j^* *is defined as:*

$$Clearness_{j^*,j'} = \frac{\sum_{e' \in D_{j'}} \frac{freq_{j^*,j'}(e^*,e')}{freq_{j'}(e')}}{Purity\_diff_{j^*,j'}(e^*)}$$

Figure 5 shows some examples. In column 6 we have 4 *P*'s in block 3 and 10 in total in the column. So the *Clearness* value for column 6, block 3 is 4/10 = 0.4. In contrast to this we have 4 *C*'s in column 14 in block 3 and none in the rest of the column. So the *Clearness* value for column 14, block 3 would be 4/4 = 1. Column 18, block 3 shows an example for multiple equivalence class values in one column. Since the *Clearness* for *C* would be 0.5 and the *Clearness* of *M* would be 0.25 we get a total *Clearness* of (0.5 + 0.25) / 2 = 0.375.
Even as the purity values say nothing about the strength of a correlation, the *Clearness* says nothing about the purity of a column. If a column of a block contains amino acids of two different equivalence class values and no other block contains amino acids of the same equivalence class values the *Clearness* would be 1. This reflects the fact that those two equivalence class values seem to require a particular equivalence class value in the basis column and therefore are an interesting observation.

Lastly we have to consider that partitioning the sequences into blocks according to the equivalence class values of the basis column can lead to subsets with only few sequences. However, if the number of sequences becomes too small it is no longer meaningful to derive rules. Therefore a *Support* parameter is introduced which allows to define what a statistically significant subset of the alignment is.

Definition 9 (Support)
The Support of $Block_{j^*}$ *that is made up by equivalence class value* e^* *of basis column* j^* *is defined as:*
$Support_{j^*}(e^*) = \frac{\|Block_{j^*}(e^*)\|}{N}$

If the support of a block is below the given threshold all the correlations in that block are discarded since they are not statistically significant observations.

Input: *min_Support, min_Clearness, min_Purity_Ratio, max_Purity_Diff,* $Blocks_{j^*}$
Result: *Matrix C*

for each *block* in $Blocks_{j^*}$ **do**
  for each *column* in *block* **do**
    if *isBasisColumn(column)* **then**
      continue
    end if
    Purity_Ratio
    $\leftarrow \frac{frequencyOfMostFrequentEquivClass(column,\ block)}{|block|}$
    Purity_Diff
    $\leftarrow$ *numberOfEquivClasses(column, block)*
    for each *equivalence class value* in *column* **do**
      cl_sum
      $\leftarrow cl\_sum + \frac{frequencyOfEquivClass(column,block)}{frequencyOfEquivClass(column)}$
    end for
    $Clearness \leftarrow \frac{cl\_sum}{numberOfEquivClasses(column,\ block)}$
    // check if column is correlated
    if $Clearness \geq min\_Clearness$ &&
    $Purity\_Ratio \geq min\_Purity\_Ratio$ &&
    $Purity\_Diff \leq max\_Purity\_Diff$ **then**
      for each *cell* in *column* **do**
        $C[cell] \leftarrow TRUE$
      end for
    else
      for each *cell* in *column* **do**
        $C[cell] \leftarrow FALSE$
      end for
    end if
  end for
end for

for each *block* in $Blocks_{j*}$ **do**
  $Support \leftarrow \frac{|block|}{N}$
  // check if block has enough support
  if $Support < min\_Support$ **then**
    for each *cell* in *block* **do**
      $C[cell] \leftarrow FALSE$
    end for
  end if
end for

Algorithm 1: Calculate correlated columns

To finally decide if a column of a block is correlated to the basis column for each column the *Purity_Diff*, the *Purity_Ratio*, the *Clearness* and the *Support* is calculated. Only if all three values are below or above the given thresholds is the column of the specific block marked as correlated. Algorithm 1 summarizes the decision process.

4.2. The VisAlign Tool

The algorithm in section 4.1 provides a matrix in which each cell is marked as either correlated or not correlated to the basis column. Because there are no settings for the parameters that can be universally accepted as correct, the result has to be evaluated for biological relevance. This cannot be automated at the current state of cellular knowledge. The sensible thing to do would be to test different hypothesis by trying different parameter settings. So the presentation of the result has to be in a form that is easy to evaluate for scientists. Visualizing the results would surely be very helpful. In our approach we display the alignment as in every commonly used Alignment Viewer (e.g. [8, 10]) but additionally display the calculated correlations by greying out all the columns which are not correlated with the selected basis column(s). This permits the user to easily grasp the results and to perceive existing patterns. Furthermore patterns that are only valid in a specific context can also be identified.

