A Dihedral Angle Database of Short Sub-Sequences for Protein Structure Prediction

Saravanan Dayalan, Savitri Bevinakoppa, Heiko Schroder

School of Computer Science and Information Technology
RMIT University
GPO Box 2476V, Melbourne 3001, Australia

{sdayalan, savitri, heiko}@cs.rmit.edu.au

Abstract

Protein structure prediction is considered to be the holy grail of bioinformatics. Ab initio and homology modelling are two important groups of methods used in protein structure prediction. Amongst these, ab initio methods assume that no previous knowledge about protein structures is required. On the other hand homology modelling is based on sequence similarity and uses information such as classification, structure, sequence and dihedral angles for prediction.

Even though there are many databases for structural and sequence information, there are not many databases for dihedral angles that store all occurring dihedral values of sub-sequences. The existing ones have limitations like not being able to retrieve dihedral values for amino acids of a specific sub-sequence or being designed only for a specific set of proteins based on sequence identity (proteins with < 20% sequence identity). They hence have disadvantages when used in protein structure prediction based on short sub-sequences and exact matches. This paper presents a dihedral angle database for short sub-sequences up to length five. In this database dihedral angles of all proteins were extracted from the Protein Data Bank (PDB) regardless of the percent of sequence similarity. This paper also shows how the database can be used for protein structure prediction using exact matches.

Keywords: Protein structure prediction, homology modelling, dihedral angle, sub-sequence.

1 Introduction

The protein structure prediction problem is considered to be the holy grail of bioinformatics. The problem states 'How to predict the exact three dimensional structure of a protein from its one dimensional amino acid sequence?'

There are many ways by which this problem is being tackled. These methods are basically classified into two groups: 1) ab initio and 2) homology modelling (Jones 2000). Out of these, ab initio methods assume they do not require knowledge about the previously obtained protein structural information. They try to solve the problem by concentrating on the physio-chemical criteria such as the Vander walls forces, bonding characteristics and the charge of amino acid. On the other hand homology modelling is based on protein sequence similarity. Protein sequence is the sequence of amino acids that makes up the protein. There are 20 different amino acids and these take different combinations and different lengths to form a protein (Carl and John 1999). A sub-sequence of a protein is a sub section of the complete protein sequence. A short sub-sequence can be considered as sub-sequences of amino acids up to length 5. Homology modelling uses previous knowledge like protein structure details, fold characteristics, family classifications and dihedral angle values for structural prediction of proteins. Homology modelling is based on the idea that similar sequences (>20% identity) may have similar structures (Jones 2000; Doolittle, R.F. 1986).

There are many databases that have detailed information about factors such as protein structure (Berman et al 1999), sequence of protein (Bairoch et al 1998; Mewes et al 1998), its family classification (Laskowski et al 1997; murzin et al 1995) and patterns (attwood et al 1998; Bairoch et al 1997). On the other hand there are not many databases dedicated for storing the dihedral values for sub-sequences of proteins. These databases would answer a user query such as 'Retrieve all the psi values of HIS and both phi psi values of CYS and phi values of ALA occurring in the sub-sequence HIS-CYS-ALA throughout the structural database'. The existing databases of dihedral angles (Sheik et al 2003; Oleg et al 1998) have limitations such not being able to retrieve dihedral values for one or more amino acids when they occur in a specific sub-sequence or being designed specifically for a set of proteins based on its sequence identity (for example, all proteins with less than 20% sequence homology). Due to these limitations the above-cited databases have disadvantages when used in protein structure prediction that is based on short sub-sequences and exact matches. The disadvantages lies in the fact that, they do not consider all the protein structures in the structural database during dihedral value

extraction and are not able extract all dihedral values for a given sub-sequence of amino acids.

