ProRegIn: A regularity index for the selection of native-like tertiary structures of proteins

Lipi Thukral, Sandhya R Shenoy, Kumkum Bhushan and B Jayaram*

Department of Chemistry and Supercomputing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110 016, India

**Corresponding author (Fax, 91-11-2658 2037; Email, bjayaram@chemistry.iitd.ac.in)*

Automated protein tertiary structure prediction from sequence information alone remains an elusive goal to computational prescriptions. Dividing the problem into three stages viz. secondary structure prediction, generation of plausible main chain loop dihedrals and side chain dihedral optimization, considerable progress has been achieved in our laboratory (*http://www.scfbio-iitd.res.in/bhageerath/index.jsp*) and elsewhere for proteins with less than 100 amino acids. As a part of our on-going efforts in this direction and to facilitate tertiary structure selection/rejection in containing the combinatorial explosion of trial structures for a specified amino acid sequence, we describe here a web-enabled tool *ProRegIn* (Protein Regularity Index) developed based on the regularity in the *Φ, Ψ* dihedral angles of the amino acids that constitute loop regions. We have analysed the dihedrals in loop regions in a non-redundant dataset of 7351 proteins drawn from the Protein Data Bank and categorized them as helix-like or sheet-like (regular) or irregular. We noticed that the regularity thus defined exceeds 86% for *Φ* barring glycine and 70% for *Ψ* for all the amino acid side chains including glycine, compelling us to reexamine the conventional view that loops are irregular regions structurally. The regularity index is presented here as a simple tool that finds its application in protein structure analysis as a discriminatory scoring function for rapid screening before the more compute intensive atomic level energy calculations could be undertaken. The tool is made freely accessible over the internet at *www.scfbio-iitd.res.in/software/proregin.jsp.*

[Thukral L, Shenoy S R, Bhushan K and Jayaram B 2006 *ProRegIn*: A regularity index for the selection of native-like tertiary structures of proteins; *J. Biosci.* **32** 71–81]

1. Introduction

In recent years, theoretical protein structure prediction techniques have advanced rapidly, providing a deeper understanding of the forces stabilizing the three-dimensional structures of proteins and the attendant energy landscapes. However, prediction of the correct folds based on *ab initio* methods remains a challenging problem. The overall tertiary structure of proteins is dictated by the backbone dihedral angles (Betancourt and Skolnick 2004). Repetitive patterns in dihedral angles are indicative of the protein secondary structures such as *α*-helices and *β*-sheets (Creighton 1996). Non-repetitive conformational regions are loops connecting regular secondary structures. In computational prediction methods, loops are considered to be a major area for improvement as they often limit the prediction quality (Jacobson *et al* 2004). They are the most difficult and error prone regions of a protein to solve by X-ray crystallography and the hardest regions to model using knowledge or energy based procedures (Donate *et al* 1996).

There have been many attempts to classify loop regions in proteins according to various common/conserved features (Sibanda and Thornton 1985; Milner-White and Poet 1986; Sibanda *et al* 1989; Efimov 1991; Ring *et al* 1992; Donate *et al* 1996; Wintjens *et al* 1996; Li and Liu 1999). Leszezynski and Rose (1986) defined a sub-class of structurally similar loops called omega (Ω)-loops. Ring *et al* (1992) categorized loops up to 20 residues in length into either linear (strap

Keywords. Dihedral angles of loops; protein data bank; protein tertiary structure selection; regularity index; scoring function

Supplementary Data pertaining to this article is available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/jan2007/pp71-81-suppl.pdf* loops), non-linear and planar (Ω)-loops or non-linear and non-planar (ζ)-loops. Martin *et al* (1995) defined loops as either open or closed depending upon whether the adjoining secondary structures are too far apart from each other to make contact or not. Donate *et al* (1996) classified loops according to their length, type of bounding secondary structures and the main chain conformation of the loops. Kwasigroch and coworkers (1996) have described a database of loops of length three to eight residues clustered according to the length of the loops. A loop prediction method based on metrics has been described by Wojcik *et al* (1999). Their analyses show that there are distinct preferences for residues close to the adjacent secondary structures with residues in the middle of the loop having greater variation in both sequence and structure.

Many methods have been described that improve the accuracy of loop predictions. These include systematic searches of conformational space (Bruccoleri *et al* 1988; Sudarsanam *et al* 1995), searching for fragments which fit the end points of the secondary structures (Jones and Thirup 1986; Sutcliffe *et al* 1987; Blundell *et al* 1988; Claessens *et al* 1989), energy based methods (Bruccoleri and Karplus 1987), molecular dynamics (Bruccoleri and Karplus 1990) and combinations of these methods (Martin *et al* 1989).

According to the number of amino acid residues, turns, a subset of loops in proteins can be categorized into *δ*-turn formed by two residues, *γ*-turn by three residues, *β*-turn by four residues, *α*-turn by five residues and *π*-turn by six residues (Richardson 1981). Nearly 80% peptides comprise *β*-turns that are associated with irregular dihedral angle values (Guruprasad *et al* 2003). The *β*-turns (Venkatachalam 1968), have been classified on the basis of backbone dihedral angles. Such turns have been explored in depth and the positional preferences for each amino acid are well defined, both statistically and experimentally (Chou and Fasman 1974; Rose *et al* 1985; Sibanda and Thornton 1985; Dyson *et al* 1988; Milburn *et al* 1987; Wright *et al* 1988; Falcomer *et al* 1992; Sibanda and Thornton 1993; Hutchinson and Thornton 1994; Scully and Hermans 1994; Guruprasad and Rajkumar 2000). *β*-turns represent the largest category of nonrepetitive secondary structures (Rose *et al* 1985). Several classes of *β*-turns have been categorized (Lewis *et al* 1971; Kuntz 1972; Chou and Fasman 1977; Richardson 1981; Ramakrishnan and Soman 1982; Kabsch and Sander 1983; Wilmot and Thornton 1988, 1990; Efimov 1993) with an N-H(i) O=C(i-3) hydrogen bond. The polypeptide chain reverses its direction on adopting this motif, a frequent occurrence in globular proteins.

Pursuing the idea that a specification of all dihedrals in a loop can lead to a coarse grained native-like structure of proteins – optimization of side chain dihedrals leading to a native-like structure with a better resolution, we posed a question as to how nature chooses the main chain loop dihedrals. We then examined the loop dihedrals with the hypothesis that loops are made up of *α*-helix-like and *β*-sheet-like dihedrals, which constitute the *minima* in the conformational space of polypeptide chains, following the seminal work of Ramachandran *et al* (1963). Any value that fell into the *Φ*, *Ψ* space of the repetitive secondary structures was categorized as regular and all other values as irregular. We noticed that the regularity index calculated for all the loop dihedrals touches 86% barring glycine for *Φ* space and 70% for the *Ψ* space. This study attempts a computation of regularity in the *Φ*, *Ψ* dihedral angles of the amino acids in the loop region.

2. Methodology

The loop dihedral angles from 7351 protein structures that share less than 50% sequence identity and were determined by X-ray crystallography, at a resolution of 2.5 Å or better were extracted from the Protein Data Bank (PDB; Berman *et al* 2000). The regions outside the helix and strand as annotated in the PDB files were defined as loops. STRIDE (Frishman and Argos 1995) was also employed in loop categorization and analysis.

The frequency of occurrence of *Φ* and *Ψ* in loop regions for all the 7351 proteins considered are depicted in figure 1 and frequencies of occurrence of Ramachandran angles for each amino acid are presented in Supplementary Data. The dihedral values for *α*-helix and *β*-sheet are conventionally taken to be (-60°,-40°) and (-120°, +120°) respectively. In case of *Φ* a maximum is observed at -60° and -120° while for *Ψ* a clustering around -15° and +150° is clearly discernible from figure 1. This prompted us to redefine the mean values for a classification of loop dihedrals into helix-like and sheet-like regions. Thus for computing the regularity index of loop dihedrals *Φ* and *Ψ* by classifying into helix-like and sheet-like regions, the values of -60°, -15°; -120°, +150° were adopted respectively. Loop dihedrals were categorized into either helix-like (H) or sheet-like (S) classes with an allowable margin of ± 30°. Values that do not fall into either of the above categories were considered to be irregular (I).

The Regularity Index (RI) for any amino acid N in a protein can then be computed (for *Φ* and *Ψ* separately) as follows:

$$\mathrm{RI} = \frac{\text{Number of loop dihedrals of N with Regular values (H+S)}}{\text{Number of occurrences of amino acid N in the loops}} \text{ x } 100.$$

Further analysis was carried out to set a threshold for acceptance/rejection of decoy (non-native) structures. Thresholds for irregular *Φ* and *Ψ* were calculated by normalizing the proteins with respect to the number of amino acids in a protein.

$$\text{Irregular } \Phi/\Psi\ (\%) = \frac{\text{Number of irregular } \Phi/\Psi}{\text{Total number of amino acids}} \text{ x } 100.$$

We have analysed 21326 decoy structures for 25 protein sequences obtained from three different decoy sets (Fisa, Four-state reduced and Rosetta). The irregular loop dihedral percentage would act as a discrete value to select native-like tertiary structures.

