https://doi.org/ 10.33472/AFJBS.6.9.2024.4580-4594

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Insilico physiochemical and structural characterization of protein sequence of novel strain of *Rhodopseudomonas faecalis*

Aysha Sherieff, M.Mohibbe Azam, K.Sesha Maheswaramma

1. Research Scholar, JNTUA, Anantapur, Andhra Pradesh,India
2. Indian Institute of Rice Research (ICAR-IIRR), Rajendra Nagar, Hyderabad, Telangana State, India
3. Department of Chemistry, JNTUACEP, Pulivendula, Kadapa, Andhra Pradesh, India

Article History: Volume 6,Issue 9, 2024

Received: 30 Apr 2024

Accepted : 17 May 2024

doi: 10.33472/AFJBS.6.9.2024.4580-4594

Abstract

The insilico physio-chemical and structural characteristics of the partial protein sequence of novel strain of *Rhodopseudomonas faecalis* was performed. The physio-chemical characteristics were analysed using EMBOSS Pepstats, EMBOSS Pepinfo and ExPASy ProtScale. Whereas the primary and secondary structures were analysed by using Chou Fasman program, GOR program, Statistical Analysis of Protein Sequences (SAPS), ProSA-web and PROCHECK. The literature review suggests that the characterization of aminoacid sequences by insilico methods were not done for *Rhodopseudomonas faecalis*

Introduction

The insilico analysis of physio-chemical and structural model of a protein is one of the keys for understanding biological processes at molecular level. The computed protein concentration and extinction coefficients are used in quantitative study of protein-ligand and protein-protein interactions in solution (Gill S.C,1989). While computed isoelectric point (pI) is applied for developing buffer systems for purification of recombinant proteins through isoelectric focusing method (Adhikari S et al.,2010). By using amino acid analysis data, a rapid, simple and accurate method was developed for determination of molar ratio of proteins (by calculation of protein concentration) in protein-protein conjugates through calculated extinction coefficients. This method is used in determining protein concentration in any peptide-protein, protein-protein conjugates and peptide-peptide interactions (Zhu D et al.,2013).

Hydropathy properties of of proteins has a crucial role in many fields of computational biology, such as biomolecular interaction, drug design and folding prediction of proteins. Many descriptors were devised to evaluate the hydrophobicity of aminoacid side chains of proteins. Both hydrophobic and hydrophilic effects are dominant driving forces for several

biochemical processes like binding, molecular recognition, protein folding and nucleic acid stability (Tanford, 1972; Brooks et al., 1998; Aftabuddin and Kundu, 2007; Moret and Zebende, 2007; Miotto et al., 2018). The hydropathy index is a number representing the hydrophobic or hydrophilic properties of its side chain was firstly proposed in 1982 by Kyte et al. (Kyte and Doolittle, 1982). It is an effective measure of the interaction between water and amino acids.

Attributing a single number, the hydropathy index to each amino acid, is useful for studying the physio-chemical and structural properties of proteins. Many hydrophobicity and hydrophilicity scales based on both experimental and theoretical approaches have been defined. These schematizations have proven their usefulness in the characterization of protein regions and the development of computational methods (Chothia, 1974; Jones, 1975; Kyte and Doolittle, 1982; Sweet and Eisenberg, 1983; Rose et al., 1985; Wilce et al., 1995). For example, hydrophobicity and hydrophilicity values for 20 amino acids is used for prediction of transmembrane regions in protein structure modeling (Deber et al., 2001). The source of polarity like internal polar or ionizable group if present then internal water molecules cluster around the groups. Unless there is a source of polarity ,it appears unlikely that water molecules alone or in small clusters can be stabilized in small or flexible hydrophobic cavities in the protein interior.

Protein flexibility is one of the factors responsible for disorder, multiple hydration patterns and partial occupancies of water binding sites (Damjanović A et al.,20007). The GOR method formalizes the problem of secondary structure prediction within an information-theoretical framework (Ma Y et al.,2018). The Chou-Fasman algorithm for the prediction of protein secondary structure is one of the most widely used predictive method because of its relative simplicity and reasonable high degree of accuracy (Prevelige. P and Fasman G.D,1989). Validation is an integral part in obtaining three-dimensional models of macromolecules in cryo- Electron Microscopy (Henderson et al., 2012) and X-ray crystallography (Read et al., 2011). It is used in interpreting quality of models from Protein Data Bank (Burley et al., 2019), since there is no formal structural quality requirement for acceptance in this repository. A key quality metric used in validation of quality of atomic models of proteins is Ramachandran plot (Ramachandran et al., 1963). Ramachandran plot describes the 2D distribution of torsion angles of protein backbone (φ, ψ). They have been used for validation of protein backbone conformations in PROCHECK (Laskowski et al., 1993).

In this work, the primary and secondary structure along with physio-chemical characteristics of a protein of novel strain of *Rhodopseudomonas faecalis* was elucidated with some of the finest algorithms, methods and analytical tools.

Methodology

Source of the sequence

Previously a novel strain of *Rhodopseudomonas faecalis* bacterium was isolated from the samples of the Sewage Treatment Plant (STP) of Association of Lady Entrepreneurs of India (ALEAP) Industrial Park in Hyderabad of Telangana State, India. It was sequenced by 16s rRNA methodology at Centre for Environment, Institute of Science and Technology (IST), Jawaharlal Nehru Technological University (JNTU), Hyderabad, Telangana State, India. The

organism, partial sequence was obtained in FASTA format and the same was used to convert in to an aminoacid sequence. The ORF finder was run to obtain the ORF protein sequences and it was used for the present study.

Physico-chemical characteristics

To analyze the physical and chemical characteristics such as molecular weight, theoretical pI, number of residues, average residue weight, molar extinction coefficient, Dayhoff stat, Improbability of expression in inclusion bodies and properties of the protein with mole percentage was computed by EMBOSS Pepstats(Madeira F et al.,2022). Pepinfo (EMBOSS) (Madeira F et al.,2022) was used to create a variety of plots that display different amino acid properties like hydropathy or charged residues and their position in the sequence. To display protein sequence profiles as 2D plot, including hydropathicity, refractivity, hydrophobicity, average flexibility index, polarity and trasmembrane tendency, etc the ProtScale tool (Gasteiger E et al.,2005) was used.

Secondary structure prediction

To predict the secondary structure of the protein, Chou-Fasman server (CFSSP) and GOR (Garnier J et al.,1996) was employed. The method implemented the secondary structure predictions based on the analysis of relative frequencies of each amino acid in helices, sheets and turns anchored in the solved X-ray crystallographic protein template (Kumar T A, 2013). Secondary structure prediction method by GOR is one of the popular method utilizing information theory. It predicts the locations of alpha-helix and beta-strand from an amino acid sequence.

Statistical Analysis of Protein Sequences (SAPS) (Madeira F et al.,2022) evaluates a wide variety of properties of protein sequence using statistical analysis. The properties considered include runs of charge, compositional biases, clusters and amino acid types, different kinds and extents of repetitive structures, locally periodic motifs, and anomalous spacings between identical residue types.