This paper proposes a database of dihedral angles for short sub-sequences of amino acids up to length five. This database has dihedral values of all proteins in the Protein Data Bank (PDB Release #101 July 2002). The proteins were selected regardless of the degree of sequence similarity between proteins. All proteins were considered because this database concentrates on short sub-sequences and hence exact matches in similarity searches during homology modelling. Section 2 gives background details on homology modelling, dihedral angles, protein databases and the need for dihedral databases. Section 3 explains related work with respect to dihedral databases along with their limitations. Section 4 presents the design and implementation details of the proposed dihedral angle database for short sub-sequences of amino acids up to length five. Finally an example is shown on how the proposed dihedral database can be used for protein structure prediction using exact matches.

2 Background

This section describes homology modelling, dihedral angles of amino acids, protein databases and finally the need to have a dihedral database and how they differ from structural databases that also holds dihedral values of amino acids.

2.1 Homology modelling

Homology modelling is one of the important methods used in protein structure prediction. When two proteins are said to be homologous, it means that both these proteins share a common evolutionary history (ed. Andreas and Francis 2001). Homology modelling is done based on sequence similarities. Sequence similarity represents the degree to which two sequences are similar. This is important because if two sequences are similar, then it is assumed that they could have derived from a common ancestor and might have similar structure functions (ed. Andreas and Francis 2001). Sequence similarity is done by aligning two sequences and looking for similarities. There are many alignment techniques such as the ones proposed by Smith et al 1981, Dayhoff et al 1978 and Needleman et al 1970. In alignment techniques substitution scores and gap penalties are often introduced to obtain degree of similarity between two sequences. There are database similarity searching methods such as BLAST (Altschul et al 1990) and FASTA (Lipman et al 1985) that uses different alignment techniques to search for similarities in sequence and structure databases for a given sequence.

Protein structure prediction is done using homology modelling by taking a sequence and looking into similarities of multiple sequences in the sequence and structure databases. Once the sequences with a good degree of similarity have been retrieved, then the prediction method uses previous obtained knowledge about factors such as structure of proteins, fold characteristics of sub-sequences, classifications of proteins and dihedral angles to make the prediction. Out of the factors used by the prediction method, dihedral angles are of importance because they define the backbone structure of a protein.

2.2 Dihedral angles

Proteins are made up of building blocks called amino acids. There are 20 different amino acids and these take different combinations and different lengths to constitute proteins. A single amino acid residue is made up of a central carbon atom called the C-alpha to which is bonded an amine group, a carboxyl group, a hydrogen atom and a side chain. The central carbon atom, hydrogen atom, amine group and the carboxyl group are the same for all amino acids. It is the difference in the side chains R that differentiates between all 20 amino acids as shown in figure 1 (Carl and John 1999).

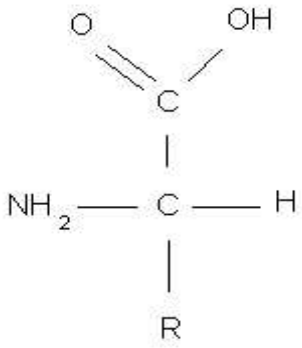

Figure 1: A single amino acid residue

Amino acids bond with each other by releasing a water molecule resulting in a bond between N of one amino acid and C of the other amino acid. This bond between N and C of two different amino acids is called as a peptide bond (Carl and John 1999). In this way multiple amino acids can bond with each other resulting in a large chain of amino acids joined by the peptide bond. The resulting large chain of amino acids would constitute a protein.

Figure 2 shows a single amino acid residue where the central carbon atom C-alpha is bonded with N and C. The angle of the bond between N and C-alpha is known as Phi and the angle of the bond between C and C-alpha is known as Psi.

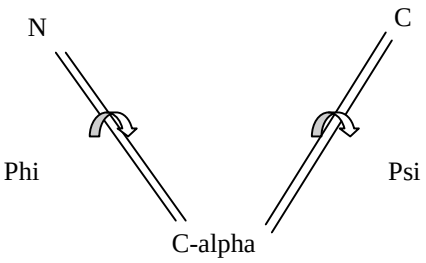

Figure 2: Amino acid residue with dihedral angles

Both these angles phi and psi together form the dihedral angles. These angles could take different values for the same amino acid in different occurrences in the same proteins as well as in different proteins. Since the dihedral angles together with omega (angle of the peptide bond) could define the entire conformation of a protein, these values are considered to be important for protein structure prediction.

