PFstats User Guide
Aspartate/ornithine carbamoyltransferase Case Study

Neli Fonseca
Contents

Overview 3

Obtaining An Alignment 3

Methods 4

Alignment Filtering 4
Reference Sequences 5
Load PDB Structure 6
Conservation 7
MRsA - Minimum Representative Sub-Alignment 7
Correlation 8
Uniprot Look-Up 9

Analyzing Results 9

View Alignment 10
Taxonomic View 10
Structure-mapped Conservation Visualization 11
Pair-wise Correlation List 11
Correlation Network 13
Communities List 13
Structure-mapped Communities Visualization 14
Correlation Tables 14
Adherence Matrix 15
Correlation Between Communities 15
Uniprot Look-up Results 16
Conserved Residues and Residues of Communities 17
Overview

PFstats is a software aimed to extract useful information from protein families through conservation and coevolution analysis. The only required input is a multiple sequence alignment (MSA) representing the protein family. All results obtained by these analysis can be both viewed and exported in many file formats.

In the first window of PFstats, as can be viewed in figure 1, there is a button called Start Wizard. By pressing this button, the user will be guided through a basic pipeline of the available methods. However, some functions can only be accessed by navigating through the Methods menu.

![Figure 1: Main Window](image)

Obtaining An Alignment

The Pfam website (www.pfam.org) provides many different options to reach a protein domain family. In the keyword search one can look for key terms such as hexokinase, globin or superoxide dismutase. In Jump to field, the user can directly enter an accession code – either from a Pfam family (as PF00068) or for a given uniprot entry (as PA2GA_HUMAN or P14555) or PDB code (1BBC). It is also possible to search by a sequence through the sequence search field. Since release 30.0, Pfam provides several types of MSAs for a given protein family (E.g. seed, full, RPs, uniprot and ncbi). In this study case, we are going to work with the family of the Aspartate/ornithine carbamoyltransferase (Pfam: PF00185) using the Uniprot alignment approach.

The PFstats can load alignments directly from the Pfam without needing to access its website. There are many parameters that affect what kind of multiple sequence alignment will be downloaded. In this example you can follow the settings shown in figure 2. The accession field indicates the Pfam family identifier, we are going to use PF00185 for the case study, but the field could also be set only by the numbers, as 00185 or only 185. The field alignment represents what type of Pfam’s alignment will be downloaded, which can be left on Full. The order field only affects the ordering of sequences and can be kept on the alphabetical.
The *type* field is related to what kind of filtering is going to be used on the next step. For filtering based in a reference sequence, the basic alignment could be used. This alignment will contain only upper case letters to represent amino acids and dashes to represent gaps. The HMM profile alignment contains upper and lowercase letters for amino acids and dashes and dots to indicate gaps, both are used to distinguish well aligned positions to insertions that couldn’t be aligned. This type of alignment is preferred for cases where the user does not want to use a reference sequence for filtering, and will be used in this case study. Finally, the *format* field indicates what kind of MSA format is going to be downloaded. It can be left on *Stockholm*. Click on fetch and wait for the MSA to be downloaded and loaded. **To be able to use the methods that connects to databases, the sequence names must be in a UniprotKb recognized format (E.g. OTC_HUMAN or P00480).**

![Alignment Filter Screen](image)

**Figure 2: Obtaining the alignment.**

**Note:** User made multiple sequences alignments or from another sources can also be used in PFstats. However, it must be in one of the readable file formats (Selex, Stockholm or Fasta).

**Methods**

**Alignment Filtering**

Even high quality alignments such as those available in Pfam usually need some filtering in order to eliminate or minimize known problems. There are four available filters on PFstats. The taxonomic filter discards sequences that do not belong to the informed taxon. It is useful for performing the analysis separately for different clades. The minimum coverage filter is used to remove fragments from the MSA. It discards sequences with length below a coverage value in relation to the reference sequence or the HMM profile. The minimum identity filter is only available when using the reference sequence filter method. This filter removes sequences that do not have a minimum fraction of identity in relation to the reference sequence. Finally, the maximum identity, compares all sequences amongst each other, removing the smallest one whenever a pair's
identity is higher than an informed value. Maximum identity filters are used to reduce the phylogenetic bias caused by accumulation of very similar sequences in the alignment. The values used in the case study can be seen in figure 3. The MSA used in the date of this study has 9815 sequences, the coverage filter removed 269 sequences and the maximum identity removed 4815, remaining 4731 sequences.

