Profrager Web Server: Fragment Libraries Generation for Protein Structure Prediction

Karina B. Santos[1], Raphael Trevizani[2], Fábio L. Custódio[1*] and Laurent E. Dardenne[1]
[1]Dept. of Comp. Mechanics - National Laboratory for Scientific Computing (LNCC), Petrópolis, RJ, Brazil
[2]Oswaldo Cruz Foundation (Fiocruz), Fortaleza, CE, Brazil
Email: karinabs@lncc.br, raphael.trevizani@fiocruz.br, *flc@lncc.br, dardenne@lncc.br
*Corresponding author

Abstract— *This paper describes Profrager, a new flexible web server for generation of protein fragment libraries. These libraries have widespread use amongst modern protein structure prediction methods. Profrager offers several options for generating customized libraries, e.g., the users can choose between three options of structural databases to generate the libraries and can also define the number of fragments per position and the fragments lengths. The selection of fragments can be guided by three scoring strategies: (i) use only sequence similarity to the target sequence; (ii) use a weighted sum of the sequence similarity score and a secondary structure score; (iii) use a Pareto Efficiency strategy with the two scores. The software outputs useful statistics about the fragments in addition to files fully compatible with the GAPF and Rosetta protein structure prediction programs. Profrager is available at* `http://www.lncc.br/sinapad/Profrager/` *as a web service.*

Keywords: protein fragment libraries; protein structure prediction

1. Introduction

The prediction of protein structures is a central challenge of modern computational biology [1]. The use of fragment libraries is one of the basic strategies employed by several successful protein structure prediction (PSP) methods [2], [3]. The objective is to simplify the complexity of PSP by reducing the conformational search space [4]. Fragment libraries are assembled from a database of experimentally determined structures and are specific to each target protein sequence. These libraries can be understood as a selected collection of possible fragments which are used to construct segments of a target sequence. Information contained within the fragments is used to build the whole tridimensional structure of the target protein [5]. Commonly, libraries are constructed by similarity between the amino acids sequences of the fragments and the target protein [6]. However, other criteria may be used, e.g., the agreement between the observed fragments secondary structure and the predicted secondary structure of the target [7].

Robust fragment libraries should allow the reconstruction of the correct protein folding using only the fragments from non-homologous structures [8]. Therefore, programs for fragment libraries generation should present several options to guide the choice of fragments in order to improve the prediction capacity of PSP methods, e.g., the amino acid substitution matrix used to select the fragments and the database of experimental structures from which to extract the fragments.

A web server provides a user friendly interface to generate the libraries without the need to install any programs, and avoids the lengthy process of creating a geometry database that are used for the fragment construction. An example of a web server, which enables users to create fragment libraries, is the Robetta server [9] (`http://robetta.bakerlab.org/`). The Rosetta method uses, as one the initial steps in its PSP protocol, the generation of fragment libraries for a specific target sequence [10]. Libraries generated by Robetta are specifically formatted to be used with the Rosetta PSP software.

The objective of this work was the development of a flexible program for creating customized PSP fragment libraries, using different databases, amino acids substitution matrices and scoring criteria for fragment selection. This program is available to the scientific community in the form of an interactive web server called Profrager (`http://www.lncc.br/sinapad/Profrager/`).

2. Implementation

Profrager creates fragment libraries from a selected database of known protein structures. This database is a subset of the Protein Data Bank (PDB) [11] extracted using PISCES (Protein Sequence Culling Server) [12]. At present, the user can choose from two different PISCES databases: (i) one comprising 5387 sequence entries, with no more than 20% identity between the sequences and resolution up to 2.0 angstroms, or (ii) one with 17342 sequence entries, with no more than 50% identity between the sequences and resolution up to 2.5 angstroms. These two databases have structures elucidated by X-ray crystallography (R-factor bellow 0.3) and NMR. Additionally, a third database option is available for the users, the Rosetta's Vall database, with 16800 sequence entries.

Profrager is capable of generating libraries with fragments of any length. Furthermore, in addition to the fragment length the user can also define the number of fragments per position. A "position" refers to the residue in the target sequence where the fragment starts. Thus, a fragment of three residues from the first position contains the structural information of residues 1, 2 and 3. The fragments overlap in consecutive positions, e.g., the second position appertain to residues 2, 3 and 4, the third 3, 4 and 5 and so forth until the end of the target sequence is reached. The default options are 200 fragments per position and libraries with three and nine residues length.

