Mechanism of Amino Acids Substitution and Substitution Model: Theory and Practice by Bioinformatics Approach

THESIS SUBMITTED TO THE UNIVERSITY OF BURDWAN
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN BIOTECHNOLOGY

BY

PARTH SARTHI SEN GUPTA

DEPARTMENT OF BIOTECHNOLOGY
THE UNIVERSITY OF BURDWAN
BURDWAN, WEST BENGAL, INDIA
2017
Dedicated
to my
Father
ACKNOWLEDGEMENT

First and foremost, I express my sincerest gratitude to Dr. Amal Kumar Bandyopadhyay, Assistant Professor, Department of Biotechnology and Dr. Inul Ansary, Assistant Professor, Department of Chemistry, The University of Burdwan for their constant encouragement, advice and insightful scientific expertise throughout the Research. Without their active participation and encouragement, this research work would not have materialized.

I am highly grateful to the Head, Department of Biotechnology, and Coordinator of DBT, Govt. of India, Dept. of Biotechnology and all the faculty members for their generous help and support. I am also thankful to all non teaching staff members of the department for their time to time helps.

I am grateful to the Dean, Faculty Council for PG Studies in Science, Prof. Bimalendu Ray, Secretary, Faculty Council of Science, Mr. Subha Prasad Nandi Mazumdar, and all other academic and administrative members of the University of Burdwan for their support and help.

I am thankful for computational facility laboratory funded by DBT, Govt. of India, in the Department of Biotechnology, The University of Burdwan.

I would like to thank the Higher Education Department, West Bengal, India, for the award of the fellowship.

I would like to thank my lab-mates Shyamashree Banerjee and Arnab Nayek for their nice company, exchanges of knowledge and skills during my Ph.D program.

I would also like to thank Prof. Nathan Baker, Pacific Northwest National Laboratory for his suggestions during the development of software ADSETMEAS.

I would like to thank from the core of the heart to my wife Mrs. Madhuri Sengupta for her constant love, support and encouragement.

Last but not the least, I express my gratitude to the family members for their constant love and support.

August, 2017

Department of Biotechnology
The University of Burdwan
Burdwan, West Bengal, India

Parth Sarthi Sen Gupta
# Table of contents

<table>
<thead>
<tr>
<th>Chapters</th>
<th>Title</th>
<th>Page number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbreviation</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Preamble</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Synopsis</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>General Introduction</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>Development of computational tools</td>
<td></td>
<td>35</td>
</tr>
<tr>
<td>Chapter 1</td>
<td>Understanding the Role of Amino Acids Substitution in the Evolution of Homologous Protein Families</td>
<td>74</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>An Insight into the Variability, Conservation and Substitutions of Protein BLOCKs</td>
<td>100</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Substitutions in common shared region of plant-type ferredoxins in Relation to Salt Adaptation of Halophilic Ferredoxin</td>
<td>121</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Are Amino acid Substitution matrices really Universal</td>
<td>138</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Screening of Disease Associated Mutations in HGO proteins and its Structural Consequences using Computational Methods</td>
<td>149</td>
</tr>
<tr>
<td>Chapter 6</td>
<td>Molecular Modeling of Synthesized Novel N-acyl Substituted Benzoazetinones and Their Docking Simulation on Antifungal Modeled Target.</td>
<td>167</td>
</tr>
<tr>
<td>List of Publications</td>
<td></td>
<td>177</td>
</tr>
</tbody>
</table>
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>APBS</td>
<td>Adaptive Poison-Boltzmann Solver</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root mean square deviation</td>
</tr>
<tr>
<td>RMSF</td>
<td>Root mean square fluctuation</td>
</tr>
<tr>
<td>Rg</td>
<td>Radius of gyration</td>
</tr>
<tr>
<td>MSA</td>
<td>Multiple Sequence Alignment</td>
</tr>
<tr>
<td>NAMD</td>
<td>Nano scale molecular dynamics</td>
</tr>
<tr>
<td>VMD</td>
<td>Visual molecular dynamics</td>
</tr>
<tr>
<td>hFD</td>
<td>Halophilic Ferredoxin</td>
</tr>
<tr>
<td>FD</td>
<td>Ferredoxin</td>
</tr>
<tr>
<td>GDH</td>
<td>Glucose 1-dehydrogenase</td>
</tr>
<tr>
<td>GPL</td>
<td>General Public License</td>
</tr>
<tr>
<td>CS</td>
<td>Conservative substitution</td>
</tr>
<tr>
<td>Kcal</td>
<td>Kilo Calorie</td>
</tr>
<tr>
<td>NCS</td>
<td>Non Conservative substitution</td>
</tr>
<tr>
<td>pFD</td>
<td>Plant Ferredoxin</td>
</tr>
<tr>
<td>MP</td>
<td>Mesophilic</td>
</tr>
<tr>
<td>cFD</td>
<td>Cyanobacterial Ferredoxin</td>
</tr>
<tr>
<td>MW</td>
<td>Molecular Weight</td>
</tr>
<tr>
<td>NDK</td>
<td>Nucleoside Diphosphate Kinase</td>
</tr>
<tr>
<td>PCNA</td>
<td>Proliferating Cell Nuclear Antigen</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>pI</td>
<td>Isoelectric Point</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic local alignment search tool</td>
</tr>
<tr>
<td>SIFT</td>
<td>Scale invariant feature transform</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SDM</td>
<td>Site directed mutator</td>
</tr>
<tr>
<td>SASA</td>
<td>Solvent Accessible Surface Area</td>
</tr>
<tr>
<td>PhD-SNP</td>
<td>Predictor of human Deleterious Single Nucleotide Polymorphisms</td>
</tr>
</tbody>
</table>
PREAMBLE

Protein evolves by means of the mechanisms of substitution, deletion and insertion, under structural and functional constraints, to produce diverse orthologous sequences of identical function in different species that includes representatives from all domains of life. The wealth of molecular information in these sequences and structures, has been the source of bioinformatics studies of which some of the recent focus is to understand i] role of substitution in evolution ii] site-specific mechanistic details, iii] role of substitution in functionally equivalent codes working under diverse environmental conditions, iv] BLOSUM scores and its utility in defining relatedness among these sequences, v] role of lethal substitutions in diseases and vi] target-ligand interactions in pathogenic context. While above aspects of studies are promising, the field made very little. Some conjectures though were made in that the non-conservative and conservative substitutions play crucial role in site specific manner by utilizing some hitherto unknown principle of evolution. With a view, to achieve the detailed milestones for these challenging issues of bioinformatics, the present thesis work has been undertaken. The realization that fully automated procedures would be the first step for accomplishment of the work, a series of efficient and user-friendly softwares were developed whose implementation along with other web-based GPL tools demonstrate for the first time that i] the site-specific, unlike BLOCK-specific, evolutionary rate vary greatly and that acts as determinant for structural and or functional modulation, ii] the observed change in codes with reference to the normal, are postulated to be neutralized when kept in its respective environment, iii] the score of BLOSUM matrices are not fundamental but are heavily dependent on BLOCK parameters (such as length, width of BLOCK etc) and thereby highlighting the need for the development of matrices de novo, iv] β-lactum derivatives could be potential drug candidates for inhibiting pathogenic fungi and v] lethal mutations are screened and characterized using novel designed in silico procedure. Taken together, the thesis work performs in-depth analyses of bio-materials (sequences and structures of proteins) using our novel developments along with other authentic GPL tools to address for the first time some front-lined fundamental problems, the result of which seems to have far-reaching and long-term applications in the field of bioinformatics and computational biology.
SYNOPSIS

Mechanism of Amino Acids Substitution and Substitution Model: Theory and Practice by Bioinformatics Approach

Introduction

Bioinformatics involves retrieval of information using integration of computers, Computational tools, Statistical methods and databases for addressing biological questions involving DNA and proteins. Proteins which are formed due to divergence of two genes after a speciation event are orthologous whereas paralogous proteins are resulted due to duplication (Fitch W, 1970). The former evolved from the same ancestor and normally retain the same function during evolution. During evolution, these sequences accommodate many changes by the mechanism of insertion, deletion and substitution. Substitution is an evolutionary mechanism wherein amino acid changes at locus specific positions may results into the modulation of structure and function. Thus major concern is to understand the reason behind this modulation.

Due to advent in the databases of DNA and protein, thousands or even millions of DNA and protein sequences are available to gain the knowledge of many substitutions or mutations. These databases are rapidly increasing day by day due to the advancement of genome sequencing technology and the improvement of X-ray crystallography and NMR spectroscopy methods. However, for intricate balance between the entry of sequences and structures in the database and to understand the biological function of these proteins, needs efficient and batch analyzing software. So in this context, it would be a major challenge to develop need specific efficient and mass scale analyzing software to catch the rapidity of databases as well as for the analyses of site specific codes.

Site-specific codes in sequence in the presence of appropriate environmental conditions form functional state of proteins (Anfinsen CB, 1973; Chotia C, 1984; Liberles et. al,
Different types and levels of substitutions are acquired at different sites of homologous sequence (Reidhaar Olson and Sauer, 1988) of which invariant and conservatively substituted sites were found to be crucial for structure and function. Non-conservative substitutions on the other hand were found to have little or no such roles (Bowie, et. al., 2015). The fact that both level and type of substitutions along with site-specific residue’s contribution to variability and conservation play crucial role in protein’s adaptation, robust all-in-one method would be useful for the establishment of BLOCK specific mechanistic model for substitutions (Echave, et. al., 2016).

PAM (Dayhoff et. al. 1978) and BLOSUM (Henikoff and Henikoff 1992) are two fundamental matrices of which the former is explicit and the later follow implicit in nature. BLOSUM62 is the best among all BLOSUM series of matrices as it works better for distantly related sequences and thus it gained much popularity in bio-computations. It is the most widely used matrices for the purpose due to its better performances relative to other ones. Bit-score was obtained for all 210 possible substitution pairs of the 20 standard amino acids. The procedure involves series of steps. Firstly, a large number of sequence blocks (>2000) are to be made maintaining a level of similarity (say ≥62%). Secondly each block is to be analyzed for obtaining scores of 210 pairs of which 20 are homo-substitution-pairs and 190 are hetero-substitution pairs. Finally, the scores thus obtained for all sequence BLOCKs are averaged to get the bit score matrix (Henikoff and Henikoff 1992). The final matrix is the BLOSUM matrix, used in conjunction with many popular web based programs such as CLUSTALW (Thompson et al. 1994), BLAST (Altschul et al. 1999), etc. Are the score of amino acid Substitution matrices really universal for all the thermophilic, mesophilic and halophilic proteins?

The knowledge of substitution in protein sequence is of great importance in understanding protein evolution and adaptation in a given environment. While mesophilic ferredoxin (mFD) and halophilic ferredoxin (hFD) belong to plant-type family, the former takes part in oxygenic photosynthetic reaction in membrane bound form and the later in contrast is soluble in the cytoplasm and participates in oxidative
decarboxylation reaction in the presence of high salt. Rao et. al, 1971 hypothesized a novel “additional protein moiety” in mFD for its effective interaction with photosynthetic lamellar systems. Similar insertion in the N-terminal region of hFD was proposed for its high salt adaptation (Oren A, 2002; Britton et. al, 1998). However, the differential role of amino acid substitutions in the common shared region (CSR) of these molecules remains to be worked out.

Substitution of amino acid can affect the protein dramatically. It could effect on stability, hydrogen bonding, interaction, conformation, activity, cellular localization and other physico-chemical properties of the protein (Stefl et al, 2013). It can also influence drug resistance and susceptibility to diseases resulting into serious pathological condition. Understanding the effect of substitution or mutation is utmost needed when one considers genetic diseases. It is well known that how the substitution of one residue in a protein lead to several diseases such as Sickle cell anaemia, Alzheimer’s, Parkinson’s and Creutzfeldt-Jakob diseases (Capriotti et. al, 2006) etc. Thus it would be interesting problem to understand the mechanisms of the effects of these substitution.

**Objective of the Study**

1) To develop softwares and algorithms for mass scale sequence and structure analysis.
2) To gain insight into the mechanistic details of site specific substitutions.
3) To understand the role of substitution in functionally equivalent codes.
4) To understand the role of substitution in adaptation and molecular evolution.
5) To check the consistency of score in substitution matrices.
6) To understand the role of substitution in diseases using molecular dynamics and docking.

**Plan of the work**

1. Extraction of large datasets with relevant annotations from Sequence database (UNIPROT).
2. Extraction of atomic structures along with relevant structural, biological, fold and domain (SCOP, CATH) parameters from the Protein Database (RCSB).

3. Development of mass scale analyzing efficient bioinformatics tools for the sequence and structure analyses.

4. Preparation of BLOCK from multiple sequence alignment file using our home built software.

5. In silico simulation to check the consistency of amino acid substitution score.

6. Preparation of difference matrix for the common block containing all taxonomic subgroups and preparation of phylogenetic tree with branch length correction to understand the phylogenetic rank of hFD in relation to plant and cyanobacterial ferredoxins.

7. Development of novel parameters to understand the diversity of amino acid substitution in evolution.

8. Extraction of salt bridges from crystal structure and their electrostatic energy calculation using our home built developed software and Adaptive Poisson Boltzmann Solver (APBS).


10. Molecular modeling, docking and dynamics study to understand the effect of substitution in diseases.

**Methodology in brief and significant findings**

**Development of Bioinformatics tools**

The automated standalone, multipurpose procedure for analysis of protein sequences and crystal structure of protein are developed. The program works from UNIX like environment including CYGWIN and are interpreted by AWK programming language. They work on various forms of FASTA files (Input) and crystal structures (input). Based on the form of FASTA-input and crystal structure, relevant outputs are produced in easy to use excel format.
The programs are designed such that large database of sequences and structures can be analyzed in a single run. Overall they provide countable advantages over the existing web-tools. I have categorized the bioinformatics tools into sequence based and structure based software (Table 1).

Table 1: name of the tools, its availability and suitable citations for sequence and structural analysis are presented.

<table>
<thead>
<tr>
<th>SI. No.</th>
<th>Name of the tool</th>
<th>Availability</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PHYSICO</td>
<td>Available through Email</td>
<td>Gupta et. al, 2014</td>
</tr>
<tr>
<td>3</td>
<td>APBEST</td>
<td><a href="https://sourceforge.net/projects/apbest/">https://sourceforge.net/projects/apbest/</a></td>
<td>Gupta et. al, 2017</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SI. No.</th>
<th>Name of the tool</th>
<th>Availability</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>SBION</td>
<td>Available through Email</td>
<td>Gupta et. al, 2014</td>
</tr>
<tr>
<td>6</td>
<td>ADSBET</td>
<td><a href="https://sourceforge.net/projects/adsbet/">https://sourceforge.net/projects/adsbet/</a></td>
<td>Nayek et. al, 2015</td>
</tr>
<tr>
<td>7</td>
<td>ADSBET2</td>
<td><a href="https://sourceforge.net/projects/adsbet2/">https://sourceforge.net/projects/adsbet2/</a></td>
<td>Nayek et. al, 2015</td>
</tr>
</tbody>
</table>

**Mechanistic details of site specific substitutions**

A total of thirty homologous protein families with known divergence rate (Dayhoff and Schwartz, 1981) are taken in the present study. These families were chosen in such a way that their divergence rate gives a wide coverage. For example H2A has 0.5% per 100/mYr and that for Kappa-casein is 33% per 100/mYr (Dayhoff and Schwartz, 1981). Family specific sequences were obtained from UNIPROT (UniProt Consortium, 2008) and other inter-linked databases. Obtained sequences were then aligned using ClustalW2 (Larkin et al., 2007) for each of the protein families for the construction of BLOCK FASTA file using automated block preparation tool of PHYSICO2 (Banerjee et al., 2015).

Relation between developed evolutionary parameter R \([\text{conservative (CS)/non-conservative substitution (NCS)}]\) and Diversity is shown in Figure 1. How much diversity could be introduced in protein during the course of evolution? Interestingly,
we have found that non-conservative substitution should be less than conservative substitution i.e R<1 (Figure 1).

The analysis of the work has been done by using APBEST (Gupta et. al, 2017). APBEST uses many equations and formulae for its implementation (figure 2).

Figure 1 Figure showing relationship between R [conservative (CS)/non-conservative substitution (NCS)] and divergence rate (DV) for thirty orthologous protein families

Figure 2 Formulae, equations implemented in APBEST and their clarity
Substitution in functionally equivalent codes

Preparation of phylogenetic tree with branch length correction is done using MEGA 7 software to understand the phylogenetic rank of Halophilic ferredoxin (hFD) in relation to plant (pFD) and cyanobacterial ferredoxins (cFD).

We have found that hFD is distantly related to Mesophilic ferredoxin (pFD and cFD) (Figure 3).

![Figure 3 Phylogenetic relationship between hFD (violet color), Universal Shock protein (USP, red color), pFD and cFD](image)

From figure 3 following results are obtained:

- CSR of hFD differs 70% from mFD
- USP & Chaperon (CHP) form sister-clade with hFD
- USP appears earlier than hFD
- hFD branched out from eukaryotic parent

hFD establishes designed long-range and networked salt bridges involving halophilically substituted and conserved acidic and basic residues both at N-terminal and C-terminal regions (Figure 4). In mFD although salt bridges are detected for both the regions, the design suffers from positional residues conservation and hence universality of these salt bridges. Unlike mFD, hFD introduces new salt bridges using halophilically conserved and substituted residues. In active site the salt bridge remains
unaffected for both hFD and mFD. It thus indicates these design salt bridges in the “common region” of the sequence play crucial role in adaptation of hFD.

Figure 4 Comparison of salt-bridge/ion-pair pattern at N-terminal (h1 vs m1), internal (h2 vs m2), C-terminal (h3 vs m3) and active site region (h4 vs m4) for hFD and mFD. In each case salt-bridge/ion-pair residues are located both in the structure and aligned BLOCK using different colored lines. Green, red, blue and pink colored lines are used to indicate N-terminal, internal, C-terminal and active site salt-bridge/ion-pair respectively. The BLOCK highlights both the residue of interest and their domain specific positional conservation using Shannon Entropy criteria. Secondary structural regions are also highlighted using similar color code as Figure 1. Halophically (h), plant (p) and cyanobacteria (c) specific conservation of (Shannon Entropy <=1.0) residue were obtained by PHYSICO2 analyses using domain specific BLOCK FASTA files (sequence >75). Non-conserved residues are shown as n and globally conserved residues are shown as g.
Universality of BLOSUM matrices

Home Build procedures were written:

a. To simulate BLOCKs parameters such as length, width etc

b. To obtain average matrix from 5000 natural sequence BLOCKs.

c. Programs run in UNIX environment and are interpreted by GAWK and C Programming languages

d. It implements following equations (Henikoff and Henikoff, 1992):

Observed probability of occurrence of each \(i,j\) pair is

\[
q_{ij} = \frac{f_{ij}}{\sum_{i=1}^{20} \sum_{j=1}^{i} f_{ij}}
\]

Let the total no. of amino acid \(i,j\) pairs \((1 \leq j \leq i \leq 20)\)

Expected probability of occurrence \(e_{ij}\) for each \(i,j\) pair is then \(p_{ij}\)

For \(i = j\) and \(p_{ipj} + p_{jpi} = 2p_{ij}\) for \(i = j\).

\[
p_i = \left( q_i + \sum_{j=1}^{20} \frac{q_{ij}}{2} \right); \quad e_{ij} = p_i p_j (i = j); \quad e_{ij} = 2p_i p_j (i \neq j)
\]

Log odd Score is

\[
\text{Odd Score: } s_{ij} = 2 \log_2 \left( \frac{q_{ij}}{e_{ij}} \right)
\]

After simulation of BLOCKs, we have found that BLOSUM matrices are not universal as the score of the matrices changes with BLOCK parameter such as length and width (Figure 5).

Figure 5: Effect of width and length on W1, L1) observed probability W2, L2) expected probability W3, L3) bit score W4, L4) ratio
Effect of substitution in diseases

To understand the role of substitution in diseases, we adopt a method shown in figure 6. MDS and docking results (not shown) showed stability loss in mutant HGO protein. Due to mutation, HGO protein became more flexible and alters the dynamic property of protein which might affect the interaction with target peptide. The results obtained from this study would facilitate wet-lab researches to develop potent drug therapies against Alkaptonurea.

![Figure 6: Flowchart describing the workflow implemented in this study](image)

References


Oren A. Adaptation of halophilic archaea to life at high salt concentrations. InSalinity: Environment-Plants-Molecules 2002 (pp. 81-96). Springer Netherlands.


Publications


Analyses of Salt-Bridges from Multiple Structure Files. Bioinformation. 2014,10 (3), 164


Symposium/Seminars Presentation

- **Parth Sarthi Sen Gupta** and Amal K Bandyopadhyay. Odd Amino Acid Substitution Matrices Are Really Odd. FIMB 2012, IISER Kolkata.
- **Parth Sarthi Sen Gupta, Rifat Nawaz Ul Islam**, Vishma Pratap Sur, Sahini Banerjee and Amal K Bandyopadhyay. New Method to Assess Substitutions and
Variability of Different Sites of Protein. 21st CRSI national symposium in chemistry with American Chemical Society (ACS). July 14-16, 2017

- **Parth Sarthi Sen Gupta** and Amal K Bandyopadhyay. Alteration of Codes in the Common Shared Region of Halophilic Ferredoxin for its Adaptation in High Salt. 21st CRSI national symposium in chemistry with American Chemical Society (ACS). July 14-16, 2017
General Introduction

Orthologous proteins are formed due to divergence of two genes after a speciation event whereas paralogous proteins are resulted due to duplication (Fitch W, 1970). They evolved from the same ancestor and normally retain the same function during evolution. During evolution, these sequences accommodate many changes by the mechanism of Substitution, Insertion and Deletion (Figure 1). Mechanism of substitution is the changes of amino acids at locus specific position and is the central theme of the thesis. Amino acid substitution is an evolutionary mechanism wherein amino acid changes at locus specific positions may results into the modulation of structure and function. Thus major concern is to understand the reason behind this modulation. Since the determination of protein sequence and structure, it has been clear that to understand biological processes, positioning and properties of amino acids is of major significance (Pal et al., 2006).

Suppose an amino acid changes in locus specific position of a protein then what could be the effect on its overall structure and function. Take the example of Haemoglobin (first determined structure), a single amino acid substitution of Glutamate in normal individual with Valine, resulting into the disease sickle cell anaemia. The substitution leads to increase in hydrophobicity and resulting into the sickle shaped cell that give the disease its name. Due to advent in the Databases of DNA and protein, thousands or even millions of DNA and protein sequences are available to gain the knowledge of many mutations. Substitutions may occur at a wide variety of evolutionary distances within one species, or between species. During evolution, there are two types of substitution, conservative and non-conservative occurs of which most of the later changes are deleterious. These amino acid exchanges are constrained by the requirement of maintenance of structure for the functionality of the protein. Certain amino acid changes tend to have little or no effect on the overall protein structure or function (conservative substitution) and observed more frequently than the changes that have influence (non-conservative substitution). Whether certain substitutions effects on protein function or not is often unknown; however, many efforts are under way to quantify these effects (Karchin et al., 2005; Ng et. al, 2006).
The goal of protein sequence comparison for understanding the effect of substitution on its adaptation, Evolution and modulation of its structure and/or function is to take a protein sequence and search the protein database to find the homologous sequences. Searching homologous sequences means to identify protein sequences that shared common ancestor. Homologous protein sequences can be divided into two groups: (1) orthologous sequences, that differ because they are found in different species; and (2) paralogous sequences, that differ due to gene duplication event. Orthologous protein sequences accumulate both neutral and naturally selected substitutions at locus specific position in the course of evolution (McDonald et. al, 1991).

**Substitution or Mutation and Evolution**

According to the neutral theory of molecular evolution; stochastic fixation of mutations is the most important driving force behind substitutions. Kimura suggested that the majority of genetic substitutions were the result of random fixation of neutral or nearly neutral
mutations (Kimura, 1968). Positive selection does operate, but in general the size of effective population is so small in comparison to the magnitude of the selective forces that the contribution of positive selection to evolution is too weak to shape the genome significantly.

Table 1. Evolutionary forces affecting speciation

<table>
<thead>
<tr>
<th>Force</th>
<th>Variation within population</th>
<th>Variation between populations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutation</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Inbreeding or genetic drift ($N_e$ is small)</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Migration ($\leftrightarrow$ isolation)</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Selection</td>
<td>Positive/negative</td>
<td>+/-</td>
</tr>
<tr>
<td>Balancing</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Incompatible</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

$N_e$: effective population size.

According to the theory, only a small minority of mutations become fixed because of positive selection. Organisms are, in general, so well adapted to the environment that many non-synonymous mutations are deleterious and, therefore, quickly removed from the population by negative selection. Stochastic events predominate and substitutions are mainly the results of random genetic drift. Evolutionary forces which affect the speciation event are shown in Table 1 (Salemi et. al, 2009).

Protein sequence (arrangement of amino acid residues) contains codes (e.g. helix, sheet, active site, binding site etc) that help to form its functional structure. Sequences that are evolved from common parent but exist in different organism and perform similar function are known as homologous or orthologous. Paralogous sequences are present in same organism as multiple copies (e.g. isozymes) and also evolve from same parents; thus homologous in nature. Conceptually protein sequences that are evolved over billions of years from a common parent keep one thing constant over the period of its evolution is the functionality for their existence. In other words orthologous sequences evolve under functional constrain (Fay et. al, 2003). It means all these sequences (of same parent) have
to preserve similar function and also to incorporate changes in amino acid sequence (substitution) to accommodate environmental pressure over the period of its evolution for its existence.

Let takes the example of glyceraldehyde-3- phosphate dehydrogenase (GAPDH) from Yeast and Thermotoga maritime is orthologous enzyme that shows optimal activity at mesophilic (~37C) and thermophilic (~85C) conditions respectively. If the thermophilic GAPDH is operated at mesophilic condition then its activity is far less than the Yeast one (Wrba et al, 1990). Comparison of Yeast (ID P00359) and Thermotoga (ID P17721) GAPDH show 52.2% identity, 70.1% similarity and 1.5% GAP. Here GAP is the INDEL. High temperature adaptation (i.e. shift of activity optima) is related with acquired substitutions and INDEL. Thus acquisition of substitutions and insertion and deletion (INDELs) are available mechanisms in sequence evolution for its adaptation in a given environment.

Neutral or Silent or Synonymous mutations are mutations (also known as non-adaptive mutations or non-Darwinian mutations) that do not change protein sequence relative to its parent. It was proposed in 1968 by Kimura (Kimura et al, 1968) and in 1969 by King and Jukes (King et al, 1969). Following example shows polymorphism or divergence in the segment of a DNA for protein (dipeptide) (Figure 2). There are a total of 36 segments possible. The first codon is for Ser (Yellow; 6 codons) and the second is for Leu (Cyan; 6 codons). Thus a total of 36 segments is the divergence. However, all this DNA segments will produce only one protein segment i.e. ----Ser-Leu----. Now as the amino acid sequence remains the same which is the determinant for 3D structure, the function of protein will remain unaffected. The mutation in the DNA here is known as Neutral or Silent or Synonymous.

![Figure 2. Divergence in the segment of a DNA for protein (dipeptide)](image-url)
Now for a tetra peptide made up of ----Ser-Val-The-Leu---- may have a total of 6x4x4x6 = 576 silent mutations in DNA as codon for Ser, Val, Thr and Leu are 6, 4, 4 and 6 respectively. Thus for a large segment of fully functional protein neutral mutations or DNA polymorphism would be astronomically high. Vast majority of these mutations are non-function or non-Darwinian type. These mutations do not have any effect on protein structure and function. These mutations are resulted by random event (Stochastic process) and thus can be analyzed statistically.

Mutation is random changes in genetic material occurred due to replication errors or induced by radiation, mutagenic chemicals etc. Viruses, transposones also induce mutation. These mutations occur irrespective of adaptive benefit. It is acquired by random event but not due to response to any environmental stress. Majority of the evolutionary biologist accept that:

- Mutations are random
  - For example: let consider the evolution of RNA;
    - Total possibility = 3! = 3*2*1 = 6
    - RAN; RNA; ARN; ANR; NAR; NRA red indicate harmful and green is beneficial and blue is neutral
    - Now say a gene of 500 nt
      - It can have a total possibility
        - \(4^500\) copies
        - \(= 1.0715 \times 10^{301}\) copies
      - Here this random means all these copies are produced in the form of off-springs. In the biological evolution all these are produced and tested individually for their effect on biological function and adaptation (Merlin et. al, 2010). Note that there is a random component i.e. all possible mutation and a non-random component i.e. natural selection which select only advantageous (to organism functionality and adaptability) one.
Mutation occurs by chance i.e. all positions are equally probable i.e. each position has equal probability for the change as A/T/G/C.

- Mutations are independent i.e. each locus is independent and not affected by other one.
- Most mutations are harmful (biologically) and are eliminated (negative mutation) i.e. of all these genes (1.075*10^301 most of them [~70% say] are harmful). Removal is a cruel decision of evolution.
- Mutations are a disorder process
- Few mutations are beneficial or advantageous and are selected (positive mutation) i.e. of remaining 30% majority are neutral and are of no effect. Only a very low percentage is beneficial.
- Positive mutation transfer to next generation (acquired substitutions useful in adaptation)
- This drive the evolution in the forward direction

Primary sequence of protein contains information in the form of codes (such as Helix, Strand, coil, active site, binding site etc). Sequence Database (such as UNIPROT) preserves the wealth of these sequence information in highly systematic manner. Homologous sequences possesses a common parent from which they diverge into different group of organisms (Bacteria, Achaea and Eukarya) and different environmental conditions [such as high temperature (Thermophilic), High salt (Halophilic), mesophilic etc]. Homologous proteins perform common function and possess similar fold of tertiary structure. Active site and substrate specificity for homologous proteins are also similar. For example, all Hexokinase binds to hexoses irrespective of source organism and environment of adaptation. For many protein sequences the evolutionary history can be traced back 1-2 billion years (Pearson et. al, 2000).

Why study of homologous protein is highly informative? Protein sequences forming a homologous set are in fact representatives of successors of natural selection. Accepted mutations in these sequences are tested and allowed by the process of Natural Selection. Largely most of these mutations may be considered due to advantageous mutations. Conceptually each sequence may not be free of mutational defects. Because natural
selection test the fitness of randomly mutated sequences for their adaptation and functional requirements. Poorly adapted and functional set still may pass the test (because ill-functioned off-spring still may survive and pass selection).

Homologous proteins undergo substitutions at locus specific positions during the course of evolution. When we have homologous proteins in BLOCK format (where homologous positions are aligned) each homologous position is comparable independently for amino acid changes (i.e. substitution or mutations). This comparison allows us to gain insight into great deals of information on the mechanism of substitutions. Evolution of protein is the result of insertion or deletion and substitution mechanisms (Figure 3). Insertion is addition of one or more residues at certain region of sequence relative to parental one. Deletion, in turn is the removal of one or more residues. As insertion and deletion are lacking information (residues) in parental and off-spring sequence respectively, they are not directly comparable. Only substitutions that occur at locus specific positions are directly comparable.

Amino acid substitution indicates change of original (parental) residue by another one. Substitution occurs at locus specific positions. In principle a given residue can be mutated by one of nineteen amino acid residue. Are all really equally probable? In practice we see only certain substitutions are preferred at high frequency based on the original residue and other are either occurs at very low frequency or not at all. Amino acid substitution also may have certain site specificity. For example if the site is active site of an enzyme (crucial for enzymes activity) then substitutions should be different than other site.

**Substitution Matrices**

To understand the amino acid substitution, primary goal should be the sequence alignment
for the construction of protein BLOCKs. The alignment of sequence is a process of arranging the sequences of protein to identify the segments of similarity that may be a consequence of structural, functional or evolutionary mechanism (Mount DM, 2004). Pairwise sequence alignment involves comparison of two sequences; alignment of more than two sequences i.e. multiple sequence alignment is used to gain insight into the functional, structural and/or evolutionary relationships between the sequences studied using scoring function (i.e. substitution matrices).

Substitution matrices are two dimensional matrices with score values describing the probability of substitution of one amino acid with another during sequence evolution. Since DNA sequences containing non-coding regions that should be avoided during similarity/homology searches, it is better to align at protein level, if ORF (open reading frame) exists. Moreover, 20x20 amino acid exchange table more accurately expressed the evolutionary relationship. To date, several different scoring schemes have been proposed, the substitution data matrices of Dayhoff and Eck (1968) and the substitution matrices of Henikoff & Henikoff (1992) are not only popular choice for the evolution of residue similarity, but also the basis of phylogenetic relationship and amino acid classification (Murphy et al., 2001). Many efforts have been made to optimize the above matrices by applying the iterative approach (Gonnet et al, 1992), and also including evolutionary information (Koshi et. al, 1995). Usually homologous sequences are aligned column by column for generating these scoring schemes but scoring of matrices is complicated as it has to reflect:

- Physicochemical/biophysical properties of amino acids i.e. amino acids mostly differ from each other in size and charge (acidic and basic). Some amino acids contain aromatic side chains.
- Conservativeness/non-conservativeness of residues being substituted among homologous protein BLOCKs. Substitution of amino acids with other amino acids having same properties is less likely to affect the protein’s structure or function than replacement of amino acid with different property.
- Evolutionary divergence, scoring of substitution matrices which parameterized with sequence divergence is needed.
Conserved amino acids with similar properties can be more easily substituted as it preserves structure and or function among the proteins. On the other hand, non-conservative or disruptive substitution have less chance of selection during evolution as they can make protein non functional (Ng and Henikoff, 2001).

**PAM (Percent Accepted Mutation) matrices**

Late Margaret Dayhoff was a pioneer in the field. A wide set of substitution matrices based on a model of protein evolution were developed by her and coworkers (Dayhoff et. al, 2001). These widely used substitution matrices are frequently called Dayhoff, MDM (Mutation Data Matrix), or PAM (Percent Accepted Mutation) matrices. These matrices were derived from the global alignments of closely related sequences and matrices for greater evolutionary distances are extrapolated from those for lesser ones. She used seventy one protein families from the database to construct hypothetical phylogenetic trees and from each branch of the tree recorded the number of observed substitution in a 20x20 matrix. The sequence identity within each group was 85% or more. One PAM unit is 1% of amino acids positions that have been mutated. There are series of PAM matrices i.e. PAM10 to PAM250 which are extrapolated from PAM1 matrix. One PAM is 1 accepted mutation per 100 amino acids and PAM120 means 120 mutations per 100 amino acids so 1.2 accepted mutations per amino acid.

