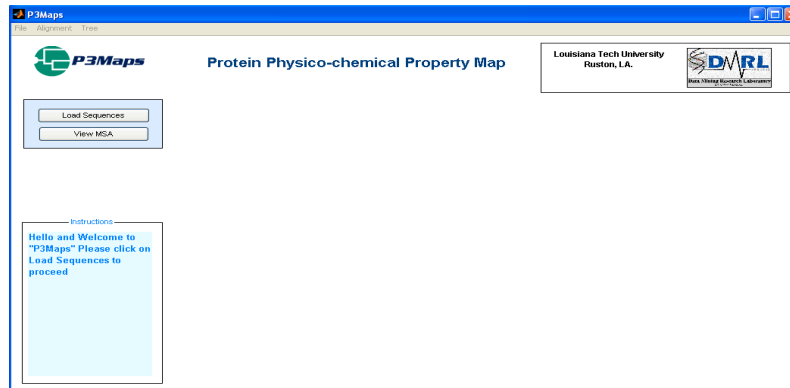


P3Maps: Protein Physico-chemical Property Maps



OVERVIEW

Proteins are not rigid bodies; they are flexible and constantly change shape and form to perform their biological roles. While we are intuitively aware of their constantly changing nature, we have little understanding of how the inter-residue interactions are encoded in the protein sequence and, therefore, have little understanding as to how they drive structural flexibility. To address this knowledge gap, we propose a tool to predict and analyze those regions over the primary sequence of the protein that are of functional importance using a myriad of physico-chemical properties.

Many protein regions (subsequences), over the body of a protein, are intrinsically conserved across related proteins. These intrinsically conserved regions are crucial to the function of many proteins, especially those involved in signaling, recognition, and regulation [1]. This study was motivated by several empirical observations relevant to protein structure, function, and evolution and by previous studies that addressed the relationship between the impairment of protein function and the resulting disease [2], [3]. There is also weak evidence that protein impairment and disease sensitivity are correlated with the physico-chemical difference between evolutionary constraint, functional impairment, and disease severity [2]. The goal of this work is to develop a (cyber)tool that overcomes the challenges inherent to the identification of those protein impairment and disease insensitivity characteristics, and functions of conserved regions within a protein by analyzing its effect under various physio-chemical conditions.

Our tool-based analysis assumes:

- (a) That evolutionary variation among orthologs in the affected position is a sample of the physico-chemical properties that are tolerated at that position and,
- (b) That correlated mutations of physico-chemical interactions between residues reveal evolutionary residue conservation patterns that reflect conserved structural domains. By using these two ideas as a premise, we develop the following tool.

We hypothesize that there is a correlated characteristic among the physico-chemical properties of a protein that can be used to predict functionally significant regions using machine-learning approaches. The proposed tool contributes to our understanding of the sequence and physico-chemical relationship and paves the way for us to identify local sequence property modulations that impact protein function without changing the protein structure.

PROPOSED METHODOLOGY

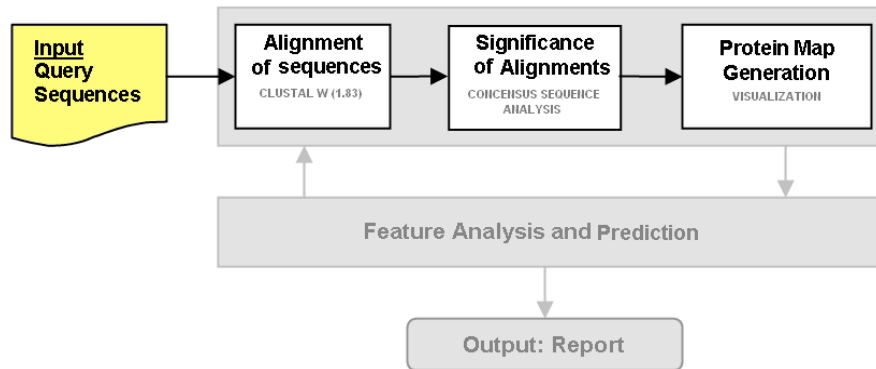


Fig.1. The proposed model.