3. Results

We have extracted loops from 7351 non redundant proteins obtained from RCSB. The dihedral angles of the residues in the loops were calculated and were divided as regular (helix-like and sheet-like) and irregular for the 20 amino acids. The distribution according to amino acids is given in table 1a for *Φ* and table 1b for *Ψ*. The regularity index for the *Φ* and *Ψ* dihedrals for each amino acid obtained from the data set of 7351 proteins is depicted in figure 2. We have carried out the analysis on all the 7351 proteins with STRIDE as well and found the results to be comparable with those obtained with the secondary structure information taken directly from the PDB. The *Φ* values obtained with

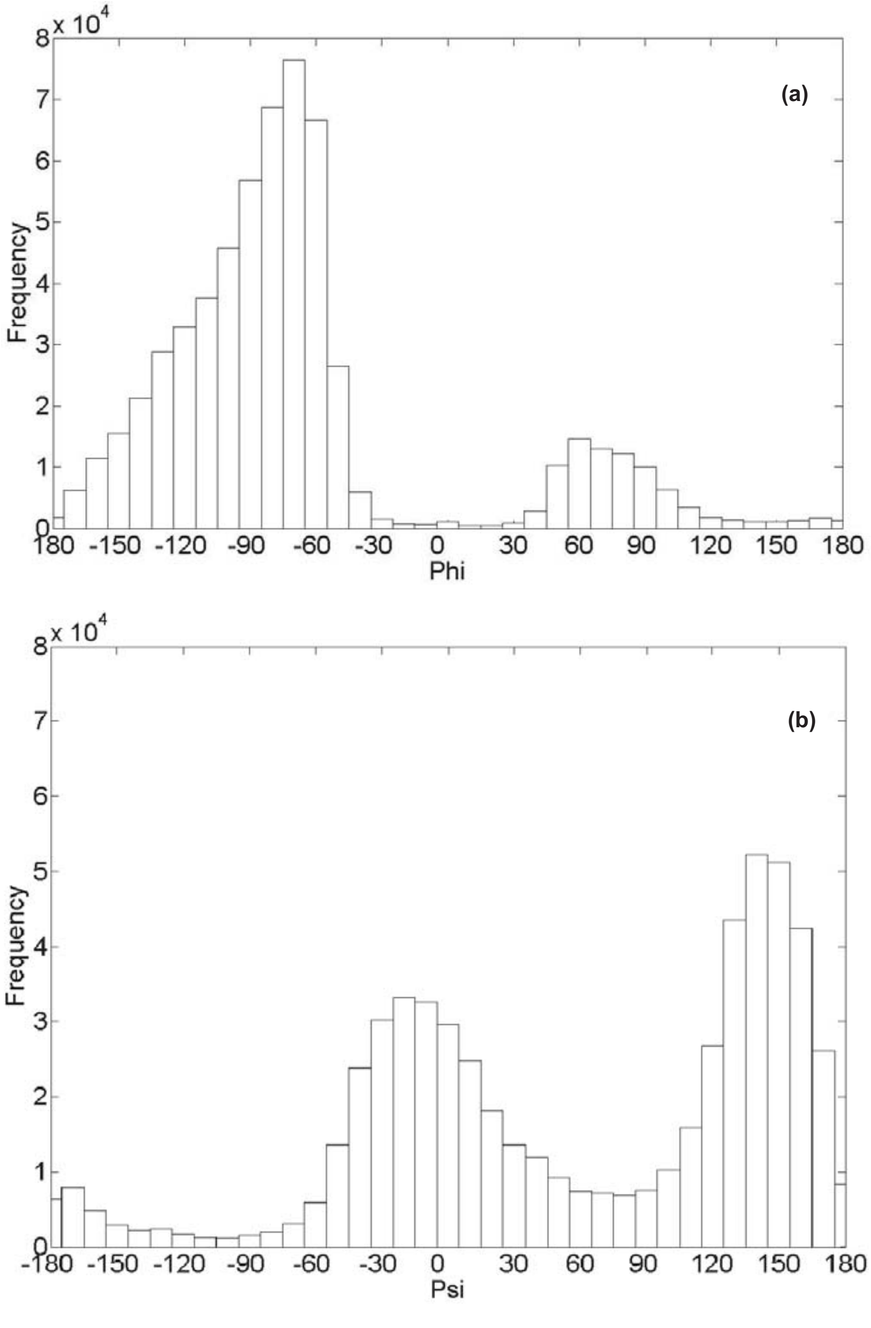


Figure 1. Plot of frequency versus (**a**) *Φ* for 7351 proteins and (**b**) *Ψ* for 7351 proteins.

Table 1a. The distribution of *Φ* dihedral angle in the helical, sheet and irregular region in the loop regions for 7351 proteins.

Amino acid	Total occurences in loops	Total *Φ* in helical range	Percent *Φ* in helical range	Total *Φ* in sheet range	Percent *Φ* in sheet range	Total *Φ* in irregular range	Percent *Φ* in irregular range
ALA	37033	23061	62.27	8834	23.85	5138	13.87
ARG	27066	12243	45.23	10682	39.47	4141	15.30
ASN	34317	11251	32.79	13525	39.41	9541	27.80
ASP	44971	21592	48.01	15757	35.04	7622	16.95
CYS	8535	3494	40.94	3517	41.21	1524	17.86
GLN	19431	8674	44.64	7727	39.77	3030	15.59
GLU	34717	19155	55.17	11181	32.21	4381	12.62
GLY	71875	12118	16.86	8168	11.36	51589	71.78
HIS	14500	5502	37.94	6133	42.30	2865	19.76
ILE	21924	9177	41.86	11530	52.59	1217	5.55
LEU	36508	18782	51.45	14723	40.33	3003	8.23
LYS	34299	16433	47.91	12870	37.52	4996	14.57
MET	8198	3827	46.68	3224	39.33	1147	13.99
PHE	19073	7538	39.52	8763	45.94	2772	14.53
PRO	48953	46992	95.99	1350	2.76	611	1.25
SER	40529	20058	49.49	13426	33.13	7045	17.38
THR	34639	13461	38.86	18273	52.75	2905	8.39
TRP	6650	3206	48.21	2636	39.64	808	12.15
TYR	17350	6769	39.01	8046	46.37	2535	14.61
VAL	29483	12213	41.42	15558	52.77	1712	5.81

STRIDE and PDB are comparable but a small difference is observed in the percentage of irregular *Ψ* obtained. The results obtained with STRIDE are shown in Supplementary Data.

The dihedral values for loops for *Φ* space are regular for all the amino acids (except glycine) with an average of 86%. Proline that has a restricted *Φ* region shows an obvious high of 98% because its side chain is linked to the backbone. *Ψ* values were also found to be regular with an average of 70% for all the amino acids. The above mentioned values (86% for *Φ* and 70% for *Ψ*) were calculated from averaging the summation of entries in column 4 and 6 from table 1a and table 1b for *Φ* and *Ψ* respectively. The loop dihedrals appear to be more regular with a mixture of both helix-like and sheet-like dihedral values.

Further analyses appeared warranted to understand why loops are unable to form regular secondary structures despite a high regularity percentage. Our analysis (figure 3) on the loop dihedral dataset reveals that consecutive occurrence of regular *Φ* and *Ψ* values in loops is limited. The formation of helix requires *i* to *i*+ 4 hydrogen bonds but as observed in figure 3 the uninterrupted occurrence of helix-like dihedrals is limited to four amino acids. The 3_{10} helices are considered as a part of helices and are not included in the loop database. The occurrence of *n* = 3 can be explained based on the *γ*-turns present in the loop regions. The occurrence of a few cases with 11 and more residues in helical conformation as observed in figure 3 is due to the broad range selected for *Φ* and *Ψ*. The residues are not present in helical conformation but are selected as helices according to our classification of helix-like dihedrals. Similarly, sheets are known to form *i* to *i*+2 hydrogen bonds and sheet-like values in loops are not found consecutively for more than two amino acids.

To set a threshold for acceptance/rejection of decoy structures, we have analysed all the 7351 native proteins from PDB with *ProRegIn* using the formula for irregular *Φ*/*Ψ* percentage explained in §2. It was observed that ~ 85% of proteins in our non-redundant protein dataset were included if the irregular *Φ* and *Ψ* percentage threshold was restricted to 1.1% and 4.3% respectively as shown in figure 4. The standard deviation associated with *Φ* is 1.10 and *Ψ* is 1.97 respectively.

Table 1b. The distribution of Ψ dihedral angle in the helical, sheet and irregular region in the loop regions for 7351 proteins..

Amino acid	Total occurences in loops	Total Ψ in helical range	Percent Ψ in helical range	Total Ψ in sheet range	Percent Ψ in sheet range	Total Ψ in irregular range	Percent Ψ in irregular range
ALA	37033	10838	29.265	17783	48.01	8412	22.71
ARG	27066	7976	29.47	11462	42.34	7628	28.18
ASN	34317	9038	26.33	7830	22.81	17449	50.85
ASP	44971	15041	33.44	11039	24.54	18891	42.01
CYS	8535	1810	21.20	4052	47.47	2673	31.32
GLN	19431	5723	29.45	8107	41.72	5601	28.82
GLU	34717	12046	34.69	13620	39.23	9051	26.07
GLY	71875	27827	38.71	12754	17.74	31294	43.54
HIS	14500	3784	26.09	5803	40.02	4913	33.89
ILE	21924	3929	17.92	11611	52.96	6384	29.12
LEU	36508	9701	26.57	17857	48.91	8950	24.51
LYS	34299	10934	31.87	13725	40.01	9640	28.10
MET	8198	2033	24.79	3955	48.24	2210	26.95
PHE	19073	4182	21.92	9304	48.78	5587	29.29
PRO	48953	13883	28.35	28790	58.81	6280	12.83
SER	40529	12893	31.81	18188	44.87	9448	23.31
THR	34639	11660	33.66	15465	44.64	7514	21.69
TRP	6650	1870	28.12	3034	45.62	1746	26.25
TYR	17350	3988	22.98	8511	49.05	4851	27.96
VAL	29483	5284	17.92	15828	53.68	8371	28.39

The threshold numbers give a lower limit to consider a structure as native-like. We have further examined the loop dihedrals in 25 publicly available decoy sets comprising 21326 decoys vis-à-vis their native structures. Table 2 shows the number of structures accepted/rejected based on the threshold values.

The number of decoys which were rejected based on the Φ and Ψ threshold values is 48.5% and 58.8% as seen from columns 5 and 8 respectively of table 2. The threshold for Ψ rejects a larger number of decoys in comparison to the Φ threshold. Overall these results indicate that the regularity index could be of considerable value in assessing protein tertiary structures for their native-like conformation.

3.1 ProRegIn *as a Web-tool*

Based on the observations presented in figure 2, a Web-tool has been created and made freely available at *www.scfbio-iitd.res.in/software/proregin.jsp*. The user inputs the PDB file of a protein. A comprehensive output is presented to the user on the screen with a list of regular/irregular amino acids in the given protein on the basis of regularity index. A snapshot of the front-end of *ProRegIn* is shown in figure 5.

4. Discussion

The connecting regions between helices and sheets are not clearly defined conformationally, despite the considerable time and work devoted towards this difficult research topic. Attempts however have been made to analyse short loops connecting repetitive structures (Fourrier *et al* 2004) and to characterize geometry of repetitive structures in proteins (Bansal *et al* 2000). This investigation focuses on defining loop dihedrals as helix-like or sheet-like. Our study on the loop dihedrals in native proteins reveals that a majority of loop dihedrals are regular i.e. they assume predominantly helix or sheet-like values. Repetitive patterns in dihedral values of Φ and Ψ lead to regular secondary structures. It is the interruption of this repetition, which appears to lead to loops.