<table>
<thead>
<tr>
<th>PAM number</th>
<th>Observed substitution rate (%)</th>
<th>Sequence identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>99</td>
</tr>
<tr>
<td>30</td>
<td>25</td>
<td>75</td>
</tr>
<tr>
<td>80</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>110</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>200</td>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>250</td>
<td>80</td>
<td>20</td>
</tr>
</tbody>
</table>

The scores derived from PAM series of Matrices are an accurate description of the relative entropy or information content of an alignment (Altschul et. al, 1991).
PAM1 implies evolution to about 1 million years. PAM120 has the largest information content among others of the PAM matrix series and considered “best” for general sequence alignment. PAM250 is the most traditional and popular matrix and considered “best” for detecting distant sequence similarity.

**BLOSUM (BLOcks Substitution Matrices)**

BLOSUM series of matrices were devised by Steve Henikoff and Jorja Henikoff in 1992. These series of matrices were based on a much larger dataset (≥2000) BLOCKs from 500 or more group of proteins than the PAM matrices. Comparison of PAM and BLOSUM series of matrices is shown in table 3. BLOSUM series of matrices were derived from local, un-gapped alignments (BLOCKs) of distantly related sequences. All series of matrices were calculated directly by comparing observed frequencies of each pair to expected frequencies, then the Log-odds matrix, no extrapolations were used. Henikoff used clustering rather than an explicit evolutionary model, in order to account for different degrees of sequence divergence as well as sample biasness.

BLOSUM series of matrices are BLOSUM30, 35, 40, 45, 50, 55, 60, 62, 65, 70, 75, 80, 85, and 90. Based on the similarity or identity levels, number after the matrix (BLOSUM30) refers to the minimum percent identity of the blocks used to construct the matrix (all blocks have ≥30% sequence identity). Similarly BLOSUM62: sequence identity < 62%, and BLOSUM80: sequence identity < 80%. Here high number indicates closely related sequence and vice-versa. Among all BLOSUM62 is popular and most widely used.

During the construction process of BLOSUM matrix, counts of amino acid pair are obtained directly from columns in the conserved blocks. For the construction of a BLOSUMn matrix, according to the identity, the sequences in each block were first grouped into clusters of sequences that are at least n% identical. To obtain the amino acid pair counts, for every pair of clusters, amino acids pairs consisting of one amino acid from each cluster were tabulated. Amino acids pair within the same cluster was not tabulated. Since, some of the clusters are bigger than others; counts were normalized by the number
of sequences in the clusters. Ideally, this could be viewed as considering a cluster as an “average sequence”.

- Positive score implies that the frequency of amino acid substitutions found in the high confidence alignments is greater than occurred by random chance.
- Zero score implies that the frequency is equal to that expected by chance.
- Negative score implies that the frequency is less to that expected by chance.

### Table 3 Comparison of PAM and BLOSUM Matrices

<table>
<thead>
<tr>
<th></th>
<th>PAM</th>
<th>BLOSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evolutionary model</td>
<td>Explicit evolutionary model</td>
<td>None</td>
</tr>
<tr>
<td>Data</td>
<td>Full length MSAs of closely related</td>
<td>Conserved blocks in protein</td>
</tr>
<tr>
<td></td>
<td>sequences.</td>
<td></td>
</tr>
<tr>
<td>Bias correction</td>
<td>Trees</td>
<td>Clustering</td>
</tr>
<tr>
<td>Evolutionary distance</td>
<td>From Markov model of sequence</td>
<td>From clustering of sequences.</td>
</tr>
<tr>
<td></td>
<td>evolution.</td>
<td></td>
</tr>
<tr>
<td>Matrices</td>
<td>Transition and log odds scoring matrices</td>
<td>Log odds scoring matrix only i.e.</td>
</tr>
<tr>
<td></td>
<td>i.e. PAM1, PAM120, PAM250 etc.</td>
<td>BLOSUM30, BLOSUM62 etc.</td>
</tr>
<tr>
<td>Parameter n</td>
<td>Distance increases with n</td>
<td>Distance decreases with n</td>
</tr>
<tr>
<td>Biophysical properties</td>
<td>Derived indirectly from data</td>
<td>Derived indirectly from data</td>
</tr>
</tbody>
</table>
| Dataset                  | Smaller dataset (71 families)             | Larger (more than 500 groups of protein)

### Substitution and Adaptation

Changes in expression and interaction networks of proteins could lead to the adaptation of organisms (Wagner, 2012). Due to error-prone nature of the DNA replication process, proteins can mutate/substitute between generations that may have beneficial, deleterious or neutral effect on the fitness of the individual organism. The most important question of protein evolution is that how much change in amino acid sequence is required for the adaptation of protein in stress environment and wherein the sequence. In A4-LDH, the amino acid substitutions those are responsible for the conservation of its enzymatic property have been identified for several species (Holland et. al, 1997; Johns et. al, 2004; Field et. al, 2005). In a study Yamayoshi, et.al found that two critical amino acid substitutions in the PB2 protein (glutamic acid to lysine at position 627 and aspartic acid...
to asparagines at position 701) of A (H7N9) viruses are responsible for the mammalian adaptation (Yamayoshi et.al, 2014). Further, in recent studies it was identified that a single change in a crucial protein that enabled some crucifer species to acquire improved salt tolerance (Ali et.al, 2016). Similarly, a single amino acid substitution in non-structural protein was found to be Mediator of the adaptation of foot-and-mouth disease Virus in Guinea Pig (Nunez et.al, 2007). Amino acid substitutions also found to be involved in the adaptation of a highly pathogenic novel H5N2 avian influenza virus in mice. Amino acid substitutions were found in the MA-DK19 PB2 (E627K), PB1 (I181T), HA (A150S), NS1 (seven amino acid extension “WRNKVD” at the C-terminal), and NS2 (E69G) proteins. These substitutions are involved in the increased replication efficiency and virulence of H5N2 AIVs in mammals (Wu et.al, 2016).

Although adaptive mutation raises controversy over Darwinian evolution recently, it was anticipated much earlier than Darwinian natural selection (Berg, 1940). Adaptive mutations are not random occur in response to specific stresses (Hall, 1998). It was first proposed by Cairns in 1988 in his classical paper where he claimed that adaptive mutations are advantageous and occurs in effect to overcome environmental challenge. It is adaptive as such mutation allows selectively advantageous or non-neutral type effect on function. Adaptive mutations are distinguished from random ones by its specificity characteristics. Here specificity means it is directed to that gene which suffers selective pressure and not directed to other genes which are not under such pressure. Directed mutations occur at certain gene or at certain area of certain gene i.e. mutation rate vary along a gene/genome (there are hot-spots which are more mutation prone) (Hall; 1998). Again such directed mutations can occur in non-dividing cells. The mutations that occur are more beneficial and specific to the given stress, instead of random and not a response to anything in particular. Here stress refers to any change in environment, such as temperature, nutrients, population size. It was shown that more of the mutations observed after a given stress was effective at dealing with the stress than chance alone would suggest is possible (Spector and Foster, 1993; Sniegowski et. al, 1995). This theory of adaptive mutation was first brought to academic attention in the 1980s by John Cairns (Cairns J, 1998). Adaptive mutations are seen to occur both in prokaryotic, Eukaryotic
microorganism in different loci of the target gene and the type of mutations can be various different types (non-specific to a given residue). It is remarkable that adaptive mutations were foreseen by the father of molecular evolution in his saying “I predict that a mechanism for just this will be found” (Fitch, 1982).

**Substitution and diseases**

Changes of amino acid in locus specific position can affect the protein dramatically. It could effect on Stability, hydrogen bonding, interaction, conformation, activity, cellular localization and other physico-chemical properties of the protein (Stefl et al, 2013). It can also influence drug resistance and susceptibility to diseases resulting into serious pathological condition. The genetic variations may occur due to change in several nucleotides or only one, the latter is known as “Single Nucleotide Polymorphism” i.e. SNP. SNPs occur approximately every 1200 bases (Mooney, 2005; Capriotti et. al, 2006; Collins et. al, 1997) and are the most common genetic variations in human population (Capriotti et. al, 2006; Sherry et al, 2001). Generally it is found that SNPs occur in the non-coding region (Taillon-Miller et al, 1998) but if present in coding region may result in the change in sequences either through substitutions (nsSNPs) or by nonsense mutations (Mooney, 2005).

Understanding the effect of Mutation is utmost needed when one considers genetic diseases. It is well known that how the substitution of one residue in a protein lead to several diseases such as Sickle cell anaemia, Alzheimer’s, Parkinson’s and Creutzfeldt-Jakob diseases (Capriotti et. al, 2006). Similarly, autosomal mutations can lead to cancers and germline mutations lead to hereditary diseases (Table 4).

**Table 4 List of disease-related mutations**

<table>
<thead>
<tr>
<th>Gene/Protein</th>
<th>Disease</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP-binding cassette transporter (ABCA1)</td>
<td>Tangier disease (reduction of HDL cholesterol in plasma)</td>
<td>N1800H, W590S</td>
</tr>
<tr>
<td>Niemann–Pick C1-Like protein 1 (NPC1L1)</td>
<td>Hypercholesterolemia (reduction of LDL cholesterol in the plasma)</td>
<td>T61M, S881L, G402S, R1268H</td>
</tr>
<tr>
<td>Thiazide-sensitive sodium-chloride cotransporter (SLC12A3)</td>
<td>Gitelman syndrome (salt wasting and low blood pressure)</td>
<td>G439S, G741R</td>
</tr>
</tbody>
</table>
Deciphering the mechanisms of the effects of these mutations (Satoh et al, 2000) is made possible by the analysis of Missense mutations that help us to understand the relationships between protein structure and function.

References

Berg DP. Classification of fishes, both recent and fossil. Trav Inst Zool Acad Sci USSR. 1940;5:87-517.


Fields PA, Dong Y, Meng X, Somero GN. Adaptations of protein structure and function to temperature: there is more than one way to ‘skin a cat’. Journal of Experimental Biology. 2015 Jun 1;218(12):1801-11.

Fitch WM. The challenges to Darwinism since the last centennial and the impact of molecular studies. Evolution. 1982 Nov 1;36(6):1133-43.


Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome research. 2001 May 1;11(5):863-74.


Spector MP, Foster JW. Starvation-stress response (SSR) of Salmonella typhimurium. In Starvation in Bacteria 1993 (pp. 201-224). Springer US.


Development of Computational Tools

Motivation

Advancement of genome sequencing technology and the improvement of X-ray crystallography and NMR spectroscopy methods is enriching the sequence and structural databases of proteins very fast. However, for intricate balance between the entry of sequences and structures in the database and to understand the biological function of these proteins, needs efficient and batch analyzing software. For understanding the mechanism of substitution (theme of my thesis work), mass scale analysis of both the sequences and structures are required. To fulfill the requirement, I needed various sequence and structural analyzing efficient computational tools. I have found many authentic GPL software and servers at the internet during the investigation of my work and realized that they were not sufficient. So, we have not only decided to develop the necessary tools required for the work but also made it freely available to academic users worldwide.

We have classified our developed tools into two categories Table 1:

Table 1 name of the tools, its availability and suitable citations for sequence and structural analysis are presented.

<table>
<thead>
<tr>
<th>SI. No.</th>
<th>Name of the tool</th>
<th>Availability</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PHYSICO</td>
<td>Available through Email</td>
<td>Gupta et. al, 2014</td>
</tr>
<tr>
<td>3</td>
<td>APBEST</td>
<td><a href="https://sourceforge.net/projects/apbest/">https://sourceforge.net/projects/apbest/</a></td>
<td>Gupta et. al, 2017</td>
</tr>
<tr>
<td>4</td>
<td>SBION</td>
<td>Available through Email</td>
<td>Gupta et. al, 2014</td>
</tr>
<tr>
<td>6</td>
<td>ADSBET</td>
<td><a href="https://sourceforge.net/projects/adsbet/">https://sourceforge.net/projects/adsbet/</a></td>
<td>Nayek et. al, 2015</td>
</tr>
<tr>
<td>7</td>
<td>ADSBET2</td>
<td><a href="https://sourceforge.net/projects/adsbet2/">https://sourceforge.net/projects/adsbet2/</a></td>
<td>Nayek et. al, 2015</td>
</tr>
</tbody>
</table>

1. Development of tools for sequence analyses

1.1 PHYSICO and PHYSICO2: An UNIX based Standalone Procedure for Computation of Physicochemical, Window-dependent and Substitution
Based Evolutionary Properties of Protein Sequences along with Automated Block Preparation Tools

1.1.1 Background

Amino acids, building blocks and determinant of ensemble characteristics of proteins, largely mediate physicochemical properties of proteins by their side chains. A stretch of amino acids in protein sequence might constitute a code and that may eventually direct proteins to achieve structure and function. In orthologous set, candidate sequences diverge from a parent and thus their studies on the position-related substitutions might reveal their true evolutionary characteristics. In this context voluminous sequence databases act as study materials for extraction and application of above information. At present there are sequences from of about 6000 genomes, and more are coming over time. To undergo systematic extraction of sequence information (both physicochemical and code-related), compare mean properties among intra and inter group of orthologous protein families and to extract evolutionary characteristics, potential computational tools would be necessary.

At present public domain tools mostly perform per-sequence analysis (e.g. physicochemical properties) that not only slow the process but also cause incorporation of manual management error. For example, we performed PHYSICO2 (Banerjee et. al, 2015) based analyses on representative candidate sequences from two taxonomic groups (metazoan: 31 and cyanobacteria: 32 sequences) of “cytochrome c family” in Intel(R) core™ i3 CPU M330 @2.13 GHz PC-CYGWIN (32 bit) environment. We also performed same analysis using “PROTPARAM” for physicochemical (8 properties [6] and their averages) and “PROTScale” (individual and average profile) for window-dependent [6] using “Alliance Broadband: PRIME (54Mbps) package” internet connection. Efficient use of these tools took ≥ 10 hours for obtaining these results in excel. On the other hand, only 6 minutes was sufficient to obtain the above and other additional properties in PWS-format.
Therefore, we have developed PHYSICO (Gupta et. al, 2014) and PHYSICO2, (a higher version of PHYSICO) that performs rapid analysis of sequences at proteomic scale. The operational principle including user flexibility in choosing hydrophobic (HP), hydrophilic (HL) and polar uncharged (PU)-residue classes, iso-electric point (pI)-methods and Shannon threshold, input sequence format (such as RAW or BLOCK FASTA), itemized output with property information are also been reported. The outputs include analysis of composition and class composition (both row and column wise), variability, positional substitutions, and positional distribution of maximum diverse residues, pI and titration profile, 52 GRAND AVERAGE properties and 46 window dependent properties. The excel format of output is user friendly and thus plotting, comparing among results are easily achieved.

A novel procedure i.e. Automated Block Preparation Tools (i.e. ABPT: Details in documentation of PHYSICO2) that helps in preparation of sequence-BLOCK from UNIPROT-FASTA-sequences are also presented. An extensive documentation along with demonstration that covers almost all methods and capabilities of the program is also provided as link for clarity and understanding. Overall, PHYSICO2 performs fast and accurate analysis of any forms of FASTA file for relevant clusters of properties which has potential in the field of sequence bioinformatics.

1.1.2 Feature and Implementation of the program

1.1.2.1 Features of the program

An automated standalone, multipurpose procedure for analysis of protein sequences in FASTA format. The program works from UNIX like environment including CYGWIN and is interpreted by AWK programming language. It works on various forms of FASTA files (Input). Based on the form of FASTA-input, relevant outputs are produced. The program is designed such that large database of sequences can be analyzed in a single run. Overall it provides countable advantages over the existing web-tools.
In a nutshell:
1) It works on any form of FASTA format file (RAW or BLOCK, multiline or single lined) with any number of sequences in it.
2) Input parameters can either be set as default or from user choice.
3) Program also includes AUTOMATED BLOCK PREPARATION TOOL (ABPT) - an extension that allows users to prepare sequence BLOCK FASTA file which is otherwise cumbersome to prepare by manual effort.
4) Multi items outputs on:
   • Composition and class-composition for sequences and for positions (for BLOCK FASTA only)
   • Positional variability by Shannon, Simpson and Wu Kabat methods.
   • Positional substitution (contains information on invariant line, conservative and non conservative substitutions) along with mentioning of conserved positions (variability ≤ Shannon threshold)
   • Positional distribution of first two most diverse (in BLOCK) and highest conserved residues.
   • pI and pI profile using six different pI scales.
   • Extinction co-efficient, Aliphatic Index, GRAVY, Molecular weight and mean molecular weight for all sequences present in FASTA file. Results are arranged in systematic manner in excel column.

1.1.2.1.1 Physico-chemical Properties

The program PHYSICO2 includes 52 GRAND AVERAGE physicochemical properties based on the scale devised either by experimental or empirical means. AAindex database (Kawashima, et. al 2008) has enlisted many (~1000) important physicochemical scales of amino acids and inter links among each other. We have included the database IDs for these 52 properties for the purpose of their easy access by users, comprehensive understanding of these properties and
improvement of the present effort by suggestions. Amino acids being building
block of proteins, these quantitative scales are very useful in

- Differentiating among protein sequences
- Global and binary comparison among sequences
- Locating site for protein-protein interaction
- Determination of mean secondary structure dominance
- Determination of mean membrane property
- Determination of mean antigenicity
- Development of substitution matrices
- Proteomic studies and prediction of mobility in 2D PAGE

Most of these index values are also included for determination of window dependent
profile that helps to identify sequence segments with

- hydrophobicity and trans membrane helix or strand
- Hydrophilicity and antigenicity
- Prediction of secondary structures (α helix, β sheet, turns)
- Proteomic profile comparison and identification of novel site of interactions

**Grand Average Property** (Table 2) is calculated using the below
formulæ:

\[
grand\ property = \frac{\sum_{i=1}^{20} \text{count of amino acid}_i \ast \text{amino \textendash} \text{acid index value}_i}{\text{total amino acid in sequence}}
\]

A total of 52 properties (Table 2) will be computed for each sequence
and that will appear in column wise manner in excel file. If relevant
their mean and SD can also be determined.

<table>
<thead>
<tr>
<th>SE</th>
<th>ID</th>
<th>Property</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>FASG760101</td>
<td>molecular weight</td>
<td>Fasman 1976, Handbook of Biochemistry and Molecular Biology, 3rd ed, Proteins</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mean molecular weight</td>
<td>~ Vol. 1, CRC Press, Cleveland Biochemistry Text Book</td>
</tr>
<tr>
<td>3)</td>
<td>--</td>
<td>Aliphatic Index</td>
<td>Ikai A, J Biochem, 88, 1895, 1980</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>--</td>
<td>Extinction coefficient when SH of Cys are in free form</td>
<td>Pace, et al, Protein Sci, 11, 2411, 1995</td>
</tr>
<tr>
<td>5</td>
<td>--</td>
<td>Extinction coefficient when SH are in -S-S- form</td>
<td>Pace, et al, Protein Sci, 11, 2411, 1995</td>
</tr>
<tr>
<td>7</td>
<td>HOPT810101</td>
<td>Hopp and Woods Hydrophilicity</td>
<td>PNAS-USA 78, 3824, 1981</td>
</tr>
<tr>
<td>8</td>
<td>KYTJ820101</td>
<td>Kyte and Doolittle Hydrophobicity</td>
<td>JMB 157, 105, 1982</td>
</tr>
<tr>
<td>10</td>
<td>EISD840101</td>
<td>Eisenberg Hydrophobicity</td>
<td>PNAS-USA 81, 140, 1984</td>
</tr>
<tr>
<td>11</td>
<td>EISD860101</td>
<td>Eisenberg-McLachlan Solvation free-energy</td>
<td>Nature 319, 199, 1986</td>
</tr>
<tr>
<td>12</td>
<td>PRAM900101</td>
<td>Prabhakaran Hydrophobicity</td>
<td>Biochem J 269, 691, 1990</td>
</tr>
<tr>
<td>13</td>
<td>WIMW960101</td>
<td>Wimley-White Experimental Hydrophilicity</td>
<td>Nat Struct Biol 3, 842 1996</td>
</tr>
<tr>
<td>15</td>
<td>vonHeijnes</td>
<td>von Heijne G Hydrophobicity</td>
<td>JMB 225, 487 1992</td>
</tr>
<tr>
<td>16</td>
<td>GRAR740102</td>
<td>Grantham Polarity</td>
<td>Science 185, 862, 1974</td>
</tr>
<tr>
<td>17</td>
<td>GRAR740101</td>
<td>Grantham Composition</td>
<td>Science 185, 862, 1974</td>
</tr>
<tr>
<td>18</td>
<td>ZIMJ680102</td>
<td>Zimmermann Bulkiness</td>
<td>J Theor Biol 21, 170, 1968</td>
</tr>
<tr>
<td>20</td>
<td>ZIMJ680104</td>
<td>Zimmermann Isoelectric point</td>
<td>J Theor Biol 21, 170, 1968</td>
</tr>
<tr>
<td>21</td>
<td>CHOC750101</td>
<td>Chothia av volume of buried residue</td>
<td>Nature 254, 304, 1975</td>
</tr>
<tr>
<td>22</td>
<td>CHOC760101</td>
<td>Chothia residue accessible surface area in tripeptide</td>
<td>JMB 105, 1, 1976</td>
</tr>
<tr>
<td>23</td>
<td>CHOC760102</td>
<td>Chothia residue accessible surface area in folded protein</td>
<td>JMB 105, 1, 1976</td>
</tr>
<tr>
<td>24</td>
<td>CHOC760103</td>
<td>Chothia Proportion of residues 95% buried</td>
<td>JMB 105, 1, 1976</td>
</tr>
<tr>
<td>25</td>
<td>CHOC760104</td>
<td>Chothia Proportion of residues 100% buried</td>
<td>JMB 105, 1, 1976</td>
</tr>
<tr>
<td>29</td>
<td>LEVM780102</td>
<td>Levitt Normalized frequency of beta-sheet</td>
<td>Biochemistry 17, 4277, 1978</td>
</tr>
<tr>
<td>30</td>
<td>LEVM780101</td>
<td>Levitt Normalized frequency of alfa-helix</td>
<td>Biochemistry 17, 4277, 1978</td>
</tr>
<tr>
<td>31</td>
<td>LEVM780103</td>
<td>Levitt Normalized frequency of reverse-Turn</td>
<td>Biochemistry 17, 4277, 1978</td>
</tr>
<tr>
<td>32</td>
<td>LEVM760102</td>
<td>Levitt Dist between C-alpha and centroid of side chain</td>
<td>JMB 104, 59, 1976</td>
</tr>
<tr>
<td>33</td>
<td>LIFS790103</td>
<td>Lifson and Sander Anti-parallel-Beta</td>
<td>Nature 282, 109, 1979</td>
</tr>
<tr>
<td>34</td>
<td>LIFS790102</td>
<td>Lifson and sander Parallel-Beta</td>
<td>Nature 282, 109, 1979</td>
</tr>
</tbody>
</table>
Some values are expressed as percent (to follow similar convention as others). Most of these properties contain ID value. This ID when entered in the web page of AAindex: [http://www.genome.jp/aaindex/](http://www.genome.jp/aaindex/), further details can be procured.

- Users can redefine hydrophobic (hence hydrophilic) residues and polar uncharged residue-classes (Table 3) over default settings. Default HP and PU are displayed during run of the program. If user wishes to change default values, residues can be entered in capital, single letter code in continuous form i.e. GLVAPC. It is worth noting that there is no strict boundary between these classes of amino acids. Therefore, it is sometime important to change the default definition for HP, HL and PU in context dependent manner. However, other classes like acidic, basic etc are fully defined and thus are not opted for change.
Table 3 Different residues classes as obtained from the run of the program. As BLOCK FASTA file is iso-width, these classes are computed both for positions and sequences. It would be extremely important to compute mean along with standard deviation from the excel output and compare them among different families of protein.

<table>
<thead>
<tr>
<th>Residue Class</th>
<th>remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP</td>
<td>HYDROPHEOBIC residues, by default they are VAFILMCPG</td>
</tr>
<tr>
<td>HL</td>
<td>HYDROPHILIC residues, by default they are STNQDERKHWY</td>
</tr>
<tr>
<td>PU</td>
<td>POLAR-UNCHARGED residues, by default they are STNQYW</td>
</tr>
<tr>
<td>DE</td>
<td>Total Negative charge at neutral pH</td>
</tr>
<tr>
<td>KR</td>
<td>Total Positive Charge at Neutral pH</td>
</tr>
<tr>
<td>TOT</td>
<td>Total Charge at neutral pH i.e. D+E+K+R</td>
</tr>
<tr>
<td>NET</td>
<td>Net Charge at neutral pH i.e. D+K-R</td>
</tr>
<tr>
<td>FYW</td>
<td>Total aromatic residue i.e. F+Y+W</td>
</tr>
<tr>
<td>NQ</td>
<td>Total amide residues i.e. N+Q</td>
</tr>
<tr>
<td>ST</td>
<td>Total alcoholic residues i.e. S+T</td>
</tr>
<tr>
<td>PG</td>
<td>Total Flexibility and bend promoting residues i.e. P+G</td>
</tr>
<tr>
<td>CW</td>
<td>Total rare residues i.e. C+W</td>
</tr>
<tr>
<td>MC</td>
<td>Total thiol containing amino acids i.e. M+C</td>
</tr>
<tr>
<td>DIS</td>
<td>Total strong disorder forming residues i.e. S+E+P+R</td>
</tr>
<tr>
<td>ORR</td>
<td>Total strong order forming residues i.e. C+F+Y+W</td>
</tr>
<tr>
<td>BB</td>
<td>Total basic residues i.e. H+R+K</td>
</tr>
</tbody>
</table>

These classes are computed both for i. per-sequence (row wise) and ii. **Position** (column wise) analyses. In this BLOCK FASTA file both are relevant.

### 1.1.2.2 Implementation of the Program

The program works on any form of input FASTA (RAW or BLOCK) file (Figure1: F1 and F2). Upon its execution, Users has option (residue classes, pI-method and Shannon-threshold) which allows them to change the default input-parameters (DPAR) by users one (UPAR) and then it enters into first phase (Figure1: P1) of analyses. In contrast to BLOCK-FASTA (F2), input having RAW-FASTA (F1) generates only one output file (Figure 1: R1) as non-comparable homologous positions i.e. column specific analyses (that generate additional three outputs: B2, B3 and B4 in the former one) are skipped. Extracted sequences from one or more taxonomic groups i.e. RAW-FASTA-file (F1) is converted into BLOCK-FASTA-file/ files of identical width using Automated Block Preparation Tool i.e. ABPT (F3) of the program.
The program performs window-dependent property analysis in the second phase of computation (P2). In case of BLOCK-FASTA (where homologous positions are aligned) as an input, output of all sequence specific profiles are redirected into an excel file (R5) for easy computation of mean along with standard deviation (Documentation) otherwise output of each sequence specific profile (R2₁, R2₂ etc) is saved in named directory (Documentation) separately
Instruction for the execution of the program:

1. Download PHYSICO2 ("PHYSICO2.zip") from its home page:
   http://sourceforge.net

2. Unzip the content - README (this file)

3. 'cd' to "WORK" directory.

3. To run PHYSICO2 with FASTA file (figure 2):
   Type >./PHYSICO2.exe FASTA-file-name [enter and follow steps thereafter]
   To run abpt (automated Block Preparation Tool)
   Type >./block/abpt.exe [enter and follow step thereafter]

Figure 2: figure shows the execution of the program from command line, after analyzing, the program asks user whether want to calculate Window dependent property.
RAW-FASTA file can be subjected for sliding window dependent sequence property analysis (Figure 3). There are a total of 46 properties. One can compute one property at a time. The property will be computed for each sequence present in the FASTA file. Upon selection of a property to be analyzed (e.g. Kyte and Doolittle hydrophobicity present at second position of the list) an appropriate WINDOW SIZE is to be provided. The WINDOW SIZE is always an odd number (e.g. 7, 9, 11 or 21 etc). However the exact value depends on the context and property to be analyzed.

Figure 3: Execution of the program having option to put the user desired property calculation. Different window dependent properties and options used for computation. Only one of 46 properties could be calculated per run.

- The program is downloaded by the worldwide users, dominant user are from India.
1.2 APBEST: Analysis of Protein Block and Evaluation of Substitution Type

1.2.1 Background

Day by day, the size of sequence database has becomes very large and with improved technology molecular sequence data are becoming increasingly easy to study. Determination of physico-chemical properties (Gupta et al. 2014; Gasteiger et al. 1999 etc.), construction of phylogenetic tree is now a routine approach for accurate and fast establishment of cladistic relationship among candidate sequences. The approach is well documented in the literature with few conceptual models and the same have been implemented in a number of famous automated user-friendly software tools (Kumar et al. 2001; Kosakovsky et al. 2005; Felsenstein 1981; Huson and Bryant 2006; Jobb et al. 2004). Although these pioneer studies have widened our understanding for the field, the governing principles of protein evolution in relation to observed hetero-pairs usage patterns still an enigma.

Mechanism of substitution i.e. acquisition of different amino acids at locus specific positions, play vital role in this process. In other words substitutions cause formation
of hetero-pairs at the cost of homo-pairs. As far as type and frequency of substitution is concerned, these could either be conservative or non-conservative, dominant or auxiliary. In these aspects extracting information regarding residue and their class specific diversity is challenging and the knowledge of it could further be important for understanding evolutionary mechanisms. At this end, our development APBEST analyze BLOCK FASTA file for extraction of quantitative information on substitutions and diversity to redirect itemized (6 items) output in results on types, frequencies, probability distribution and usage pattern of different substitution-hetero-pairs. Moreover, Implication and formulation of parameter that relate divergence rate is also been implemented. Overall the program extracts evolutionary parameter that are non-overlapping with the conventional phylogenetic ones, the knowledge of which has potential in understanding molecular evolution in relation to substitution mechanism.

1.2.2 Features and implementation of the program

APBEST, interpreted by AWK and C++ programming language, is an efficient and user friendly program. It works either from Unix/Linux/cygwin or Window command prompt modes. The program takes BLOCK-FASTA file as input and performs detailed statistical analyses in homologous position specific manner to redirect output in excel format. It is freely available at http://sourceforge.net/projects/APBEST/ for academic users.

APBEST operational flow chart is presented in Figure 5. Upon start of the program it checks for the FASTA format of the input. If it is not in BLOCK format, the program loops out. However, if the input is in BLOCK format, the program needs user confirmation for redirection of output for distribution of hetero-pairs substitution. An output of named HPSD.xls is redirected on affirmation. However, for NO it looks for input of variability threshold for final output.
1.2.2.1 Analyses of BLOCK FASTA File and extraction of evolutionary parameters

Analysis of BLOCK FASTA files were performed using in house procedure APBEST. \(D, R, E, MDR, RD\) and \(CD\) were computed using relevant observed frequency of SHP (see below). For a given BLOCK there are 190 hetero-pairs and 20 homo-pairs (Henikoff and Henikoff, 1992). The program computes SHP and substitution-homo-pair (SMP) frequencies first. Observed SHP frequency for a given pair (e.g. \(XY\)) is computed for each column position \((m)\) of BLOCK and then summed over all positions \((w)\) to obtain BLOCK \((B)\) frequency.

\[
f_{XY}^B = \sum_{m=1}^{w} f_{XY}^m = \sum_{m=1}^{w} \left[ n_x C_1 * n_y C_1 \right]_m = \sum_{m=1}^{w} \left[ n_x * n_y \right]_m \tag{1a}
\]
\[ f_{XY}^{RPF} = \frac{f_{XY}^B \times 100}{f_{SHP}^B} \quad [1b] \]

Where \( n_X \) and \( n_Y \) are frequencies of amino acid \( X \) and \( Y \) respectively for a given column position (\( m \)) of block \( B \) of width \( w \). Block specific observed frequency (\( f_{XY}^B \)) for a given pair (\( XY \)) is the sum of frequencies of all column positions. \( f_{XY}^{RPF} \) is the relative percentile frequency or probability for the hetero-pairs \( XY \).

Similarly the observed frequency of SMP for a given pair (say \( ZZ \)) is computed as

\[ f_{ZZ}^B = \sum_{m=1}^{w} f_{ZZ}^m = \sum_{m=1}^{w} \left[ p_z C_2^m \right] = \sum_{m=1}^{w} \left[ \frac{n_Z(n_Z-1)}{2} \right] \quad [2a] \]

\[ f_{ZZ}^{RPF} = \frac{f_{ZZ}^B \times 100}{f_{SMP}^B} \quad [2b] \]

Where \( n_Z \) is the count of amino acid \( Z \) for a given column position (\( m \)).

SHP parameters such as \( D, R \) and \( E \) are computed based on its observed frequency (Equation 1a). \( D \) indicates SHP whose frequency is highest. It is automatically extracted using the program. All SHP (190 types) are divided into two categories namely conservative and non-conservative for computation of \( R \). Conservative does not alter overall protein’s properties (such as \( D \Leftrightarrow E \)) and non-conservative part is the one that does change protein physico-chemical properties (such as \( D \Leftrightarrow K \)).

\[ R = \frac{f_{HB-HL}^B \times 100}{(f_{HB-HB}^B + f_{HL-HL}^B)} \quad [3] \]

Thus, \( R \) is defined as non-conservative to conservative substitution ratio, \( f_{HB-HL}^B \):

Hydrophobic to hydrophilic, \( f_{HB-HB}^B \); hydrophobic to hydrophobic and \( f_{HL-HL}^B \); hydrophilic to hydrophilic SHP frequencies. \( E \) (the usage parameter for hetero-pair) is the normalized frequency of SHP for a given BLOCK which is computed as:

\[ E = \frac{f_{SHP}^B \times 100}{f_{SHP}^B + f_{SMP}^B} \quad [4] \]
\( f_{\text{SHP}}^B \): Substitution-hetero-pair frequency and \( f_{\text{SMP}}^B \): Substitution-homo-pair frequency. Superscript \( B \) indicate sum of all SHPs frequency for a given BLOCK. \( E \) is 0 when hetero-pair frequency is absent and no evolution of sequence BLOCK i.e. \( f_{\text{SHP}}^B \) is 0. On the other hand, if \( f_{\text{SMP}}^B \) is 0 then \( E \) is 1 which means maximum evolution of sequence BLOCK.

