Plant Tissue Cult. & Biotech. **19**(2): 217-226, 2009 (December)

PTC&B

Sequence Specific Dihedral Angle Distribution: Application in Protein Structure Prediction and Evaluation

S. M. Minhaz Ud-Dean[1] and Mahdi Muhammad Moosa*

Department of Genetic Engineering and Biotechnology, University of Dhaka, Dhaka-1000, Bangladesh

Key words: Torsion angle, ϕ-ψ distribution, Sequence specific ramachandran plot, Ramasekharan, Protein structure appraisal

Abstract

Protein structure prediction and evaluation is one of the major fields of computational biology. Estimation of dihedral angle can provide information about the acceptability of both theoretically predicted and experimentally determined structures. Here we report on the sequence specific dihedral angle distribution of high resolution protein structures available in PDB and have developed Sasichandran, a tool for sequence specific dihedral angle prediction and structure evaluation. This tool will allow evaluation of a protein structure in pdb format from the sequence specific distribution of Ramachandran angles. Additionally, it will allow retrieval of the most probable Ramachandran angles for a given sequence along with the sequence specific data.

Introduction

Protein structure prediction remains one of the greatest unsolved challenges of computational biology. While physical scientists are interested about the mechanism of the formation of complex 3D structure from linear primary sequence, biologists are more interested about its implication in the function of the biomolecule. Structure prediction is used both for assessing the quality of newly determined structures and predicting the structures of proteins whose sequences are newly determined. Massive amount of sequence data generated by several ongoing genome sequencing projects is making protein structure and function prediction even more important.

Secondary structure prediction is the start point of several *de novo* structure prediction methods (Lee et al. 2004a, Rohl et al. 2004, Sayalan et al. 2006). Large scale secondary structure prediction is used for function prediction, determination of the trend of different secondary structures in specific sub-cellular

*Author for correspondence. <mahdi.moosa@gmail.com>. [1]Present address: Optics Research Group, TNW/IST, TU Delft, Lorentzweg 1, 2628 CJ Delft, The Netherlands.

locations or among related organisms, identification of protein folds or enumeration of fold classes, identification of domains and estimation of the proportion of unfolded proteins in a given genome (Carter et al. 2003; Clare et al. 2006, Kelley et al. 2000, Lee et al. 2004b, Liu and Rost 2001, Montgomeire et al. 2006, Szafron et al. 2004). Besides, patterns of secondary structures can help in the identification of the overall protein folds which can aid in function prediction (Petsko and Ringe 2004). Protein secondary structure prediction is also important for molecular biologists in deciding sites of *affinity tag addition* as well as *fragmentation and joining* for partial and chimeric protein expression, respectively (Grasselli et al. 2003, Montgomeire et al. 2006, Szafron et al. 2004). Several applications of protein secondary structure prediction are enlisted in the article by Montgomeire and co-workers (Montgomeire et al. 2006).

Ramachandran et al. (1963) introduced the concept of allowed dihedral angles (ϕ, ψ) based on stereochemical hindrance as a parameter to describe protein backbone structure or secondary structure. The plot of these angles has been extensively used in determining protein structure (Kleywegt and Jones 1996, Morris et al. 1992), in defining protein secondary structure folds (Chou and Fasman 1974, Ho et al. 2003, Munoz and Serrano 1994) and in the verification of experimentally determined protein structures (Hovmöller et al. 2002, Laskowski et al. 1993, Morris et al. 1992).