2.3 Protein Databases

Biological databases are databases that store biological data. Some of the types of data that a biological database would contain are genome information of different species, DNA sequences, protein sequences, structural details of proteins and family classification details. Protein databases are the ones that store details relating to proteins. The different types of protein databases are primary sequence databases, composite protein sequence databases, structure databases, structure classification databases and pattern databases (Attwood and Parry 1999).

Protein sequence databases stores sequence information of proteins. Since sequencing a protein or in other terms finding out the amino acid sequence that the protein is make of takes much less than determining the three dimensional structure of a protein (Attwood and Parry 1999), the sequence databases are much larger than the structure databases. Some of the important sequence databases are MIPS (Mewes and Apweiler 1998), SWISS-PORT (Bairoch et al 1996) and TrEMBL (Mewes and Apweiler 1998).

Composite protein sequence databases extract information from various sequence databases and aim at having non-redundant data. These databases extract data from sequence databases on a regular interval. An example of a composite sequence database is OWL (Bleasby et al 1994). Protein structure databases hold the three dimensional structural information of proteins whose structures were determined by crystallographic process and nuclear magnetic resonance methods (Berman et al 1999). PDB (Berman et al 1999) is an example of structural database that holds nearly 20,000 protein structures as of August 2003[1].

Structure classification databases contain details on the classification of proteins. SCOP (Murzin et al 1995) is a classification database that classifies proteins based on different factors such as evolutionary relationships, structural and functional characteristics and common fold. Similarly CATH (Orengo et al 1997) is another database that does classification of proteins based on structure, topology, homology and sequence. Pattern databases contain details about patterns like highly conserver areas that occur in proteins. PRINTS (Attwood ET AL 1998) is an example of a pattern database.

[1] http://www.rcsb.org/pdb/holdings.html#holdings

3 Need for dihedral databases

Even though there are many databases for proteins as specified in section 2.3, there are not many databases that store dihedral values. Since dihedral values are an important factor in the prediction process, there are many methods that uses dihedral angles in the protein structural prediction problem such as the works done by Mark et al 2003, Rooman et al 1991, Gabriel et al 1999 and Simon et al 1998. A homology modelling method for structural prediction that concentrates on dihedral angles would make an analysis of these angles for all proteins in the structural database. Even though the dihedral values are indirectly stored in structure databases, answering a query like 'Retrieve all the psi values of HIS and both phi psi values of CYS and phi values of ALA occurring in the sub-sequence HIS-CYS-ALA throughout the structural database' would be time consuming. This is because in the structure databases the information that is stored is the sequence of amino acids that make up the protein and the position of all atoms of amino acids in 3D space.

To answer the above query using a structural database a program should be run to extract all the dihedral angles of all occurrences of amino acids in nearly 20,000 protein structures. Then the program will have to look at the long list of extracted dihedral angles and search for the sub-sequence HIS-CYS-ALA. With nearly 20,000 protein structures, the above-specified sub-sequence is likely to occur hundreds of times. Finally after looking at all the occurrences of the sub-sequence, the program would answer the query. A protein structure prediction program using dihedral angles would initiate hundreds and thousands of such queries during the prediction process especially if the prediction process concentrates on exact matches of sub-sequences. Clearly using a structural database for such prediction methods will be time consuming. To answer such queries efficiently a separate database is needed that stores all occurring dihedral angles of sub-sequences for all proteins in the structural database. To answer a query such as mentioned above using a dihedral database, a program would directly look at the appropriate entry of the sub-sequence of question and obtain a long list of all occurring dihedral values.

4 Related work

Two important dihedral angle databases are the Conformation Angles DataBase of proteins (CADB) of Sheik et al 2003 and Conformational Database for Amino Acid Residues in Protein Structures of Oleg et al 1998. In this section, both these databases are explained along with their limitations.

