Computational analysis of hydrophobicity across six enzyme classes revealing relative contribution of aliphatic and aromatic residues

A.R. Chowdhury (Chakravarty)\(^1\), A. Seal\(^2\), H.G. Nagendra\(^1\)

\(^1\)Department of Biotechnology, Sir M. Visvesvaraya Institute of Technology, Bangalore, 562157, Karnataka, India; Email: nagshaila@gmail.com
\(^2\)Department of Biochemistry and Biophysics, University of Kalyani, Kalyani, 741235, West Bengal, India

ABSTRACT

One of the fundamental contributing factors, essential in folding of a linear chain of amino acids into a three dimensional protein structure is hydrophobicity and this information for protein to fold into specific geometry lies with the primary sequence itself. Thus, mapping of the hydrophobic amino acids in a protein molecule often provide ideas about the folding patterns, secondary structural arrangements, tertiary properties etc., all of which lead to functional attainment of the molecule. The present investigation intends to determine the hydrophobic constituents in various classes of enzymes obtained from different organisms. The analysis has revealed a notable point of similarity in the percentage distribution of aromatic and aliphatic hydrophobic amino acids among the different enzyme classes. The common distribution pattern of hydrophobic percentage of amino acids across the enzyme classes possibly argue in favour of a structural need, might be for folding of the peptide chains of the enzymes, even when enzyme classes are meant for diverse functions. Phenylalanine and Leucine contribute the most to aromatic and aliphatic hydrophobic content respectively across the six classes; interestingly, both happen to be at the second position in hydrophobic scale of their categories. Understanding of the hydrophobic distribution in detail seems to offer certain basic tenets and better success of protein engineering experiments.

Keywords: hydrophobicity, aromatic, aliphatic, Fauchere-Pliska scale, phenylalanine, leucine

INTRODUCTION

Distribution of twenty amino acids primarily contributes towards structural and functional properties of a protein. The side chain properties of the amino acids allow their broad classification as hydrophobic and hydrophilic. Interaction of side chains with water and neighbouring residues leads to the native conformation of the protein molecule. In his classical work, Anfinsen showed that the information for spontaneous folding of the protein lies with amino acid sequence [1]. The final conformation of the protein is supposed to be thermodynamically most stable [2]. Tanford et al. [3] have highlighted the important contribution of hydrophobicity in folding and self assembly of protein molecules. Lesk and Chothia examined hydrophobicity as one of the many factors that appeared to constrain amino acid choice in different structural environments [4,5]. The hydrophobic side chains bury themselves in the interior of a protein and provide an important driving force for protein folding [6-8] and thermodynamic stability to the molecule. It has been demonstrated that hydrophobic segments of á-helix and/or â strands get buried to form the core of globular proteins [9,10]. Distribution of hydrophobic residues is central for different structural motifs of protein molecules such as in coiled-coil strands [11,12] and in transmembrane helices [13]. Both and Sleigh [14] compared hydrophobicity of the haemaglutinnins of different viral strains for identification of the antigenic sites. Hopp and Woods [15] used the technique to predict other antigenic determinants. Kyte and Doolittle [16] mapped the hydrophobic sequences of
soluble and membrane bound proteins and introduced the method of prediction of interior and exterior regions of the molecules. A related analysis of hydrophobic moments to distinguish transmembrane from globular proteins was shown by Eisenberg [17] in 1984. This method has been used by Engelman et al. [18] to identify transbilayer helices in membrane proteins.

The distribution of hydrophobic residues has the imprint for folding and function of proteins. The present investigation intends to analyze the percentage distribution of hydrophobicity in reference to aromatic and aliphatic amino acids in several members of each of the six enzyme classes from different organisms. The contribution to total hydrophobicity of individual amino acid was calculated after assigning values to the amino acids based on Fauchere-Pliska hydrophobic scale [19]. This scale reflects hydrophobicity in the context of denatured proteins and/or small synthetic peptides [20]. Hydrophobicity of a residue is represented by the free energy of transfer by one mole of residue from an apolar solvent to water [21]. The values of this hydrophobic scale is the free energy of the transfer of amino acid analogues from octanol to water [17, 19]. Investigators like Kurgan et al. [22] used this scale for computational biology related to prediction and analysis of secondary structure of proteins. Margalit et al. [23] used the scale for predicting immunodominant helper T cell antigenic sites.

