Université Paris Cité

M2 BIOINFORMATIQUE

LONG PROJECT

Year 2024-2025

Optimisation of the *SA-conf* program and application to the study of the Bcl-Xl protein

---

Eliott TEMPEZ

Referent teacher:

Leslie REGAD

HTTPS://GITHUB.COM/ELIOTT-TEMPEZ/M2_SA-CONF_OPTIMISATION

1 Introduction

Proteins are flexible molecules, and this property plays a key role in their interactions with partners. Studying their three-dimensional structure is therefore essential to better understand their function. In the PDB (Protein Data Bank) database, there is much redundancy when a single protein is taken into account, meaning that there are often several conformations available for the same molecule. This is explained by the wide range of parameters used for structure solving, such as different methodologies, the presence or absence of partners, the study of wild-type or mutated proteins, or even the resolution quality. This redundancy is an advantage; the study of different structures for the same protein can provide key information on the variability and function of the protein. Thus, the SA-conf tool has been designed to analyse the different conformations of a single protein based on experimentally resolved multiple target conformations (MTC) [1]. It provides an analysis of variability at three levels: protein sequence, secondary structures, and three-dimensional structure.

SA-conf aims to identify regions of flexibility in proteins and to provide insight on their origin: partner binding, mutations, or intrinsic properties for example. It has been used to explore these types of parameters in proteins such as p53 and HIV-2 protease [1]. Although SA-conf can be applied to diverse sets of MTC, is effective to quantify the structural variability of a MTC set, and to localise the structural variable regions of the target, its accuracy heavily depends on the quality and diversity of the datasets, which are influenced by experimental conditions or modelling assumptions. Thus, the goal of this project is to improve SA-conf so as to have a more detailed analysis of MTC structural properties, and to apply these improvements to experimental data, in order to characterise the different conformations of a protein with resolved structures.

2 Materials and methods

2.1 SA-conf

2.1.1 SA-conf steps

SA-conf is written in Python 2 and R, and takes as input a text file with the list of the PDB IDs of interest. It can also accept a directory path that contains PDB files, and it returns a set of files with data and graphs regarding the structural analysis. SA-conf runs in several steps:

1. Description of each structure among the MTC: experimental approach, resolution, number of chains, chain length, etc.

2. Protein sequence extraction and multiple sequence alignment (MSA) using ClustalW (default) or T-coffee. A pre-determined multiple alignment can also be given as input.

3. Translation according to a structural alphabet and alignment:

   SA-conf is based on the HMM-SA structural alphabet [2], a classification of 4 alpha carbon-long (4-αC) fragments of proteins according to their geometry, obtained using hidden Markov models (HMM). This classification contains 27 classes, named structural letters (SL) labelled [a, A-Z]. Of these letters, 4 correspond to alpha helix conformations, 5 to beta strands, and the remaining 18 to loops. Using HMM-SA, a structure is simplified into a sequence of structural letters, where each SL corresponds to the fold of a 4-αC fragment.

   In SA-conf, HMM-SA is used to extract the local conformation for each residue. To do so, HMM-SA is used to simplify each 3D structure into a SL sequence, where each SL corresponds to the fold of a 4-αC fragment. After this step, the MSA is translated into a multiple structural letters alignment (MSLA) by replacing each amino-acid (AA) letter by the corresponding SL.

4. The fourth step consists in comparing the AA and SL of each position within the different structures, to locate :

   - the conserved positions in terms of each sequence and structure, i.e the positions having the same AA or SL for each structure
   - the variable positions in terms of sequence or structure, i.e the positions having different AA or SL within the structure.

   To quantify the structure conservation, SA-conf computes the neq_{SL} that measures the number of major SL per position. To quantify the sequence conservation, SA-conf computes the neq_{AA} that measures the number of major AA per position:

   - if $neq(i) = 1$: letter (AA or SL) strictly preserved at position i.
   - if $neq(i) \in]1, 1.5[$: slightly variable position
   - if $neq(i) \geq 1.5$: variable position

5. Output generation

2.1.2 SA-conf outputs

- Global and alignment information:
  - *dataset_composition.csv*: list of parameters for each submitted PDB ID (methodology, resolution, number of chains, ...)
  - in the */PDB* folder : the PDB files automatically downloaded
  - *Mutation_res.txt* : information of all mutations in the structures
  - the MSA and MSLA files in several formats, including *fasta*
- Information based on the neq:

  - *position_description.csv* : variability indexes for each MSA position for all the structures
  - *Count_position_type.txt* a text file with the neq interpretation for each position in the alignment
  - *Structural_Variable_position_res.txt* : list of all variable positions
  - *graph/Neq_graph.pdf* : visualisation of the neq_{AA} and neq_{SL} for each position, coloured by type of secondary structure (fig. 3)
  - *script_pymol_.plm* : PyMol script with the structure of the first target protein in the multiple alignment file, coloured accordingly to the variable positions

- Heatmaps:
  - *graph/AA-alignment.pdf* : Heatmap of the MSA (fig. 1a)
  - *graph/SL-alignment.pdf* : Heatmap of the MSLA (fig. 1b)

2.2 Illustration

2.2.1 Data presentation

Bcl-Xl (extra large B-cell lymphoma) is a protein of the Bcl-2 family, which includes inhibitors and inducers of cell death [3]. Within this family, Bcl-Xl is an anti-apoptosis protein; it prevents the release of mitochondrial contents such as cytochrome c, which then leads to programmed cell death [4]. It also has anti-autophagy properties, inhibiting the initiation of the autophagosome formation [5]. Bcl-Xl is overexpressed in many cancers, and its inhibitors have shown good therapeutic effects [6].

