DAVI: a tool for clustering and visualising protein domain architectures

Paul Saighi[1], **Chadi Jaouadi**[1], **Fabio R.J Vieira**[2,✉], **and Juliana Silva Bernardes**[3,✉]

[1]Sorbonne Université, UFR 919, 75005 Paris, France
[21]Institut de Biologie de lÉcole Normale Supérieure (IBENS), 75005 Paris, France
[3]Sorbonne Université, laboratoire de Biologie Computationnelle et Quantitative, CNRS UMR7238, 75005 Paris, France

The characterization of protein functions is one of the main challenges in bioinformatics. Proteins are often composed of individual units termed domains, motifs that can evolve independently. The domain architecture of a given protein is the particular order and the content of its numerous domains. Some computational approaches predict the most likely domain architecture for a set of proteins. However, a few numbers of visualization tools exist, and most of them are unavailable. Here we present DAVI, an efficient and user-friendly web server for protein domain architecture clustering and visualization. DAVI accepts the output of most used domain architecture prediction tools and also produces domain architectures for a set of protein sequences. It provides a rich visualization for comparing, analyzing, and visualizing domain architectures.
Availability: `http://genome.lcqb.upmc.fr/Domain-Architecture-Viewer`

protein annotation | domain architectures | clustering | visualisation
Correspondence: ***juliana.silva_bernardes@sorbonne-universite.fr***

Introduction

The identification of functional domains is essential for protein sequence analysis. Domains are the building blocks of all proteins; they are sequence fragments that can be independently stable and folded, have a specific function and occur alone or in groups. The majority of proteins, especially in complex organisms, are composed of one or more domains. The arrangement of domain units in a protein forms its domain architecture and determine, to a large extent, the protein function. Therefore, accurate detection of domain architectures is extremely useful for protein function prediction and structural analyses. Many approaches have been proposed to predict the most likely domain architecture for a set of proteins (1–3). However, a few numbers of visualisation tools in the form of web server exists. Often, such graphical tools are particular to a group of organisms and cannot be used in general. Moreover, there is no possibility to group proteins according to their domain architecture similarities.

Here we propose DAVI, a web server for protein Domain Architecture clustering and VIsualisation. DAVI accepts the output of most used domain architecture prediction tools such as BMC (best match cascading), DAMA (1), and dPUC (2). However, it is also possible to run those tools by inputting FASTA sequences or a set of identified domains. Users can also choose between two clustering algorithms (hierarchical or spectral) to group proteins with similar domain architectures. Once input files are processed and proteins clustered, DAVI provides a set of visualisation features, including domain architecture groups, compact view, statistics and the possibility of exporting data in several formats.

Methods

Figure 1 shows the flowchart of DAVI; it has two main steps: producing domain architectures and clustering proteins with similar domain architectures.

Producing domain architectures. We have constructed a pipeline to produce domain architectures for a set of proteins automatically. For that, users should input sequences in FASTA format or a set of potential domains obtained after running *hmmscan* program (4) on the Pfam database (5). From sequences in FASTA format, DAVI runs *hmmscan*. The output of *hmmscan* (generated by DAVI or provided by the user) is sent to some domain architecture tool, previously informed: DAMA, dPUC or BMC.

Clustering proteins with similar domain architectures. To group proteins by their domain architectures, we used two clustering algorithms: hierarchical (6) and spectral (7). Both algorithms require a distance matrix that contains pairwise distances for each protein pair. To compute such a matrix, we first encode each protein domain architecture by a sequence of letters representing domains. Next, the pairwise distance between two proteins, saying P_1 and P_2 was obtained with Levenshtein metric (8) that computes the minimum number of single-character edition (insertions, deletions or substitutions) required to transform P_1 into P_2. Finally, we input the distance matrix in some clustering algorithms to obtain groups.

Web server

Inputs. Users can input three types of files: sequences in FASTA format, *hmmscan* output, or protein domain architectures provided by any software. If sequences or *hmmscan* output are inputted, users need to choose a domain architecture tool: DAMA, dPUC or BMC. Next, they should choose a clustering algorithm: hierarchical or spectral. After submitting their input file, users need to guide clustering algorithms to determine the number of clusters. For hierarchical clustering, a dendrogram is displayed to help users set a specific height and cut the tree into clusters. For spectral clustering,

