Nucleic Acids Research, 2024, 1–8
https://doi.org/10.1093/nar/gkae384
Web Server Issue

PROSCA: an online platform for humanized scaffold mining facilitating rational protein engineering

Xiaona Wang[1,2,†], Yintao Zhang[3,†], Zengpeng Li[4], Zixin Duan[1], Menghan Guo[1], Zhen Wang[2], Feng Zhu[3,*] and Weiwei Xue[1,*]

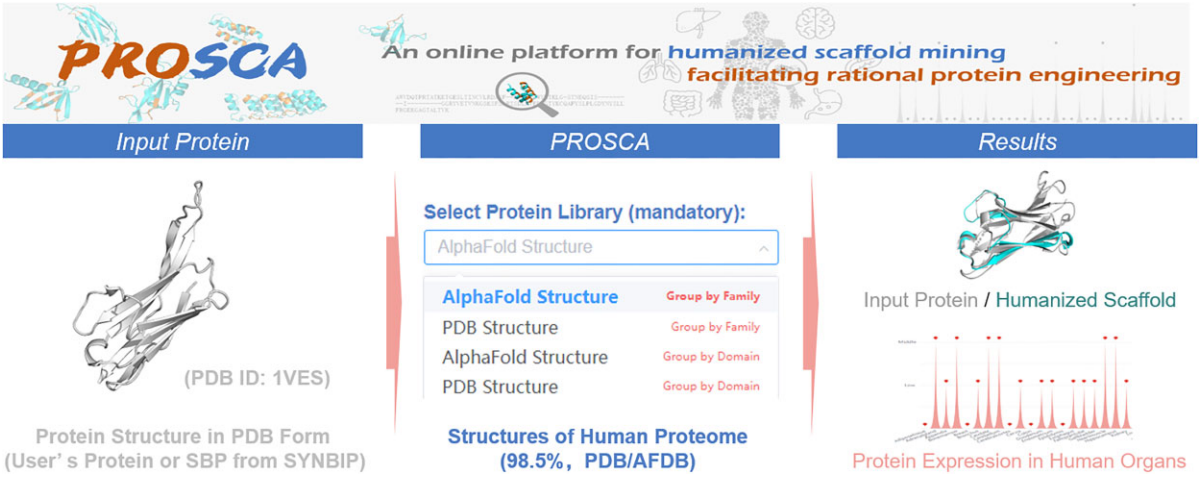
[1]Chongqing Key Laboratory of Natural Product Synthesis and Drug Research, School of Pharmaceutical Sciences, Chongqing University, Chongqing 401331, China
[2]Department of Intensive Care Medicine, Army Medical Center of PLA, Chongqing 401331, China
[3]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China
[4]State Key Laboratory Breeding Base of Marine Genetic Resources, Third Institute of Oceanography Ministry of Natural Resources, Xiamen 361005, China

[*]To whom correspondence should be addressed. Tel: +86 187 0236 4293; Fax: +86 023 6567 8450; Email: xueww@cqu.edu.cn
Correspondence may also be addressed to Feng Zhu. Email: zhufeng@zju.edu.cn
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Protein scaffolds with small size, high stability and low immunogenicity show important applications in the field of protein engineering and design. However, no relevant computational platform has been reported yet to mining such scaffolds with the desired properties from massive protein structures in human body. Here, we developed PROSCA, a structure-based online platform dedicated to explore the space of the entire human proteome, and to discovery new privileged protein scaffolds with potential engineering value that have never been noticed. PROSCA accepts structure of protein as an input, which can be subsequently aligned with a certain class of protein structures (e.g. the human proteome either from experientially resolved or AlphaFold2 predicted structures, and the human proteins belonging to specific families or domains), and outputs humanized protein scaffolds which are structurally similar with the input protein as well as other related important information such as families, sequences, structures and expression level in human tissues. Through PROSCA, the user can also get excellent experience in visualizations of protein structures and expression overviews, and download the figures and tables of results which can be customized according to the user's needs. Along with the advanced protein engineering and selection technologies, PROSCA will facilitate the rational design of new functional proteins with privileged scaffolds. PROSCA is freely available at https://idrblab.org/prosca/.

Graphical abstract



Introduction

Protein scaffolds with small size, high stability and low immunogenicity represent the next generation protein-based research, diagnostics, and therapeutics (1). In addition to conventional evolutionary or mechanically-themed methods (2,3), the development of new and effective approaches to mining such scaffolds from massively known 3D structures in RCSB Protein Data Bank (RCSB PDB) (4) and AlphaFold

Received: February 29, 2024. **Revised:** April 23, 2024. **Editorial Decision:** April 26, 2024. **Accepted:** April 29, 2024

Protein Structure Database (AlphaFold DB) (5) is extremely essential in the field of rational protein engineering (6) and design (7–9). However, no relevant server has been reported yet.

