Fiscore Package: Effective Protein Structural Data Visualisation and Exploration

Auste Kanapeckaite

Abstract Lack of bioinformatics tools to quickly assess protein conformational and topological features motivated to create an integrative and user-friendly R package. Moreover, *Fiscore* package implements a pipeline for Gaussian mixture modelling making such machine learning techniques readily accessible to non-experts. This is especially important since probabilistic machine learning techniques can help with a better interpretation of complex biological phenomena when it is necessary to elucidate various structural features that might play a role in protein function. Thus, *Fiscore* package builds on the mathematical formulation of protein physicochemical properties that can aid in drug discovery, target evaluation, or relational database building. Moreover, the package provides interactive environments to explore various features of interest. Finally, one of the goals of this package was to engage structural bioinformaticians and develop more R tools that could help researchers not necessarily specialising in this field. Package *Fiscore* (v.0.1.2) is distributed via CRAN and Github.

Introduction

Fiscore R package was developed to quickly take advantage of protein topology/conformational feature assessment and perform various analyses allowing a seamless integration into relational databases as well as machine learning pipelines (Kanapeckaitė et al., 2021). The package builds on protein structure and topology studies which led to the derivation of the Fi-score equation capturing protein dihedral angle and B-factor influence on amino acid residues (Kanapeckaitė et al., 2021). The introduced tools can be very beneficial in rational therapeutics development where successful engineering of biologics, such as antibodies, relies on the characterisation of potential binding or contact sites on target proteins (Kanapeckaitė et al., 2021; Du et al., 2020). Moreover, translating structural data features into scores can help with target classification, target-ligand information storage, or integration into machine learning pipelines (Kanapeckaitė et al., 2021; Du et al., 2020). As a result, Fi-score, a first-of-its-kind *in silico* protein fingerprinting approach, created a premise for the development of a specialised R package to assist with protein studies and new therapeutics development (Kanapeckaitė et al., 2021).

Fiscore package allows capturing dihedral angles and B-factor effects on the protein topology and conformation. Since these physicochemical characteristics could help with the identification or characterisation of a binding pocket or any other therapeutically relevant site, it is important to extract and combine data from structural files to allow such information integration (Kanapeckaitė et al., 2021; Fauman et al.; Faraggi et al., 2009). Protein dihedral angles were selected as they contain information on the local and global protein conformation where protein backbone conformation can be highly accurately recreated based on the dihedral angles (Kanapeckaitė et al., 2021; Faraggi et al., 2009). Moreover, since Ramachandran plot, which provides a visualisation for dihedral angle distributions, namely ϕ (phi) and ψ (psi), allows only a holistic description of conformation and cannot be integrated with traditional parametric or non-parametric density estimation methods, a specific transformation was required to use this data. An additional parameter, specifically the oscillation amplitudes of the atoms around their equilibrium positions (B-factors) in the crystal structures, was also used. B-factors contain a lot of information on the overall structure; for example, these parameters depend on conformational disorder, thermal motion paths, associations with the rotameric state of amino acids side-chains, and they also show dependence on the three-dimensional structure as well as protein flexibility (Kanapeckaitė et al., 2021; Faraggi et al., 2009). Normalised dihedral angles (standard deviation scaling to account for variability) and scaled B-factors (min-max scaling) (Eq.1) were integrated into the Fi-score equation (Eq.2). It is important to highlight that B-factors need to be scaled so that different structural files can be compared and the dihedral angle normalisation transforms angular data into adjusted values based on the overall variability (Kanapeckaitė et al., 2021).

$$B_{norm} = \frac{B_i - B_{min}}{B_{max} - B_{min}}$$

Equation 1. Min-max normalisation and scaling of B-factor where Bi-norm is scaled B-factor, Bi - B-factor for C_α, Bmax - largest B-factor value for the total protein B-factors for all C_α, Bmin - smallest B-factor value for the total protein B-factors for all C_α. B-factor normalisation is based on the full length protein.

$$Fi_{score} = \frac{1}{N}\sum_i \frac{\phi_i \psi_i}{\sigma_{\phi_i}\sigma_{\psi_i}} B_{norm}$$

Equation 2.Fi-score evaluation where N- is the total number of atoms for which dihedral angle information is available, ϕ and ψ values represent dihedral angles for an C_α atom, σ_{ϕ_i} and σ_{ψ_i} represent corresponding standard deviations for the torsion angles and Bi-norm is a normalised B- factor value for the C_α atom. B-factor, σ_{ϕ_i} and σ_{ψ_i} normalisation is based on the full length protein.

In order to identify meaningful clusters based on the structural complexity, Gaussian mixture models (GMM) were selected as a primary machine learning classifier (Kanapeckaitė et al., 2021). The strength of GMM lies in the probabilistic model nature since all data points are assumed to be derived from a mixture of a finite number of Gaussian distributions with unknown parameters (Kanapeckaitė et al., 2021; Reynolds, 2015). Consequently, the soft classification of GMM where a data point has a probability of belonging to a cluster is much more suitable to assess biological parameters compared to other hard classification techniques in machine learning, such as k-means, which provide only a strict separation between classes. GMM pipeline offers a number of benefits to categorise protein structural features and the information can be used to explore amino acid grouping based on their physicochemical parameters. The designed GMM implementation takes care of information criterion assessment to fine tune the number of clusters for modelling and predicts the best suited model for the expectation-maximisation (EM) algorithm to maximise the likelihood of data point assignments (Kanapeckaitė et al., 2021; Reynolds, 2015).