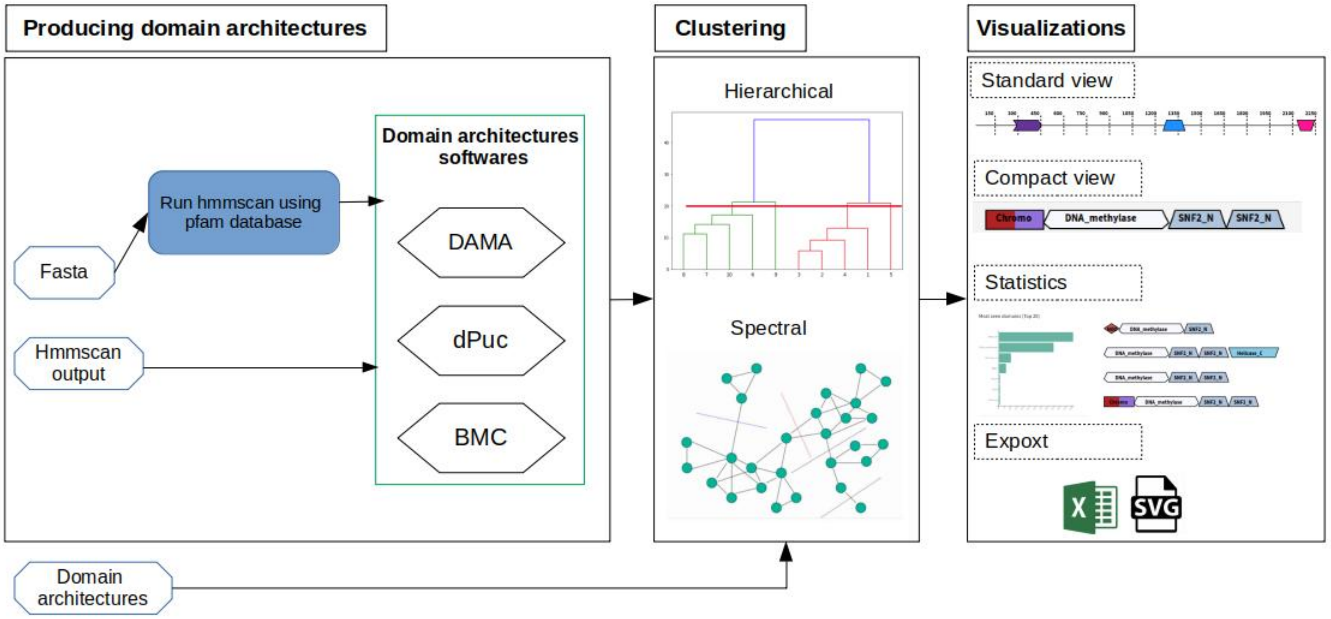

Fig. 1. DAVI flowchart. **Producing domain architectures** - users input sequences in FASTA format or *hmmscan* output. From sequences, we run *hmmscan* program using Pfam database. From *hmmscan* output, we launch a tool to obtain domain architectures: DAMA, dPUC or BMC. **Clustering** - the output of some domain architecture tool is processed to build a distance matrix used to obtain clusters with hierarchical or spectral algorithms. **Visualisations** - users can visualise groups of proteins with similar domain architectures with four different modules: standard view, compact view, statistics and export.

users should input the number of clusters directly. Before visualisation, users can choose between a contracted visualisation (proteins are scaled) or an expanded visualisation (protein length are shown).

Visualization. DAVI has four visualization modules implemented for different objectives: the standard view, compact view, statistics and export files, see the right panel in Figure 1.

The *standard view* displays groups of proteins with similar domain architectures. Initially, each group reveals only a representative protein, but users can expand and navigate to explore all protein architectures of a given group.

The *compact view* was mainly designed to obtain a condensed view of protein domain architectures. In this view, only domain order is displayed, protein lengths are masked. By hiding domain positions, users can easily compare a large set of proteins of a given group.

The *statistics* is a module that displays quantitative information about user's data such as the number of groups, sequences, mean of domain per protein and number of distinct domains. It also provides a histogram with the distribution of architecture sizes and the most seen domains/architectures.

Finally, *export* allows users to save their data in different formats.

Conclusion

DAVI is a unique web server for clustering and visualizing protein domain architectures. It is simple, fast, and deal with protein sequences or their domain hits. It provides a direct way to quickly group, analyze and study domain information of a set of proteins. DAVI can be useful for a multitude of purposes: compare domain architectures and find homologous relationships, identify/distinguish promiscuous (very frequent) domains architectures, and study evolutionary events involving domains: insertion, deletion, and conservation.

References

1. Juliana S Bernardes, Fabio Rocha Jimenez Vieira, Gerson Zaverucha, and Alessandra Carbone. A multi-objective optimization approach accurately resolves protein domain architectures. *Bioinformatics*, 32(3):345–353, 2015.
2. Alejandro Ochoa, Manuel Llinás, and Mona Singh. Using context to improve protein domain identification. *BMC bioinformatics*, 12(1):90, 2011.
3. Corin Yeats, Oliver C Redfern, and Christine Orengo. A fast and automated solution for accurately resolving protein domain architectures. *Bioinformatics*, 26(6):745–751, 2010.
4. Sean R. Eddy. Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998.
5. Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, et al. The pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427–D432, 2018.
6. Robin Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, 16(1):30–34, 1973.
7. Bojan Mohar, Y Alavi, G Chartrand, and OR Oellermann. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications*, 2(871-898):12, 1991.
8. Vladimir Iosifovich Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.