We embedded this visualization in a framework that provides further support for the exploration endeavor. Figure 6 shows a screenshot of our VisAlign tool. At the moment it consists of five main components: the Alignment Viewer, the Parameter Window, the Mapping Window, the Properties Window, and the 3D-Viewer (not yet implemented in the prototype).

The **AlignmentViewer** depicts the alignment and the basis columns can be selected. Each amino acid is displayed in the color of its equivalence class value. Furthermore, as already stated above, the result of the calculation is visualized by greying out all the cells which are not correlated to the basis columns.

The algorithm that determines the correlated columns expects maximum or minimum values of the parameters which can be used as a threshold to decide whether a column is correlated or not. In the **Parameter-Window** the user can set those values and easily vary them to test different settings. Changing the parameters results in immediate recalculation of the correlations and visual feedback in the AlignmentViewer.

The **Mapping-Window** enables the user to group the amino acids into equivalence classes. Thereby similar properties of the amino acids can be taken into account or hypothesis in which the user is only interested in a specific property can be tested. Furthermore the colors can be choosen in which the amino acids of each equivalence class value should be displayed.

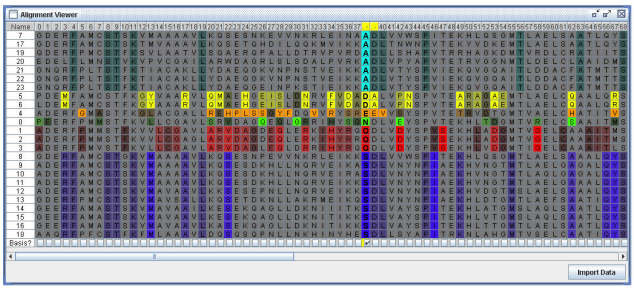


Figure 7: *Block coloring scheme* to visualize the different strength of the correlations. The brighter the color of a column is the stronger the correlation is.

The **Properties-Window** allows changes in the visualization such as fading the invariant columns in and out, displaying the names of the proteins and of the amino acids and to change the size of the cells and therefore zoom in the alignment. Depending on the task which has to be accomplished this can help to focus on the most important parts of the alignment.

In the **3D-Viewer** the three-dimensional structure of one of the proteins is shown. It is linked to the AlignmentViewer and allows to determine the positions of the columns in the three-dimensional molecule and to select interesting basis columns.

Additionally, the displayed correlations can be filtered if information about the three-dimensional structure is available. To test hypothesis that are supposed to be only valid in a specific context, the search for correlations can be restricted to a specific type of secondary structure (not yet implemented in the prototype) or to a certain distance around a basis column. Then the tool will display only those correlations that are within the defined context. The rest of the calculated correlations are faded out such that they are still distinguishable from the areas where no correlations have been found. This permits to check how significant the hypothesis is for the specific context.

Sometimes it also can be helpful to lower the threshold for the Correlation value. This is especially important if columns are only partially correlated to the basis column. However, since in general higher clearness implies something more worthy of attention, a special type of visualization was developed which also shows the strength of the correlations. Instead of assigning the colors according to the equivalence class values the amino acids belong to in this visualization the colors are assigned blockwise and the strength of the correlation is mapped to the intensity of the color. Figure 7 shows a screenshot of the AlignmentViewer when this alternative coloring scheme is used.

