ENDscript 2.2

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 "**MMSB**" laboratory (UMR5086 **CNRS** / **Univ. Lyon 1**). ENDscript is an application supported by **SBGrid**.

This documentation contains all the information you need to use the ENDscript Web server, whether you are a beginner or an advanced user.

The following two 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 user-friendly web server that extracts and renders a comprehensive analysis of primary to quaternary protein structure information in an automated way from a PDB or CIF query file.**

- ENDscript is a tool of choice for biologists and structural biologists, allowing them to generate a set of detailed, high-quality figures and interactive 3D representations of their proteins of interest with just a few mouse clicks.

- The ENDscript Web server is fast and convenient:

  - No special knowledge of bioinformatics is needed to obtain comprehensive and relevant illustrations.
  - The user is guided through the process by tooltips and detailed help topics (this documentation), which are accessible at any time.
  - Thanks to its automated pipeline and a parallel programming, ENDscript can deliver results in one click and within a minute.
  - Advanced or expert users can change settings to fine-tune ENDscript to their needs.
  - ENDscript produces publication-quality illustrations in most common file formats (PDF, PostScript, PNG and TIFF) and sizes (US Letter, A4, A3, A0 and the gigantic 'Tapestry' format).
  - ENDscript is accessible from any 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

ENDscript's automated pipeline chains together multiple sequence and structure analysis software:

- **MAXIT**, to convert structure files from CIF to PDB format.
- **SPDB**, a homemade program to check residue numbering and chainIDs from the query structure 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 for protein homologues using the sequence of the PDB/CIF 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 with known 3D structures on the PDB/CIF query.
- **PyMOL**[13], to generate scripts and session files to display sequence and structure conservation with interactive 3D representations.
- **NCBI's Tree Viewer** web server, to display a phylogenetic tree.
- **JalviewJS or Jalview Desktop**[14], for multiple sequence alignment editing, visualisation and analysis.

These software are executed sequentially in three successive phases:

• Phase 1

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

In the first field of the ENDscript interface (`Query PDB or CIF file`), the form must be filled in by at least:

- clicking on the PDB icon and entering the PDB entry code (*e.g.* `2CAH`) of your protein structure (NMR and crystallographic structures are supported),
- or uploading your own PDB or CIF file by clicking on the `Browse` button (or equivalent depending on your browser language).

Click on SUBMIT in the buttons frame.

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

A SPDB output file is generated and passed 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 that provides the following information for each monomeric sequence contained in your query structure:

- Secondary structure elements and residues in alternate confirmation are shown above the 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 query structure is performed against a chosen sequence database (PDBAA by default) to detect protein homologues.

The result is passed 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.
- Additional 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 called 'Cartoon':**

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

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

  - It shows a variable tube representation of the Cα trace of the query structure.
  - In this goal, all homologous protein structures are superimposed on the query structure using ProFit, and the size of the tube is proportional to the per residue r.m.s. deviation between Cα pairs.
  - The same white-to-red color ramp is used to visualize sequence conservation.
  - By combining these two pieces of 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 a range of additional data via the PyMOL control panel:

- 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 color code.

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

All the resulting files from phases 1 to 3 can be visualized with a mouse click or downloaded with the right button / "Save As".

3 Phase 1 in details

- **Result:** a first ENDscript flat figure is produced with information about each monomeric sequence contained in the query structure:

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

• SPDB

- **Main function:** checks and cleans up the query structure before it enters the ENDscript automated pipeline.

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

- If necessary, SPDB reassigns chainIDs from A to Z, a to z and 0 to 9.
- First model is retained for multiple NMR models.
- First conformer is kept for alternate residues.
- Second oxygen atom of C-terminus main chain is removed (atom OXT).

> In the case of a query structure with multiple chains, the user can specifically select the chain to be processed by ENDscript

> (Chain ID option). Warning: this option is case sensitive.

> If the query structure contains selenomethionine residues (three-letter residue name MSE), the user can replace the latter with 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 query structure. To do so, the user must enter their names in the Keeping unsupported hetero-compounds tabular form (up to 10 names of 2-3 characters per column and one name per row).