The irregularity observed in the loop regions may be attributed to the influence of neighbouring residues and

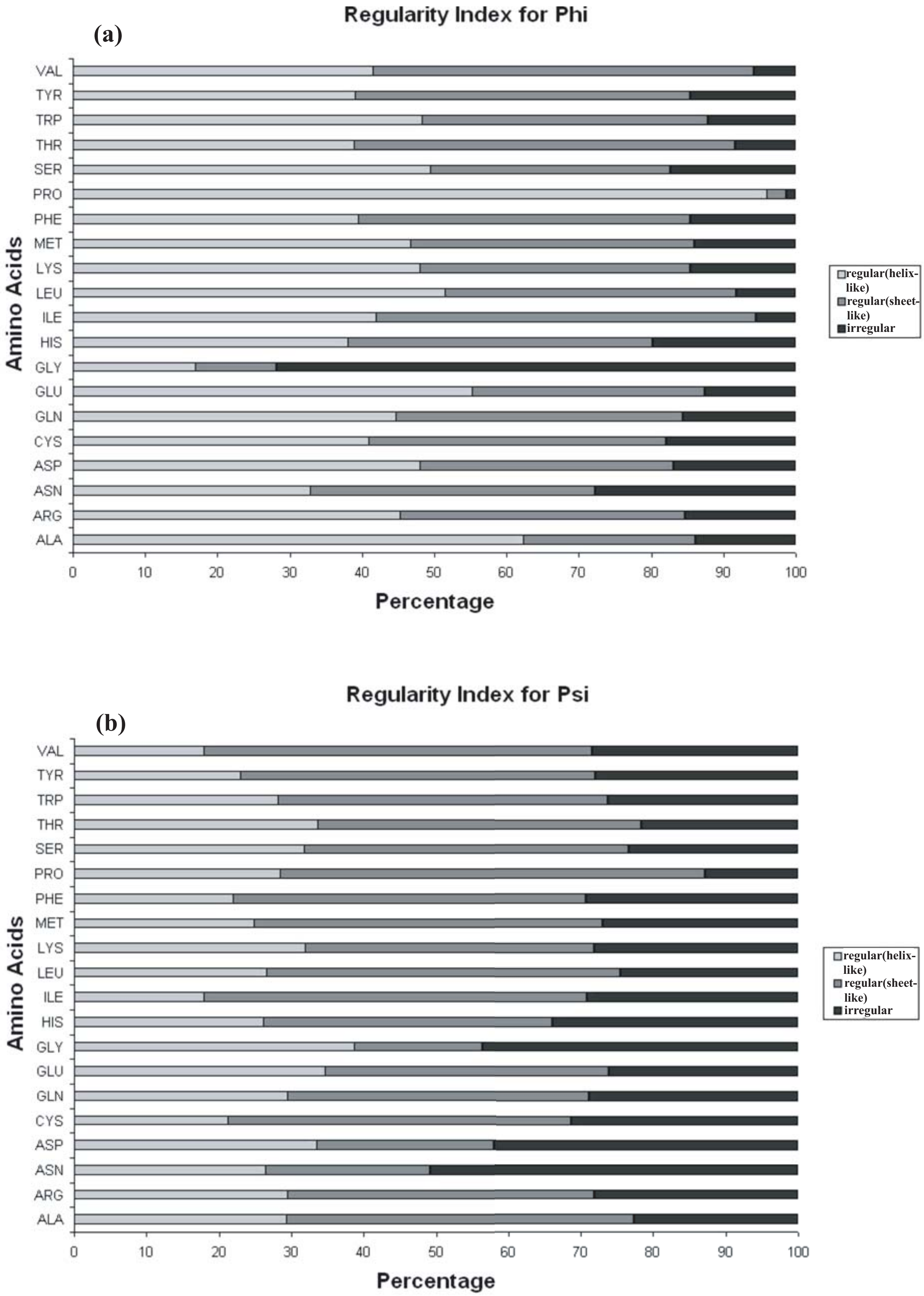


Figure 2. Regularity Index for (**a**) Φ and (**b**) Ψ for all the twenty amino acids.

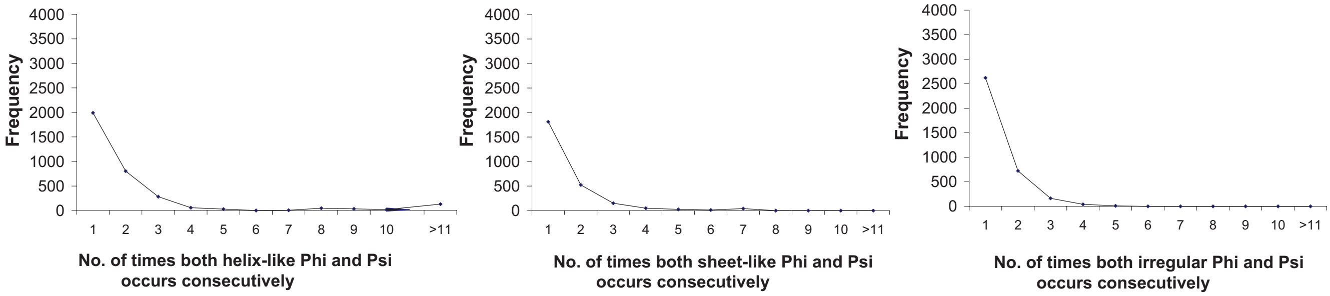


Figure 3. Plot of frequency vs. number of consecutive occurrences of H/S-like and irregular loop dihedrals.

Table 2. Performance appraisal of *ProRegIn* on 21326 decoys (shown as percentage above or below the set thresholds)

		Φ (threshold 1.1%)			*Ψ* (threshold 4.3%)		
Protein ID	Number of decoys	Native	<Threshold	≥Threshold	Native	<Threshold	≥Threshold
Fisa2cro[#]	501	1.5	2.8	97.2	9.2	11.4	88.6
Fisa1hddC[#]	500	0.0	65.6	34.4	10.5	44.4	55.6
Fisa4icb[#]	500	1.3	0	100.0	2.6	89.4	10.6
Fisa1fc2[#]	501	2.3	7.6	92.4	11.6	20.4	79.6
4state1r69[±]	676	0.0	31.1	68.9	1.6	33.6	66.4
4state1sn3[±]	660	1.5	6.5	93.5	3.1	2.4	97.6
4state3icb[±]	654	1.3	13.6	86.4	6.7	81.2	18.8
4state4rxn[±]	352	1.9	0	100.0	1.9	7.4	92.6
1aa2[*]	999	1.0	98.7	1.3	0.0	90.1	9.9
1ail[*]	999	0.0	24.4	75.6	3.0	28.6	71.4
1ayj[*]	999	0.0	38.7	61.3	16	25.6	74.4
1c5a[*]	999	0.0	38.1	61.9	10.8	54.5	45.5
1ddf[*]	999	1.6	45.0	55.0	6.3	66.9	33.1
1fbr[*]	998	0.0	68.2	31.8	6.5	73.2	26.8
1hev[*]	999	4.7	0	100.0	11.6	30.6	69.4
1kte[*]	999	0.0	87.9	12.1	5.0	85.8	14.2
1mbd[*]	999	0.6	93.8	6.2	2.0	99.8	0.2
1nxb[*]	999	1.6	30.1	69.9	6.5	43.7	56.3
1svq[*]	999	0.0	46.2	53.8	4.3	72.8	27.2
1r69[*]	999	0.0	91.2	8.8	6.6	57.1	42.9
1utg[*]	999	1.6	26.9	73.1	0.0	64.3	35.7
1wiu[*]	999	1.1	46.7	53.4	5.4	80.7	19.3
2ezh[*]	999	1.5	72.1	27.9	4.6	48.4	51.6
2gdm[*]	999	0.0	95.8	4.2	3.3	99.8	0.2
2ptl[*]	999	3.8	59.0	41.0	3.8	74.0	26.0
	21326		51.5	48.5		41.2	58.8