2.1 Profrager in Use

From a target sequence the program scans the chosen database building a list of candidate fragments for each position. The choice of which fragments will be included in the final library is guided by a ranking score. Each candidate fragment has its sequence similarity, to the corresponding segment on the target sequence, evaluated using an amino acids substitution matrix. The user may choose to use BLOSUM62 (default), BLOSUM45, PAM30 or PAM80 matrices. Sequence similarity identifies the probability of an amino acid being replaced by another in the protein sequence. The sequence similarity score is given by the sum of values from the matrix comparing the fragment sequence with the target segment sequence.

The selection of fragments can be augmented by comparing the predicted secondary structure for the target sequence, using PSIPRED [13] (or other program by providing a secondary structure file in the horizontal format), and the secondary structure for the proteins in the database detected using STRIDE [14]. The score for this comparison is calculated using the confidence given by PSIPRED for each residue. When the predicted secondary structure for a position on the target sequence is the same as the detected in the corresponding position in a fragment, the confidence score is added to the score of that fragment. Otherwise, the confidence is subtracted from the score. The secondary structure score is added to the similarity score and the final score is used for fragment ranking. Moreover, an important customization aspect is that the secondary structure score can be multiplied by a weight defined by the user (1.0 by default).

Another fragment selection option implemented in Profrager, is the use of a multi-objective Pareto Efficiency strategy [15]. This strategy avoids the choice of a particular value to weight the two scores (amino acid and secondary structure similarities). Pareto Efficiency employs the concept of dominance where fragments which have the best scores for at least one criterion are classified as non-dominated and make up the Pareto Front. Successive fronts are used to build the fragment libraries until the desired number of fragments per position is fulfilled. In general, these are the fragments that have the best values for at least one evaluation criteria.

Users have access to other advanced options during the creation of their libraries. The minimum score a fragment need to obtain to be included in the library can be controlled. Furthermore, fragments might be extracted from: (i) any protein in the database, (ii) only non-homologous proteins to the target sequence or (iii) exclusively from homologous proteins to the target sequence. In these two last cases, homology is detected using PSI-BLAST [16] via the E-Value.

2.2 Output

The default output format is compatible with the GAPF PSP suite developed in our group (http://www.gmmsb.lncc.br) [17], [18], [19]. The fragment libraries files contain, for each residue at each position, the following information in separate columns: (1) PDB code and chain of the structure which originated the fragment, (2) type of amino acid (one letter code), (3) type of secondary structure, (4) position in the target sequence, (5) position in the sequence from the PDB, (6) backbone dihedral angles ϕ, ψ and ω, (7) main chain bond angles defined by N-Cα-C, Cα-C-N and C-N-Cα, (8) the score from sequence similarity, (9) the score from secondary structure agreement and (10) the total score. Each file is a fragment library of a particular length and all target sequence positions are marked with a header line containing the position number and the number of fragments at that position. Moreover, allowing for a wide range of applications for the libraries generated by Profrager, files in a format compatible with Rosetta are created by default. Another useful output is an automatically generated plot depicting the fragments' secondary structure distribution, per position, for each library created.

It has been shown that when using backbone angles from experimental structures with idealized (and fixed) bond geometries, e.g., in *de novo* and *ab initio* PSP, the resulting structures can present large deviations from the original structure, for longer sequences this is more severe [8]. This can be solved by freezing bonds under idealized geometries and then optimizing the backbone dihedral angles to recreate structures as close as possible to the originals. Alternatively, the libraries can include the backbone geometry bond angles. The first option has the disadvantage of requiring a preliminary processing of all structures contained in the database, in addition to changing the actual values of the original backbone angles. Profrager generated libraries have backbone torsion angles values extracted directly from experimental structures. A different choice can be found in libraries generated by Robetta, which have backbone dihedral angles calculated from structures with idealized bond geometries [20]. Nevertheless, Profrager users' have the option of using Rosetta's geometries database to generate the fragment libraries. In this case, Profrager libraries will

contain recalculated dihedral angles and fixed idealized main chain bond angles.