Hetero-compound type	Name	Symbol
Nucleotides	ADE ADP AMP ATP CYT FAD FMN GDP GTP GUA NAD NAH NAP NDP THY URI	*
	A T G C U DA DT DG DC	/
Porphyrin groups	BCL BCB BPH CLA CHL HEC HEM MQ7	:
Sugars	FUC GAL GLC MAN NAG SIA XYL	"
Ions	CA CL CO CU FE K MG MN NA ZN	+
Fatty acids & miscellaneous	CIT MYR OLA PLM	^

> 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 user's assignment in the Keeping unsupported hetero-compounds tabular form.

> By default, these symbols are shown in red if the distance of the protein:hetero-compounds contacts is less than 3.2 Å and in black if it is greater.

• DSSP

- **Main function:** 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 disulfide bridges are identified.

• CNS

- **Main function:** calculates protein:protein and protein:ligand contacts.
  - CNS calculates both crystallographic and non-crystallographic contacts between each protein molecule.
  - It also calculates contacts between protein residues and hetero-compounds, if the 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 or CIF file).
  - The upper limit for the 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 used for each residue.

• ESPript

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

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

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

> α-, 3_{10}- and π-helices are shown above the 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 indicated by the letters TTT and TT, respectively.

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

- The relative accessibility, calculated by DSSP in the previous step, is shown by a blue-colored bar below the 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 edges is highly exposed (A > 1). A red box means that relative accessibility is not calculated for the residue, because it is truncated. Note: only molecules located in the crystallographic asymmetric unit are taken into account by DSSP in its accessibility calculation. Thus, you can find 'highly accessible' residues involved in contacts with crystallographic neighbors according to the ESPript figure. These residues are actually buried in the crystal lattice.

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

- Disulfide 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 disulfide bridges below the bar of hydropathy. The shortest intermolecular distance is taken for each residue. Corresponding contact symbols (see **paragraph above**) 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, a to z and 0 to 9 means that the amino acid residue in question has a contact with an amino acid residue in the chain of the letter/number displayed (*e.g.* this amino acid residue is involved in a non-crystallographic interface).

> *A* to *Z*, *a* to *z*, *0* to *9* **in italic** means that the amino acid residue in question has a crystallographic contact with an amino acid residue in the chain of the letter/number displayed (*e.g.* this amino acid residue is involved in a crystallographic interface).

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

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

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

- Further information is given with colors:

> A yellow background indicates 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 indicates 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 function:** finds sequences homologous to that of the query structure.

> 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*
CANLF	Complete proteome from *Canis lupus 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*
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*
ORYSI	Complete proteome from *Oryza sativa subsp. indica*
ORYSJ	Complete proteome from *Oryza sativa subsp. japonica*
PANTR	Complete proteome from *Pan troglodytes*
PIG	Complete proteome from *Sus scrofa*
RABIT	Complete proteome from *Oryctolagus cuniculus*
RAT	Complete proteome from *Rattus norvegicus*
XENTR	Complete proteome from *Xenopus tropicalis*

YEAST	Complete proteome from *Saccharomyces cerevisiae*
SWISSPROT	SwissProt database from UniProt Knowledgebase
ALPHAFOLD_DB	AlphaFold Protein Structure Database (AI-driven predictions) (Model Organism Proteomes, Global Health Proteomes & Swiss-Prot subsets - **see details here**)
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

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 multiple 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 (especially PDBAA).

• Multiple sequence alignment

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

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 is selected, a dendrogram is calculated. It is used, in the RESULTS pop-up window, to construct and view a phylogenetic tree using the **NCBI's Tree Viewer** server.

You can examine the ENDscript results with the online JalviewJS viewer [(14)] available in the RESULTS pop-up window. This tool allows editing, visualization, and analysis of multiple sequence alignments. 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 software - 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 were aligned from the guide tree (select 'aligned'). They can also be displayed in the same order in which they were identified by the BLAST+ search, from the lowest to the highest *E*-value (select 'input', default).

When the Display all known structures option 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.