[#] *http://dd.stanford.edu/ddownload.cgi?fisa.*

[±] *http://dd.stanford.edu/ddownload.cgi?4state_reduced.*

[*] *http://www.bakerlab.org.*

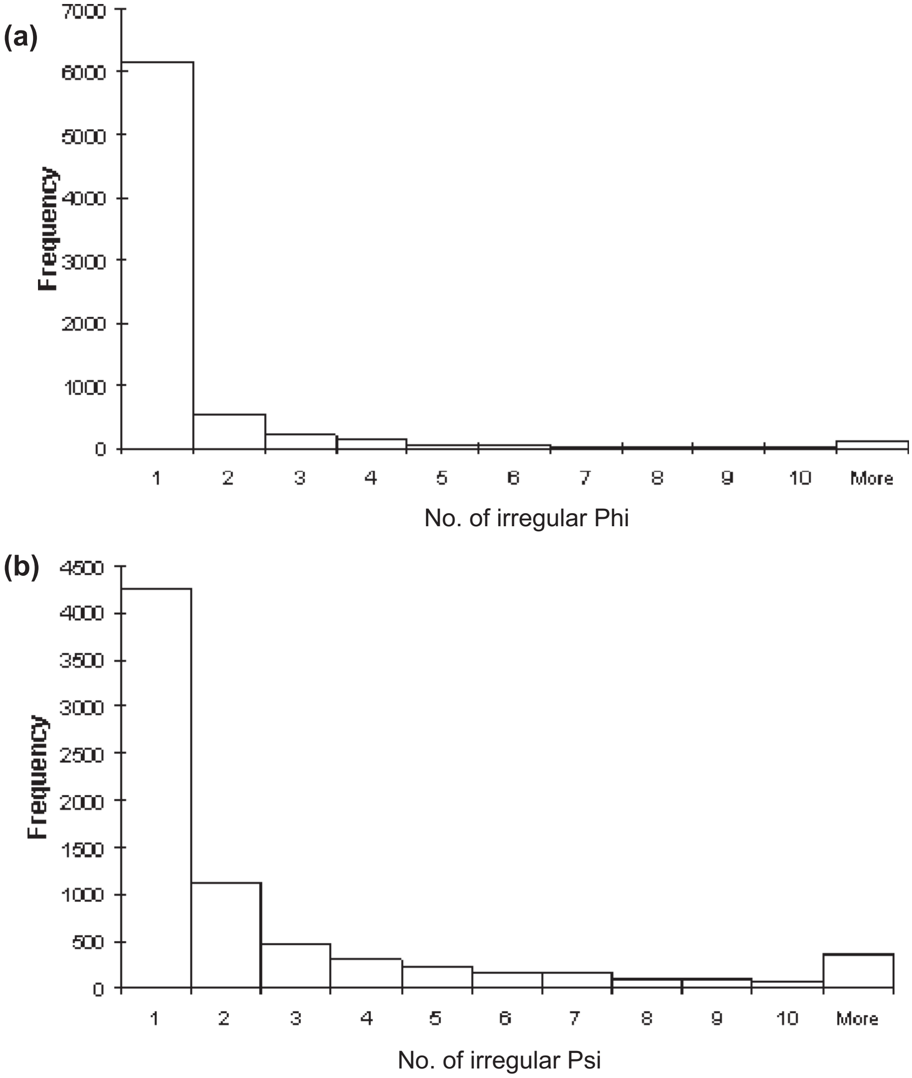
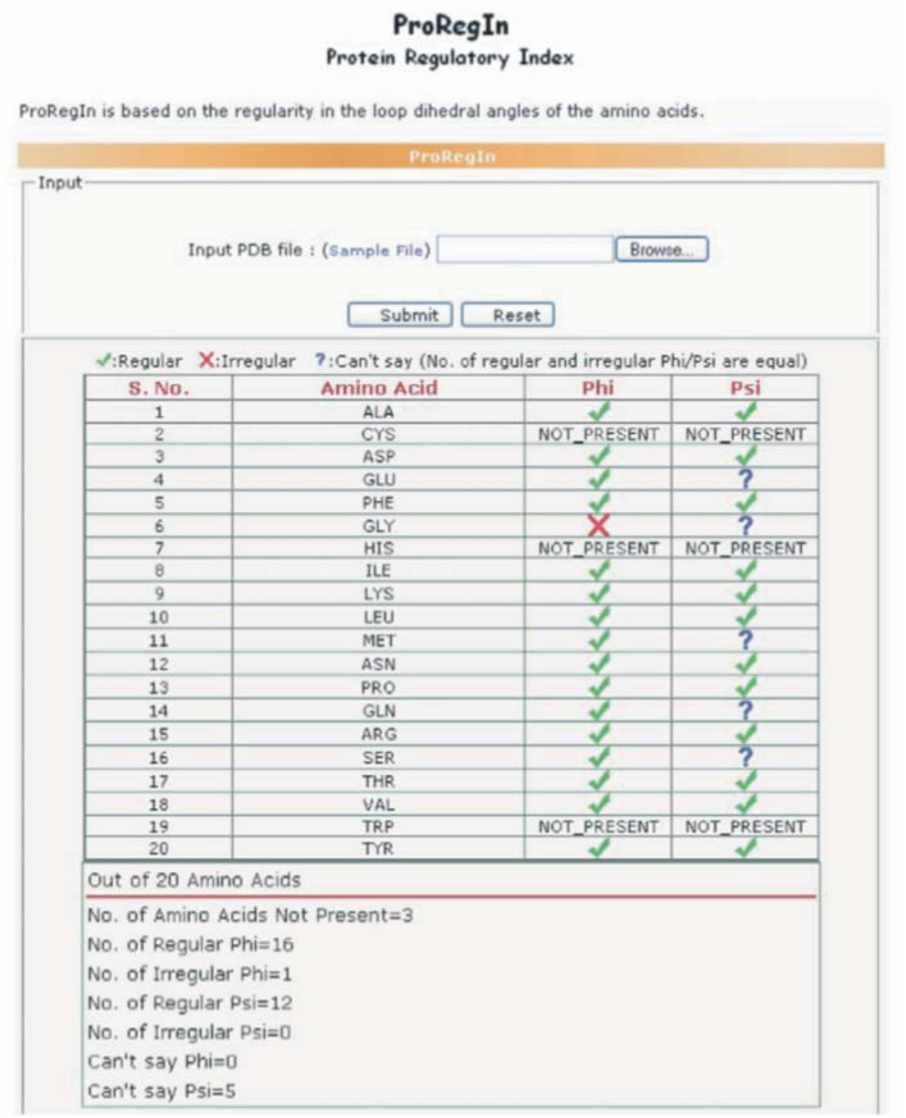


Figure 4. Number of irregular (**a**) *Φ* per 100 and (**b**) *Ψ* per 100 amino acids in 7351 proteins.

environmental conditions, propelling the loops to form connectors between helices and sheets with *Φ* and *Ψ* deviating from either helix-like or sheet-like values. Our findings are consistent with earlier reports in the literature, which demonstrate that although irregular; loops have been shown by many studies not to have completely random backbone conformations (Edwards *et al* 1987; Srinivasan *et al* 1991; Sowdhamini *et al* 1992; Sun and Blundell 1995).

The *ProRegIn* tool presented here could facilitate trial structure generation/acceptance/rejection during modelling of tertiary structures of proteins. This tool in a way is complementary to other protein structure validation tools such as PROCHECK (Laskowski *et al* 1993), Squid (Oldfield 1992), WHATCHECK (Hooft *et al* 1996) and PROVE (Pontius *et al* 1996).

We have previously developed an energy based protein tertiary structure prediction software suite christened *Bhageerath* (*http://www.scfbio-iitd.res.in/bhageerath/index.jsp*) for narrowing down the search space of tertiary structures of small globular proteins (Narang *et al* 2005, 2006). It combines physics based potentials with biophysical filters to arrive at 100 plausible candidate structures starting from sequence and secondary structure information. This is a viable pathway for small globular proteins. For larger proteins however, additional filters are required to narrow



ProRegIn
Protein Regulatory Index

ProRegIn is based on the regularity in the loop dihedral angles of the amino acids.

✓:Regular ✗:Irregular ?:Can't say (No. of regular and irregular Phi/Psi are equal)

S. No.	Amino Acid	Phi	Psi
1	ALA	✓	✓
2	CYS	NOT_PRESENT	NOT_PRESENT
3	ASP	✓	✓
4	GLU	✓	?
5	PHE	✓	✓
6	GLY	✗	?
7	HIS	NOT_PRESENT	NOT_PRESENT
8	ILE	✓	✓
9	LYS	✓	✓
10	LEU	✓	✓
11	MET	✓	?
12	ASN	✓	✓
13	PRO	✓	✓
14	GLN	✓	?
15	ARG	✓	✓
16	SER	✓	?
17	THR	✓	✓
18	VAL	✓	✓
19	TRP	NOT_PRESENT	NOT_PRESENT
20	TYR	✓	✓

Out of 20 Amino Acids

No. of Amino Acids Not Present=3
No. of Regular Phi=16
No. of Irregular Phi=1
No. of Regular Psi=12
No. of Irregular Psi=0
Can't say Phi=0
Can't say Psi=5

Figure 5. A snap-shot of the front-end of web-enabled *ProRegIn.*

down the search space. A tool such as *ProRegIn* should be of considerable value in generating reasonable structures and discarding improbable structures in search of the native. We have found that application of *ProRegIn* followed by topological equivalence allows us to bring down the 100 plausible structures to 10 candidates for the native for small proteins. This option is now integrated with the Bhageerath suite.

Because regularity index spans a large percentage of well-defined Φ, Ψ space, the challenge therefore is to correlate amino acid residue preferences in the context of neighboring residues for assuming helix-like and sheet-like values. The irregular loops could then be fixed using energy based approaches. The regularity index presented here is to assist the researchers to visualize loops from a different perspective and to propose newer strategies for pinning down loop dihedrals.

Acknowledgements

Funding from the Department of Biotechnology, New Delhi is gratefully acknowledged. The authors wish to thank Mr Shailesh Tripathi for web-enabling *ProRegIn* and Ms Sandhya Ganesan for validating *ProRegIn*. LT is a project-trainee from Banasthali Vidyapith, Rajasthan, India. KB is a recipient of the Senior Research Fellow award from the Council of Scientific and Industrial Research (CSIR), New Delhi.

References

Bansal M, Kumar S and Velavan R 2000 HELANAL: A program to characterize helix geometry in proteins; *J. Biomol. Struct. Dyn.* **17** 811–819

Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 The Protein Data Bank; *Nucleic Acids Res*. **28** 235–242

Betancourt M R and Skolnick J 2004 Local propensities and statistical potentials of backbone dihedral angles in proteins; *J. Mol. Biol.* **342** 635–649

Blundell T L, Carney D, *et al* 1988 18th Sir Hans Krebs Lecture. Knowledge-based protein modeling and design; *Eur. J. Biochem*. **172** 513–520

Bruccoleri R E and Karplus M 1987 Prediction of the folding of short polypeptide segments by uniform conformational sampling; *Biopolymers* **26** 137–168

Bruccoleri R E, Haber E, *et al* 1988 Structure of antibody hypervariable loops reproduced by a conformational search algorithm ; *Nature* (*London*) **335** 564–568

Bruccoleri R E and Karplus M 1990 Conformational sampling using high temperature molecular dynamics; *Biopolymers* **29** 1847–1862

Chou P Y and Fasman G D 1974 Conformational parameters for amino acids in helical, beta-sheet and random coil regions calculated from proteins; *Biochemistry* **13** 211–222

Chou P Y and Fasman G D 1977 *β*-turns in proteins; *J. Mol. Biol.* **115** 135–175

Claessens M, Van Cutsem E, *et al* 1989 Modeling the polypeptide backbone with 'spare parts' from known protein structures; *Protein Eng*. **2** 335–345

Creighton T E 1996 *Proteins: Structures and Molecular Properties* 2nd edition (New York: W H Freeman)

Donate L E, Rufino S D, Canard L H J and Blundell T L 1996 Conformational analysis and clustering of short and medium size loops connecting regular secondary structures. A database for modeling and prediction; *Protein Sci*. **5** 2600–2616

Dyson H J, Rance M, Houghten R A, Lerner R A and Wright P E 1988 Folding of immunogenic peptide fragments of proteins in water solution 1. Sequence requirements for the formation of a reverse turn; *J. Mol. Biol.* **201** 161–200

Edwards M S, Sternberg M J E and Thornton J M 1987 Structure and sequence patterns in the loops of *βαβ* units; *Protein Eng*. **1** 173–181

Efimov A V 1991 Structure of alpha-alpha hairpins with short connections; *Protein Eng*. **4** 245–250

