# Sequence Analysis to Predict Protein Active Sites using SSHM

**P.Satheesh**[1], **B.Srinivas**[2], **M.Prasada Rao**[3], **Col.Prof.Allam Apparao**[4], **and G.Charles Babu**[5]
[1] Associate Professor, CSE Department, MVGR College of Engineering, Vizianagaram, Andhrapradesh, India
[2] Assistant Professor, CSE Department, MVGR College of Engineering, Vizianagaram, Andhrapradesh, India
[3] Lecturer, CSE Department, JNTU kakinada, Kakinada, Andhrapradesh, India
[4] Vice-Chancellor, JNTU kakinada, Kakinada, Andhrapradesh, India
[5] Professor, CSE Department, Vidya Vikas Institute of Technology, Chevella, Andhrapradesh, India

**Abstract**— *The advancement of bioinformatics is remarkable after the analysis of human genome is completed. The functions of the protein coded from the genome are compared with the sequence of amino-acid of unknown proteins and the sequence of the protein that is already known. It can find the similar sequence, but the global relations among the sequences cannot be extracted. Initially two protein sequences are taken one among which is the known amino acid sequence(seq A) and the other one is the unknown amino acid sequence(seq B). By Sequence mapping , comparison is made between the two sets. This method searches for the matched and the frequent set of unknown amino acid in the known amino acid sequence. The algorithm takes protein sequence as input and does the mapping. If there are zero matches of seq B in seq A then there exists noise, which is to be eliminated.*

**Keywords:** Sequence Mapping, Amino-acid, Frequency Set, Sequence Search Hill Climbing Algorithm(SSHM)

## 1. Introduction

Proteins are made up of chains known as amino acids which are bind together by the peptide bonds. Protein structure is determined by the nucleotide sequence of that protein. Amino acids are made up of carbon, hydrogen, oxygen and nitrogen which will combine to form different types of proteins which are required by the body. PHP is a scripting language which is used for web development. PHP can run on any existing platforms based on the same code. Using PHP as the scripting language fastens the execution, moreover it supports several database and HTTP server interfaces. We implemented heuristic search based on Hill Climbing Algorithm in PHP. Heuristic search method is used to find the best solution in least possible time. Hill Climbing Algorithm is an Iterative algorithm which starts with a solution and then incrementally finds a better solution by changing a single element of the solution. The matched amino acid sequence is searched and frequency set is generated.

## 2. Scripting Languages

Specialised scripting languages include: PHP (Hypertext PreProcessor). It is popular scripting language which has more than thousand inbuilt function support. And has nested, associative arrays features.
Perl (Practical Extraction and Report Language). This is a popular string processing language for writing small scripts for system administrators and web site maintainers. Much web development is now done using Perl. newline Hypertalk is another example. It is the underlying scripting language of HyperCard.
Lingo is the scripting language of Macromedia Director, an authoring system for develop high-performance multimedia content and applications for CDs, DVDs and the Internet.
AppleScript, a scripting language for the Macintosh allows the user to send commands to the operating system to, for example open applications, carry out complex data operations.
JavaScript, perhaps the most publicised and well-known scripting language was initially developed by Netscape as LiveScript to allow more functionality and enhancement to web page authoring that raw HTML could not accommodate. A standard version of JavaScript was later developed to work in both Netscape and Microsoft's Internet Explorer, thus making the language to a large extent, universal. This means that JavaScript code can run on any platform that has a JavaScript interpreter.
VBScript, a cut-down version of Visual Basic, used to enhance the features of web pages in Internet Explorer.

### 2.1 PHP scripting

PHP is an HTML-embedded scripting language. Much of its syntax is borrowed from C, Java and Perl with a couple of unique PHP-specific features thrown in. The goal of the language is to allow web developers to write dynamically generated pages quickly."
This is generally a good definition of PHP. However, it does contain a lot of terms you may not be used to. Another way to think of PHP is a powerful, behind the scenes scripting language that your visitors won't see!
When someone visits your PHP webpage, your web server

processes the PHP code. It then sees which parts it needs to show to visitors(content and pictures) and hides the other stuff(file operations, math calculations, etc.) then translates your PHP into HTML. After the translation into HTML, it sends the webpage to your visitor's web browser.