CADB by Sheik et al stores main-chain and side-chain conformation angles of protein structures. CADB has functionalities like visualizing conformation angles (main-chain and side-chain) for a particular amino acid residue, users being able to study the interrelationship between main-chain and side-chain conformation angles and a World Wide Web interface. CADB concentrates on protein structures in two datasets with sequence identities of 25% and 90%. In total CADB

has used about 7000 protein structures out of more than 15,000 protein structures in the PDB. Since CADB does not use all available protein structures from the PDB, there is a limitation in its use when a program wants to analyse dihedral angles of all protein structures. This is especially important if the prediction program uses short sub-sequences of exact matches in which case the prediction program will be analysing the values for all proteins regardless of the percent of sequence similarities between proteins. Another limitation of CADB is, even though it allows visualizing conformation angles for a particular amino acid, it will not be able to answer a query that requires dihedral values of multiple amino acids based on their positions with each other, like obtaining all dihedral values for HIS and ALA when they appear in the sub-sequence CYS-HIS-ALA-GLY.

The conformational database (CDB) by Oleg et al 1998 has data from 473 high-quality non-homologous protein structures from the PDB. For each amino acid residue CDB holds information such as the PDB code, amino acid code, one letter code for residue, values of dihedral angles, details of fractional area, energy by residue and more. CDB has data about 473 proteins that have less than 20% sequence identity. CDB has got a World Wide Web interface to access the database and its interface has got the option of selecting values for multiple parameters such as resolution, crystallographic R-factor and pattern for amino acid sequence. Since CDB has data about just 473 proteins out of more than 15,000 proteins in the PDB, it has got a strong limitation when used in structural prediction programs using short sub-sequences and exact matches. The program using CDB will not be able to analyse all dihedral values occurring in the PDB and hence the result of an analysis would be inconsistent.

These are the limitations of the existing databases of dihedral angles. The database design proposed in this paper overcomes these limitations and hence can be used in structure prediction programs efficiently.

5 Dihedral Angle Database (DAB)

For the design of the proposed dihedral angle database, all proteins from the Protein Data Bank were considered regardless of the degree of the sequence identity between them. This is due to the fact that, the proposed database is designed to store all occurring dihedral values of short sub-sequences, that is sequences of amino acids up to length 5. This database would answer protein structure prediction program queries efficiently that uses homology modelling methods using short sub-sequences and exact matches.

5.1 Design

DAB contains all occurring dihedral values of sub-sequences of amino acids from length one to five. Since there are 20 amino acids, a sub-sequence of length one can have 20 different combinations (20^1). Similarly a sub-sequence of length two can take 400 different combinations (20^2). Sub-sequences of length three, four and five would have 8000, 160,000 and 3.2 million different combinations respectively in a similar fashion. DAB has entries containing all occurring dihedral angles for each of a total of 3,368,420 combinations (20+400+8,000+16,000+3,200,000). This method of storing dihedral values for all combinations would ensure faster retrieval when compared to retrieving dihedral values using structure databases due to reasons specified in section 3.

DAB was built in two phases as shown in figure 3. In the first phase, dihedral angles (Phi, Psi and Omega) were extracted for all proteins in the PDB (release #101 July 2002) using a molecular graphics program RasMol (Sayle and Bissell 1992) and the extracted results were stored in a resultant file (RF) in the following format. RasMol extracts the dihedral angles of one protein at a time and appends RF with the extracted values. In RF, when the dihedral values are appended, the protein sequence order is preserved. Hence the sequence of amino acids in RF would be similar to the amino acid sequence of PDB file from which the angles were extracted. After extracting dihedral angle values of all combinations up to length 5, RF has dihedral values of more than 8 million amino acid residues from nearly 20,000 proteins. For proteins in the PDB with several polypeptide chains, all the chains were considered during the dihedral angle extraction process.