To fill this gap, we developed PROSCA, an online platform for humanized scaffold mining facilitating rational protein engineering. The server accepts any protein structure in PDB format as the input data. Through AFalign and (or) PDBalign algorithms (see Materials and methods section) based on TM-align method (10), the input structure will be automatically aligned with selected proteins from homo species. Then, the aligned data will be processed to give candidate proteins as output which have similar architectures with the input one. On the result pages, a table can be downloaded, and the detailed information of the target proteins will be displayed including protein families or domains, sequences, structures, and expression profiles in 45 human tissues (11). Therefore, our PROSCA is unique in (i) automated-identification of scaffolds with innate low immunogenicity from the space of the entire human proteome by using extended structural alignment algorithms; (ii) expression-level-based assessment of the candidate scaffolds' stability in human body for protein engineering; (iii) structural alignment can be conducted on a certain class of protein structures including the human proteome either from experientially resolved or AlphaFold2 predicted structures, and the human proteins belonging to specific families or domains and (iv) optional query of synthetic binding proteins (SBPs) is provided for users by integrating the structural information collected in our SYNBIP database (12).

To demonstrate the utility of PROSCA, the structure of a variable new antigen receptor (vNAR, a single domain antibody from shark) was inputted as an example, and resulting 307 protein fragments corresponding to 103 different proteins in human body. These proteins have low immunogenicity and designated structure features which could be selected as new scaffolds for protein engineering and design (13–15). Particularly, one of them is an Ig domain identified from human Neural Cell Adhesion Molecule 1 (hNCAM1) (16). The scaffold has been used for *i*-body drugs development to attenuate renal fibrosis (17), represented by AD-214 in phase I clinical trials (NCT04415671). Compared with classical biochemical techniques (17), PROSCA runs much more quickly and efficiently (about 5 h). Moreover, the rate of success is relatively higher due to the large space of available protein scaffolds. Therefore, PROSCA may has a potential to facilitate the development of new functional proteins in biomedical science.

Materials and methods

Dataset preparation

Protein 3D structures

PROSCA provides two classes of protein structures from homo sapiens (i.e. experimentally resolved structures and AlphaFold2 predicted protein structures). The details of how to obtain aforementioned structures were as follows: for AlphaFold2 predicted structures, the organism of 'Homo sapiens' was chosen as a filter to download the data in AlphaFold DB, resulting 2 391 human protein structures which cover 98.5% of human proteome. These structures have a median backbone accuracy of 0.96 Å Cα RMSD at 95% residue coverage (18). For experimentally resolved structures, data were collected through a two-step process: Initially, protein structures pertaining to the organism 'Homo sapiens' were selected from the RCSB PDB, yielding a total of 65 959 PDB files. Subsequently, a rigorous filtering process was undertaken to exclude engineered structures, repetitive structures with identical sequences, and those without accompanying PDB files. This filtering resulted in the retention of 40 129 structure files. The experimental structures from RCSB PDB, accounting for 32% of the human proteome, were resolved through X-ray diffraction, Cryo-EM, or NMR, which have relative high-quality (4). The two libraries were used as the basic datasets for structural alignment calculation in the server.

In addition, 2430 different domains were collected as the third structure library. The data processing was as follows: (i) these proteins were annotated with the UniprotKB domain information (19); (ii) 17 871 AlphaFold domian structures and 4489 PDB domain structures were extracted based on sequence similarity, sequence length, and position information of domains. For the same protein, the sequence from RCSB PDB and the one from AlphaFold DB may be different, and this is because some amino acid residues may be added or removed during the preparation of experiment resolved structure; (iii) to further identify the SCOP types of domain structures, we counted the numbers of secondary structure in each domain through labels (e.g. helix, sheet or strand) which already existed or were generated by PyMOL scripts in the structure files. All collected domain structures were divided into four classes to construct sub-datasets. The structure classes include 'All-Alpha Domain', 'All-Beta Domain', 'All-Loop Domain' and 'Alpha and Beta Domain'; and (iv) in SBP section, the query structures were also annotated with the SCOP classification for quick and precise identification of desired candidates.

Protein families

A total of 15 471 human proteins' family information was successfully retrieved from UniProtKB (19). Here, if the numbers of a specific protein family's members in both the AlphaFold and PDB datasets are <100, the corresponding protein will not be labeled separately. As a result, five most common protein families were separately labeled and were provided as sub-libraries for user to select. They are Dynein heavy chain family, G-protein coupled receptor superfamily, Immunoglobulin superfamily, Krueppel C2H2-type zinc-finger protein family, and Protein kinase superfamily.