### 2.2 Advantages of PHP

PHP is an extremely popular scripting language. It was originally created in 1995 and designed for the web. It is free of charge and can be used on almost every operating system. There are over 20 million websites and over 1 million web servers running PHP and those numbers are growing every day. The reason why PHP is so popular and it is continuously growing is because it offers many advantages. These are:
Fast - PHP was created to develop dynamic Web Pages so it is fast on websites. The PHP code is embedded in HTML and the time it takes to process and load the browser with HTML and create a full web page is very quick.
Free - PHP is released under the PHP License. This licence is compatible with the GNU General Public License or GPL. Thus making PHP free software. This means that anybody can download it and use it 100
Easy - The syntax of PHP is very easy to use and learn. PHP is usually mixed in with HTML and can be easily included in HTML files.

## 3. Materials and Methods

The unknown protein sequence (Q08392) is given as input to our algorithm .The known sequence which he used is (1ML6). Our algorithm does the Heuristic Search and generates the frequency set. The frequency set gives the pattern matches respective times. Algorithm is developed based on Hill Climbing algorithm and implemented in PHP. It is implemented in PHP script which generates the frequency set. We used Docking Tool to predict the protein structure.

### 3.1 SEQUENCE SEARCH HILLCLIMBING ALGORITHM (SSHM)

Step 1: Start
Step 2: Take Unknown Sequence Sk
Step 3:Take Known KSs Sequence set As a Search Space
Step 4:Apply Hill Climbing Technique to match Sk with KSs
Sk[i,j](Intersection)KSs[i,j]>Sk[i,k](Intersection) KSs[i,k]
Then Sk[i,j] (Intersection) KSs[i,j] is the Result
Step 5:If Matches are not found with KSs Then Sk is New Sequence
Step 6:Stop
Known Protein Sequence 1ML6:
AGKPVLHYFNARGRMECIRWLLAAAGVEFEEKFIQSPE
DLEKLKKDGNLMFDQVPMVEIDGMKLVQTRAILNYIA
TKYDLYGKDMKERALIDMYTEGILDLTEMIGQLVLCPP
DQREAKTALAKDRTKNRYLPAFEKVLKSHGQDYLVGN
RLTRVDVHLLELLLYVEELDASLLTPFPLLKAFKSRISSL
PNVKKFLQPGSQRKPPLDAKQIEEARKVFKF
Unknown Protein Sequence Q08392:
MSGKPVLHYANTRGRMESVRWLLAAAGVEFEEKFLEK
KEDLQKLKSDGSLLFQQVPMVEIDGMKMVQTRAILNY
IAGKYNLYGKDLKERALIDMYVEGLADLYELIMMNVV
QPADKKEEHLANALDKAANRYFPVFEKVLKDHGHDFL
VGNKLSRADVHLLETILAVEESKPDALAKFPLLQSFKAR
TSNIPNIKKFLQPGSQRKPRLEEKDIPRLMAIFH