Apart from above frequency dependent parameters (\( D, R, E \)), APBEST also determine diversity parameters (such as \( \text{MDR}, \text{RD} \) and \( \text{CD} \)). \( \text{RD} \) indicates sum of frequencies of all nineteen hetero-pairs that associate residue \( X \). Thus,

\[
\text{RD}_X = \sum_{k=1}^{19} f_{Xk}
\]  

[5]

Here \( k \) is any of 20 amino acids except \( X \). Maximum and minimum diversities (\( \text{RD}_{\text{MAX}} \) and \( \text{RD}_{\text{MIN}} \) respectively) are also determined. \( \text{MDR} \) refer maximally diverse residue i.e. \( \text{RD}_{\text{MAX}} \).

\( \text{CDs} \) (classes or \( \text{CL}: \) acidic, basic, non-polar, hydrophobic and hydrophilic) are calculated as:

\[
\text{CD}_{\text{CL}} = \sum_{j=1}^{n_{\text{CL}}} \left( \sum_{k=1}^{r} f_{j,k} \right) - \left( \sum_{j=1}^{n_{\text{CL}}} \left( \sum_{t=j+1}^{n_{\text{CL}}} f_{j,t} \right) \right)
\]  

[6]

The above equation contains two terms. The first term is for computation of the sum of diversities of candidate residues belong to the class and the second term is the sum of frequencies of overlapping frequency of SHP formed by candidate residues of the class. Total number of overlapping hetero-pairs is obtained by \( n_{\text{CL}}C_2 \) where \( n_{\text{CL}} \) is the number of class residues. \( \text{CL} \) is the class and \( n_{\text{CL}} \) is the total count of class residues (for basic residues HRK, \( \text{CL} = \text{basic class}, \ n_{\text{CL}} = 3 \)). For computation of diversity for each residue of a class, that residue is excluded from the list of 20 amino acids (e.g. ACDEFGHIKLMNPQRSTVWY). Thus, \( r \) is always constituted
by 19 residues except the one for which SHP diversity is considered to avoid inclusion of its homo-pair. For example, basic class diversity:

\[ DV_{HRK} = (f_{\text{hetero-pair of } H} + f_{\text{hetero-pair of } R} + f_{\text{hetero-pair of } K}) - (f_{HR} + f_{HK} + f_{RK}) \]

\[ f_{\text{hetero-pair of } H} = f_{HA} + f_{HC} + f_{HD} + f_{HE} + f_{HF} + f_{HG} + f_{HI} + f_{HK} + f_{HL} + f_{HM} + f_{HN} + f_{HR} + f_{HS} + f_{HT} + f_{HV} + f_{HW} + f_{HY} \]

\[ j=1; k=1 \text{ to } 19, \text{ that exclude } H \text{ from } ACDEFGHIKLMNPQRSTVWY \]

\[ f_{\text{hetero-pair of } R} = f_{RA} + f_{RC} + f_{RD} + f_{RE} + f_{RF} + f_{RG} + f_{RK} + f_{RL} + f_{RM} + f_{RN} + f_{RP} + f_{RQ} + f_{RS} + f_{RT} + f_{RV} + f_{RW} + f_{RY} \]

\[ j=2; k=1 \text{ to } 19, \text{ that exclude } R \text{ from } ACDEFGHIKLMNPQRSTVWY \]

\[ f_{\text{hetero-pair of } K} = f_{KA} + f_{KC} + f_{KD} + f_{KE} + f_{KF} + f_{KG} + f_{KH} + f_{KL} + f_{KM} + f_{KN} + f_{KP} + f_{KQ} + f_{KS} + f_{KT} + f_{KV} + f_{KW} + f_{KY} \]

\[ j=3; k=1 \text{ to } 19, \text{ that exclude } H \text{ from } ACDEFGHIKLMNPQRSTVWY \]

Residues constituting a class can be written in any given order (e.g. HRK or RHK or KHR etc for basic class) to assign the value of \( j \). Each will return identical value of diversity. The second term always contains \( n^{CL}C_2 \) hetero-pairs whose frequencies are summed twice (\( HR, HK \text{ and } RK \)) in the net sum of first term. Therefore sum of the frequency of these terms is deducted once from the first term.

Apart from the above computations, APBEST also extracts other position specific BLOCK parameters. Shannon entropy (Shannon, 1948) is computed using the following formula:

\[ H = -\sum_{i=1}^{M} P_i log_2 P_i \]  \[ \text{[7]} \]

Where \( P_i \) is the fraction of residues of amino acid types \( i \), and \( M \) is the number of amino acid types (20 in number). \( H \) ranges from 0 (only one residue in present at that position) to 4.322 (all 20 residues are equally represented in that position). Typically, positions with \( H \geq 2.0 \) are considered variable, whereas those with \( H \leq 2 \) are consider conserved. Highly conserved positions are those with \( H \leq 1.0 \).

BLOCK positions undergo different types of substitutions. Different positions of BLOCK are also assessed based on residue types. If there is only one type of amino
acid in a given position then it is marked as **invariant**. If substituted then qualitatively positional substitutions are assessed as different categories such as hydrophobic-hydrophobic, hydrophilic-hydrophilic and hydrophobic to hydrophilic etc.

APBEST keep all these results in itemized form in excel output ([https://sourceforge.net/projects/apbest/files/](https://sourceforge.net/projects/apbest/files/)) for analyses.

➢ **Instruction for the execution of the program:**

A] Run from window command prompt—

1. Window should have gawk/Awk binary; can be downloaded from following link or elsewhere [http://gnuwin32.sourceforge.net/packages/gawk.htm](http://gnuwin32.sourceforge.net/packages/gawk.htm)

2. Unzip in a proper location eg C:\AWK

3. Include the bin directory in window-path variable: (>==click; >>right click) *->start- >>computer->properties->Advance system settings->environment variables->user variable-<new> [Note: if path not exist then type <new>; if it does just click edit highlighting path]

4. Variable name-- path variable value-- C:\AWK\bin

5. Click ok

6. Open a command prompt

7. Type Awk it will print many things if it does then it is done else repeat the above steps (1 to 7) correctly.

8. Now in a command prompt type APBEST block.fasta_input_file (provided as ldh_m-bl._inline._BLOCK.fasta)

9. Locate output (name ldh_m-bl._inline._BLOCK.xls in the present directory and also provided in the OUTPUT folder)

B] For execution of APBEST in UNIX/LINUX SHELL:
1. Put the APBEST in a directory in your drive, do chmod (executable) and chown (ownership) if needed.

2. In that directory type. /APBEST block_fasta_input_file

3. As step 9 (see above) [note: alternatively the program can be kept in some system or user specific bin dir whose path to be included in .login etc file. in this case program can be called from any dir]

➢ The program is downloaded by the worldwide users (figure 6), dominant user are from India followed by Spain.

![Figure 6: Download statistics of the software APBEST from Worldwide users.](image)

2. Development of tools for structural analyses

2.1 SBION and SBION2 : Analyses of Salt Bridges and its Binary Details from Multiple Structure Files

2.1.1 Background
Salt Bridges are specific electrostatic interactions that forms between oppositely charged side chain atoms of acidic (acceptor atoms) and basic (donor atoms) residues that are within 4Å distance (Xu et. al, 1997). In protein folding, these interactions play crucial role in nucleation process and provide overall stability of the protein. Due to advancement of method such as NMR and X-ray crystallography, the growth of structural database become rapid and it has been increasing day by day. To catch this rapidity, there is an urgent need of efficient and batch analyzing software. Bioinformatics tools for determination of atomic intra- and inter-chain salt-bridges and their secondary structural distribution for single input at a time (Grusaran et. al, 2014) and for multiple structure files per run along with additional details on networked SB and missing residues information (Gupta et al., 2014) are now available. However, further development is urgently needed for details on residue specific SB (but not atomic ones as above) with binary details as it has been seen that a salt-bridge could have many binary details, such as it may be buried (in core) or exposed (in surface), networked or isolated, hydrogen-bonded or non-hydrogen bonded, present in secondary structure or in coil i.e. intra- or inter-helical/strand/coil, local or non-local (if local, i→(i+n) typing), formed by single bond or multiple bonds (bonds ≥ 2). Each of the mentioned binary details is known to modulate the energy contribution of salt-bridge [Rose et al., 1994]. Therefore, Extraction of not only salt-bridge but along with these binary details is also important. Since helix and strand are permanent dipole in protein and thus to understand the structural stability as well as packing, determination of orientation for intra type SBs (i.e. basic residue is at N-terminal [orientation-I] or acidic residue is at N terminal [Orientation-II]) would be useful (Rose et al., 1994).

For the analyses of above mentioned binary details, we have developed SBION (Gupta et al., 2014) and SBION2 (Gupta et al., 2015) (improved version of SBION), which can extracts all the binary details from multiple structures file in a single run. The output of the software includes ready to use compact supplementary table. The program SBION2 runs in three alternative modes as per user requirement. Each three mode could works on multiple numbers of structure files with any number of chains and at any given atomic distance of ion-pair. Our development, SBION2 has been the
first of its kind which not only display all the above details in systematic manner along with others (Gupta et al., 2014) but also store and redirect these data in excel format for comprehensive post run statistical analyses.

Availability

SBION2 is freely available at http://sourceforge.net/projects/sbion2/ for academic users.

2.1.2 Features and implementation of the program

- Improvement of SBION2 (Gupta et al., 2015) over SBION (Gupta et al., 2014)

| FEATURES: |
| Common a] structures | any number with .pdb extension |
| Common b] chains | all chains in structures |
| Common c] Sb distance | user choice |
| NEW d] Local threshold | user choice |
| NEW e] core threshold | user choice |
| Common f] ugly PDBs | corrected |
| Common g] NMR files | screened out |
| Common h] missing res | reported |
| NEW i] Sb distance | averaged over Sb partners interactions |
| NEW j] Binary outputs | in excel format |
| Improved k] Res. distribution | acidic (D, E) and basic (K, H, and R): |
| NEW l] intervene Res. Dist. | outputs for all, isolated and networked |
| NEW m] supplementary table | for direct use in publication purpose |
| NEW n] versatile ASA calc | either with surface racer or Naccess |
| NEW o] General output | missing res, networked Sb and |
| | Protein- Protein interaction |
| NEW p] Binary analysis: KD, KE, HD, HE, RD and Residue specific |
| Isolated vs. Networked |
| Sb in Core vs. Sb in Surface |
Methodology

A flow-chart with detailed functioning of the program SBION2 is presented in Figure 7.

Upon start of the program it screens out NMR files if they are there in the working directory followed by making a list of X-ray structure files. The program verify for accessibility calculation program (ASA) in specified directory. If absent (i.e. NO), the program obtains atomic salt bridge pairs (step-1) and then converts them in residue specific salt bridge pairs (step-2). At this level four outputs are redirected (O-1).

Figure 7: Flowchart of the functioning of the program SBION2.
Output details of isolated \((O-2)\) and networked salt-bridges \((O-3)\) are separately reported \((STEPS\ 3\ to\ 6)\). A supplementary table \((O-4)\) is prepared which include both isolated and networked salt-bridge information \((STEPS\ 5\ and\ 6)\). If ASA \((STEP\ 7)\) is \textit{YES}, SBION2 performs step 1 through 6 that include \textit{core} and \textit{surface} location information of salt bridges along with earlier ones. \textit{Freq}: Frequency; \textit{SMFS}: Single bond multiplicity frequency and sum; \textit{MMF}: Multiple bond multiplicity frequency; \textit{MMS}: Multiple bond multiplicity sum; \textit{HB\_SB}: Hydrogen bonded salt-bridge frequency; \textit{non\_HB}: non-Hydrogen bonded salt-bridge frequency; \textit{SSD\_SB}: secondary structure (helix, sheet and coil) distribution of salt-bridge; \textit{L\_SB}: local salt-bridges; \textit{non\_L}: non-local salt bridges; \textit{II\_HSCI}: intra- and inter-helix, sheet and coil interactions; \textit{int\_T}: intervening residue table.

![SBION2 Execution](image)

**Figure**: 8 SBION2 execution (A), its outputs (B) and display of few of its output details on salt-bridges (C: 1 through 12) are presented.
Execution of the program with option “h” shows three modes of its operation (A). A total of 27 kinds of outputs relevant for presentation and statistical calculation are obtained (B). The program extracted salt-bridge interactions such as **INTRA-HELIx** (C1 and C2), **COIL-HELIx** (C3 and C4), **STRAND-COIL** (C5), **INTRA-STRAND** (C6), **INTER-HELIx** (C7), **INTER-STRAND** (C8), **ACID-NETWORK** (C9), **BASE-NETWORK** (C10), **HELIx-STRAND** (C11) and **LOCAL-COIL** (C12) are shown with orientation-I (C1 and C3 with basic residue is at N-terminal) and orientation-II (C2 and C4 with acidic residue is at N-terminal) for representative candidate pairs. Bond multiplicity (1, 2, 4 etc; white line), Average distance, local/non-local (**L/\text{nL}**) (C13).

Execution of the program with different modes

The program runs in three different modes such as mode-1, mode-2 and mode-3. Purpose of the first two modes is extraction of surface and core location of salt bridges (**SBs**) by two alternative paths. Execution of Mode-3 is different and it runs in different path (Figure 7). If the program is run with option 1 or 2, it’s upon listing of PDB files also checks for existence of Accessible Surface Area (ASA) software in the specified directory (Detailed at [http://sourceforge.net/projects/sbion2/files/README.txt/download](http://sourceforge.net/projects/sbion2/files/README.txt/download)) for extraction of surface and core location of salt bridges (see below).

As input users need to provide PDB files, ion-pair distance, ASA and Local/non-Local threshold (see screenshots at [http://sourceforge.net/projects/sbion2/](http://sourceforge.net/projects/sbion2/)). NMR files are identified and screen out and only crystal structures are processed. Execution of the program allows user to avail three different modes (identified as **1/2/3**) (Figure 8 A). User needs to provide ion-pair distance (for salt bridge, the distance is \( \leq 4 \) Å), ASA threshold and local/non-Local threshold for mode 1 and 2. For mode 3, ion-pair distance and local/non-local threshold are to be provided during the execution.

As output, Firstly the program lists results in binary items 11 of which are for isolated **SBs** (Figure 7, **O-2**) and 11 outputs are for networked **SBs** (Figure 7, **O-3**).
Secondly, for a given chain of protein, \textit{SBs} with INTRA-helical/strand/coil or INTER-helical/strand/coil with \(i \to i+n\) type of connections are reported in excel files (Figure 7, \textit{O-4}). Representative \textit{SBs} are presented in Figure 8 (\textit{C1} through \textit{C12}).
Figure 10: Execution of the program with Mode 2
Figure 11: Execution of the program with Mode 3
Execution of the program SBION 2 (Gupta et. al, 2015) with different modes (MODE-1, MODE-2 and MODE-3) is shown in figures 9, 10 and 11 respectively. In figure 12 it is observed that the most popular user with maximum downloads of SBION2 are from United States.

2.2 ADSBET, ADSBET2 and ADSETMEAS: Automated Determination of Salt-bridge Energy Terms and Micro Environment from Atomic Structures using APBS method

2.2.1 Background
Salt bridges are electrostatic interactions occurs between groups of two or more opposite charges. Net energy of a salt-bridge can be calculated by summing bridge
desolvation ($\Delta \Delta G_{\text{dsolv}}$) and protein ($\Delta \Delta G_{\text{prot}}$) energy terms. $\Delta \Delta G_{\text{brd}}$ energy and $\Delta \Delta G_{\text{dsolv}}$ energy originated due to the direct Coulombic attraction and desolvation of opposite charges respectively. On the other hand, $\Delta \Delta G_{\text{prot}}$ is originated due to spatial proximity and interaction of salt-bridge partners with other charged or dipole atoms in protein-chain. Estimation of Salt-bridge energy has to be worked out by computational means as calculation of $\Delta \Delta G_{\text{dsolv}}$, $\Delta \Delta G_{\text{prot}}$ and hence $\Delta \Delta G_{\text{net}}$ are not possible by the experimental one. Thus, the calculation of energy terms by computational means is the only choice.

Although structural database is increasing rapidly, very limited number of crystal structures ($\leq 0.01\%$) is worked out for salt-bridge energy terms, their contribution in protein stability remains an enigma. To gain further insight into the observation that they could either be stabilizing or destabilizing or neutral, an efficient algorithm or software that would analyze large dataset would be useful. To the best of our knowledge, such software is truly lacking in public domain.

In this end, our fully automated development ADSETMEAS [improvement over ADSBET (Nayek et. al, 2015) and ADSBET2 (Nayek et. al, 2015) uses Adaptive-Poisson-Boltzmann-Solver (APBS) method to compute component as well as net energy-terms of salt-bridges and redirect results in easy to use excel-format. The procedure provides users choice for selecting parameters such as a] model for component energy terms to-date available in the literature along with additional one and b] method (default or advanced) with parametric optimization for APBS calculations. The program processes all crystal structures available in the working directory with any number of salt-bridges (atomic-distance $\leq 4\AA$) or ion-pair in them for the calculation of energy terms. Overall our development ADSETMEAS, provides accurate details on salt-bridges energetic and find application in the field of structural and Computational biology.

### 2.2.2 Features and implementation of the program

Automated Determination of Salt-Bridge Energy Terms and Micro Environments from Atomic Structures, is a program that uses capability of Adaptive Poisson-Boltzmann Solver (APBS: an open source all purpose PBE solver; Baker et. al, 2001) for
computation of salt-bridge or ion-pair energy-terms. ADSETMEAS is the improved version of the software ADSBET (Nayek et. al., 2015) and ADSBET2 (Nayek et. al, 2015). It also partitions $ddG_{prot}$ energy-terms into 4 different sub-terms that help in understanding micro environment of salt-bridge. Crystal structures in RCSB PDB file format are used as inputs (that are kept in the working directory) by the program. It works on any numbers of structure file with any numbers of salt-bridges present in them and redirect output in excel format. The program is freely available to academic and commercial users. It comes under GNU GPL License version 2.0. The program works from any UNIX like environment including CYGWIN.

The program ADSETAMES uses:

A. "APBS v1.4 (binary)", as PBE solver (Baker, et. al, 2001),

B. "PDB2PQR v1.9 (binary)" for generation of pqr file (Dolinsky, et. al, 2001) and

C. "SBION2"/"SBION" for obtaining salt-bridges in residue form (Gupta et. al, 2014 & 2015) for functioning.

Salt-bridge energy and different energy terms

Both experimental and computational methods are used to estimate energy of salt-bridge. In experiment, pKa and double mutation cycle methods are popularly used. However, these methods could not determine desolvation and background energy terms. Thus, computational method is the only choice for obtaining estimate of desolvation ($\Delta\Delta G_{desolv}$), bridge ($\Delta\Delta G_{brd}$) and background ($\Delta\Delta G_{prot}$) energy-terms (Bosshard, et. al, 2004). However, appropriate modeling of these energy terms and their estimate by continuum electrostatic model are necessary. The $\Delta\Delta G_{tot}$ is obtained by summing these component energy terms.

$$\Delta\Delta G_{tot} = \Delta\Delta G_{desolv} + \Delta\Delta G_{brd} + ddG_{prot} \quad [1]$$

Desolvation energy ($\Delta\Delta G_{desolv}$) is due to desolvation of charges forming salt-bridges. It is always costly (Hendsch and Tidor, 1994). Computationally it is estimated by equation 3 (see below).

Background energy ($\Delta\Delta G_{prot}$) could be modeled in different ways. It is the energy
due to background interaction of salt-bridge that originates due to side chain charges other than salt-bridge residues (Hendsch and Tidor, 1994; Kumar and Nussinov, 1999) or due to side-chain of permanent dipoles such as HELIX, STRAND and non-dissociable polar amino acids (Bosshard, et. al., 2004). This energy term can either be contributing or costly. If the later is taken into consideration for the estimation of $\Delta \Delta G_{\text{prot}}$, then rest of the residues in protein-chain are used to estimate other effects. Computationally this energy can be estimated by equation 4 (see below).

**Bridge energy** ($\Delta \Delta G_{\text{brd}}$) is the direct interaction of salt-bridge partners in the folded state of protein. It is always contributing as it originates from interaction between negatively charged acidic and positively charged basic residues. Computationally it is estimated by equation 1 and 2 (see below)

**Why ADSETMEAS would be useful?**

- Note that APBS (ADAPTIVE POISSON-BOLTZMANN SOLVER) is a general purpose PBE solver that can be used for a number of electrostatic calculations such as generation of electrostatic surface, obtaining charge density, determination of binding energy of biomolecular complexes etc. Thus, APBS is popularly used with molecular graphics programs such as Pymol, VMD for computation and visualization of electrostatic potential of biomolecules. These plugin-based methods also provide easy use of APBS for other similar purposes.

- Careful investigation of literatures as to use of APBS (or other similar solver such as Delphi) for computation of salt-bridge energy-terms compare to its other applications, show that the former is very less. Overall, about $\leq 2\%$ of the database crystal structures are worked out for the purpose till date.

- Such low growth in salt-bridge or ion-pair energy-terms might be due to the fact that it needs multi-steps involved processes such as:

- Determination of residue specific salt-bridges or ion-pairs of all proteins present in the working directory
Determination of protein specific center of mass would be needed for desolvation energy computation

Determination of protein specific grid points (along x, y and z dimension) to be used in APBS calculation

Preparation of pqr (charge-radius) file from pdb file

Mutation of residues in original pqr file, by their respective hydrophobic isosteres (5 extra files per PDB file) is needed for computation of electrostatic field energy calculation.

Preparation of APBS input file based on method (default or Advanced iterative focus calculation) demanded by users

Flexibility in choice of model (based on side-chain definition of salt-bridge residues and inclusion of background residues for computation of $\Delta \Delta G_{\text{prot}}$) to date available in the literature

Partitioning of $\Delta \Delta G_{\text{prot}}$ based on different classes of amino acids (polar, dipolar, positively-charged, negatively-charged, hydrophobic) to understand effect of Micro environment around salt-bridge.

Redirection of all energy terms in compact excel format along with relevant documentation for further uses.

ADSETMEAS performs all the above tasks for any number of proteins with any number of salt-bridges/ion-pairs in them, in a single run.

Model for computation of energy terms

Let us consider a salt bridge $[P_{ms}^i]-[N_{ms}^j]$ where $P_{ms}^i$ and $N_{ms}^j$ are positive and negative charge residues at position $i$ and $j$ respectively. Here, $m$ and $s$ indicate main and side chain atoms respectively.

Bridge energy model

$\Delta \Delta G_{\text{brd}}^{p_i N_j}$ is resulted from interaction of side chain atoms of positive ($P^s$) and negative charge ($N^s$) residues in the folded state of protein. According to the working model, in the native protein-charge-radius (PQR) file, charges of either $P^s$ or $N^s$ is retained and others are mutated by hydrophobic isosteres. The generated PQR file is then
subjected for manually configured multigrid Poison Boltzmann calculation under single Debye-Huckel boundary condition (mPBsDH) by APBS method\(^9\). The atomic potential file thus obtained is used for computation of bridge energy term ($\Delta \Delta G_{brd}^{pi/Ni}$) using following formula:

$$\Delta \Delta G_{brd}^{pi/Ni} = [\sum_s(q_{Ni}^s * \phi_{Ni}^s)]$$  \[When atomic charges of $P^s$ is used for mPBsDH\]  

Equation 1

Or

$$\Delta \Delta G_{brd}^{pi/Ni} = [\sum_s(q_{Ps}^s * \phi_{Ps}^s)]$$  \[When atomic charges of $N^s$ is used for mPBsDH\]  

Equation 2

**Background energy model**

$\Delta \Delta G_{prot}^{pi/Ni}$ is the interaction energy of $P^s$ and $N^s$ with $O^s$ (other side chain atoms in protein i.e. all except $P^s$ and $N^s$) in the folded state of protein\(^5\). In this case mPBsDH is solved on a generated PQR file that retains atomic charges only for $P^s$ and $N^s$. The potential file thus obtained was used for calculation of $\Delta \Delta G_{prot}^{pi/Ni}$ using the following formula:

$$\Delta \Delta G_{prot}^{pi/Ni} = [\sum_s(q_{Os}^s * \phi_{Os}^s)]$$  \[When atomic charges of $P^s$ and $N^s$ are used for mPBsDH\]  

Equation 3

**Desolvation energy model**

It is calculated separately on $P^s$ and $N^s$ for both folded and unfolded state of protein\(^5\) using the following formula:

$$\Delta \Delta G_{dsolv}^{pi/Ni} = \left[\left(G_{n_{sol}}^{P^s} - G_{n_{vac}}^{P^s}\right) + \left(G_{n_{sol}}^{N^s} - G_{n_{vac}}^{N^s}\right)\right] - \left[\left(G_{u_{sol}}^{P^s} - G_{u_{vac}}^{P^s}\right) + \left(G_{u_{sol}}^{N^s} - G_{u_{vac}}^{N^s}\right)\right]$$

Equation 4

$G_{n_{sol}}^{P^s}$ is the reaction field energy of side chain atoms of $P^i$ in native solvated condition

$G_{n_{vac}}^{P^s}$ is the reaction field energy of side chain atoms of $P^i$ in native vacuum condition

$G_{n_{sol}}^{N^s}$ is the reaction field energy of side chain atoms of $N^j$ in native solvated condition

$G_{n_{vac}}^{N^s}$ is the reaction field energy of side chain atoms of $N^j$ in native vacuum condition
In unfolded state of protein (right part of right hand side equation) terms has similar meaning.

In this calculation, PQR file possess \( m \) of \((i-1)\), both \( m \) and \( s \) of \( i \) and \( m \) of \((i+1)\) with atomic charges only on \( s \). In this case \((i-1)\) and \((i+1)\) acts as local backbone for the residue \((P^i)\). Similar consideration is given on other salt bridge partner \((N^j)\)\(^5\). Calculations for both folded and unfolded states were performed at identical center of grid points.

**Net salt bridge energy model**

Net salt bridge energy was obtained by summing component energy terms\(^7\) i.e.

\[
\Delta \Delta G_{net} = \Delta \Delta G_{brd} + \Delta \Delta G_{prot} + \Delta \Delta G_{dsov}
\]

**Equation 5**

**Residue specific salt bridges determination**

Side chain of acidic (acceptor: D & E) and basic (donor: R, K, H) residues in folded protein form salt bridges. ADSBET adapts features of SBION and SBION2 to obtain residue specific 1:1 salt bridges for computation of energy terms. Networked and multivalent salt bridge contribution was resolved based on their architecture during post run analysis.

**Program input:**

Poison Boltzmann Equation solvers (such as APBS) require multi-parameters input for generation of reaction field energy. Thus, apart from PDB-files in the working directory users need to input parameters (Figure 13: FIII: User pars) such as model for side-chain \((F1)\), ASA method, grid-spacing, pH, mobile ion concentration (in Molar), protein-dielectric constant of protein, salt-bridge or ion-pair distance and force-field. Protein specific parameters (such as dime, gcent and grid-points) are auto-generated by the program and other parameters are used as default (FIII: Default pars) if not mentioned otherwise.
Program output:

Details of SBS-energetics and accessibility for any number of SBS in any number of PDBs are redirected into two different named-outputs. Unlike ADSBET, ADSBET2 redirects most relevant model-based additional output (Figure 13: FIII: Green color parts) on overall and normalized energetics in monomer, ISBS and NSBS format along with their core and surface location information.

Figure 13: Model-based side-chains (F1), different SBS (FII) and flow-chart (FIII) for ADSBET2. MDL-1 and MDL-2 are used in model-1 and model-2 respectively. S1 (energy E1) and N1 (energy E3 and E4) are single-bonded isolated and networked SBS. S2 (energy E2) and N2 (energy E5 and E6) are multiple-bonded isolated and networked SBS. P shows core and
surface SBS. In FIII, $N-A^S$ and $U-A^U$ are acidic side-chains in folded and unfolded state of protein respectively. Additional features in ADSBET2 are shown in green color.

- The program is downloaded by the worldwide users (figure 6), dominant user are from India followed by Spain.

![Figure 14: Download statistics of the software ADSBET2 from Worldwide users.](image)

**References**

window-dependent and substitution based evolutionary properties of protein sequences

- Bhaskaran R, PonnuSwamy PK. Positional flexibilities of amino acid residues in globular

- Bosshard HR, Marti DN, Jelesarov I. Protein stabilization by salt bridges: concepts,
experimental approaches and clarification of some misunderstandings. Journal of

- Carvalho GR, Hauser L. Molecular genetics and the stock concept in fisheries. Reviews in
Fish Biology and Fisheries. 1994 Sep 1;4(3):326-50.

- Chothia C. The nature of the accessible and buried surfaces in proteins. Journal of

- Deleage G, Roux B. An algorithm for protein secondary structure prediction based on class


- Eisenberg D, Weiss RM, Terwilliger TC. The hydrophobic moment detects periodicity in
protein hydrophobicity. Proceedings of the National Academy of Sciences. 1984 Jan
1;81(1):140-4.

- FAUCHÈRE JL, Charton M, Kier LB, Verloop A, Pliska V. Amino acid side chain
parameters for correlation studies in biology and pharmacology. Chemical Biology & Drug

- Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach.

- Gasteiger E, Hoogland C, Gattiker A, Duvaud SE, Wilkins MR, Appel RD, Bairoch A.

- Gupta PS, Banerjee S, Islam RN, Mondal S, Mondal B, Bandyopadhyay AK. PHYSICO:
An UNIX based Standalone Procedure for Computation of Individual and Group Properties

- Gupta PS, Banerjee S, Islam RN, Sur VP, Bandyopadhyay AK. Substitutional Analysis of

- Gupta PS, Mondal S, Mondal B, Islam RN, Banerjee S, Bandyopadhyay AK. SBION: A
program for analyses of salt-bridges from multiple structure files. Bioinformation.
2014;10(3):164.


CHAPTER 1
Understanding the Role of Amino Acids Substitution in the Evolution of Homologous Protein Families

Summary
Homologous/Orthologous proteins emerged due to divergence from their parent. They largely have similar secondary structure sequence and topology thus perform similar function under different environmental conditions. Substitution of amino acids at locus specific positions in BLOCK i.e. formation of hetero-pairs, play vital role in its evolution. However, differential contribution of hetero-pairs (total 190 types) in a given functional clade (i.e. a BLOCK of sequences) remains to be understood. Here we involve thirty homologous protein BLOCKs of known divergence rate to work out a general view on evolution. Our procedure APBEST uses BLOCK FASTA file to extract BLOCK specific evolutionary parameters like dominantly used hetero-pair (DHPs), usage of hetero-pairs (E), non-conservative to conservative substitution ratio (R), maximally-diverged residues (MDRs), maximally conserve residues (MCRs), residue (RD), class specific diversity (CD) and others. It is observed that DHPs is largely (88%) conservative in nature pointing towards restoration of function. Inter-relationship among MDRs, DHPs and MCRs reveal secondary structure, its burial and stability for BLOCKs. Distinct MCRs have characteristic functional role. While E sets the upper-limit of usage of hetero-pairs, the positive, linear correlation between R and divergence-rate indicates that the later is directly dependent on non-conservative substitutions. The observation that MDRs and MCRs occupy alternate positions in BLOCK indicates both divergence and conservation are important. While Positional diversity/heterogeneity occurs due to substitution, protein function demands preservation of observed conservation i.e. parental form. Thus a balance between diversity and conservation seems to be the strategy in evolution of homologous proteins.
**Introduction**

Homologous proteins, emerged due to speciation event, are structurally and functionally similar (Betts and Russell, 1993). Evolution accommodates changes in these sequences. Amino acid changes are mostly achieved by substitution, deletion and insertion mechanisms of which earlier is the result of accumulation of changes at locus specific positions. In evolution, two types of substitutions namely conservative (CS) and non-conservative (NCS) occur of which most of the later changes are deleterious. Thus these are eventually eliminated through purifying selection. Beneficial ones (both conservative and non-conservative) are restored in sequence population and thus contribute to species differentiation (Ng and Henikoff, 2006). Comparison among homologous sequences of database reveal sequences of closely related species (e.g. human and mouse) are more similar than that of distantly related species (human vs. bacteria). When homologous positions (column-wise in a BLOCK) are fixed, it would be seen that each of these positions bears characteristic details. While some are invariant (INV-line), other is either conservative (CS) or non-conservative (NCS) types of substitutions (Ng and Henikoff, 2006). Henikoff and Henikoff (1992) pioneered the concept of BLOCK of sequences. A BLOCK contains homologous sequences whose allelic positions are fixed. These types of BLOCKs of different level of sequence similarity were used to develop different series of average BLOSUM matrices (Henikoff and Henikoff, 1992).