Efimov A V 1993 Standard structures in proteins; *Prog. Biophys. Mol. Biol*. **60** 201–239

Falcomer C M *et al* 1992 Chain reversals in model peptides: studies of cystine-containing cyclic peptides 3. Conformational free energies of cyclization of tetrapeptides of sequence Ac-Cys-Pro-X-Cys-NHMe; *J. Am. Chem. Soc*. **114** 4036–4042

Fourrier L, Benros C and de Brevern A G 2004 Use of structural alphabet for analysis of short loops connecting repetitive structures; *BMC Bioinformatics* **5** 58

Frishman D and Argos P 1995 Knowledge-based protein secondary structure assignment; *Proteins* **23** 566–579

Guruprased K and Rajkumar S 2000 *α* and *β*-turns in proteins revisited: A new set of amino acid turn-type dependent positional preferences and potentials; *J. Biosci*. **25** 143–156

Guruprasad K, Rao M J, Adindla S and Guruprasad L 2003 Combinations of turns in proteins; *J. Pept. Res*. **62** 167–174

Hooft R W W, Sander C, Vriend G and Abola E E 1996 Errors in protein structures; *Nature* (*London*) **381** 272

Hutchinson E G and Thornton J M 1994 A revised set of potentials for *β*-turn formation in proteins; *Protein Sci.* **3** 2207–2216

Jacobson M P, Pincus D L , Rappa C S, Day T J F, Honig B, Shaw D E and Friesner R R 2004 A hierarchial approach to all atom protein loop prediction; *Proteins Struct. Funct. Bioinformat.* **55** 351–367

Jones T A and Thirup T 1986 Using known substructures in protein model building and crystallography; *EMBO J*. **5** 819–822

Kabsch W and Sander C 1983 Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features; *Biopolymers* **22** 2577–2637

Kuntz I D 1972 Protein Folding; *J. Am. Chem. Soc.* **94** 4009–4012

Kwasigroch J M, Chomilier J, *et al* 1996 A global taxonomy of loops in globular proteins; *J. Mol. Biol*. **259** 855–872

Laskowski R A, MacArthur M W, Moss D S and Thornton J M 1993 PROCHECK: a program to check the stereochemical quality of protein structures; *J. Appl. Crystallogr.* **26** 283–291

Leszczynski J F and Rose G D 1986 Loops in globular proteins: a novel category of secondary structure; *Science* **234** 849–855

Lewis P N, Momany F A and Scheraga H A 1971 Folding of polypeptide chains in proteins: A proposed mechanism for folding; *Proc. Natl. Acad. Sci.USA* **68** 2293–2297

Li W and Liu Z 1999 Protein loops on structurally similar scaffolds: database and conformational analysis; *Biopolymers* **49** 481–495

Martin A C R, Cheetham J C, *et al* 1989 Modeling antibody hypervariable loops: a combined algorithm; *Proc. Natl. Acad. Sci. USA* **203** 9268–9272

Martin A C R, Toda K, *et al* 1995 Long loops in proteins; *Protein Eng*. **11** 1093–1101

Milburn P J, Konishi Y, Meinwald Y C and Scheraga H A 1987 Chain reversals in model peptides: studies of cysteine-containing cyclic peptides 1. Conformational free energies of cyclization of hexapeptides of sequence Ac-Cys-X-Pro-Gly-Y-Cys-NHMe; *J. Am. Chem. Soc*. **109** 4486–4496

Milner-White E J and Poet R 1986 Four Classes of beta-hairpins in proteins; *J. Mol. Biol*. **238** 733–747

Narang P, Bhushan K, Bose S and Jayaram B 2005 A computational pathway for bracketing native-like structures for small alpha helical globular proteins; *Phys. Chem. Chem. Phys.* **7** 2364–2375

Narang P, Bhushan K, Bose S and Jayaram B 2006 Protein structure evaluation using an all-atom energy based empirical scoring function; *J. Biomol. Str. Dyn.* **23** 385–406

Oldfield T J 1992 SQUID: A program for the analysis and display of data from crystallography and molecular dynamics; *J. Mol. Graphics* **10** 247–252

Pontious J, Richelle J and Wodak S 1996 Deviations from standard atomic values as a quality measure for protein measure for protein crystal structure; *J. Mol. Biol.* **264** 121–136

Ramachandran G N, Ramakrishnan C and Sasisekharan V 1963 Stereochemistry of polypeptide chain configurations; *J. Mol. Biol.* **7** 95–99

Ramakrishnan C and Soman K V 1982 Identification of secondary structures in globular proteins - A new algorithm; *Int. J. Peptide Protein Res*. **20** 218–237

Richardson J S 1981 The anatomy and taxonomy of protein structure; *Adv. Protein Chem*. **34** 1–109

Ring C S, Kneller D G, Langridge R and Cohen F E 1992 Taxonomy and conformational analysis of loops in proteins; *J. Mol. Biol.* **224** 685–699

Rose G, Gierasch L and Smith J 1985 Turns in peptides and proteins; *Adv. Protein Chem.* **37** 1–109

Rufino S D, Donate L E, Canard L and Blundell T L 1996 *BioComputing: Proceedings of the 1996 Pacific Symposium* (eds) Lawrence Hunter and Teri Klein (Singapore: World Scientific)

Scully J and Hermans J 1994 Backbone flexibility and stability of reverse turn conformation in a model system; *J. Mol. Biol.* **235** 682–694

Sibanda B L and Thornton J M 1985 *β*-hairpin families in globular proteins; *Nature* (*London*) **316** 170–174

Sibanda B L, Blundell T L and Thornton J M 1989 Conformation of beta-hairpins in protein structures. A systematic classification with applications to modeling by homology, electron density fitting and protein engineering; *J. Mol. Biol*. **206** 759–777

Sibanda B L and Thornton J M 1993 Accommodating sequence changes in *β*-hairpins in proteins; *J. Mol. Biol*. **229** 428–447

Sowdhamini R, Srinivasan N, Ramakrishnan C and Balram P 1992 Orthogonal *ββ* motifs in proteins; *J. Mol. Biol*. **223** 845–851

Srinivasan N, Sowdhamini R, Ramakrishnan C and Balram P 1991 Analysis of short loops connecting secondary structural elements in proteins; in *Molecular conformation and biological interactions* (eds) C Ramakrishnan and P Balram (Bangalore: Indian Academy of Sciences) 59–73

Sutcliffe M J, Haneef I, *et al* 1987 Knowledge based modeling of homologous proteins, Part I: three dimensional frameworks derived from the simultaneous superposition of multiple structures; *Protein Eng*. **1** 377–384

Sudarsanam S, DuBose R F, *et al* 1995 Modeling protein loops using a Φ_{i+1}, Ψ_i dimmer database; *Protein Sci.* **4** 1412–1420

Sun Z and Blundell T L 1995 *The pattern of common supersecondary structure (motifs) in protein database* (Proceedings of the 28th annual Hawaii international conference on system sciences, USA)

Venkatachalam C M 1968 Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units; *Biopolymers* **6** 1425–1436

Wilmot C M and Thornton J M 1988 Analysis and prediction of the different types of *β*-turns in proteins; *J. Mol. Biol.* **203** 221–232

Wilmot C M and Thornton J M 1990 *β*-turns and their distortions: A proposed new nomenclature; *Protein Eng.* **3** 479–493

Wintjens R T, Rooman M J and Wodak S J 1996 Automatic classification and analysis of alpha-alpha turn motifs in proteins; *J Mol. Biol.* **255** 235–253

Wojcik J, Mornon J P, *et al* 1999 New efficient statistical sequence dependent structure prediction of short to medium sized protein loops based on an exhaustive loop classification; *J. Mol. Biol.* **289** 1469–1490

Wright P E, Dyson H J and Lerner R A 1988 Conformation of peptide fragments of proteins in aqueous solution: implications for initiation of protein folding; *Biochemistry* **27** 7167–7175

*e*Publication: 18 December 2006

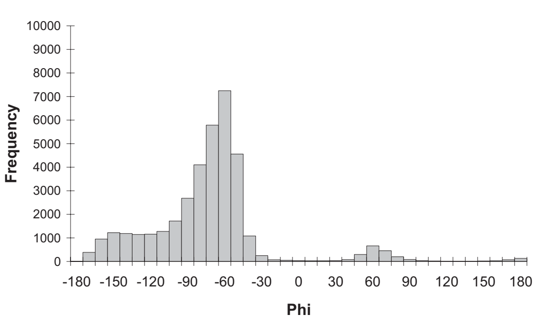
ProRegIn: A regularity index for the selection of native-like tertiary structures of proteins

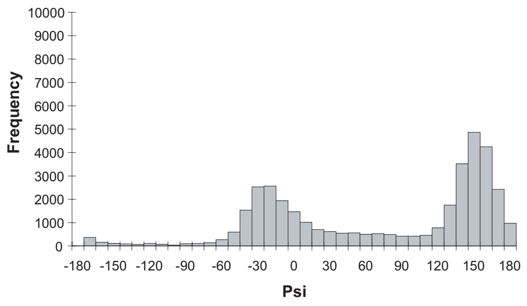
LIPI THUKRAL, SANDHYA R SHENOY, KUMKUM BHUSHAN and B JAYARAM

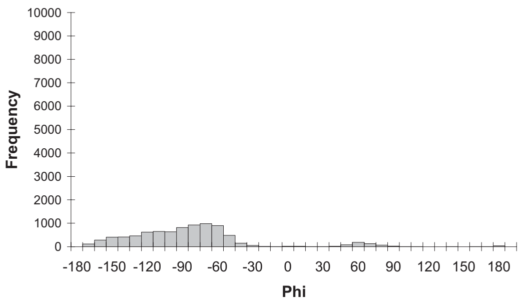
***J. Biosci.* 32(1), January 2007, 71–81, © Indian Academy of Sciences**

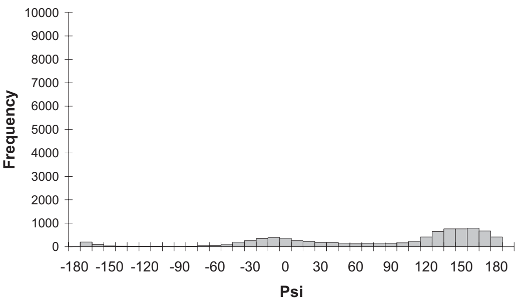
Supplementary Data

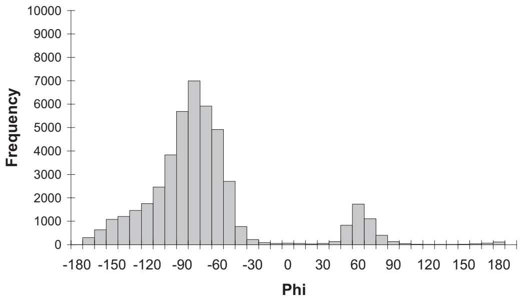
Frequency plots of loop dihedrals of all the amino acids in the entire 7351 non-redundant protein dataset.

