ENDscript 2.0

User Guide

Preamble

This user guide documents the ENDscript Web server developed by **Patrice GOUET and Xavier ROBERT** in the "**Retroviruses and Structural Biochemistry**" research team of the "**Molecular Microbiology and Structural Biochemistry**" laboratory (UMR5086 **CNRS** / **Lyon University**). ENDscript is an **SBGrid** supported application.

This documentation contains all the information you need to use the ENDscript Web server as a beginner or advanced user.

The two following notation conventions are used to draw your attention to certain important pieces of information:

> If the option Display all known structures is activated *via* the interface (default), an automatic search is performed to check if a sequence name can be related to a known 3D structure.

○ The program identifies α-helices (shown by medium squiggles), 3_{10} helices (small squiggles), π-helices (large squiggles), β-strands (arrows), strict α-turns (`TTT` letters) and β-turns (`TT` letters) from the 3D structure.

Table of contents

1 Introduction

- ENDscript is a friendly Web server, which extracts and renders a comprehensive analysis of primary to quaternary protein structure information in an automated way.

- ENDscript is a tool of choice for biologists and structural biologists, which allows generating with a few mouse clicks a set of detailed high quality figures and 3D interactive representations of their proteins of interest.

- ENDscript Web server is fast and convenient:

  - No particular knowledge in bioinformatics is needed to obtain comprehensive and relevant illustrations.
  - The user is guided through the process by tooltips and detailed help topics (the present documentation) accessible at any time.
  - Thanks to its automated pipeline and a parallel programming, ENDscript can deliver results in one click and within one minute.
  - Demanding or expert users can modify settings to fine-tune ENDscript at their needs.
  - ENDscript produces publication-quality illustrations in most common file formats (PostScript, PDF, PNG, and TIFF) and sizes (US letter, A4, A3, A0 and the gigantic 'Tapestry' format).
  - ENDscript is accessible with any modern Web browser equipped with a PDF reader. To take advantage of the 3D interactive representations, the PyMOL software (**free open-source** or **commercial version**) is required.

2 Overview of the ENDscript automated pipeline

The ENDscript automated pipeline involves numerous sequence and structure analysis programs:

- **SPDB**, a homemade program to check residue numbering and chainIDs from the query PDB file.
- **DSSP** (1,2), to extract secondary structure elements, disulfide bridges and solvent accessibility per residue.
- **CNS** (3), to calculate non-crystallographic and crystallographic protein:ligand and protein:protein contacts.
- **BLAST+** (4), to search protein homologues using the sequence of the PDB query against a chosen sequence database.
- **Clustal Omega** (5), **MAFFT** (6), **MSAProbs** (7) or **MultAlin** (8), to perform multiple sequence alignments.
- **ESPript** (9-11), to render all this information with flat figures.
- **ProFit** (12), to superimpose all homologous proteins of known 3D structures on the PDB query.
- **PyMOL** (13), to generate scripts and session files to display sequence and structure conservation with 3D interactive representations.
- **PhylodendronWeb**, to build a phylogenetic tree.
- **Jalview Desktop** (14), for multiple sequence alignment editing, visualisation and analysis.

All these programs are launched sequentially in three succeeding phases:

• Phase 1

To run the first phase, ENDscript uses as query either a four digit **PDB** (15) identifier or an user-uploaded coordinate file in **PDB format**.

On the first box of the ENDscript interface (Query PDB file), fill up the form by at least:

- clicking on the PDB icon and typing the PDB entry code (*e.g.* `2CAH`) of your protein structure (NMR and crystallographic structures are supported),

- or uploading you own PDB file by clicking on the Browse button (or equivalent depending on your browser language).

Click on SUBMIT in the buttons frame.

The PDB query is processed with SPDB and the amino acid sequence is extracted.

A SPDB output file is generated and given to DSSP to extract secondary structure elements, disulfide bridges and solvent accessibility per residue. The same SPDB output file is then used by CNS to determine non-crystallographic and crystallographic protein:ligand and protein:protein contacts.

At this point, an ESPript figure is generated, giving the following information on each monomeric sequence contained in your PDB query:

- Secondary structure elements and residues in alternate confrimation are shown above sequence query.
- Accessibility and hydropathy scales, protein-protein and protein-ligand contacts and possible disulfide bridges are shown below.