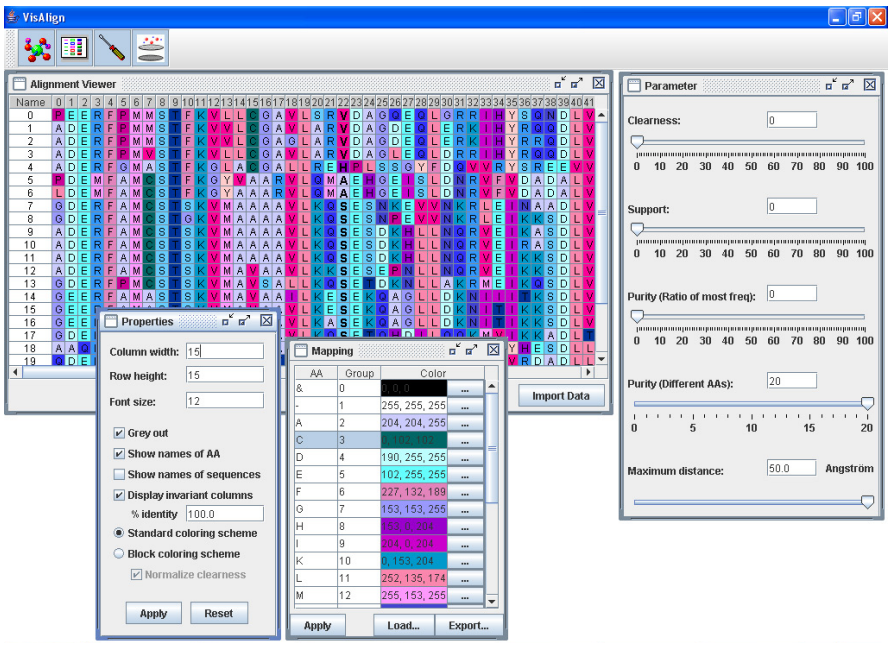


Figure 6: *The VisAlign Tool.* The main component of the tool is the AlignmentViewer. Here the alignment is visualized and the calculated correlations are displayed. The other components support the user in the visual exploration process by allowing to filter the correlations, to conveniently vary the parameters that define correlations and to adapt the visualization to personal preferences and the exploration task.

5. Application

In the following section a sample explorative session with the VisAlign tool is shown. The dataset that is used consists of 24 β-lactamase sequences (part of the data used in [5]). They were aligned using the ClustalW algorithm with default settings [12]. Of course this is not a statistically significant number of sequences if all possible correlations are to be considered. But since they are unambigiously alignable and significantly different it serves as a nice example dataset.

One of the most important properties of an amino acid is hydrophobicity. It is vital for the stability of the protein structure that mostly hydrophobic amino acids are in the core of globular proteins. If a hydrophobic amino acid in the core is mutated into a hydrophilic one, it can be highly destabilizing for the protein structure. This is why we would like to examine our sequence alignment with respect to the hydrophobicity of the amino acids.

The first step is to group the amino acids into equivalence classes according to their hydrophobicity. In the visualization we display all the hydrophobic amino acids in dark green and the hydrophilic ones in light green. Figure 8a shows the alignment. In order to see how many columns have either a hydrophobic or a hydrophilic amino acid we set the parameters Purity_Ratio and Clearness to 100% and select one of the invariant columns. In the resulting visualization all the columns not satisfying our criteria are greyed out (see figure 8b). 65 of our 153 columns are determined by this property. By setting the Purity_Ratio value to 95% we see that 13 columns more appear colored, those with only one deviating amino acid. In our further exploration we would now like to have a closer look at the columns which are variant with respect to hydrophobicity. Therefore we grey out all the invariant columns and only display the rest (see figure 8c). We select another column as basis column and the tool immediately recalculates the correlations and updates the visualization. This is repeated to test different basis columns. Figure 8d shows one of the correlations we found. Armed with potentially interesting observations, generality can be tested by including additional beta-lactamases, and biological explanation can be sought. As already stated in section 3 most approaches assume that correlations mostly occur between columns which are close to each other in the folded three-dimensional structure. As a test, the structure data of one of the used sequences was loaded, which is based on x-ray crystallography [7]. This allows one to calculate the distances between the columns. To find out how far apart our amino acids are we now gradually reduce the maximum distance. It is noteworthy that the four columns are not all near to the basis column. The distances are as follows: column 17 is approximately 21 Ångström from the basis column, for column 67 the value is 15 Ångström and column 43 is the nearest one being sep-

arated by only 10 Ångström. The 3D-Viewer that is being planned will additionally show the location of the amino acids in the three-dimensional structure.

6. Future Work

The number of possible combinations of basis columns, properties and parameter settings is enormous. Therefore filtering mechanisms will be very helpful to the user. At this time only filtering by distance to the basis columns has been implemented. We plan to also allow the correlations to be restricted to portions of the overall sequences, such as protein domains which generally interact weakly with each other.