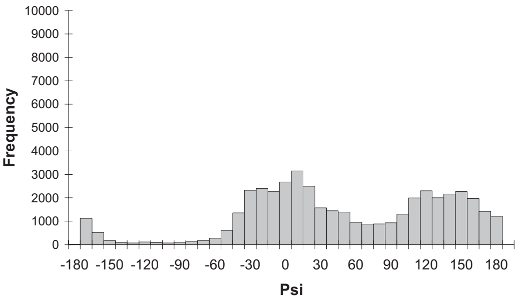

Alanine

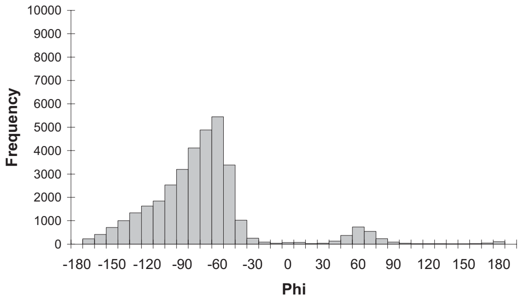

Alanine

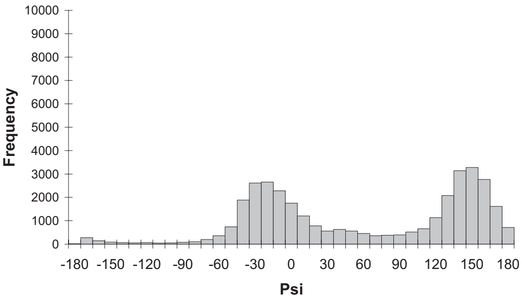

Cysteine


Cysteine

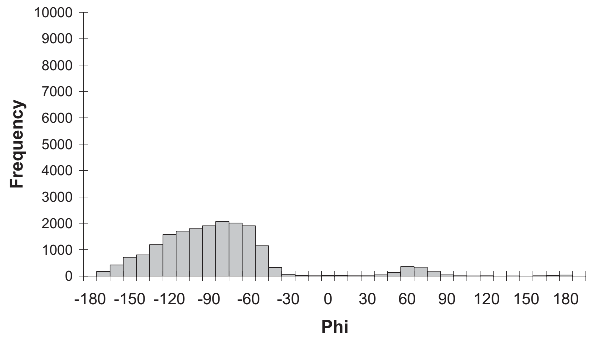

Aspartate


Aspartate

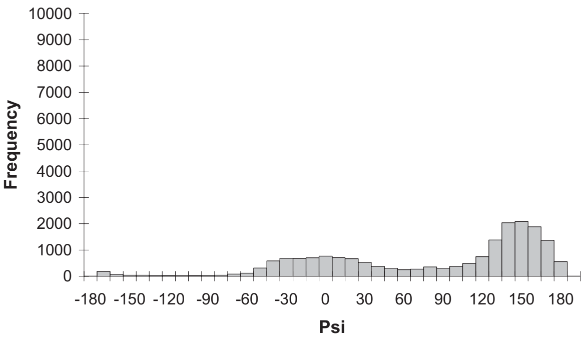

Glutamate


Glutamate

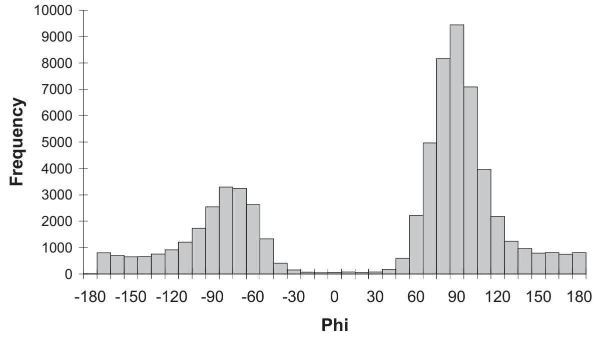
Phenylalanine



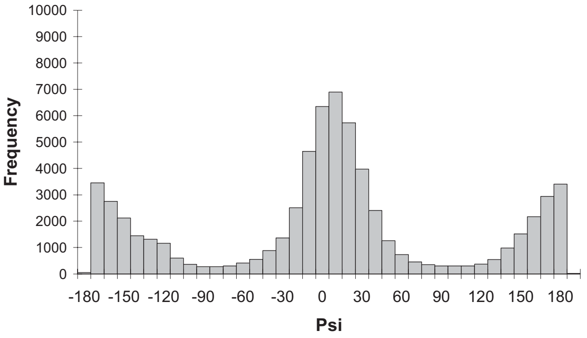
Phenylalanine



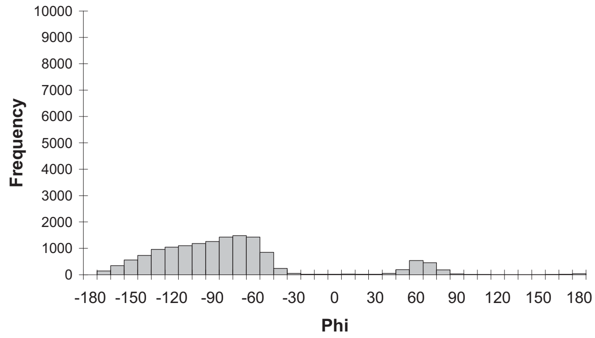
Glycine



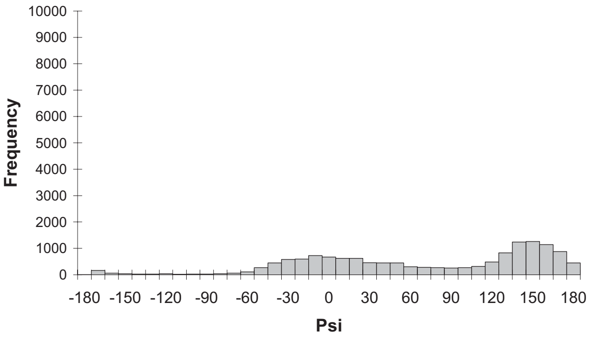
Glycine



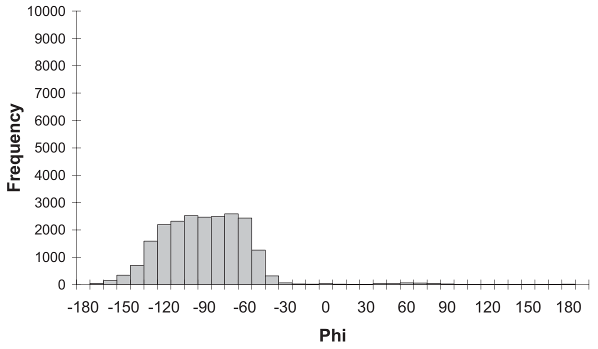
Histidine



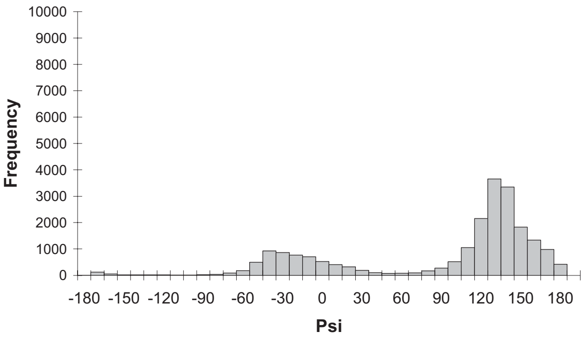
Histidine



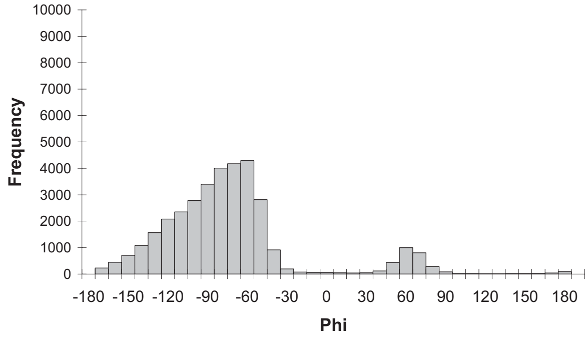
Isoleucine



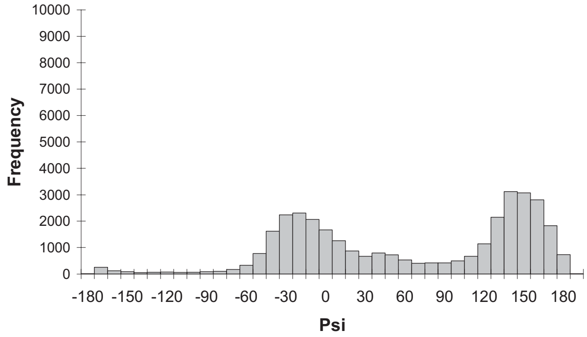
Isoleucine



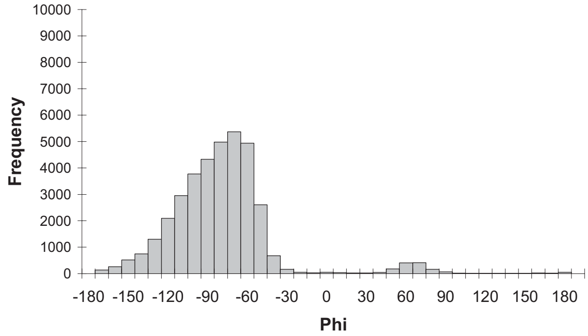
Lysine



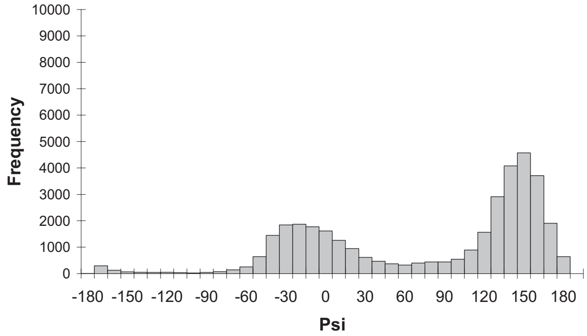