Protein expression levels

The Human Protein Atlas (HPA) is a public database that provides tissue and cellular distribution information of human proteins (20). In this database, researchers used highly specific antibodies and detailed immunoassay techniques to detect the expression of human proteins in normal tissues and displayed their information maps. This protein expression profile reflects the most likely histological distribution and relative expression levels of each protein, which is generally considered to be related to *in vivo* stability and side effects (11). It is one of the important factors considered in the clinical translation process of proteins. Ultimately, we obtained 10 788 human proteins and their expression information in 45 tissues from the HPA database. Here, the protein expression levels were assigned values for visualization, and the values of 3, 2, 1 and 0 represent high-level expression, moderate expression, low-level expression, and no expression data (undetected or not expressed), respectively.

Data entry

The server provides three intuitive sectors to accept input data in PDB format, which are 'Submit a Protein Structure', 'Select a Synthetic Binding Protein' and 'Retrieve a History Alignment' (Figure 1). To figure out the type of input data, an example can be loaded.

First, in the panel of 'Submit a Protein Structure', the input can be structure data of any protein that users are interested in, but must contain coordinate information. These data can be accepted by the website through copying and pasting or uploading as a file.

Second, in the panel of 'Select a Synthetic Binding Protein', PROSCA realizes automatic connection with SYNBIP Database (12) which contains 1337 advantaged structures of SBPs corresponding to 56 scaffolds. Through the drop-box, the input protein is selected based on its name, scaffold or target and then the structure file will be uploaded into the server automatically without data preparation. Notably, these SBPs are also classified according to SCOP type and the related information is added after the protein names.

Third, in the panel of 'Retrieve a History Alignment', PROSCA accepts previous mission IDs which appear on the progress page or the user's email. If the user inputs a correct ID, the previous results will reappear without waiting for a long time (usually in a few seconds) and no new ID be generated. It should be noted that the IDs are valid within 10 days by default.

Library selection for structural alignment

PROSCA offers libraries of AlphaFold structures or PDB structures from human proteome for users to align with the input. Furthermore, the above two libraries are subdivided into multiple sub-libraries according to protein family and domain information. If the user selects one sub-library, the running time will be significantly reduced, for example a few minutes. PROSCA suggests to use the item 'ALL' for high hit ratio.

AFalign and PDBalign algorithms

The web server works basically by running two algorithms: AFalign and PDBalign. The algorithms are proposed based on TM-align algorithm, but implement more functions such as result filtering and complete structure extraction. AFalign runs the structural alignment of the import protein and predicted structures from AlphaFold DB (5). While PDBalign aligns the input structure with experimental structures from RCSB PDB (4). Therefore, PROSCA covers the space of the entire human proteome through the 'Predicted Structures' and 'Experimental Structures' below, and provides protein fragments structurally similar to the input (TM-score $\geq$ 0.5) and their detail information.

AFalign

The implementation of AFalign mainly consists of the following five steps: (i) automatically connecting the input protein with the human protein accurately predicted by the AlphaFold2 algorithm; (ii) calculating the structural similarity between two proteins; (iii) filtering out human proteins with TM score less than 0.5 that do not have structural similarity, and sorting them based on the degree of similarity; (iv) extracting the complete structural fragments of the alignment region of the human protein and (v) establishing connections with protein expression dataset and other relevant information, and then outputting the detailed information. In general, structural comparisons between input proteins and all human protein structures predicted by AlphaFold can be completed within one day. Emails with the job ID will be sent to the users' inbox when the job is submitted and finished in PROSCA. It is worth noting that the longer the protein length, the longer the running time.

PDBalgin

PDBalgin is proposed for the comparison of experimental structures and the input protein. The original structure file may embody multiple proteins from different species, or multiple chains of a protein. For example, PDBID 4MGI contains human Ras related protein Rap-1b and mouse Ras guanine nucleotide exchange factor 4, while PDBID 1A3N has A and B chains of human hemoglobin. PDBalign can exactly locate the target chain of human proteins. The algorithm firstly applies structural similarity methods to mine suitable scaffolds, and then uses sequence similarity methods to determine human protein and locate target chain, ultimately achieving accurate recognition and extraction of target fragments from human proteins. If the user chooses all experimental structures as the mining objects, PDBalgin will run, and the results will be generally returned within a few hours.

Results and discussion

Running progress

After submitting the mission, a progress bar shows how far the task has progressed. Above the bar, more detailed information including the mission ID, protein name and protein library is displayed and updated. Below the bar, a hyperlink of this progress page is showed. In addition, if the users close the running page and (or) don't inform PROSCA the email address, they can also get the calculation results through the ID number or the hyperlink.