PROSA web server (Wiederstein & Sippl, 2007 and Sippl, M.J. 1993) is used to validate the modeled protein structure with available protein structure which can be derived from PDB or uploading an input pdb extension file of an aminoacid sequence on the basis of z-score.

The stereo chemical quality of the predicted models and accuracy of the protein model was evaluated after the refinement process using Ramachandran Map calculations computed with the PROCHECK (Laskowski et al., 1993).

Discussion and Results

The aminoacid partial sequence of novel strain of *Rhodopseudomonas faecalis* was obtained from ORF finder tool previously (Fig 1). It has a total of 210 aminoacids which may have potential genes of expression for a functional protein.

```
>Protein
MISHIGTETRPKLLREAAVGNIGQWAQA
MAEVSGTASVEVKFVDIRKNTSGEGGSLGHY
MSRTGRRDVTFSSEPGAQVLHGCRQLVS
MGASLIQPCRVSDEGPRVVKLFCAGR
MLKALAAPLVSKPTNGWHSSFTAWTTRVSNPVCSPRFRASASVMAQ
MQHLCSRLRREGHISATGPGHVKGW
MLHRLCGPPSIPLSFNLATVLPRRNA
```

Fig 1: Aminoacid sequence of the strain of *Rhodopseudomonas faecalis*

Physico-chemical characteristics

EMBOSS Pepstats was used to know the following functions. The Molecular weight of the total aminoacids are 22683.23, 210 Residues, Average Residue Weight is 108.015, Charge value is 19.0, Isoelectric Point value is 11.4109, A280 Molar Extinction Coefficients is 23490 (reduced); 23865 (cystine bridges), A280 Extinction Coefficients are 1mg/ml is 1.036 (reduced) ; 1.052 (cystine bridges) and Improbability of expression in inclusion bodies is 0.827. The individual aminoacid mole% and Dayhoffstat values along with different properties were given in Fig2.

Residue	Number	Mole%	DayhoffStat
A = Ala	19	9.048	1.052
B = Asx	0	0.000	0.000
C = Cys	6	2.857	0.985
D = Asp	3	1.429	0.260
E = Glu	8	3.810	0.635
F = Phe	6	2.857	0.794
G = Gly	20	9.524	1.134
H = His	8	3.810	1.905
I = Ile	7	3.333	0.741
J = ---	0	0.000	0.000
K = Lys	7	3.333	0.505
L = Leu	17	8.095	1.094
M = Met	8	3.810	2.241
N = Asn	6	2.857	0.664
O = ---	0	0.000	0.000
P = Pro	13	6.190	1.190
Q = Gln	7	3.333	0.855
R = Arg	19	9.048	1.846
S = Ser	22	10.476	1.497
T = Thr	12	5.714	0.937
U = ---	0	0.000	0.000
V = Val	17	8.095	1.227
W = Trp	4	1.905	1.465
X = Xaa	0	0.000	0.000
Y = Tyr	1	0.476	0.140
Z = Glx	0	0.000	0.000

Property	Residues	Number	Mole%
Tiny	(A+C+G+S+T)	79	37.619
Small	(A+B+C+D+G+N+P+S+T+V)	118	56.190
Aliphatic	(A+I+L+V)	60	28.571
Aromatic	(F+H+W+Y)	19	9.048
Non-polar	(A+C+F+G+I+L+M+P+V+W+Y)	118	56.190
Polar	(D+E+H+K+N+Q+R+S+T+Z)	92	43.810
Charged	(B+D+E+H+K+R+Z)	45	21.429
Basic	(H+K+R)	34	16.190
Acidic	(B+D+E+Z)	11	5.238

Fig2: Different characteristics of protein of the strain of *Rhodopseudomonas faecalis* using EMBOSS Pepstats

When EMBOSS Pepinfo was run, the histogram of physiochemical properties of the protein and Hydropathy Plots of the protein of *Rhodopseudomonas faecalis* were obtained that specifies the presence of residues having nine characteristic featured types such as tiny, small, aliphatic, aromatic, non-polar, polar, charged, positive and negative residues (Fig 3).

From Fig 4, Using Kyte and Doolittle hydropathy parameter, hydropathy plot of 210 residues were plotted for hydropathy value from -2 to +2 against residue number from 0 to 200 which has shown highest peak at +2 for residue number 90 whereas at -2 highest peak was shown for 50 and 63 residue numbers.

Using OHM hydropathy parameters (Sweet & Eisenberg) hydropathy plot of 210 residues were plotted for hydropathy value from -0.5 to +0.4 against residue number from 0 to 200 which has shown highest peak at +0.4 for residue number 202 whereas at -0.5 highest peak was shown for 50 residue number. Using consensus parameters (Eisenberg et al), hydropathy plot of 210 residues were plotted for hydropathy value from -0.5 to +0.5 against residue number from 0 to 200 which has shown highest peak at +0.5 for residue number 90 whereas at -0.5 highest peak was shown for 62 residue number.

