Properties of pore lining residues of transporter protein families in different Staphylococcus Aureus strains

Suprevisor: *Prof. Dr.* Volkhard Helms
Advisor: Duy Nguyen
Co-advisor: Rahmad Akbar
Presented by: Jacob Marcus Ambat

OUTLINE

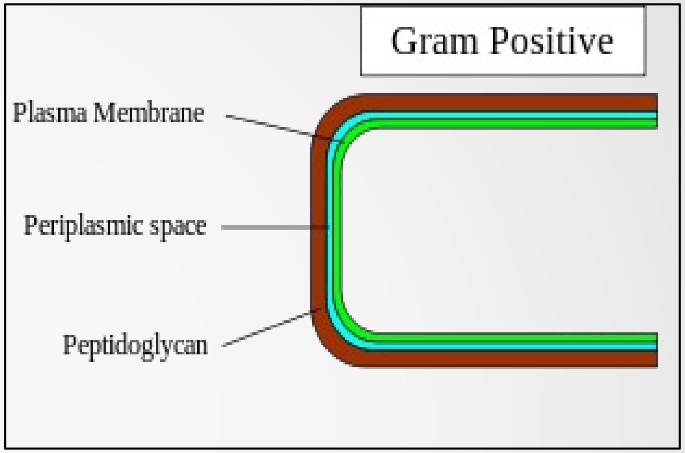
- **MOTIVATION**
- **FPRA PIPELINE**
- **PROGRAMS AND METHODS**
- **RESULTS**
- **REFERENCES**

MOTIVATION

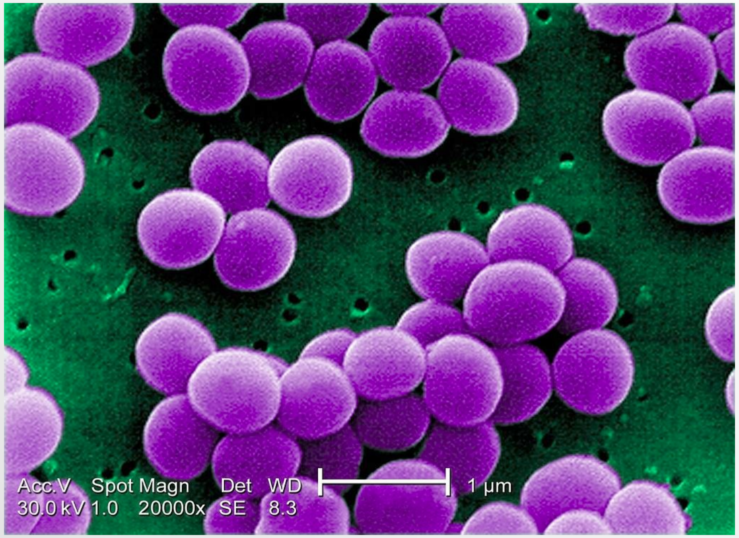
- I'm interested in transport of substrate molecules through transport proteins and channels.
- Identification of Pore Lining Residues (PLRs) of Transmembrane proteins with unknown 3D structures.
- Exploit the central hypothesis in biology which is “Functionally important residues are conserved”.
- Aim : A tool that determines the conservation of PLRs.
- Compare the degree of conservation (DOC) of the PLRs within different transporter families or within *Staphylococcus aureus subspecies*.

STAPHYLOCOCCUS AUREUS

- *Staphylococcus aureus* is a gram positive coccal bacterium.
- It is found on the respiratory tract and skin.
- Also known as "Golden Staph" or "Oro Staphira".



Source: Wikipedia



- Associated with a wide range of diseases which are nasocomial and community-acquired.
- Our study is based on antibiotic resistant *S. aureus* strains COL and N315.