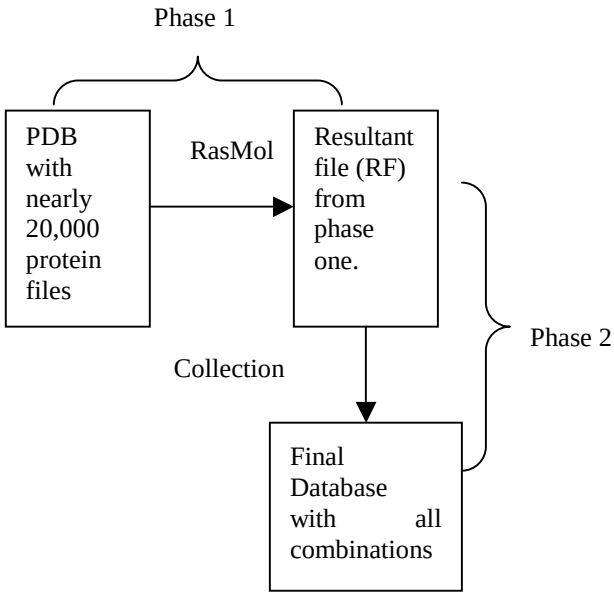

Figure 3: Dihedral Database design phases

In the second phase, RF was scanned for the occurrence of every combination of amino acid sub-sequences from length one to five (for all 3,368,420 combinations). The collection of dihedral angle values for sub-sequence one was done as follows. RF was scanned for the occurrence of XXX-AAA-XXX, where XXX indicates any amino acid and AAA is the amino acid to which all occurring dihedral values are to be collected. For each of this occurrence the following information was stored:

1. PDBCode, which is the unique id of the protein file in which the amino acid AAA occurs.
2. One letter code of the polypeptide chain (if there were multiple chains).
3. Amino acid sequence number, which is the position of occurrence of the considered amino acid AAA along the protein sequence.
4. Phi value of AAA.
5. Psi value AAA.
6. Omega value of AAA.

The exceptional case would be when AAA occurs at the start or the end of a polypeptide chain in which case only one of either Phi or Psi was stored and the other was assigned the value 0.

Similarly the collection process for dihedral angle values of sub-sequence length two was done as follows. RF was scanned for the occurrence of XXX-AAA1-AAA2-XXX, where XXX indicates any amino acid, AAA1 indicates the first amino acid that occurs before the second amino acid AAA2 in the protein sequence. An example of a sub-sequence of amino acid length two is, XXX-CYS-HIS-XXX, where XXX in XXX- CYS indicates that CYS could be preceded by any amino acid. Similarly XXX in HIS-XXX indicates that HIS could be succeeded by any amino acid. CYS-HIS indicates that amino acid CYS should precede HIS in the protein sequence. The occurrence of CYS followed by HIS would be searched in RF and the dihedral information (Phi, Psi values of both CYS and HIS) stored. The only difference between the information stored in sub-sequence of length one and sub-sequence of length two were the Phi, Psi and Omega values. In case of length one, Phi, Psi and Omega values were collected for just AAA, whereas in case of length two, Phi, Psi and Omega values were collected for both AAA1 and AAA2 when they occur in the order AAA1 followed by AAA2. Similarly for sub-sequence of length three, four and five, the Phi, Psi and Omega values were collected for three, four and five amino acids respectively.

Even though the database has 3,368,420 entries, one for each combination, not all of these entries hold data. This is because, if sub-sequence of length 5 is considered, it can take 3.2 million combinations, but not all these combinations of amino acid sequences occur in proteins. For example, the sub-sequence CYS-CYS-CYS-CYS-CYS might not occur in proteins. Similarly for sub-sequence of length 4 not all combinations of amino acid sequences is found in proteins. But for sub-sequence of length one, two and three, all combinations occur in proteins because all three together form only 8420 combinations. For the proposed database DAB, sub-sequences of length up to five were considered because the longer sub-sequence, the lesser number of instances it would occur. For example, the number of occurrences of the sub-sequence HIS-PHE-VAL-LYS (length 4) is 47♦ and the number of occurrences of the sub-sequence GLY-GLY-GLN-SER-SER (length 5) is 11♦. Sub-sequences of length 6 or more are not considered for DAB because they would have very less number of occurrences. Hence a collection of dihedral values of sub-sequences of length 6 or more will not be of good significance for a prediction program.

♦Except for the cases where the sequence has CYS on both ends

5.2 Examples queries to DAB

This section explains a scenario that shows the use of the proposed dihedral database for a protein structure prediction program that queries dihedral angle values of sub-sequences.