Note: Sometimes discarding sequences can reduce the amount of information in the alignment. Thus, another option to maximize the representativeness is to apply entropy weightening to sequences when calculating the correlations.

Reference Sequences

The reference sequence window is used to choose one or a group of sequences that are going be compared in the results. The reference sequences are a way to show the residues found by conservation and correlation analysis with their actually sequence numbering and amino acid. An example can be seen in the figure 20. This helps understanding the biological significance of the results by looking for positions in academic papers and databases. There is also the possibility to select a set of sequences by a database search. The user can select all proteins that have a tridimensional structure deposited in the PDB, or the proteins that have a minimum preestablished annotation score in the UniprotKb, or proteins from a specific clade. For the case study, select the sequences that has pdb structures.

Note 1: Don’t forget to press save before advancing to the next step.
Note 2: If a sequence stored as reference sequence was removed during the filtering step, it are going to be reinserted in the working sub-alignment. So, if you are used a filtered alignment, select less as possible reference sequences.
Load PDB Structure

Click in **Methods** menu and then in **Load PDB Structure**.

At this window, the user can load PDB files and align them with their respective sequences from the MSA using the Needleman-Wunsch algorithm. The Pfam’s Stockholm alignments usually contains information about available structures, PDB identifier and chain. If it is the case, those information are going to be displayed in the **recommended PDBs** list when the corresponding sequence are selected in the **Sequence** field. If it is not the case, the user can still get automatically structures information by clicking on the **Search PDBs** button. For the tutorial, select the PYRB,ECOLI sequence and the structure 1RAI, chain A, and click on **Fetch From Internet**. After load, a window will ask if the downloaded structure should be aligned with the informed sequence and click Yes. If everything went correctly, the **Current Structures** list will now contain the 1RAI item (Figure 5).
**Note:** A custom pdb file can also be provide. Just select the corresponding sequence in the *Sequence* field and use the *Choose from File* button to upload the file.

**Conservation**

Click the *Methods* menu and then in *Conservation*.

At this window, the user can calculate a series of frequency entropy and stereochemistry statistics. The *Alpha* and *Beta* fields defines the amount of entropy weight and stereochemistry weight are going to be considered in the final calculus. Defining both weights with 1.0 implies in considering both weights as equal strength. The minimum conservation field defines the minimum conservation score value in order to a residue be considered conserved. For the case study, just leave all the fields as default and double click the 1RAI loaded structure to use it in a later visualization (make sure that *generate visualization by struture* is checked). Click *calculate*. After the calculation, choose the path where the generated pdb file will be saved.

![Conservation window](image)

Figure 6: Conservation window.

**MRsA - Minimum Representative Sub-Alignment**

Click the *Methods* menu and then *Minimum Representative Sub-Alignment*.

This program implements a variation of the procedure described in (Dima and Thirumalai 2006). It calculates the average entropy for the full alignment and then for multiple sub-alignments with decreasing sizes (from 100% to 1% of its size). This method is used to find the minimum sub-alignment that maintains the full alignment representativeness. A network cutoff value is chosen by a manual analysis on the generated graph. To generate the MRsA graph, click on calculate and wait until the program runs. The generated graph can be seen in figure 7, and for this case, a reasonable value to use would be between 0.10 and 0.20.
When a small number of sequences are removed from the alignment (right to left side of the graph), there is not much variation in the average entropy. As the sub-alignment size reaches 5-15% of the original size, there is a very steep rise in the average conservation measurement, meaning that alignments with a lower number of sequences will not be statistically significant representations of the original alignment. It is advisable to avoid using MRsA values in this range and use a safer value for sub-alignment size (e.g., 25% in this case), which will be used in the correlation method.

**Correlation**

Click the *Methods* menu and then *Correlation*.