Source: Centers for Disease Control and Prevention's Public Health Image Library (PHIL)

DATASET

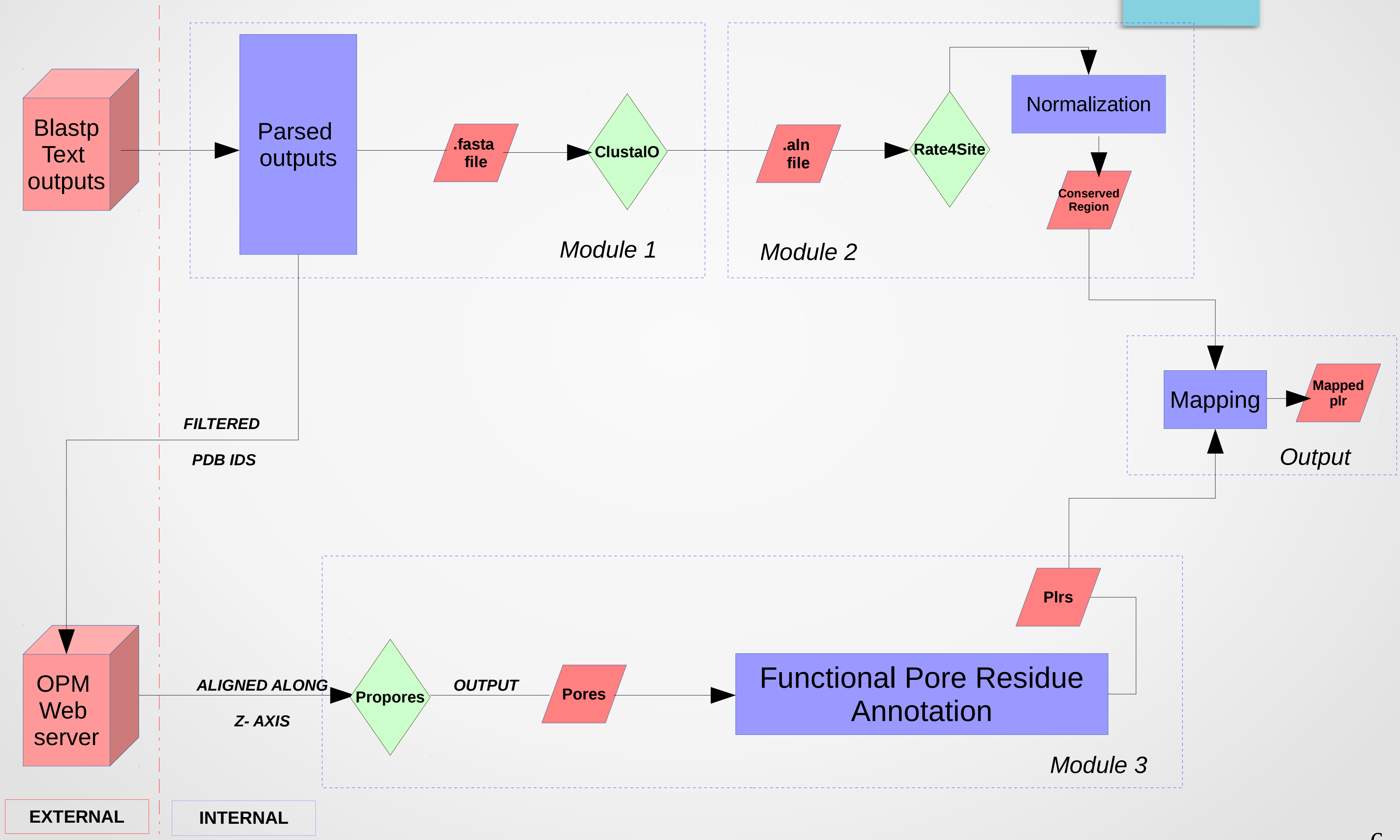
- SACOL and SAN315 transporter protein query sequences were obtained from transportDB - *http://www.membranetransport.org/*

- Structural information for all known alpha-helical transporter proteins from Stephen White laboratory - *http://blanco.biomol.uci.edu/*

- NCBI repository(nr and pdbaa).