As a case example, Nur77 protein was used to demonstrate various package functionalities. Nuclear receptor subfamily 4 group A member 1 (NR4A1; also termed Nur77/TR3/NGFIB), is a member of the nuclear receptor superfamily and regulates the expression of multiple target genes (Wu and Chen, 2018). This nuclear receptor is classified as an orphan receptor since there is no known endogenous ligand. Nur77 has the typical structure of a nuclear receptor which consists of an N-terminal domain, a DNA binding domain, and a ligand-binding domain. This regulatory protein plays many potentially therapeutically relevant roles regulating cell proliferation and apoptosis (Wu and Chen, 2018). Consequently, the Nur77 protein is an excellent example to highlight how in-depth structural analysis and classification could be beneficial in better understanding protein functions and finding potentially druggable or ligand binding sites.

Based on the need to develop integratable and specialised tools for protein analyses, a specialised *Fiscore* package was developed to assist with a wide spectrum of research questions ranging from exploratory analyses to therapeutic target assessment 1. The introduced set of new tools provides an interactive exploration of targets with an easy integration into downstream analyses. Importantly, the package and associated tools are written to be easy to use and facilitate analyses for non-specialists in structural biology or machine learning.

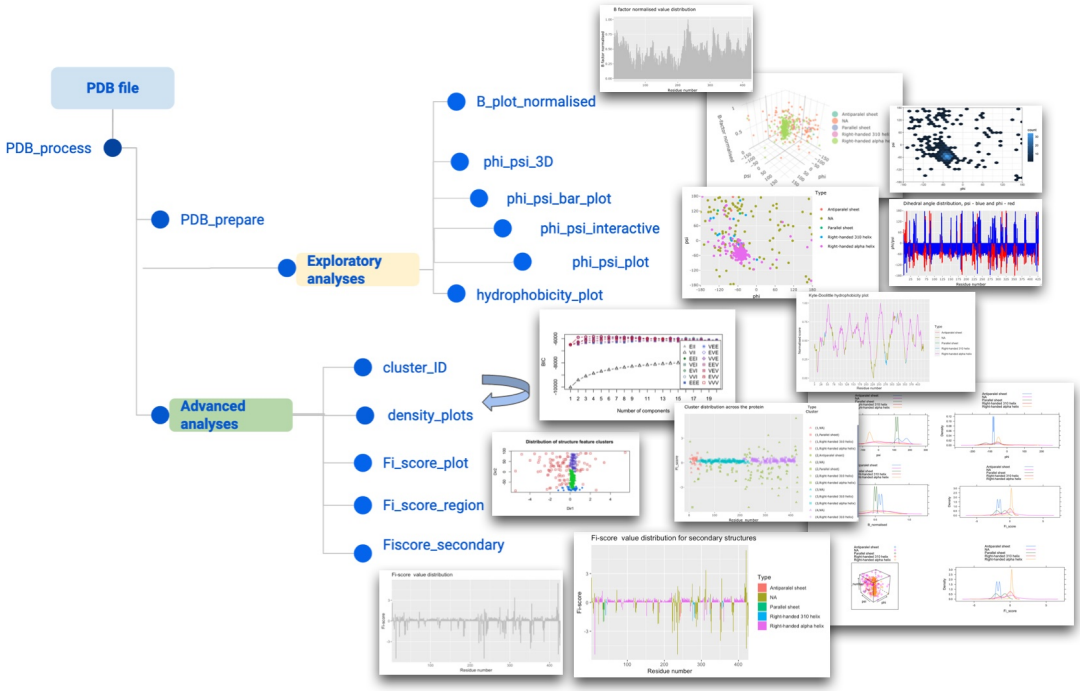

Figure 1: Schematic visualisation of the package features.

Methods

Fiscore package architecture is divided into exploratory and advanced functions 1 with a detailed vignette to guide the user (Kanapeckaite, 2021). Several key packages, such as ggplot2 (Wickham, 2016), Bio3D (B.J. et al., 2006), plotly (Sievert, 2020) and mclust (Scrucca et al., 2016), are also employed to create an easy-to-use analytical environment where a user-friendly machine learning pipeline of GMM (Kanapeckaitė et al., 2021) allows for a robust structural analysis. GMM implementation is designed to include the optimal cluster number evaluation (Bayesian information criterion), automatic model fitting in the EM phase of clustering, model-based hierarchical clustering, density estimation, as well as discriminant analysis (Kanapeckaitė et al., 2021; Kanapeckaite, 2021). Researchers also have an option to perform advanced exploratory studies or integrate the package into their development pipelines. *Fiscore* also takes care of raw data pre-processing and evaluation with optional settings to adjust how the analyses are performed. The package was built using functional programming principles with several R S3 methods to create objects for PDB files (Chambers, 2014). *Fiscore* is accompanied by documentation and vignette files to help users with their analyses (Kanapeckaite, 2021). Since PDB files are typically large, the documentation provides a compressed testing environment as well as a detailed tutorial. Additional visualisations provided with PyMol (DeLano, 2021). Nur77 protein was retrieved from the Protein Data Bank database (ID: 6KZ5) (PDB, 2021).

Data preparation

The workflow begins with the PDB file preprocessing and preparation. The user should also generally assess if the the structure is suitable for analysis, i.e. has a good resolution, no or a minimal number of breakages. Function *PDB_process* takes a PDB file name which can be expressed as '6KZ5.pdb' or 'path/to/the/file/6KZ5.pdb'. One of the function's dependencies is package Bio3D (B.J. et al., 2006), this useful package provides many tools to begin any PDB file analysis. *PDB_process* function can also take a 'path' parameter. Path can point to a directory where to split PDB into separate chain files (necessary for the downstream analysis) or you can leave this option to create a folder in your working directory by default. Function *PDB_process* returns a list of split chain names. If you split multiple PDB files in a loop it will be continuously added to the same folder. It is important to highlight that for the downstream processing PDB files need to be split so that separate chains can be analysed independently.