Several methods of protein backbone ϕ-ψ angle prediction have been developed so far (Bystroff et al. 2000, de Brevern 2004, de Brevern et al. 2000, Dor and Zhou 2007, Karchin et al. 2003, Kuang et al. 2004, Mooney et al. 2006, Wood and Hirst 2005, Wu and Zhang 2008, Zimmermann and Hansmann 2006). Most of these methods use training algorithms like Neural Network (Dor and Zhou 2007, Kuang et al. 2004, Mooney et al. 2006, Wood and Hirst 2005), Support Vector Machine (Kuang et al. 2004, Zimmermann and Hansmann 2006) and Hidden Markov Model (Bystroff et al. 2000, Karchin et al. 2003) to predict the ϕ-ψ torsion angle. Composite Neural Network and Support Vector Methods were also developed by Wu and Zhang (2008). Despite the success of these methods in predicting local protein structure (Boomsma et al. 2008), to our knowledge, no probabilistic prediction model based on dihedral angle frequency of different amino acid combinations has been developed so far.

To reveal the relationship between dihedral angles and amino acid sequences and subsequent development of tools for backbone torsion angle prediction and structure verification, we have carried out comprehensive analysis of C_α geometry of all proteins available in the Research Collaboratory for Structural Bioinformatics Protein Data Bank (PDB) with the resolution of 1.5 Å or higher (Berman et al. 2000). Based on the frequency of di, tri and tetra-amino acid combinations and their respective most frequently appearing ϕ-ψ torsion angles, we have developed a probabilistic web tool for the prediction of protein secondary structure. The tool can be accessed from the URL: http://sasichandran.8bhuj.com. Additionally a theoretically predicted or experimentally determined structure can be evaluated by this tool. An offline version of the tool along with the source code is available for download from the website. We evaluated 97 proteins structures with resolution between 1.5 and 1.7 Å and 96 proteins with resolution >2.5 Å with our appraisal tool. It was found that among evaluated proteins, those with better resolution had better scores for most residues compared to low resolution ones.

Materials and Methods

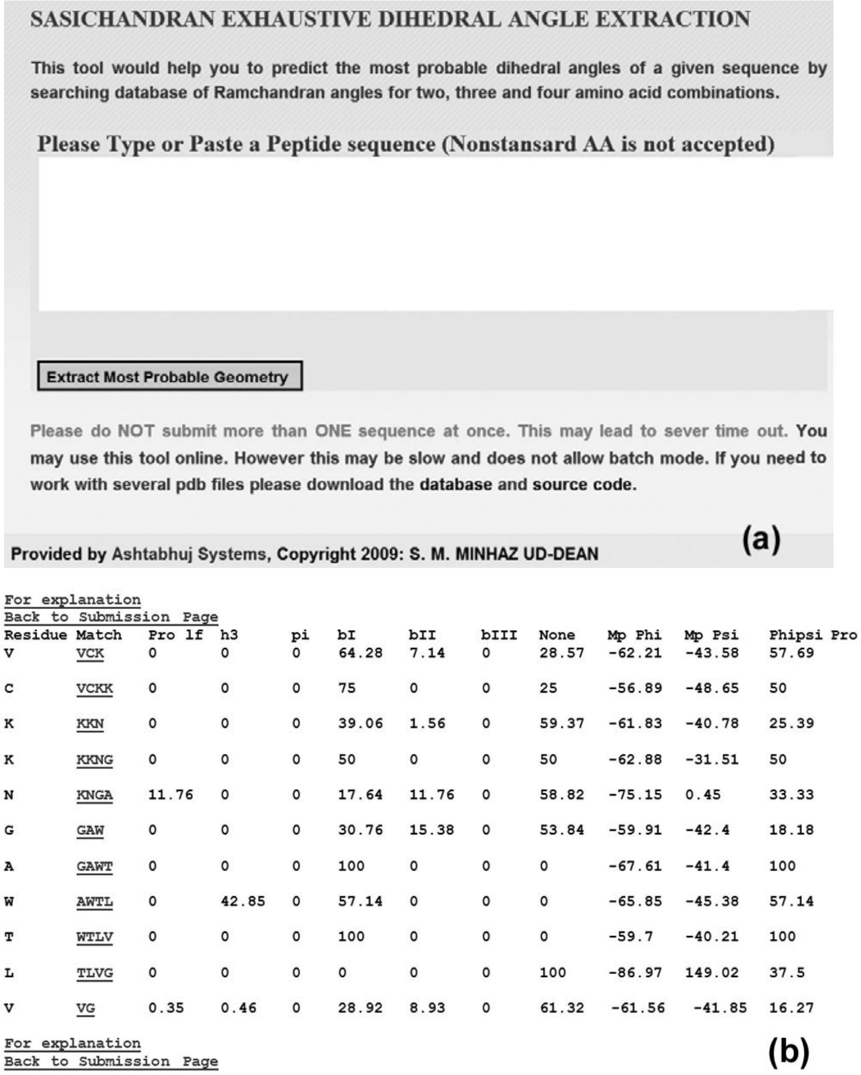
(a) *Database:* PDB files of proteins with resolution better than 1.5 Å were downloaded and sequences with greater than 95% homology were discarded. Non standard amino acids were excluded by splitting the polypeptide chains into two fragments whenever they were encountered. ϕ and ψ values for each amino acids were calculated using a script adapted from Carroll (2005). For all possible two, three and four amino acid combinations, respective values of ϕ-ψ pairs were computed. For each of the ϕ- ψ combinations of each AA combination, ϕ-ψ instances within 10^0 degree (both ϕ and ψ) were counted. Then the maximal ϕ-ψ combinations within the acceptable range were considered most probable. The probability of most probable combination was computed by dividing the number of instances within the range with the total number of instances. The probability was expressed as per cent.