SCRIPT:
<TABLE border=1>
<? php
$str1="AGKPVLHYFNARGRMECIRWLLAAAGVEFEEKFI
QSPEDLEKLKKDGNLMFDQVPMVEIDGMKLVQTRAIL
NYIATKYDLYGKDMKERALIDMYTEGILDLTEMIGQLV
LCPPDQREAKTALAKDRTKNRYLPAFEKVLKSHGQDY
LVGNRLTRVDVHLLELLLYVEELDASLLTPFPLLKAFK
SRISSLPNVKKFLQPGSQRKPPLDAKQIEEARKVFKF";
$str2="MSGKPVLHYANTRGRMESVRWLLAAAGVEFE
EKFLEKKEDLQKLKSDGSLLFQQVPMVEIDGMKMVQ
TRAILNYIAGKYNLYGKDLKERALIDMYVEGLADLYE
LIMMNVVQPADKKEEHLANALDKAANRYFPVFEKVL
KDHGHDFLVGNKLSRADVHLLETILAVEESKPDALAK
FPLLQSFKARTSNIPNIKKFLQPGSQRKPRLEEKDIPRL
MAIFH";
? >
<TR>
<TD>String length</TD>
<TD> <?php print_r(strlen($str1)); ?></TD>
<TD><?php print_r($str1); ?></TD>
</TR>
<TR>
<TD>String length</TD>
<TD><?php print_r(strlen($str2)); ?></TD>
<TD><?php print_r($str2); ?></TD>
</TR>
</table>
<table style="float:left" border=0>
<?php
$chars = array(”);
$chars1 = array();
$red= array();
$c=0;
$high=0;
for($l = 0; $l<=strlen($str2); $l++)
for($k = 0; $k<=strlen($str2)-$l; $k++)
$string = substr($str2,$l,$k);
//echo substr($str2,$l,$k). $l. ’->’. $k . "<br/>";
$chunk = substr($str2,$l,$k); if(strlen($chunk)>0)$cnt = substr_count($str1,$chunk);
if($cnt>0)
if(isset($red[$string]))
$c++;

```
if($cnt>$high)
$high=$cnt;
$red[$string]=$cnt;
}
}
}
}
}
foreach($red as $i => $value)
?>
<tr><td><b><?php print_r($i); ?></b></td>
<TD> <?php echo $red[$i]; ?> </TD>
<?php
echo " <TD >".(($red[$i]/sizeof($red))*100)."
for($j=0;$j<$red[$i];$j++)
echo " <TD bgcolor='blue'>.</TD> ";
echo "</tr>";

?>
<tfoot></tfoot>
</table>
<?php
echo sizeof($red)."<span style='float:left'><b>Total match count is ".$c." with highest frequency as ".$high."</b></span>";
?>
```

# 4. Results

The important of this study relates to the importance of dissimilar residues between any two proteins under study. In this case Q08392 and 1ML6 consider to prepare frequency table chart based on the designed algorithm. Owing to the importance of this analysis, a amino acid frequency chart representative of single and double amino acids are reported. The below tabulated values gives percentage matches with single and double amino acids matches.

From the above tables frequency of single amino acid residue between Q08392 and 1ML6 reported to contain 51.66% number of matches. The amino acid with highest frequency was found to be "Leucine"(L) with 32 number of matches (7.58%).Most of the percentage matches were in the range 1.42% to 2.6%.The basic amino acid "Histinde"(H) was at 0.71% .Considering the two amino acids matches ,most of the observes matches between these two proteins are not more than 2 to 3 matches with percentage number of matches been 0.5%. The overall success rate with single amino acid and double amino acid was 51.66% and 21.92% respectively.

# 5. Conclusions

The importance of this coding and subsequently results from this study is to emphasis the crucial amino acid residues responsible for functional attributes can be detected used molecular docking technology. In other words the presence of either single or double amino acids with in or nearer active site region leads to gain insights towards the functional relevance of Glutathione S transferase (GST).Hence a homology modeling protein was undertaken to build the protein and subsequently molecular docking studies are initiated.

Table 1: single and double amino acids matches with percentages

| Amino Acid | No. of matches | % of matches |
|---|---|---|
| M | 7 | 1.66 |
| S | 7 | 1.66 |
| G | 11 | 2.61 |
| K | 22 | 5.21 |
| P | 12 | 2.84 |
| V | 13 | 3.08 |
| L | 32 | 7.58 |
| H | 3 | 0.71 |
| Y | 8 | 1.91 |
| A | 16 | 3.79 |
| N | 6 | 1.42 |
| T | 8 | 1.90 |
| R | 13 | 3.08 |
| E | 17 | 4.03 |
| F | 10 | 3.77 |
| D | 14 | 3.72 |
| Q | 9 | 2.13 |
| I | 10 | 2.13 |
| | Success rate | 51.66 |

# References


[1] Aebersold R.H., Leavitt J., Saavedra R.A., Hood L.E., and Kent S.B. *"Internal amino acid sequence analysis of proteins separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose.*, Proc. Natl. Acad. Sci. Vol.84: pp.6970-6974, 1987.

[2] 2) Bergman T. and Jörnvall H., *"Electroblotting of individual polypeptides from SDS/polyacrylamide gels for direct sequence analysis."*, Europian Journal of Biochem. Vol.169, pp 9-12, 1987.

[3] online Manual on PHP available online, *http://php.net/manual/en/faq.general.php*