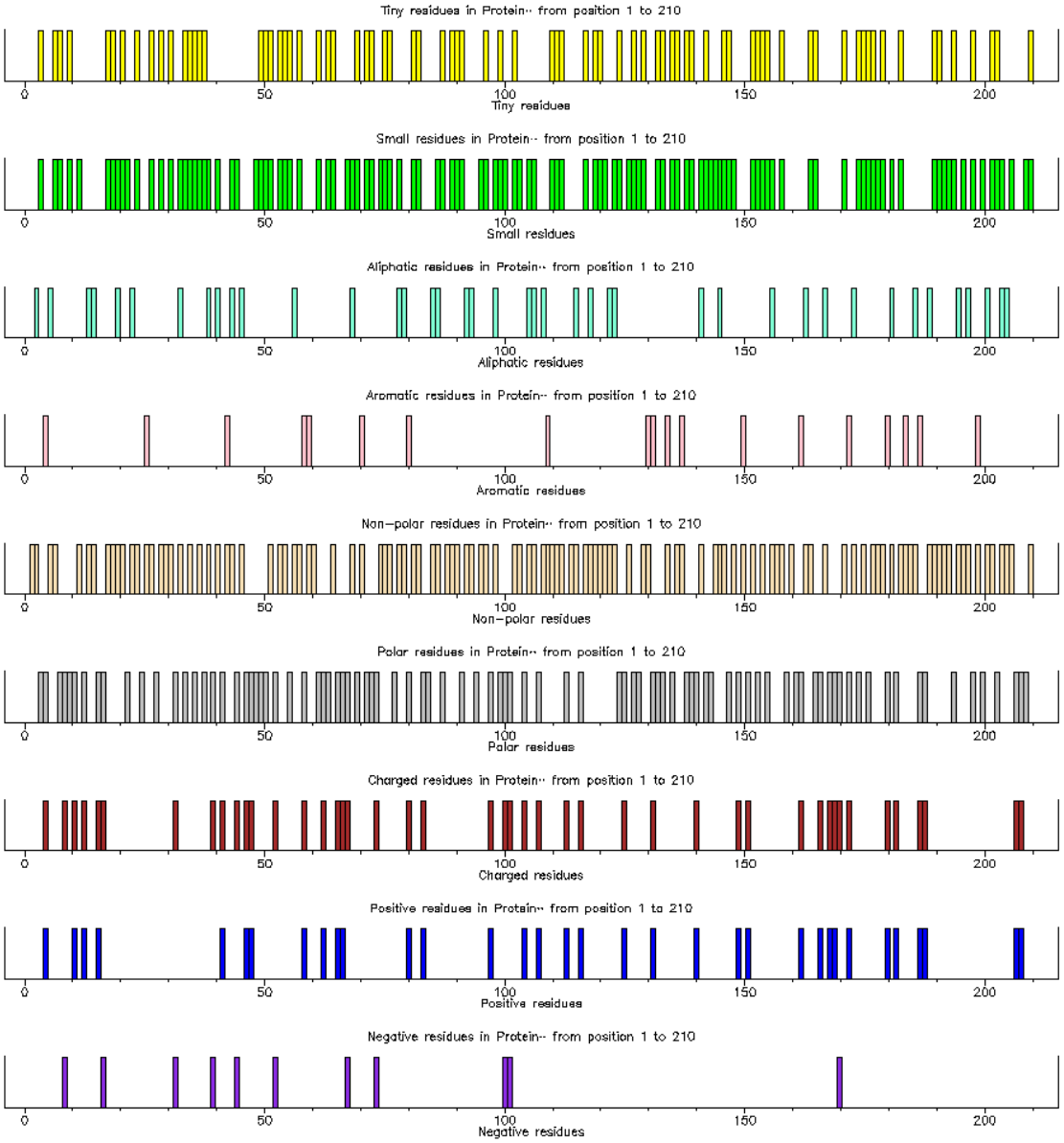


Fig3: Histogram of physiochemical properties of the protein strain of *Rhodopseudomonas faecalis* using EMBOSS Pepinfo

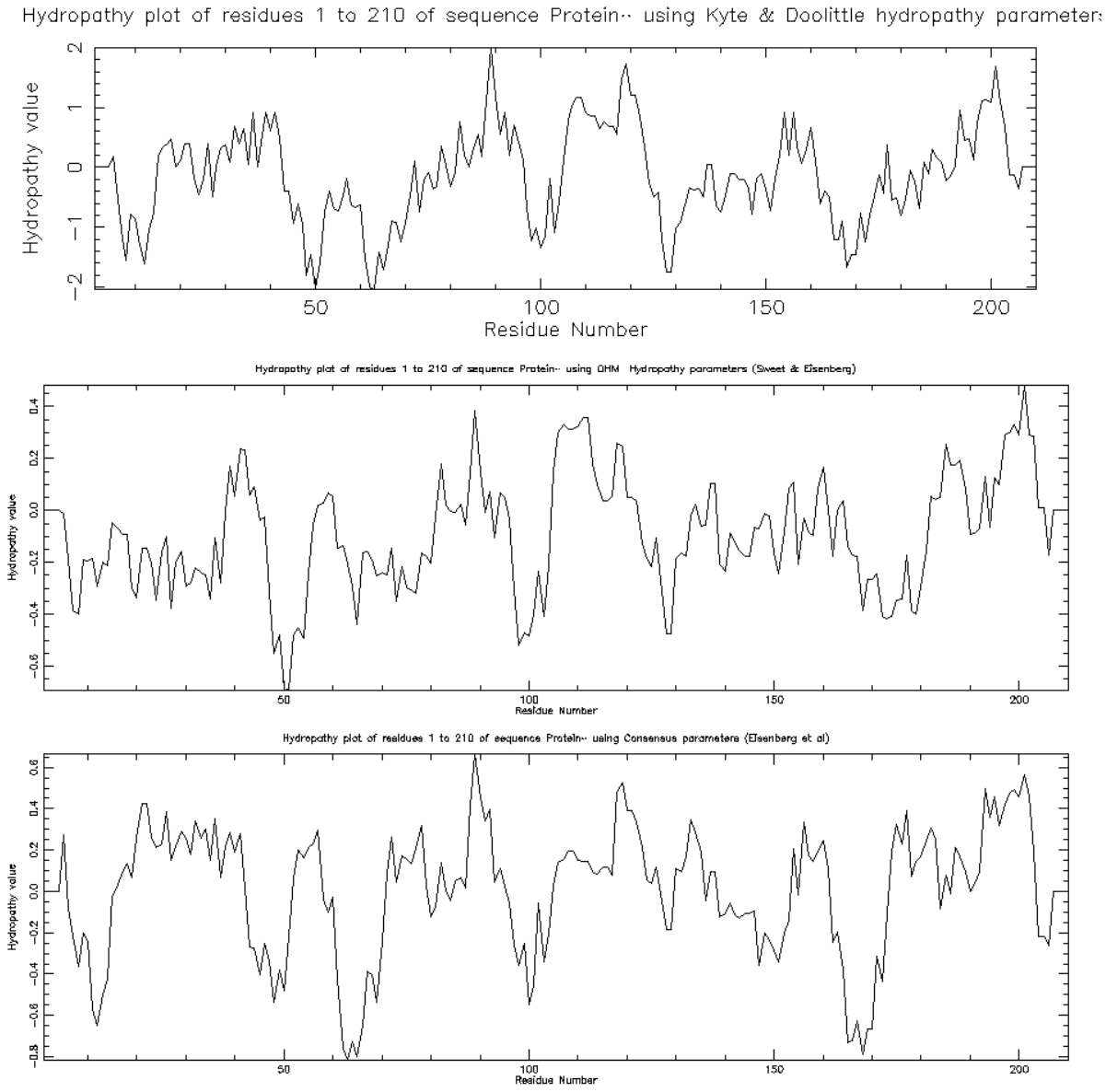


Fig4: Hydropathy Plots of the protein of strain of Rhodopseudomonas faecalis using EMBOSS Pepinfo

S.NO	Aminoacid	Polarity	Hydropathicity	Refractivity	Hydrophobicity (delta G1/2 cal)	Transmembrane Tendency	Average flexibility index
1	Ala	0.000	1.800	4.340	0.440	0.380	0.360
2	Arg	52.000	-4.500	26.660	-2.420	-2.570	0.530
3	Asn	3.380	-3.500	13.280	-1.320	-1.620	0.460

4	Asp	49.700	-3.500	12.000	-0.310	-3.270	0.510
5	Cys	1.480	2.500	35.770	0.580	-0.300	0.350
6	Gln	3.530	-3.500	17.560	-0.710	-1.840	0.490
7	Glu	49.900	-3.500	17.260	-0.340	-2.900	0.500
8	Gly	0.000	-0.400	0.000	0.000	-0.190	0.540
9	His	51.600	-3.200	21.810	-0.010	-1.440	0.320
10	Ile	0.130	4.500	19.060	2.460	1.970	0.460
11	Leu	0.130	3.800	18.780	2.460	1.820	0.370
12	Lys	49.500	-3.900	21.290	-2.450	-3.460	0.470
13	Met	1.430	1.900	21.640	1.100	1.400	0.300
14	Phe	0.350	2.800	29.400	2.540	1.980	0.310
15	Pro	1.580	-1.600	10.930	1.290	-1.440	0.510
16	Ser	1.670	-0.800	6.350	-0.840	-0.530	0.510
17	Thr	1.660	-0.700	11.010	-0.410	-0.320	0.440
18	Trp	2.100	-0.900	42.530	2.560	1.530	0.310
19	Tyr	1.610	-1.300	31.530	1.630	0.490	0.420
20	Val	0.130	4.200	13.920	1.730	1.460	0.390