After the file or files are preprocessed Function *PDB_prepare* can be used to prepare a PDB file to generate Fi-score and normalised B-factor values as well as secondary structure designations. The function takes a PDB file name that was split into chains, e.g.'6KZ5_A.pdb' where a letter designates a split chain. The file is then cleaned and only the complete entries for amino acids are kept for the analysis, e.g., amino acids from the terminal residues that do not contain both dihedral angles are removed. The function returns a data frame with protein secondary structure designation *Type*, Fi- score values *Fi_score*, as well as normalised B-factor values *B_normalised* 2. Extracting protein secondary structure element information, i.e. *Type*, helps to prepare a file with all the relevant information about a target and this information could be supplied into cheminformatics or other bioinformatics pipelines where structural features are important to assess protein sites and amino acid composition.

```
##                 phi       psi       chi1      chi2 chi3 chi4 chi5 df_resno
##  31.A.ALA  54.94701  53.80667         NA        NA   NA   NA   NA       31
##  32.A.ASN -60.91976 -18.01379 -157.93838 -76.10748   NA   NA   NA       32
##  33.A.LEU -64.18792 -45.94048  176.74103  46.17107   NA   NA   NA       33
##  34.A.LEU -65.44501 -38.46630  -88.61643 158.49518   NA   NA   NA       34
##  35.A.THR -67.68541 -43.43184   75.18019        NA   NA   NA   NA       35
##  36.A.SER -76.88914 -17.40880  157.45159        NA   NA   NA   NA       36
##           df_res B_factor B_normalised   Fi_score                     Type
##  31.A.ALA    ALA    30.07   0.35506724 0.37663226 Right-handed alpha helix
##  32.A.ASN    ASN    15.68   0.18227666 0.07176640 Right-handed alpha helix
##  33.A.LEU    LEU     7.79   0.08753602 0.09261096 Right-handed alpha helix
##  34.A.LEU    LEU     9.85   0.11227185 0.10140389 Right-handed alpha helix
##  35.A.THR    THR    20.79   0.24363593 0.25696344 Right-handed alpha helix
##  36.A.SER    SER    23.69   0.27845821 0.13372748 Right-handed alpha helix
```

Figure 2: PDB file processing output.

It is important to comment on the use of dihedral angles, B-factor, and the derivation of Fi-score. To capture protein structural and topological features that could help with the identification or characterisation of a binding pocket or any other relevant site, it is necessary to extract and combine

data from structural files to allow such information integration (Kanapeckaitė et al., 2021; Fauman et al.; Faraggi et al., 2009). Protein dihedral angles were used as they contain information on the local and global protein conformation where protein backbone conformation can be highly accurately recreated based on the dihedral angles (Kanapeckaitė et al., 2021; Fauman et al.; Faraggi et al., 2009). Similarly, B-factors provide a lot of information on the overall structure and can be used to infer conformational disorder, the rotameric state of amino acids side-chains, and protein flexibility (Kanapeckaitė et al., 2021; Fauman et al.; Faraggi et al., 2009). In combination, these physicochemical parameters are integrated into the Fi-score (Eq.2) (Kanapeckaitė et al., 2021) for further analyses.

Function calls are simple and user-friendly:

```
#general function for preprocessing raw PDB files

pdb_df<-PDB_process(pdb_path)

#cleaning and prepration of PDB file

pdb_df<-PDB_prepare(pdb_path)

#explore the output

head(pdb_df)

#the package allows to call test data directly for Nur77 example file

pdb_path<- system.file("extdata", "6kz5.pdb", package="Fiscore")
```

Exploratory analyses

The scope of the exploratory analyses can be split into principal and advanced analyses 1 where general exploration involves physicochemical parameters, such as dihedral angles, B-factors or hydrophobicity scores, and their distribution analyses.

Basic analyses are accessed through simple function calls to explore how dihedral angles and B-factors are distributed. These analyses offer interactive and easy visualisations of key parameters that are currently not offered in any other package.

```
#calling a Ramachandran plot function

phi_psi_plot(pdb_df)

#Visualisation of dihedral angle juxtaposed distributions

phi_psi_bar_plot(pdb_df)

#B plot value visualisation

B_plot_normalised(pdb_df)

#Interactive plot to map amino acids in 2D distribution to precisely see what parameters an individual amino

phi_psi_interactive(pdb_df)

#3D visualisation of dihedral angles and B-factor values

phi_psi_3D(pdb_df)
```

Especially useful functionality is hydrophobicity visualisation with the superimposed secondary structure elements. To the author knowledge, there are currently no tools implementing such a visualisation 3. The package provides an easy to use wrapper:

```
hydrophobicity_plot(pdb_df,window = 9,weight = 25,model = "linear")

#alternatively an exponential model can be selected

hydrophobicity_plot(pdb_df,window = 9,weight = 25,model = "exponential")
```

Nur77 protein contains a relatively large number of helices with majority showing a hydrophobic profile, i.e., the larger the score, the more hydrophobic the region. Some likely disordered regions can be seen spanning 50-70 amino acids 4. Another interesting region is around the 131 amino acid, since these amino acids undergo significant shifts in their hydrophilicity and hydrophobicity. Thus, studying these regions could provide hints at potential functional sites or sites of structural importance 3 4.