• Phase 2

A BLAST search using the sequence of the PDB query is performed against a chosen sequence database (PDBAA by default) to detect protein homologues.

The result is piped to a multiple sequence alignment software (Clustal Omega, MAFFT, MSAProbs or MultAlin).

A second figure is then generated by ESPript:

- It shows the aligned sequences colored according to their degree of similarity.
- In addition, each homologous sequence of known 3D structure is adorned with its secondary structure elements extracted by DSSP.
- Further information is presented below the alignment as in phase 1.

• Phase 3

Two PyMOL session files are generated. They can be downloaded and interactively examined with the molecular 3D visualization program PyMOL installed on the user's computer.

- **The first PyMOL representation is named 'Cartoon':**

  - This is a ribbon depiction of the PDB query colored as a function of similarity scores calculated from the previous multiple sequence alignment.
  - This color ramping from white (low score) to red (identity) allows to quickly locate regions of weak and strong sequence conservation on the structure of the query.

- **The second PyMOL representation is named 'Sausage':**

  - It shows a variable tube representation of the Cα trace of the PDB query.
  - In this goal, all homologous protein structures are superposed onto the PDB query with ProFit and the size of the tube is proportional to the r.m.s. deviation per residue between Cα pairs.
  - The same white to red color ramping is used to visualize sequence conservation.
  - By combining these two information, the user can identify areas of weak and strong structural conservation and correlate this with sequence conservation.

If applicable, the two PyMOL representations can also display, *via* the PyMOL control panel, an assortment of supplementary data:

- Biological assembly,
- Multiple NMR models,
- Disulfide bridges,
- Nucleic acids / ligands / monatomic elements and their contacting residues,
- Strictly conserved residues,
- PDB SITES markers,
- Solvent-accessible surface mapped with the sequence conservation coloring code.

All these features are fully user-editable thanks to the PyMOL control panel and publication-quality pictures can rapidly be ray-traced (please refer to PyMOL documentation or **PyMOLWiki**).

All the resulting files from phases 1 to 3 can be visualized by a mouse click or retrieved on your computer with the right button / "Save as" option of your browser.

3 Phase 1 in details

- **Result:** A first ENDscript flat figure is produced with information on each monomeric sequence contained in the PDB query:

  - Secondary structure elements and residues in alternate confirmation are shown above the sequence of the PDB query,
  - Accessibility and hydropathy scales, protein-protein and protein-ligand contacts and possible disulfide bridges are shown below.

• SPDB

- **Main role:** checks and cleans the PDB query before entering the ENDscript automated pipeline.

SPDB (and by extension ENDscript) supports structure files from the Protein Data Bank or resulting directly from any program conforming to the PDB format.

- If necessary, SPDB re-assigns chainIDs from A to Z, 0 to 9, and a to z.
- First model is kept for multiple NMR models.

   - First conformers are kept for alternate residues.
   - Second oxygen atom of C-terminus main chain is removed (atom `OXT`).

   > In case of a PDB query with multiple chains, the user can specifically select the one he wants to process with ENDscript (Chain ID option). Warning: this option is case sensitive.

   > If the PDB query contains selenomethionine residues (three-letter residue name `MSE`), the user can substitute these last to methionine residues (`MET`) by selecting the Substitute MSE residues to MET option.

- ENDscript has the ability to determine and depict contacts between protein residues and hetero-compounds, if present:

   > Several common hetero-compounds are automatically kept (see table below) and are subsequently depicted by given symbols on the flat figures. The user can manually keep non-supported hetero-compounds contained in its PDB query. Hence, he must type their names in the Keeping unsupported hetero-compounds tabular form (up to 10 names of 2-3 characters per column and one name per line).

Hetero-compound type	Name	Symbol given
Nucleotides	`ADE GUA CYT THY URI A G C T U DA DG DC DT`	*
Porphyrin groups	`HEM BCL BPH MQ7`	:
Sugars	`GLC GAL MAN NAG FUC SIA XYL`	"
Miscellaneous	`NAD NAH NDP NAP FMN`	^
Modified amino acids	Regardless of their names, as long as they contain main chain atoms N, Cα, C, O	@

   > Contacts between protein residues and automatically or manually kept hetero-compounds are shown in the phase 1 flat figure. In this goal, the symbols `* : " ^ @ < >` / are used according the assignment of the user in the Keeping unsupported hetero-compounds tabular form.