Lysine



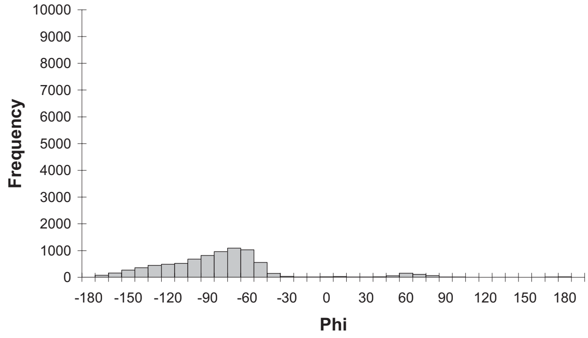
Leucine



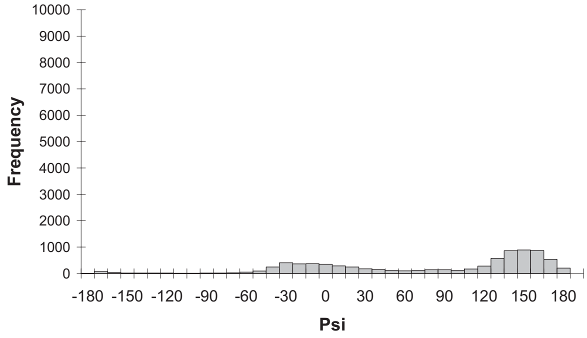
Leucine



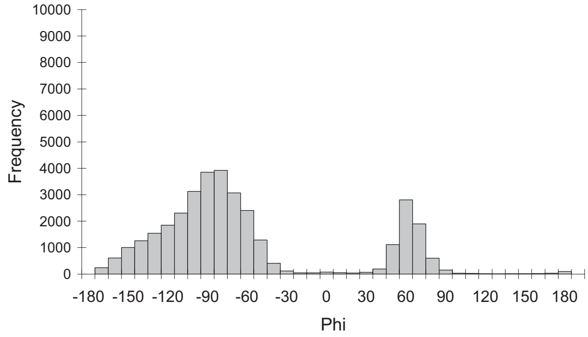
Methionine



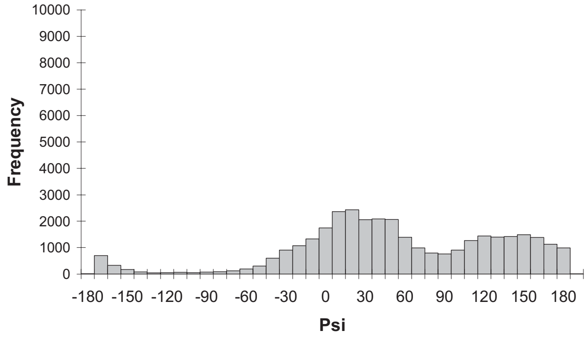
Methionine



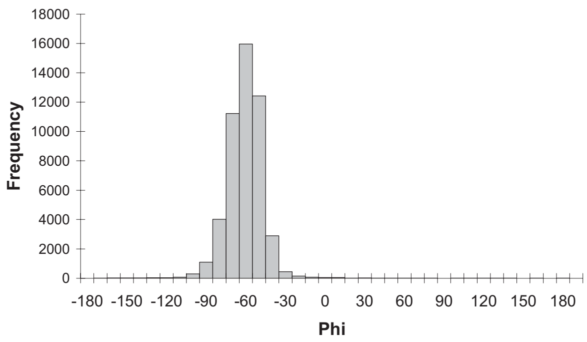
Asparagine



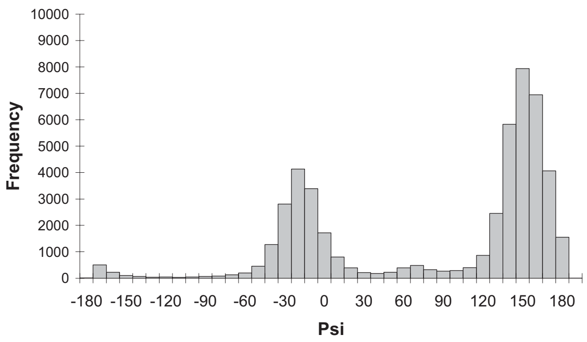
Asparagine



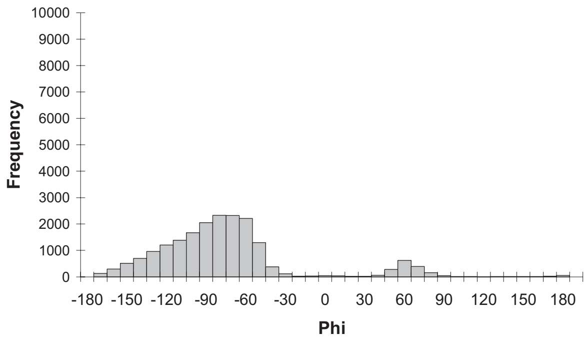
Proline



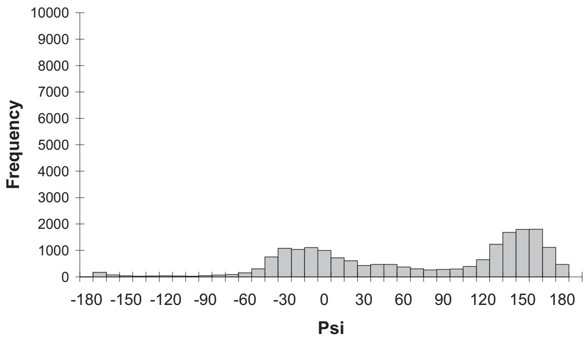
Proline



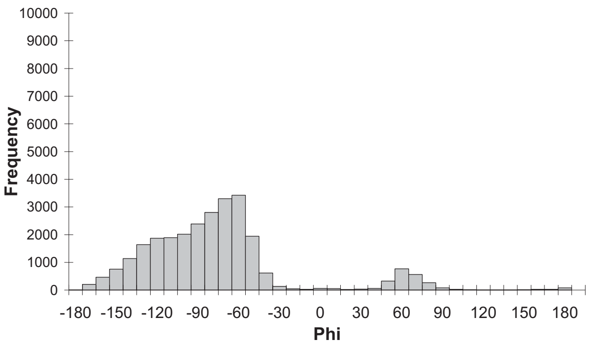
Glutamine



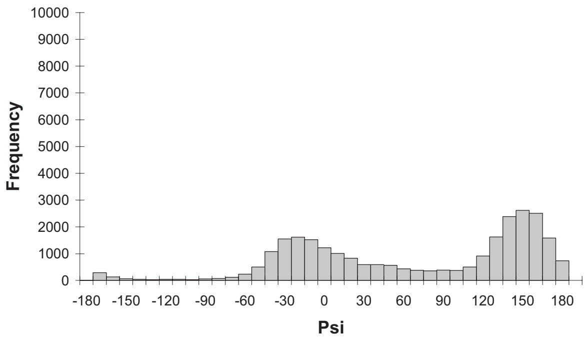
Glutamine



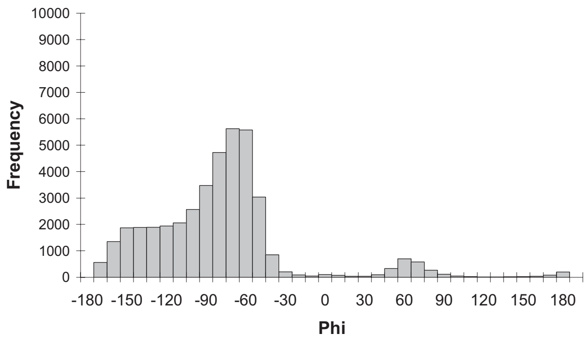
Arginine



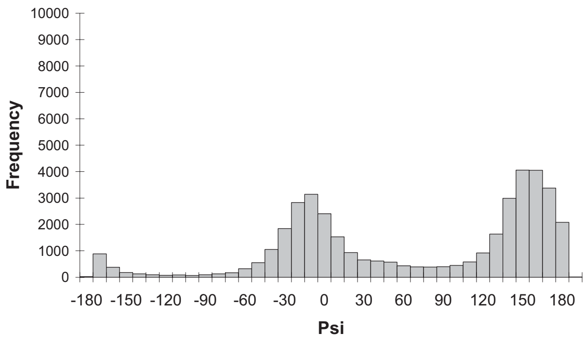
Arginine



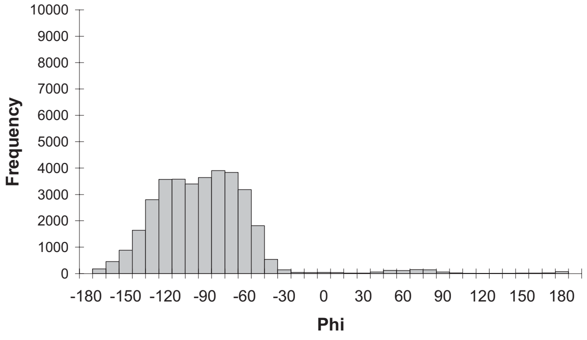
Serine



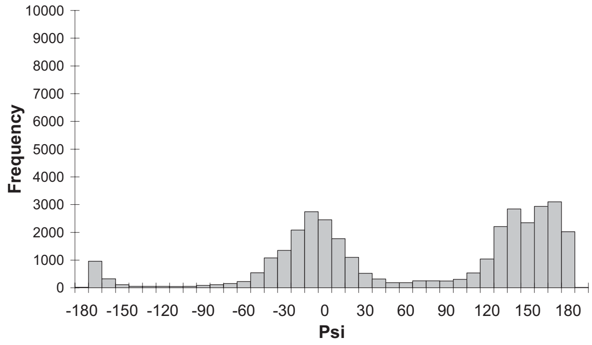
Serine



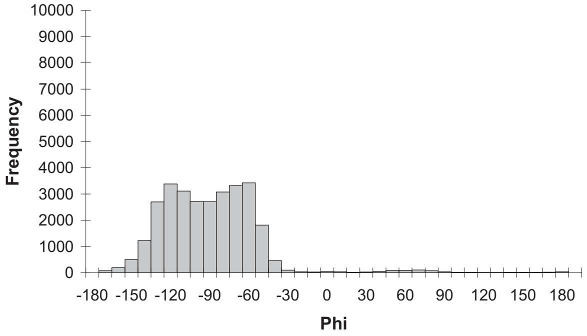
Threonine