The dataset comprises structural data of Bcl-Xl structures obtained from the PDB, exclusively resolved by X-ray diffraction. It includes 65 unique entries (149 chains), with resolutions ranging from 1.3 Å to 3.45 Å. These structures were obtained from the UniProt ID Q07817, by filtering out chains that were too small, non-Bcl-xL sequences, or unrelated to the target, ensuring consistency in sequence length (between 130 and 155 residues). The structures include ligand-bound and ligand-free states, and several of them contain multiple models or crystallised heteroatoms such as sulphate, glycerol, or other small molecules.

2.2.2 Structure characterisation

We compared our results with physical properties of the structures, extracted from the PDB file:

- The total buried surface area, which refers to the surface area of the protein that is ”buried” or not exposed to the solvent

- The surface area of the complex, meaning the total surface area of the entire protein that is exposed to the solvent

- The change in solvent free energy, which is the change computed as the difference in the system's energy, due to the interaction with the solvent

These parameters are computationally generated, and integrated into the PDB file by the PISA software, an interactive tool for the exploration of macromolecular interfaces.

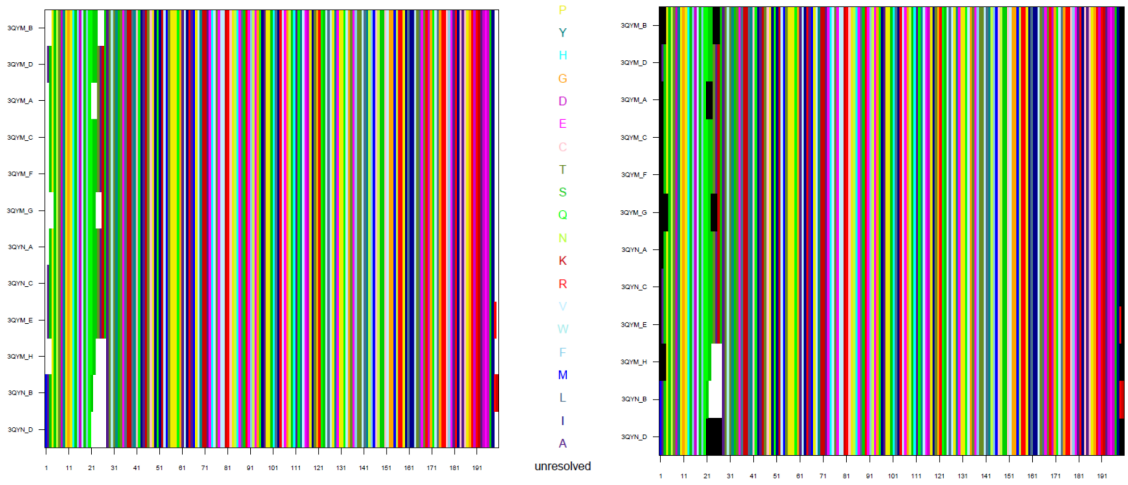
3 Results

3.1 Program improvements

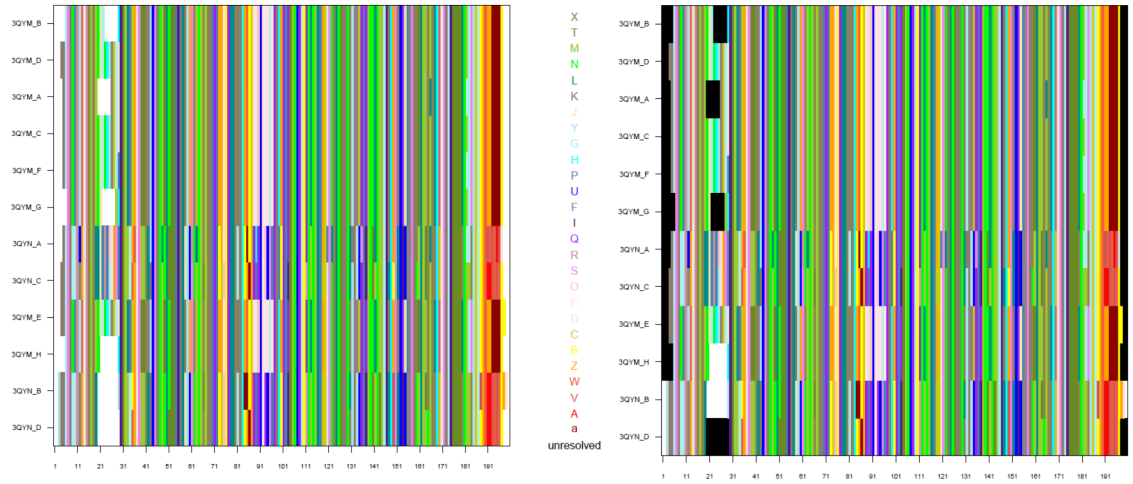
The program was first updated to Python 3, to make it more compatible with current code standards, and to rid it of deprecated methods and libraries. We then added improvements focusing on more detailed characterisation of protein structural analysis:

Missing residues

During the experimental resolution of structures, residues that are present in the protein are sometimes not modelled for various reasons: they may have a conformation that is too flexible for their position to be resolved, or they may have been degraded by the crystallisation or purification process. The current heatmap visualisation of the multiple alignments does not allow us to differentiate between unmodelled residues and gaps introduced during alignment. We improved the MSA and MSLA visualisation by implementing the differentiation of these two different types of gaps using distinct colours (fig. 1). To do this, the algorithm checks whether the residues surrounding a gap have consecutive numbers, and if not, whether the numbers between these two ends have corresponding residues in the protein sequence. If the gap is made up of a mixture of non-elucidated residues and gaps introduced by the alignment, the black colour will be centred in the gap in question. Similarly, as the translation into a SL sequence is based on HMM-SA taking into account 2 residues before and 1 residue after the iterated residue, the gaps within the alignment from the structural letters (fig. 1b) will always be $+3$ longer than those from the amino acids (fig. 1a). These three complementary missing residues will be coloured the same colour as the corresponding gap edges in the amino acid alignment.



(a) MSA



(b) MSLA

Figure 1: Heatmaps of the MSA (top) and MSLA (bottom) of the human p63. The new version (right) differentiates unmodelled residues in black, whereas the old version (left) left all gaps in white, regardless of their nature.

Uncertain coordinates

Sometimes, some residues have been modelled, but their modelling quality is inferior compared with the rest of the model. The quality of the model can be measured by the RSRZ score, that quantifies the quality of the fit between part of a model (in this case, a residue) and the data in real space [7]: for a $RSRZ > 2$, the residue is considered as an outlier. As it is important to identify the residues with an uncertain conformation, we decided to represent on the heatmaps the positions in the multiple alignment for which more than half of the residues are outliers (fig. 2). To do this, we extracted the RSRZ of the alpha carbon for each residue from the XML file available on PDB. A text file is also created with information on the positions in question.

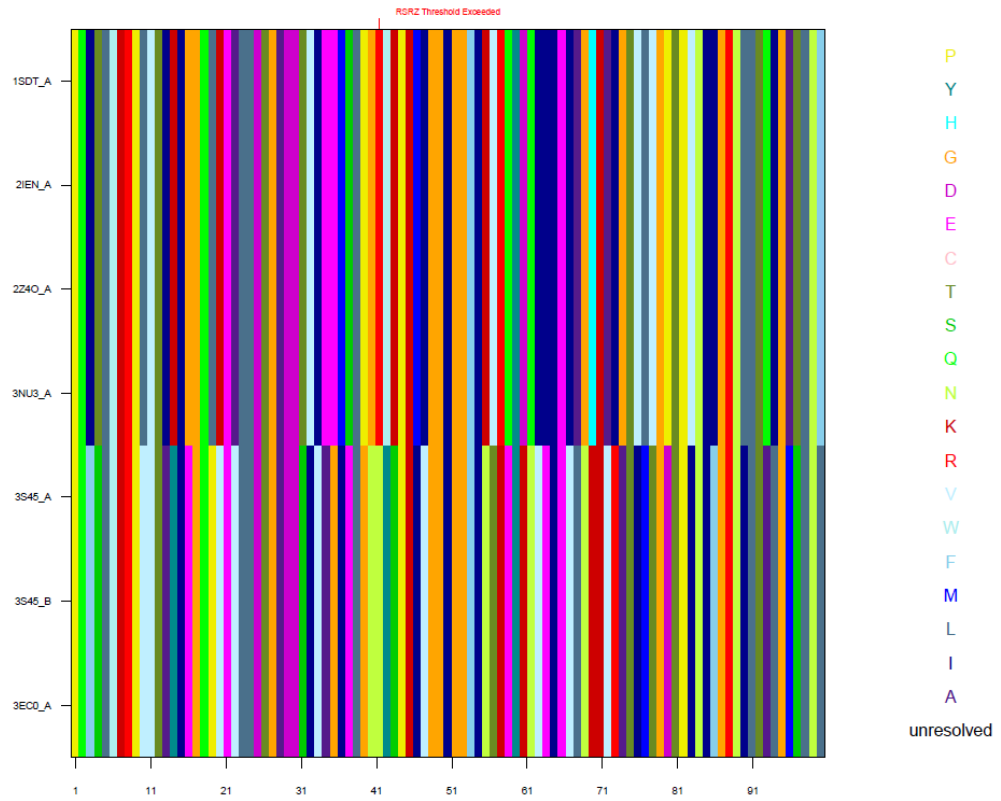


Figure 2: Heatmap of the MSA for several structures of the human HIV protease. We can see a vertical segment at the top at the position 41, where more than 50% of the residues have a $RSRZ > 2$. The same segment will also be incorporated into the MSLA heatmap, not depicted here. If there is no outlier, the sentence 'RSRZ Threshold exceeded' will not be displayed.

Residue flexibility

The B-factor is a parameter which quantifies the degree of thermal movement and static disorder of an atom in the crystalline structure of a protein [8]. The greater the degree, the more variable the position of the atom, and therefore the more flexible the residue. Thus, relating the neq to the flexible nature of residues at a particular position is relevant for studying the intrinsic nature of the protein's flexibility, in relation to possible sequence variations [9]. To do this, we extracted the B-factor for each of the alpha carbons of the residues, and we centred and reduced these values by conformation, taking into account the B-factor for each Cα in the structure in question. We then calculated the average B-factor value for each position in the multiple alignment. An average B-factor > 0 indicates a flexible position, and an average B-factor < 0 a rigid position. This information was added to the neq graph (fig. 3).