   > By default, this mark is shown in red if the distance of the protein: hetero-compounds contacts is less than 3.2 Å and in black if it is in the range 3.2-5.0 Å.

• DSSP

- **Main role:** calculates secondary structure elements.

   - The program identifies α-helices (shown by medium squiggles), 3_{10} helices (small squiggles), π-helices (large squiggles), β-strands (arrows), strict α-turns (`TTT` letters) and β-turns (`TT` letters) from the 3D structure.
   - Accessibility by residue is calculated.
   - Only coordinates of protein residues are taken into account.
   - Cystein residues involved in disulphide bridges are identified.

• CNS

- **Main role:** calculates protein-protein and protein-ligand contacts.

   - CNS calculates both crystallographic and non-crystallographic contacts between each protein molecule.
   - Contacts between protein residues and hetero-compounds are also calculated, if these latter have been automatically or manually kept.
   - If available, cell parameters and space group are extracted for crystallographic structures.
   - Hydrogen atoms are deleted and thus excluded from distance calculation.
   - Main chain atoms (N, Cα, C, O) can also be excluded from distance calculation, by enabling the Use side chains only option in the first box of the form (Query PDB file).
   - Upper limit for calculation of molecular contacts is 3.7 Å by default and can be changed with the Contacts up to option. The shortest intermolecular distance is taken for each residue.

• ESPript

- **Main role:** generates the first ENDscript flat figure.

- The protein sequence of each chainID contained in PDB query is displayed.

- Secondary structure elements have been calculated by DSSP in the previous step and:

   > α-, 3_{10}- and π-helices are shown above sequence as medium, small and large squiggles with α, β and π labels, respectively,
   >
   > β-strands are shown as arrows labeled β,
   >
   > Strict α- and β-turns are marked by `TTT` and `TT` letters, respectively.

- Residues in an alternate conformation are highlighted by a grey star above sequences.

- Relative accessibility, calculated by DSSP in the previous step, is shown by a blue-colored bar below sequence. White is buried (A < 0.1), cyan is intermediate (0.1 ≤ A ≤ 0.4), blue is accessible (0.4 < A ≤ 1) and blue with red borders is highly exposed (A > 1). A red box means that relative accessibility is not calculated for the residue, because it is truncated. Remark: only molecules located in the crystallographic asymmetric unit are taken into account by DSSP in its calculation of accessibility. Thus, you can find 'highly accessible' residues involved in contacts with crystallographic neighbors according to the ESPript figure. These residues are in fact buried in the crystal lattice.

- Hydropathy is calculated from the sequence according to the algorithm of Kyte & Doolittle [16] with a windows of 3. It is shown by a second bar below accessibility: pink is hydrophobic (H>1.5), grey is intermediate (-1.5 ≤ H ≤ 1.5) and cyan is hydrophilic (H < 1.5).
- Disulphide bridges, identified by DSSP in the previous step, are shown by green pairs of digits (1 1, 2 2 ...) below the bar of hydropathy.
- Protein-protein and protein-ligand contacts, calculated by CNS in the previous step, are displayed along with disulphide bridges below the bar of hydropathy. The shortest intermolecular distance is taken for each residue. Corresponding contact symbols (see **above paragraph**) are written in red if the distance is less than 3.2 Å and in black if the distance is in the range 3.2-5.0 Å.
- Main information is given according to the written marks, which shows protein-protein and protein-ligand contacts:

> A to Z, 0 to 9 or a to z means that the concerned amino acid residue has a non-crystallographic contact with an amino acid residue of the Chain A to Z, 0 to 9 or a to z (*e.g.* this amino acid residue is involved in a non-crystallographic interface).

> *A to Z, 0 to 9, a to z* **in italic** means that the concerned amino acid residue has a crystallographic contact with an amino acid residues of the Chain A to Z, 0 to 9 or a to z (*e.g.* this amino acid residue is involved in a crystallographic interface).

> # identifies a contact between two amino acid residues having the same names and numbers (*e.g.* along a 2-fold symmetry axis).

> `* : " ^ @ < > /` means that the concerned amino acid residue has a contact with a ligand (*i.e.* an automatically kept or a chosen hetero-compound - see **above paragraph**).

> *`* : " ^ @ < > /`* **in italic** means that the concerned amino acid residue has a crystallographic contact with a ligand (*i.e.* an automatically kept or a chosen hetero-compound - see **above paragraph**).