(b) *Implementation:* The database is provided for both online and offline use. The online interface runs on a Windows NT server, the scripts are written in Perl and plots are generated by JavaScript. The online version includes an interface for calculation of dihedral angles from pdb files, an interface for extracting the most probable dihedral angles of a given sequence and appraisal of a protein structure in pdb format.

All of the above mentioned functionalities are also available in the offline version. Additionally the offline version allows computation in batch mode for multiple files.

Features

(a) *Sasichandran dihedral angle extractor:* Prediction of the most probable two, three or four amino acid blocks can be carried out by going to respective pages; an option of exhaustive search is also available. Fig. 1 shows an example output of exhaustive search.



SASICHANDRAN EXHAUSTIVE DIHEDRAL ANGLE EXTRACTION

This tool would help you to predict the most probable dihedral angles of a given sequence by searching database of Ramchandran angles for two, three and four amino acid combinations.

Please Type or Paste a Peptide sequence (Nonstansard AA is not accepted)

Please do NOT submit more than ONE sequence at once. This may lead to sever time out. You may use this tool online. However this may be slow and does not allow batch mode. If you need to work with several pdb files please download the database and source code.

Provided by Ashtabhuj Systems, Copyright 2009: S. M. MINHAZ UD-DEAN

```
For explanation
Back to Submission Page
Residue Match  Pro lf h3     pi   bI     bII    bIII None   Mp Phi  Mp Psi  Phipsi Pro
V       VCK    0      0      0    64.28  7.14   0    28.57  -62.21  -43.58  57.69

C       VCKK   0      0      0    75     0      0    25     -56.89  -48.65  50

K       KKN    0      0      0    39.06  1.56   0    59.37  -61.83  -40.78  25.39

K       KKNG   0      0      0    50     0      0    50     -62.88  -31.51  50

N       KNGA   11.76  0      0    17.64  11.76  0    58.82  -75.15  0.45    33.33

G       GAW    0      0      0    30.76  15.38  0    53.84  -59.91  -42.4   18.18

A       GAWT   0      0      0    100    0      0    0      -67.61  -41.4   100

W       AWTL   0      42.85  0    57.14  0      0    0      -65.85  -45.38  57.14

T       WTLV   0      0      0    100    0      0    0      -59.7   -40.21  100

L       TLVG   0      0      0    0      0      0    100    -86.97  149.02  37.5

V       VG     0.35   0.46   0    28.92  8.93   0    61.32   -61.56  -41.85 16.27

For explanation
Back to Submission Page
```

Fig. 1. (a) The input box for searching most probable ϕ, ψ angles. (b) Output page of exhaustive prediction of most probable ϕ, ψ angles for first ten amino acid residues of human β-globin chain. The output also gives the probability (as per cent) of observing suggested parameters. Each of the short subsequence provides link to the database of the dihedral angle for that sequence. A Ramachandran plot for the sub-sequence is also displayed.