For example, consider a prediction program that analyses the effect of ALA preceding and succeeding ALA. Or in other terms, the sub-sequences under analysis would be ALA-ALA-ALA. For clarity sake if numbers are added to the above representation of the sub-sequence it could be written as ALA1-ALA2-ALA3. First the program would make a query to understand the distribution of dihedral values (for this example, consider only Psi) if ALA1 occurs by itself. The query could be termed as 'Extract all the Psi values of XXX-ALA2-XXX', where XXX represents any amino acid. For this query the results obtained from the proposed database DAB is shown in figure 4.

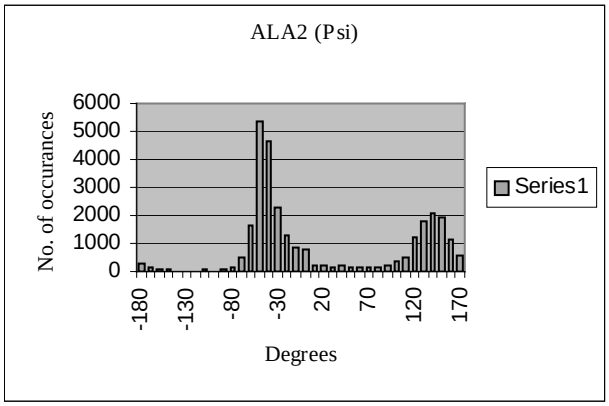

Figure 4: Psi of ALA in XXX-ALA-XXX

This graph shows the distribution of the Psi values of ALA2 in XXX-ALA2-XXX. This graph shows that majority of the occurrences of Psi lies from –70 degrees to –20 degrees and from 120 degrees to 160 degrees. Next, the program would want to know the impact of having ALA3 succeeding ALA2. Thus the next query would be 'Extract all occurring Psi values of ALA2 in the occurrence XXX-ALA2-ALA3-XXX. The result obtained from the dihedral database DAB is shown in figure 5.

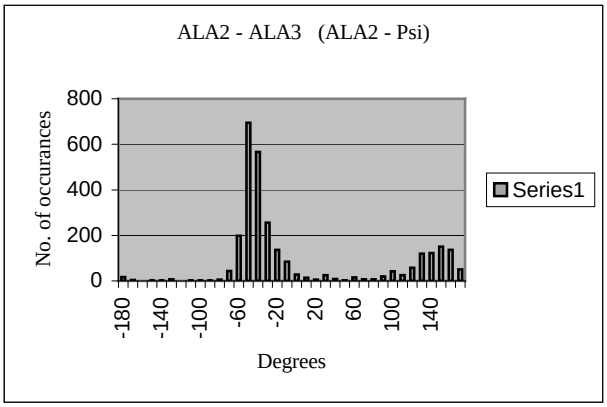

Figure 5: Psi of ALA2 in

XXX-ALA2-ALA3-XXX

This graph shows that even though the number of occurrence has reduced from more than 5000 to around 700, the distribution pattern remains the same except for some areas like 150 degrees to 170 degrees. Finally the program would query 'Extract all Psi values of ALA2 in the sub-sequence XXX-ALA1-ALA2-ALA3-XXX'. The result obtained from the dihedral database DAB is shown in figure 6.

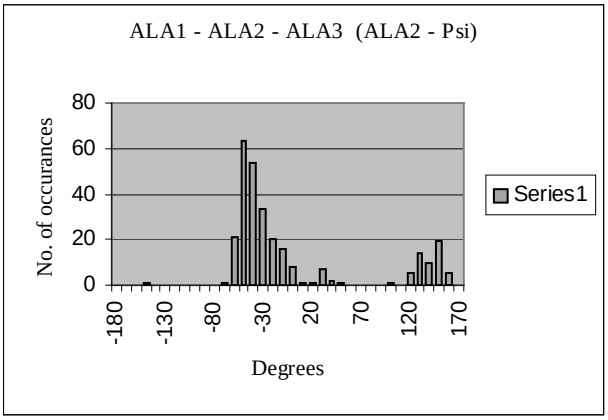

Figure 6: Psi of ALA2 in

XXX-ALA1-ALA2-ALA3-XXX

This graph clearly shows a continuing pattern in the Psi value distribution of ALA 2 in spite of being surrounded by ALA1 and ALA3. From these graphs the prediction program could conclude that the Psi value of ALA2 is unaffected by the presence of ALA1 and ALA3 and could use this information to assign the Psi value of ALA2.