3. Methods

For the validation of the libraries and to demonstrate their compatibility with the Rosetta suite, a set of 48 proteins ranging from 54 to 148 residues was selected from the CASP9 experiment (Table 1). For each target, three differents fragment libraries were generated by Profrager with the Rosetta's Vall database using: (I) only sequence similarity, given by Blosum62, (II) sum of the similarity score and the secondary structure score (weight=1.0) and (III) Pareto Efficiency strategy. Each generated library has 200 fragments with three residues (3-mers) and nine residues (9-mers) for each position of the target sequence. For each type of generated fragment libraries we perform a protein structure prediction protocol using Rosetta (version 3.4). The default *ab initio*-relax protocol was used and 1000 models for each sequence were generated. The quality of generated structures was evaluated with the TM-Score program [21]. This program gives the GDT-TS criterion (Global Distance Test Total Score) and only the best model (i.e., with greater GDT-TS value) was considered during comparisons. Models with GDT-TS $\geq 50\%$ indicate good predictive ability [22]. For the sake of comparison, all tests were also performed against fragment libraries generated by the Robetta server (using secondary structure prediction and sequence similarities scoring schemes), which are the default libraries for Rosetta predictions.

4. Results and Discussion

Table 1 shows the GDT-TS values of the best models generated using fragment libraries from Robetta server and using fragment libraries from Profrager server - Library I (based on sequence similarity alone), Library II (based on secondary structure prediction and sequence similarity) and Library III (based on Pareto Efficiency).

Library I showed the worst results and for the majority of sequences the best models generated using the Robetta libraries have GDT-TS values similar to those generated using Profrager libraries II and III.

The number of sequences that had models with good quality generated, GDT-TS > 50%, was: Library I: 3, Library II: 9, Library III: 9 and Robetta Library: 11. Thus, the models created by Rosetta using libraries generated by the Robetta server have a small advantage over those created using libraries generated by Profrager. By comparing the distribution of secondary structures between different libraries (Fig. 1) it becomes apparent that there are considerable differences between those from Profrager libraries, which uses PSIPRED, and those from Robetta libraries. For example, Profrager provides mainly coil fragments between residues 45 and 55, while Robetta provides mainly helical fragments.

Table 1: GDT-TS score (%) of the best models.

CASP9 ID	PDB	Library I	Library II	Library III	Robetta*	Length
T0522	3nrd	28.85	41.54	39.42	42.31	134
T0523	3mqo	27.48	40.77	35.36	42.79	120
T0527	3mr0	24.61	29.33	28.35	32.09	142
T0530	3npp	33.43	44.77	47.97	58.43	115
T0531	2kjx	33.85	41.54	40.00	43.46	65
T0538	2l09	60.08	84.27	79.44	82.66	54
T0539	2l0b	22.25	21.70	21.7	20.60	81
T0540	3mx7	46.67	57.50	54.44	57.50	90
T0541	2l0d	28.95	40.13	32.46	36.40	106
T0544	2l3w	33.74	35.66	30.07	46.50	135
T0546	2l5q	27.29	32.39	30.46	29.93	134
T0548	3nnq	36.96	51.09	45.92	45.38	106
T0549	2kzv	40.49	50.00	46.74	55.16	84
T0551	3obh	28.13	28.52	28.52	31.25	74
T0552	2l3b	29.81	33.85	36.73	41.73	122
T0553	2ky4	28.69	40.10	31.71	40.60	141
T0555	2l06	28.23	27.42	25.16	34.19	148
T0557	2kyy	26.31	32.03	32.19	34.15	145
T0559	2l01	47.40	66.88	64.94	77.92	69
T0560	2l02	49.70	61.28	61.59	68.60	74
T0562	2kzx	32.44	37.98	34.35	39.89	123
T0564	2l0c	26.80	41.49	40.98	41.75	89
T0567	3n70	21.07	24.82	24.82	26.07	145
T0569	2kyw	37.93	38.51	42.53	38.79	79
T0572	2kxy	25.00	26.50	26.75	29.00	93
T0574	3nrf	26.49	32.18	35.64	41.83	126
T0579	2ky9	19.32	21.78	20.83	22.92	124
T0580	3nbm	35.89	47.28	45.54	47.52	105
T0581	3npd	40.77	38.96	34.91	43.69	136
T0586	3neu	28.91	36.52	36.09	33.48	125
T0590	2kzw	24.14	18.28	18.62	26.90	137
T0592	3nhv	19.70	19.89	19.13	17.99	144
T0594	3ni8	22.68	22.50	21.25	20.54	140
T0600	3nja	25.48	25.48	25.00	26.92	125
T0602	3nkz	41.40	47.04	50.27	50.27	123
T0605	3nmd	53.85	62.50	60.58	62.50	72
T0612	3o0l	33.41	36.45	36.92	40.89	129
T0614	3voq	24.35	23.71	22.20	24.57	135
T0616	3nrt	32.78	39.17	38.89	39.72	103
T0617	3nrv	27.38	34.13	36.11	37.50	148
T0619	3nrw	36.52	29.17	31.13	34.31	111
T0622	3nkl	40.21	60.83	51.25	57.71	138
T0624	3nrl	41.04	51.49	44.78	55.97	81
T0630	2kyt	26.40	31.20	30.80	37.40	132
T0634	3n53	25.00	34.48	39.44	42.46	140
T0637	2x3o	22.33	26.53	25.19	23.47	146
T0639	3nym	28.54	32.08	36.50	34.96	128
T0643	3nzl	51.43	58.21	63.93	63.21	83
Average		32.38	38.75	37.66	41.33	