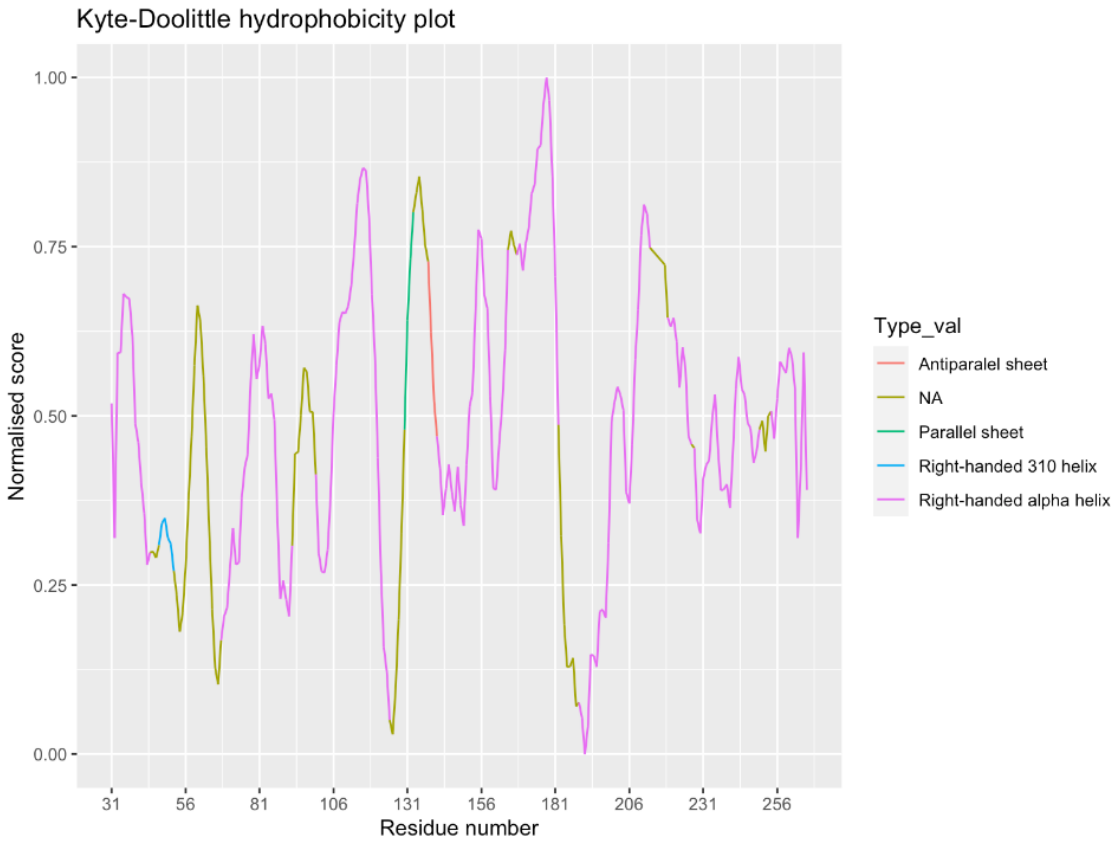

Figure 3: Hydrophobicity with secondary structure superimposition plot.

The function builds on Kyte-Doolittle hydrophobicity scale (Kanapeckaitė et al., 2021; Kyte and Doolittle, 1982) to detect hydrophobic regions in proteins. Regions with a positive value are hydrophobic and those with negative values are hydrophilic. This scale can be used to identify both surface-exposed regions as well as transmembrane regions, depending on the window size used. However, to make comparisons easier, the original scale is transformed from 0 to 1. The function requires a PDB data frame generated by PDB_prepare and the user needs to specify a 'Window' parameter to determine the size of a window for hydrophobicity calculations. The selection must include odd numbers between 3 and 21 with default being 21.' Weight' is another parameter that needs to be supplied to the function to establish a relative weight of the window edges compared to the window center (%); default=100%. Finally, a 'Model' parameter provides an option for weight caluclation; that is, the selection determines whether weights are calculated linearly $y = k \cdot x + b$ or exponentially $y = a \cdot b^x$; default setting is "linear". The function evaluates each amino acid in a selected window where a hydrophobicity influence from surrounding amino acids is weighed in. Thus, the terminal amino acids that cannot be included into the window for centering and weighing are assigned unweighted values based on the Kyte-Doolittle scale (Kyte and Doolittle, 1982). The plot values are all scaled from 0 to 1 so that different proteins can be compared without the need to convert or other adjustments.

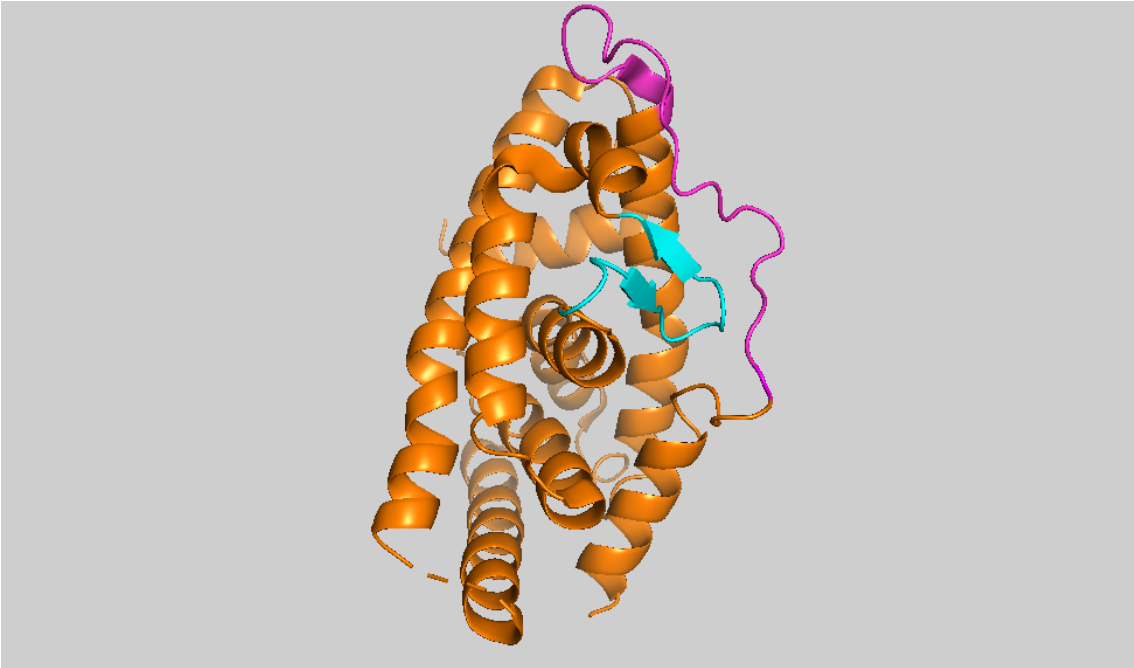

Figure 4: Nur77 protein where magenta highlights are used to define a likely disordered region between 50 and 70 amino acids and cyan indicates a region between 127 and 140 amino acids.