As shown in the above example scenario, a prediction program would make hundreds of such queries during the prediction process. The proposed dihedral database would deal with these queries efficiently because all occurring dihedral angle values of all combinations up to amino acid length five are already stored in the database and only needs retrieving. Even though DAB has millions of entries, when a query is made, the program can directly reach the desired entry without spending time performing a sequential search. This is achieved as follows. Each amino acid is assigned a number and their combinations stored according to an order. For example, if a 3 sub-sequence combination were considered the combinations would be stored in the order 1-1-1, 1-1-2, 1-1-3 etc, and the last entry would be 20-20-20. Hence if the query has an amino acid sequence, it would translate to a number, for example the sub-sequence HIS-CYS-ALA might translate to 239. The program would now compute the dihedral values collection entry position as:

$$20^2 * (2-1) + 20^1 * (3-1) + 20^0 * (9)$$

which equals to 449. Hence the program could jump to the entry 449 to extract the dihedral values. A general formula for three combinations could be written as: If x1, x2, x3 is considered to be the amino acid sequence, then the dihedral angle collection could be calculated by,

$$20^2 * (x1-1) + 20^1 * (x2-1) + 20^0 * (x3)$$

As shown in this section, the proposed database DAB can efficiently retrieve results for queries of prediction programs. DAB overcomes the limitations of the existing databases (Sheik et al 2003 and Oleg et al 1998) by extracting dihedral values from all protein structures of the Protein Data Bank and by extracting all occurring dihedral values one of more amino acids for a given sub-sequence of amino acids, which would be used in prediction programs as shown in this section. The constant growth of PDB could be easily dealt with because the database has already entries for all combinations and hence the new dihedral values will just needed to be appended to the correct entry.

6 Conclusion

This paper proposed a dihedral angle database of short sub-sequences up to length 5. The proposed database would handle protein structure prediction program queries efficiently that is based on short sub-sequences and exact matches. Existing dihedral databases have limitations such as not being able to retrieve dihedral values for one or more amino acids occurring in sub-sequences or designed for a specific set of proteins based on its sequence identity. The database proposed in this paper overcomes these limitations by considering all proteins of PDB during dihedral angle extractions and by extracting dihedral values of one or more amino acids that occur in a specific sub-sequence.

This database could be extended in the future by including different factors of amino acids such as hydrophobic characteristics and charge. DAB can also be designed to extract values based on protein family classifications. Another important future work will be to make this database available on the World Wide Web. Including interactive graphics program where the user will be able to select the desired dihedral values for different amino acids and make analysis of the resulting dihedral values by interacting with graphs could also enhance this database.

Acknowledgments

We would like to acknowledge Victorian Partnership for Advanced Computing (VPAC),

Melbourne, for providing us with their facility. We would like to thank John Thangarajah for going through this manuscript.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990), 'Basic local alignment search tool'. *Journal of Molecular Biology*, 215(3):403—410.

Andreas, D., and Francis, B.F., 2001, *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 2nd edn, Wiley-Interscience, NY.

Attwood, T. K., Beck, M. E., Flower, D. R., Scordis, P. & Selley, J. (1998). 'The PRINTS protein fingerprint database in its fifth year', *Nucleic Acids Research*, 26, 304-308.

Attwood, T.K., and Parry, D.J., 1999, *Introduction to bioinformatics*, Prentice Hall, UK.