Correlation will run a series of methods in order to calculate the pairwise correlations, generate and decompose the network into communities. The amino acids coevolution networks are formed by a set of nodes and links, in which the nodes represent amino acids and their respective positions in the MSA, and the links connect a pair of nodes reflecting a significantly correlation between both of them. The correlation parameters determine the cutoffs for network building. The first parameter, *Minimum Score*, is related to the binomial probability associated to the frequency shift for that correlation. For this case study, the *Minimum Score* can be kept as 5, meaning that the network will include edges connecting nodes on which the presence of one residue alters the frequency of the other and the associated probability for that frequency shift occurs at random is lower than $10^{-5}$. The *MRsA fraction* field represents the fraction of the alignment to use in correlation calculation. This parameter is picked in the previous step by analysing the MRsA graph. We are going to use 0.20. Finally, the *Delta Frequency* represents the minimum frequency variation to be considered a valid correlation between each pair of nodes. Set this value as 0.25, that is, a correlation will only be considered a link of the network if both residues of each correlated pair increase the frequency of each other for more than 75%, or, in the case of anti-correlation, when it decreases to less than 25%. More-
Figure 8: Correlation window.

over, make sure that *generate structural visualization* box is checked and double click on 1RAI to generate the correlation structural visualization. There’s also a *generate communities subalignments* box that can be checked in order to create sub-alignments for each network community containing all sequences that matches the community residues. Figure 8 shows how this window is set.

**Uniprot Look-Up**

Click on menu *Methods* and then *Uniprot Look-Up*.

This method consists in search for position specific annotations at UniprotKb among those from the co-evolved and conserved sets found in the previous analysis. Click the + button to select all the sequences then click *search*.

**Note:** Make sure that the *conserved residues* and *communities* fields are checked. They can only be checked if you already performed the conservation and correlation analysis.

**Analyzing Results**

At this step, we are going to navigate through the results menu and see each of the results visualizations windows available in PFstats. At this point you can click file and save results if you want to continue the analysis afterwards.

It is important to know that you can navigate through the two lists on the left (Alignment and Sub-Alignment). It is possible to work with multiple alignments and the currently selected MSA can be changed any time by clicking on the the respective item in the alignment list. Each alignment has its respective sub-alignments that are located in their below Sub-alignments list. The user can change the current sub-
alignment any time by navigating through this list, but each sub-alignment has its own statistical data. Therefore, to analyze conservation or correlation in a different sub-alignment, these methods need to be performed using the desired sub-alignment as input. Each sub-alignment can also store multiple correlation networks data, and they can be navigated through the bottom level items in the sub-alignments list.

Besides the results data visualization in the GUI itself, it is also possible to export all the results data in a several file formats (txt, csv, html and xml). To export results, just navigate through the export menu.

**View Alignment**

The *View Alignment* window is a useful option to realize a general inspection in the MSA. There is possible to visualize conservation patterns by residue coloring and logos, in addition to sort, filter and select sequences and columns. An interesting analysis is to navigate through the full full alignment and the communities sub-alignments to see how the conserved positions changes between these sub-alignments. Let’s start visualizing the MSA conserved positions. Click *results* and then *View Alignment*. Now navigate through the sub-align list and see how the conserved positions changes. The communities sub-alignment have positions a bit more conserved and specific in relation to the filtered sub-alignment. Figure 9 shows the absolutely conserved positions for the filtered, the community 1 and community 2 sub-alignments.

![View of the filtered alignment.](image)

![View of the community 1 sub-alignment.](image)

![View of the community 2 sub-alignment.](image)

![View of the community 3 sub-alignment.](image)

**Figure 9: Alignment view**

**Note:** If you need to copy a specific sequence in text format, just use the *List Sequences* visualization.

**Taxonomic View**

The *taxonomic view* window contains a sunburst plot visualization in order to illustrate the taxonomic diversity of the alignments. Sometimes a community of residues can be related to a specific taxonomic clade, in this example, as can be seen in figure 12, the filtered alignment is well distributed between bacteria, eukaryota and archaea. However, the subalignment of sequences that has all residues of community 1 is
restricted to bacteria and archaeas, and the subalignment of community 2 is fully restricted to bacteria organisms. These kind of particularities can be easily seen by clicking in the expand button and navigating through the communities sub-alignments. Figure 10 shows the expanded window for visualizing the taxonomic distributions.

(a) View of the filtered alignment.  
(b) View of the community 1 sub-alignment.  
(c) View of the community 2 sub-alignment.