Thus, the hydrophobicity analysis can be especially useful when preparing to engineer proteins for various expression systems as the superimposition of structural features and hydrophobicity scores can help with deciding if a protein region or domain is likely to be solvent exposed or prefer hydrophobic environments. For example, assessing the N or C terminal amino acid hydrophobicity and structural milieu can help in deciding which terminal should be tagged. Moreover, this tool could be broadly applied in drug discovery for explorations involving protein-protein interactions, protein-nucleic acid interactions, and membrane association events.

Advanced analyses

Advanced analyses provide an opportunity to evaluate Fi-score distributions and take advantage of a streamlined GMM pipeline. The development of this pipeline was motivated by making the functions and valuable data modelling accessible to non-experts. More advanced users can supply custom parameters for GMM and extract probabilities from the output to use scores in other analyses or integrate the values in their own discovery pipelines.

```
#Fi-score distribution plot to explore scores for corresponding amino acids
Fi_score_plot(pdb_df)

#Fi-score for a selected region, this value for multiple sites can be stored in relational databases

Fi_score_region(pdb_df,50,70)

#Plot with Fi-score values with superimposed secondary structures
Fiscore_secondary(pdb_df)
```

Fi-score captures several interesting regions in Nur77 and an especially intriguing site is around the 50 amino acid 5. Some other regions are also picked up which should be studied in more detail based on the amino acid composition and 3D conformations.

Function *cluster_ID* can be used to group structural features using the Fi-score via Gaussian mixture modelling since this tool will also automatically select an optimal number of clusters and a model to be fitted during the EM phase of clustering for GMM 6. The output of the function summarises cluster information and also provides plots to visualise the identified clusters based on the Fi-score 7.

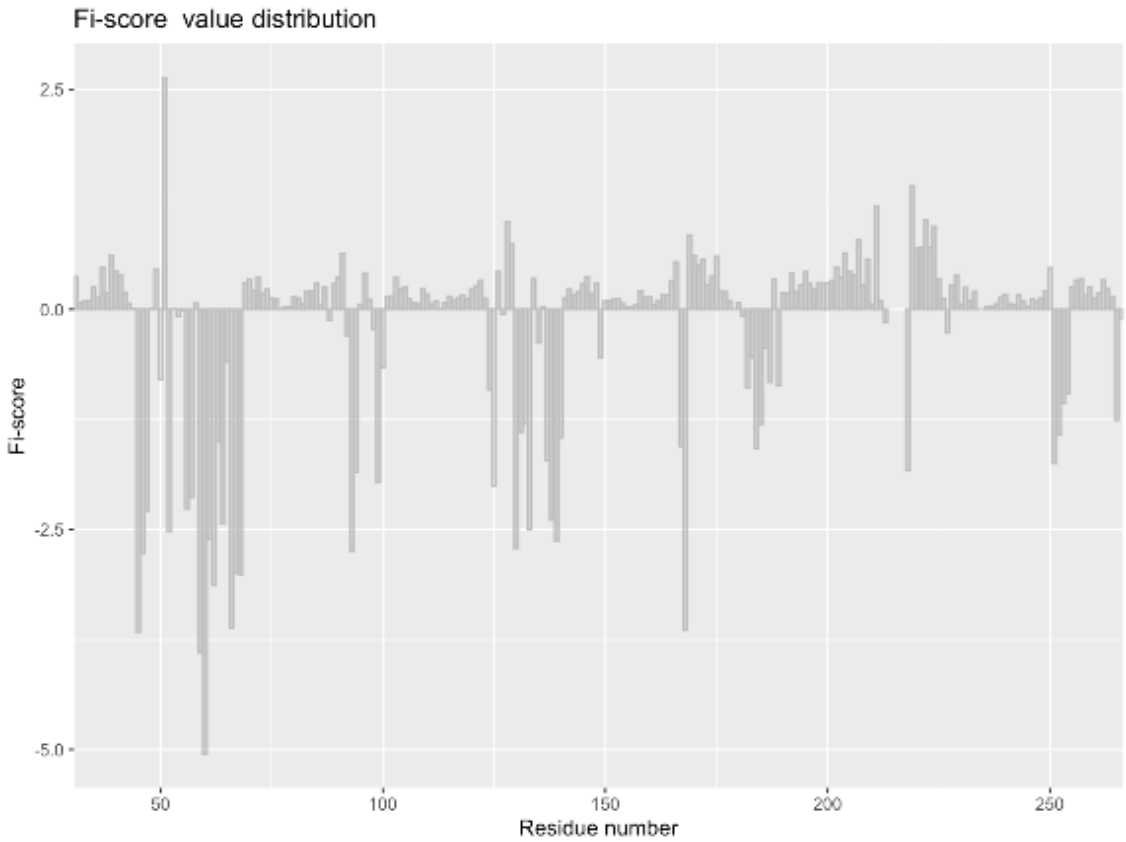

Figure 5: Fi-score distribution.

```
df<-cluster_ID(pdb_df)

#user selected parameters

df<-cluster_ID(pdb_df,clusters = 5, modelNames = "VVI")
```

The users are advised to set seed for projects to reproduce the results. *cluster_ID* takes a data frame containing a processed PDB file with Fi-score values as well as a number of clusters to consider during model selection; by default 20 clusters ("max_range") are considered. In addition, a "secondary_structures" parameter is needed to define whether to include the information on secondary structure elements from the PDB file when plotting; default value is TRUE. Researchers also have an option to select a cluster number to test "clusters" together with "modelNames". However, it is important to stress that both optional entries need to be selected and defined, if the user wants to test other clustering options that were not provided by the automated Bayesian information criterion (BIC) output. This is an advanced option and the user should assess the BIC output to decide which model and what cluster number he or she wants to try out.