Table 1: Scale Value of individual aminoacids of protein of strain of *Rhodopseudomonas faecalis* using ExPASy ProtScale

ExPASy ProtScale was run for the aminoacid sequence of strain of *Rhodopseudomonas faecalis*, wherein the parameters shows certain scale values of aminoacids observed for Polarity: a maximum of 52.000 for Arg and minimum of 0.000 for Gly and Ala. Hydropathicity : maximum of 4.500 for Ile and minimum is -4.500 for Arg. Refractivity :a maximum of 42.530 is for Trp and minimum of 0.000 for Gly. Hydrophobicity (delta G1/2 cal) : a maximum of 2.560 is for Trp and minimum of -2.540 for Lys. Transmembrane Tendency: maximum of 1.980 for Phe and minimum of -3.270 for Asp. Average flexibility index: a maximum of 0.540 for Gly and minimum of 0.300 for Met respectively as shown in Table1.

Secondary structure predictions of protein

The Chou Fasman program reveals that there are out of 210 residues 113 are helix, 127 are sheets and 29 are Turns with 53.8%, 60.5% and 13.8% respectively as shown in Fig5.

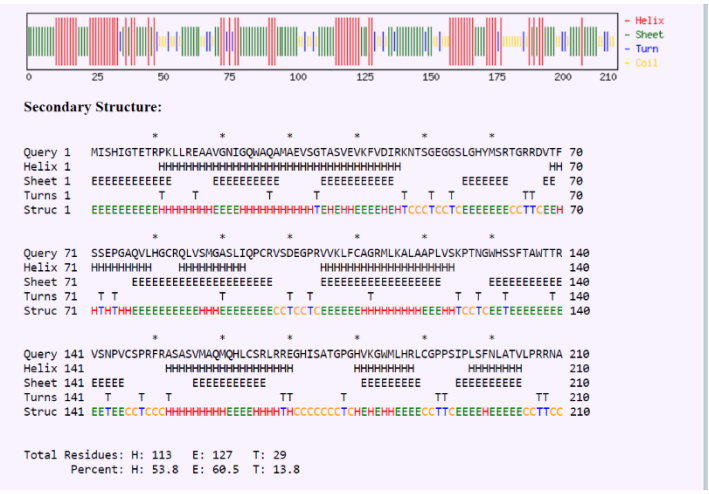


Fig 5: Chou-Fasman program of the aminoacid sequence of strain of *Rhodopseudomonas faecalis*

From Fig 6, the GOR program have generated the prediction of the structure of alpha helix and beta sheets for the entire sequence with scale value high for sheets at -450, -380 and 240, 200 below and above the '0' point of the graph respectively.

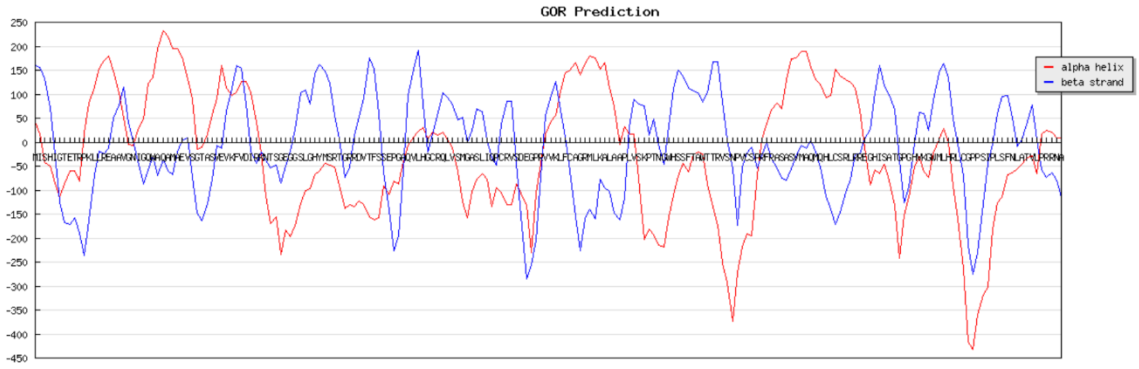
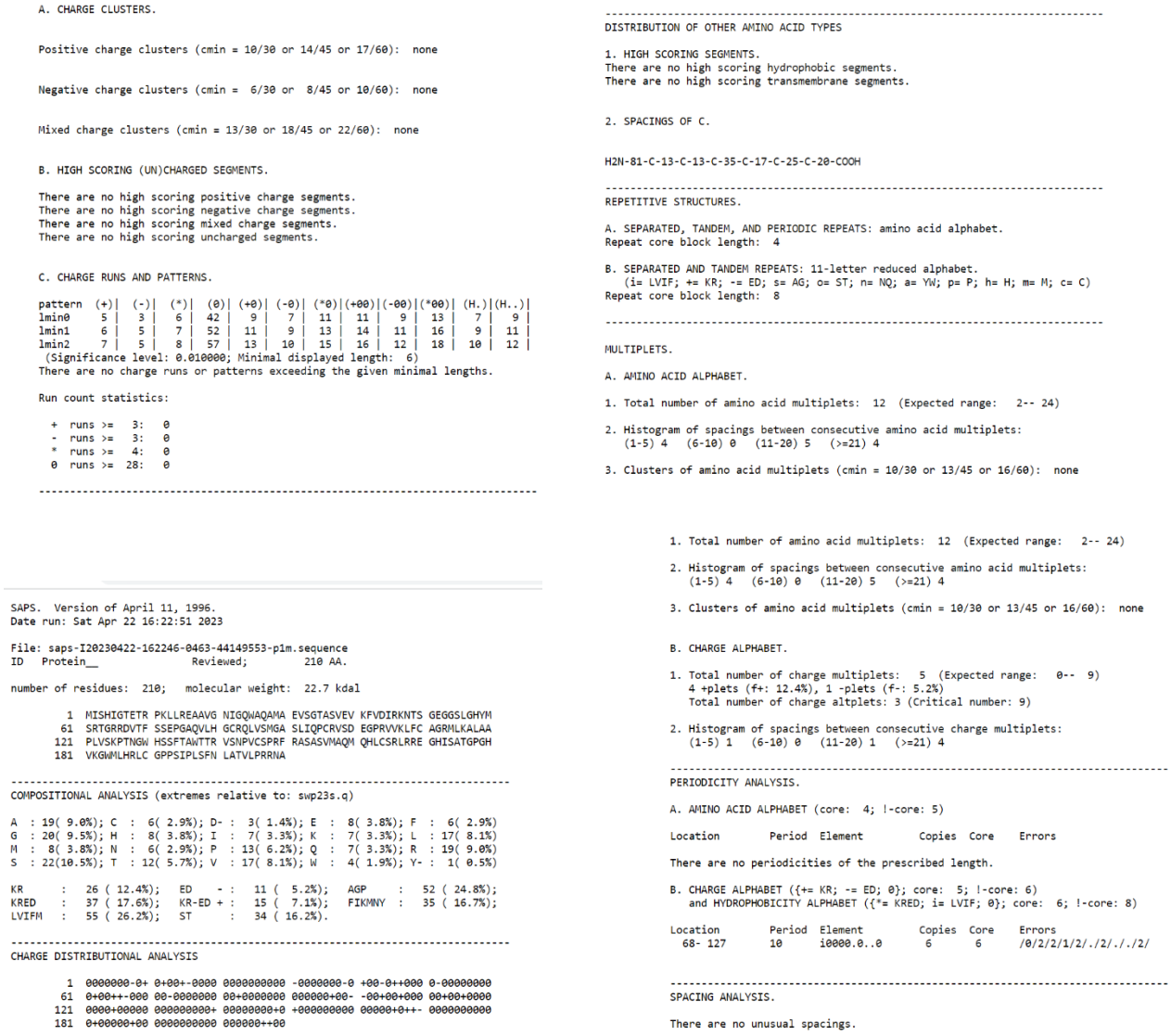