Figure 10: Expanded Taxonomic View

Structure-mapped Conservation Visualization
At this window the user can visualize the chosen protein structure with their residues colored according to the conservation scores. Residues with “colder” colors (blue) are those with lower conservation value, in contrast, the residues with “hotter” colors are more highly conserved. This window allows the rotation of the structure, as well as changing representation types (sticks, lines, cross, sphere, cartoon and trace) and zooming in and out. The user also has the possibility to load the structure in their preferred software, such as PyMol or VMD. For doing this, use the software functionality for coloring a structure by b-factors values.

Pair-wise Correlation List
A positive correlation between an amino acid \( X \) at position \( i \) and an amino acid \( Y \) at position \( j \) can be rephrased as: the presence of \( X_i \) implies an increase in frequency of the amino acid \( Y_j \). At the same time, a pair \( X_i-Y_j \) is considered to be anti-correlated (i.e., presenting a negative correlation score) when the presence of \( X_i \) implies a decrease in the frequency of \( Y_j \). The correlation list window shows the list of all valid correlations and anti-correlations of the network. In other words, all the links of the network with their respective linked nodes and correlation scores (weight). By default, the positions numbering are set according with the MSA columns, but it is possible to change it according to any sequence of the alignment or loaded PDB. Thus, exporting this data might be useful, as edge list is the mainly file format used in network science software such Cytoscape and Gephi.
Figure 11: Structure-mapped Conservation Visualization

Figure 12: Pair-wise Correlation List
Correlation Network

The correlation network window contains a dynamic network visualization. It is also possible to expand the network to gain access to more functions: hide anti-correlation links and the isolated nodes, scaling nodes by hub size, scaling links by correlation scores and to color node by communities. Figure 13 shows the expanded network visualization window.

![Correlation Network](image)

Figure 13: Expanded correlation network

Communities List

This window is analog to the pair-wise correlation list window, but listing residues from each of the communities. As in the previous case, it is also possible to change the numbering pattern and export the data.

![Communities List](image)

Figure 14: Communities List
Structure-mapped Communities Visualization

This window is identical to the structural conserved residues, but residues are now colored according to their communities. Residues from the same community will have the same colors.

Correlation Tables

The correlation tables show the frequency variation of a residue when another residue from the same community is present. The tables can be seen by both frequency percentage and by correlation score. In the case of frequency percentage tables, the first column, “All”, consist on residue frequency in the input MSA. The others columns indicate the frequency of each residue, when the corresponding row residues are present. For example, the correlation table on Figure 16 shows that lysine in position 291 has a frequency of 26.71% in the filtered alignment, but for sequences presenting another glycine in position 384, that frequency increases to 98.59%. Similarly, the log p table, represents the correlation score for each pair of residues.
Adherence Matrix

The adherence matrix is used to quantify how much each sequence is related to each community. The higher the adherence value for a sequence, the more it represents a given community. Adherence matrices can be useful to identify the biological meaning of the communities by bibliographic reviews, guiding which sequences are more representative of each community. For example, if most proteins with a given characteristic adhere to one community while those without it present low or zero adherence to that community, it is likely that such characteristic is related to the residues in that community. If that is the case, adherence can be used for gene annotation - novel sequences with high adherence to a community that is related to a given characteristic may be annotated as putatively presenting such characteristic. In the figure 17, we can seen the adherence tables sorted by the communities 1 and 2.

![Adherence Matrix](image1)

(a) Sorted by community 1.  (b) Sorted by community 2.

Figure 17: Adherence Matrix

Correlation Between Communities

In addition to analyzing individual communities, the software allows studying their relationship with each other. This window shows a network visualization containing communities as nodes. The network also takes into account isolated anti-correlated residues.

![Correlation network between communities](image2)

Figure 18: Correlation network between communities
Uniprot Look-up Results

There are two ways to visualize the annotation search results. By proteins (Figure 19a), and by communities (Figure 19b). In this case study, by looking at the first community, it is possible to observe its relationship to the catalytic activity. Six residues are present in this community, two of them being cysteines interacting by disulfide bond, another being the catalytic Histidine and the three other residues being linked to calcium binding.

Figure 19: Uniprot look results
Conserved Residues and Residues of Communities

These two windows are very similar. Both show conserved or correlated residues from the alignment on the column header and the corresponding residues for each of the reference sequences. If you selected a good representative set of reference sequences, interesting patterns can be seen in the residues of communities. Figure 20 shows the conserved residues window and figure 21 shows the residues of communities 1 to 4.

Figure 20: Conserved residues

Figure 21: Residues of the communities