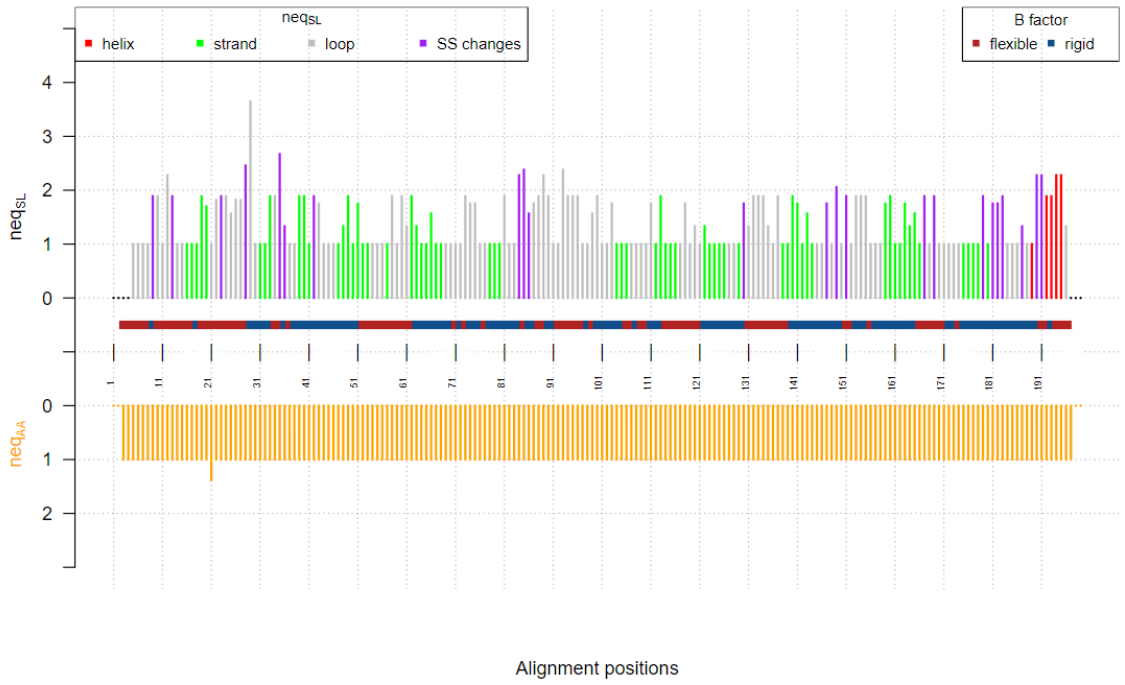


Figure 3: Graph of neq_{SL} and neq_{AA} for several structures of the p63. Information on the B-factor has been added above the position numbers, and the legend has been added to the top right.

Clustering

Another interesting question concerns the similarity of the structures: as SA-conf allows for the quantification of the structural and sequence variability, we can compute the structures that most resemble among a MTC set. Thus, we decided to implement a clustering of the different structures, based on the similarity of the SL: first, a distance matrix between the SL was computed using the Hamming distance. We then performed a hierarchical classification based on this matrix, with Ward's aggregation method. From this hierarchical classification, clusters of structures were extracted after cutting the dendrogram with a height threshold. To determine the optimal number of clusters, we used the silhouette method. The graphical output returns the heatmap of the structural sequences according to the identified clusters, and the hierarchical classification tree (fig. 4).