Fig 6: GOR prediction for aminoacid sequence of strain of *Rhodopseudomonas faecalis* showing alpha helix and beta strands



```
A. CHARGE CLUSTERS.

Positive charge clusters (cmin = 10/30 or 14/45 or 17/60):  none

Negative charge clusters (cmin =  6/30 or  8/45 or 10/60):  none

Mixed charge clusters (cmin = 13/30 or 18/45 or 22/60):  none

B. HIGH SCORING (UN)CHARGED SEGMENTS.

There are no high scoring positive charge segments.
There are no high scoring negative charge segments.
There are no high scoring mixed charge segments.
There are no high scoring uncharged segments.

C. CHARGE RUNS AND PATTERNS.

pattern  (+)|  (-)|  (*)|  (0)| (+0)| (-0)| (*0)|(+00)|(-00)|(*00)| (H.)|(H..)|
lmin0     5 |   3 |   6 |  42 |   9 |   7 |  11 |  11 |   9 |  13 |   7 |   9 |
lmin1     6 |   5 |   7 |  52 |  11 |   9 |  13 |  14 |  11 |  16 |   9 |  11 |
lmin2     7 |   5 |   8 |  57 |  13 |  10 |  15 |  16 |  12 |  18 |  10 |  12 |
 (Significance level: 0.010000; Minimal displayed length:  6)
There are no charge runs or patterns exceeding the given minimal lengths.

Run count statistics:

  +  runs >=   3:   0
  -  runs >=   3:   0
  *  runs >=   4:   0
  0  runs >=  28:   0

--------------------------------------------------------------------------------
DISTRIBUTION OF OTHER AMINO ACID TYPES

1. HIGH SCORING SEGMENTS.
There are no high scoring hydrophobic segments.
There are no high scoring transmembrane segments.

2. SPACINGS OF C.

H2N-81-C-13-C-13-C-35-C-17-C-25-C-20-COOH

--------------------------------------------------------------------------------
REPETITIVE STRUCTURES.

A. SEPARATED, TANDEM, AND PERIODIC REPEATS: amino acid alphabet.
Repeat core block length:  4

B. SEPARATED AND TANDEM REPEATS: 11-letter reduced alphabet.
   (i= LVIF; += KR; -= ED; s= AG; o= ST; n= NQ; a= YW; p= P; h= H; m= M; c= C)
Repeat core block length:  8

--------------------------------------------------------------------------------
MULTIPLETS.

A. AMINO ACID ALPHABET.

1. Total number of amino acid multiplets:  12  (Expected range:   2-- 24)

2. Histogram of spacings between consecutive amino acid multiplets:
   (1-5) 4   (6-10) 0   (11-20) 5   (>=21) 4

3. Clusters of amino acid multiplets (cmin = 10/30 or 13/45 or 16/60):  none

SAPS.  Version of April 11, 1996.
Date run: Sat Apr 22 16:22:51 2023

File: saps-I20230422-162246-0463-44149553-p1m.sequence
ID   Protein__                Reviewed;         210 AA.

number of residues:  210;   molecular weight:  22.7 kdal

        1  MISHIGTETR PKLLREAAVG NIGQWAQAMA EVSGTASVEV KFVDIRKNTS GEGGSLGHYM
       61  SRTGRRDVTF SSEPGAQVLH GCRQLVSMGA SLIQPCRVSD EGPRVVKLFC AGRMLKALAA
      121  PLVSKPTNGW HSSFTAWTTR VSNPVCSPRF RASASVMAQM QHLCSRLRRE GHISATGPGH
      181  VKGWMLHRLC GPPSIPLSFN LATVLPRRNA

--------------------------------------------------------------------------------
COMPOSITIONAL ANALYSIS (extremes relative to: swp23s.q)

A  : 19( 9.0%); C  :  6( 2.9%); D- :  3( 1.4%); E  :  8( 3.8%); F  :  6( 2.9%)
G  : 20( 9.5%); H  :  8( 3.8%); I  :  7( 3.3%); K  :  7( 3.3%); L  : 17( 8.1%)
M  :  8( 3.8%); N  :  6( 2.9%); P  : 13( 6.2%); Q  :  7( 3.3%); R  : 19( 9.0%)
S  : 22(10.5%); T  : 12( 5.7%); V  : 17( 8.1%); W  :  4( 1.9%); Y- :  1( 0.5%)

KR      :   26 ( 12.4%);   ED    - :   11 (  5.2%);   AGP     :   52 ( 24.8%);
KRED    :   37 ( 17.6%);   KR-ED + :   15 (  7.1%);   FIKMNY  :   35 ( 16.7%);
LVIFM   :   55 ( 26.2%);   ST      :   34 ( 16.2%).

--------------------------------------------------------------------------------
CHARGE DISTRIBUTIONAL ANALYSIS

        1  0000000-0+ 0+00+-0000 0000000000 -0000000-0 +00-0++000 0-00000000
       61  0+00++-000 00-0000000 00+0000000 000000+00- -00+00+000 00+00+0000
      121  0000+00000 000000000+ 00000000+0 +000000000 00000+0++- 0000000000
      181  0+00000+00 0000000000 000000++00

1. Total number of amino acid multiplets:  12  (Expected range:   2-- 24)

2. Histogram of spacings between consecutive amino acid multiplets:
   (1-5) 4   (6-10) 0   (11-20) 5   (>=21) 4

3. Clusters of amino acid multiplets (cmin = 10/30 or 13/45 or 16/60):  none

B. CHARGE ALPHABET.

1. Total number of charge multiplets:   5  (Expected range:   0--  9)
   4 +plets (f+: 12.4%), 1 -plets (f-: 5.2%)
   Total number of charge altplets: 3 (Critical number: 9)

2. Histogram of spacings between consecutive charge multiplets:
   (1-5) 1   (6-10) 0   (11-20) 1   (>=21) 4

--------------------------------------------------------------------------------
PERIODICITY ANALYSIS.

A. AMINO ACID ALPHABET (core:  4; !-core: 5)

Location        Period  Element         Copies  Core    Errors

There are no periodicities of the prescribed length.

B. CHARGE ALPHABET ({+= KR; -= ED; 0}; core:  5; !-core: 6)
   and HYDROPHOBICITY ALPHABET ({*= KRED; i= LVIF; 0}; core:  6; !-core: 8)

Location        Period  Element         Copies  Core    Errors
  68- 127       10      i0000.0..0       6       6      /0/2/2/1/2/./2/././2/

--------------------------------------------------------------------------------
SPACING ANALYSIS.

There are no unusual spacings.
```