The present study aims to depict the contribution of hydrophobic residues in polypeptides across enzyme classes obtained from same and different species/organisms and to understand the maximum and minimum contribution of any particular residue towards hydrophobicity. The occurrence of such hydrophobic distribution can provide a better basis for the advancement of protein engineering experiments.

MATERIALS AND METHODS

Creation of non-redundant sequence database

Protein sequences were downloaded from Protein Data Bank (PDB) [24]. A non-redundant sequence database was created with the aid of Representative Protein Database [25]. PDB, an authentic structural database contains the structural information of these sequences, and thus, enables the correlation between hydrophobicity and the three dimensional geometry of the protein molecule as well. Sequences of length greater than 40 residues were only chosen for the database. To have full length original sequences, mutants and fragments were excluded from the database. Further, to avoid duplication of sequences in the database, only x-ray crystallographic structures were considered. Identity among the sequences in the database was restricted to less than 30% to maintain non-redundancy of the database. Finally, the non-redundant sequence database was created with 3551 sequences in it.

Program outline

A C++ program calcu.cpp was devised to find the hydrophobicity due to aromatic and aliphatic amino acids of a protein sequence, considering the sequence to be a linear array of string characters i.e. amino acids. The hydrophobicity scale determined by Fauchere and Pliska [19] for amino acids has been employed here to calculate the hydrophobicity of the protein sequence. Basic algorithm of the program to find the hydrophobicity of each sequence due to both aromatic and aliphatic residues is stated below and the flowchart showcasing the summary of these steps is shown in figure 1. Appropriately, a protein sequence was taken from the non-redundant database created. Each amino acid of the protein sequence was read one by one by the program and through a decision making process, each amino acid was identified. A numeric variable, as per the single letter code of each amino acid, was assigned for each of the hydrophobic aromatic and aliphatic amino acids to count the number of different amino acids in each protein sequence. So, hydrophobic content due to a particular amino acid = number of that amino
acid multiplied by hydrophobic scale value of that particular amino acid in Fauchere-Pliska hydrophobic scale, i.e., WH is the hydrophobicity due to Tryptophan residues in a sequence. In a similar manner, hydrophobicities due to other aromatic and aliphatic residues were calculated. Classification of Rees et al. [26] and Arthur Lesk [27] for aromatic, aliphatic hydrophobic amino acids has been followed in this work. The program assigned \( \text{TotAroH} \) the value of total hydrophobicity in a sequence due to all aromatic residues and calculates it as the summation of hydrophobicities due to Tryptophan (WH), Phenylalanine (FH), Tyrosine (YH) and Histidine (HH), i.e., \( \text{TotAroH} = WH + FH + YH + HH \). Similarly, \( \text{TotAliH} \) was the variable which calculates the summation of hydrophobicity due to the hydrophobic aliphatic residues, Isoleucine (IH), Leucine (LH), Cysteine (CH), Methionine (MH), Valine (VH), Proline (PH) and Alanine (AH), i.e., \( \text{TotAliH} = IH + LH + CH + MH + VH + PH + AH \). Finally, all the numeric variables involved in the calculation of hydrophobicity were initialized to zero and the whole process of counting each amino acid and their respective hydrophobicity as well as the total hydrophobicity under aromatic and aliphatic category of a sequence gets repeated for all the sequences in the database. The results of all the sequences were saved in the file data.txt.