The offline version works on the same principle. In the command line one can enter the intended peptide sequence and then select the mode (exhaustive, 2AA, 3AA or 4AA). The result is both displayed on the screen and saved in a .tsv file.

This tool calculates dihedral angle of structures from PDB database (Berman et al. 2000). PDB ID of respective files should be submitted to calculate ϕ and ψ dihedral angles. The offline version of Ramasekharan would take a folder containing .pdb files and return .tsv files containing ϕ-ψ angles for each of the residues except nonstandard AA. Additionally the calculated angles are also displayed on the command line.

(c) *Sasichandran dihedral angle appraiser:* The tool gives the option of appraising dihedral angle. This can be done both by uploading protein structure file in pdb format or pasting the formatted file in the input window. The tool first calculates dihedral angles of each amino acid residues. The input is then split into overlapping ten amino acid fragments. The probability of each observed dihedral angle is calculated for each fragment and given as output.

```
Thank you for using our service. Your intended appraisal result is following:
Your sequence block: LRIDEGLRLK
    Residue        Match           Phi            Psi         Probability
    L              LRI           -63.36         -29.93            20
    R              LRID          -68.24         -30.34            9.09
    I              RIDE          -73.46         -42.15           83.333
    D              IDEG          -68.40         -39.94            50
    E              DEGL          -98.48         -25.52           7.142
    G              EGLR           76.55        -172.01           13.333
    L              GLRL         -141.41         127.91            25
    R              LRLK         -138.19         130.50             0
    L              LK            -99.39           5.85           2.714
                          O indicates the origin of the peptide.
```

Fig. 2. Output of Sasichandran dihedral angle appraiser. The probability values of dihedral angle of each amino acid residues are appraised.

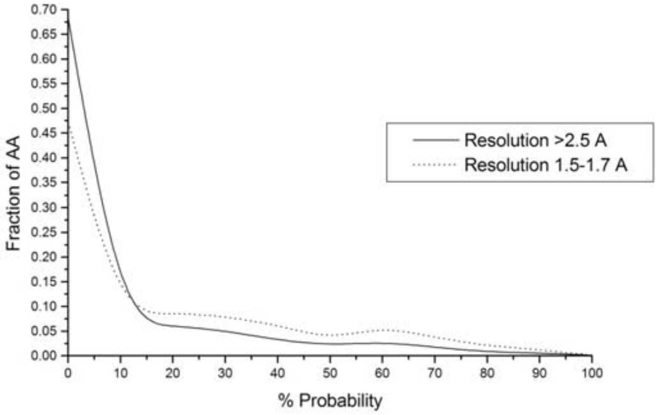


Fig. 3. Fraction of amino acids in different probability range. From the graph, it is to be noted that low resolution structures have greater deviation compared to higher resolution structures.

The offline version provides more options. This can be used in batch mode for multiple pdb files. First the name of the folder containing pdb files is provided to Ramasekharan, which produces .tsv files containing the calculated dihedral angles. Then the .tsv files may be copied into a separate folder and

Sasichandran appraiser would appraise the dihedral angles. The result is both displayed on the command line and saved in .tsvappraised.tsv files. These files may be viewed by either a text editor or a spreadsheet program.

Evaluation of the tools: To evaluate the performance of Sasichandran Appraiser, we analyzed reported structures from PDB which were not used in the calculation of dihedral angle frequency of di, tri and tetra-amino acid combinations. We divided the protein structures into two categories based on crystal resolution. The analysis showed that proteins with better resolution had greater probability of having the most probable backbone torsion angle predicted from amino acid combination database (Fig. 3).