- Further information is given with colors:

> A yellow background identifies a non-crystallographic contact.
>
> An orange background identifies an amino acid involved in both a crystallographic and a non-crystallographic contact.
>
> A blue frame identifies an amino acid involved in both a protein-protein and a protein-ligand contact.
>
> A red letter identifies a contact < 3.2 Å.
>
> A black letter identifies a contact between 3.2 Å and 5.0 Å.

4 Phase 2 in details

- **Result:** A second ENDscript flat figure is produced. It displays:
  - A multiple sequence alignment of homologous proteins colored according to residue conservation,
  - The secondary structure elements of each homologous sequence of known structure.

To generate this second flat figure, the following program pipeline is called by ENDscript:

• BLAST search

- **Main role:** finds sequences homologous to that of the PDB query.

> If the option Enable the BLAST search is activated (default), a BLAST+ search is performed against a chosen sequence database (defined by the Choose a database option):

APIME	Complete proteome from *Apis mellifera*
ARATH	Complete proteome from *Arabidopsis thaliana*
BOVIN	Complete proteome from *Bos taurus*
CAEEL	Complete proteome from *Caenorhabditis elegans*
CANFA	Complete proteome from *Canis familiaris*
CAVPO	Complete proteome from *Cavia porcellus*
CHICK	Complete proteome from *Gallus gallus*
DANRE	Complete proteome from *Danio rerio*
DROME	Complete proteome from *Drosophila melanogaster*
FELCA	Complete proteome from *Felis catus*
HORSE	Complete proteome from *Equus caballus*
HUMAN	Complete proteome from *Homo sapiens*
MAIZE	Complete proteome from *Zea mays*
MOUSE	Complete proteome from *Mus musculus*
ORYBR	Complete proteome from *Oryza brachyantha*
ORYGL	Complete proteome from *Oryza glaberrima*

ORYSI	Complete proteome from *Oryza sativa subsp. indica*
ORYSJ	Complete proteome from *Oryza sativa subsp. japonica*
PANTR	Complete proteome from *Pan troglodytes*
PDBAA	Sequences derived from PDB protein structures (default)
PDBAA50	PDBAA with clustering of protein chains at 50% sequence identity
PDBAA70	PDBAA with clustering of protein chains at 70% sequence identity
PDBAA90	PDBAA with clustering of protein chains at 90% sequence identity
PDBAA95	PDBAA with clustering of protein chains at 95% sequence identity
PIG	Complete proteome from *Sus scrofa*
RABIT	Complete proteome from *Oryctolagus cuniculus*
RAT	Complete proteome from *Rattus norvegicus*
SWISSPROT	SwissProt database from UniProt Knowledgebase
TREMBL	TrEMBL database from UniProt Knowledgebase
XENTR	Complete proteome from *Xenopus tropicalis*
YEAST	Complete proteome from *Saccharomyces cerevisiae*

> The user can change the threshold for retaining sequence matches identified by the BLAST+ search (E-value option, default: 1e-6). The E-value gives an indication of the statistical significance of a given pairwise alignment. The lower the E-value is (or the closer it is to zero), the more significant the match is.

> The Discard identical seq. option, if enabled (default), allows ENDscript to keep only a single representative sequence when several identical sequence hits are found by the BLAST+ search. This option is useful to discard sequences of proteins with multiple identical chains or when the BLAST search is performed against a redundant database (notably PDBAA or TrEMBL).

• Multiple sequence alignment

- **Main role:** aligns all the sequence hits identified by the BLAST+ search with that of the PDB query.

This multiple sequence alignment can be performed by Clustal Omega (default), MAFFT, MSAProbs or MultAlin (Multiple seq. alignment program option).

If Clustal Omega or MAFFT are chosen, a dendrogram is calculated. It will be used, in the RESULTS pop-up window, to build and view a phylogenetic tree with the external PhylodendronWeb server.

> You can examine the ENDscript results with the online Jalview Desktop viewer (14) available in the RESULTS pop-up window. This tool allows multiple sequence alignment editing, visualisation and analysis. A secondary structure consensus, calculated by ENDscript, is included. In this consensus, the most present conformational state is reported for each residue. Finally, a downloadable file in **Stockholm format** allows to import ENDscript results in your own Jalview Desktop program - for more information, please refer to the **Jalview website**.