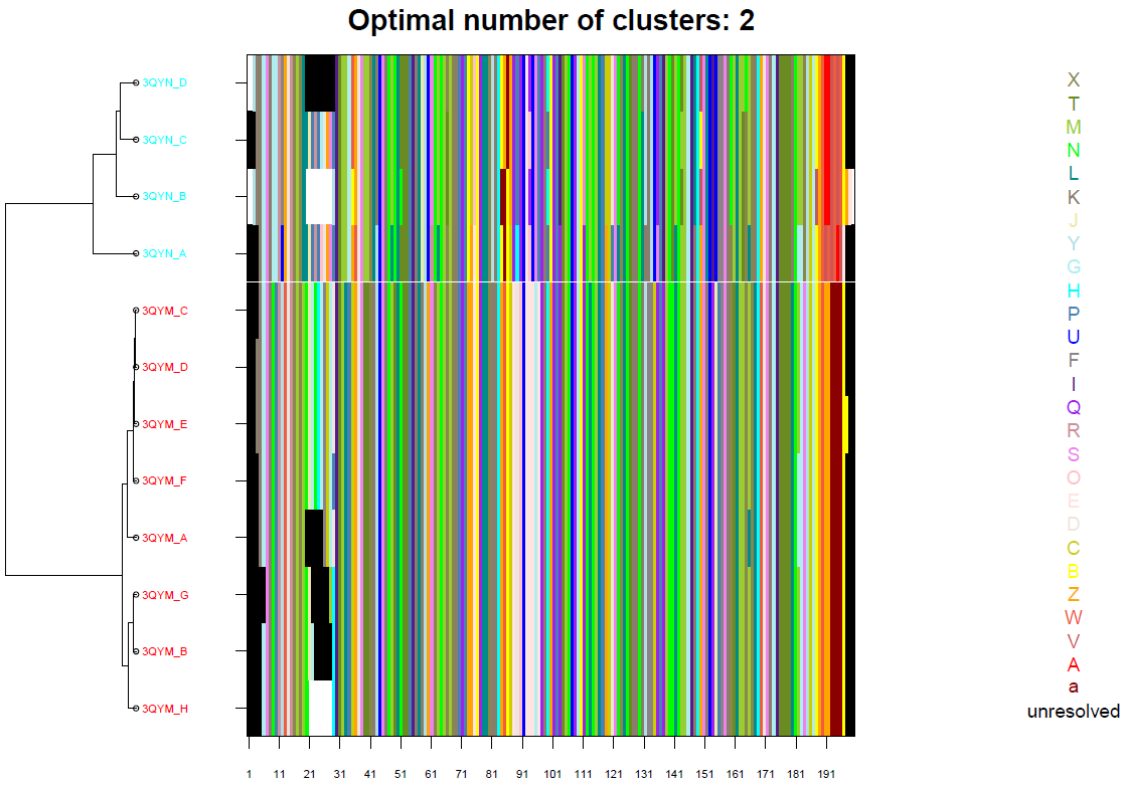


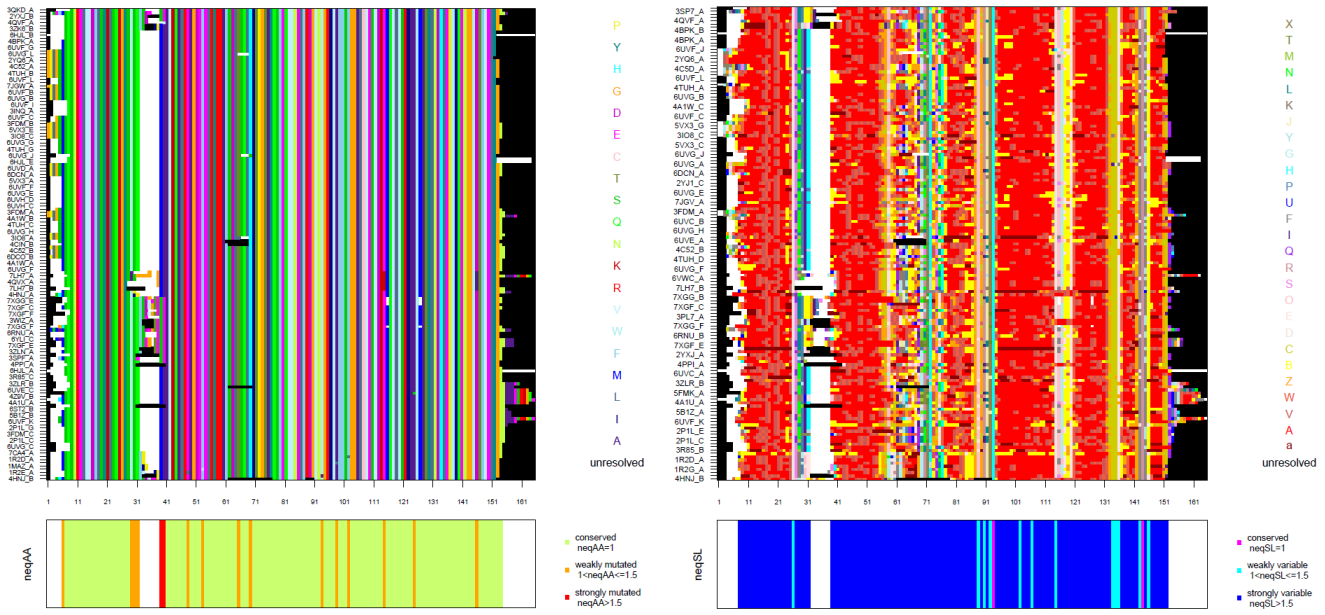
Figure 4: Graphical output showing the clusters for several conformations of the human p63. Left: the tree resulting from the hierarchical clustering. Each colour for the structure's names corresponds to a different cluster. Right: the heatmap of the MSLA, with the clusters ordered (see fig. 1b for the unordered result).

3.2 Application to Bcl-Xl

In order to study the variability of Bcl-Xl, we executed SA-conf on the 149 structures in the dataset.

Missing residues and uncertain coordinates

The first output, the MSA, allowed us to visualise the unresolved residues; they are mostly positioned on the C-ter end of the structures, as well as on the N-ter end. There seems to be a constant gap in most of the structures (from the positions 32 to 40 approximately, on the MSA), that isn't due to unresolved residues, but is probably consecutive to an insertion in some structures (fig. 5). There is no position for which more than 50% of the structures have uncertain coordinates.



(a) MSAand neq_{AA} *(b) MSLA and neq_{SL}*

Figure 5: Heatmap of the MSA with the neq_{AA} at the bottom (on the left) and heatmap of the MSLA with the neq_{SL} at the bottom (on the right) of Bcl-Xl.

Variable positions and residue flexibility

The AA positions in the MSA of Bcl-Xl are mostly conserved, with a $neq_{AA} > 1.5$ only for the positions 39 to 40 (out of 175 positions in total) (fig. 5a). In contrast, the positions in the MSLA seem highly variable in terms of SL, with a $neq_{SL} > 1.5$ for 123 out of 175 positions in total (89.8%) (fig. 5b). The variable region in the MSA doesn't seem to particularly coincide with a highly variable region in the MSLA (fig. 6). This variability in the SL sequences are associated with secondary structure changes, except in a few positions where alpha helices are conserved (fig. 6).