The concept divergence rate has become an important tool in the assessment of mechanisms of diversification in sequence evolution (Hendry and Kinnison, 2001). Table values of divergence rates of few of these protein families are available (Dayhoff and Schwartz, 1978). Although different homologous proteins possess different divergence rate (Dickerson 1971; Dayhoff and Schwartz, 1978), for a given family, it is seen to be constant (dos Reis et. al, 2016). For example, fibrino-peptide, a blood clotting factor, has the highest and histone, a DNA binding protein, has the lowest divergence rate (Dickerson, 1971; Dayhoff and Schwartz, 1978). The variability in these rates was related to structural and functional requirements of these molecules (Tourasse and Li, 2000). In this aspect, great deals of studies and developments are available (dos Reis et. al, 2016; Dayhoff and Schwartz, 1978). Understanding the mechanism of substitutions largely
involve comparison of locus-specific positions (Marini et al., 2010) for its effect on physicochemical properties (Andreas et al., 2001) and identity (Larkin et al., 2007) or similarity (Altschul et al., 1990). Similarity or identity scores are used for pair-wise comparison of sequence that eventually helps their alignment, finding relatedness (Altschul et al., 1990), obtaining functional significance and constructing phylogenetic trees (Andreas et al., 2001; Gabaldon and Koonin, 2013). Further sequence based studies also include analyses and extraction of information from INDEL regions of alignment. It is an additive alternative to substitution-mechanism for understanding protein evolution (Ajawatanawong and Baldauf, 2013). While these studies have widened our understanding in different aspects of molecular evolution of protein sequences, the governing principles of evolution for homologous protein families in relation to acquired substitutions (i.e. the usage of observed hetero-pairs) still remain an enigma. Fundamental question concerning the non-conservative substitutions, as to how these are managed in these functionally similar proteins when they are known to be deleterious (Ng and Henikoff, 2006; Chun and Fay, 2009) remain to be answered. Further it remains a challenge to understand the relationship of maximally diverse residues (MDRs) and maximally conserved residues (MCRs) with the restoration of secondary structure sequence for similar topology of homologous proteins in BLOCK.

In this work, we report results on HPs (hetero-pairs) for thirty protein BLOCKs of known divergence rate (Dayhoff and Schwartz, 1978), that were used to develop a general model of evolution for homologous proteins. BLOCK parameters (DHPs, R, E, MDRs, MCRs, RD and CD) are computed efficiently using APBEST. Importantly the role of R, having high correlation with divergence rate, and MDRs and MCRs, positional distribution, are highlighted for the first time in this work. Computed diversities (both residue and class specific diversity) and their implication are also been worked out in the study.

Materials and Methods

Collection of data and retrieval of sequences
A total of thirty homologous protein families (downloadable from https://sourceforge.net/projects/apbest/files/) with known divergence rate (Dayhoff and
Schwartz, 1978) are taken in the present study. These families were chosen in such a way that their divergence rate gives a wide coverage. For example H2A has 0.5% per 100/mYr and that for Kappa-casein is 33% per 100/mYr (Dayhoff and Schwartz, 1978). Family specific sequences were obtained from UNIPROT (UniProt Consortium, 2008) and other inter-linked databases. Duplicate IDs if they were present in the sequence BLOCK is subjected for verification by the program APBEST (Gupta et. al, 2017) and terminate analysis until removed. Obtained sequences were then aligned using ClustalW2 (Larkin et al., 2007) for each of the protein families for the construction of BLOCK FASTA file.

**Preparation of BLOCK FASTA files**

BLOCK-FASTA files were prepared using automated block preparation tool of PHYSICO2 (Banerjee et al., 2015). As the method involve manual step during removal of partial sequences, care was taken such that maximal sequence information is restored in the BLOCK. The BLOCK FASTA file thus produced was used as input for APBEST. An example input BLOCK fasta file can be downloaded at [http://sourceforge.net/projects/apbest/files/](https://sourceforge.net/projects/apbest/files/). All BLOCK fasta files used in the study are also downloadable from [http://sourceforge.net/projects/apbest/files](http://sourceforge.net/projects/apbest/files).

**Analyses of BLOCK FASTA file and Extraction of evolutionary parameters**

Analysis of BLOCK FASTA files were performed using in house procedure APBEST. The program is written in AWK-programming language that runs in CYGWIN-UNIX like environment. It is efficient, error free and user-friendly. A compact itemized (Item A through I) output is redirected in excel format. The program and example output file is freely available at [http://sourceforge.net/projects/APBEST/](http://sourceforge.net/projects/APBEST/) for academic users. Of all items in the output file, item-2 lists evolutionary parameters. These are 1] sum of hetero-pairs (SHP), 2] homo-pairs (SMP) and 3] total pairs (TP) frequencies; 4] sum of hydrophobic(HB)-to-hydrophobic (HBS), 5] hydrophilic (HL)-to-hydrophilic (HLS) and 6] hydrophobic-to-hydrophilic (NCS), 7] minimally diverse residue, 8-10] maximally diverse residue (MDR in MDR1, MDR2 and MDR3 format), 11-13] dominantly used hetero-pair (DHR in DHR1, DHR2 and DHR3 format), 14-16] maximally conserved
residue (MCR in MCR1, MCR2 and MCR3 format), 17-21] Class-specific diversity (CD: acidic, basic, hydrophobic, hydrophilic and polar), 22] non-conservative to conservative ratio (R) and 23] ratio of hetero-pair to total pair (E) are computed using relevant observed frequency of homo-pair and hetero-pair (output file). In this item (Item-2) program also list non-frequency parameters such as 24] unused hetero-pair count, 25] positional counts of invariant (INV-line), 26] hydrophobic (HB-line), 27] hydrophobic-hydrophilic (HB-HL line) and 28] hydrophilic lines (HL line).

A hetero-pair harbors different types of amino acids in a locus of a given pair of homologous proteins. They originate due to substitution in the course of evolution. Homo-pair in turn is a pair that harbors identical type of amino acid at any locus of a given pair of homologous proteins. For a given BLOCK, there are 190 hetero-pairs and 20 homo-pairs (Henikoff and Henikoff, 1992). The program first computes hetero-pair (HP) and homo-pair (MP) frequencies. Frequency of Observed heteropair (HP) for a given pair (e.g. XY) is computed for each column position (m) of BLOCK and then summed over all positions (w) to obtain BLOCK (B) frequency.

\[ f_{XY}^B = \sum_{m=1}^{w} f_{XY}^m = \sum_{m=1}^{w} \left[ n_x C_1 * n_y C_1 \right]_m = \sum_{m=1}^{w} \left[ n_x n_y \right]_m \]  \[ 1a \]

\[ f^{RPF}_{XY} = \frac{f_{XY}^B * 100}{f_{SHP}^B} \]  \[ 1b \]

Where \( n_x \) and \( n_y \) are frequencies of amino acid \( X \) and \( Y \) respectively for a given column position (m) of block B of width w. Block specific observed frequency (\( f_{XY}^m \)) for a given pair (XY) is the sum of frequencies of all column positions. \( f_{XY}^{RPF} \) is the relative percentile frequency or probability for the hetero-pairs \( XY \).

Similarly the observed frequency of SMP for a given pair (say ZZ) is computed as

\[ f_{ZZ}^B = \sum_{m=1}^{w} f_{ZZ}^m = \sum_{m=1}^{w} \left[ n_z C_2 \right]_m = \sum_{m=1}^{w} \left[ \frac{n_z (n_z - 1)}{2} \right]_m \]  \[ 2a \]

\[ f^{RPF}_{ZZ} = \frac{f_{ZZ}^B * 100}{f_{SMP}^B} \]  \[ 2b \]

Where \( n_z \) is the count of amino acid \( Z \) for a given column position (m).
BLOCK parameters (item 2: output file) are computed based on observed frequencies of homo-pairs and hetero-pairs (Equation 1a and 2a). Sum of frequencies of all 190 hetero-pairs give $SHP$ and that for 20 homo-pair gives $SMP$. $TP$ is the sum of $SHP$ and $SMP$.

\[ SHP = \sum_{i=1}^{20} \sum_{j \neq i} f_{X_i,Y_j} \]  
\[ SHP\% = \frac{SHP \times 100}{TP} \]  
\[ SMP = \sum_{i} f_{Z_i,Z_i} \]  
\[ SMP\% = \frac{SMP \times 100}{TP} \]

Where $TP = SHP + SMP$

Hetero-pairs are divided into three categories such as HB-HB (hydrophobic-hydrophobic), hydrophilic-hydrophilic (HL-HL) and hydrophobic-hydrophilic (HB-HL) types. Sum of hetero-pair frequencies of hetero-pair types of each of these categories give us $HBS$ (Hydrophobic substitutions), $HLS$ (hydrophilic substitutions) and $NCS$ (non-conservative substitutions i.e. hydrophobic to hydrophilic substitutions) parameters respectively. Sum of $HBS$ and $HLS$ produce conservative substitutions ($CS$). Conservative ($CS$) does not alter overall protein’s properties (such as DE) whereas NCS may alter protein’s physico-chemical properties (such as DK).

$DHP$ indicates $HP$ whose frequency is highest.

Each of 20 residues diversities are computed by summing up of hetero-pair frequencies for a given residue ($X$).

\[ RD_{X} = \sum_{k=1}^{19} f_{X_k} \]  

Here $k$ is any of 20 amino acids except the residue $X$. Maximum and minimum diversities ($MDR$ and $MiD$ respectively) are determined from this set of residues diversities. Maximally conserved residue ($MCR$) is the residue whose homo-pair frequency exceeds that of others. Instead of only one, first three top scores in descending order of each of $DHP$ (as $DHP1$, $DHP2$ and $DHP3$), $MDR$ (as $MDR1$, $MDR2$ and $MDR3$) and $MCR$ (as $MCR1$, $MCR2$ and $MCR3$) are extracted from BLOCK.
All SHP (190 types) are divided into two categories namely conservative (CS) and non-conservative (NCS) for computation of $R$.

$$R = \frac{f_B^{B-HL} \times 100}{(f_B^{B-HB} + f_B^{B-HL})} = \frac{NCS \times 100}{CS}$$ \hspace{1cm} [5]\]

Thus, $R$ is defined as non-conservative to conservative substitution ratio, $f_B^{B-HL}$: Hydrophobic to hydrophilic, $f_B^{B-HB}$: hydrophobic to hydrophobic and $f_B^{B-HL}$: hydrophilic to hydrophilic SHP frequencies.

$E$ (the usage parameter for hetero-pair) is the normalized frequency of SHP for a given BLOCK which is computed as:

$$E = \frac{f_B^{B-HP} \times 100}{f_B^{B-HP} + f_B^{B-MP}} = \frac{SHP \times 100}{TP}$$ \hspace{1cm} [6]\]

$f_B^{B-HP}$: Substitution-hetero-pair frequency and $f_B^{B-MP}$: Substitution-homo-pair frequency.

Superscript $B$ indicate sum of all HPs frequency for a given BLOCK. $E$ is 0 when hetero-pair frequency is absent and no evolution of sequence BLOCK i.e. $f_B^{B-HP}$ is 0. On the other hand, if $f_B^{B-MP}$ is 0 then $E$ is 1 which means maximum evolution of sequence BLOCK.

$CD$s (classes or CL: acidic, basic, hydrophobic, hydrophilic and polar) are calculated as:

$$CD_{CL} = \sum_{j=1}^{n_{CL}} \left( \sum_{k=1}^{r} f_{j,k} \right) - \left( \sum_{j=1}^{n_{CL}} \left( \sum_{i=j+1}^{n_{CL}} f_{j,i} \right) \right)$$ \hspace{1cm} [7]\]

The above equation contains two terms. The first term is for computation of the sum of diversities of candidate residues belong to the class and the second term is the sum of frequencies of overlapping frequency of HP formed by candidate residues of the class.

Total number of overlapping hetero-pairs is obtained by $n_{CL}C_2$. $CL$ is the class and $n_{CL}$ is the total count of class residues (for basic residues HRK, CL=basic class, $n_{CL}$=3). For computation of diversity for each residue of a class, that residue is excluded from the list of 20 amino acids (e.g. ACDEFGHIKLMNPQRSTVWY). Thus, $r$ is always constituted by 19 residues except the one for which HP diversity is considered to avoid inclusion of its homo-pair. For example, basic class diversity:

$$DV_{HRK} = (f_{\text{hetero-pair of H}} + f_{\text{hetero-pair of R}} + f_{\text{hetero-pair of K}}) - (f_{HR} + f_{HK} + f_{RK})$$

$f_{\text{hetero-pair of H}} = f_{HA} + f_{HC} + f_{HD} + f_{HE} + f_{HH} + f_{HS} + f_{HT} + f_{HV} + f_{HW} + f_{HV} + f_{HW}$
Residues constituting a class can be written in any given order (e.g. HRK or RHK or KHR etc for basic class) to assign the value of $j$. Each will return identical value of diversity. The second term always contains $n_{CL}C_2$ hetero-pairs whose frequencies are summed twice ($HR$, $HK$ and $RK$) in the net sum of first term. Therefore sum of the frequency of these terms is deducted once from the first term.

$MDR$ and $MCR$ are sum of maximally diverse and conserved residues. To know the distribution of these ensemble frequencies along different positions of BLOCK, homologous position specific frequencies of these parameters are determined (Item C, D for $MDR1$ and $MDR2$ and item E for $MCR1$ and $MCR2$: output file).

APBEST further computes observed probability range specific types and frequency of HPs in ready table format (Item G, output file) suitable for plot.

Finally, the program extracts position specific BLOCK parameters (Item I; output file). Shannon entropy (Shannon, 1948) is computed using the following formula:

$$H = -\sum_{i}^{M} P_{i} \log_{2} P_{i}$$  \[8\]

Where $P_{i}$ is the fraction of residues of amino acid types $i$, and $M$ is the number of amino acid types (20 in number). $H$ ranges from 0 (only one residue in present at that position) to 4.322 (all 20 residues are equally represented in that position). Typically, positions with $H \geq 2.0$ are considered variable, whereas those with $H \leq 2$ are considered conserved. Highly conserved positions are those with $H \leq 1.0$.

BLOCK positions undergo different types of substitutions. Different positions of BLOCK are also assessed based on residue types. If there is only one type of amino acid in a given position then it is marked as invariant. If substituted then qualitatively positional substitutions are assessed as different categories such as hydrophobic-hydrophobic, hydrophilic-hydrophilic and hydrophobic to hydrophilic etc.
APBEST keep all these results in itemized form in excel output (https://sourceforge.net/projects/apbest/files/) for analyses.

**Results**

To explore evolutionary, structural and functional significance of HP (heteropair) for any given homologous protein family, we have analyzed thirty homologous protein BLOCKs of known divergence rate with the help of APBEST. Details about BLOCKs used in the present study are provided in supplementary material that includes FASTA file name (downloadable from https://sourceforge.net/projects/apbest/files/), BLOCK-width, number of sequences in BLOCKs, mean properties such as $\overline{HP}$, $\overline{HL}$, $\overline{AL}$, $pI$, $\text{GRAVY}$ of sequence segments in BLOCKs, PDB file used for secondary structure assessment, helix and sheet content in these structures. Other analytical results are presented in Table 1 through 3 and Figure 1 through 2 (see below). A typical output of the program could be freely downloaded at (https://sourceforge.net/projects/apbest/files/) for clarity. In the output, there are details of nine different items (Item A through I). Items A to H compute quantitative results on substitutions and Item I provides qualitative and quantitative insight into positional mutations and variability respectively.

APBEST performs extensive analysis on each BLOCK and lists homo-pair and hetero-pair related 28 types of evolutionary parameters in item-2 of outputs. Frequency of conservative ($\text{HBS}$, $\text{HLS}$) and non-conservative ($\text{NCS}$) substitutions, Ratio-parameters ($\text{R}$, $\text{E}$ and $\text{N}$) and class-specific diversities ($\text{DVac}$, $\text{DVbs}$, $\text{DVhb}$ and $\text{DVhl}$) are presented in Table 1. Maximally diverse residues (in MDR1:MDR2:MDR3 format), dominant hetero-pairs (in DHR1; DHR2; DHR3 format) and maximally conserved residues (in MCR1:MCR2:MCR3 format) are presented in Table 2. Table 3 contains homologous position specific mutation and SHANON-

**Table 1: Analyses on residue type at homologous positions and class-specific diversity**

<table>
<thead>
<tr>
<th>Protein name</th>
<th>DV*</th>
<th>HBS</th>
<th>HLS</th>
<th>NCS</th>
<th>R</th>
<th>E</th>
<th>N</th>
<th>AcD</th>
<th>BaD</th>
<th>HBD</th>
<th>HLD</th>
</tr>
</thead>
</table>
conservation details. Details of formulation of each of these parameters are described in the Material and Method section.

| Protein                | Histone H2A | Histone H2B | GDH α | Troponin C | Plant ferredoxin | Cytochrome C | Corticotrophin | TPI | LDH | Plastocyanin | Insulin | α Crystalline B | Troponin C | Plant ferredoxin | Cytochrome C | Corticotrophin | TPI | LDH | Plastocyanin | Insulin | α Crystalline B | Troponin C | Plant ferredoxin | Cytochrome C | Corticotrophin | TPI | LDH | Plastocyanin | Insulin | α Crystalline B | Troponin C | Plant ferredoxin | Cytochrome C | Corticotrophin | TPI | LDH | Plastocyanin | Insulin | α Crystalline B | Troponin C | Plant ferredoxin | Cytochrome C | Corticotrophin | TPI | LDH | Plastocyanin | Insulin | α Crystalline B | Troponin C | Plant ferredoxin | Cytochrome C | Corticotrophin | TPI | LDH | Plastocyanin | Insulin | α Crystalline B | Troponin C | Plant ferredoxin | Cytochrome C | Corticotrophin | TPI | LDH | Plastocyanin | Insulin | α Crystalline B | Troponin C | Plant ferredoxin | Cytochrome C | Corticotrophin | TPI | LDH | Plastocyanin | Insulin | α Crystalline B | Troponin C | Plant ferredoxin | Cytochrome C | Corticotrophin |
|------------------------|-------------|-------------|-------|------------|------------------|--------------|----------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|--------------|------------------|-----|-----|-------------|---------|-----------------|------------|------------------|percent/100 MYr; DV: Divergence rates (second column of the table) for protein BLOCKs (first column) are taken from (Marks, 1988; Dayhoff and Schwartz, 1978). LDH: Lactate dehydrogenase; GDH: Glutamate dehydrogenase; TPI: Triose Phosphate Isomerase; HBS: Hydrophobic Substitution; HLS: Hydrophilic Substitution; NCS: Non-Conservative Substitution

*percent/100 MYr; DV: Divergence rates (second column of the table) for protein BLOCKs (first column) are taken from (Marks, 1988; Dayhoff and Schwartz, 1978). LDH: Lactate dehydrogenase; GDH: Glutamate dehydrogenase; TPI: Triose Phosphate Isomerase; HBS: Hydrophobic Substitution; HLS: Hydrophilic Substitution; NCS: Non-Conservative Substitution
Figure 1 show a typical presentation of off-diagonal hetero-pairs (total 190), with its frequency at the lower half and types at the upper half of the diagonal. These hetero-pairs (HPs) are divided into three classes: hydrophobic-hydrophobic (HBS i.e. HB-HB; upper: black shade region; total 36 elements), hydrophilic-hydrophilic (HLS i.e. HL-HL; lower, white region; total 55 elements) and hydrophobic-hydrophilic (NCS i.e. HB-HL; middle, gray shade region; total 99 elements). Upper and lower classes (HBS and HLS) together form conservative (CS) and the middle alone (HB-HL) forms the non-conservative substitutions (NCS). Diversity or divergence for residue Q is highlighted (Horizontal and vertical Grey strips, Figure 1) to show frequencies for all 19 hetero-pairs (off-diagonal) along with its types. Total diversity for each of 20 residues is presented in the last row.

![Figure 1](image_url)

Figure: 1 190 SHP types (upper-half of diagonal) and observed frequencies (lower-half of diagonal) are shown. Substitution-homo-pair frequencies (i.e. 20) are at the diagonal position. Both these types and their frequencies divided into three categories: a] HB-HB category: total 36 (upper dark shade), b] HL-HL category: total 55 (lower white shade) and HB-HL category: 99 in number (middle gray shade region). Residue Q is shown by gray-strip for explanation of the calculation of diversity of a given hetero-pair.
Class specific diversities (due to acidic i.e. AcD, basic i.e. BaD, hydrophobic i.e. HLD, hydrophilic i.e. HLD and polar PoD) are computed by the use of equation 7.

**TABLE 2: BLOCK specific quantitative parameters for MDRs, DHPs and MCRs as obtained by APBEST analysis.**

<table>
<thead>
<tr>
<th>BLOCK name</th>
<th>H</th>
<th>S</th>
<th>MDR1</th>
<th>MDR2</th>
<th>MDR3</th>
<th>DHP1</th>
<th>DHP2</th>
<th>DHP3</th>
<th>MCR1</th>
<th>MCR2</th>
<th>MCR3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histone H2A</td>
<td>48</td>
<td>5</td>
<td>K(18.7)</td>
<td>V(18.3)</td>
<td>I(18.2)</td>
<td>IV(10.5)</td>
<td>KR(7.9)</td>
<td>LI(4.9)</td>
<td>L(15.5)</td>
<td>A(14.1)</td>
<td>G(10.3)</td>
</tr>
<tr>
<td>Histone H2B</td>
<td>48</td>
<td>6</td>
<td>S(30.8)</td>
<td>K(26.2)</td>
<td>R(22.0)</td>
<td>KR(13.8)</td>
<td>IV(9.1)</td>
<td>SA(7.8)</td>
<td>K(11.5)</td>
<td>S(9.6)</td>
<td>A(8.9)</td>
</tr>
<tr>
<td>GDH</td>
<td>48</td>
<td>13</td>
<td>I(16.9)</td>
<td>V(16.1)</td>
<td>A(15.8)</td>
<td>IV(5.0)</td>
<td>Li(3.6)</td>
<td>AG(2.4)</td>
<td>G(13.4)</td>
<td>A(9.9)</td>
<td>V(7.8)</td>
</tr>
<tr>
<td>Alpha Crystalline B</td>
<td>6</td>
<td>52</td>
<td>L(20.6)</td>
<td>S(19.2)</td>
<td>V(16.1)</td>
<td>IV(4.7)</td>
<td>ED(4.6)</td>
<td>KR(4.4)</td>
<td>E(12.9)</td>
<td>L(10.0)</td>
<td>R(8.8)</td>
</tr>
<tr>
<td>Troponin C</td>
<td>75</td>
<td>3</td>
<td>E(26.2)</td>
<td>K(21.0)</td>
<td>D(17.1)</td>
<td>ED(6.0)</td>
<td>KE(4.6)</td>
<td>KR(3.2)</td>
<td>E(16.6)</td>
<td>L(13.0)</td>
<td>D(12.5)</td>
</tr>
<tr>
<td>Plant ferredoxin</td>
<td>20</td>
<td>33</td>
<td>E(24.2)</td>
<td>V(22.7)</td>
<td>T(19.5)</td>
<td>ED(8.2)</td>
<td>IV(6.2)</td>
<td>TS(4.4)</td>
<td>D(13.0)</td>
<td>L(9.1)</td>
<td>G(8.5)</td>
</tr>
<tr>
<td>Cytochrome C</td>
<td>41</td>
<td>1</td>
<td>A(24.5)</td>
<td>L(24.3)</td>
<td>K(23.2)</td>
<td>LV(7.5)</td>
<td>Li(5.4)</td>
<td>TA(5.0)</td>
<td>G(16.5)</td>
<td>K(14.0)</td>
<td>L(12.9)</td>
</tr>
<tr>
<td>Corticotrophin</td>
<td>-</td>
<td>-</td>
<td>A(30.6)</td>
<td>V(24.1)</td>
<td>S(22.4)</td>
<td>SA(19.5)</td>
<td>YF(13.7)</td>
<td>ED(13.0)</td>
<td>P(12.8)</td>
<td>E(11.6)</td>
<td>K(11.2)</td>
</tr>
<tr>
<td>TPI</td>
<td>34</td>
<td>16</td>
<td>A(25.4)</td>
<td>V(19.2)</td>
<td>E(16.0)</td>
<td>IV(5.8)</td>
<td>Li(3.8)</td>
<td>EA(3.1)</td>
<td>A(13.5)</td>
<td>G(12.0)</td>
<td>E(10.6)</td>
</tr>
<tr>
<td>LDH</td>
<td>46</td>
<td>19</td>
<td>A(20.7)</td>
<td>V(19.3)</td>
<td>L(17.0)</td>
<td>IV(5.7)</td>
<td>Li(3.5)</td>
<td>LV(3.1)</td>
<td>G(12.9)</td>
<td>V(11.2)</td>
<td>L(9.5)</td>
</tr>
<tr>
<td>Plastocyanin</td>
<td>15</td>
<td>42</td>
<td>S(21.3)</td>
<td>V(21.2)</td>
<td>A(19.5)</td>
<td>IV(7.7)</td>
<td>SA(4.4)</td>
<td>TS(4.1)</td>
<td>G(16.3)</td>
<td>V(10.6)</td>
<td>N(7.6)</td>
</tr>
<tr>
<td>Insulin</td>
<td>66</td>
<td>9</td>
<td>L(25.8)</td>
<td>A(21.2)</td>
<td>E(16.1)</td>
<td>VA(5.1)</td>
<td>ED(4.7)</td>
<td>QL(3.2)</td>
<td>L(19.9)</td>
<td>C(11.2)</td>
<td>G(7.8)</td>
</tr>
<tr>
<td>Melanotropin beta</td>
<td>-</td>
<td>-</td>
<td>R(30.6)</td>
<td>P(23.0)</td>
<td>G(21.6)</td>
<td>KR(11.2)</td>
<td>RP(10.2)</td>
<td>DG(7.1)</td>
<td>P(15.1)</td>
<td>G(13.1)</td>
<td>K(9.3)</td>
</tr>
<tr>
<td>Trypsin</td>
<td>10</td>
<td>34</td>
<td>S(20.7)</td>
<td>A(19.8)</td>
<td>V(17.8)</td>
<td>IV(4.8)</td>
<td>SA(3.4)</td>
<td>LV(3.2)</td>
<td>G(18.1)</td>
<td>V(9.2)</td>
<td>L(7.8)</td>
</tr>
<tr>
<td>Acid proteases</td>
<td>9</td>
<td>48</td>
<td>S(19.7)</td>
<td>V(17.1)</td>
<td>L(17.1)</td>
<td>IV(3.5)</td>
<td>TS(3.4)</td>
<td>Li(3.2)</td>
<td>G(19.6)</td>
<td>D(11.1)</td>
<td>L(9.1)</td>
</tr>
<tr>
<td>Amyloid A</td>
<td>77</td>
<td>0</td>
<td>A(18.8)</td>
<td>K(16.1)</td>
<td>R(15.7)</td>
<td>SA(4.3)</td>
<td>VA(3.9)</td>
<td>YF(3.5)</td>
<td>A(16.3)</td>
<td>G(13.8)</td>
<td>R(9.5)</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>79</td>
<td>0</td>
<td>A(23.5)</td>
<td>K(20.1)</td>
<td>L(18.1)</td>
<td>KA(4.5)</td>
<td>Li(4.0)</td>
<td>AG(3.1)</td>
<td>L(14.3)</td>
<td>K(14.0)</td>
<td>A(11.5)</td>
</tr>
<tr>
<td>Hemoglobin alpha</td>
<td>78</td>
<td>0</td>
<td>A(29.3)</td>
<td>L(17.6)</td>
<td>S(17.6)</td>
<td>SA(5.5)</td>
<td>AG(5.5)</td>
<td>VA(3.6)</td>
<td>L(13.2)</td>
<td>A(10.8)</td>
<td>V(8.6)</td>
</tr>
<tr>
<td>Carbonic anhydrase</td>
<td>15</td>
<td>25</td>
<td>L(19.0)</td>
<td>V(16.4)</td>
<td>S(15.9)</td>
<td>LV(3.0)</td>
<td>PL(2.7)</td>
<td>IV(2.5)</td>
<td>H(10.3)</td>
<td>V(10.0)</td>
<td>G(9.5)</td>
</tr>
<tr>
<td>Lutropin alpha</td>
<td>4</td>
<td>51</td>
<td>K(21.6)</td>
<td>V(20.2)</td>
<td>A(19.9)</td>
<td>KR(11.4)</td>
<td>IV(8.8)</td>
<td>TM(6.0)</td>
<td>C(12.7)</td>
<td>K(12.4)</td>
<td>T(11.3)</td>
</tr>
<tr>
<td>Phospholipase A2</td>
<td>43</td>
<td>6</td>
<td>K(18.3)</td>
<td>L(14.0)</td>
<td>V(14.0)</td>
<td>IV(4.0)</td>
<td>KR(3.2)</td>
<td>FL(2.5)</td>
<td>C(26.1)</td>
<td>G(12.3)</td>
<td>Y(11.5)</td>
</tr>
<tr>
<td>Haptoglobin alpha</td>
<td>9</td>
<td>29</td>
<td>S(22.3)</td>
<td>N(21.1)</td>
<td>E(15.4)</td>
<td>NH(7.0)</td>
<td>TL(6.4)</td>
<td>KR(6.3)</td>
<td>Y(10.4)</td>
<td>D(9.4)</td>
<td>N(9.1)</td>
</tr>
<tr>
<td>Lipotropin beta</td>
<td>-</td>
<td>-</td>
<td>E(24.4)</td>
<td>S(20.6)</td>
<td>A(19.9)</td>
<td>KR(10.2)</td>
<td>ED(7.1)</td>
<td>YH(6.7)</td>
<td>K(16.1)</td>
<td>G(9.2)</td>
<td>P(7.3)</td>
</tr>
<tr>
<td>Ribonuclease</td>
<td>23</td>
<td>39</td>
<td>S(25.4)</td>
<td>R(16.9)</td>
<td>T(16.4)</td>
<td>TS(5.3)</td>
<td>IV(4.7)</td>
<td>SP(4.4)</td>
<td>C(15.7)</td>
<td>V(9.3)</td>
<td>T(8.9)</td>
</tr>
<tr>
<td>Somatotropin</td>
<td>45</td>
<td>0</td>
<td>S(22.0)</td>
<td>L(18.5)</td>
<td>E(16.3)</td>
<td>ED(3.4)</td>
<td>SA(3.4)</td>
<td>LV(2.9)</td>
<td>L(17.2)</td>
<td>K(8.7)</td>
<td>S(8.3)</td>
</tr>
<tr>
<td>Complement C3a</td>
<td>1</td>
<td>1</td>
<td>K(27.0)</td>
<td>A(20.1)</td>
<td>R(18.6)</td>
<td>KR(9.0)</td>
<td>YF(5.7)</td>
<td>ED(4.6)</td>
<td>K(12.3)</td>
<td>R(12.1)</td>
<td>C(10.9)</td>
</tr>
<tr>
<td>IgL C region</td>
<td>3</td>
<td>22</td>
<td>V(19.3)</td>
<td>E(19.1)</td>
<td>T(18.9)</td>
<td>LV(4.2)</td>
<td>QE(3.3)</td>
<td>DG(2.7)</td>
<td>T(10.9)</td>
<td>P(8.1)</td>
<td>V(7.8)</td>
</tr>
<tr>
<td>Lutropin beta</td>
<td>-</td>
<td>-</td>
<td>A(22.6)</td>
<td>P(21.9)</td>
<td>R(20.3)</td>
<td>RP(5.4)</td>
<td>TS(4.9)</td>
<td>VA(3.5)</td>
<td>P(18.3)</td>
<td>C(17.8)</td>
<td>L(9.2)</td>
</tr>
<tr>
<td>IgG C region</td>
<td>9</td>
<td>45</td>
<td>S(21.1)</td>
<td>T(15.6)</td>
<td>K(14.8)</td>
<td>YF(3.4)</td>
<td>SK(2.4)</td>
<td>SN(2.2)</td>
<td>P(13.0)</td>
<td>V(8.5)</td>
<td>S(8.4)</td>
</tr>
<tr>
<td>Kappa casein</td>
<td>-</td>
<td>-</td>
<td>V(26.4)</td>
<td>P(23.2)</td>
<td>T(23.0)</td>
<td>VA(9.6)</td>
<td>SP(5.9)</td>
<td>PL(5.6)</td>
<td>P(15.6)</td>
<td>T(13.4)</td>
<td>A(9.7)</td>
</tr>
</tbody>
</table>

S*: Secondary structure information is extracted using one of the representatives of PDB structure. MDR: Maximally Diverse Residue; DHP: Dominant Hetero-Pair; MCR: Maximally Conserved Residue; LDH: Lactate dehydrogenase; GDH: Glutamate dehydrogenase; TPI: Triose Phosphate
Several points are noteworthy from Figure 1 and Tables (1, 2 and 3). First, type specific normalized hetero-pair frequencies are seen to be non-identical for BLOCKs (Figure 1). Three categories of substitutions (as observed in Figure 1 as HB-HB, HL-HL and HB-HL; or as HBS, HLS and NCS in Table 1 column 3-5) varies among each other of which NCS is seen to be always lower than the sum of HBS and HLS (conservative substitution) i.e. CS (Figure 2: A4). Further the usage of hetero-pair frequencies (E or DV) for different BLOCKs are seen to be different (Table 1: column 7). Second, dominantly used hetero-pairs (DHP) are mostly seen to be conservative type with normalize frequency ranging between 2.5 to 11.0 (Table 2: column 7-9). Third, residue (Figure 1 and Table 2: column 6-7) and class-specific (Table 1: column 9-12) diversities (RD and CD respectively) show BLOCK specific variation. Fourth, unlike minimum diversity class (Item B parameter: 7 of output file), frequency and types of maximally diverse residues (MDRs) i.e. MDR1, MDR2 and MDR3 (Table 2: column 4-6) are seen to be more versatile with frequencies 26 - 31, 15 - 24 and 14 - 23 and types KRESAVLI, KRESTNAVLPS and KRESTAVLIG for MDR1, MDR2 and MDR3 respectively. Fifth, maximally conserved residues (MCRs) i.e. MCR1, MCR2 and MCR3 are also versatile as MDRs. Most conserved residues are observed as LKGP (frequency range: 11-26), LKGV (frequency range: 8-14) and LGAV (frequency range: 7-12) for MCR1, MCR2 and MCR3 respectively. Finally, ratio parameters (R, E and N) also show BLOCK specific variation (Table 1 column 6-8).