> The Sequences output order option allows the multiple sequence alignment program to present the sequences in the same order as they have been aligned from the guide tree (choose 'aligned'). They can also be displayed in the same order as they were identified by the BLAST+ search, from the lowest to the highest E-value (choose 'input', default).

> If the option Display all known structures is activated *via* the interface (default), an automatic search is performed to check if a sequence name can be related to a known 3D structure. This option has no effect in phase 1 and is functional when a BLAST+ search is performed against all databases but TREMBL.

> If the option Residue conservation rescaling is activated (default), residue conservation is rescaled according to the Global score threshold . If untick, this last is forced to 0. The default option allows a better color ramping (from white - low score - to red - identity) of the sequence conservation on the interactive 3D representations of the PDB query.

Known secondary structure elements of each matching sequence are displayed in turn in the ESPript figure.

• ESPript

- **Main role:** generates a second flat figure with a multiple sequence alignment adorned with secondary structure elements of each homologous sequence of known structure.
- Secondary structure elements have been calculated by DSSP in the previous step and:

> α-, 3_{10}- and π-helices are shown above sequence as medium, small and large squiggles with α, β and π labels, respectively,
>
> β-strands are shown as arrows labeled β,
>
> Strict α- and β-turns are marked by TTT and TT letters, respectively.

- Residues in an alternate conformation are highlighted by a grey star above sequences.
- Relative accessibility, calculated by DSSP in the previous step, is shown by a blue-colored bar below sequence. White is buried ($A < 0.1$),

cyan is intermediate (0.1 ≤ A ≤ 0.4), blue is accessible (0.4 < A ≤ 1) and blue with red borders is highly exposed (A > 1). A red box means that relative accessibility is not calculated for the residue, because it is truncated. Remark: only molecules located in the crystallographic asymmetric unit are taken into account by DSSP in its calculation of accessibility. Thus, you can find 'highly accessible' residues involved in contacts with crystallographic neighbors according to the ESPript figure. These residues are in fact buried in the crystal lattice.

- Hydropathy is calculated from the sequence according to the algorithm of Kyte & Doolittle [16] with a windows of 3. It is shown by a second bar below accessibility: pink is hydrophobic (H>1.5), grey is intermediate (-1.5 ≤ H ≤ 1.5) and cyan is hydrophilic (H < 1.5).

- Disulphide bridges, identified by DSSP in the previous step, are shown by green pairs of digits (`1 1`, `2 2` ...) below the bar of hydropathy.

- Protein-protein and protein-ligand contacts, calculated by CNS in the previous step, are displayed along with disulphide bridges below the bar of hydropathy. The shortest intermolecular distance is taken for each residue. Corresponding contact symbols (see **above paragraph**) are written in red if the distance is less than 3.2 Å and in black if the distance is in the range 3.2-5.0 Å.

- Main information is given according to the written marks, which shows protein-protein and protein-ligand contacts:

  > A to Z, 0 to 9 or a to z means that the concerned amino acid residue has a non-crystallographic contact with an amino acid residue of the Chain A to Z, 0 to 9 or a to z (*e.g.* this amino acid residue is involved in a non-crystallographic interface).

  > *A* to *Z*, *0* to *9*, *a* to *z* **in italic** means that the concerned amino acid residue has a crystallographic contact with an amino acid residues of the Chain A to Z, 0 to 9 or a to z (*e.g.* this amino acid residue is involved in a crystallographic interface).

  > # identifies a contact between two amino acid residues having the same names and numbers (*e.g.* along a 2-fold symmetry axis).

  > `* : " ^ @ < > /` means that the concerned amino acid residue has a contact with a ligand (*i.e.* an automatically kept or a chosen hetero-compound - see **above paragraph**).

  > *`* : " ^ @ < > /`* **in italic** means that the concerned amino acid residue has a crystallographic contact with a ligand (*i.e.* an automatically kept or a chosen hetero-compound - see **above paragraph**).

- Further information is given with colors:

  > A yellow background identifies a non-crystallographic contact.
  >
  > An orange background identifies an amino acid involved in both a crystallographic and a non-crystallographic contact.
  >
  > A blue frame identifies an amino acid involved in both a protein-protein and a protein-ligand contact.
  >
  > A red letter identifies a contact < 3.2 Å.
  >
  > A black letter identifies a contact between 3.2 Å and 5.0 Å.