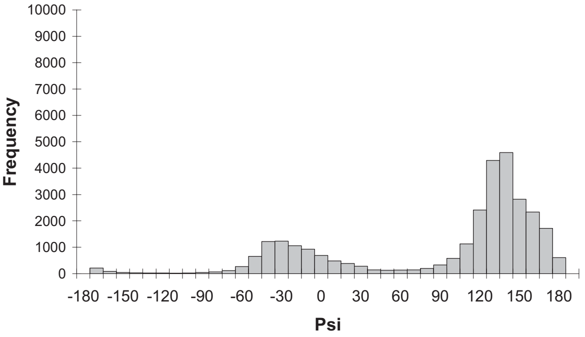
Threonine



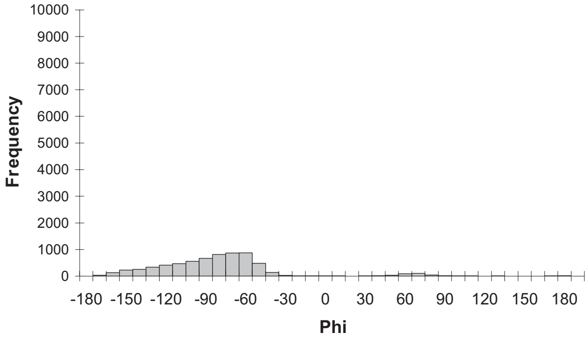
Valine



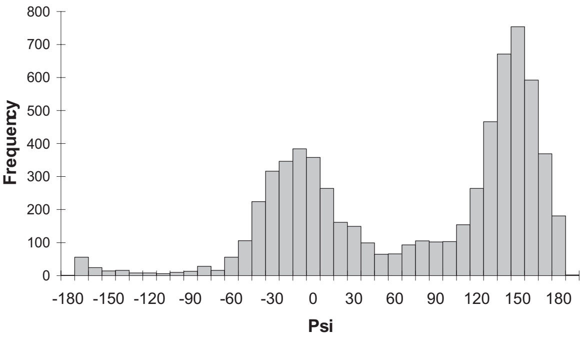
Valine



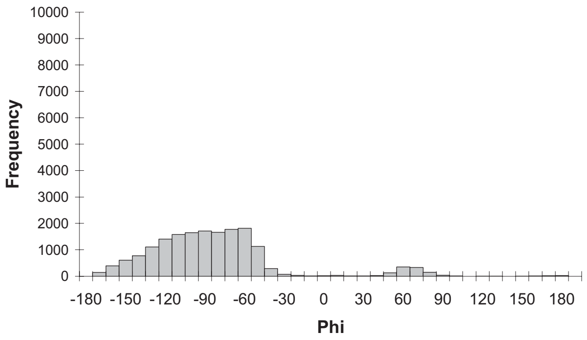
Tryptophan



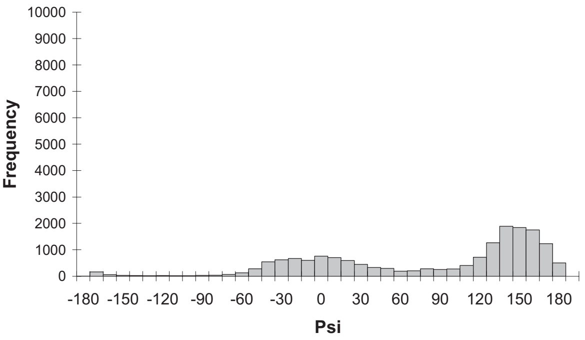
Tryptophan



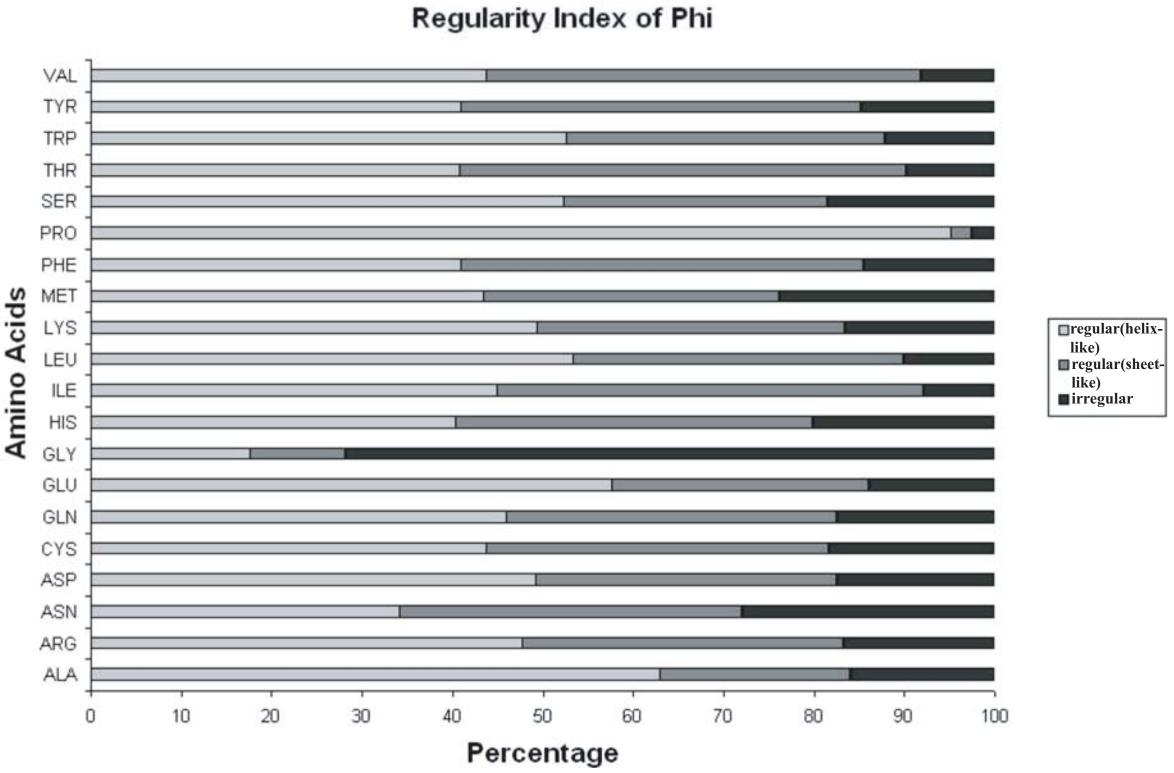
Tyrosine

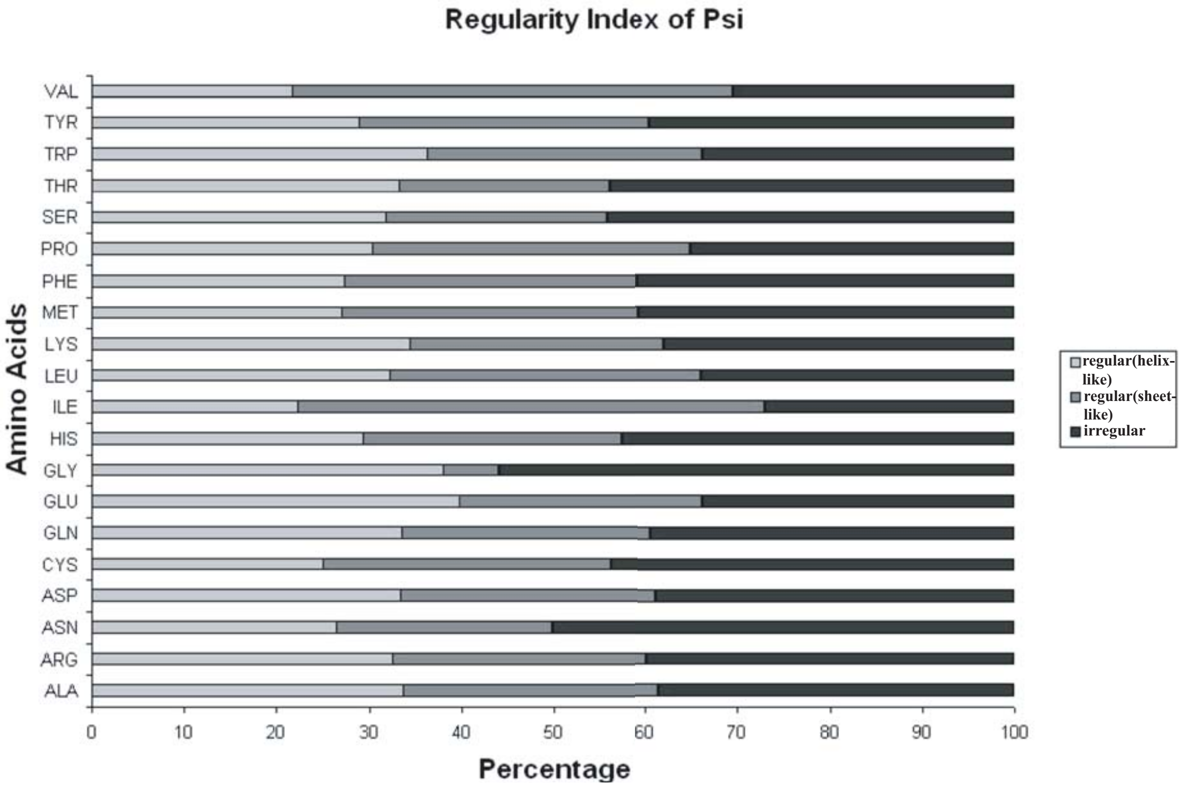


Tyrosine



Regularity Index for *Φ* and *Ψ* for all twenty amino acids using STRIDE structural assignment method.


Regularity Index of Phi
VAL
TYR
TRP
THR
SER
PRO
PHE
MET
LYS
LEU
ILE
HIS
GLY
GLU
GLN
CYS
ASP
ASN
ARG
ALA
Amino Acids
Percentage
regular(helix-like)
regular(sheet-like)
irregular


Regularity Index of Psi
VAL
TYR
TRP
THR
SER
PRO
PHE
MET
LYS
LEU
ILE
HIS
GLY
GLU
GLN
CYS
ASP
ASN
ARG
ALA
Amino Acids
Percentage
regular(helix-like)
regular(sheet-like)
irregular