Results and Discussion

A total of 1527 proteins met the criteria set in the methods section. After removing nonstandard amino acids, further analysis was carried out on 482764 amino acid residues of 2579 polypeptide chains. All possible 400 (20^2) two amino acid combinations were found to be present in the dataset. On the contrary, 106479 of 160000 (20^4) and 7978 of 8000 (20^3) possible four and three amino acid combinations were respectively observed.

Smaller amino acids were found to be less stringent in their C_α geometry. Analysis of the effect of neighboring amino acids in C_α geometry revealed that charged or large amino acid residues affect the geometry considerably. As for example, presence of Gly in a given sequence decreases the predictability (Fig. 4). However, the presence of charged group like Asp increases the level of stringency and results in confinement in two symmetrically distributed regions with few offsets (Fig. 4d). This represents the effect of neighboring asymmetric groups on the symmetry of Gly. For rare amino acids it is possible to predict the C_α geometry with greater than 75% confidence.

Although Dayalan et al. (2006) developed a database of ϕ-ψ dihedral angles of different amino acid combinations (of 5227 non-redundant high resolution protein structures) they did not provide any option for prediction or evaluation of the backbone dihedral angle. The database calculated ϕ-ψ angle frequency of different amino acid combinations independently. However, it is expected that ϕ-ψ dihedral angles are interdependent (Ramakrishnan and Ramachandran, 1965). Conformational Angles DataBase (CADB) of proteins by Sheik et al. (2003) also provides a similar database of ϕ-ψ angles of 6146 protein chains. Like the Dayalan et al. (2006) database, this neither provides option for prediction or evaluation of protein structures even though it provides the option of retrieving side chain conformations. Our analysis of 1527 proteins with 1.5 Å or higher resolution revealed that the presence of charged amino acids poses restriction on the allowed region of neighbouring amino acids. However, presence of large nonpolar AA does not pose that much restriction. In accordance with previous theoretical analysis (Ramachandran et al. 1963), smaller amino acids have larger conformational freedom. Nevertheless, presence of neighbouring charged groups significantly restricts the freedom even in small amino acids.