- Similarities between the PDB query sequence of the chosen chainID (chain A by default - redefinable in Chain ID option) and homologous sequences aligned are rendered by a boxing in color. A score is calculated for each column of residues, according to a matrix based on physicochemical properties.

- By default, residue names are written in black if score is below 0.7 (low similarity); they are in red and framed in blue if score is in the range 0.7-1 (high similarity); they are in white on a red background in case of strict identity.

- You can switch to other scoring matrices once a first run of ENDscript has been done. These setting are available in the Sequence similarities depiction parameters box of the ENDscript form.

  - A percentage of Equivalent residues (%Equivalent option, default) can be calculated considering either physicochemical properties (`HKR` are polar positive ; `DE` are polar negative ; `STNQ` are polar neutral ; `AVLIM` are non polar aliphatic ; `FYW` are non polar aromatic ; `PG` ; `C`) or similarities used in MultAlin (`IV` ; `LM` ; `FY` ; `NDQEBZ`).

  - Risler PAM250 BLOSUM62 and Identity are other possibilities of scoring matrix (check **Appendix**). The Risler matrix gives usually an excellent rendering.

- Sequences can be removed or their order can be changed by using the box Defining group and the following syntax:

  - `1-3 6-10`, removes sequences 4 and 5 from a 10 sequences alignment.
  - `1 3 2 4 5`, swaps the order of sequences 2 and 3 from a 5 sequences alignment.
  - `2 all`, display sequence 2 first than all the others.
  - Warning: query sequence (sequence 1) must be kept otherwise ENDscript produces an error.

> With the ESPRIPT button, you can export your ENDscript results to the ESPript server. There, you will have a better grip on the layout and you will be able to edit and enhance your sequence illustrations and save your session on your own computer.

5 Phase 3 in details

- **Result:** produces two interactive 3D PyMOL representations of the PDB query.

• ProFit

- **Main role:** superposes all identified homologous structures onto the PDB query.

So as to superpose each known structure onto the PDB query, information on zones of equivalent residues must be known. This can be achieved by two distinct methods, controlled by the Pairwise 3D structures superposition option.

If enabled (default), ProFit performs a 3D superposition of the PDB query with each homologous protein based using a pairwise Needleman & Wunsch sequence alignment as guide.

If disabled, the global sequence alignment of the PDB query with each homologous protein is used instead.

> Enabling this option is recommended because it improves the structural alignment and the calculation of the r.m.s deviation per residue. Disabling this option is only recommended in case of highly similar sequences hits and/or for multiple sequence alignments with few gaps.

For both methods, each mobile structure is fitted onto the reference structure (the PDB query) by using Cα pairs.

> Fitted structures are written in turn in a `.zip` file archive, downloadable from the RESULTS pop-up window.

Finally, a mean r.m.s. deviation per residue is calculated using all fitted Cα pairs. It will be used afterwards in the PyMOL 'Sausage' representation.

• PyMOL-ScriptMaker

- **Main role:** generates 3D interactive 'Cartoon' and 'Sausage' representations.

The program PyMOL-ScriptMaker gathers all previously calculated information and prepares two PyMOL session files:

- The first PyMOL representation is named 'Cartoon':
  - This is a ribbon depiction of the PDB query colored as a function of similarity scores calculated from the previous multiple sequence alignment.
  - This color ramping from white (low score *i.e.* %equivalent limit, 0.7 by default) to red (identity) allows to quickly locate areas of weak and strong sequence conservation on the structure of the query.
  - A solvent-accessible surface can be mapped with the same coloring code *via* the PyMOL control panel.

- The second PyMOL representation is named 'Sausage':
  - It shows a variable tube representation of the Cα trace of the PDB query.
  - For this drawing, all homologous protein structures were superposed onto the PDB query with ProFit and the size of the tube is proportional to the mean r.m.s. deviation per residue between Cα pairs.
  - The same white to red color ramping is used to visualize sequence conservation.
  - Hence, the user can identify areas of weak and strong structural conservation and correlate this result with sequence conservation.