Data output

The first result page contains two main parts: the uploaded protein information and the result table of calculation. In the first part, the sequence and structure are displayed for the query. And additional information ('Molecular Weight', 'Tm', 'Parent Scaffold', 'Binding Target', etc.) is also provided if the input is a molecule of SBPs. In the second part, PROSCA outputs a table containing a series of candidates of human protein fragments structurally similar to the query. The columns are 'Protein Name', 'Length of the Protein Fragments', 'Coverage', 'TM-score' and 'Maximum Expression Level in Human Tissues' of the candidates. The first four columns achieve the prior sorting functions by several clicks, and the data of expression in fifth column can be used for filtering. The protein name includes recommended name and location information (for example, Tim3.18 Fab Heavy Chain and 7KQL_Chain A). The coverage is the portion of structurally aligned amino acids divided by the sequence length of query protein. For the TM-score, the higher value demonstrates the higher similarity between the structures of candidate and query. Remarkably, the table can be downloaded and be used to create customized tables.

To further understand a candidate, users can click the button 'Detail' in the table and the second result page will appear to display the information. The family or domain, protein type and sequence length data are firstly introduced, while the

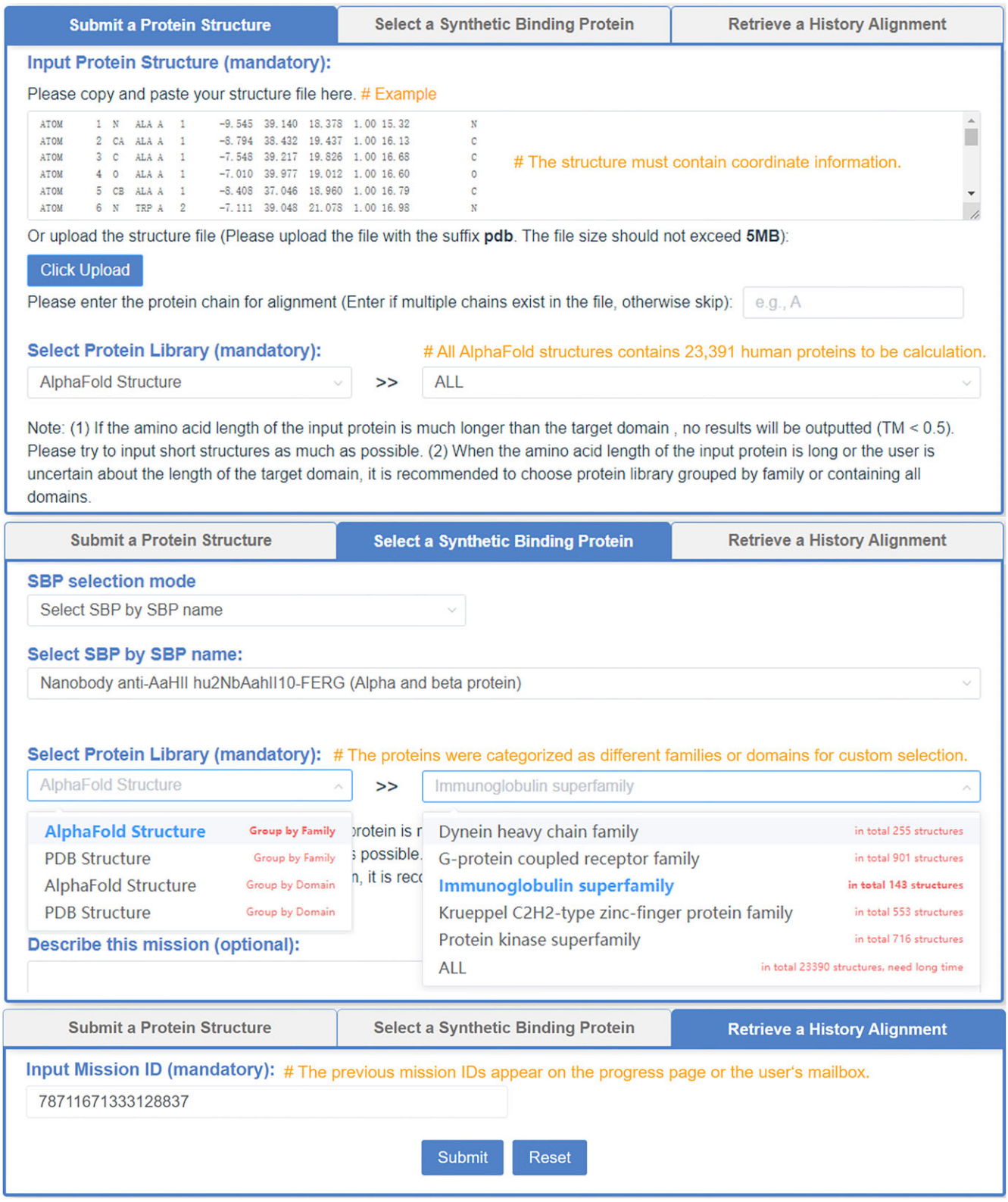


Figure 1. Three panels in PROSCA to receive input data. The website allows the user to import data through (i) copying and pasting the structure information, (ii) selecting a structure of SBP from SYNBIP database and (iii) filling in previous mission IDs.