When the Residue conservation rescaling option is enabled (default), the residue conservation is rescaled according to the Global score threshold . If unticked, the latter is forced to 0. The default option allows a better color ramp (from white, low score to red, identity) of the sequence conservation in the interactive 3D representations of the query structure.

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

• ESPript

- **Main function:** 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 the 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 indicated by the letters TTT and TT, respectively.

○ Residues in an alternate conformation are highlighted by a grey star above the sequence.

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

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

○ Disulfide 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 disulfide bridges below the bar of hydropathy. The shortest intermolecular distance is taken for each residue. Corresponding contact symbols (see **paragraph above**) 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, a to z and 0 to 9 means that the amino acid residue in question has a contact with an amino acid residue in the chain of the letter/number displayed (*e.g.* this amino acid residue is involved in a non-crystallographic interface).

> *A* to *Z*, *a* to *z*, *0* to *9* **in italic** means that the amino acid residue in question has a crystallographic contact with an amino acid residue in the chain of the letter/number displayed (*e.g.* this amino acid residue is involved in a crystallographic interface).

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

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

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

- Further information is given with colors:

> A yellow background indicates 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 indicates a contact between 3.2 Å and 5.0 Å.

- Similarities between the query structure sequence of the selected chainID (chain A by default - redefinable in Chain ID option) and aligned homologous sequences 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 the score is less than 0.7 (low similarity); they are in red and framed in blue if the score is in the range 0.7-1.0 (high similarity); they are in white on a red background if they are strictly identical.

- You can switch to other scoring matrices after a first run of ENDscript. These settings 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 possible scoring matrices (see **Appendix**). The Risler matrix usually gives an excellent representation.

- Sequences can be removed or their order can be changed by using the Defining group box and the following syntax:
  - `1-3 6-10` removes sequences 4 and 5 from an alignment of 10 sequences.
  - `1 3 2 4 5` swaps the order of sequences 2 and 3 from in a 5 sequence alignment.
  - `2 all` display sequence 2 first, then all the others.
  - Warning: the query sequence (sequence 1) must be kept otherwise ENDscript will produce an error.

> The ESPRIPT button allows you to 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 query structure.

• ProFit

- **Main function:** superimposes all identified homologous structures onto the query structure.

In order to superimpose any known structure on the query structure, information about equivalent residue zones must be known. This can be done in two different ways, controlled by the Pairwise 3D structures superposition option.

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

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

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

For both methods, each mobile structure is fitted to the reference structure (the PDB/CIF query) using Cα pairs.

> The fitted structures are written in turn to a zip file archive, which can be downloaded 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 function:** generates 3D interactive 'Cartoon' and 'Sausage' representations.

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

- The first PyMOL representation is called 'Cartoon':
  - This is a ribbon depiction of the query structure colored as a function of the similarity scores calculated from the previous multiple sequence alignment.
  - This color ramp 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 query structure.
  - A solvent-accessible surface can be mapped with the same color code using the PyMOL control panel.

- The second PyMOL representation is called 'Sausage':
  - It shows a variable tube representation of the Cα trace of the query structure.
  - For this representation, all homologous protein structures were superimposed on the query structure using 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 ramp is used to visualize sequence conservation.
  - This allows the user to 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&stick representation,
- Contacting residues in pale green stick representation,
- Monatomic elements in sphere representation,
- Identical residues in dark pink ball&stick representation,
- 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.

> Advanced users can also download a zip file archive containing the PyMOL .pml script and the necessary files for manual editing (see the PyMOL documentation or **PyMOLWiki**).

6 Alignments output layout and file formats

- The following options control the layout of the ENDscript flat figures generated in 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 interactive PyMOL 3D 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 row (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 when using the `300 dpi` or `600 dpi` options. Therefore, high dpi formats (>150 dpi) are only recommended for publication quality images. The PDF format is recommended for viewing ENDscript flat figures.

- PDF and PostScript files can be edited with **Adobe Illustrator™**. PDF files can be viewed and printed using **Adobe Reader™**.

7 Appendix

• Similarity scores

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

- **in-Group Score** (*ISc*) is a classical calculation 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 into 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.
Colors are selected using 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
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
```

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: November 18, 2025

SBGrid Consortium