**Sequence Category**

The list of sequence IDs under the six enzyme classes - Ligase, Isomerase, Lyase, Oxidoreductase,
Transferase and Hydrolase were downloaded from PDB [24], and saved as six separate text files. The program filter.cpp (Figure 2) was framed to extract the details of the calculation of hydrophobicity, due to hydrophobic aromatic and aliphatic amino acids from data.txt, for those sequences whose ID matched the ID of a particular sequence under a specific enzyme class from the PDB downloaded list mentioned above. At the end of the execution of the program filter.cpp, separate files, namely, lig.txt, isom.txt, lya.txt, oxidored.txt, transf.txt and hydrol.txt, were obtained for the six enzyme classes respectively. Total Hydrophobicity% (TH%) is calculated according to the formula TH% = (Total Sum of hydrophobic scale values for all amino acids present in a sequence/Length of Sequence) X 100. It was found that maximum number of sequences lie in the TH% range 20%-50% which agreed with the distribution shown in figure 3. So, we have analyzed here those sequences from the various enzyme classes which lie in this TH% range and the number in the different classes has been listed below in table 1. The enzyme classes are arranged in the ascending order of their number in the class.

**Calculation of hydrophobicities**

The hydrophobicities of all the sequences due to the hydrophobic aromatic and aliphatic residues were calculated separately as below:

\[ \text{TotAroH}\% = (\text{TotAroH/Length of a Sequence}) \times 100 \]  

(1)

<table>
<thead>
<tr>
<th>Enzyme Classes</th>
<th>Number of Sequences in TH% range 20-50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ligase</td>
<td>71</td>
</tr>
<tr>
<td>Isomerase</td>
<td>118</td>
</tr>
<tr>
<td>Lyase</td>
<td>186</td>
</tr>
<tr>
<td>Oxidoreductase</td>
<td>363</td>
</tr>
<tr>
<td>Transferase</td>
<td>430</td>
</tr>
<tr>
<td>Hydrolase</td>
<td>571</td>
</tr>
</tbody>
</table>

Table 1. Number of sequences analyzed in each enzyme class.

Figure 3. Variation of total hydrophobicity (%) of all the sequences in the database.
where TotAroH is the hydrophobic content in a sequence due to the hydrophobic aromatic residues.

\[
\text{TotAliH} = \left( \frac{\text{TotAliH}}{\text{Length of a Sequence}} \right) \times 100 \quad \ldots \ldots \quad (2)
\]

where TotAliH is the hydrophobic content in a sequence due to the hydrophobic aliphatic residues. The data for distributions of TotAroH% and TotAliH% in the six enzyme categories are in Table 2. The individual amino acid like Tryptophan’s contribution to aromatic hydrophobic content has been calculated as

\[
\text{WH} = \left( \frac{\text{WH}}{\text{TotAroH}} \right) \times 100 \quad \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldot
The first and third quartile values of the respective distributions have been mentioned (table 2) as the concentration of sequences is maximum in this extent. The mean and standard deviation values of these distributions are also provided. From the mean values and the values for the standard deviation (SD) mentioned in table 2, it is evident that the hydrophobic percentage of both aromatic and aliphatic residues of six different enzyme classes are within a similar range. The hydrophobic contribution of aliphatic residues in all protein sequences analyzed here is more than that of the aromatic residues.

There is a correlation between the numbers of aromatic residues and aliphatic residues in a sequence, the correlation coefficient between them in six classes of enzymes are given in table 3. In the figure 4 and 5, the hydrophobic contributions from respective aromatic and aliphatic residues have been illustrated. Here also, we have chosen the first and third quartile values to indicate the extent of variation for WH %, FH %, YH % and HH % as well as for IH %, LH %, CH %, MH %, VH %, PH % and AH % as the majority of the sequences lie in this range. To supplement this fact, we have also provided the mean and standard deviation (SD) values for the respective contributions from all the hydrophobic aromatic and aliphatic residues for all the enzyme classes. The higher standard deviation values as usual indicate that some of the values are spread over a wider range. The dots in the figure indicate the mean values for each of the bars which correspond to the extent of variation for WH%, FH%, YH% and HH% respectively in each of the plots of the aromatic category in figure 4 and for IH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %, MH %, LH %, CH %,
Figure 5. Percentage distribution to show the contribution from different hydrophobic aliphatic residues to total aliphatic hydrophobic content in six different classes of enzymes.