Bairoch, A. and Apweiler, R. (1996). 'The SWISS-PROT protein sequence data bank and its new supplement TrEMBL'. *Nucleic Acids Research*, 24(1):21-25

Bairoch, A., Bucher, P., & Hofmann, K. (1997). 'The PROSITE database, its status in 1997', *Nucleic Acids Research*, 25, 217-221.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000) 'The Protein Data Bank'. *Nucleic Acids Research*, 28, 235-242.

Bleasby AJ, Akrigg D, Attwood TK. (1994) 'OWL - a non-redundant composite protein sequence database'. *Nucleic Acids Research*; 22(17): 3574-3577.

Carl, B., and John, T., 1999, *Introduction to Protein Structures*, 2nd edn, Garland Publishing Inc, NY

Dayhoff, M.O , R. M. Schwartz, and B. C. Orcutt (1978), 'Atlas of Protein Sequence and Structure' (M. O. Dayhoff ed.), Vol.5, suppl. 3, pp. 345--352, *National Biomedical Research Foundation*, Washington D. C.

Doolittle, R.F. 1986, *Of URFs and ORFs: A Primer on How to Analyse Derived Amino Acid Sequences*. University Science Books, Mill Valley, CA.

Gabriel Cornilescu, Frank Delaglio, and Ad Bax (1999), 'Protein backbone angle restraints from searching a database for chemical shift and sequence homology', *Journal of Biomolecular. NMR*, 13, 289-302

Jones, D. T. (2000). A practical guide to protein structure prediction. *Methods in Molecular Biology*, 143**:**131-154.

Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L., and Thornton, J. M. (1997). 'PDBsum: A Web-based database of summaries and analyses of all PDB structures'. *Trends in Biochemical Science*, 22, 488-490.

Lipman, D.J. and Pearson, W.R. (1985), 'Rapid and sensitive protein similarity searches,' *Science*, Vol. 227, pp. 1435-1441.

Mark A. DePristo, Paul I.W. de Bakker, Simon C. Lovell, and Tom L. Blundell (2003), 'Ab Initio Construction of Polypeptide Fragments: Efficient Generation of Accurate, Representative Ensembles', *PROTEINS: Structure Function, and Genetics*, 51:41-55.

Mewes, H.W., Hani,J., Pfeiffer, F. and Frishman, D. (1998) 'MIPS: a database for protein sequences and complete genomes'. *Nucleic Acids Research*, 26(1), 33-37.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). 'SCOP: a structural classi- cation of proteins database for the investigation of sequences and structures', *Journal of Molecular Biology*, 247, 536-540.

Needleman, S. B., and Wunsch, C. D. (1970), 'A general method applicable to the search for similarities in the amino acid sequence of two proteins'. *Journal of Molecular Biology*, 48:443--453.

Oleg, S., Vaisman, I., Shats, A., and Sherman, S (1998), 'Conformational database for amino acid residues in protein structures'. *Fifth Electronic Computational Chemistry Conference (ECCC-5)*, November 2-30.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). 'CATH-a hierarchic classification of protein domain structures'. *Structure* 5, 1093-1108.

Rooman, M. J., Kocher, J-P. A. & Wodak, S. J. (1991). 'Prediction of Protein Backbone Conformation Based on Seven Structure Assignments: Influence of Local Interactions', *Journal of Molecular Biology*, 221, 961-979.

Sayle, R. and Bissell, A. (1992). 'RasMol: A Program for Fast Realistic Rendering of Molecular Structures with Shadows', Proceedings of the 10th Eurographics UK '92 Conference, University of Edinburgh, Scotland

Sheik, S.S, Ananthalakshmi. P, Ramya Bhargavi.G, and K. Sekar, (2003), 'CADB: Conformation Angles DataBase of proteins', *Nucleic Acids Research*, Vol. 31, No. 1 448-451.

Simon Sherman, Stanley L. Sclove, Oleg Shats and Leonid Kirnarsky (1998). 'Dihedral Probability Cluster Monte Carlo Procedure for Conformational Analysis of Proteins'. *Internet Journal of Chemistry*, 1, Article 22.

Smith, T.F. and Waterman. M.S. (1981), 'Identification of common molecular subsequences'. *Journal of Molecular Biology*, 147, *195-197.*