As shown above individual heteropairs (HP) frequency vary from one another for a given BLOCK (Figure 1) and among BLOCKs its normalized usage (E) also vary (Table 1: column 7), we have presented hetero-pair frequency against observed probability in Figure 2 (plot A1 and A2). It is seen in the figure that overall distribution pattern and region specific details of observed hetero-pair types vary greatly for BLOCKs. At low probability range, observed hetero-pair frequency is very high and non-selective. As we move towards higher probability ranges the frequency and corresponding type of hetero-pair become narrower and selective. For example, at highest probability range, the sole and lone observed hetero-pair is LV and ED for plot A1 and A2 respectively (Figure 2). It is worth noting here that both are conservative type with former being hydrophobic and the later hydrophilic.
In evolution functionally similar sequences (homologous/Orthologous sequences forming BLOCK) are the result of substitution in the parental one. While conservation of sequence positions as parental one is the prerequisite for functionality, evolution demands substitutions (i.e. formation of HPs) at homologous positions. Lethal substitutions cause malfunctioning of proteins (Ng and Henikoff, 2006; Chun and Fay, 2009). What are the lower and upper limits of usage of HPs? To check this, we have plotted $E$ for different BLOCKs (Figure 2: Plot A6). Theoretically $E$ varies between 0 to 1 (Equation 4). The plot shows that the lower and upper limit of $E$ are 0.13 and 0.7 respectively in studied BLOCKs. Interestingly kappa-casein, seen to possess highest divergence rate (Table 1: column 6) compared to other protein-BLOCKs, shows lower $E$ value (0.33). Similar is the case for Somatotropin-BLOCK (E 0.48). Thus the parameter $E$ is largely non-linearly correlated with divergence rate.

Is there any BLOCK specific parameter that would correlate divergence rate? In figure 2 (A3), $R$ parameter is plotted and fitted against the divergence rates (Marks, 1988; Dayhoff and Schwartz, 1978). Notably, $R$ is the ratio of non-conservative to conservative substitution (Equation 5). The plot shows that $R$ is positively and linearly correlated with divergence rate having correlation coefficient of 0.93.

Such strong correlation of $R$ with divergence rate indicates that it could be a valuable parameter in the study of substitution related molecular evolution of protein-BLOCK.

Note that $R$ is a ratio of non-conservative (NCS) to conservative (CS) substitutions. How CS and NCS are related to each other? In other words, does there exist any general tendency of these parameters for BLOCKs? The plot A4 (Figure 2) shows data and fitted curve for NCS and CS. It is clear from the plot that they are inversely correlated to each other whose BLOCK (an Orthologous/homologous set) specific outcome depends on the divergence rate. For all out BLOCKs, NCS is seen to be much lower than CS. However, extrapolation of fitted lines shows that they meet each other (i.e. NCS equals to CS) around a divergence rate of 45 unit beyond which NCS is lower than CS. i.e. NCS>CS. Many factors might affect BLOCK-positional divergence or diversity. Some of these factors are positional entropy (Shannon, 1948), position specific physicochemical characteristics of BLOCKs.
Figure: 2 Plot of observed hetero-pair frequency vs. probability range (A1, A2), $R$ vs. $DV$ (A3), NCS and CS vs. DV (A4), $E$ vs. DV (A6), INV and HB-HL vs. DV (A7), HB-HB and HB-HL vs. DV (A8), HL-HL and HB-HL vs. DV (A9). A5 is segment of crystal structure of Somatotropin-receptor interaction. Maximally conserved residue i.e. $MCR$ (Lys) is shown in red. A11 is the segment of crystal structure of active site of Trypsin. Active site and substrate specificity site residues with highlight on $MCR$ (Gly). A13 shows segment of crystal structure of plant
ferredoxin with Glu + Asp residue on the surface in red (A12) and MCR i.e. Leu (blue), Gly (red dot) in A13. A10 of Figure 2 is the plot of MDR1, MDR2 and MCR1, MCR2 of kappa casein.

APBEST computes above details of which few are listed in table 3 and plotted in Figure 2, A7, A8 and A9 for comparison of HB-HL-line vs. INV-line, HB-HL line vs. HB-Hb-line and HB-HL line vs. HL-HL-line respectively. It is observed that in all these plots HB-HL homologous positions exceed than that of INV, HB-HB and HL-HL. Several points are noteworthy from the table. a) Majority of sequence positions in BLOCKs contains mixed type (HB+HL) amino acid. Thus, HB+HL-type dominate over others such as HB-HB, PU+PC etc. b). All but myoglobin and carbonic anhydrase C contains invariant lines with highest for corticotrophin BLOCK. Invariant line doesn’t evolve over time and are largely involved in the preservation of structure and or function of BLOCK as parental one. c) Shannon entropy is the measure of positional conservation. A value ≤1.0 indicate highly conserved positions. Details of conserved positions are shown in table 3 (column 14).

Highest and lowest conservation is seen in the case of haptoglobin α (90.9%) and lowest carbonic anhydrase C (2.3%) respectively.

To check positional distribution of maximally (diverse and conserve) residues i.e. MDRs and MCRs, we have plotted APBEST extracted data (Item C, D, E of output) for both diversity (MDR1 and MDR2) and conservation (MCR1 and MCR2) in Figure 2, A10. The plot shows amplitudes of MCRs are greater than MDRs and are distributed differentially along BLOCK positions. Divergence occupies less position than conservation.

Table: 3 Positional analysis of BLOCKs for invariant line (only one type of residue), conserved position (Shannon entropy ≤1.0) and type of amino acid classes (such as HB, Ac, Bs, PC, ST, HB+HL and PU+PC). Normalized values (in %) are presented for comparison among BLOCKs

<table>
<thead>
<tr>
<th>Protein name</th>
<th>INV-line</th>
<th>HB-HB</th>
<th>HB-HL</th>
<th>PU</th>
<th>Ac</th>
<th>Bs</th>
<th>MC</th>
<th>PG</th>
<th>ST</th>
<th>NQ</th>
<th>PU+PC</th>
<th>CONV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histone H2A</td>
<td>4.4</td>
<td>18.7</td>
<td>67.1</td>
<td>2.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.1</td>
<td>-</td>
<td>-</td>
<td>5.5</td>
<td>69.2</td>
</tr>
<tr>
<td>Histone H2B</td>
<td>10.3</td>
<td>15.5</td>
<td>53.6</td>
<td>-</td>
<td>1</td>
<td>2.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>15.5</td>
<td>77.3</td>
</tr>
<tr>
<td>GDH</td>
<td>10.4</td>
<td>13.9</td>
<td>67.1</td>
<td>0.5</td>
<td>0.3</td>
<td>-</td>
<td>0.3</td>
<td>0.3</td>
<td>-</td>
<td>-</td>
<td>7.8</td>
<td>31.8</td>
</tr>
<tr>
<td>α Crystalline</td>
<td>B</td>
<td>3.9</td>
<td>14.6</td>
<td>68.9</td>
<td>-</td>
<td>1.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10.7</td>
<td>66</td>
</tr>
<tr>
<td>Troponin C</td>
<td>3.3</td>
<td>12.5</td>
<td>58.3</td>
<td>0.8</td>
<td>3.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>20.8</td>
<td>20.8</td>
</tr>
<tr>
<td>Plant fd</td>
<td>13.3</td>
<td>15.6</td>
<td>58.9</td>
<td>1.1</td>
<td>3.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.1</td>
<td>-</td>
<td>7.8</td>
<td>62.2</td>
</tr>
<tr>
<td>Cytochrome C</td>
<td>5.7</td>
<td>14.3</td>
<td>77.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.9</td>
<td>-</td>
<td>2.9</td>
<td>42.9</td>
</tr>
<tr>
<td>Corticotrophin</td>
<td>52.8</td>
<td>11.1</td>
<td>25</td>
<td>2.8</td>
<td>2.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.8</td>
<td>88.9</td>
</tr>
<tr>
<td>TPI</td>
<td>6.3</td>
<td>8.7</td>
<td>80.3</td>
<td>-</td>
<td>0.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.9</td>
<td>27.6</td>
</tr>
<tr>
<td>LDH</td>
<td>1.3</td>
<td>6.7</td>
<td>89.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.5</td>
<td>14.6</td>
</tr>
<tr>
<td>Plastocyanin</td>
<td>18.8</td>
<td>17.6</td>
<td>58.8</td>
<td>-</td>
<td>1.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.5</td>
<td>52.9</td>
</tr>
<tr>
<td>Insulin</td>
<td>11.5</td>
<td>13.8</td>
<td>63.2</td>
<td>2.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8</td>
<td>48.3</td>
</tr>
<tr>
<td>Melanotropin</td>
<td>β</td>
<td>23.5</td>
<td>5.9</td>
<td>47.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>11.8</td>
<td>64.7</td>
</tr>
<tr>
<td>Trypsin Acid proteases</td>
<td>12.1</td>
<td>10.6</td>
<td>73.7</td>
<td>0.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.5</td>
<td>-</td>
<td>3</td>
<td>33.8</td>
</tr>
<tr>
<td>Amyloid A</td>
<td>35.4</td>
<td>6.1</td>
<td>38.4</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>18.2</td>
<td>65.7</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>-</td>
<td>15.3</td>
<td>80.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.2</td>
<td>23.7</td>
</tr>
<tr>
<td>Hemoglobin α</td>
<td>7</td>
<td>14</td>
<td>72.9</td>
<td>0.8</td>
<td>-</td>
<td>0.8</td>
<td>-</td>
<td>0.8</td>
<td>-</td>
<td>-</td>
<td>3.9</td>
<td>49.6</td>
</tr>
<tr>
<td>Carbonic anhydrase C</td>
<td>-</td>
<td>4.5</td>
<td>94.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.1</td>
<td>2.3</td>
</tr>
<tr>
<td>Lutropin α</td>
<td>26.8</td>
<td>12.2</td>
<td>46.3</td>
<td>1.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9.8</td>
<td>84.1</td>
</tr>
<tr>
<td>Phospholipase A2</td>
<td>-</td>
<td>4.7</td>
<td>84.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.1</td>
<td>42.2</td>
</tr>
<tr>
<td>Haptoglobin α</td>
<td>38.2</td>
<td>9.1</td>
<td>32.7</td>
<td>1.8</td>
<td>1.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.8</td>
<td>-</td>
<td>12.7</td>
<td>90.9</td>
</tr>
<tr>
<td>Lipotropin β</td>
<td>44.1</td>
<td>11.9</td>
<td>25.4</td>
<td>3.4</td>
<td>1.7</td>
<td>-</td>
<td>-</td>
<td>1.7</td>
<td>3.4</td>
<td>-</td>
<td>6.8</td>
<td>83.1</td>
</tr>
<tr>
<td>Ribonuclease A</td>
<td>17.5</td>
<td>11.2</td>
<td>68.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.5</td>
<td>47.5</td>
</tr>
<tr>
<td>Somatotropin</td>
<td>12.7</td>
<td>8.2</td>
<td>65.8</td>
<td>1.3</td>
<td>1.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.6</td>
<td>-</td>
<td>8.2</td>
<td>34.8</td>
</tr>
<tr>
<td>Compt. C3a</td>
<td>9.5</td>
<td>2.4</td>
<td>85.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.4</td>
<td>46.6</td>
</tr>
<tr>
<td>Ig λ chain C</td>
<td>5.7</td>
<td>5.7</td>
<td>82.9</td>
<td>1.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.3</td>
<td>25.7</td>
</tr>
<tr>
<td>Lutropin β</td>
<td>28.2</td>
<td>13.6</td>
<td>52.4</td>
<td>2.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.9</td>
<td>-</td>
<td>2.9</td>
<td>75.3</td>
</tr>
<tr>
<td>Ig γ chain C</td>
<td>10</td>
<td>12.7</td>
<td>62.7</td>
<td>1.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>12.1</td>
<td>40.4</td>
</tr>
<tr>
<td>Kappa casein</td>
<td>1</td>
<td>15.2</td>
<td>76.8</td>
<td>0.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5.1</td>
<td>36.3</td>
</tr>
</tbody>
</table>

**INV** Invariant position; **CONV** Conserved position; **HB** position contains only hydrophobic amino acids; Similarly **Ac** acidic, **Bs** basic, **PC** Polar charge, **ST** serine plus threonine; **HB+HL** position contains hydrophobic and hydrophilic amino acids; similarly **PU+PC** polar uncharged and polar charged; - absent.

**Discussions**

Functionally similar proteins (i.e. Orthologous protein BLOCK) evolve to create amino acid divergence at locus specific positions (Gaucher et. al, 2001; Wang and Gu, 2001). Different BLOCKs evolve at different rate but for a given family (or BLOCK) it is constant (Dickerson 1971; Dayhoff and Schwartz, 1978). The divergence although alters primary sequence, secondary structural sequence largely remains unaltered, the one that
acts as determinant for functional form or state of homologous protein family. The fact that not all hetero-pairs (total 190) contribute equally for evolution of a BLOCK, characterizing their differential role in relation to its function is of central importance in the study evolution. It also remains a challenge to assess acquired positional mutations in primary sequence and maintenance of secondary structural sequence for functional state. Our study is a first time attempt to gain insight into the mechanism of substitution based on observed hetero-pairs and its positional diversity in relation to protein’s secondary structure and function.

**Hetero-Pairs (HP) and Dominantly used Hetero-Pair (DHP)**

BLOSUM series of fundamental matrices utilize observed hetero-pairs for computation of odd-score (Henikoff and Henikoff, 1992). However, its use in relation to BLOCK (protein family) evolution is rare. In the course of evolution observed heteropairs (HPs), the source of diversity in different sites in BLOCKs, emerge in expense of homo-pairs from ancestral proteins. When identical amino acid residues are present at a given homologous position of a pair of homologous proteins, that position is said to form a homo-pair. A hetero-pair on the other hand involve non-identical residues. A total of 20 homo-pairs (diagonal) and 190 hetero-pairs (off-diagonal) participates in this process. The later may be subdivided into three groups: hydrophobic-hydrophobic (i.e. HB-HB i.e. **HBS**), hydrophilic to hydrophilic (HL-HL i.e. **HLS**) and hydrophobic to hydrophilic (HB-HL i.e. **NCS**) based on the type of substitution. HB-HB and HL-HL are considered conservative (**CS**) and HB-HL is non-conservation (**NCS**) type. In contrast to the later, the former may preserve secondary structures and thus functionality of proteins (Ng and Henikoff, 2006).

In our study with thirty functionally distinct protein BLOCKs, BLOCKs parameters (Item B in output) show BLOCK specific variations. Largely conservative (~88%) nature of dominantly used hetero-pair (parameter **DHP**) might indicate its role in the preservation of secondary structural sequence in BLOCK. While non-conservative substitutions act as a means to modulate proteins function (Shen, Jones and Mohrenweiser, 1998) needed in its evolution, it could also be deleterious for cellular processes (Ng and Henikoff, 2006; Chun and Fay, 2009). Conservative nature of dominantly used hetero-pair for all
BLOCKs support protein evolution under structural and or functional constraints (Tóth-Petróczy and Tewfik, 2011), (see below).

**Maximally diverse residues (MDRs) in BLOCK evolution**

Understanding evolution using intrinsic codes of protein has got new horizon in recent times. Proteins evolve to overcome structural and functional constraints in a given environment (Tóth-Petróczy and Tewfik, 2011; Chakrabarti and panchenko, 2010; Dessailly et. al, 2010). Divergence or diversity is the measure of the evolution. Although all amino acids residues contribute to the overall diversity for a given BLOCK (Table 1) their contribution vary over a wide range of frequencies.

BLOCK specific variation of residue diversity \(RD\) (Table 1) indicates that maximally diverse residue \(MDR\) and dominant heteropair \(DHP\) may not be unique but varies among functionally diverse BLOCKs. To gain detailed insight into these parameters (\(MDR\) and \(DHP\)) in relation to structure and function of BLOCKs, we considered three values of each such as MDR1, MDR2 and MDR3 for MDR and DHP1, DHP2 and DHP3 for \(DHP\). For the purpose, three maximally conserved residues (\(MCRs\)) as MCR1, MCR2 and MCR3 were also taken into consideration (Table 2). At this point it is worth raising the question as to which condition makes a residue maximally diverse or maximally conserved. If in a BLOCK, positional count of the residue exceeds over others and that position also possesses maximum heterogeneity (i.e. maximum type of residues) then that residue cause maximum positional diversity. Maximal conservation in turn is achieved when a position is like \(MDR\) but devoid of heterogeneity. Further is it also important to understand condition for dominant hetero-pair (\(DHP\)). However, when the above condition for \(MDR\) for a residue is met in a position and that position also possesses a second residue whose count is greater than others and also greater than one but less than \(MDR\), then \(MDR\) and the second residue together form \(DHP\). In other words, from the knowledge of \(DHP\) and \(MDR\), it might be possible to understand principle strategy of amino acid exchange and their role in secondary structure. \(\alpha\)-helix, \(\beta\)-sheet and coil contribution of \(MDR/MCR\) residues could be understood from their propensity values in that E>A>L>M>Q>K>R>H have \(\alpha\)-helical, V>I>Y>C>W>F>T have \(\beta\)-strand and G, N, P, S, D have coil propensity. It is also noted that L>M (from helix list) and I>W>F (from
strand list) are somewhat amphipathic in nature. It means although L and M have higher propensity for α-helix, they also possess β-strand propensity. Similarly I, W and F possess helical propensity (Williams et al, 1987) Structural significances of MDRs were understood in relation to DHPs and MCRs for all 30 families in our study. It is apparent that:

1. Nature of MDRs (along with MCRs) help to interpret overall secondary structural contents of BLOCKs.

2. Interestingly, of three, at least one (mostly two) MDRs/MCRs are hydrophobic in nature indicated their contribution to core formation.

3. It may also points that secondary structure formation is driven by burial and hence closeness of peptide segments harboring hydrophobic MDRs/MCRs as isolated solvent exposed strand or helix are known to be less stable.

4. Conservative nature of amino acid exchanges between MDRs and DHPs indicate restoration of physicochemical properties of residue in question (mentioned above) and thus overall structural properties of proteins.

5. Most interestingly, MCRs not identical as MDRs could be identified as distinct MCRs. These distinct MCRs are interpreted to have impactive functional and or structural role (see below).

In light of the above points and BLOCKs specific data on MDRs, DHPs and MCRs, following examples are noteworthy.

Somatotropin is α-helix rich protein with helix content 45% (1HGU). Its maximally diverse residues (MDRs): L (Leu) and E (Glu) along with MCR of L (Leu) seems to contribute to its helical content. Again MDR1 which is S (Ser), secondary structurally neutral, is in pool with MCR3 i.e. S (Ser) and is in exchange with DHP2 (SA i.e. Ser to Ala). A (Ala) possess high helical propensity (Williams et al, 1987) contributes to helical content of the protein. Other DHPs (such as LV and ED) seems to restore physicochemical behavior of MDRs. Most interestingly MCR2 which is K (Lys) is not in pool with MDRs is identified as distinct MCR (see above). What could be the role of this highly conserved Lysine? Detailed investigation of the structural-complex of Somatotropin with its receptor shows that the interaction site of Somatotropin with its
receptors are lined up with many Lysine residues (Figure 2, A5) whose contribution in this aspects is well established (Wells, 1996). Interestingly, like Somatotropin, other all alpha proteins such as Hemoglobin, Myoglobin possess **MDRs** (A, K and L), **DHPs** (AK, LI, VA and SA) and **MCRs** (L, K and A etc) that have strong helical propensity (Table 2). Trypsin is a protease that contains both α-helix and β-sheet (Table 2) with later has higher contribution. Its **MDRs**: A (Ala), V (Val) and MCRs: V (Val) and L (Leu) contribute to both these secondary structures. At this point it is noteworthy that although Luecine has higher propensity for α-helix, it also has reasonable propensity for β-strand (Williams et. al, 1987). Further critical investigations of crystal structures support the above contention. Interestingly, amino acid exchanges such as SA lower helical content with enhanced hydrophobicity and thus may contribute to burial of such peptide segments. The amino acid exchanges VI and VL (which is in pool with **MCRs**: V and L) largely contribute to the restoration of β-content, their burial and thus stability. MCRI which is Glycine is not in the pool of amino acid exchange and thus seems a distinct residue (see point 5, above). The importance of Gly is seen in the active site of the enzyme (Figure 2: A11). Active site as well as the substrate specificity site contain many conserve Gly residues and thereby providing required flexibility in the site and catalysis.

Plant ferredoxin is a [2Fe -2S] chromophore containing protein that possesses β-grasp fold. It contains both β-sheet and α-helix with the earlier makes major contribution. **MDRs** V, T (Table 2) have strong β-propensity (Williams et. al, 1987) with earlier exchanges with Ile which also has high propensity for β (Table 2). At the same time plant ferredoxin possesses MDR1 (i.e. E) which has α-helical propensity. Interestingly, there are two MCRs (i.e. L and G) which are not in the pool with **MDRs** or **DHPs**. Are there any structural and or functional significance of these **MCRs**? Closer investigation of the chromophore region (1A70.pdb) shows that there are 8-9 Leucines and 5-6 Glycine (Figure 2: A13). The former contributes to the core formation of the chromophore region. It may also direct burial of peptidic segments involved in secondary structure. As far as Glycine is concerned, it may impart flexibility to the [2Fe-2S] region. [2Fe- 2S] center coordinates with distantly spaced four Cysteines residues that may require sufficient structural flexibility for the region.
Such highly conserved Gly (as MCRs) in Cytochrome C, Triose Phosphate Isomerase, Lactate Dehydrogenase, plastocyanin etc may contribute similar flexibility as discussed (above).

How these MDRs and MCRs distributed along BLOCK positions? It would be interesting to address the question as the former incorporate divergence due to sequence evolution and the later restore their parental forms.

As far as distribution of MDRs and MCRs is concerned closer look shows that such unique positions are limited in sequence BLOCKs. For Kappa casein BLOCK distribution of MDR1 (V, blue), MDR2 (P, light green) and MCR1 (P, red), MCR2 (T, pink) are shown in Figure 2, A10. The fact that MCRs having higher amplitudes covers greater portion of width of sequence BLOCK than MDRs. Thus it could be said for example that diversity of the BLOCK of kappa-casein is contributed in major by Valine and proline (Table 2) and is achieved by substitution at few selected positions in BLOCK (Blue for V and Red for P peaks; Figure 2, A10). It is, therefore, reasonable to think that to achieve a given divergence rate (i.e. 33 percent/100 mYr for kappa casein) in evolution such limited strategic locus specific substitutions would be sufficient. How then other positions contribute to protein’s structure and or functionality when limited BLOCK positions would be sufficient for divergence rate? As far as remaining BLOCK positions are concerned, intuitively it seems, they remain open for second kinds of substitution mechanism: such as maintaining invariant positions (that contains one type of residue), conservative substitutions (that retains functionality) (light green for MCR1:P and pink for MCR2:T), positional entropic conservation (Shannon Entropy <1.0) etc. Overall, it seems there exist two distinct parts in sequence BLOCK: one contributes to divergence rate and other maintains conservation. While the former seems to be a prerequisite for neutralization of environmental adaptation pressure the later largely retains protein structure and or function as ancestral one.

*Physical meaning of E and R*

Hetero-pairs are the result of substitution of homo-pairs whose usage limit (E) varies between 0 (no hetero-pair, all are homo-pairs) and 1 (no homo-pair and all are hetero-pairs) for a BLOCK. In our study the observed range of E between 0.3 and 0.7 (Figure 2:
A5) indicates that both homo and hetero pairs are essential for BLOCK evolution. Non-linear dependent of $E$ on divergence rate (Figure 2:A5) pointing to the strategy of preferential usage of hetero-pairs. In other words while some hetero-pairs are used at high frequency, others are either used at low frequency or not used at all.

Divergence in protein sequences allows versatility to functionality in wide range of earth’s atmospheres. Although both conservative (CS) and non-conservative (NCS) substitution (together constitute hetero-pairs) contributes to the divergence rate (Dickerson 1971; Dayhoff and Schwartz, 1978) their relative contributions remain to be understood. Our analyses show that NCS and CS are reciprocally dependent on divergence rate. This would mean that either of three situations could define the divergence rate of a BLOCK. They are a] $\text{NCS} < \text{CS}$ b] $\text{NCS} = \text{CS}$ and $\text{NCS} > \text{CS}$. Our analyses with 30 protein families show that they all satisfy condition a (Figure 2: A4). At around DV 45, NCS is equal to CS i.e. non-conservative and conservative are in equal balance (Figure 2, A4). At higher divergence rate (such as 80 for example) non-conservative substitution is far greater than conservative one. In our study with thirty protein BLOCKs the observation that $R$ is always less than unity indicates in evolution, conservative substitution exceeds the non-conservative one. The fact that the latter is the means to modulate protein function in evolution (Shen et. al, 1998) excessive accumulation may lead to deleterious effect (Ng and Henikoff, 2006; Chun and Fay, 2009). The divergence rate is directly related to $R$ (Figure 2: A3) and hence to the non-conservative substitution (Equation 5) lead us to hypothesized that it is the non-conservative heteropairs (HPs) that act as determinant of the divergence rate.

At this point it is worth mentioning that Kappa-casein BLOCK with highest divergence rate and highest $R$ value shows high positional conservation (64%; Shannon entropy≤1.0). This apparent contrast of high divergence rate and high conservation of kappa casein BLOCK could be resolved by the observation that non-conservative substitutions (determinant of divergence rate) occurs only at limited and unique BLOCK positions (Figure2: A10). Such limit might allow protein to use rest of the BLOCK positions for conservation to retain structures (Creamer et. al, 1998) and or function. Again, the above mentioned strategy which is seen in case of kappa-casein is also observed in other
BLOCKs. Thus, it seems to be a protein specific evolutionary strategy wherein divergence is relating to overcome structural and or functional constraints that utilize limited BLOCK positions and conservation for maintenance of sequence-structure as ancestor one use rest of BLOCK positions.

**Conclusion**

APBEST-based, a home-built fully automated procedure, analyses with 30 protein BLOCKs of known divergence rate shows that observed hetero-pairs type and their corresponding frequencies bears characteristic BLOCK specific evolutionary signatures as revealed in the distribution pattern, non-equivalent hetero-pair frequencies and ratio parameters such as $DHPs$, $E$ and $R$, $MDRs$, $MCRs$, $RD$ and $CD$. These parameters are used for understanding structural and or functional evolution of BLOCKs. $E$ is suitable for understanding usage limit of hetero-pairs and $R$ is directly related with the divergence rate. Non-conservative substitution is directly and conservative one is inversely related with the divergence rate. It is non-conservative but not conservative substitutions that act as determinant for the divergence rate. $DHP$, $MDRs$ along with $MCRs$ secondary structure content, core stability and functional significance of distinctly conserved residues. $MDRs$ also contribute to $CD$ parameter and divergence rate. The fact that it occurs only at unique and limited BLOCK positions, the divergence rate utilizes smaller part of the total width of BLOCK. It would means that a BLOCK with high conservation can still have high divergence rate. This strategy of limited yet unique use of positions for divergence rate is postulated for the purpose of incorporation of other important mechanisms of substitutions such as conservation.

**References**

- Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome research. 2001 May 1;11(5):863-74.
CHAPTER 2

An Insight into the Variability, Conservation and Substitutions of Protein BLOCKs

Summary

The knowledge of site-specific variability, conservation and substitutions is central in the study of homologous BLOCKs. Observed-hetero-pair (OHP) method has been efficient in estimating ensemble-positional-variability (epv) and the residues’ contribution. For the latter purpose, Shannon-entropy (SE) method is seen to be anomalous. OHP method further extracts site-specific details on maximally-variable-residues (MVRs), maximally-conserved-residues (MCRs) and non-conservative-substitutions (NCSs) to conservative-substitutions (CSs) ratio (R). The difference, of sites with R>1.0, between coiled and secondary structures is much lower (5%) than that of surface and core (23%) indicates the core has much higher selectivity to NCSs. 1/4th of the sites of MVRs have R>1.0. These sites are seen to be populated in helix, strand and core regions may indicate NCSs-induced structural tuning. Investigation of the candidate sequences harboring these sites for functions, sub-cellular locations and isoforms revealed that the acquisition of NCSs is induced by environmental stress. Unlike MVRs, almost all sites (95%) of MCRs have R≤1.0 of which very few are invariant-types pointing to the importance of CSs in site-specific conservation. Distinctly-conserved-residues (dMCRs) assist to the discovery of site-specific structural and/or functional constraints. Overall, the study seems to have general applications in site-specific characterization of BLOCKs.

Introduction

Site-specific codes in sequence in the presence of appropriate environmental conditions form functional state of proteins (Anfinsen CB, 1973; Chotia C, 1984; Liberles et. al, 2012; Bandyopadhyay AK, 2015). In this aspect sites of BLOCKs are studied in great details (Bowie, et. al., 2015). Different types and levels of substitutions are acquired at different sites of homologous sequence (Reidhaar Olson and Sauer, 1988) of which
invariant and conservatively substituted sites were found to be crucial for structure and function. Non-conservatively on the other hand were found to have little or no such roles (Bowie, et. al., 2015). Substitutions at binding site, active site, interaction site etc were shown to impose restrictions (Nelson and Sauer, 1986; Shenkin, et. al., 1991; Garnier and Levin, 1991, Bowie, et. al., 2015) and any alternative to these have devastating effects (Shortle D, 1989; Alber, et. al., 1987). Substitutions largely modulate local interactions and thereby change overall stability of proteins (Shortle D, 1989; Tzul, et. al., 2015). In orthologous set overall variability due to substitutions (Hellinga, et. al., 1992) is largely constant, although their position-specific details vary greatly (Guindon, et. al., 2004). Such positional variability in aligned orthologous BLOCK indicated enormous capacity of modulation of codes in its respective environment (Matthews, 1987). While conservative substitutions (CSs) (French and Robson, 1983) restore, the non-conservative ones (NCSs) modulate structure and function (Ng and Henikoff, 2001). The later could also be deleterious and thus majority of it are eliminated from selection (Ng and Henikoff, 2001). Stewart et. al. (1997) used SE method to show larger portion of Igs and TCRs are variable with little of it responsible for maintenance of structure and stability. In contrast high positional conservation was observed in highly diverse kappa-casein BLOCK. The fact that both level and type of substitutions along with site-specific residue’s contribution to variability and conservation play crucial role in protein’s adaptation, robust all-in-one method would be useful for the establishment of BLOCK specific mechanistic model of substitutions (Echave, et. al., 2016).

In this work, we compare results of simulation on designed BLOCKs using SE and OHP methods. The study also highlights novel strategy of site specific incorporation of maximal variability and conservation along with their implications.

**Materials and Methods**

**Observe hetero-pair (OHP) method**

Henikoff and Henikoff (1992) generated BLOSUM bit-score using observed and expected frequency of hetero-pairs in BLOCKs (Henikoff and Henikoff, 1992). There are 190 non-equivalent hetero-pair. For example DE and ED are equivalent. So, one of them is taken into consideration.
Let us consider a BLOCK of width $w$ with $m$ positions. Each position has $k$ type of residues (A, M, C, D and etc) with frequency $f_k$. Now variability for residue $X$ (where $X\neq k$) at a given position $m$ would be

$$OHP_X^m = \frac{1}{2} f_X \left( \sum_{k \neq X} f_k \right) \quad [1]$$

Similarly variability of a given residue $X$ for a BLOCK is the sum of $OHP_X^m$ for all positions i.e.

$$OHP_X = \sum_{m=1}^{w} OHP_X^m \quad [X = 20 \text{ residues}] \quad [2]$$

Out of 20 $OHP_X$, first three top-scores variability are identified as $MVR1$, $MVR2$ and $MVR3$ where $MVR1 > MVR2 > MVR3$. Physicochemical type, frequency and types of hetero-pairs formed by each of these MVR are also determined.

Total variability for the position $m$ for all residues ($X$s) is

$$OHP^m = \sum_{X} OHP_X^m \quad [3]$$

Variability for BLOCK is the sum of all positions and for all residues

$$OHP = \sum_{m=1}^{w} OHP^m \quad [4]$$

Similarly site-specific observed conservation of residue ($Z$) can be determined using homo-pair frequency of BLOCK as

$$OCR_Z^m = \frac{1}{2} \left( f_Z C_Z \right) \quad [5]$$

Similarly conservation of a given residue $Z$ for a BLOCK is the sum of $OCR_Z^m$ for all positions i.e.

$$OCR_Z = \sum_{m=1}^{w} OCR_Z^m \quad [Z = 20 \text{ residues}] \quad [6]$$

Out of 20 $OCR_Z$, first three residue-level conservation are identified as $MCR1$, $MCR2$ and $MCR3$ where $MCR1 > MCR2 > MCR3$. Physicochemical type, frequency and types of homo-pairs formed by each of these MCR are also determined.
Total conservation for the position \( m \) for all residues (\( Z_s \)) is

\[
OCR^m = \sum_{Z_s} OCR^m_{Z_s} \tag{7}
\]

Variability for BLOCK is the sum of all positions and for all residues

\[
OCR = \sum_{m=1}^{w} OCR^m \tag{8}
\]

Site-specific and BLOCK-specific non-conservative substitutions (NCSs) to conservative substitutions ratio (R)

There are 20 amino acids in natural proteins. They are divided into two classes: hydrophobic and hydrophilic. A, V, F, I, L, M, C, G, P are taken as hydrophobic and rest 11 are hydrophilic.

These residues are considered as \textit{hydrophobic} because first seven residues (i.e. \textit{AVFILMC}) show positive values in Kyte-Doolittle hydropathy scale (Hoop and Woods, 1981) and last two (i.e. \textit{PG}) show 0.0 value in their Hopp-Woods hydrophilicity (Hoop and Woods, 1981) scale. Gly are also found in the core of solved structures of proteins. Pro although may occur in surface, it is hydrophobic (Betts and Russell, 2003).

Total 190 hetero-pairs form 36 \textit{HP-HP} (hydrophobic-hydrophobic), 55 \textit{HL-HL} (hydrophobic-hydrophilic) and 99 \textit{HP-HL} pairs.

\textbf{Site-specific R (} \( R^m \) \textbf{) can be calculated as}

\[
R^m = \frac{\sum f^m_{HP-HL}}{\sum f^m_{HP-HP} + \sum f^m_{HL-HL}} \tag{9}
\]

\[
R^{BLOCK} = \sum_{m=1}^{w} R^m \tag{10}
\]

\textbf{Shannon method:}

Shannon ensemble entropy (Shannon C.E, 1981) is compute using the following formula:

\[
H = -\sum_{i=1}^{M} P_i \log_2 P_i \tag{11}
\]

For residues contribution the same formula was used without summation.
Here, \( P_i \) is the fraction of residues of amino acid types \( i \), and \( M \) is the number of amino acid types (20 in number). \( H \) (ensemble positional variability) ranges from 0 (only one residue in present at that position) to 4.322 (all 20 residues are equally represented in that position). Typically, positions with \( H \geq 2.0 \) are considered variable, whereas those with \( H \leq 2 \) are consider conserved. Highly conserved positions are those with \( H \leq 1.0 \).