TOOL FEATURES

Clustal W (1.83):

Clustal W is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. General documentation is available from <http://www.ebi.ac.uk/clustalw/>. The current tool uses Clustal W (version 1.83) to align multiple sequences provided through the input fasta format file. The input and output files both have the .aln extension.

Sample Input:

```

>1A0H:E|PDBID|CHAIN|SEQUENCE
IVEGQDAEVGLSPWQVMLFRKSPQELLCGASLISDRWVLTAAHCLLYPPWDKNFTVDDLVRIGKHSRTRYERKVEKIS
M
LDKIYIHPRYNWKENLDRDIALLLKLRPIELSDYIHPVCLPDKQTAAKLLHAGFKGRVGTWGNRRRETWTTSVAEVQPSV
L
QVVNLP LVERPVCKASTRIRITDNMFCAGYKPGEGKRGDACEGDSGGPFVVMKSPYNNRWYQMGIVSWGEGCDRDGKYGF
Y
THVFRLLKWIQKVIDRLGS
>1A0H:B|PDBID|CHAIN|SEQUENCE
IVEGQDAEVGLSPWQVMLFRKSPQELLCGASLISDRWVLTAAHCLLYPPWDKNFTVDDLVRIGKHSRTRYERKVEKIS
M
LDKIYIHPRYNWKENLDRDIALLLKLRPIELSDYIHPVCLPDKQTAAKLLHAGFKGRVGTWGNRRRETWTTSVAEVQPSV
L
QVVNLP LVERPVCKASTRIRITDNMFCAGYKPGEGKRGDACEGDSGGPFVVMKSPYNNRWYQMGIVSWGEGCDRDGKYGF
Y
THVFRLLKWIQKVIDRLGS
>1A0H:D|PDBID|CHAIN|SEQUENCE
SPLLETCPDRGREYRGLAVTHGSRCLAWSSEQAKALSKDQDFNPAVPLAENFCRNPDGDEGAWCYVADQPQDFEY
C
DLNYCEEPVDGDLGRGDDPDAIEGRTSEDFHQPFFNEKTFGAGEADCGLRPLFEKKVQDQTEKELFESYIEGR
>1A0H:A|PDBID|CHAIN|SEQUENCE...
  
```

Sample Output:

```

CLUSTAL W (1.83) multiple sequence alignment

1A0D_B|PDBID|CHAIN|SEQUENCE  PYFDNISTIAYEGPASKNPLAFKFNPEEKVGDKTMEEHLRFSVAYWHTF
1A0D_C|PDBID|CHAIN|SEQUENCE  PYFDNISTIAYEGPASKNPLAFKFNPEEKVGDKTMEEHLRFSVAYWHTF
1A0D_D|PDBID|CHAIN|SEQUENCE  PYFDNISTIAYEGPASKNPLAFKFNPEEKVGDKTMEEHLRFSVAYWHTF
1A0D_A|PDBID|CHAIN|SEQUENCE  PYFDNISTIAYEGPASKNPLAFKFNPEEKVGDKTMEEHLRFSVAYWHTF
1A0D_D|PDBID|CHAIN|SEQUENCE  -----MHLTPEEKS-----AVTALWGK
1A0D_B|PDBID|CHAIN|SEQUENCE  -----MHLTPEEKS-----AVTALWGK
1A0D_C|PDBID|CHAIN|SEQUENCE  -----VLSPADKT-----NVKAAWGK
1A0D_A|PDBID|CHAIN|SEQUENCE  -----VLSPADKT-----NVKAAWGK
1A0H_E|PDBID|CHAIN|SEQUENCE  -----IVEGQDAEVGLSPWQVMLFRKSPQE-----LLCGASL
1A0H_B|PDBID|CHAIN|SEQUENCE  -----IVEGQDAEVGLSPWQVMLFRKSPQE-----LLCGASL
1A0H_D|PDBID|CHAIN|SEQUENCE  -----SPLLETCPDRGRE-----YRGLAV
1A0H_A|PDBID|CHAIN|SEQUENCE  -----SPLLETCPDRGRE-----YRGLAV
  
```

Phylogenetic Trees:

The Phylogenetic Tree generated by the tool is a graphical user interface (GUI) that allows you to view and explore phylogenetic tree data.