If applicable, these two PyMOL representations can display an assortment of supplementary information compiled by ENDscript:

- Biological unit in grey Cα trace representation,
- All NMR models in light pink Cα trace representation,
- Disulfide bridges in yellow stick representation,
- Side chains in line representation colored as a function of the conservation score,
- Nucleic acids in cartoon representation,
- Ligands in ball and stick representation,
- Contacting residues in pale green stick representation,
- Monatomic elements in dotted sphere representation,
- Identical residues in dark pink ball and stick representation and highlighted,
- PDB SITES markers in blue mesh representation,
- Solvent-accessible surface colored as a function of the conservation score,
- Sequence viewer.

These two representations can be downloaded and interactively examined with the molecular 3D visualization program PyMOL installed on the user's computer.

> Expert users can also download a `.zip` file archive containing PyMOL `.pml` script and associated necessary files to manually edit them (please refer to PyMOL documentation or **PyMOLWiki**).

6 Alignments output layout and file formats

- The following options controls the layout of ENDscript flat figures generated during phases 1 and 2. You can render these figures in a variety of output formats and sizes. These settings have no effect on the two PyMOL 3D interactive representations.

  - Font size : font size in points (monospaced '`Courier`' font for sequence names and residues) (default: 6).
  - Number of columns : number of residue columns per line (default: 140).
  - Color scheme :
    - Normal: standard color scheme (default).
    - Flashy: flashy colors, similar residues are written with black bold characters and boxed in yellow.
    - Thermal: colored with all letters in bold, ideal for article figures.
    - Slide: light cyan background, ideal for slides.
    - B&W: a grey scale is used.
  - Orientation : Portrait (default) or Landscape.
  - Paper size : A4, A3 (default), A0, US letter or Tapestry (width: 0.8m x height: 3.3 m).

> Rendering PNG or TIFF images may take some time, especially if you use the 300 dpi or 600 dpi options. Hence, high dpi formats (>150 dpi) are only recommended for publication-quality figures. For examining the ENDscript flat figures, PDF format is recommended.

- PostScript and PDF files can be edited with **Adobe Illustrator™**. PDF files are viewable and printable from **Adobe Reader™**.

7 Appendix

• Similarity scores

If Risler BLOSUM62 PAM250 or Identity , several scores are calculated:

- **in-Group Score** (*ISc*) is a classical computation of a similarity score within each group.

  For a column made of 3 residues ACD:
  ISc = (AC+AD+CD) ÷ 3

- **Cross-Group Score** (*XSc*) is the similarity score average for every sequence pair, where each sequence belongs to a different group.

  For a column made of 6 residues divided in 3 groups (ACD)(DE)(G):
  XSc = [(AD+AE+CD+CE+DD+DE)÷6+(AG+CG+DG)÷3+(DG+EG)÷2] ÷ 3

- **Total Score** (*TSc*) is the mean of **in-Group Score** and **Cross-Group Score**:

  TSc = (*ISc* + *XSc*)÷2

The user specifies a threshold for **in-Group** (*ThIn*) and **Diff-Group** (*ThDiff*) scores.
Colours are chosen according to the following rule:

A Red box, white character → Strict **identity**.

Y Red character (or black bold character with color scheme "Flashy") → **Similarity in** a group: *ISc* > *ThIn*.

T Blue frame (filled in yellow with color scheme "Flashy") → **Similarity across** groups: *TSc* > *ThIn*.

Q Green fluo box → **Differences** between conserved groups: (*ISc*-*Xsc*)÷2 > *ThDiff*.