Fig 7 : Statistical Analysis of Protein Sequences (SAPS) for the aminoacid sequence of Rhodopseudomonas faecalis

The Statistical Analysis of Protein Sequences (SAPS) reveals several parameters like Compositional analysis of aminoacids which is maximum for aminoacid S: 22(10.5%) and aminoacids LVIFM :55 (26.2%), Charge distributional analysis with no high scoring hydrophobic and transmembrane segments. Repetitive structures with repeat core block length is 4 and for separated, tandem and periodic repeats & separated and tandem repeats. There are 12 total number of aminoacid multiplets and 5 charged multiplets. Periodic analysis shows that there no periodicities of the prescribed length. There are no unusual spacings as shown in Fig 7.

The ProSA-Web shows the Z-Score as -0.95 as shown in Fig 8 (Z -Score vs Number of residues) and the sequence position is shown in Fig 9 (Knowledge based energy vs Sequence position) along with its structure generated in JSmol visualized for the protein of the strain of *Rhodopseudomonas faecalis* as shown in Fig 10. The The z-score of this model is far too high for a typical native structure. So it requires energy distribution to be optimized for a model with in a range of native conformation by studies using modeling and simulation methods.

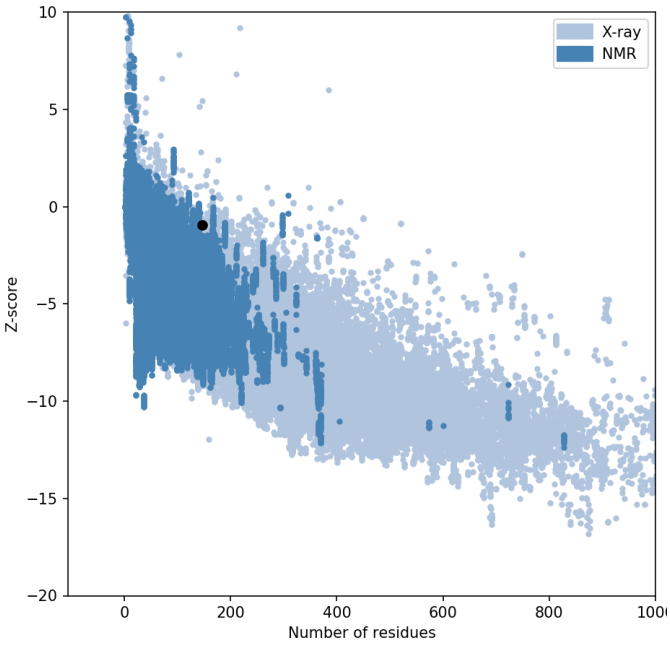


Fig 8: ProSA-web-Overall model quality with Z-score(-0.95)

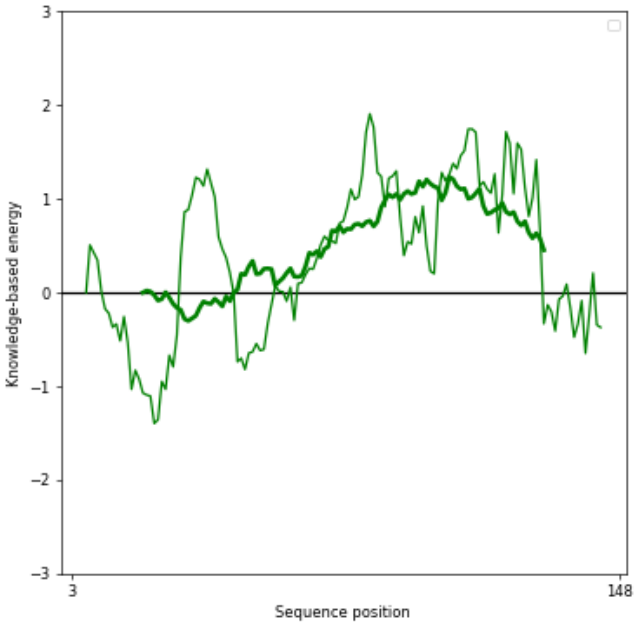


Fig 9 : ProSA-Web-Local Model Quality

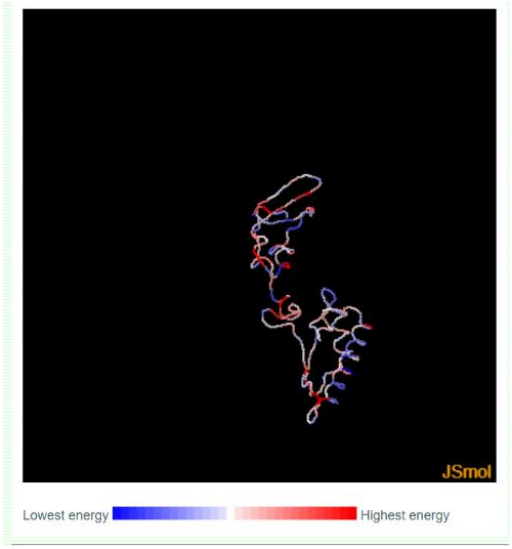


Fig 10: ProSA-web -Visualisation of the protein

PROCHECK was run to show results for ERRAT program with overall quality factor is 36.220. The generally accepted range is >50 for a high quality model. So, the structure need to be simulated through energy minimization. Ramachandran plot with Psi degrees and Phi degrees shows that residues in most favored regions are 82%, residues in additional allowed regions are 13.1% , residues in generously allowed regions are 3.3% and residues in disallowed regions are 1.6%. Further the number of non-glycine and non-proline residues are 122, number of end residues are 1, number of glycine residues are 15 and number of proline residues are 8 as shown in Fig 11 & Fig 12 respectively.