There is a segment of flexible residues (pictured in red), from position 55 to 75, which is visually identifiable because of its corresponding highly variable positions (fig. 6) and has an average neq_{SL} of 5.98. Most of this segment is composed of changing secondary structures, with constant loops at its end. This seems to correspond to a disordered region [10] that contains a loop required for the interaction with NLRP1, a nucleotide-binding oligomerisation domain-like receptor that plays a role in innate immunity [11]. Its flexibility and variable SL indicates a highly-adaptable region, that likely facilitates diverse protein-protein interactions, allowing Bcl-Xl to engage with multiple partners, such as NLRP1.

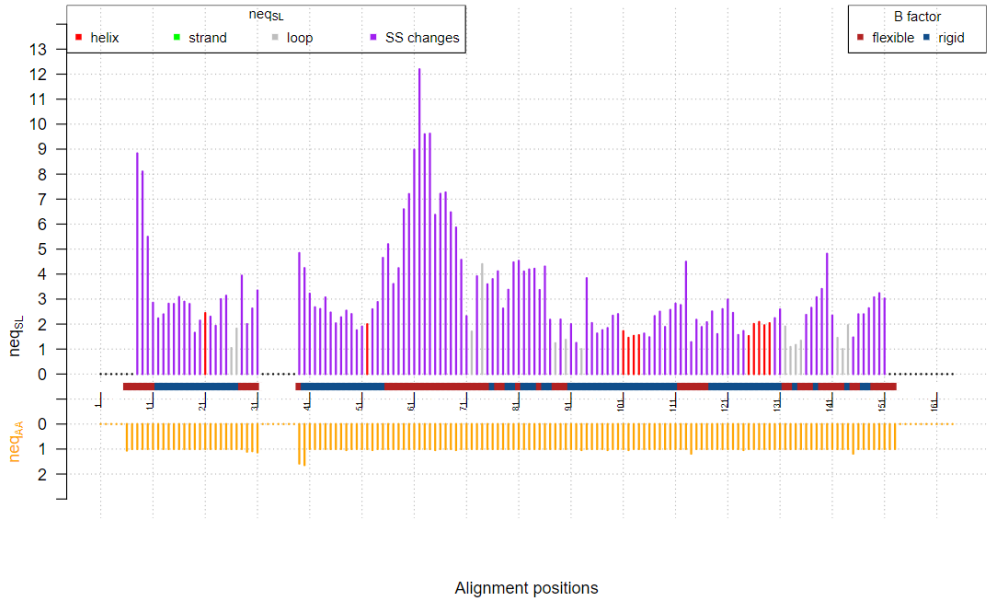


Figure 6: Graph of the neq_{SL} and neq_{AA} for Bcl-Xl

Clustering

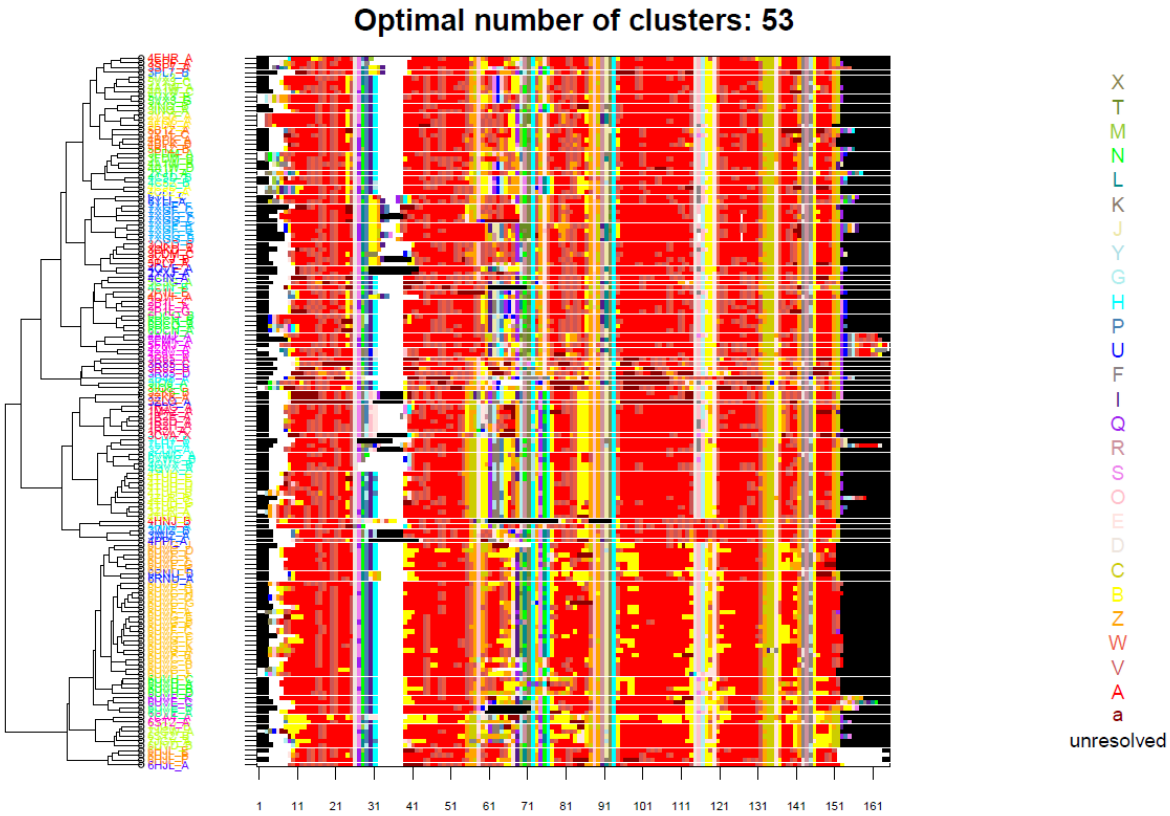
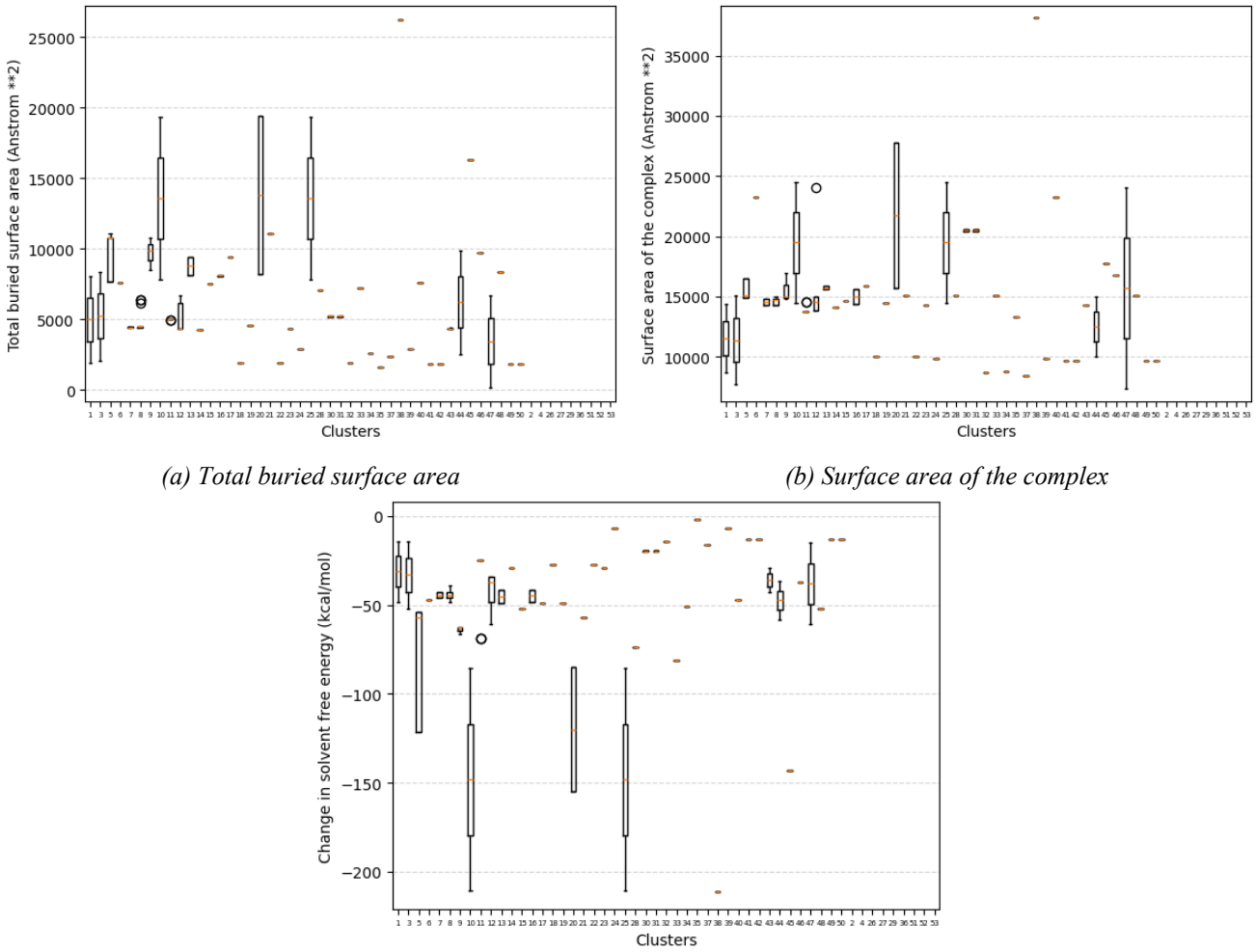


Figure 7: Graphical output showing the clusters for Bcl-Xl

Our clustering method returned 53 different clusters (fig. 7), of which 19 included only one structure. In order to explore those clusters, we compared the distribution of some physical properties of the structures included in each of them. A certain heterogeneity can be observed visually on the distributions of the thermodynamic and structural properties of the structures included in each cluster (fig. 8); this is promising regarding the fact that the clustering method based on the MSLA allows us to differentiate between the structures based on their conformations and interactions with their different partners.



(a) Total buried surface area

(b) Surface area of the complex

(c) Change in solvent free energy

Figure 8: Distribution of the physical properties the structures in each cluster computed from Bcl-Xl.