• Similarity scores matrices

Risler matrix (17)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	.
A	22	-15	2	17	6	6	-6	17	14	13	10	13	-2	18	15	20	19	20	-9	2	-30
C	-15	22	-17	-15	-16	-17	-18	-16	-16	-15	-16	-16	-18	-14	-15	-13	-14	-14	-18	-11	-30
D	2	-17	22	10	-3	-4	-13	0	1	-2	-5	8	-12	6	-1	7	0	0	-14	-4	-30
E	17	-15	10	22	6	3	-6	15	14	9	6	14	-1	21	19	18	16	16	-10	2	-30
F	6	-16	-3	6	22	-4	-11	10	1	10	-2	4	-11	7	4	5	3	8	-9	20	-30
G	6	-17	-4	3	-4	22	-12	0	-1	-2	-4	2	-12	2	1	7	2	1	-13	-2	-30
H	-6	-18	-13	-6	-11	-12	22	-8	-10	-9	-12	-3	-16	-5	-4	-4	-9	-7	-17	-8	-30
I	17	-16	0	15	10	0	-8	22	10	21	9	9	-6	14	14	16	16	22	-7	4	-30
K	14	-16	1	14	1	-1	-10	10	22	7	4	10	-7	17	21	14	12	12	-11	5	-30
L	13	-15	-2	9	10	-2	-9	21	7	22	18	8	-8	11	12	13	12	20	-8	5	-30
M	10	-16	-5	6	-2	-4	-12	9	4	18	22	0	-12	12	11	6	8	8	-13	-2	-30
N	13	-16	8	14	4	2	-3	9	10	8	0	22	-10	16	12	19	11	11	-11	-1	-30
P	-2	-18	-12	-1	-11	-12	-16	-6	-7	-8	-12	-10	22	-6	-3	-3	-5	-6	-16	-12	-30
Q	18	-14	6	21	7	2	-5	14	17	11	12	16	-6	22	20	18	17	15	-10	5	-30
R	15	-15	-1	19	4	1	-4	14	21	12	11	12	-3	20	22	20	19	15	-8	8	-30
S	20	-13	7	18	5	7	-4	16	14	13	6	19	-3	18	20	22	21	18	-8	4	-30
T	19	-14	0	16	3	2	-9	16	12	12	8	11	-5	17	19	21	22	16	-10	3	-30
V	20	-14	0	16	8	1	-7	22	12	20	8	11	-6	15	15	18	16	22	-7	3	-30
W	-9	-18	-14	-10	-9	-13	-17	-7	-11	-8	-13	-11	-16	-10	-8	-8	-10	-7	22	-6	-30
Y	2	-11	-4	2	20	-2	-8	4	5	5	-2	-1	-12	5	8	4	3	3	-6	22	-30
.	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	0

PAM250 matrix (18)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	.
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-4	1	1	1	-6	-3	0	-15
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-15
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-4	-1	1	0	-4	-2	-2	-15
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	-15
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-15
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	-15
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	-15
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	-1	1	0	-7	-5	-1	-15
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	-15
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-15
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-15
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	-15
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-15
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-15
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-15
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	-15
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	-15
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-15
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-15
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-15
.	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	0

BLOSUM62 matrix (19)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	.
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-4

Identity matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	.
A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

```
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -4
. -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```

```
V  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0
.  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

8 References

1. Kabsch, W., and Sander, C. (1983) *Biopolymers* **22**(12), 2577-2637.
2. Joosten, R. P., te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., Sander, C., and Vriend, G. (2012) *Nucleic Acids Res.* **39**(Database issue), D411-419.
3. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) *Acta Cryst. D***54**, 905-921.
4. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009) *BMC bioinformatics* **10**, 421.
5. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D., and Higgins, D. G. (2011) *Molecular systems biology* **7**, 539.
6. Katoh, K., and Standley, D. M. (2013) *Mol. Biol. Evol.* **30**, 772-780.
7. Yongchao, L., and Bertil, S. (2014) *Methods Mol. Biol.* **1079**, 211-218.
8. Corpet, F. (1988) *Nucleic Acids Res.* **16**(22), 10881-10890.
9. Gouet, P., Courcelle, E., Stuart, D. I., and Metoz, F. (1999) *Bioinformatics* **15**(4), 305-308.
10. Gouet, P., and Courcelle, E. (2002) *Bioinformatics* **18**(5), 767-768.
11. Gouet, P., Robert, X., and Courcelle, E. (2003) *Nucleic Acids Res.* **31**(13), 3320-3323.
12. Martin, A. C. R., and Porter, C. T. (2009) ProFit 3.1 *Ed. Martin, A.C.R., London.*
13. Schrödinger, LLC. (2013) The PyMOL Molecular Graphics System, *www.pymol.org*
14. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009) *Bioinformatics* **25**(9), 1189-1191.
15. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., and Zardecki, C. (2002) *Acta Cryst. D***58**, 899-907.
16. Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.* **157**(1), 105-132.
17. Risler, J. L., Delorme, M. O., Delacroix, H., and Henaut, A. (1988) *J. Mol. Biol.* **204**(4), 1019-1029.
18. Dayhoff, M. (1978) Atlas of protein sequences and structure, *National Biomedical Research Foundation, Washington, D.C.*
19. Henikoff, J. G., and Henikoff, S. (1996) *Methods in enzymology* **266**, 88-105.

User guide last revision: August 8, 2017