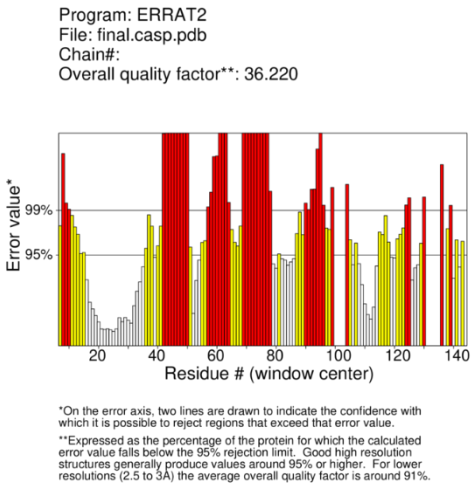


Fig 11: ERRAT Program of the protein

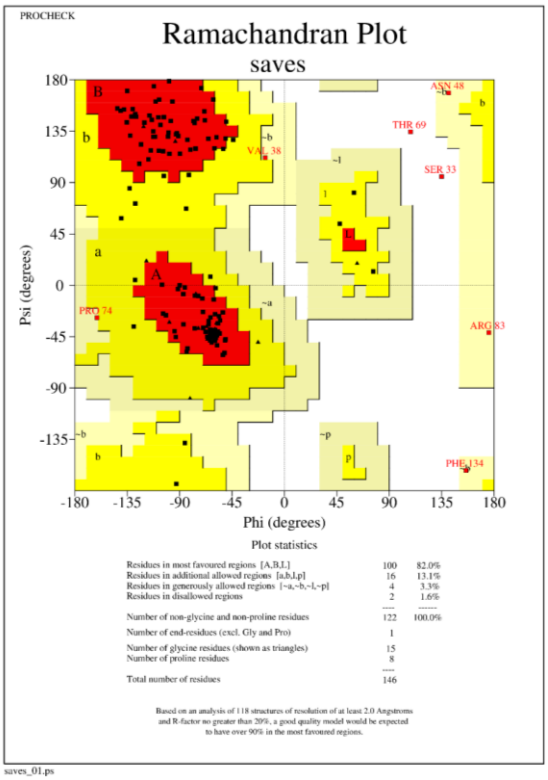
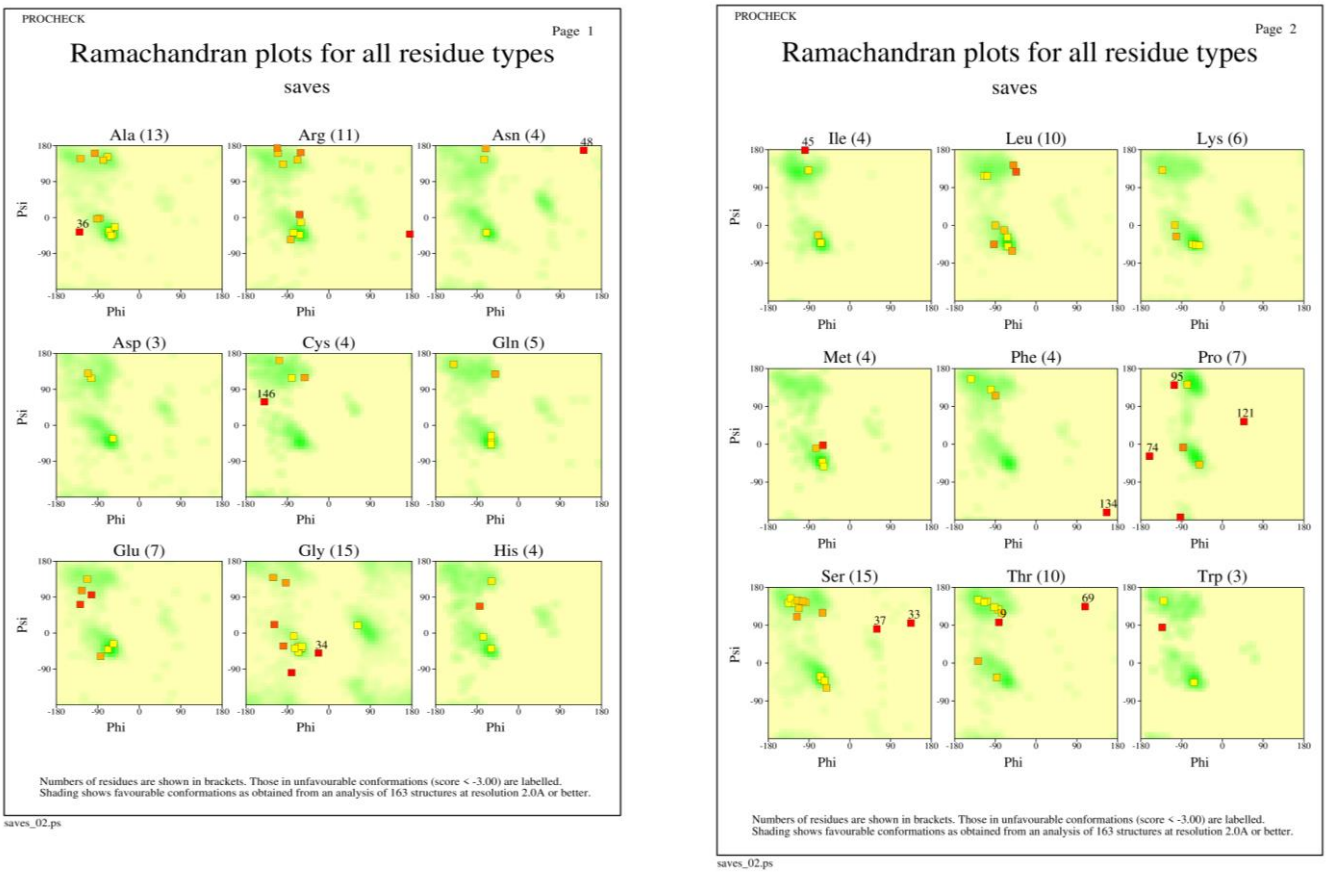


Fig 12: Ramachandran Plot of the protein

The Ramachandran Plots for individual aminoacids of the protein of *Rhodopseudomonas faecalis* were shown in Fig 13 along with their residues showing favourable and unfavourable conformations.



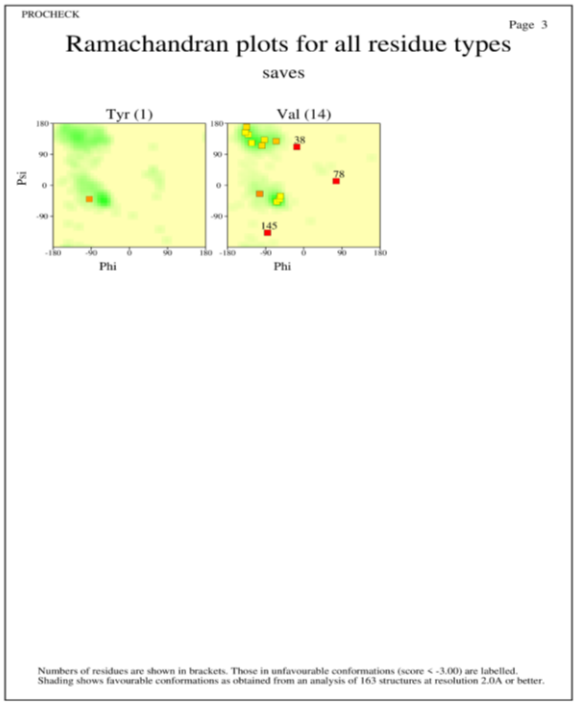


Fig 13 : Ramachandran Plots for the individual aminoacids of the protein

Conclusion

The physio-chemical and structural characterization of the protein of novel strain of *Rhodopseudomonas faecalis* was performed with various bioinformatics tools. The secondary

structure of protein shows that it has maximum number of beta sheets as revealed by Chou Fasman and GOR Programs. Using SAPS, the compositional analysis have shown maximum scale value for Serine, Leucine, Valine, Isoleucine, Phenylalanine and Methionine. Few of the same aminoacids have shown the characteristics like Average flexibility index (Met), Transmembrane tendency (Phe) and hydropathicity (Ile) as evident by ExPASy ProtScale. Besides, Serine has maximum mole% as shown by EMBOSS Pepstats. Ramachandran plot reveals that maximum percentage of the residues are in most favored regions.