A dimension reduction method for visualising the clustering or classification structure is also automatically provided 8. Dimension reduction is a useful technique to explore multidimensional biological data through the key eigenvalues that define the largest information content of the explored features (Scrucca et al., 2016). In other words, one can infer how well the explored characteristics define the data and if the classification is sufficient for downstream analyses. For example, in the case of Fi-score clustering, this analysis allows assessing whether the number of clusters separates data points based on their distribution features. Nur77 has six distinct clusters based on indvidual distribution profiles which might indicate functionally and structurally similar regions in the target protein. It appears that the data points are separated into groups well accounting for the different variability.

However, one of the most valuable features of this set of functions is to generate clusters based on amino acids by providing information with an option to also include secondary structure information 9 10.The produced interactive plots enable researchers to explore structural characteristics of a protein of interest 9 10. Protein structure subdivision based on the physicochemical features offers a

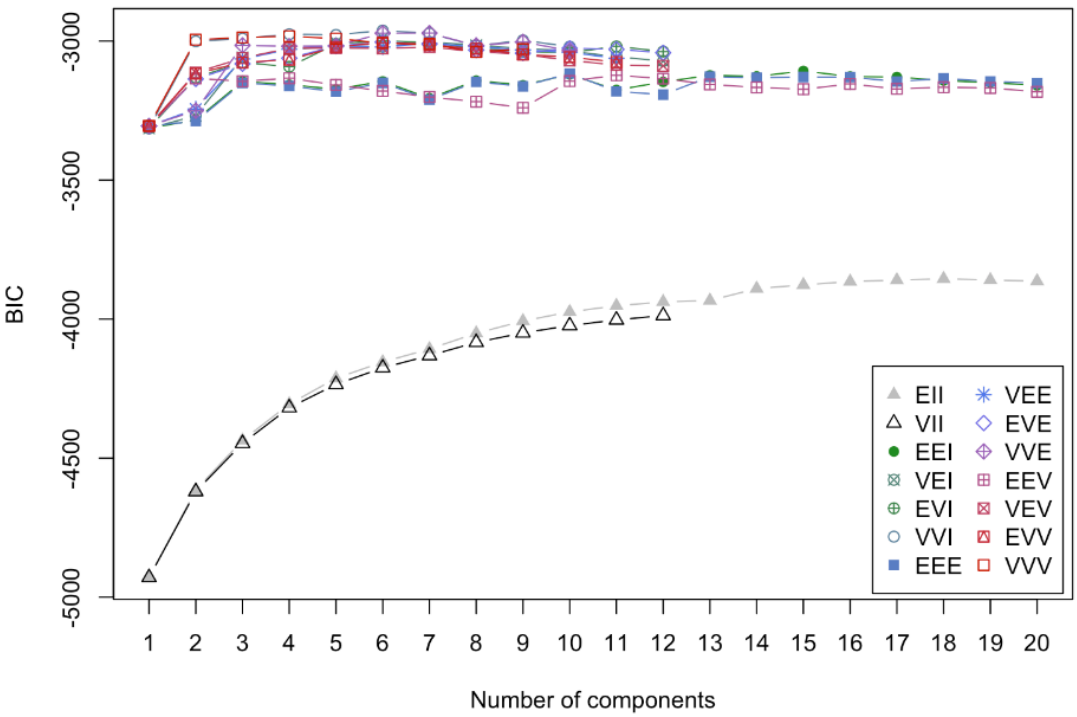

Figure 6: Gaussian mixture modelling output showing Bayesian information criterion evaluation.

```
## Best BIC values:
##          VVI,6        VVE,7        VVE,6
## BIC      -2962 -2971.310677 -2971.756243
## BIC diff     0    -9.310508    -9.756073
## ----------------------------------------------------------------
## Dimension reduction for model-based clustering and classification
## ----------------------------------------------------------------
##
## Mixture model type: Mclust (VVI, 6)
##
## Clusters  n
##        1 15
##        2 16
##        3 41
##        4 87
##        5 39
##        6 34
##
## Estimated basis vectors:
##                      Dir1      Dir2
## Residue_number 0.00047634  0.056189
## Fi_score       0.99999989 -0.998420
##
##                 Dir1      Dir2
## Eigenvalues  0.74861   0.55583
## Cum. %      57.38945 100.00000
```

Figure 7: Output table for Gaussian mixture modelling evaluation.