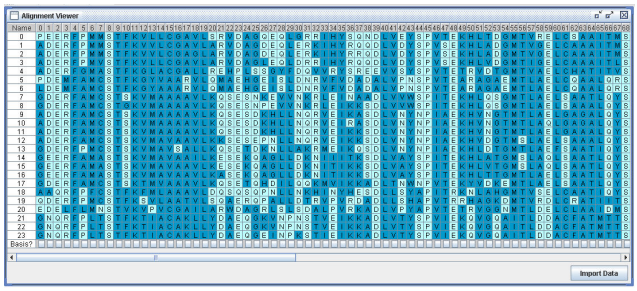
Furthermore, we would like to integrate known amino acid patterns which are used by proteins to communicate with other biomolecules or determine their location in the cell (or whether they are to be sent outside the cell membrane).

Databases such as Prosite [9] will be of much value for this purpose. These function-based patterns place constraints on the location of the amino acids, and will influence greatly how the alignments should be assembled. Identification of these experimentally established patterns will often reveal the basic principle of many correlations which our tools will identify. Our goal is to draw attention to patterns whose cause cannot yet be ascertained through other tools or algorithms. This would help the user to see what is already knowable and to concentrate on the protein parts which still need explaining.

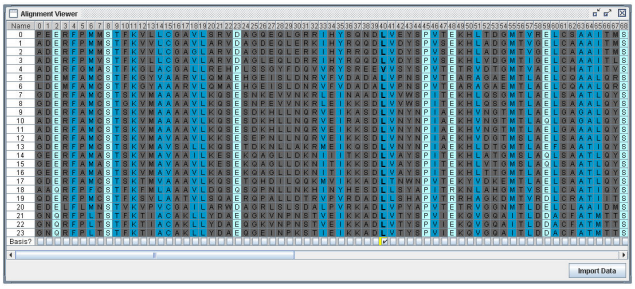
Strong correlations between columns which are not caused by features identifiable by protein pattern-based databases nor secondary structures, such as alpha-helices and beta-sheets, would point out the possibility of hidden information whose meaning needs to be extracted. Furthermore, the validity of the claimed patterns could be checked against large datasets. Preliminary tests (not reported here) show that some reported beta-lactamase sequences lack a Prosite pattern which almost all the others in the dataset have, although the function is indeed surely present.

Another challenge is to distinguish between statistical artefacts which may be due to common descent and not enough time has transpired for mutations to occur, and correlations necessary for real functional reasons. With this in mind we intend to calculate the probability that a correlation is due to chance and include this in the visualization to help the user judge how much effort should be invested in searching for underlying reasons. We also plan to calculate the probability that known functional patterns may have arisen by random mutations.

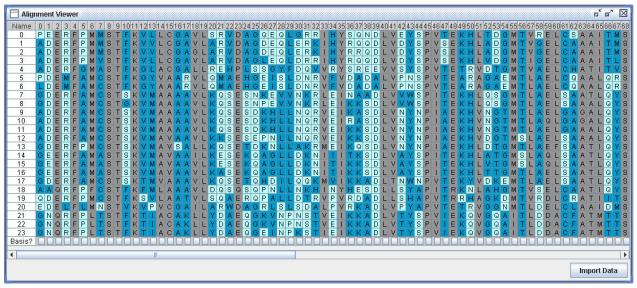
As ever more structural and informational patterns are identified, our visualization tool will permit a more thorough overview of a protein family and its component fea-



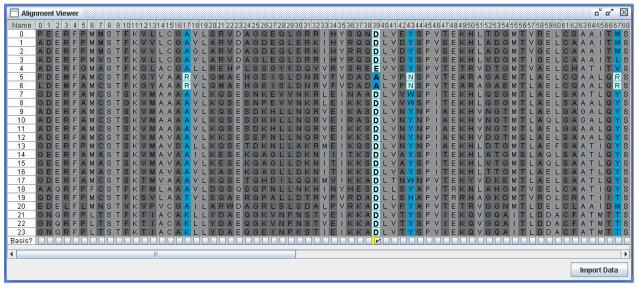
(a) Hydrophobic amino acids are displayed in dark green and hydrophilic ones in light green.



(b) Only the columns that are either purely hydrophobic or purely hydrophilic appear colored.



(c) Only the columns that are variant with respect to hydrophobicity appear colored.