Table 1. Comparison of PROSCA with other related tools

Tools/methods	Purpose	Features
TM-align/CE	Structural alignment between two proteins	Search based on structural similarity; Need to prepare comparative structural data in advance; Generate a large number of result files of pairwise structures; Compared with CE and Dali, TM-align runs faster and has higher accuracy and coverage of the alignment region.
Dali/Foldseek	Structural alignment between the input protein and a structure library (e.g. AlphaFold structures and PDB structures)	Search based on structural similarity; Dali requires uncontrollable time to queue up after task submission, and don't provide more related information about target proteins to users; The running speed of Foldseek is fast. But it is necessary to filter the results based on the species, and Foldseek outputs insufficient information about candidates.
SCOP2/CATH/InterPro	To search for proteins with certain structural features or evolutionary relationships	Search based on known protein structure classification; The results are often some proteins with a certain type of structural feature; Due to the lack of specific scores for structural alignment, similarity cannot be compared intuitively.
BLAST/SWISS-MODEL	To search for homologous proteins	Search for homologous proteins based on sequence similarity.
Hu-mAb	Humanization of non-human antibodies	Search based on machine learning methods; Suitable for humanization of non-human antibodies.
Llamanade	Humanization of Nanobody proteins	Search based on sequence similarity; Suitable for humanization of Nanobody proteins.
PROSCA	To search for human scaffolds with similar structures to input proteins	Search based on structural similarity; Provide several search approaches (e.g. predicted structures, experiment resolved structures, even specific to the structures of a certain family or domain of proteins and the human proteome) and 56 types of advantageous scaffolds from the SYNBIP database; More information about input protein and target proteins; Connect with other protein database such as UniprotKB and RCSB PDB in order to provide more information.

relationship between the structurally aligned protein fragment and the full-length protein is clearly showed through color-coded amino acid sequences and visual 3D structures. In addition, PROSCA offers candidate proteins' expression information in 45 human tissues, the levels of which are usually associated with the stability and side effect *in vivo* (11). The figure of expression visualization can be download for presentation or publication usage. Finally, four indicators ('Coverage', 'TM-score', 'Aligned Residues' and 'Aligned Structures') are provided for the evaluation of the alignment quality between the candidate and query. To improve the comprehensiveness of search results, PROSCA achieves interactive connections with other protein database such as UniprotKB (19) and RCSB PDB (4).

Comparison with other related tools

Until now, some tools have been developed for protein structure comparison such as TM-align (10), CE (21), Dali (22) and Foldseek (23). Few tools connect the humanization method with the human proteome directly based on structural similarity except for Dali and Foldseek. Table 1 lists the comparison of PROSCA with other related tools. Compared with CE and Dali, TM-align runs faster, and its resulting structure alignments have higher accuracy and coverage (10). However, the user needs to prepare the structures of all human proteins before using TM-align. At the same time, the results need to be further processed because a result file only contains the structural alignment information of two proteins, which is a challenge for users without computational expertise. Although Dali can align the input protein with AlphaFold proteins and PDB proteins, tests of the tool show that the queue time is uncertain and uncontrollable after task submission (it may take several minutes or even days), and the result information about target proteins is insufficient. Foldseek significantly decreases computation times (within a few minutes), but did not pre-treat protein structures from RCSB PDB (4) and requires further result screening based on species. And it only outputs the names, origins, alignment information of target proteins. What's more, the results obtained from the above tools lacked the important information of protein expression level in human body. In addition, we also examined the accuracy and the number of scaffolds identified. A structure file with ID 1VES from the RCSB PDB was used as the input. The results indicated that there was a significant difference in the number of scaffolds identified among the various tools. PROSCA outputted 2226 predicted human proteins with similar structures to the input protein, Foldseek outputted 378, and Dali outputted 1780. PROSCA uniquely identified 416 structures that were not recognized by Foldseek or Dali. As a result, PROSCA identified the highest number of human protein structures similar to the input protein structure.

In addition, some other tools are also used to evaluate the similarity of two structures or find similar protein structures, but there are still limitations on the identification of privileged scaffolds from human proteome. Protein databases, for example SCOP2 (24), CATH (25) and InterPro (26), provide built-in tools for users to search similar structures by dividing proteins into several categories according to structural characteristics like topology and domain. After importing a specific structure, the user will get result data without the evaluation of the similarity between the input structure and identified