We first build a multiple alignment of multiple protein sequences (orthologs or closely related paralogs; distant paralogs) are excluded to avoid including evolutionary variations that specify functional differences. The sequences' evolutionary relationships are inferred using the phylogenetic tree, which also yields the branch lengths in substitutions per sequence of the tree. Based on the topology and branch lengths of the tree, weights are calculated for each sequence. These weights are a control for phylogenetic correlation among the sequences.

Given the input sequences we extract the consensus sequence and perform the analysis of regions of interest (highly conserved regions) at each residue location specified.

Physico-chemical properties:

Enlisted are the different physico-chemical properties that are part of the tool.

- Hydrophobicity Index (Argos et al., 1982)
- Polarity (Grantham, 1974)
- Free Energy in alpha helical conformation (Munoz-Serrano, 1994)
- Side Chain volume (Krigbaum-Komoriya, 1979)
- Polarizability parameters (Charton-Charton, 1982)
- Average non-bounded energy per atom (Oobatake-Ooi-1997)
- Residue Volume (Bigelow, 1967)

Protein Maps:

The contribution of conserved residues toward bio-chemical function is determined by the interactions formed with substrates, cofactors, and other residues. Traditional sequence based techniques of homology transfer are sensitive and unreliable, and forcing researchers to venture into structure alignment and structure pattern matching techniques. Though more effective, the dependence of this method on 3D coordinate information make it computationally expensive on larger datasets. Protein Maps provides a visual representation of correlated mutations of physico-chemical interactions between residues reveal evolutionary residue conservation patterns that are unique to homologous proteins.

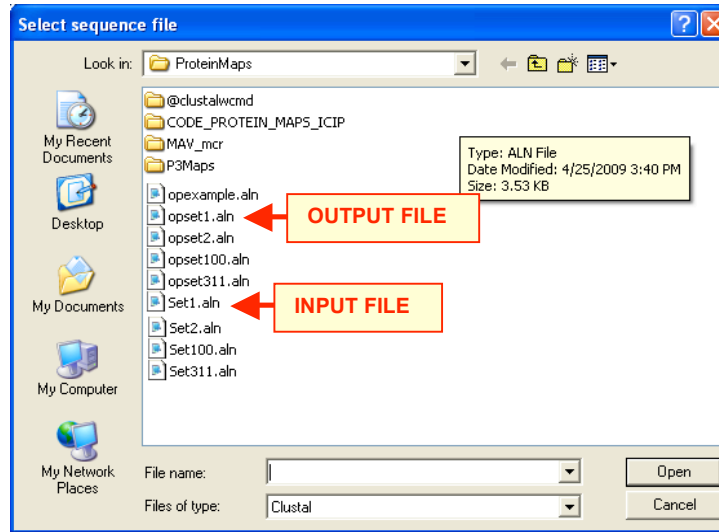
TOOL USAGE:

Step 1: Loading Input Sequences

The screenshot shows the P3Maps application window. The 'File' menu is open, with 'Input Alignment File...' highlighted by a red arrow. The 'Protein Physico-chem' panel is visible on the right, showing input and output file name fields and alignment parameters. The 'Instructions' panel at the bottom is circled in red and contains the following text:

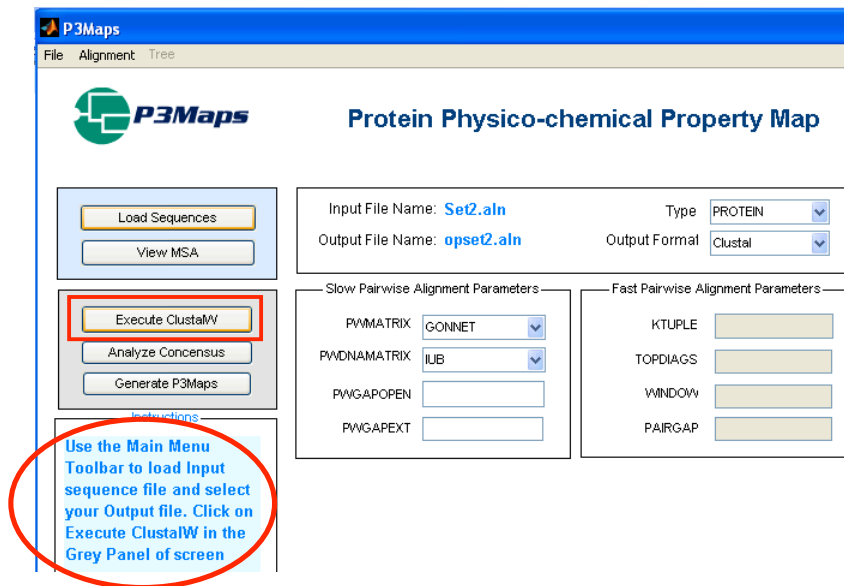
Step by Step Instructions

Use the Main Menu Toolbar to load Input sequence file and select your Output file. Click on Execute ClustaW in the Grey Panel of screen



The output file is created and saved by the user using the File>Output Alignment option. This is independent of the Input Alignment file option.

Step 2: Execution of Clustal W



This executes clustalw on the given input sequences and the output is stored in the output file for analysis. In order to view the alignment and the resultant phylogenetic tree, click on View MSA. Again step by step instructions are provided in the instruction panel at the bottom left of the screen.

Step 3: Viewing Multiple Sequence Alignment

The screenshot displays two windows from the P3Maps application. The left window, titled 'Louisiana Tech University Ruston, LA.', shows the results of a ClustalW multiple sequence alignment. It lists 12 sequences with their respective lengths and provides pairwise alignment scores for various combinations of sequences. The right window, titled 'Aligned Sequences', shows a text-based representation of the multiple sequence alignment, with residues aligned vertically. Below the alignment, a phylogenetic tree (Figure 2) is shown, illustrating the evolutionary relationships between the input sequences based on their alignment.

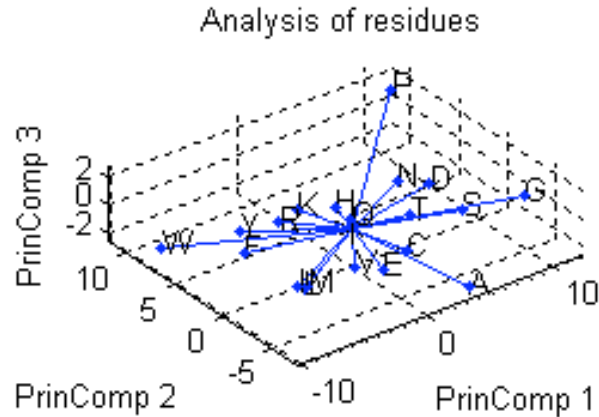
Above are the screen shots of the alignments and the resultant phylogenetic tree that captures the evolutionary relationship between the input sequences.



The output file is created and saved by the user using the File>Output Alignment option, stores the resultant alignment obtained from ClustalW

Step 4: Analysis of Consensus Sequence

The screenshot shows the 'Protein Physico-chemical Property Map' interface in P3Maps. The interface includes a header with the P3Maps logo and the text 'Protein Physico-chemical Property Map'. Below the header, there are several control panels. On the left, there are buttons for 'Load Sequences', 'View MSA', 'Execute ClustalW', 'Analyze Consensus', and 'Generate P3Maps'. In the center, there are input fields for 'Input File Name' (Set2.aln) and 'Output File Name' (opset2.aln), along with dropdown menus for 'Type' (PROTEIN) and 'Output Format' (Clustal). Below these are sections for 'Slow Pairwise Alignment Parameters' and 'Fast Pairwise Alignment Parameters'. On the right, there is a 3D plot titled 'Analysis of residues' showing the distribution of residues in a 3D space defined by PrinComp 1, PrinComp 2, and PrinComp 3. A note at the bottom left states: 'The default region for Consensus Analysis is between residues 80 to 100. Next Click Generate P3Maps in Grey Panel'.



From the input sequences, the consensus sequence is extracted and the analysis of the conserved region of interest is carried out.

The process involves the multiplication of the weights with the fraction of sequences carrying a particular amino acid to get the alignment summary (matrix summary) which we interpret by using a matrix of physico-chemical property scales. The result is an estimate of the physico-chemical constraints on each position based on the mean and variance of the property distributions observed in its alignment.