Library I: built using only sequence similarity. Library II: built using sequence similarity score and secondary structure prediction agreement. Library III: built using a Pareto Efficiency Strategy. Robetta: libraries built using Robetta server.

This is because the Robetta server uses a weighted average of various secondary structure prediction methods [23]: PSIPRED, JUFO [24], SAM [25] and PROF [26]. In general, the use of a consensus strategy improves the accuracy of the prediction [27].

It is interesting to note that the use of secondary structure prediction information greatly reduces fragment diversity within the libraries to the point of some positions having

only one type of secondary structure represented (Fig. 1). This is a desirable trait as it reduces the search space for PSP algorithms, allowing that better models be generated. However, good predictions become extremely dependent of accurate secondary structure predictions.

5. Conclusions

Profrager web server for protein fragment libraries generation presents a wide range of flexible and useful options, not present in other similar services. Beyond the basic options of number of fragments per position and fragment length, the web server allows the creation of libraries from different pre-built databases (20% or 50% identity cut-off) or even Rosetta's Vall database. The different strategies for fragment selection offered by Profrager can be used to generate distinct fragment libraries for PSP programs. Moreover, the facilities provided by Profrager can be interesting to enable the development of new PSP strategies by other research groups in this field.

To further improve the quality of the libraries, an option to use several secondary structure prediction methods is being implemented in Profrager.

Acknowledgment

Funding: This work was supported by CNPq - Contract grant n°. 307062/2010-4 and FAPERJ - Contract grant n°. E26/102.443/2009. SINAPAD (`https://www.lncc.br/sinapad/`) team for hosting and web server and Eduardo Krempser for drafting the first versions of the server.

References

[1] D. Baker and A. Sali, "Protein structure prediction and structural genomics," *Science*, vol. 294, no. 5540, pp. 93–96, 2001.

[2] S. Raman, R. Vernon, J. Thompson, M. Tyka, R. Sadreyev, J. Pei, D. Kim, E. Kellogg, F. DiMaio, O. Lange, *et al.*, "Structure prediction for casp8 with all-atom refinement using rosetta," *Proteins: Structure, Function, and Bioinformatics*, vol. 77, no. S9, pp. 89–99, 2009.

[3] Y. Zhang, "I-tasser: Fully automated protein structure prediction in casp8," *Proteins: Structure, Function, and Bioinformatics*, vol. 77, no. S9, pp. 100–113, 2009.

[4] R. Kolodny, P. Koehl, L. Guibas, and M. Levitt, "Small libraries of protein fragments model native protein structures accurately." *Journal of Molecular Biology*, vol. 323, no. 2, pp. 297–307, 2002.

[5] R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. Strauss, and D. Baker, "Rosetta in casp4: progress in ab initio protein structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 45, no. S5, pp. 119–126, 2001.

[6] S. Li, D. Bu, X. Gao, J. Xu, and M. Li, "Designing succinct structural alphabets," *Bioinformatics*, vol. 24, no. 13, pp. i182–i189, 2008.