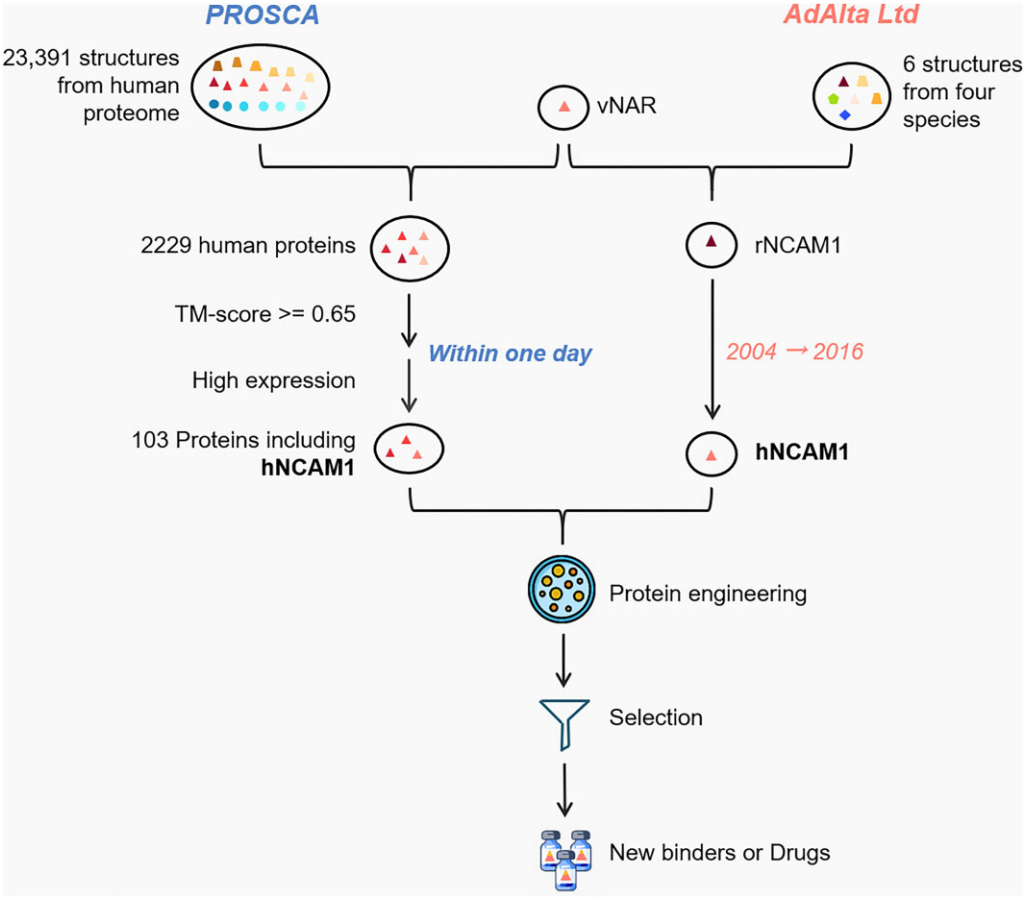


Figure 2. The workflows of both PROSCA and AdAlta to identify hNCAM1 as a privileged scaffold for *i*-body development. The same shape of the figures indicates that these protein structures are similar. If the color of a human protein is closer to the one of input protein, the structural similarity is higher.

structures, even without providing the annotations of protein origins. So far, there are some tools that can be used for protein humanization based on sequence similarity such as protein BLAST (27) and SWISS-MODEL (28), which are usually used to find homologous proteins based on sequence similarity with the input sequences. Recently, two specific tools (Hu-mAb (29) and Llamanade (30)) for humanization of non-human antibodies were reported. Hu-mAb was developed based on machine learning method and Llamanade was developed based on homologous proteins. However, these humanization methods explore limited structural space of human proteome or have limited applicability.

Therefore, PRCOSCA is a solution for mining new scaffolds from the space of the entire human proteome automatically based on AFalign and PDBalign algorithms, and provides quantitative analysis data of protein expression level in 45 tissues of human. The platform offers 23 391 AlphaFold predicted structures and 40 129 PDB structures from homo species to reduce the workload of users to prepare data. PRCOSCA allows the input protein to structurally align with almost all (98.5%) human proteins (18), and the probability of finding similar proteins is significantly increased.

Case study: identification of hNCAM1 as a privileged scaffold for *i*-body development

To further demonstrate the reliability of PROSCA, an Example.pdb was used as the input (Figure 2). The structure file was retrieved from RCSB PDB (PDBID: 1VES (31)). It represents a kind of scaffold structure called vNAR, which has some advantaged characteristics such as small size, high stability, good tissue permeability and high affinity and specificity to therapeutic target CXCR4. Subsequently, we selected all of AlphaFold structures to structurally align with vNAR protein, wrote an email address and submitted. After about 5 hours, the result link and mission ID appeared in the email. Through clicking the link, we got 2229 results which were structurally aligned protein fragments. To obtain structures with high similarity, 7 results with TM-score values of about 0.5, 0.55, 0.6, 0.65, 0.7, 0.75 and 0.8 were selected to compare with each other. According to customized acceptance criteria, we determined to take the results with TM-score values above 0.65 and high expression levels, resulting 307 protein fragments corresponding to 103 human proteins (Figure 2). These proteins can be redesigned or engineered to develop new functional proteins.