(d) The visualization shows one of the correlations with respect to hydrophobicity that has been found with the tool. Assumedly, a otherwise necessary hydrophilic amino acid in column 39 is compensated by hydrophobic amino acids in colums 17, 43, and 67.

Figure 8: Different steps of the exploration process with an alignment of 24 β-lactamase sequences.

tures. Interestingly, preliminary tests already suggest that some amino acids share multiple design constraints. The exact nature of these constraints needs to be elucidated. However, we predict that these particular amino acids are likely to be intolerant to substitution via random mutations, and this may explain why less variability is found at some positions than anticipated.

7. Conclusion

We have proposed a novel correlation measure for protein sequences which offers adjustable parameters. In contrast to previous approaches found in the literature we avoided statistical simplifications but tried to capture the underlying complex biological relationships. We recognize that the physicochemical and functional basis for correlations are not universal and fixed, but rather depend on context. Typically, only a few specific amino acids are involved mutually to solve a particular protein requirement, and alternative solutions will work. The nature of the correlations therefore depends on the protein feature involved and the quality and quantity of the data available. Furthermore, we introduced our tool VisAlign which allows the visual exploration of alignments of functionally equivalent protein sequences. By integrating the automated calculation of correlations and a rapid visual representation of the results, the tool supports the user in the search for promising patterns and in testing hypothesis.

Acknowledgement

We would like to express our gratitude to Siegfried Scherer (Technical University of Munich) for many stimulating discussions and insights.

References

[1] http://www.hhmi.swmed.edu/Labs/rr/SCA.html. Supplementary Material to [14].

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 1994.

[3] S. F. Altschul, J. C. Wootton, E. M. Gertz, A. Agarwala, R.and Morgulis, A. A. Schaffer, and Y. K. Yu. Protein database searches using compositionally adjusted substitution matrices. *FEBS J.*, 272(20):5101–5109, 2005.

[4] D. D. Axe. Extreme functional sensitivity to conservative amino acid changes and enzyme exteriors. *J. Mol. Biol.*, 301(3):585–595, 2000.

[5] D. D. Axe. Estimating the prevalence of protein sequences adopting functional enzyme folds. *J. Mol. Biol.*, 341(5):1295–1315, 2004.

[6] G. Chelvanayagam, A. Eggenschwiler, L. Knecht, G. H. Gonnet, and S. A. Benner. An analysis of simultaneous variation in protein structures. *Protein Eng.*, 10(4):307 – 316, 1997.

[7] J. Chen et al. MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.*, 31(1):474–477, 2003.

[8] M. Clamp, J. Cuff, S. M. Searle, and G. J. Barton. The Jalview Java alignment editor. *Bioinformatics*, 20(3):426–427, 2004.

[9] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. Sigrist, K. Hofmann, and A. Bairoch. The PROSITE database, its status in 2002. *Nucleic Acids Res.*, 30:235–238, 2002.

[10] N. Galtier, M. Gouy, and C. Gautier. SEA VIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *CABIOS*, 12(6):543–548, 1996.

[11] U. Gobel, C. Sander, R. Schneider, and Valencia A. Correlated mutations and residue contacts in proteins. *Proteins*, 18:309–317, 1994.

[12] D. Higgins, J. Thompson, T. Gibson, J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.

[13] S. M. Larson, A. A. Di Nardo, and A. R. Davidson. Analysis of Covariation in an SH3 Domain Sequence Alignment: Applications in Tertiary Contact Prediction and the Design of Compensating Hydrophobic Core Substitutions. *J. Mol. Biol.*, 303(3):433 – 446, 2000.

[14] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437:512 – 518, 2005.

[15] S. V. Taylor, K. U. Walter, P. Kast, and D. Hilvert. Searching sequence space for protein catalysts. *Proc. Natl. Acad. Sci. U.S.A.*, 98(19):10596–10601, 2001.

[16] W. R. Taylor. The classification of amino acid conservation. *J. Theor. Biol.*, 119:205–218, 1986.

[17] R. Truman. Protein mutational context dependence: a challenge to neo-darwinian theory: part 1. *Technical J.*, 17(1):117–127, 2003.

[18] R. Truman. The ubiquitin protein: chance or design? *Technical J.*, 19(3):116–127, 2005.

[19] S. Vicatos, B. V. B. Reddy, and Y. Kaznessis. Prediction of distant residue contacts with the use of evolutionary information. *Proteins*, 58:935 – 949, 2005.