We have chosen it for BLOCK preparation.

**Sequence retrieval and BLOCK preparation**

Ferredoxin is well studied, functionally diverse metallo-protein which is available in wide range of organisms including bacteria and green plants. It was used as a probe for understanding the origin of life and biological evolution (Hall and Rao, 1971). Highly curated-sequence and structural database of the protein are also available. With a view of its suitability in gaining insight into the mechanism of substitution we have chosen this protein for its BLOCK analysis. Total of 104 (plant 67, cyanobacteria 31 and others 6) sequences were procured after removal of partial sequences. These sequences are subjected for structural alignment using “Tcoffee expresso” (Armougom et. al, 2006) method. INDELs were manually removed. Thus a total of 93 positions were retained.

**Determination of secondary and coiled structure region**

Following crystal structures that are available in the RCSB database were downloaded (Table S2).

<table>
<thead>
<tr>
<th>Organism</th>
<th>RCSB ID</th>
<th>UNIPROT ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyanobacteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nostoc sp.</td>
<td>1EWY</td>
<td>P0A3C8</td>
</tr>
<tr>
<td>Nostoc sp.</td>
<td>1FRD</td>
<td>P11053</td>
</tr>
<tr>
<td>Aphanothece sacrum</td>
<td>1FXI</td>
<td>P00250</td>
</tr>
<tr>
<td>Nostoc sp.</td>
<td>1J7A</td>
<td>P0A3C7</td>
</tr>
<tr>
<td>Synechocystis sp.</td>
<td>1OFF</td>
<td>P27320</td>
</tr>
<tr>
<td>Mastigocladus laminosus</td>
<td>1RFK</td>
<td>P00248</td>
</tr>
<tr>
<td>Leptolyngbya boryana</td>
<td>3B2G</td>
<td>Q51577</td>
</tr>
<tr>
<td>Spirulina platensis</td>
<td>4FXC</td>
<td>P00246</td>
</tr>
</tbody>
</table>

Table S2. Details of crystal structures used for secondary structure analyses.
In case of duplicate RCSB IDs against a single UNIPROT ID, best structures were considered. The location of secondary structures (strand and helix) are manually collected from RCSB displayed “sequence-chain view” page use DSSP (Kabsch and Sander, 1983) definition of secondary structure. This secondary structural information was also checked in the original structure file. The secondary structural regions thus obtained were localized in the alignment file (Fig. S1).

![Fig. S1. Details of strands (yellow) and helix (cyan) on sequences whose structures are known. Two violet spots (red sequence ID) are excluded as they have deletion. Out of 16, only 14 are used as two IDs (red-colored) harbor deletion (pink-shade).](image)

**Determination of core and surface residues**

NACCESS (Hubbard and Thornton, 1993) program is run using structure files as input (Table S2). The output with extension “.rsa” was used for determination of surface and core residues. The highlighted field was used for the purpose. The residues are taken as core-residues when the relative (REL) accessibility is less than or equal to 25.
Results

As far as site-specific substitutions are concerned four different situations may arise along the width of a BLOCK: $H = 0$ (invariant), $H = 4.32$ (fully-variable), $4.32 < H < 1.0$ (variable) and $1.0 \leq H > 0.0$ (conserved). How are substitutions incorporated in the latter two cases? These sites may be of interest in the understanding of variability and conservation of BLOCK.

Shannon entropy ($SE$) method is widely used to estimate ensemble positional variability ($epv$) of BLOCKs (20). Are $OHP$ extracted $epv$ and residues’ variability similar to Shannon method? To check this, simulations are performed on designed BLOCKs (Table-S1). For simulation-I, seventeen BLOCKs (Table-S1) are used that contain invariant, variable and conserved positions.

Table S1. BLOCKs and their compositions used in the simulation studies. Out of three designed BLOCKs used in the study only two (I and II) are shown. Composition of site 6 and 5 of plant type ferredoxin (pFD) BLOCK are also included.

<table>
<thead>
<tr>
<th></th>
<th>BL</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMULATION-I</td>
<td>V</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIMULATION-II</td>
<td>V</td>
<td>7</td>
<td>28</td>
<td>38</td>
<td>43</td>
<td>60</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>88</td>
<td>69</td>
<td>57</td>
<td>51</td>
<td>32</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIMULATION-III</td>
<td>V</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The first BLOCK possesses 10 Valines. In the second, one of it is substituted by Isoleucine. The substitution of Valine by Isoleucine continues till BLOCK-5. In BLOCK-6 to 8, Isoleucines are successively substituted using Phenylalanine, Serine and Threonine. BLOCK-9 has 5 Valines (dominant-residue) and 5 Isoleucines (second dominant-residue). BLOCKs-10 to 17 are obtained by successive substitutions of Isoleucine and then Valine by other residues. BLOCK-17 is therefore completely variable. BLOCK-2 to BLOCK-16 has partial level of variation. Simulation is also performed using BLOCKs of length 100 with increasing and decreasing frequency of residues (Simulation-
II, Table-S1). In simulation-III, length of BLOCKs is made multiples of twenty where each 20 positions are occupied by 20 different residues. Site-6 and site-5 of plant type ferredoxin BLOCK (pFD) are also analyzed (Table-S1). All BLOCKs are analyzed using OHP (Gupta et. al, 2017) and SE methods.

Will epv and its residues’ contribution be similar by SE and OHP methods? To check this, figure 1 is presented. Several points emerge from the figure. First, epv is calculated using SE (A1) and OHP (A2) methods, show similar increasing patterns. Second, residues’ variability profile are also similar for dominant (Figure-1:B1&B2), second-dominant (Figure-1:C1&C2) and other residues (Figure-1:D1&D2). Third, the maxima of residues’ variability are seen in BLOCK-5 and BLOCK-9 (C1&C2).

Simulation-II is designed with increasing (Valine and Isoleucine) and decreasing (Phenylalanine) residue contents. Variability profiles for Phenylalanine and Valine by SE method do not match with that of the OHP method (Figure-1:E1&E2). Unlike OHP variability, SE variability of Phenylalanine for the first four BLOCKs is seen to be lower than that of Valine, although the former has higher frequency (Figure-1:E1). Again for the last two BLOCKs, lower SE variability is observed for Valine when its frequency is higher than Phenylalanine.

Figure 1. Comparison of variability-profile of different simulation sets (I,II&III) as obtained by SE with OHP methods. Set-I uses seventeen designed-BLOCKs. Variability-
profile by SE and OHP methods, for epv (A1 & A2), dominant-residue (B1 & B2), second-dominant-residue (C1 & C2) and other residues (D1 & D2) are shown in upper and lower panel respectively. Each of set-II and set-III use 5 BLOCKs of constant and variable length respectively. For set-II, residues’ variability profiles, that are obtained by SE and OHP methods, are shown in E1 and E2 respectively. BLOCK-specific residues’ frequency that are shown in E1, are identical to that of E2. For set-III, epv-profile of SE (F1) and OHP (F2) methods are compared.

In simulation-III, variability by SE method is independent of residue frequency and length of BLOCKs (Figure-1:F1). In contrast, by OHP method, variability increases linearly (Figure-1:F2).

To check the effect in natural protein BLOCK, pFD-BLOCK of length 104 and width 93 is used (Figure-2: pFD). pFD is one of the most primitive [Fe2-S2]-chromophore containing proteins and is abundant in oxygenic-photosynthetic cyanobacteria and green-plants. Due to availability of highly curated sequences and structures, it is used for understanding mechanistic details of substitutions. However, profiles of epv is similar for SE (Figure-2:A1) and OHP (A2) methods. To check site-specific residues’ contribution, site-6 and site-5 are assessed using SE (Figure-2:B1&C1 respectively) and OHP (Figure-2:B2&C2 respectively) methods. For these sites are 2.0565 and 0.468 respectively. For site-6, SE method gives lower variability for Isoleucine (Figure-2:B1) than that of Valine. Similarly, although Leucine has highest abundance, its variability by SE method is seen to be much lower than Isoleucine (Figure-2:C1). Such anomalous behavior is not seen in the case OHP method (Figure-2:B2&C2 and Figure-1:E2).

Figure 2. Comparison of ensemble positional (A1 vs A2) and site-specific (B1 vs B2 and C1 vs C2) variability. Upper (A1, B1, C1) and lower (A2, B2 and C2) panels are
computed by SE and OHP methods respectively. Values against points (B1,C1,B2 and C2) indicate site-specific frequencies. pFD-BLOCK was prepared using 104 sequences. In this alignment only highly curated and complete sequences were used that includes 67 from plants, 31 from cyanobacteria and 6 from others of “plant-type ferredoxin family” of UNIPROT (pFD) which constitute about 84% of the database sequences. 14 structures were used for extraction of structural information (pFD). Duplicate and INDEL-containing structures are excluded. Alignment was performed using “TCOFFEE-expresso” method that performs structural alignment. Residue level variability was computed using above methods for site-6 (B1&B2) and site-5 (C1&C2) of the BLOCK. Highly conserved (H≤1.0) chromophore region (residue:32-48) is amplified (D1) along with details on site-specific dominant-residues and NCS-sites (total three, that are shown as blue vertical bars). The region also contains three of four cysteines (shown in red color) in CxxxxCxxC universal format that are involved in chromophore coordination.

Are substitutions equally likely for all positions of BLOCK? R for a given BLOCK is seen to be constant and less than unity (Gupta et. al, 2017). It is 0.63 for pFD-BLOCK. However, it shows wide variation along the width of the BLOCK. 26 of 93 positions (28%) show R>1.0. Figure-3 shows BLOCK-wide distribution of R (Plot-B) along with α-helix, β-sheet, coil, core and surface sites (Plot-C).

Figure 3. Aligned BLOCK of pFD with highlight of sites of (A) dMCRs. Leucine (blue-shade) and Glycine (green-shade) along with four cysteins (yellow-shade) coordinating [Fe2-S2] center, (B) NCSs (N) and CSS (c) and (C) strands (Yellow-shaded), helix (cyan-shaded), core (red and blue residues) and surface (grey residues). Surface and core residues are extracted using NACCESS
procedure using available atomic structures of pFD (further details in S2_Materials and Methods.docx section). Cysteine-3 (C45) and cysteine-4 (C75) that coordinate one of the two Fe-atoms of the chromophore ([Fe2-S2]) center are spaced 29 residues apart in primary sequence which in folded structure come in close proximity (<4.0 Å). The space is highly populated by dMCRs: three Glycines (G47, G52 and G70) and two leucines (L62, L73). It also have three strands and one helix. 8 of 26 sites of non-conservative substitutions (N) are present in this long stretch of sequence of pFD.

As far as secondary structural sites of the BLOCK are concerned, 25% of secondary structures and 30.6% of coil are seen to possess R>1 (Figure-4:M2). Again 15% and 38% of the sites of the core and the surface, respectively, possess R>1 (Figure-4:N2).

Site-specific variability emerges due to substitutions. What is the incorporation strategy for maximum variability to occur? Analyses of the pFD-BLOCK show that except invariant and fully variant sites, substitutions in other sites are somewhat restricted. Both by SE and OHP criteria, it is seen that for a given position, if the abundance of a residue exceeds over others, that residue shows maximal positional variability (Figure-1:C1&C2). OHP and SE derived site-specific details of first three MVRs of pFD-BLOCK (E, V and T) are compared (Figure-4:R1vsH1, R2vsH2 and R3vsH3 respectively). Here, following points are noteworthy:

![Figure 4. Site-specific details of R for MVRs (upper-paired-panel:R1vsH1, R2vsH2 and R3vsH3) and MCRs (middle-paired-panel:C1vsG1, C2vsG2 and](image)
MVRs incorporate maximum positional variability. *SE* criteria shows majority (~70%, Figure-4:H1&H2&H3) of the sites of MVRs are variable (*H*≥1.0). In contrast, majority (~80%) of these positions have *R*≤1.0 (Figure-4:R1&R2&R3) and thus restore parental attributes. Rest of the sites of MVRs (~20%) have *R*>1.0 (i.e. NCSs>CSs) and are directed in α-helix, β-sheet and core sites which are crucial for maintenance of structure and stability of proteins.

Unlike MVRs, MCRs trend to resist positional variability. First three MCRs of *pFD-BLOCK* are *D, L* and *G* (Figure-4:C1&G1 for *D*, Figure-4:C2&G2 for *L* and Figure-4:C3&G3 for *G*). *SE* criteria shows ~70% of the sites of *D* are conserved (Figure-4:G1&G2&G3). In turn, *OHP* method estimates ~95% sites have CSs>NCSs of which ~40% are in secondary structure ((Figure-4:C1&C2&C3). Sites of *L* and *G* are occupied by few residue types. Almost all have *R*=0 i.e. NCSs are absent (Figure-4:C2&C3). Further all the sites of *L* are seen to be present in the core with half of it being present in the secondary structural region (Figure-4:C2). More than 50% sites of *G* are present in the core and secondary structure sites (Figure-4:C3).

For the *pFD-BLOCK*, MVRs, MCRs and DHPs are [*E, V* and *T*], [*D, L* and *G*] and [*ED, IV* and *SA*] respectively. Here, in MCRs, *L* and *G* are not identical to either DHPs or MVRs. Such MCRs are identified as distinct MCRs (dMCRs). *L* is involved in the formation of core at two distinct regions (Figure-5:A4 and A5). *G* imparts flexibility between strands (Figure-5:A2,A3), strand & coil (Figure-5:A6), helix & strand (Figure-5:A5), helix & coil (Figure-5:A4) regions, that might be necessary for the formation of chromophore center (Figure-5:A2). Our analyses with other BLOCKs show, their dMCRs play similar roles (data not shown).
Figure 5. Plot of crystal structure (A2:1AWD.pdb) and its different regions (A3,A4,A5&A5) along with site-specific details on secondary (yellow for β and cyan for α) vs coiled (-) structure, NCSs (N) vs CSs (c) and core (red-R) vs surface (black-R) (A1) along with highlight of dMCRs: six leucines (L5,L23,L33,L62,L73&L93) and six Glycines (G10,G30,G40,G47,G52&G70). G10 (A3), G30 (A4), G52(A5) and G70 (A6) are present between two levels of structure. G47 is present near C45&C75 and G40 is present near C37&C42. Each of these cysteine-pairs coordinate with one of two Fe-center of chromophore. Two hydrophobic patches are formed by Leucines: one by L5,L23&L33 (A4) and other by L62,L73&L93 (A5). These patches are spaced on different sites of [Fe2-S2]-center (A2).

Discussion

Proteins orchestrate functional diversity of cell. Homologous proteins have similar structure with enormous diversity at sequence level. Except invariant-sites, others accommodate variability at locus specific positions. While evolution is the cause of variability in codes, under different environmental conditions, it is however constrained for the maintenance of properties of BLOCK as the parental sequence (Bowie et. al, 1990). Thus, the assessment of epv, its residues’ contribution and substitutions under
homologous situation is central in gaining insight into the principle of natural engineering of sequences for similar codes.

**SE vs. OHP-method**

In aligned BLOCK, it seems, local codes (helix, strand and coil) may not be independent but connected via inter-codes interactions for folding. A method will be useful that performs comprehensive characterization of these sites with respect to physicochemical nature of the residues and substitutions in relation to protein’s function.

Although epv by SE and OHP methods are similar (Figure-1:A1vsA2; Figure-2:A1vsA2), the former showing anomalous behavior for residues’ contribution, might be due to independent assessment of variability. OHP method in turn uses half the sum of the frequency of all possible hetero-pairs formed by a residue, in a site and therefore, is dependent on the other-types. The SE method, which uses 20 residues as basis may underestimate epv when length is >20 and multiple copies of 20 residue-types are present.

In contrast, OHP method uses 190 elements as its basis and thus is more versatile (Figure-1:F1vsF2). In the alignment of natural BLOCKs, >20 sequences are generally likely. Specific sites of BLOCK (e.g. interaction-sites) are seen to harbor preferred residue types (Bartlett et. al, 2002). Assessment of these for residues’ contribution may lead to the discovery of BLOCK and site-specific incorporation strategies of residues’ variability, conservation and constraints, where SE method may not be a reliable choice (see below).

The anomalous observation of variability profiles for Valine and Phenylalanine (Figure-1:E1) are due to the characteristic bell-shaped curve of SE variability against frequency. The problem is not observed in the case of Isoleucine whose frequencies for BLOCKs falls on one side of the peak of the curve. In the pFD-BLOCK (Figure-2:C1&C2) site-5 has 95 Leucines, 8 Isoleucines and 1 Valine. The lower variability of Leucine than that of Isoleucine is a case of underestimation by SE method due to the above mentioned reason. Similar is the situation with Isoleucine for site-6 (Figure-2:B1).

**Condition for MVR**

In evolution, invariant positions maintain absolute conservation (Zuckerkandl et. al, 1965), hyper variable-sites impart complete flexibility (Ayres et. al, 2016), and partially
resistant positions in turn could either be conserved or variable. Highest variability is seen to be achieved when at least half of the total positions are occupied by a residue, followed by others. In BLOCK-9, both Valine and Isoleucine are present in equal counts where the variability is maximum either for Valine (Figure-1:B2&B1) or for Isoleucine (Figure-1:C2&C1). Notably, epv for BLOCK-9 is 1.0 (conserved) and that for BLOCK-6, it is 1.3 (variable), where dominant and second-dominant residues exist with others (Table-S1). In pFD-BLOCK, such sites are abundant, may indicate different levels of resistance from dominant residue which is likely to be the parental type.

**Environmental stresses induce NCSs**

In the pFD-BLOCK, epv shows that sequence between 32-48 are conserved (Figure 2: D1: \( H \leq 1.0 \)). Except six invariant sites (Figure-2 D1-P34-S36-C37-G40-C42-C45), other positions show limited and selective substitutions wherein each position is highly dominated by specific residue type as D32-L33-(invariant)-Y35-(invariant)-(invariant)-R38-A39-(invariant)-A41-(invariant)-S43-S44-(invariant)-A46-G47-K48. By SE and OHP criteria, it is observed that epv receives maximum contribution from each of these residues and therefore seems to be the parental type. While both methods reveal epv, the latter extracts parametric details for sites. For example, position 39 has 94 Alanines, 8 Serines, 1 Glycine and 1 Threonine. Here, MVR is Alanine and dominant hetero-pair (DHP) is AS, with the apparent direction of substitution being A to S. The position is dominated by NCSs (\( R > 1.0 \)) although conserved (\( H = 0.55 \)) (Figure-2:D1). Thorough investigation of annotation of candidate sequences that contains S (instead of A), reveals that 4 sequences participate in N2-fixation in heterocyst, 3 are from different isoforms (FD2=2 and FD4=1) and one is chloroplastic from green algae. A to T substitution occurs in FD2 and A to G substitution is observed in ferredoxin that functions in apicoplast (in *P. falciparam*). Hence, it clearly suggests that the functional environment of the protein acts as the driving force for site-specific substitutions. It also indicates that paralogous sequences have different selective pressure. Here, site-specific substitution in green algae is not clearly understood. In evolution of homologous protein BLOCKs, such situations are more realistic where the original parental residue tends to maintain maximum positional variability by allowing decisive substitutions either more preferably by a
second or few similar types of residues (CSs). Thus, a second dominant residue seems to be preferred and so on.

**R is marker for NCSs-sites**

Although for BLOCKs, $R$ is less than unity, its site-specific details vary. In the $pFD$-BLOCK, 28% of sites have $R>1.0$ indicating site-specific preference for NCSs. While CSs restore parental properties, NCSs modulate it. The latter could be also deleterious (Ng and Henikoff, 2001). It may not be the primary sequence, but the sequence of helix, strand, coil, core and surface that acts as a determinant for a specific fold. Nearly identical sites of ordered secondary (25%) and coiled (30%) structures with $R>1$, may indicate equal compatibility of these for the accommodation of evolutionary pressure. Notably, coil-sites were reported to harbor crucial interaction sites (Sanz et. al, 2003). As far as the core and surface are concerned, the former which is known to be crucial for protein folding is far more conserved.

**MVRs incorporate variability in secondary and core structures**

Site-specific residues’ variability in BLOCK is the consequence of evolution, whose assessment is central in gaining insight into site-specific tolerance of mutations in relation to protein function (Echave et. al, 2016). For a given BLOCK, sites of $MVR$s with $R>1.0$ indicates dominance of NCSs over CSs. For the $pFD$-BLOCK, the first three $MVR$s are $E$, $V$ and $T$. The former is acidic with highest helical propensity. The latter two are β-structure promoters with $V$ being hydrophobic and $T$ being hydrophilic (Williams et. al, 1987). Plant type ferredoxin is highly acidic in nature with characteristic β-grasp fold (Burroughs, et. al, 2007) that match with physicochemical characteristics of $MVR$s. Lower $R$ value for majority of the sites of $MVR$s indicates substitutions in the restoration of positional properties. Bowie, et. al., (2015) pointed out that NCSs are less important for structure and function of BLOCKs (Bowie, et. al, 2015). Contrary to that, most surprisingly, among all the sites of $MVR$s, very few sites (20-30%) are involved in the modulation of helix, strand and core that are crucial for protein folding (Figure-4:R1,R2&R3). In this respect, $H$ values are non-specific (Figure-4:H1,H2&H3). Hence, these sites of NCSs seem to be more effective in the neutralization of environmental stress.
related structural changes. This unique observation might point towards the robustness and novel strategy of structural modulation in relation to protein adaptation.

**MCRs and distinct MCRs**

Lower $R$ for most of the sites of MCRs indicate dominance of CSs over NCSs. Notably, sites of MCRs are not invariant indicating highly selective CSs (Figure-4:C1,C2&C3). These sites are not only located in the core and ordered structures, but also in coil (Figure-4:C1), indicating crucial conservation may be equally important for the latter that harbors binding, active and interaction sites. Due to participation in substitutions, MCRs can be identical with either MVRs or DHPs. For example, it is seen that $D$ (MCR1) substitutes $E$ (MVR1) therefore forming dominant hetero-pair (DHP1) $ED$, which acts to preserve the acidic nature of the protein. In case where MCRs are not shared with MVRs and DHPs, they form dMCRs. The latter seems to have crucial structural and or functional roles (see below).

In *pFD*-BLOCK, six dMCR:L are directed in two distinct sites (Figure-5:A4&A5) of the core. L5 and L23 are also in strand and helix respectively, which allows these secondary structures to come in close proximity. Therefore it seems, dMCR:L not only induces core formation, but also promotes inter-code interactions. $G$ is another dMCR of which 3 are present in the core. *pFD* being a metallo-protein, requires attachment of [2Fe-2S]-chromophore with its four invariant Cysteines (Figure-5:A7), in sequence these are distantly positioned, with C45 and C75 being 29 residue apart (Figure-5:A1). In folded structure, they are seen to be within 2.4Å of Fe-centers (Figure-5:A7). Therefore, the presence of six $G$s (dMCRs) near the [Fe2-S2]-center, between $\beta$-$\beta$ (Figure-5:A3&A6), $\beta$-coil (Figure-5:A2&A6), $\alpha$-$\beta$ (Figure-5:A5) and $\alpha$-coil (Figure-5:A4) regions seems to provide junctional flexibility and thereby helping appropriate coordination of the chromophore with the protein.

**Conclusion**

By the use of designed and functional protein BLOCKs, we demonstrate that in comparison to SE method, OHP method is more accurate, exhaustive and versatile. While
both these methods are capable of efficient extraction of epv from BLOCK. SE method is anomalous in the extraction of residues’ contribution. OHP method in comparison to SE, has multiple additional features for parametric characterization of site-specific evolutionary and conservation properties of BLOCK. Maximally variable residues (MVRs) incorporate maximum site-specific variability in discontinuous fashion with 70-80% sites having \( R \leq 1.0 \). The rest 20-30% having \( R > 1.0 \) are directed in sites of helix, sheet and core residues, which are crucial for the structure and stability. MCRs use similar distribution as MVRs. In this case about 95% of sites have \( R \leq 1.0 \) with majority are forbidden for NCSs. MCRs that are not common with MVRs and dominant hetero-pairs (DHPs) are distinct MCRs (dMCRs). dMCRs have potential role in the structure and stability for parental fold.

References


Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome research. 2001 May 1;11(5):863-74.


CHAPTER 3
Substitutions in common shared region of plant-type ferredoxins in Relation to Salt Adaptation of Halophilic Ferredoxin

Summary

The knowledge of amino acid changes in protein sequence is of great importance in understanding protein evolution and adaptation in a given environment. While mesophilic ferredoxin (mFD) and halophilic ferredoxin (hFD) belong to plant-type family, the former takes part in oxygenic photosynthetic reaction in membrane bound form and the later in contrast is soluble in the cytoplasm and participates in oxidative decarboxylation reaction in the presence of high salt. Hall et. al, (1971) hypothesized a novel “additional protein moiety” in mFD for its effective interaction with photosynthetic lamellar systems. Similar insertion in the N-terminal region of hFD was proposed for its high salt adaptation (Oren A, 2002; Britton et. al, 1998; Marg et. al, 2005). However, the differential roles of amino acid changes at locus specific positions in the common shared region (CSR) of these molecules remain to be worked out. Here we show domain of lives specific conserved substitutions in the CSR of mFD and hFD play crucial role in differential adaptation in their respective environments. Although hFD and mFD belong to the same family, their CSR are distantly related indicating domain of lives specific evolution. Highly conserved substitutions are identified from ferredoxin-Oxidoredctase and photosynthetic membrane interaction zones of CSR that are helping to gain insight into environment specific differential adaptation of mFD and hFD for these zones. Moreover, CSR of hFD have acquired highly conserved substitutions relative to mFD which induce long-range and networked salt-bridges, a halophilic design observed in other halophilic proteins (Eisenberg et. al, 1992). Our results demonstrate changes in the rally of codes in protein sequences for its function and stability in their respective environments. We anticipate our
approach involving differentially adapted protein molecules is applicable for many similar systems evolved under varieties of stressed environments. The study can be combined involving site directed mutagenesis and other in vitro approaches to comprehend the knowledge of natural protein engineering and molecular adaptation.

**Introduction**

Evolution of protein relay on many factors and mechanisms such as substitutions, deletions and insertions for its functioning in a given environment (Pál et. al, 2006). Homologous or Orthologous proteins may bear crucial signatures of these mechanisms at the sequence and structural level understanding of which is the major concern in natural protein engineering (Zhang et. al, 2002).

[2Fe-2S] ferredoxins belong to plant-type family where the chromophore center attached with four conserved Cysteines residues (Hall et. al, 1971). The family is ubiquitous that are abundant in specific families of the domains Bacteria, Archaea and Eukarya. The protein is also functionally diverse as it takes part in electron transfer reaction in varieties of biochemical processes such as photosynthesis, Nitrogen fixation and oxidative decarboxylation reaction (Kerscher and Oesterhelt, 1981). The cyanobacterial ferredoxin (cFD) and plant ferredoxin (pFD) together form mesophilic ferredoxins (mFD) that takes part in oxygenic photosynthesis in partially membrane bound form (Hall et. al, 1971) whereas halophilic ferredoxin (hFD) is representative from archaeal domain of lives of family halobacteriaceae that takes part in oxidative decarboxylation reaction in soluble form in the cytoplasm that harbor saturated salt solution (Kerscher and Oesterhelt, 1981). Apart from these, [2Fe-2S] ferredoxins are also found among representatives of proteobacteria (PrFD) that operates in extreme of salt environments (Oren A, 2006) and participates in nitrogen fixation (Kerscher and Oesterhelt, 1981). Ferredoxin is a co-factor protein that interacts with ferredoxin-NADP+-Oxidoreductase in case of mFD and different Oxidoreductase enzymes in the case of hFD (Kerscher and Oesterhelt, 1981).

Extra region is seen to acquire in sequences of proteins that are anticipated for additional function in their evolution. Hall et. al, (1971) identified a novel “additional protein moiety” in the mFD when compared it with anoxygenic photosynthetic bacterial
ferredoxins. It was considered to help the earlier to interact effectively with “photosynthetic lamellar systems”. Similarly comparison between mFD and hFD showed an insertion of some 26 amino acids at N-terminal end of the later (Marg et. al, 2005) which was considered to contribute its adaptation in high salt environment (Marg et. al, 2005; Eisenberg et. al, 1992). hFD is thus well studied whose high salt adaptation has many different views. While biophysical studies demonstrate adaptation involve the entire protein (Bandyopadhyay AK, 2000; Bandyopadhyay et. al, 2001; Bandyopadhyay et. al, 2007) atomic structural studies highlight the insertion domain and its preferential acidic substitutions (Marg et. al, 2005; Frolow et. al, 1996). The insertion domain which was postulated to be due to lateral gene transfer between halobacteria and cyanobacteria (Pfeifer et. al, 1993) was found to behave as “intrinsic chaperon” that help to maintain folded state of the protein in high salt (Hirota et. al, 2005). Statistical comparison between the sequences of hFD and mFD showed the former reduce bulkiness while keep hydrophobicity same as later, an observation that contradict general compositional bias of other halophilic proteins (DasSarma et. al, 2006; Karan et. al, 2012; Paul et. al, 2008). However, it was proposed that hFD competes with cytosolic salt for water through its excess carboxylates contributed by Asp and Glu (Rao and Argos, 1981).

Hall et. al., (1971) considered ferredoxins for understanding evolution and the origin of life. While ferredoxin of anaerobic fermentative bacteria was considered as root, green-plant one was taken as final step of its evolution. Although two intermediates states namely green and red photosynthetic bacterial ferredoxins were proposed to be involved, a big gap in the transition from red-photosynthetic bacteria to green-plant was postulated that raises the possibility of additional intermediate state in the evolution (Zaccai and Eisenberg, 1990). At the time of the above studies, representative sequences were very few in the literature and hFD was unavailable. At present database possesses representative sequences of ferredoxins from all domains of lives including halophilic archaea. However, the fundamental questions as to what is the evolutionary position of hFD in the tree proposed by Hall, et. al., (1971) and what are the differential roles of amino acid changes at locus specific positions of mFD and hFD in the common shared region (CSR) remain to be answered.
In this study, we worked out a comprehensive model of evolution of plant-type ferredoxins including representatives from all domains of lives. We also report on substitutions in the CSR of hFD and mFD that may act as determinant for differential subcellular location and interactions with Oxidoredctase in these molecules. Unlike mFD, deployment of redesigned salt bridges for maintenance of stability of hFD in high salt is also presented in this article.

**Materials and Methods**

**Sequence retrieval**
Primary sequences of ferredoxins were retrieved from Uniprot databases. The primary sequences of these proteins were also obtained in fasta format from PDB sources ([http://www.rcsb.org/pdb/home/home.do](http://www.rcsb.org/pdb/home/home.do)). Structures of these proteins were retrieved from RCSB.

**Phylogenetic tree construction**
The NR-PKS sequences and PT sequences were aligned with ClustalW2, respectively. Phylogenetic analyses were conducted using MEGA version 7 by the bootstrap neighbor joining method52. The evolutionary distances were computed using the Poisson correction method and were in the units of the number of amino acid substitutions per site. The phylogenetic tree was displayed by iTOL53.

**Salt bridge determination**
Pair wise sequence alignment was performed by using the ClustalW2 using default parameters as supplied. Tertiary structural alignment was performed using the method developed by Zhang and Sholnick (2005), which is faster and much accurate and independent to the length of the sequences. Extractions of salt bridges for their analyses were done using SBION2 (Gupta et. al, 2015).

**Sequence hydrophobicity determination**
Hydrophobicity determination for their analysis in common shared region was done using PHYSICO (Gupta et. al, 2014) and PHYSICO2 (Banerjee et. al, 2015).
Ion-pair determination

Ion-pair distance calculation was made following three methods. Firstly protein structures were loaded in VMD v 1.8.3 (available at http://www.ks.uiuc.edu/Research/vmd/) and supplied plugin was used directly for identification of the pairing candidates. In this case a broad cut-off value was chosen for identification of all possible pairs. Protein structures were further displayed by Swiss PDB viewer (available at http://spdbv.vital-it.ch/) and for each BAA-AAA pair individual atom-pair distances were computed. Further, for residue and networked salt bridges extraction SBION2 (Banerjee et. al, 2015). Finally, manual verification were made of each individual pair as necessary using the PDB coordinates for atoms.

Result and Discussion

Both mFD (pFD and cFD) and hFD belong to plant-type ferredoxin family of which the former takes part in photosynthesis and the later in oxidative decarboxylation reaction in cytosol that harbor saturated salt. The alignment of hFD with pFD and cFD shows three extra (INDEL) regions (Figure 1a) of which N-terminal one is the longest. About 50% of it is constituted by acidic residues. X-ray crystallographic (Frolow et. al, 1996) and NMR (Marg et. al, 2005) studies showed that this segment forms amphipathic helix domain that contributes to the halophilic property of hFD. The same region was shown to play role in the maintenance of folded form of the protein in high salt (Hirota et. al, 2005). Salt dependent spectroscopic studies showed global structural loss at low salt including tertiary, secondary and chromophore structures (Bandyopadhyay AK, 2000; Bandyopadhyay et. al, 2001).

General characteristic of “common region” of hFD

Phylogenetic tree generated using the sequence BLOCK (Figure 1c) and presented in Figure 1, b1 and b2 along with selected proteobacterial ferredoxins and halophilic UspA proteins. It is seen that “common region” of hFD differs by 60-70% from mFD indicating large evolutionary differences of the former for the region. Anoxyogenic photosynthetic and salt-loving proteobacterial ferredoxin (PrFD) forms an out-group (Figure 1 b2) in the tree. Usp A, highly diverse, most ancient and ubiquitous protein,
responses to a variety of environmental stresses such as high salinity, high temperature, nutrient starvation etc (Zaccai and Eisenberg, 1990) forms a sister group with hFD but appear earlier than PrFD. Halophilic Usp A branch out much earlier than hFD indicating it is more ancient than the later. Cladestic relationship of Usp A with hFD further indicates that the “common-region” of the later might have similar structural element as the earlier to resist deleterious effect of salt in the cytosol. However, it appears that hFD is intermediate between PrFD and mFD (pFD and cFD) raising the possibility of horizontal gene transfer between hFD and PrFD. Further, such transfer is also facilitates between UspA and hFD as they form clade.