It is worth mentioning that the protein hNCAM1, which is one of the identified candidates, has been used for protein engineering for *i*-body drugs development (16). In 2004, Streltsov found that vNAR was structurally similar with rat NCAM1 (rNCAM1) through comparing the structures of vNAR and 6 proteins (Nanobody, rat NCAM, human VH, human VL, Human TCR V alpha, and Telokin) (31), and they found that rat NCAM1 was similar with human NCAM1 in structure (Figure 2). Based on the above conclusions, AdAlta Ltd tried to graft two loops of vNAR onto hNCAM1 in 2016 and has successfully produced *i*-bodies (16). One of *i*-bodies,

named AD-214, has successfully completed the Phase I safety study (17).

In this case, both methods including PROSCA and AdAlta identified the hNCAM1, which is structural similar with vNAR, as privileged scaffold for protein engineering (Figure 2). Nevertheless, AdAlta used limited numbers of proteins to structurally align with vNAR and relied heavily on the experience of previous research. PROSCA explores the structural space of proteome and runs quickly and efficiently. Thus, PROSCA is a powerful tool to solve immunogenicity caused by non-human species and has potential applications in drug research and development.

Conclusion

Therapeutic proteins always undergo humanization before entering clinical trials. With the rapid development of protein design and protein structure prediction technology in the last few years, it is now possible to identify protein scaffolds that have advantages for rational protein engineering. Here, PROSCA was designed to be an effective and interactive platform for mining human scaffolds with low immunogenicity, designated structure and expression characteristics. Unlike sequence similarity search, it works based on structural similarity. By submitting a structural file of protein, PROSCA will automatically and quickly display the results through sorting and filtering functions, tables and visualizations, which removes computational barriers for scientist that work in a wet lab. It is believed that, along with the advanced protein engineering techniques and selection methods, PROSCA will play an increasingly important role in functional protein design.

Data availability

The source code for PROSCA v1.0 is open for academic usage and available in the FigShare repository at https://figshare.com/articles/software/PROSCA/22219291. The PROSCA web server is available at https://idrblab.org/prosca/.

Acknowledgements

We would like to thank the users who using PROSCA and providing us with valuable suggestions for further improvement of this server.

Funding

Natural Science Foundation of Chongqing [2023NSCQ-MSX0140]; Technology Innovation and Application Demonstration Project of Chongqing [cstc2018jscx-msybX0287]; Entrepreneurship and Innovation Support Plan for Chinese Overseas Students of Chongqing [cx2020127]; Open Project of Central Nervous System Drug Key Laboratory of Sichuan Province [230012-01SZ]. Funding for open access charge: Natural Science Foundation of Chongqing.

Conflict of interest statement

None declared.