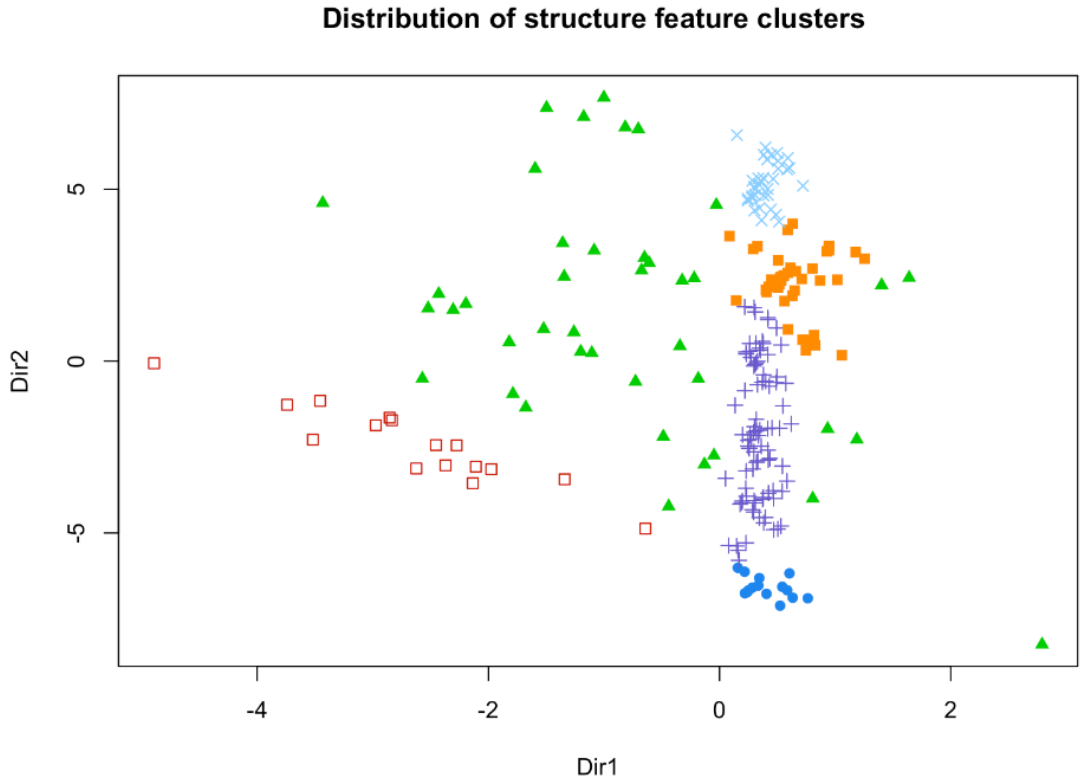

Figure 8: Dimension reduction plot for identified clusters.

new way to detect novel functional sites or protein regions important for structural stability. Figures 9 and 10 clearly indicate that some structural elements in Nur77 are likely similar in their function and physicochemical characteristics, but the regions around the 50 amino acid as well as N and C terminus are distinct and might be sites where the receptor mediates specific functions, such as binding of co-activators or repressors (Wu and Chen, 2018).

All previous analyses tie in with a function *density_plots* which provides a density plot set for ϕ/ψ angle distributions, Fi-score, and normalised B-factor. 3D visualisation of dihedral angle distribution for every residue is also included. The plots can be used for a quick assessment of the overall parameters as well as to summarise observations. Density plots are optimal to evaluate how well the selected features or scores separate protein structural elements and if, for example, a protein structure is of good quality (i.e., dihedral angles, B-factors, or Fi-scores capture distinct features). The function also gives another reference point to establish if the selected number of clusters separated amino acids well based on the secondary structure elements. In order to get this information, the user is only required to supply the output from *cluster_ID* function 11.

```
# data summary and evaluation

density_plots(pdb_df)


# data summary and evaluation including GMM outputs

cluster_IDs<-cluster_ID(pdb_df)

density_plots(cluster_IDs)
```

Figure 9: Protein cluster identification with secondary structure elements.

Figure 10: Protein cluster identification.

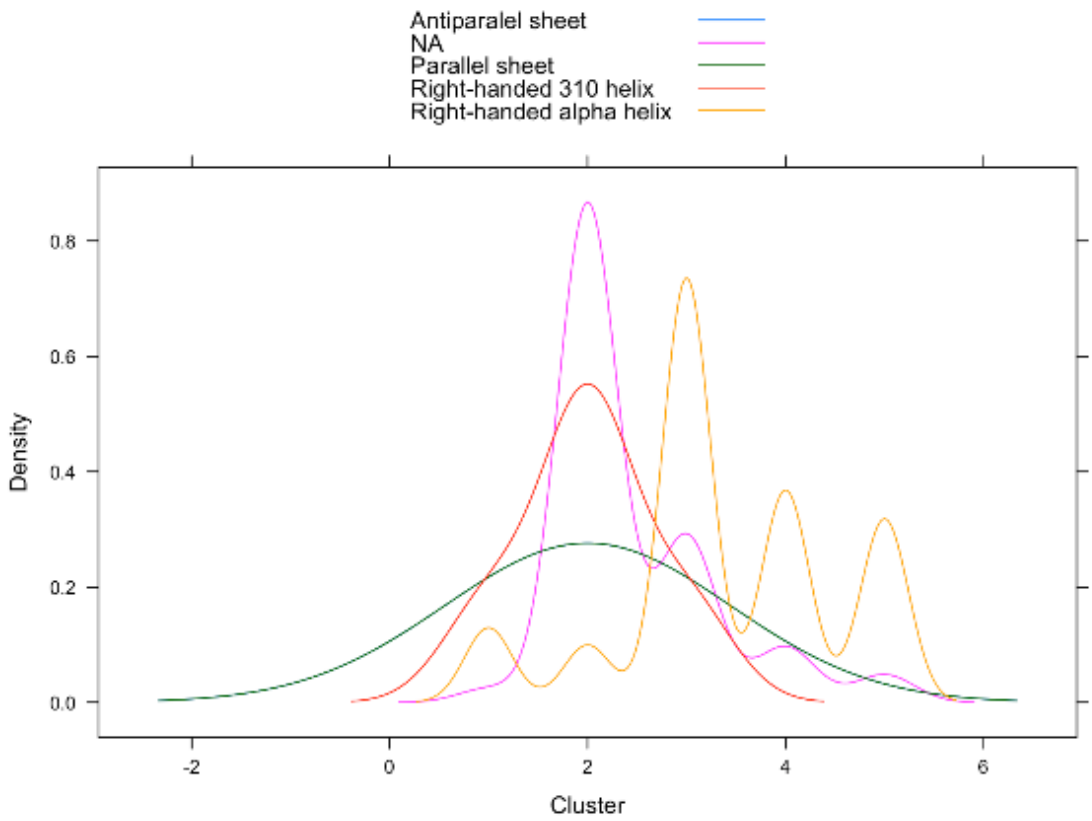

Figure 11: Protein cluster density plots.