[7] D. Chivian, D. Kim, L. Malmström, J. Schonbrun, C. Rohl, and D. Baker, "Prediction of casp6 structures using automated robetta protocols," *Proteins: Structure, Function, and Bioinformatics*, vol. 61, no. S7, pp. 157–166, 2005.

[8] J. Holmes and J. Tsai, "Some fundamental aspects of building protein structures from fragment libraries," *Protein science*, vol. 13, no. 6, pp. 1636–1650, 2004.

[9] D. Kim, D. Chivian, and D. Baker, "Protein structure prediction and analysis using the robetta server," *Nucleic Acids Research*, vol. 32, no. suppl 2, pp. W526–W531, 2004.

[10] P. Bradley, D. Chivian, J. Meiler, K. Misura, C. Rohl, W. Schief, W. Wedemeyer, O. Schueler-Furman, P. Murphy, J. Schonbrun, *et al.*, "Rosetta predictions in casp5: successes, failures, and prospects for complete automation," *Proteins: Structure, Function, and Bioinformatics*, vol. 53, no. S6, pp. 457–468, 2003.

[11] H. Berman, K. Henrick, H. Nakamura, and J. Markley, "The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D301–D303, 2007.

[12] G. Wang and R. Dunbrack Jr, "Pisces: a protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, 2003.

[13] L. McGuffin, K. Bryson, and D. Jones, "The psipred protein structure prediction server," *Bioinformatics*, vol. 16, no. 4, pp. 404–405, 2000.

[14] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," *Proteins: Structure, Function, and Bioinformatics*, vol. 23, no. 4, pp. 566–579, 1995.

[15] A. Charnes, W. Cooper, B. Golany, L. Seiford, and J. Stutz, "Foundations of data envelopment analysis for pareto-koopmans efficient empirical production functions," *Journal of Econometrics*, vol. 30, no. 1, pp. 91–107, 1985.

[16] S. Altschul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[17] F. L. Custódio, H. J. Barbosa, and L. E. Dardenne, "Genetic algorithm for finding multiple low energy conformations of poly alanine sequences under an atomistic protein model," *Advances in Bioinformatics and Computational Biology*, pp. 163–166, 2007.

[18] ——, "Full-atom ab initio protein structure prediction with a genetic algorithm using a similarity-based surrogate model," in *Evolutionary Computation (CEC), 2010 IEEE Congress on*. IEEE, 2010, pp. 1–8.

[19] ——, "A multiple minima genetic algorithm for protein structure prediction," *Applied Soft Computing*, vol. 15, pp. 88–99, 2014.

[20] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, "Protein structure prediction using Rosetta," *Methods Enzymol.*, vol. 383, pp. 66–93, 2004.

[21] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function, and Bioinformatics*, vol. 68, no. 4, pp. 1020–1020, 2007.

[22] J. Xu and Y. Zhang, "How significant is a protein structure similarity with tm-score= 0.5?" *Bioinformatics*, vol. 26, no. 7, pp. 889–895, 2010.

[23] R. Das, B. Qian, S. Raman, R. Vernon, J. Thompson, P. Bradley, S. Khare, M. D. Tyka, D. Bhat, D. Chivian, *et al.*, "Structure prediction for casp7 targets using extensive all-atom refinement with rosetta@home," *Proteins: Structure, Function, and Bioinformatics*, vol. 69, no. S8, pp. 118–128, 2007.

[24] J. Meiler, M. Müller, A. Zeidler, and F. Schmäschke, "Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks," *Molecular modeling annual*, vol. 7, no. 9, pp. 360–369, 2001.

[25] K. Karplus and B. Hu, "Evaluation of protein multiple alignments by sam-t99 using the balibase multiple alignment test set," *Bioinformatics*, vol. 17, no. 8, pp. 713–720, 2001.

[26] M. Ouali and R. D. King, "Cascaded multiple classifiers for secondary structure prediction," *Protein Science*, vol. 9, no. 06, pp. 1162–1176, 2000.

[27] Y. Wei, J. Thompson, and C. Floudas, "Concord: a consensus method for protein secondary structure prediction via mixed integer linear optimization," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. The Royal Society, 2011, p. rspa20110514.