References

1. Gebauer,M. and Skerra,A. (2020) Engineered protein scaffolds as next-generation therapeutics. *Annu. Rev. Pharmacol. Toxicol.*, **60**, 391–415.
2. Binz,H.K., Amstutz,P. and Pluckthun,A. (2005) Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.*, **23**, 1257–1268.
3. Kang,S., Davidsen,K., Gomez-Castillo,L., Jiang,H., Fu,X., Li,Z., Liang,Y., Jahn,M., Moussa,M., DiMaio,F., *et al.* (2019) COMBINES-CID: an efficient method for de novo engineering of highly specific chemically induced protein dimerization systems. *J. Am. Chem. Soc.*, **141**, 10948–10952.
4. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Christie,C.H., Dalenberg,K., Di Costanzo,L., Duarte,J.M., *et al.* (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.
5. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
6. Golinski,A.W., Holec,P.V., Mischler,K.M. and Hackel,B.J. (2019) Biophysical characterization platform informs protein scaffold evolvability. *ACS Comb. Sci.*, **21**, 323–335.
7. Liang,T., Chen,H., Yuan,J., Jiang,C., Hao,Y., Wang,Y., Feng,Z. and Xie,X.Q. (2021) IsAb: a computational protocol for antibody design. *Brief Bioinform.*, **22**, bbab143.
8. Cao,L., Coventry,B., Goreshnik,I., Huang,B., Sheffler,W., Park,J.S., Jude,K.M., Markovic,I., Kadam,R.U., Verschueren,K.H.G., *et al.* (2022) Design of protein-binding proteins from the target structure alone. *Nature*, **605**, 551–560.
9. Cao,L., Goreshnik,I., Coventry,B., Case,J.B., Miller,L., Kozodoy,L., Chen,R.E., Carter,L., Walls,A.C., Park,Y.J., *et al.* (2020) De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science*, **370**, 426–431.
10. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
11. Uhlen,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A., *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
12. Wang,X., Li,F., Qiu,W., Xu,B., Li,Y., Lian,X., Yu,H., Zhang,Z., Wang,J., Li,Z., *et al.* (2022) SYNBIP: synthetic binding proteins for research, diagnosis and therapy. *Nucleic Acids Res.*, **50**, D560–D570.
13. Koide,S., Koide,A. and Lipovsek,D. (2012) Target-binding proteins based on the 10th human fibronectin type III domain ((1)(0)Fn3). *Methods Enzymol.*, **503**, 135–156.
14. Tian,P., Lemaire,A., Senechal,F., Habrylo,O., Antonietti,V., Sonnet,P., Lefebvre,V., Isa Marin,F., Best,R.B., Pelloux,J., *et al.* (2022) Design of a protein with improved thermal stability by an evolution-based generative model. *Angew. Chem. Int. Ed Engl.*, **61**, e202202711.
15. Tian,P., Louis,J.M., Baber,J.L., Aniana,A. and Best,R.B. (2018) Co-evolutionary fitness landscapes for sequence design. *Angew. Chem. Int. Ed Engl.*, **57**, 5674–5678.
16. Griffiths,K., Dolezal,O., Cao,B., Nilsson,S.K., See,H.B., Pfleger,K.D.G., Roche,M., Gorry,P.R., Pow,A., Viduka,K., *et al.* (2016) i-bodies, Human single domain antibodies that antagonize chemokine receptor CXCR4. *J. Biol. Chem.*, **291**, 12641–12657.
17. Cao,Q., Huang,C., Yi,H., Gill,A.J., Chou,A., Foley,M., Hosking,C.G., Lim,K.K., Triffon,C.F., Shi,Y., *et al.* (2022) A single-domain i-body, AD-114, attenuates renal fibrosis through blockade of CXCR4. *JCI Insight*, **7**, e143018.

18. Tunyasuvunakool,K., Adler,J., Wu,Z., Green,T., Zielinski,M., Žídek,A., Bridgland,A., Cowie,A., Meyer,C., Laydon,A., *et al.* (2021) Highly accurate protein structure prediction for the human proteome. *Nature*, **596**, 590–596.
19. UniProt, C. (2023) UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
20. Uhlen,M., Oksvold,P., Fagerberg,L., Lundberg,E., Jonasson,K., Forsberg,M., Zwahlen,M., Kampf,C., Wester,K., Hober,S., *et al.* (2010) Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.*, **28**, 1248–1250.
21. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
22. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
23. van Kempen,M., Kim,S.S., Tumescheit,C., Mirdita,M., Lee,J., Gilchrist,C.L.M., Söding,J. and Steinegger,M. (2024) Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.*, **42**, 243–246.
24. Andreeva,A., Kulesha,E., Gough,J. and Murzin,A.G. (2020) The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.*, **48**, D376–D382.
25. Sillitoe,I., Bordin,N., Dawson,N., Waman,V.P., Ashford,P., Scholes,H.M., Pang,C.S.M., Woodridge,L., Rauer,C., Sen,N., *et al.* (2021) CATH: increased structural coverage of functional space. *Nucleic Acids Res.*, **49**, D266–D273.
26. Paysan-Lafosse,T., Blum,M., Chuguransky,S., Grego,T., Pinto,B.L., Salazar,G.A., Bileschi,M.L., Bork,P., Bridge,A., Colwell,L., *et al.* (2023) InterPro in 2022. *Nucleic Acids Res.*, **51**, D418–D427.
27. Boratyn,G.M., Camacho,C., Cooper,P.S., Coulouris,G., Fong,A., Ma,N., Madden,T.L., Matten,W.T., McGinnis,S.D., Merezhuk,Y., *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.*, **41**, W29–W33.
28. Waterhouse,A., Bertoni,M., Bienert,S., Studer,G., Tauriello,G., Gumienny,R., Heer,F.T., de Beer,T.A.P., Rempfer,C., Bordoli,L., *et al.* (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.*, **46**, W296–W303.
29. Marks,C., Hummer,A.M., Chin,M. and Deane,C.M. (2021) Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics*, **37**, 4041–4047.
30. Sang,Z., Xiang,Y., Bahar,I. and Shi,Y. (2022) Llamanade: an open-source computational pipeline for robust nanobody humanization. *Structure*, **30**, 418–429.
31. Streltsov,V.A., Varghese,J.N., Carmichael,J.A., Irving,R.A., Hudson,P.J. and Nuttall,S.D. (2004) Structural evidence for evolution of shark ig new antigen receptor variable domain antibodies from a cell-surface receptor. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 12444–12449.

Received: February 29, 2024. **Revised:** April 23, 2024. **Editorial Decision:** April 26, 2024. **Accepted:** April 29, 2024