References

1. Adhikari S., Manthena P.V., Sajwan K., Kota K.K., Rabindra R. A unified method for purification of basic proteins. Anal Biochem. 2010;400(2):203–206.
2. Aftabuddin M., Kundu S. (2007). Hydrophobic, hydrophilic, and charged amino acid networks within protein. Biophys. J. 93, 225–231. 10.1529/biophysj.106.098004
3. Brooks C. L., Gruebele M., Onuchic J. N., Wolynes P. G. (1998). Chemical physics of protein folding. Proc. Natl. Acad. Sci. United States 95, 11037–11038.
4. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Costanzo LD, Christie C, Duarte JM, Dutta S, Feng Z, Ghosh S, Goodsell DS, Green RK, Guranovic V, Guzenko D, Hudson BP, Liang Y, Lowe R, Peisach E, Periskova I, Randle C, Rose A, Sekharan M, Shao C, Tao Y-P, Valasatava Y, Voigt M, Westbrook J, Young J, Zardecki C, Zhuravleva M, Kurisu G, Nakamura H, Kengaku Y, Cho H, Sato J, Kim JY, Ikegawa Y, Nakagawa A, Yamashita R, Kudou T, Bekker G-J, Suzuki H, Iwata T, Yokochi M, Kobayashi N, Fujiwara T, Velankar S, Kleywegt GJ, Anyango S, Armstrong DR, Berrisford JM, Conroy MJ, Dana JM, Deshpande M, Gane P, Gáborová R, Gupta D, Gutmanas A, Koča J, Mak L, Mir S, Mukhopadhyay A, Nadzirin N, Nair S, Patwardhan A, Paysan-Lafosse T, Pravda L, Salih O, Sehnal D, Varadi M, Vařeková R, Markley JL, Hoch JC, Romero PR, Baskaran K, Maziuk D, Ulrich EL, Wedell JR, Yao H, Livny M & Ioannidis YE (2019). Nucleic Acids Res. 47, D520–D528.
5. Chothia C. (1974). Hydrophobic bonding and accessible surface area in proteins. Nature 248, 338–339.
6. Damjanović A, Schlessman JL, Fitch CA, García AE, García-Moreno E B. Role of flexibility and polarity as determinants of the hydration of internal cavities and pockets in proteins. Biophys J. 2007 Oct 15;93(8):2791-804.
7. Deber C. M., Wang C., Liu L. P., Prior A. S., Agrawal S., Muskat B. L., et al. (2001). Tm finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. Protein Sci. 10, 212–219.
8. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.; Protein Identification and Analysis Tools on the Expasy Server; (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005).pp. 571-607
9. Garnier, J., Gibrat, J. F. & Robson, B. GOR method for predicting protein secondary structure from amino acid sequence. Methods in Enzymology 266, 540–553 (1996).

10. Gill S.C., von Hippel P.H. Calculation of protein extinction coefficients from amino acid sequence data. Anal Biochem. 1989;182(2):319–326.
11. Henderson R, Sali A, Baker ML, Carragher B, Devkota B, Downing KH, Egelman EH, Feng Z, Frank J, Grigorieff N, Jiang W, Ludtke SJ, Medalia O, Penczek PA, Rosenthal PB, Rossmann MG, Schmid MF, Schröder GF, Steven AC, Stokes DL, Westbrook JD, Wriggers W, Yang H, Young J, Berman HM, Chiu W, Kleywegt GJ, Lawson CL (2012). Structure. 20 205–214.
12. Kyte J., Doolittle R. F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.
13. Laskowski R A, MacArthur M W, Moss D S, Thornton J M (1993). PROCHECK - a program to check the stereochemical quality of protein structures. J. App. Cryst., 26, 283-291.
14. Ma Y, Liu Y, Cheng J. Protein Secondary Structure Prediction Based on Data Partition and Semi-Random Subspace Method. Sci Rep. 2018 Jun 29;8(1):9856.
15. Madeira F, Pearce M, Tivey ARN, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research. 2022 Jul;50(W1):W276-W279.
16. Miotto M., Olimpieri P. P., Di Rienzo L., Ambrosetti F., Corsi P., Lepore R., et al. (2018). Insights on protein thermal stability: a graph representation of molecular interactions. Bioinformatics 35, 2569–2577. 10.1093/bioinformatics/bty1011
17. Moret M., Zebende G. (2007). Amino acid hydrophobicity and accessible surface area. Phys. Rev. E Stat. Nonlin Soft Matter Phys. 75, 011920. 10.1103/PhysRevE.75.011920
18. Prevelige, P., Fasman, G.D. (1989). Chou-Fasman Prediction of the Secondary Structure of Proteins. In: Fasman, G.D. (eds) Prediction of Protein Structure and the Principles of Protein Conformation. Springer, Boston, MA. https://doi.org/10.1007/978-1-4613-1571-1_9
19. Ramachandran GN, Ramakrishnan C & Sasisekharan V (1963). Journal of Molecular Biology. 7, 95–99.
20. Read RJ, Adams PD, Arendall WB, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissinel EB, Lütteke T, Otwinowski Z, Perrakis A, Richardson JS, Sheffler WH, Smith JL, Tickle IJ, Vriend G & Zwart PH (2011). Structure. 19, 1395–1412.
21. Rose G. D., Geselowitz A. R., Lesser G. J., Lee R. H., Zehfus M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. Science 229, 834–838.
22. Sippl, M.J. (1993). Recognition of Errors in Three-Dimensional Structures of Proteins. *Proteins* 17, 355-362.
23. Sweet R. M., Eisenberg D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. J. Mol. Biol. 171, 479–488.
24. Tanford C. (1972). Hydrophobic free energy, micelle formation and the association of proteins with amphiphiles. J. Mol. Biol. 67, 59–74. 10.1016/0022-2836(72)90386-5
25. Wiederstein & Sippl (2007), ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research 35, W407-W410.
26. Wilce M. C. J., Aguilar M.-I., Hearn M. T. W. (1995). Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid

hydrophobicity coefficients derived from rp-hplc of peptides. Anal. Chem. 67, 1210–1219.

27. Zhu D, Qian F, Wu Y, Jones DS, Rowe C, Narum DL, Duffy P, Miller LH, Saul A. Determination of protein concentration for protein-protein conjugates using ultraviolet absorption. J Immunol Methods. 2013 Jan 31;387(1-2):317-21.