4 Conclusion and discussion

For this project, we successfully enhanced the SA-conf tool, expanding its capabilities for more detailed structural analysis of multiple target conformations. Key improvements included the differentiation between unresolved residues and alignment gaps, the identification of uncertain coordinates using the RSRZ score, and the integration of B-factor data to characterise residue flexibility. Additionally, clustering of structural data was implemented to group conformations based on structural variability, offering insights into protein conformational diversity.

Regarding the clustering, additional improvements could help fine-tune its capabilities to separate the data into more relevant classes. Firstly, the distance computation could do with a more precise approach: indeed, the Hamming distance does not take into account the overall distribution of letters within sequences, nor the similarity between certain letters in relation to others (letters designating the same secondary structure for example). It could be interesting to use a metric including these factors, which could for example be based on a substitution matrix. A machine learning algorithm could also be implemented for this, with the training data being MTC sets divided into subgroups according to relevant biological parameters, based on protein functions for example.

It could also be interesting to conduct a more in-depth analysis of the clusters obtained from the Bcl-Xl dataset with integration of other properties, such as the presence/absence of a partner and its nature, or the different mutations present in each cluster. Additionally, a statistical analysis of these results would allow us to conclude with certitude on the differentiation power of the clustering.

In conclusion, while the improvements made to the SA-conf tool mark significant progress in structural analysis and clustering of protein conformations, further refinement in clustering methodologies and deeper statistical investigations will be essential for enhancing its accuracy and biological relevance.

Bibliography

[1] Leslie Regad et al. “Exploring the potential of a structural alphabet-based tool for mining multiple target conformations and target flexibility insight”. In: *PLOS ONE* 12.8 (Aug. 17, 2017). Publisher: Public Library of Science, e0182972. issn: 1932-6203. doi: 10.1371/journal.pone.0182972. url: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182972 (visited on 11/19/2024).

[2] A. C. Camproux, R. Gautier, and P. Tufféry. “A hidden markov model derived structural alphabet for proteins”. In: *Journal of Molecular Biology* 339.3 (June 4, 2004), pp. 591–605. issn: 0022-2836. doi: 10.1016/j.jmb.2004.04.005.

[3] J. Marie Hardwick and Lucian Soane. “Multiple Functions of BCL-2 Family Proteins”. In: *Cold Spring Harbor Perspectives in Biology* 5.2 (Feb. 2013), a008722. issn: 1943-0264. doi: 10.1101/cshperspect.a008722. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3552500/ (visited on 01/07/2025).

[4] Nidhish Sasi et al. “Regulated cell death pathways: new twists in modulation of BCL2 family function”. In: *Molecular Cancer Therapeutics* 8.6 (June 2009), pp. 1421–1429. issn: 1538-8514. doi: 10.1158/1535-7163.MCT-08-0895.

[5] Sophie Pattingre et al. “Bcl-2 antiapoptotic proteins inhibit Beclin 1-dependent autophagy”. In: *Cell* 122.6 (Sept. 23, 2005), pp. 927–939. issn: 0092-8674. doi: 10.1016/j.cell.2005.07.002.

[6] Mingxue Li et al. “Bcl-XL: A multifunctional anti-apoptotic protein”. In: *Pharmacological Research* 151 (Jan. 1, 2020), p. 104547. issn: 1043-6618. doi: 10.1016/j.phrs.2019.104547. url: https://www.sciencedirect.com/science/article/pii/S1043661819314847 (visited on 01/07/2025).

[7] G. J. Kleywegt et al. “The Uppsala Electron-Density Server”. In: *Acta Crystallographica Section D: Biological Crystallography* 60.12 (Dec. 1, 2004). Publisher: International Union of Crystallography, pp. 2240–2249. issn: 0907-4449. doi: 10.1107/S0907444904013253. url: //journals.iucr.org/paper?ba5061 (visited on 01/07/2025).

[8] D. E. Tronrud. “Knowledge-Based B-Factor Restraints for the Refinement of Proteins”. In: *Journal of Applied Crystallography* 29.2 (Apr. 1, 1996). Publisher: International Union of Crystallography, pp. 100–104. issn: 0021-8898. doi: 10.1107/S002188989501421X. url: https://journals.iucr.org/j/issues/1996/02/00/wb0026/ (visited on 01/07/2025).

[9] Dhoha Triki et al. “Characterizing the structural variability of HIV-2 protease upon the binding of diverse ligands using a structural alphabet approach”. In: *Journal of Biomolecular Structure & Dynamics* 37.17 (Oct. 2019), pp. 4658–4670. issn: 1538-0254. doi: 10.1080/07391102.2018.1562985.

[10] *BCL2L1 - Bcl-2-like protein 1 - Homo sapiens (Human) | UniProtKB | UniProt*. url: https://www.uniprot.org/uniprotkb/Q07817/entry#family_and_domains (visited on 01/11/2025).

[11] Jean-Marie Bruey et al. “Bcl-2 and Bcl-XL Regulate Proinflammatory Caspase-1 Activation by Interaction with NALP1”. In: *Cell* 129.1 (Apr. 6, 2007). Publisher: Elsevier, pp. 45–56. issn: 0092-8674, 1097-4172. doi: 10.1016/j.cell.2007.01.045. url: https://www.cell.com/cell/abstract/S0092-8674(07)00304-2 (visited on 01/11/2025).