Relative composition, class-composition and mean sequence properties of hFD for the “common region” is presented in Figure 1f, 2g and 2h respectively. hFD shows distinct compositional characteristics for the “common region” that are rarely entertained in other halophilic proteins in general. Firstly overall hydrophobicity (Figure 1g), bulkiness and VDW volume (Figure 1h) show an increase with reduction of drastic decrease of boarder-line hydrophobic residues (ST) in hFD. Secondly, acidic residues (DE), total charge (TOT), net charge (NET) at neutral pH show a decrease in hFD with an increase of basic residues (KR). Finally Glycine and Proline (PG) that impart flexibility and induce turn respectively show a decrease in hFD. The above observations for the “common region” of hFD contradict the generally accepted mode of halophilic adaptation (Paul et. al, 2008; Lanyi JK, 1974 and DasSarma et. al, 2006). It is responsible for FD-FNR interaction, [2Fe-2S] chromophore coordination with four evolutionarily conserved cysteines residues. While these structure-function requirements are common for both mFD and hFD, the later is further constrained with its adaptation in high salt environment. Again, unlike mFD which is partially bound to photosynthetic membrane system, hFD functions in soluble form in cytoplasm in the presence of high salt. Thus it seems a challenge to understand the halophilic adaptation of the “common region” of hFD whose evolution is constrained with above structure-function requirements unlike other halophilic proteins. Below we present results in this aspects.

**Hydrophobic profile of hFD for the “common region” reveals salt dependent properties**
Figure 1c present the “common-region” along with highlights for secondary structures (beta-strand as yellow and alpha-helix as cyan) and substituted amino acids whose domain specific conservation are assessed using Positional Shannon Entropy (PAN) (Figure 1e). Figure 1d show kyte-Doolittle mean Hydrophobicity profiles for the “common region”. Both pFD (green) and cFD (cyan) contain two hydrophobic regions (R1 and R3) that are connected with the sequence BLOCK using blue and green lines respectively. In mFD (pFD and cFD) R3 is more prominent than R1. R1 and R3 match with beta-3 and beta-5 region of the BLOCK respectively. In hFD (pink) R1 is more prominent than that of mFD and R3 is almost absent.

**R1 of hFD shows increased hydrophobicity**

In region R1 that overlap with beta-3 segment of BLOCK contains two crucial substitutions i.e. G47A and K48I (mFD → hFD). Both A47 and I48 are halophilically conserved as PSN for positions 47 and 48 are 0.5 and 0.0 respectively. Again G47 and K48 are also mesophilically conserved as PSN for these positions are 0.05 and 0.08 respectively. Both A47 and I48 (in hFD) are hydrophobic with highest hydrophobicity for I48. Substitution of G by A in 47 position (i.e. G47A) seems to be a desperate decision in hFD as conserved Glycine is known to play crucial structural role (Betts and Russell, 2003). However, overall increment of hydrophobicity in the region R1 in hFD is related with G47A and K48I substitutions. We raise question as to what is the structural role of this incremented hydrophobicity in R1 in hFD. Crystal structure of mFD from cyanobacteria (1EWY) and plant (1GAQ) are complex of ferredoxin and FNR (Ferredoxin-NADP+-Reductase). The interaction sites of ferredoxin and FNR for both cyanobacteria and plant (both mesophilic) were determined using NACCESS procedure and presented in figure 1d along with Kyte-Doolittle profiles. It is seen that two overlapping contiguous regions of cFD and pFD (i.e. C1 and C2 in Figure 1d) along with spatially close but distantly spaced residues in sequence are responsible for contact with FNR. It is also seen that the first hydrophobic region (R1) which is common for both hFD and mFD with greater hydrophobicity in the former shows reasonable overlap with the contact region (C1) of FD-FNR complex of mFD. The observation might indicate
hydrophobicity of R1 contributes to the mFD-FNR interaction in mFD. The fact that R1 of hFD overlap with R1 (hence C1) of mFD, we propose similar FD-Oxidoreductase interaction in hFD for the region. The enhanced hydrophobicity of R1 (in hFD) seems to be related with halophilic adaptation of hFD. Water activity in high salt medium is much lower than aqueous solution (Hasted et. al, 1948). In absence of substitutions (G47A and K48I), R1 of hFD would have lower effective hydrophobicity than mFD under halophilic situation (saturated salt solution). It therefore seems reasonable to think that to achieve similar level of hydrophobicity as mFD in high salt, R1 of hFD has acquired these crucial substitutions (G47A and K48I).

**R2 of hFD shows decreased hydrophilicity**

The second FD-FNR contact region (C2) overlaps with a hydrophilic region (i.e. R2; Figure 1d) of the Kite-Doolittle profiles of mFD and hFD. In hFD, R2 matches with residues M56, D57 and M58. The PSN was 0.0, 0.0 and 0.0 for these positions respectively indicated absolute halophilic conservation. The corresponding residues in mFD are Q56, S57 and D58 with PSN 0.05, 0.52 and 0.14 respectively indicated these positions are also mesophilically conserved. Similar hydrophobic profile of mFD and hFD for the region R2 and its overlap with mesophilic FD-FNR interaction zone (C2) might indicate similar FD-Oxidoreductase interaction for hFD. However, the substitutions of Q56, S57 and D58 of mFD by M56, D57 and M58 of hFD seem to be related with halophilic adaptation in relation to FD-Oxidoreductase interaction in the later. The observed reduction in hydrophilicity for the region R2 is due to substituted hydrophobic residues M56 and M58. Due to lower effective hydrophobic force under halophilic conditions (Hasted et. al, 1948) this decreased hydrophilicity might get adjusted at similar level as mFD in high salt where hFD function.

**R3 of hFD is hydrophobically neutral**

In mFD R3 which overlap with beta-5 of sequence BLOCK is more prominent than R1 and is almost absent in hFD (Figure 1d). This hydrophobic region is unlikely for effective FD-FNR interaction as single residue (but not a region) is seen to involve in the
interaction. What could be the role of R3 in mFD and hFD? Hall et al. (1971) anticipated that mFD has acquired an additional region in comparison to photosynthetic red-bacteria for membrane interaction in photosynthetic system. We propose that R3 is the region for FD-membrane interaction.

Figure 1 Alignment of complete sequence of hFD and mFD (a), phylogenetic tree by neighbor joining method (b1 and b2), common (BLOCK) region of hFD and mFD after removal of INDELs (c), Kyte-Doolittle hydropathy profiles using a window size of 9 (d) for hFD (pink), plant (green) and cyanobacteria (cyan) along with FD-FNR interaction region for pFD and cFD. The profiles were smoothed using negative exponential of degree 1 in SIGMA PLOT v11.0 of SYSTAT Scientific Software. Positional Shannon Entropy (e) for hFD (pink), plant (green) and cyanobacteria (cyan), relative composition (f), relative class-composition (g) and relative mean sequence property for the common region. The alignment and construction of phylogenetic tree was performed by Mega7. INDELs were removed manually to obtain common (BLOCK) region of
hFD and mFD (pFD and cFD). In the BLOCK regions for beta (beta-1 through beta-7 in yellow color) and alfa-helix (alfa-1 through alfa-3 in cyan color) are shown along with highlight of halophilically substituted acidic (red), basic (blue) and hydrophobic (black) residues and BLOCK specific conserved residues. FD-FNR interaction was determined using NACCESS procedure using available complex-crystal structure of pFD and cFD whose regions are aligned with hydropathy profiles. Shannon entropy and relative physicochemical properties were determined using PHYSICO2 procedure. Relative class composition for hydrophobic (HB), Hydrophilic (HL), polar un-charge (PU), Acidic (DE), Basic (KR), Gly-Pro (PG) residues, total charge (TOT), net charge (NET) at neutral pH (D+E-K-R) and disorder forming residues (DIS) were determined. Residue index value dependent mean sequence properties for Aliphatic Index (AI), Isoelectric Points (pI), GRAVY (KD), Zimmerman Bulkiness (ZB), Deligex-Roux Beta (beta), Deligex-Roux Alfa-Helix (alfa), Deligex-Roux Coil (coil), VDW volume (vol), Hydrogen bond donor (hb), Jenin Buried and accessible surface and Flexibility were determined. Relative values in each case were calculated using (h-m)/m formula where h is the mean value for hFD and m is that for mFD.

The region (R3) is 6A and 7A away from hydrophobic (R1) and hydrophilic (R2) regions of FD-FNR interaction zone respectively. Notably hFD which takes part in oxidative decarboxylation reaction is soluble in cytosolic salt solution and thus such hydrophobicity (R3) seems no longer be needed. The enhanced hydrophilicity of hFD for the region (R3) is observed to be due to a crucial substitution (I/V72R) corresponding to the peak of R3. The former residue has highest (I72 +4.5; V72 +4.2) and the later has the lowest (R72 -4.5) hydrophobicity in the residue hydrophobicity scale (Cornette et. al, 1987). The substitution of I72 or V72 of mFD by R72 of hFD seems to be a deliberate evolutionary consequence that may not only abolish the possibility of membrane interaction of the region (R3) in hFD but also increase solubility of the protein in cytosol. Further, Arginine is known to reduce protein aggregation a fact generally entertains in high salt (Tsumoto et. al, 2004).

We show later that R72 also form a stable salt bridge which would replenish the energetic loss due to substitution of hydrophobic residue in mFD. Soluble [2Fe-2S] ferredoxin of rhodobactor that takes part in N2-fixation also possesses Arginine at this position (Armengaud et. al, 1994). Further halophilic Usp A that functions in cytosol in soluble form also contains Arginine at this position (Kixmüller et. al, 2011). What about the domain specific residue conservation? PSN for hFD for the position is 0.0 and that for mFD 0.1 indicating the position is halophilically and mesophilically highly conserved. We propose that the “additional protein moiety” anticipated by Hall et. al., (1971) seems to be centered on the position 72 where a Valine or Isoleucine in present
case of photosynthetic ferredoxins and an arginine replace that for soluble ferredoxins. Thus targeting of the protein for membrane or cytoplasm is largely controlled by the substitution.

While crucial substitutions in the “common region” provide insight into the FD-Oxidoreductase interaction and solubility of hFD in cytoplasmic solution, the question relating to the stability of hFD in high salt remain to be understood.

Figure 2 Comparison of salt-bridge/ion-pair pattern at N-terminal (h1 vs m1), internal (h2 vs m2), C-terminal (h3 vs m3) and active site region (h4 vs m4) for hFD and mFD. In each case salt-bridge/ion-pair residues are located both in the structure and aligned BLOCK using different colored lines. Green, red, blue and pink colored lines are used to indicate N-terminal, internal, C-terminal and active site salt-bridge/ion-pair respectively. The BLOCK highlights both the residue of interest and their domain specific positional conservation using Shannon Entropy criteria. Secondary structural regions are also highlighted using similar color code as Figure 1. Halophically (h), plant (p) and cyanobacteria (c) specific conservation of (Shannon Entropy <=1.0) residue were obtained by PHYSICO2 analyses using domain specific BLOCK FASTA files (sequence >75). Non-conserved residues are shown as n and globally conserved residues are shown as g.
As shown above both Usp A (a stress response protein) and hFD form a clade (Figure 1b2) with the former branch out earlier than the later, the stress related structural elements seems to be present in the “common region”. It was shown that salt bridge or ion-pair interactions play crucial role in the stabilization of halophilic proteins in general (Nayek et. al, 2014) We have presented salt-bridge or ion-pair design of hFD in reference to mFD in Figure 2.

**hFD introduces long-range and networked ion-pair in the N-terminal region**

mFD establishes a β1(K4)—β2(E15) local connectivity (Figure 2 m1) using conserve K4 and E15 residues respectively. However, the above salt bridge is not conserve in evolution as judged by positional PSN values. While K4 is conserved (PSN ≤ 0.4) for both pFD and cFD, E15 is non-conserved (PSN ≥1.5) in these domains. Unlike mFD, hFD establishes connectivity among β1(K4), β3(E51) and β7(K65) as N-terminal design (Figure 3 h1). N-terminal ion-pair thus becomes long range and networked in the case of hFD. These three positions (4, 51 and 65) undergo substitutions with halophilically conserved acidic and basic residues namely K4E, A51E and I65K (figure 1c). It was shown that networked salt bridges increase under halophilic situation (Nayek et. al, 2014). While long range connectivity would facilitate better maintenance of characteristic beta-grasp fold structure of ferredoxin in high salt, networked salt bridge enhances overall stability. The computed ΔΔG_net was found to be -4.8 Kcal mol⁻¹, indicating that the salt-bridge is highly stabilizing. Thus, formation of newly designed N-terminal salt-bridge with the help of halophilically substituted and conserved residues seem to be related with halo-adaptation of hFD.

**hFD introduces long-range and networked ion-pair in the C-terminal region**

In mFD, α3 (K91 and D94) and β7 (E88) form a local [i→i+3 i.e. K91→D94] and [i→ i-3 i.e. K91→E88] type ion-pair interactions. E88 of β7 also forms ion-pair interaction with K52 of β3. Here the interaction may looks like: D94(α3)—K91(α3)—E88(β7)—K52(β3) (Figure 2 m3). As shown above K52 (β3) is also involved in FD-FNR interaction in mFD. While the ion-pair is networked and long-range type, the question that what is the universality of such interaction for this domain of life seems relevant.
mFD is constituted by pFD and cFD. The PSN presented in the figure shows that K52 is conserved and E94 is non-conserved for both pFD and cFD. While K91 and E88 are conserved in pFD, they are are non-conserved in cFD (Figure 2). This would mean the pattern of ion-pair as mentioned is not universally preserved in mFD. In other word this ion-pair seems to be less important for structural evolution of mFD. In contrast hFD form a newly designed ion-pair involving α2 and C-terminal extra region (Figure 2h3). In α2, E92 and K94 join in i to (i+4) fashion. Again E92 form salt-bridge with R126 residue. The salt-bridge K94(α2)—E92(α2)—R126(CT-extra region) is thus networked and long-range. Are these residues conserving in hFD? Unlike mFD, these residues are halophilically highly conserved and hence seem play important role in the stability of FD in high salt. The computed ΔΔG was found to be -1.8 Kcal mol⁻¹, indicating the salt-bridge is moderately stabilizing.

**hFD introduces novel salt-bridge in the middle region**

In hFD a novel salt bridge is introduced in that residues D79 (β4) and R99 (β5) establish β4-- β5 connection. It is long-range and isolated type. Both these positions are observed to be halophilically conserved as judged by positional Shannon Entropy (Figure 2). ΔΔG was found to be -2.4 Kcal mol⁻¹ and thus stabilizing. In hFD R99(β5) is a novel substitution at R3-region of the sequence (Figure 1c) that substitutes V99(β5) in mFD. In mFD, V99(β5) establishes hydrophobic contact with V77(β3). Substitution of V99 by R99 in hFD abolishes such hydrophobic contact. In turn salt bridge interaction between R99 (β5) and D79 (β3) would balance the above energetic loss. Overall, these halophilically substituted and conserved residues contribute to halo adaptation of hFD of which R99 play dual roles.

**Salt-bridge in active site**

In this case both hFD and mFD show similar design of salt-bridge in that it is largely E24 (α1)—E27 (α1)—R38 type. All these positions are seen to be highly conserved as judged by their PSN values (Figure 2). ΔΔG is -5.0 Kcal mol⁻¹ and thus stabilizing. Taken together hFD establishes designed long-range and networked salt bridges involving halophilically substituted and conserved acidic and basic residues both at N-
terminal and C-terminal regions. In mFD although salt bridges are detected for both the regions, the design suffers from positional residues conservation and hence universality of these salt bridges. Unlike mFD, hFD introduces new salt bridges using halophilically conserved and substituted residues. In active site the salt bridge remains unaffected for both hFD and mFD. It thus indicates these design salt bridges in the “common region” of the sequence play crucial role in halophilic adaptation of hFD.

**Conclusion**

We employed sequences and structure of plant-type ferredoxins of database to work out model of its evolution, possible reason of differential sub-cellular location and interaction-sites with Oxidoreductase under different solvent conditions. We show for the first time by employing the CSR of mFD and hFD that the later which functions in high salt in cytoplasm is an intermediate-clade between PrFD and mFD. Cladestic relationship of hFD with UspA showed the possibility of incorporation of stress related properties of the later into the former. Interaction of hFD with Oxidoreductase is achieved using similar chemistry as mFD with modulation of interaction sites by incorporation of highly conserved halophilic substitutions, might be due to alleviation of deleterious effect of salt. Similar substitutions are also incorporated in different zone of CSR of hFD for alteration of sub-cellular location of the molecule. While in high salt proteins get destabilized in general, hFD remains stable by redesigning its salt-bridges using substituted acidic and basic residues which are otherwise absent in its mesophilic counterpart. These salt bridges are largely networked and long ranged in contrast to mFD. Further, in compare to mFD, salt bridges of hFD are highly conserved in evolution. Taken together our study employing CSR of plant-type ferredoxins show differential adaptation that suits their environment and has potential in the field of protein engineering.

**References**


Oren A. Adaptation of halophilic archaea to life at high salt concentrations. InSalinity: Environment-Plants-Molecules 2002 (pp. 81-96). Springer Netherlands.


CHAPTER 4

Are Amino acid Substitution matrices really Universal

Summary

DNA sequence or its translated form as protein sequence acts as code of life. Establishment of structural, functional and evolutionary relationship from an input-set of sequence of proteins from different sources is the prime concern of functional and comparative genomics which is achieved by bioinformatics studies. In this context two series of scoring matrices such as PAM (Dayhoff et. al. 1978) and BLOSUM (Henikoff and Henikoff 1992) are most widely used. While PAM based on explicit evolutionary model, BLOSUM relay on implicit procedure. The later earned much popularity in recent days in the field of sequence bioinformatics due to its capacity for analyzing the relatedness of highly diverge sequences.

Each of the above matrices assigns a constant bit-score for each pair of amino acid of 210 possible pairs. These assigned values are then used for determining score of a given sequence with reference to a standard one. Are the amino acid score of BLOSUM series of matrices constant? To answer the question we performed simulation on various sets of blocks of sequences in that in each set, a given parameter is varied keeping others constant. Some of these important parameters which were tested in the present work are: width of blocks; length of blocks; location of homologous position in a given block. The simulation was performed using automated home-build procedure (https://sourceforge.net/projects/genblosumxx/) for fast, efficient and error free calculation of scores of amino acids for blocks. Then we have constructed score vs. parameter plot for each individual amino acid pair. Our observations in plot suggest that amino acid scores are not constant, which is expected, rather vary with the simulated parameter. For example score for a pair of amino acid vary linearly with the width of block whose correlation co-efficient is almost one. Thus, our analysis proves that
BLOSUM series of matrices are not universal in nature. This finding proves the need for development of new score matrices.

**Introduction**

In the database both number and complexity in terms of types, functions and source-organisms of protein sequences have increased in recent time. Detailed analysis and finding relatedness are generally performed by binary scoring methods (Bucher and Hofmann 1996). PAM (Dayhoff et. al. 1978) and BLOSUM (Henikoff and Henikoff 1992) are two such fundamental matrices of which the former is explicit and the later follow implicit substitution model. BLOSUM62 is the best among all BLOSUM series of matrices as it works better for distantly related sequences and thus it gained much popularity in bio-computations. It was (Henikoff and Henikoff 1992), to derive this matrix first in that conserved regions of protein families in the form of BLOCKs were considered for obtaining relative frequencies of amino acids and their substitution probabilities. The BLOSUM series are implicit substitution matrices that help in finding relatedness, obtaining alignment and constructing phylogenetic tree by binary scoring method of input sequences. It is the most widely used matrices for the purpose due to its better performances relative to other ones. Bit-score was obtained for all 210 possible substitution pairs of the 20 standard amino acids. The procedure involves series of steps. Firstly, a large number of sequence blocks (>2000) are to be made maintaining a level of similarity (say ≥62%). Secondly each block is to be analyzed for obtaining scores of 210 pairs of which 20 are homo-substitution-pairs and 190 are hetero-substitution pairs. Finally, the scores thus obtained for all sequence BLOCKs are averaged to get the bit score matrix (Henikoff and Henikoff 1992). The final matrix is the BLOSUM matrix, used in conjunction with many popular web based programs such as CLUSTALW (Thompson et al. 1994), BLAST (Altschul et al. 1999), etc.

In this work, we wanted to test whether the bit score used by many popular program and software are universal. At this juncture, a program that analyzes BLOCK-FASTA-FORMAT (blff) file(s) for bit-score would be useful in simulation as well as comparative evolutionary studies among various orthologous BLOCKs. It would further be useful for
generation of average matrix of desired series (e.g. xx) from related sequence BLOCKs. Therefore, we have developed GENBLOSUMxx (available at https://sourceforge.net/projects/genblosumxx/) along with details of its execution, outputs and usefulness not only for the simulation work but also for extraction of additional information in relation to substitution mechanism apart from the state-of-the-art analyses of sequence BLOCKs in bioinformatics and protein engineering.

**Materials and Methods**

**Development of the program for simulation**

The program is written and interpreted by AWK programming language (Aho et. al. 1977), and functions from UNIX/CYGWIN shell terminals. For each BLOCK, bit-score i.e. log odd ratio of observed to expect probability (Henikoff and Henikoff 1992), is computed (Figure 2 B) in 20x20 matrix format.

Details of operations of the program are shown in the flowchart (Figure 1).

---

**Fig. 1 Flowchart for the details of operations of GENBLOSUMxx for single or multiple BLOCK FASTA files. The program provides two outputs such as BL Prop<sub>av</sub> and Matrix<sub>av</sub>. BL: BLOCK, Prop: Properties, av: average**
The program first checks whether the input sequences are in FASTA FORMAT and are in BLOCK-FORMAT (i.e. the widths of sequences are identical). If true for both (i.e. **YES**), the program starts analysis of BLOCKs present in the current directory (7 BLOCKs in the present case; Figure 2, part-A). Upon completion of analysis, two outputs are provided, one for BLOCK properties and the other for bit-score matrix with identity series (i.e. **xx**). The above operations are performed either for single or multiple BLOCKs present in the current directory.

The output file is provided both on the screen and in excels form. While the former provide live update of run-status to the user, the later is useful for post-run analysis and presentation. In the output two subsections are there: one for the average matrix with series and the other for BLOCKs properties.

**Equations used by the Program**

Observed probability of occurrence of each $i,j$ pair is

$$q_{ij} = \frac{f_{ij}}{\sum_{i=1}^{20} \sum_{j=1}^{i} f_{ij}}$$

Let the total no. of amino acid $i,j$ pairs ($1 \leq j \leq i \leq 20$)

Expected probability of occurrence $e_{ij}$ for each $i,j$ pair is then $pipj$

For $i = j$ and $pipj + pjpj = 2pipj$ for $i = j$.

$$p_i = \left( q_i + \sum_{j \neq i} \frac{q_{ij}}{2} \right) ;$$

$$e_{ij} = p_ip_j \ (i = j), e_{ij} = 2p_ip_j \ (i \neq j)$$

Log odd Score is

$$\text{OddScore : } s_{ij} = 2 \cdot \left( \log_2 \left( \frac{q_{ij}}{e_{ij}} \right) \right)$$

**Preparation of artificial BLOCK for simulation**

We have prepared artificial BLOCKs for the simulation purpose using our In-house program GENBLOSUMxx, keeping BLOCKs parameter such as width and length in mind. BLOCKs are prepared manually for simulation keeping many points in mind such as keeping length of the column fixed and varying the width and vice-versa.

We have made many schemes; some of them are as follows:
(a) Sequence having width=1 & length=20[W=19 & S=1(fixed)]

>1
W
>2
W
>3
W
>4
W
>5
W
>6
W
>7
W
>8
W
>9
W
>10
W
>11
W
>12
W
>13
W
>14
W
>15
W
>16
W
>17
W
>18
W
>19
W
>20
S

(b) Sequence having width 150 & length 20(fixed)

>1
WHLADKPRVMIEFCYSTGNQMREYSNHKLVPGQAICTDFQNGTSYCFEIMVPRKDALHVPGQAICTD
FQNGTSYCEFIMVTRPGQYDCSNFHAEMIVLKYHRPQLMGCTNIKAYFRVQMPKIREFGYARIHM
NQVLVIPAMADHLINGKRC
>2
Result and discussion

Development of the Program

With the advent of automated sequencing procedure, database complexity increases. BLOSUM series (45, 62 etc) of matrices are obtained from sequence sets of varying identity (such as ≥45%, ≥62% etc). However, if we make exclusive choice of sequences from clade/genotypes in phylogenetic tree and produce bit-score matrices for each set; their comparison might reveal novel insight into amino acid substitution mechanisms. Further, two set of sequence BLOCKs procured from two different sources (such as mesophilic vs. thermophilic or eukaryotic vs. archaeal etc) could be analyzed for obtaining average BLOSUMxx matrices. The differential, if any, in these matrices might reveal fundamental differences between these sources. With this broad perspective we have developed GENBLOSUMxx that not only produces BLOSUM matrix of series xx from any number of BLOCK-FASTA (Lipmann and Pearson 1985), files but also provides detailed BLOCKs properties for the simulation in this study.
**Effect of Width on Block Parameters**

With the increment of width along the BLOCK, the parameter such as observed probability and expected probability both decreases (figure 1 A and B) but the rate of decrease in case of observed probability is lesser than expected probability. The BLOSUM score which is used by many popular web based software also affected by the increment of width (figure 1C).

![Graphs showing the effect of width on block parameters](image)

**FIGURE 1: Effect of width on A) observed probability B) expected probability C) bit score D) ratio**

As the width of BLOCK increases the Bit score also increases simultaneously. Similarly the ratio of the Observed probability (OP) and expected probability (EP) also linearly increases (figure 1D).

**Effect of Length on Block Parameters**

When the length of a BLOCK increases, the BLOCK parameters such as Observed probability (OP), Expected probability (EP), Bit score and the ratio of Observed probability (OP) and expected probability (EP) also affected.

Observed probability (OP) and expected probability (EP) decreases with the increment of length of the BLOCK (Figure 2 A and B). Here in contrast with the case of increment of
width, the Bit score and the ratio of Observed probability (OP) and expected probability (EP) also decreases. The reason of their decrease might be as the rate of decrease of Observed probability (OP) is

![Figure 2: Effect of length on A) observed probability B) expected probability C) bit score D) ratio.](image)

More than the decrease of expected probability (EP). In Both the cases of increment of length and width of the BLOCK, the parameters of the BLOCK are affected.

**Effect of types of amino acid or biased amino acid**

Figure 3 (A1, B1 and C1) show observed probability (OP) and Score vary at the same rate while expected probability (EP) remain constant. This situation arises when a type of amino acid changes in a column.

Figure 3 (A2, B2 and C2) show expected probability (EP) and Score varies while observed probability (OP) remain constant. This situation arises when a type of amino acid change in a row.
Conclusion

We have developed GENBLOSUMxx, a program interpreted by AWK programming language that can be executed both from C- and B-Shells for obtaining BLOSUM matrix, and important BLOCKs properties. The program is efficient in that it takes any number of BLOCK-FASTA files of any length and width of sequences as input to redirect output both in STDOUT and excel forms. While the former output is for live update of run-status the later is in ready post-run usable format. The automated BLOSUM series determination is another attribute of the program wherein it is computed from the average of BLOCK specific least of all calculated identity scores. The current version of the program provides three-point information, the knowledge of which has potential in the field of molecular evolution (R Grantham 1974), in relation to amino acid substitution. The program is basically designed for the simulation of BLOCK parameters to check its effect on score. Interestingly, our simulated result suggests that amino acid scores are not constant; rather vary with the simulated parameter. For example score for a pair of amino acid vary linearly with the width of block whose correlation co-efficient is almost one. Thus, our analysis shows that that BLOSUM series of matrices are not universal in nature. Development of new score matrices is an urgent need. Further, we are keenly working for
state of the art version of the program GENBLOSUMxx along with our other developed software PHYSICO (Gupta et. al. 2014), SBION2 (Gupta et. al. 2015), PHYSICO2 (Banerjee et. al. 2014) and SBION (Gupta et. al. 2014), based on comprehensive knowledge of the field, technical implementation of GUI and integrated online server support for output to worldwide users.

References

CHAPTER-5
Screening of Disease Associated Mutations in HGO proteins and its Structural Consequences using Computational Methods

Summary
Alkaptonura (AKU) is a rare genetic disorder caused by the mutation and impaired function in homogentisate dioxygenase (HGO), an enzyme required for the catabolism of phenylalanine and tyrosine. Among frame shift, intronic, splice-site and Missense mutations latter in the HGO gene are the most common forms of genetic variations that produces AKU. In this analysis, we implemented a computational approach to filter out the most deleterious mutations on Homogentisate-1, 2-dioxygenase protein that might be associated with the disease. We found W60G, A122D and V300G (out of 65 mutations) as most deleterious and disease associated by using SIFT, SNAP, PANTHER, SDM, PHD-SNP, Meta-SNP, Pmut and Mutpred tools. To understand the atomic arrangement in 3D space, the native and mutant (W60G, A122D and V300G) structures were modeled. Molecular docking and molecular dynamics simulation (MDS) approach was used to investigate the structural and functional behavior of HGO protein upon mutation. MDS and docking results showed stability loss in mutant HGO protein. Due to mutation, HGO protein became more flexible and alters the dynamic property of protein which might affect the interaction with target peptide. The study adopt a well designed computational methodology to analyses the disease associated nsSNPs and their molecular mechanism. The results obtained from this study would facilitate wet-lab researches to develop potent drug therapies against Alkaptonurea.

Introduction
Alkaptonurea is a rare metabolic disease (Garrod et. al; 1902, 1908) in which homogentisate, an intermediary product in the phenylalanine catabolic pathway, cannot be
further metabolized resulting into the causes of homogentisic aciduria, ochronosis, and arthritis (La Du et. al; 1995). It is one of the first conditions in which Mendelian recessive inheritance was proposed and also one of the inborn errors of metabolism (Garrod et. al; 1902). The manifestations of the disease are urine that turns dark on standing due to homogentisate, which can be spontaneously oxidized to a black pigment and alkalinization due to excretion of excessive amounts of homogentisic acid, large joint arthritis and black ochronotic pigmentation of cartilage and collagenous tissue. The homogentisate dioxygenase (HGO) gene provides instructions for making an enzyme called homogentisate oxidase. This enzyme helps break down of the amino acids phenylalanine and tyrosine, which are important building blocks of proteins. Mutations in the HGO gene impair the enzyme's role in this process and as a result, a substance called homogentisate, which is produced as a broken down of phenylalanine and tyrosine, accumulates in the body. Excess of homogentisate and related compounds are usually deposited in connective tissues and the main case of darkening of cartilage and skin. Over time, a deposition of homogentisate in the joints of a patient leads to arthritis. This is an inherited disease in which is carried as an autosomal recessive pattern, means both copies of the gene in each cell have mutations. The parents of an individual with an autosomal recessive condition do not show signs and symptoms of the condition but each carry one copy of the mutated gene.

The single nucleotide variations in the genome that occur at a frequency of more than 1% are referred to as single nucleotide polymorphisms (SNPs). In the human genome SNPs occur in just about every 3000 base pairs and the frequency of occurrence of the different alleles differs in different populations. Amino acid substitution can have many consequences on protein stability, activity, folding and splicing. It may also disrupt protein binding sites or ligand-binding pockets that are critical in the function of protein. To date as far our best knowledge stability, activity and pathogenicity of the SNPs that may modify the structure and function of the HGO gene is not clearly understood. Experimentally it is very difficult and time consuming and over the past few years, in silico studies have been improved significantly in screening the functional SNPs and predicting the conformational changes upon single amino acid substitution in proteins.
Therefore, to explore possible associations between substitution and phenotypic variations, different computational tools and algorithm like Sorting Intolerant from Tolerant (SIFT) (Ng and Henikoff, 2003), PHD-SNP (Capriotti et al., 2006), SNAP (Bromberg et al., 2008), PANTHER (Thomas et al., 2003), Site Directed Mutator (SDM) (Worth et al., 2011), Meta-SNP (Capriotti et al., 2013), Pmut (Carles et al., 2005) and Mutpred (Li et al., 2009) were used for the prioritization of high-risk non-synonymous mutations in coding regions that are likely to have an effect on the function and structure of the protein. Our analysis suggests that, P230S, V300G, M368V, G161R are most prevalent AKU mutations among 21. So, we compared the 3D structure model of above mutations with the native protein structure. We further examined the native and mutant protein structures for solvent accessibility and secondary structure analyses. Our in-silico study further suggests the presence of additional deleterious mutations in HGO gene that may affect the structure and function of proteins with apparent roles in Alkaptonuria.

Figure 1: Flowchart describing the workflow implemented in this study
Materials and method

Collection of Dataset
The protein sequence and SNP variant (total of 65 variants) information for our in-silico analysis were obtained from swissprot/UniProt database (Yip et. al, 2004).