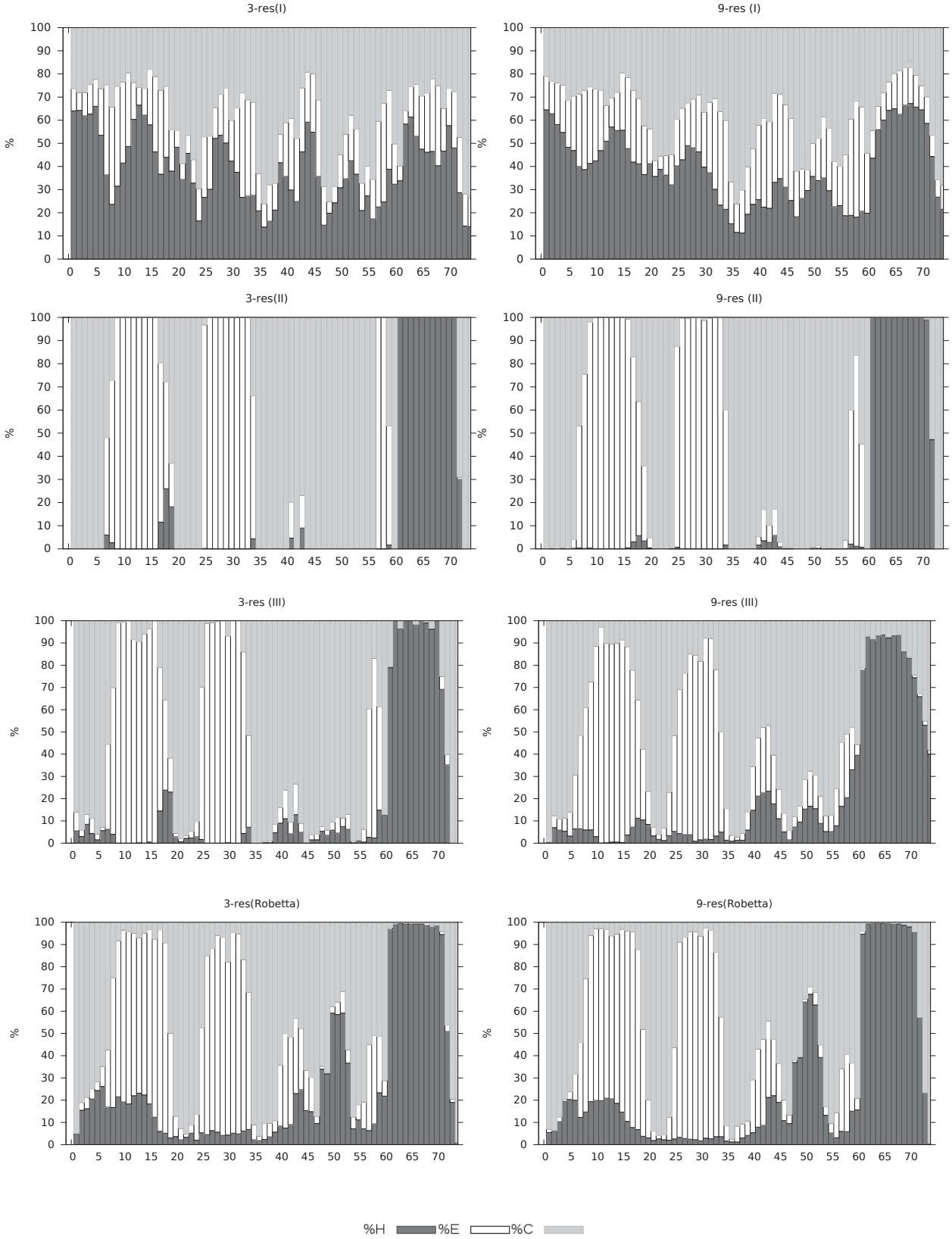

Fig. 1: Secondary structure distribution profile for generated libraries with fragments containing 3 and 9 residues, target T0551. The horizontal axis represent each residue in the sequence. Vertical axes show secondary structure percentages among the fragments at each residue (position): H = helix, C = coil and E = extended. (I) Libraries built using only sequence similarity. (II) Libraries built using sequence similarity score and secondary structure prediction agreement score. (III) Libraries built using Pareto Efficiency strategy. (Robetta) Libraries built using Robetta server. Graphics I, II and III are automatically generated by Profrager.