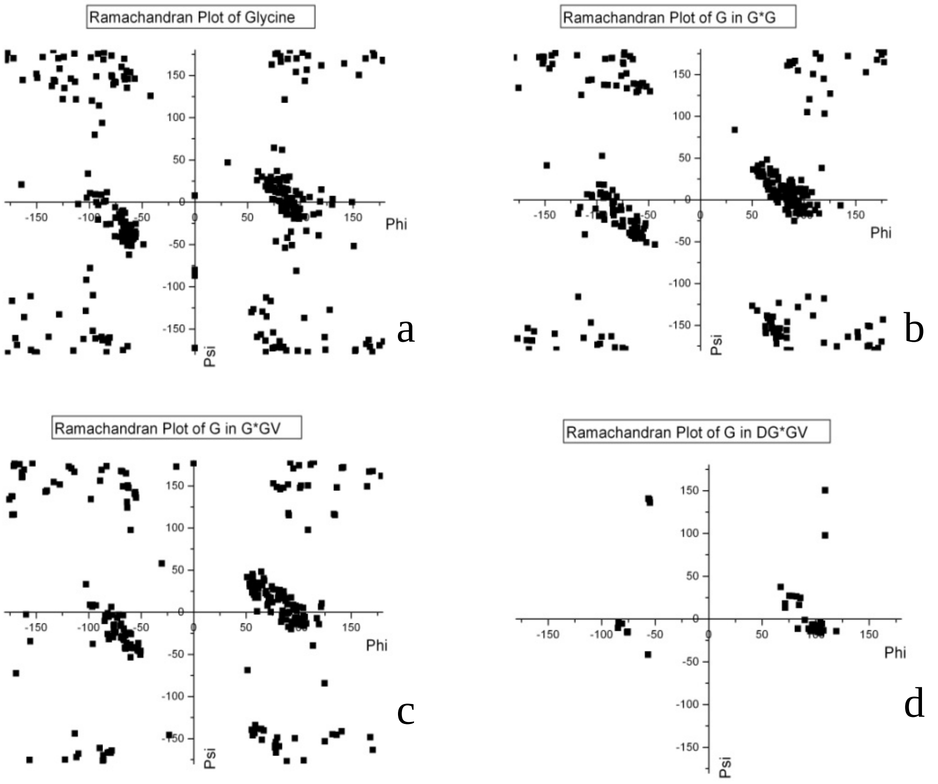


Fig. 4. (a) Only G seems to be scattered over a large range of ψ and φ angles. (b) G in GG has about the same freedom. (c) The presence of large amino acid like valine restricts G geometry little. d) However, presee of a charged group like D makes G in DGGV confined in two symmetrically distributed regions with few offsets. This represents the effect of neighbouring asymmetric groups on the symmetry of G. Similar pattern was observed for Alanine.

Further improvement of probabilistic structure prediction depends on determination of high resolution protein structures. Although we found 7978 of 8000 possible three amino acid combinations, many four amino acid combinations were missing. This does not necessarily indicate that these theoretically allowed combinations are improbable, rather they are absent from the list of high resolution structures determined till date. Since the tool uses dihedral angles of amino acid combinations in PDB structures to build the Database (section a) of Materials and Methods, presence of these rare combinations may reduce the prediction quality significantly. It is to be noted that although Dayalan et al. (2006) incorporated penta amino acid combinations, the database contains only 733000 fragments out of 3200000 (*20^5*) possible combinations.

The probabilistic model could be further developed by incorporating the effect of preoccupied volumes. By tracing the backbone through the probable dihedral angles for consecutive amino acids and excluding the volume preoccupied by the preceding amino acid, a set of probable backbone structures can be formulated. These structures may be used as seeds for energy minimization programs. This approach would significantly decrease the complexity of energy minimization based structure prediction.

Acknowledgements

The authors would like to thank reviewers for their critical comments.

References

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and **Bourne PE** (2000) The protein data bank. Nucl. Acids Res. **28**: 235-242.

Boomsma W, Mardia K, Taylor C, Ferkinghoff-Borg J, Krogh A and **Hamelryck T** (2008) A generative, probabilistic model of local protein structure. Proc. Natl. Acad. Sci. USA. **105**: 8932-8937.

Bystroff C, Thorsson V and **Baker D** (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J. Mol. Biol. **301**: 173-190.

Carroll M (2005) Universal research interchange. J. Comput. Sci. Colleges **20**: 119-119.

Carter P, Lie J and **Rost B** (2003) PEP: Preductions for entire proteome. Nucleic Acids Res. **31**: 410-413.

Chou P and **Fasman G** (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry **13**: 211-222.

Clare A, Karwath A, Ougham H and **King RD** (2006) Functional bioinformatics for Arabidopsis thaliana. Bioinformatics **22**:1130-1136.

Dayalan S, Gooneratne N, Bevinakoppa S and **Schroder H** (2006) Dihedral angle and secondary structure database of short amino acid fragments. *Bioinformation.* **1**: 78.

de Brevern A (2004) Local backbone structure prediction of proteins. In Silico Biol. **4**: 381-386.

de Brevern A, Etchebest C and **Hazout S** (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Protein Struct Funct Genet. **41**: 271-287.

Dor O and **Zhou Y** (2007) Real-SPINE: An integrated system of neural networks for real-value prediction of protein structural properties. Proteins. **68**: 76-81.

Grasselli E, Noviello G, Rando C, Nicolini C and **Vergani L** (2003) Expression, purification and characterisation of a novel mutant of the human protein kinase CK2. Mol. Biol. Rep. **30**: 97-106.