Discussion

Fiscore package 1 allows a user-friendly exploration of PDB structural data and integration with machine learning approaches. The package was benchmarked through several analytical stages that involved a diverse set of proteins (3352) (Kanapeckaitė et al., 2021). With a number of helpful functions, including secondary feature distribution analyses or hydrophobicity assessment in the context of structural elements, *Fiscore* enables the exploration of new target families and comprehensive data integration since the described fingerprinting captures protein sequence and physicochemical properties. Such analyses could be very helpful in exploring therapeutically relevant proteins or in drug repurposing studies when a chemical compound needs to be juxtaposed to a number of potential targets. In addition, provided tutorials and documentation should guide researchers through their analysis and allow adapting the package based on individual research needs (Kanapeckaitė et al., 2021). Nur77 nuclear receptor used as a case study helped to demonstrate that it is very helpful to capture physicochemical data in visual representations and using a novel scoring system as well as machine learning can lead to interesting insights about sites of structural and functional importance. Another important aspect of *Fiscore* package is the simplification of complex analytical steps so that researchers without an extensive background in structural bioinformatics or machine learning could still use the tools for their analyses, such as protein engineering, protein assessment, and data storage based on specific target sites. Thus, the interactive analytical and visualisation tools could become especially relevant in the pharmaceutical industry and drug discovery as more complex targets and protein-protein interactions need to be assessed in a streamlined fashion.

Summary

Fiscore is an R package developed to quickly take advantage of protein topology and conformational feature assessment and perform various exploratory analyses or integrate derived scores into relational databases as well as machine learning pipelines. The package builds on protein structure and topology studies which led to the derivation of the Fi-score, or a mathematical equation, that captures dihedral angle and B-factor parameter influence on amino acid residues in a protein. Since target evaluation is paramount for rational therapeutics development, there is an undeniable need for specialised analytical

tools and techniques that can be used in RD or academic research. For example, to successfully engineer therapeutic antibodies, it is necessary to characterise potential binding or contact sites on target proteins. Moreover, translating structural data into parameters that can be used to classify targets, store targeligand information, or develop machine learning applications could significantly improve our ability to assess new targets and develop novel therapeutics. As a result, *Fiscore* package was developed to aid with therapeutic target assessment and make machine learning techniques more accessible to a wider scientific audience.

Acknowledgements

The author would like to thank the anonymous reviewers and code testers for helping to improve the package with their valuable suggestions and advice.

Funding

This package development was not supported by outside funding.

Bibliography

RCSB Protein Data Bank. RCSB Protein Data Bank, 2021. [p3]

G. B.J., R. A.P.C., E. K.M., M. J.A., and C. L.S.D. Bio3d: An r package for the comparative analysis of protein structures. *Bioinformatics*, 22:2695–2696, Nov 2006. [p3]

J. Chambers. Object-oriented programming, functional programming and r. *Statistical Science*, 29, 2014. doi: 10.1214/13-STS452. URL https://projecteuclid.org/journals/statistical-science/volume-29/issue-2/Object-Oriented-Programming-Functional-Programming-and-R/10.1214/13-STS452.full. [p3]

W. DeLano. *The PyMOL Molecular Graphics System, Version 2.3.0*. The PyMOL Molecular Graphics System,Schrödinger, LLC., 2021. [p3]

J. Du, J. Guo, D. Kand, Z. Li, G. Wang, J. Wu, and et al. New techniques and strategies in drug discovery. *Chinese Chemical Letters*, 31, 2020. doi: 10.1016/j.cclet.2020.03.028. URL https://www.sciencedirect.com/science/article/abs/pii/S1001841720301479. [p1]

E. Faraggi, B. Xue, and Y. Zhou. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins*, 74, 2009. doi: 10.1002/prot.22193. URL https://onlinelibrary.wiley.com/doi/10.1002/prot.22193. [p1, 4]

E. Fauman, B. Rai., and E. Huang. [p1, 4]

A. Kanapeckaite. *Fiscore: Effective Protein Structural Data Visualisation and Exploration*. 2021. URL https://CRAN.R-project.org/package=Fiscore. R package version 0.1.2. [p3]

A. Kanapeckaitė, C. Beaurivage, M. Hancock, and E. Verschueren. Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies. *Journal of Biomolecular Structure and Dynamics*, 2021. doi: 10.1080/07391102.2020.1854859. URL https://www.tandfonline.com/doi/abs/10.1080/07391102.2020.1854859. [p1, 2, 3, 4, 5, 11]

J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982. ISSN 0022-2836. doi: https://doi.org/10.1016/0022-2836(82)90515-0. URL https://www.sciencedirect.com/science/article/pii/0022283682905150. [p5]

D. Reynolds. *Gaussian Mixture Models*. Springer, Boston, 2015. doi: 10.1007/978-1-4899-7488-4. URL http://link.springer.com/10.1007/978-0-387-73003-5_196. [p2]

L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317, 2016. URL https://doi.org/10.32614/RJ-2016-021. [p3, 7]

C. Sievert. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC, 2020. ISBN 9781138331457. URL https://plotly-r.com. [p3]

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org. [p3]

L. Wu and L. Chen. Characteristics of nur77 and its ligands as potential anticancer compounds. *Molecular medicine reports*, 2018. doi: 10.3892/mmr.2018.9515. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6236262/. [p2, 9]

Auste Kanapeckaite
1. Algorithm379
2. University of Reading, School of Pharmacy
Hopkins Building, Reading RG6 6UB
United Kingdom
0000-0001-6829-4082
auste.kan@algorithm379.com