For example these statistics are biologically significant where the mean measures hydrophobic character, and the variance measures the strength of the constraint. Deviations from the alignment column are obtained for each variant by calculating its property difference from the mean and dividing it by the square root of the variance. We can interpret this statistic as a signed measure of “*constraint deviation*”.

To compute a single score measuring the deviation of constraint across all properties, we first de-correlate the properties using Principal Component Analysis (PCA). This application gives rise to a new feature space in which each axis is a principle component, and the distance from the origin to any amino acid is the amino acid impact score.

Based on the principles of sequence alignment, the most common starting point for generating a model, we aim to capture those domains that reflect structural conservation between homologous proteins.



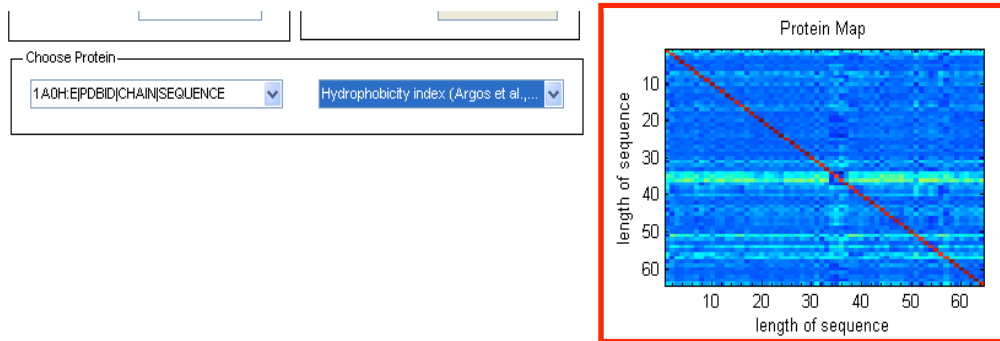
The default region of analysis is set to residue locations 80 to 100 for the given sample sequences.

Step 5: Generation of Protein Maps

Once the analysis of consensus sequence is carried out, the user clicks on Generate P3Maps

Execute ClustalW	PWMATRIX: GONNET	KTUPLE: <input type="text"/>
Analyze Consensus	PWDNAMATRIX: ILUB	TOPDIAGS: <input type="text"/>
Generate P3Maps	PWGAPOPEN: <input type="text"/>	WINDOW: <input type="text"/>
Instructions	PWGAPEXT: <input type="text"/>	PAIRGAP: <input type="text"/>
<p>The default region for Consensus Analysis is between residues 80 to 100. Next Click Generate P3Maps in Grey Panel</p>	Choose Protein: 1A0H:EPDBID CHAIN SEQUENCE	
	Hydrophobicity index (Argos et al.,...)	

The user is provided the option to choose a specific protein of interest and a specific physico-chemical property. This triggers the visual representation of the interaction of region of interest along the body of the protein.



The user could choose any sequence and property, for further analysis.

Installation instructions

Minimum System Requirements:

Processor	Intel® Pentium® CPU 2.39 Gz.
RAM	1G
OS	MicroSoft Windows XP
Software	Matlab 2007 a

Please follow the instructions provided for installation

Provided is the **P3Maps.exe.pkg** file of size 1.23 MB, can be stored anywhere on the system (for example on the **desktop**)

Step 1: Double click **P3Maps.exe.pkg** to unpack package file. You will observe a popup dos prompt, appear and disappear. Additional files created:

P3Maps.exe
P3Maps.ctf
Clustalw.exe

Step 2: Double click on **P3Maps.exe**. This would install the software and would take a couple of minutes for the GUI to appear.

Step 3: The default sample input file provided in the package can be set by following these steps:
 Click on the button **Load Sequences**. This enables the **main menu** on top of screen.

Click on the option **File>Input Alignment File...**, This brings up the Select sequence file window.

Step 4: To select the Sample **input file**, choose the following folders: **P3Maps_mcr > P3Maps > Set2.aln**.

Step 5: To create the **output file**, Click on the option **File>Output Alignment File...**, and provide an output file name. e.g.: **opSet2.aln**.

This would help you processed through the tutorial...**Have fun!!!**