Ho B, Thomas A and **Brasseur R** (2003) Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the a-helix. Protein Sci. **12**: 2508-2522.

Hovmöller S, Zhou T and **Ohlson T** (2002) Conformations of amino acids in proteins. Acta Crystallogr. **D58**: 768-776.

Karchin R, Cline M, Mandel-Gutfreund Y and **Karplus K** (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins. **51**: 504-514.

Kelley LA, MacCallum RM and **Sternberg MJE** (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. J. Mol. Biol. **299**:501-522.

Kleywegt GJ and **Jones TA** (1996) Phi/Psi chronology: Ramachandran revisited. *Structure.* **4**: 1395-1400.

Kuang R, Leslie C and **Yang A** (2004) Protein backbone angle prediction with machine learning approaches. Bioinformatics **20:** 1612-1621.

Laskowski RA, MacArthur MW, Moss DS and **Thornton JM** (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. **26**: 283-291.

Lee J, Kim S-Y, Joo K, Kim I and **Lee J** (2004a) Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. Proteins **56**: 704-714.

Lee S, Cho MK, Jung JW, Kim JH and **Lee W** (2004b) Exploring protein fold space by secondary structure prediction using data distribution method on Grid platform. Bioinformatics **20**: 3500-3507.

Liu J and **Rost B** (2001) Comparing function and structure between entire proteomes. Protein Sci. **10**: 1970-1979.

Montgomeire S, Sundararaj S, Gallin WJ and **Wishart DS** (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics **7**: 301.

Mooney C, Vullo A and **Pollastri G** (2006) Protein Structural Motif Prediction in Multidimensional ϕ-ψ Space Leads to Improved Secondary Structure Prediction. J Comput Biol. **13**: 1489-1502.

Morris AL, MacArthur MW, Hutchinson EG and **Thornton JM** (1992) Stereochemical quality of protein structure coordinates. Protein Struct Funct Genet. **12**:345-364.

Munoz V and **Serrano L** (1994) Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. Proteins. **20**: 301-311.

Petsko GA and **Ringe D** (2004) Importance and determinants of protein secondary structure, in *Protein structure and function* (Lawrence E and Robertson M eds) pp 12-13, London: New Science Press Limited.

Ramachandran GN, Ramachrishnan C and **Sasisekharan** V (1963) Steriochemistry of polypeptide chain configurations. J. Mol. Biol. **7**: 95-99.

Ramakrishnan C and **Ramachandran G** (1965) Stereochemical criteria for polypeptide and protein chain conformations: II. Allowed conformations for a pair of peptide units. Biophys J. **5**:909.

Rohl CA, Strauss CEM, Misura KMS and **Baker D** (2004) Protein structure prediction using Rosetta. Methods Enzymol. **383**: 66-93.

Sayalan S, Gooneratne ND, Bevinakoppa S and **Schroder H** (2006) Dihedral angle and secondary structure database of short amino acid fragments. Bioinformation. **1**: 78-80.

Sheik S, Ananthalakshmi P, Bhargavi G and **Sekar K** (2003) CADB: Conformation angles database of proteins. Nucleic Acids Res. **31**: 448-451.

Szafron D, Lu P, Greiner R, Wishart DS, Poulin B, Eisner R, Lu Z, Anvik J, Macdonell C, Fyshe A and **Meeuwis D** (2004) Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. Nucl Acids Res. **32**:W365-371.

Wood M and **Hirst J** (2005) Protein secondary structure prediction with dihedral angles. Proteins. **59**: 476-481.

Wu S and **Zhang Y** (2008) ANGLOR: A composite machine-learning algorithm for protein backbone torsion angle prediction. *PLoS One.* **3**: e3400. doi:10.1371/journal. pone. 0003400

Zimmermann O and **Hansmann U** (2006) Support vector machines for prediction of dihedral angle regions. Bioinformatics. **22**: 3009-3015.