%, VH %, PH % and AH % respectively in the graphs in figure 5 of the aliphatic group. From figure 4, Phenylalanine emerges as the major contributor and Histidine as the minimum whereas in figure 5, Leucine shows the largest contribution and Cysteine the least.

The values of TotAroH% and TotAliH% are in similar range in the different enzyme classes (Table 2). The similarity of the hydrophobicity% in the different classes of the enzymes might have implication for basic structural stability in them. The contribution of aliphatic amino acids towards percentage of hydrophobicity in the enzyme classes is more than that of aromatic residues. This is likely for additive values of seven aliphatic residues versus four aromatic amino acids [26,27] under consideration in this study. The probability of occurrence of seven aliphatic residues in a given length of a protein is more than the four aromatic residues. The strong positive correlation coefficient between the number of aromatic residues and aliphatic residues in sequences of the different enzyme classes (Table 3) signifies that with the increase in number of aromatic residues, the number of aliphatic residues also increases in all cases.

After studying the contribution of all aromatic and aliphatic residues to the hydrophobic percentage of six different classes of enzymes, the contribution of particular amino acids in each group revealed that, Phenylalanine notably contributes most to the total aromatic hydrophobic content and the one that contributes least is Histidine, for all enzyme classes (Figure 4). Similarly, in case of aliphatic hydrophobic content, Leucine is the highest contributor while Cysteine contributes the least amongst the enzyme classes (Figure 5). According to Fauchere and Pliska hydrophobic scale [19], used in this study, Tryptophan is the most hydrophobic aromatic amino acid, followed successively by Phenylalanine, Tyrosine
and Histidine. Though being at the second position in the scale adopted, Phenylalanine has emerged as the highest contributor to aromatic hydrophobic content, which indicates that Phenylalanine occurs more frequently than Tryptophan, in all enzyme classes.

Like wise, Histidine with the lowest value in the hydrophobic scale is also the least contributor to aromatic hydrophobic content of the enzymes (Figure 4). This might have implications for the presence of Histidines at the active site of the enzyme molecules [28-30]. With reference to the hydrophobic content of aliphatic amino acids, Isoleucine emerges at the top of the hydrophobic scale of Fauchere-Pliska, followed consecutively by Leucine, Cysteine, Methionine, Valine, Proline and Alanine. Interestingly, though Leucine is at the second place of the scale (with 0.1 value difference with Isoleucine), yet it contributes the most towards hydrophobic content of aliphatic category. Although Cysteine is at the third position in hydrophobic scale value, it appears as the least contributor to aliphatic hydrophobic content, probably owing to its limited occurrences [31].

It is noteworthy that both in case of aromatic and aliphatic residues, the residue at the second position of hydrophobic scale contribute most to the total hydrophobic percentage across all the enzyme classes, namely Phenylalanine and Leucine. Like-wise, in spite of diversity in the source organism and specific function of the six enzyme classes, the contributions of different aromatic and aliphatic residues towards hydrophobic content of the sequences are similar, highlighting the fact that the preservation of fundamental folding characteristic [7] of the polypeptide chains to attain the catalytic activity across the enzyme classes gains utmost significance. Though several authors [32-34] have proposed that the conserved characteristic of hydrophobicity provides structural stability to the protein molecules, and the choice of a particular hydrophobic residue to get replaced during the course of evolution by another amino acid similar in volume and hydrophobicity [1] is natural, our investigations have objectively indicated the statistical contributions of the individual aromatic and aliphatic amino acids towards their total aromatic and aliphatic hydrophobicities for these enzyme molecules, which would necessarily assist critical understanding of the structural and functional tenets of enzymes, enabling better design of the same for novel requirements.

Acknowledgements: We acknowledge the support extended to the work by the Department of Biochemistry and Biophysics, University of Kalyani, Kalyani and the Department of Biotechnology, Sir M. Visvesvaraya Institute of Technology managed by Sri Krishnadevaraya Educational Trust, Bangalore.

REFERENCES