Molecular modeling and Dynamic Simulation
BLASTp web server was used for the search of template. The template with PDB id 1EY2_A was used in this work for the generation of model. The alignment between the template and target sequence is performed using Clustal Omega (Sievers et. al, 2014). The model is generated using the popular and widely used program Modeler 9v11 (Webb and Sali, 2014). Five 3D models were obtained and form that we have selected the best modeled judged using Dope Score. The selected best model (wild) was then subjected for mutations (W60G, A122D and V300G) using SPDB viewer (Guex and Peitsch, 1997). The wild and mutated 3D structures were then energetically optimized Conjugate gradient energy minimization scheme of NAMD (NAnoscale Molecular Dynamics) simulation package (Phillips et. al, 2005) in presence of appropriate explicit water box covering the protein in full. The minimization was performed for 5000 steps with collection of frame at the interval of 200 steps and thus a total of 25 frames were collected. The best frames with lowest potential energy for each of the three structures were selected. The quality of model structures were validated and verified using PROCHECK, ANOLEA, ERRAT and VERIFY 3D in SAVES v4 web server (https://services.mbi.ucla.edu/SAVES/). Molecular Dynamic Simulation (MDS) was performed by using NAMD (NAnoscale Molecular Dynamics) simulation package (Phillips et. al, 2005). Structures of wild and mutant HGO protein were used as starting point for MDS. The PSF were generated using VMD (Humphrey et. al, 1996) autoPSF Plugin. System was solvated in a water box of radius 10 Angstrom. After this, whole molecular system was subjected to energy minimization for 5000 iterations by steepest descent algorithm implementing CHARMM force field. Berendsen temperature coupling method (Berendsen, 1984) was used to regulate the temperature inside the box. Electrostatic interactions were computed using the Particle Mesh Ewald method.
(Essmann et. al, 1995). Finally, the systems were subjected to MD simulation for 2ns. We then computed the comparative analysis of structural deviations in native and mutant structure. RMSD, RMSF, SASA and Rg were carried out by using VMD and scripts.

**Disease associated SNP prediction**

The SNP occurring in the protein coding region may lead to the deleterious consequences and might disturb its 3D structure. Here we used (SIFT) (Ng and Henikoff, 2003), PHD-SNP (Capriotti et. al, 2006), SNAP (Bromberg et al., 2008), PANTHER (Thomas et. al, 2003), Site Directed Mutator (SDM) (Worth et. al, 2011), Meta-SNP (Capriotti et. al, 2013), Pmut (Carles et. al, 2005) and Mutpred (Li et. al, 2009) tools in order to examine the disease associate mutations. SIFT prediction is based on the sequence homology and the physicochemical properties of amino acids which are identified by the substituted amino acid. SIFT score ≥0.05 indicates the amino acid substitution is deleterious, whereas the score ≤0.05 is predicted (Ng and Henikoff, 2003). PANTHER program predicts the frequency of occurrence of amino acid at a particular position in evolutionary related protein sequences. The threshold subPSEC score of −3 has been assigned below which the predictions are considered as deleterious (Thomas et. al, 2003). PhD-SNP is SVM based tool, trained over the million amino acid polymorphism datasets using supervised training algorithm [Capriotti et. al, 2006]. It predicts whether the given amino acid substitution leads to disease associated or neutral.

Site Directed Mutator (SDM) is a statistical potential energy function that uses environment-specific amino-acid substitution frequencies within homologous protein families to calculate a stability score, which is analogous to the free energy difference between the wild-type and mutant protein (Worth et. al, 2011). We filtered the Missense mutations that were combinedly predicted to be deleterious and damaging from Pmut (Carles et. al, 2005) and Mutpred (Li et. al, 2009) tools. PMUT allows the fast and accurate prediction of the pathological character of single point amino acidic mutations based on the use of neural networks (Carles et. al, 2005). MutPred is a web based tool, used to identify the molecular cause of disease related amino acid substitution (Li et. al,
It utilizes several attributes related to protein structure, function, and evolution. Thus by combining the scores of the servers, accuracy of prediction increases.

**Cation-\(\pi\) sites and stabilizing residues**

Cation-\(\pi\) interactions were computed using CaPTURE pro-gra(m (Gallivan and Dougherty, 1999). The CaPTURE program identifies energetically significant cation–\(\pi\) interactions within proteins in the Protein Data Bank (PDB). Cation–\(\pi\) interactions in protein structures are identified and evaluated by using an energy-based criterion for selecting significant side chain pairs.

**Identifying stabilizing residues**

Sride tool (Magyar et. al, 2005) was used to identify the stabilizing residues in HGO protein. The calculation is done by using four essential criteria which are residue hydrophobicity, the long-range order, the stabilization center by considering contact map of a protein and the conservation scores of residues. Two residues are considered in contact, if there is at least one pair of heavy atoms with a distance less than the addition of the van der Waals radii of the two atoms plus 1.0 Å. Further a contact is long-range if they are separated by at least ten residues in the amino acid sequence. Two residues are considered as SC elements when they are involved in long-range contacts and if at least one supporting residue can be found in each of the flanking tetra-peptides, in such a way that minimum of seven out of the possible nine interactions are formed between the two triplets. Combining all of mentioned criteria a stabilizing residue is reported.

**Free Energy Calculation**

Free energy was calculated using DFIRE (distance-scaled, finite, ideal-gas reference state) method (Yang et. al, 2008). They found correlation coefficients between experimentally measured protein-ligand binding affinities and those predicted by the DFIRE energy function are around 0.63 and also found that the function makes highly accurate predictions of binding affinities of protein complexes (Zhang et. al, 2005).
Molecular docking approach

We used PatchDock for docking native and mutants of HGO with target optimized Homogentisate. PatchDock is reliable as it performs docking based on molecular shape representation, surface patch matching in addition filtering and scoring (Duhovny et. al, 2002). Further its fast transformational search, which is driven by local feature matching rather than brute force searching of the six dimensional transformation spaces, makes it more significant. Moreover it speeds up the computational processing time by using advanced data structures and pattern detection techniques. After docking, it provides many complex structures based on their docking scores. The complex structures with the best docking score were selected for their further analysis. The best complex structures were further refined using FireDock web server (Mashiach et. al, 2008). FireDock is a web server which refines and rescores the rigid-body protein-protein docking solutions. The higher negative value of total interaction energy implies better interaction and vice-versa.

Result and Discussion

Prediction of deleterious nsSNPs

Out of 65 inputs Missense SNP dataset, 53 nsSNPs were found to be disease associated with the SIFT (Table1). In case of SDM server the affected Missense mutations are 21 out of 65 (Table1). 42 out of 65 Missense mutations are predicted to be disease associated with the PANTHER and 58 out of 65 in case of PHD-SNP. 56 and 57 out of 65 with SNAP and META-SNP respectively (Table1). Among these 65 Missense mutations, 15 nsSNPs (W60G, L61P, L116P, C120W, A122D, G123R, G161R, R187G, R197G, I216T, G217W, F227S, V300G, R321P and R330S)

Table 1 Deleterious and damaging nsSNPs predicted by six computational methods SIFT, SDM, PANTHER, PHD-SNP, SNAP and META-SNP in HGD gene

<table>
<thead>
<tr>
<th>Variant</th>
<th>SIFT Effect</th>
<th>SIFT Score</th>
<th>SDM Effect</th>
<th>SDM DDG value</th>
<th>PANTHER Effect</th>
<th>PANTHER Score</th>
<th>PHD-SNP Score</th>
<th>SNAP Effect</th>
<th>SNAP Score</th>
<th>META-SNP Effect</th>
<th>META-SNP Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>E3A</td>
<td>N</td>
<td>0.530</td>
<td>N</td>
<td>-0.38</td>
<td>N</td>
<td>0.212</td>
<td>N</td>
<td>0.113</td>
<td>N</td>
<td>0.265</td>
<td>N</td>
</tr>
<tr>
<td>E13K</td>
<td>N</td>
<td>0.200</td>
<td>N</td>
<td>-0.86</td>
<td>N</td>
<td>0.158</td>
<td>N</td>
<td>0.539</td>
<td>N</td>
<td>0.520</td>
<td>N</td>
</tr>
<tr>
<td>D18N</td>
<td>D</td>
<td>0.000</td>
<td>N</td>
<td>-0.10</td>
<td>N</td>
<td>0.270</td>
<td>N</td>
<td>0.310</td>
<td>N</td>
<td>0.415</td>
<td>N</td>
</tr>
<tr>
<td>L25P</td>
<td>D</td>
<td>0.000</td>
<td>N</td>
<td>-1.44</td>
<td>N</td>
<td>0.252</td>
<td>D</td>
<td>0.793</td>
<td>D</td>
<td>0.635</td>
<td>D</td>
</tr>
<tr>
<td>Q33R</td>
<td>N</td>
<td>0.090</td>
<td>N</td>
<td>0.10</td>
<td>N</td>
<td>0.145</td>
<td>D</td>
<td>0.520</td>
<td>N</td>
<td>0.470</td>
<td>N</td>
</tr>
<tr>
<td>E42A</td>
<td>D</td>
<td>0.000</td>
<td>N</td>
<td>-0.14</td>
<td>D</td>
<td>0.697</td>
<td>D</td>
<td>0.837</td>
<td>D</td>
<td>0.775</td>
<td>D</td>
</tr>
</tbody>
</table>
are screened out those are predicted to be diseases associated with all the six predicting tools (SIFT, SDM, PANTHER, PHD-SNP, SNAP and META-SNP).
These 15 nsSNPs (W60G, L61P, L116P, C120W, A122D, G123R, G161R, R187G, R197G, I216T, G217W, F227S, V300G, R321P and R330S) were further subjected for another level of screening with Pmut and Mutpred to find the most lethal Missense mutations. We have found the three Missense mutations (W60G, A122D and V300G) to be most lethal after screening with the above mentioned tools (Table2).

Table 2 The disease associated nsSNPs are predicted from Pmut and Mutpred servers

<table>
<thead>
<tr>
<th>Variant</th>
<th>Pmut</th>
<th>Mutpred</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Effect</td>
<td>Score</td>
</tr>
<tr>
<td>W60G</td>
<td>Pathological</td>
<td>0.7041</td>
</tr>
<tr>
<td>L61P</td>
<td>Neutral</td>
<td>0.1187</td>
</tr>
<tr>
<td>L116P</td>
<td>Neutral</td>
<td>0.1928</td>
</tr>
<tr>
<td>C120W</td>
<td>Pathological</td>
<td>0.7364</td>
</tr>
<tr>
<td>A122D</td>
<td>Pathological</td>
<td>0.7101</td>
</tr>
<tr>
<td>G123R</td>
<td>Pathological</td>
<td>0.6783</td>
</tr>
<tr>
<td>G161R</td>
<td>Pathological</td>
<td>0.6200</td>
</tr>
<tr>
<td>R187G</td>
<td>Pathological</td>
<td>0.8225</td>
</tr>
<tr>
<td>R197G</td>
<td>Pathological</td>
<td>0.7184</td>
</tr>
<tr>
<td>I216T</td>
<td>Neutral</td>
<td>0.4444</td>
</tr>
<tr>
<td>G217W</td>
<td>Pathological</td>
<td>0.8140</td>
</tr>
<tr>
<td>F227S</td>
<td>Pathological</td>
<td>0.5709</td>
</tr>
<tr>
<td>V300G</td>
<td>Pathological</td>
<td>0.5968</td>
</tr>
<tr>
<td>R321P</td>
<td>Pathological</td>
<td>0.7986</td>
</tr>
<tr>
<td>R330S</td>
<td>Pathological</td>
<td>0.7325</td>
</tr>
</tbody>
</table>

Mutations highlighted in bold has been predicted to show pathogenic property

These three screened Missense mutations (W60G, A122D and V300G) are further subjected to structural and functional analysis.

**Structural and functional analysis of HGD protein upon mutation**

To understand the structural and functional behavior of the screened disease associated mutations, we have performed molecular modeling and dynamics simulation for native and mutant HGD proteins. The template with PDB id 1EY2_A was used for the generation of model for the three variants (W60G, A122D and V300G). The structure validation of modeled protein suggests that the modeled HGD proteins have the quality of NMR structures. We further studied RMSD, RMSF, Rg and SASA analysis between
the native and mutant (W60G, A122D and V300G) HGD proteins to understand the structural variation.

RMSD for Wild and mutant Type Protein Structures

The MD simulation resulted in the generation of various plots from the trajectory analysis. One among them is the RMSD plot. The RMSD for the wild and mutant is calculated against the time simulation. The RMSD is a crucial parameter to analyze the equilibration of MD trajectories. It is estimated for backbone atoms by using the wild type and mutant structure of the MD simulations. The RMSD of the backbone atoms relative to the corresponding starting structures are calculated from trajectory.

In the wild type protein structure, we can notice that the average RMSD trajectory increases from 250 ps. The wild type protein structure shows deviations starting from 250 ps throughout the simulation time period. The RMSD plot of the wild type protein structure can be viewed in Figure 1.

In the mutant structures, the deviation in average RMSD can be seen for all the three variants with respect to Wild. The mutant protein W60G is seen to be decreasing near 500ps in contrast with A122D and V300G which increases, the increment in V300G is more. In case of V300G, decrease in RMSD is seen around 750 and 1500ps. After 1500ps decrease in RMSD is seen in W60G in contrast with wild, V300G and A122D. The RMSD of the trajectory is used to understand the stability of the protein. Higher RMSD value implies low stability of the protein structure.

Figure 1: Backbone RMSD are shown as a function of time for wild type (black) and mutants W60G (Cyan), A122D (Blue) and V300G (Red) at 300 K.
From the RMSD trajectory analysis of both wild and mutant protein structures, V300G has more RMSD than Wild protein throughout the trajectory that implies less stability of the mutant protein V300G than the native one. W60G shows higher stability during 600ps and after 1500ps than Wild protein. On the other hand A122D has less stability initially but the stability increases toward the end of trajectory.

RMSF for Wild and mutant Type Protein Structures

With the aim of determining and understanding whether the mutation affects the dynamic behavior of residues, the RMSF values of wild type and mutant structures were computed and plotted (Figure 2).

![RMSF of the backbone CAs of C atoms of wild type α (black) and mutant W60G (Cyan), A122D (Blue) and V300G (Red) at 300 K.]

The RMSF with respect to the average MD simulation trajectories is used as a mean of describing flexibility among residues. The backbone RMSF of each residue of wild type and three mutants are computed in order to analyze the flexibility of backbone structure. Higher RMSF value shows more flexible movements whereas low RMSF value implies limited movement during simulation.

The RMSF range of the wild type and mutant protein structure is between 0.09 nm and 0.8 nm. High fluctuation in the Wild protein structure can be seen throughout the MD
simulation time period. High fluctuation can be noticed at residues positions 50, 98, 148, 230, 248 and after 400. All the Mutant proteins (W60G, A122D and V300G) have high fluctuation than Wild type. W60G has high fluctuation at residues positions 46, 60, 210, 298, 366 and after 400. In A122D high fluctuation can be noticed at residues positions 56, 78, 122 and after 400. In case of V300G high fluctuation can be seen at residues positions 50, 98, 380 and after 400 but with highest fluctuation at reside position 300.

From the RMSF plot in Figure 2, we can notice that the wild type protein structure shows very less fluctuations at almost all residual positions when compared to the mutant protein structure. The terminal residues of both the protein structures show high fluctuations. The RMSF result shows that the mutant protein structures are more flexible i.e. less stable than the wild type structure.

**Radius of Gyration (Rg) for Wild and mutant Type Protein Structures**

Rg is performed generally in order to understand the levels of compaction of the native and mutant class of protein. The Rg is defined as the mass weighted root mean square distance of a collection of atoms from their common center of mass. Hence, this trajectory analysis for radius of gyration provides us the overall dimensions of the protein. The Rg range of the wild type protein structure is between 2.38 nm and 2.47 nm (figure 3).

![Figure 3: Radius of Gyration (Rg) are shown as a function of time for wild type (Red) and mutants W60G (Blue), A122D (Yellow) and V300G (Green) at 300 K.](image-url)
We can notice decrease of Rg in both Wild type and mutant protein throughout the simulation time. In wild protein, at 400ps there is slight increase in Rg to 2.475nm thereafter it continuously decreases and at 1250ps it again increases and then decreases till the end. In case of A1122D the Rg increases at 100ps to 4.9nm and then decreases to again increase in 800 and from 1500ps to till end of the simulation. Continuous decrease till 400ps and then increase from 500ps to 1500ps can be seen in W60G after that it decreases till the end. In V300G the Rg started from low 2.42nm and shows similar pattern like Wild type at 400ps and rest of the pattern is low but similar to A122D. From the Rg plot (figure3) we can say that the wild type protein is more stable than the mutants.

**Solvent Accessible Surface Area for Wild and mutant Type Protein Structures**

Solvent Accessible Surface Area (SASA) is to understand the solvent accessibility of the wild type and mutants HGO structures. SASA plot accounts for surface area of biomolecules that is assessable to solvent molecules i.e. the surface area of a protein which interacts with its solvent molecules. The increase in SASA value denotes relative expansion.

The SASA range of wild type of protein structure is in between 254 to 270 nm$^2$ (Figure4).

From the start of the simulation, the drop in the SASA value for both Wild type and mutant proteins (W60G, A122D and V300G) can be noticed. The fluctuation of SASA in 300ps and 1600ps can be seen in Wild type proteins. In W60G the fluctuation can be seen at 400ps. In case of A122D the fluctuation is lesser and in case of V300G the fluctuation can be seen at750ps and 1500ps. Overall, the wild type protein has decreased SASA values implies its more stability than the mutant protein structures.
Molecular Docking & binding energy calculation

From the molecular dynamics approach and trajectory analysis we examined that, due to mutation in HGO, protein lost stability and thus became more flexible. This flexibility can alter the binding phenomenon of protein with ligand molecule. To validate this, we further applied molecular docking approach to evaluate the interaction between the Wild type protein with ligand (homogentisate) and similarly mutant proteins (W60G, A122D and V300G) with the ligand for comparison (Table 3).

In depth analysis of the docking complex reveals notable features. Calculation of binding energy is very important to understand the affinity level of protein and ligand partners. Total binding energy of the complex mainly constitute of attractive and repulsive van der Waals interaction energy, atomic contact energy (ACE), hydrogen bond (HB) and area (interfacial surface area of the complex) between HGO (native and mutant) and ligand (homogentisate).

![Figure 4: Solvent accessible surface area (SASA) of wild type (Red) and mutants W60G (Blue), A122D (Yellow) and V300G (Green) are shown as a function of time at 300 K.](image)

Table 3. Molecular docking for the calculation of binding energy and its component of the Wild type and mutants (W60G, A122D and V300G) proteins

<table>
<thead>
<tr>
<th>Protein</th>
<th>Binding Energy Kcal/mol</th>
<th>Attractive vDW</th>
<th>Repulsive vDW</th>
<th>Atomic Contact Energy (ACE)</th>
<th>Hydrogen Bonds (HB)</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild</td>
<td>-22.12</td>
<td>-11.11</td>
<td>1.70</td>
<td>-4.77</td>
<td>0.00</td>
<td>338.20</td>
</tr>
<tr>
<td>A122D</td>
<td>-22.79</td>
<td>-11.21</td>
<td>0.13</td>
<td>-4.29</td>
<td>0.00</td>
<td>337.70</td>
</tr>
<tr>
<td>W60G</td>
<td>-22.22</td>
<td>-11.28</td>
<td>1.68</td>
<td>-4.50</td>
<td>0.00</td>
<td>337.70</td>
</tr>
<tr>
<td>V300G</td>
<td>-26.18</td>
<td>-11.95</td>
<td>0.05</td>
<td>-5.51</td>
<td>0.00</td>
<td>303.20</td>
</tr>
</tbody>
</table>
Less interaction has been found in Wild type when compared to mutants and this was due to more interfacial surface area and less contribution of hydrogen bond as shown in (Table3). Increase in the binding energy of the mutants (Table3) i.e. -22.22, -22.79 and -26.18 for W60G, A122D and V300G respectively than Wild type (-22.12) suggests that the flexibility of the mutant protein increase. Upon mutation the Wild type protein becomes less stable as it gains flexibility.

To examine the reason behind the loss of stability upon mutations, we further calculated the Cation-π interactions, stabilizing residues and free energy. Cation–π interactions were known to be important contributors to the stability of a protein. We observed a total of five energetically significant Cation–π interactions and total free energy (Table 4). DFIRE Total energy of mutants is -934.34, -929.32 and -875.31 for W60G, A122D and V300G respectively whereas Wild type has -931.87 of energy.

From table 4 it is observed that number of Cation–π interactional residue is five for both wild type and mutant protein implies it doesn’t change upon mutation in the wild
protein. But total Cation–π interactional energy increases in case of mutant protein i.e. -26.64, -26.62 and -27.01 for W60G, A122D and V300G respectively Whereas Wild type of protein has -26.53 of Cation–π interactional energy. This result also suggests that Wild type of protein loses its stability upon mutations.

**Conclusion**

Bioinformatics and Computational analysis has now increasingly become a modern and significant way to characterize and predict disease associated nsSNPs at atomic level. Using multiple in silico methods and molecular docking & dynamics simulation approaches, we identified the HGO (W60G, A122D and V300G) mutations as highly deleterious as well as their Structural and functional consequences. The loss of stability is clearly observed from all of or results and analysis in mutant (W60G, A122D and V300G) structure. Due to mutation, HGO protein became more flexible in nature, which is well supported by our analysis. This might modulate the structural orientation and binding phenomenon of the HGO protein as observed by molecular docking and energy calculation. Our adopted analysis suggests an important template to identify and characterize the disease associated mutation and its structural and functional consequences in HGO protein. The insight gain from this study of mutation (W60G, A122D and V300G) in HGO protein might helps researchers to develop drug therapies for HGO associated diseases.

**References**


Sievers F, Higgins DG. Clustal Omega, accurate alignment of very large numbers of sequences. Multiple sequence alignment methods. 2014:105-16.


CHAPTER-6

Molecular Modeling of Synthesized Novel N-acyl Substituted Benzoazetinones and Their Docking Simulation on Antifungal Modeled Target.

Summary

Stable N-acyl substituted benzoazetinones derivatives (series of 3a to 3i) were optimized using quantum mechanics (QM) and docked into the modeled target protein P450, class of CYP53A15, a benzoate 4-monooxygenase that are abundantly found in the genome of ascomycetes and besidomycetes classes of pathogenic fungi. Low per residue RMSD of modeled structure of the enzyme and Ramachandran plot indicates similar topology as template (4D6Z.pdb). Interaction energy judges the site specific docking of quantum mechanically optimized benzoazetinone derivatives onto the target enzyme. These results suggested that compound 3i having high negative score is best antifungal agent among others. The hydrophobic substituent present in benzoazetinone derivatives facilitated the stability of ligand-target complex. Overall, the studies further provide insight into the specificity of the interactions and thereby facilitating the development of broad-spectrum antifungal agents against these opportunistic and infectious fungi.

Introduction

The basic structural core of benzoazetinones is β-lactam fused with benzene ring. β-lactams are of considerable interest not only having a long and illustrative history of antibacterial activity (Ceric et. al., 2010) but also because of their interesting structural features, high reactivity and wide use as important starting precursors for the synthesis of biologically active various organic compounds (Bose et. al, 2000). Moreover, the importance of crucial β-lactam antibiotics and β-lactamase inhibitors has much attracted researchers to synthesize a variety of new β-lactams (Buynak et. al, 1951). Besides, the clinical applications of β-lactams as therapeutic compounds for lowering plasma
cholesterol (Burnett et al., 2004) and human leukocyte elastase inhibitory mechanisms of β-lactams have also been studied (Finke et al., 1995).

Azetidinones as synthones are known to act as potential anti fungal agents (O’Driscoll et al., 2006). However, fungal resistance to almost all classes of commercially available antifungal drugs is well known (Kanafani et al., 2008). Thus, discovery of new antifungal agents are of interest against different invasive fungal infections (Andriole et al., 2000) caused by *Candida* (Hart, 1969), *Aspergillus* (Abdin et al., 1998), *Colchilobus* (Badiee et al., 2009) and Immuno-compromised patients due to cancer chemotherapy, solid organ transplant and HIV infections (Masubuchi et al., 2003). Antifungal activity of amines, amides and azoles derivatives against the target cytochrome P450, CYP families has been investigated in details (aha et al., 2012). Different classes of cytochrome P450 (i.e. CYPs) are highly diverse in fungal genome (Crešnar et al., 2011) especially in pathogenic *Aspergillus* and *Gibberella*. For this reasons, fungal CYPs are targets of antifungal infection against animal (including human) and plant (Kelly et al., 2013). Previous study (Podobnik et al., 2008) also identified benzoate 4-monooxygenase (CYP53A15; benzoate para-hydroxylase; BPH) of the fungal pathogen Cochliobolus lunatus as a crucial enzyme involved in detoxification of benzoic acid (BA), and demonstrated that CYP53A15 activity was inhibited with natural phenolic compounds, such as isoeugenol, and revealed their antifungal potential.

We further used our synthetic compounds against the target CYP53A15 for gaining insight into the interactions at its active site. Therefore, we have selected CYP53A15 as a target for our synthesize compound which are benzoate like chemical in nature. Since the crystal structure of CYP53A15 from Cochliobolus lunatus is unavailable, we used the homology-build model of CYP53A1519 for molecular docking of the series of synthesized benzoazetinones derivatives of compounds.

**Material and Methods**

Optimization procedure for benzoazetinone
Chemical structure was drawn in Gauss View 3.0. Each of this structure was subjected to two successive optimization schemes. Semi-empirical PM3 was used first, followed by \textit{ab initio} HF/STO-3G* methods using Gaussian 03 package program. Lowest energy structure was stored for docking studies.

**Homology Modeling**

Three templates for CYP53A15 (UNIPROT ID B8QM33) were chosen using E-value and sequence identity criteria. Full length of 4D6Z_A (resolution 1.93 Å, UNIPROT ID P08684 and length 23-503) along with part of 1Z10_A (1.90 Å, UNIPROT ID P11509) and 4NY4_A (2.95 Å, UNIPROT ID P08684) were used to obtain the topology information for missing residues in the former using multiple template methods of Modeler 9v11 (Webb and Sali, 2014). Hetero atoms of template were transfer using advanced modeling method of Modeler. Five models were generated of which the best model was selected based on Discrete Optimized Protein Energy (DOPE) score. The model was refined using Conjugate gradient energy minimization scheme of NAMD (NAnoscale Molecular Dynamics) in presence of explicit water box for 5000 steps with an interval of 200 steps and thus a total of 25 frames were collected. Lowest potential energy frame was taken as final model.

Structural evaluation, validation and stereo chemical analyses of model was done by PROCHECK, ANOLEA, ERRAT and VERIFY in SAVES v4 web server (https://services.mbi.ucla.edu/SAVES/).

**Docking and virtual screening**

Autodock 4.2 (Morris, et. al, 2009) along with Autodock tools (Morris, et. al, 2009) of the Scripps Research Institute was used for docking and preparation of ligand and target protein respectively. To identify the interaction zone, we perform docking, at the site of one of the three heteroatom (Heme, GOL and PK9) of the template, at a time. The target (CYP53A15) and one of our compounds (3i) were prepared with MGLtools. Gasteiger partial charges were added. Non-polar hydrogen atoms were merged and rotatable bonds were defined. The site of the target intended for docking was placed in a grid box with
appropriate dimension and center along with 0.375 Å grid spacing. Torsions were allowed to the long side chains of the amino acid residues in the vicinity of the ligand. Docking simulations were performed using the Lamarckian genetic algorithm with default parameters. Initial position, orientation, and torsions of the ligand were set randomly. The best interacting site (Heme binding pocket of active site) thus obtained was then used for docking of all of our compounds (Table 1) using virtual screening procedure of Vina 1.1.2 (May, 2011).

**Result and discussion**

**Molecular Modeling of benzoazetinone**

Global optimizations of benzoazetinones were followed prior flexible docking of these compounds onto the active site (heme binding site) of the target-model, CYP53A15. A use of properly minimized three-dimensional structure is a necessity for effective assessment of flexible ligand-target interactions. We use semi-empirical followed by *ab initio* procedure of Gaussian 03 for the purpose. Two step procedures were necessary as only semi-empirical method was seen to be ineffective for meeting the convergence criteria (e.g. 3k; Table 1) of some of these compounds. Again, application of *ab initio* method as sole scheme, poses problem of meeting convergence criteria. Energies of each of these structures are very large and negative indicative of well-optimized global minimal structure. The energy of 3g is highly negative, may be due to substitution of R3 by Br (Table 1).

<table>
<thead>
<tr>
<th>Compounds</th>
<th>Structure</th>
<th>Energy</th>
<th>Docking score</th>
</tr>
</thead>
<tbody>
<tr>
<td>3a</td>
<td><img src="image" alt="Structure" /></td>
<td>-654.50</td>
<td>-28.56</td>
</tr>
<tr>
<td>3b</td>
<td><img src="image" alt="Structure" /></td>
<td>-892.67</td>
<td>-34.44</td>
</tr>
</tbody>
</table>

Table 1: Details of optimization and docking score of benzoazetinones.
| 3c | ![Image of 3c] | -542.10 | -28.14 |
| 3d | ![Image of 3d] | -580.66 | -31.51 |
| 3e | ![Image of 3e] | -654.50 | -28.98 |
| 3f | ![Image of 3f] | -770.23 | -31.54 |
| 3g | ![Image of 3g] | -3198.71 | -29.45 |
| 3h | ![Image of 3h] | -993.67 | -32.34 |
| 3i | ![Image of 3i] | -919.83 | -37.38 |
| 3j | ![Image of 3j] | -881.25 | -34.86 |
| 3k | ![Image of 3k] | -692.87 | -36.54 |

High negative scores for site-specific docking of this group of compounds against the model structure indicate well-formed complexes. The nature of the substituent at R3-site of ligand seems to contribute in the complex formation and hence the effectiveness in interactions. Bulky aromatic ring along with non-polar substitution at the site (Table 1) seen to produce more negative docking score (as seen in the case of 3i; Table 1). For example, compounds 3g to 3j differ only by R3 substitutions. Non-polar and polar nature of R3-substituent is more in 3i and 3g respectively. The former has highest and the later has lowest docking score (Table 1).
Homology modeling of antifungal target

Model structure of CYP53A15 is highly optimized and validated using 4-point criteria (Sen Gupta et. al, 2013, 2014) as judged by per residue RMSD comparison (Figure 1).

![Figure 1: Per residue RMSD of 4D6Z (Blue; Template used as reference), the present model (red line) and PM0075149 (green) Alignment of three structures was done using STRAMP structural alignment in VMD](image)

Accuracy in homology model depends on many factors such as choice of template, their alignment, loop optimization, and energy minimization of initial model as obtained by Modeller software. Overall, per residue RMSD of our model is better in comparison with the model PM0075149 generated in previous study (Podobnik, et. al, 2008) as shown in Figure 1.

The structure 4D6Z.pdb instead of 4LXJ.pdb was used as template. There are few advantages of the earlier than the latter. First, the former belongs to the same class as the query (CYP53A15). Both are 4-monoxygenase with the former has broad-spectrum substrate range. Second, query sequence has higher identity (26%) with 4D6Z_A than 4LXJ.pdb (19%). Third, 4LXJ.pdb although belong to same genera, it participates in demethylation reaction. Further, we use multiple templates as 4D6Z.pdb has missing residues. The missing region is modeled using the structural information of other templates (1Z10.pdb and 4NY4.pdb) keeping rest of the information from 4D6Z.pdb.
Docking studies

To identify and understand the potential antifungal agent, docking calculations and virtual screening were performed using Autodock 4.2 (Morris, et. al, 2009) respectively. Target molecule and Ligand were prepared using Autodock (Morris, et. al, 2009). The scores thus obtained are presented in Table 1. Figure 2 shows stabilizing interactions of CYP53A15 and one of these representative ligands (3i) at the heme-binding pocket. Ligand-protein interactions are largely stabilized by hydrogen bonding, van der waals, hydrophobic and π-alkyl interactions (Wang et. al, 2016). In the heme-binding pocket of the target, 3i is seen to establish number of stabilizing interactions that include i] amide-π-stacking with Ala284, ii] alkyl and π-alkyl interactions with Val429 and Ala430 respectively, iii] hydrogen bonding with Gly426, iv] covalent bond with Phe108 and v] van der Waals interactions with Thr288, Val425, Cys424, Phe108, Gln281, Leu255 and Glu433 (Figure 2).

Figure 2: 2D model of interactions between the ligand 3i with target protein; for clarity, only interacting residues are displayed. Presence of blue sphere around residues indicates accessibility.

Figure 3 (A) shows docking complex between the target protein and the ligand, 3i which has maximum interaction energy (Table 1). It is seen that the binding pocket is
appropriately fits by the ligand 3i with its hydrophobic group, MeC6H4 directed deep near the floor of the cavity. This region of the cavity is also seen to be hydrophobic (Figure 3 B; SAS scale). Interestingly the methyl group of MeC6H4 (R3 substituent) is also in contact with hydrophobic part of the pocket. Such hydrophobic stabilization is not possible with 3g, 3h and 3j structures that they have Br, OMeC6H4 and C6H5 groups respectively. Thus, additional stabilization might be related with the methyl group in 3i at its R3 site. Further, ligand structures 3a to 3f, all have hydrogen atom at R3 site (Table 1). The docking scores of these ligands are much lower than that of 3i except 3b. The R5 substituent of 3i and 3b are Me and C6H5 groups respectively. Overall, it seems hydrophobic nature of R3 and also R5 bring stabilization of the docking complex with higher effective contribution from the earlier.

![Figure 3: A, docking complex of model structure of CYP53A15 and geometry optimized ligand, 3i. B, ligand (red) binding pocket, interacting residues, and their solvent accessible surface (SAS) are shown along with SAS scale. C, a copy of ligand at identical orientation as it is present in the binding pocket to guess positions of invisible atoms. Residues labels written in blue color are not visible in the front view.](image)

**Conclusions**

In conclusion, we have modeled the target molecule and optimized the Stable N-acyl substituted benzoazetinones derivatives for molecular docking. These compounds undergo site-specific interactions in the heme site. 3i makes most effective interaction
might be due to hydrophobic stabilization and packing at the interaction site. The site is lined up with Alanine at close proximity. The score of interaction not only show these lists of compounds to be potential antifungal agents but also raising the possibility of development of potential antifungal agents specifically targeting CYP53.

References


List of Publications


➢ Symposium/Seminars Presentation

- **Parth Sarthi Sen Gupta** and Amal K Bandyopadhyay. Odd Amino Acid Substitution Matrices Are Really Odd. FIMB 2012, IISER Kolkata.

Burdwan.


- **Parth Sarthi Sen Gupta** and Amal K Bandyopadhyay. Alteration of Codes in the Common Shared Region of Halophilic Ferredoxin for its Adaptation in High Salt. 21st CRSI national symposium in chemistry with American Chemical Society (ACS). July 14-16, 2017