TransportDB		Stephen White Lab
SACOL	SAN315	Transmembrane Protein : Alpha-Helical
192	314	1572

FPRA PIPELINE



Blastp

- Proteins are modular in nature.
- Blast algorithm finds the functional domains or shorter stretches of sequence similarity.
- A higher bit score and an lower E-value indicate significant hits.
- Sequence homology using Blast standalone program.
- 2 level parsing steps to filter the hits.
- Level 1 parsing extracts < = 500 homologs from all organisms.
- Level 2 parsing extracts homologs based on the threshold.
  - E-value < = 1e – 10
  - Identity % > = 35%
  - Coverage % > = 75%

	SACOL		SAN315	
	SEQ LIST	STR LIST	SEQ LIST	STR LIST
Level 1	30896	6421	49616	8449
Level 2	28010	352	45032	436

DEFINITION OF THRESHOLD VALUES

EXPECT-VALUE

- Definition : The Expect value (E) is a parameter that describes the number of hits of the same or higher score, one can "expect" to see by chance when searching a database of given size.
- E-value < = 1e – 10 means it has a very low probability of occurring randomly.
- This in fact helps in finding highly similar sequences in closely related species.

IDENTITY

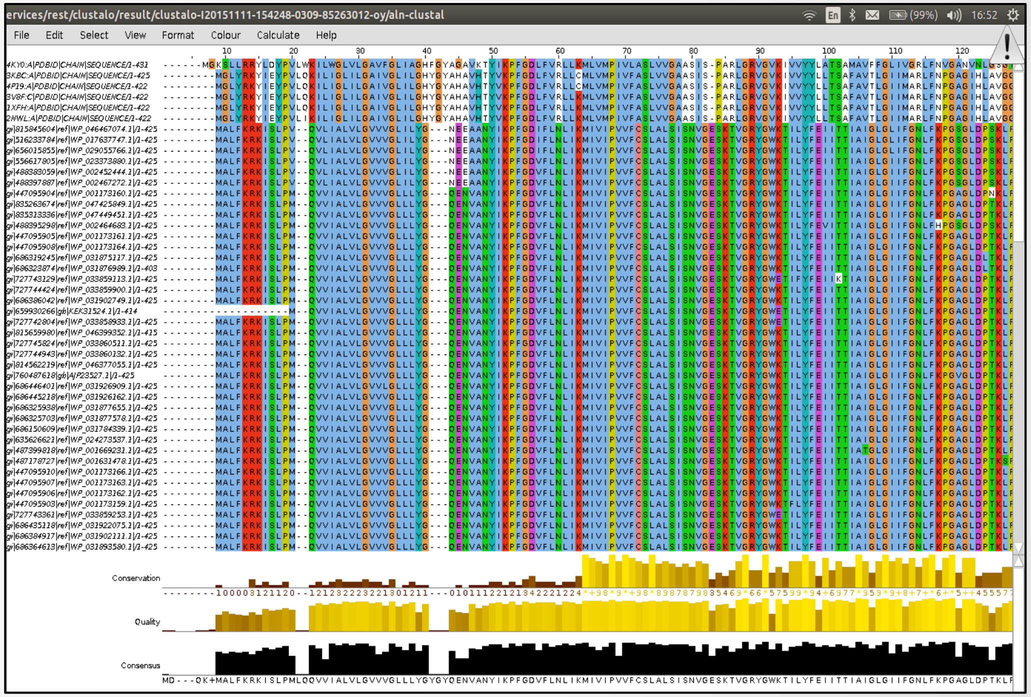
- Definition : The extent to which two (nucleotide or amino acid) sequences have the same residues at the same positions in an alignment, often expressed as a percentage.
- Identities > = 35% means that there are 35 percent or more residues that are identical with the query sequence.

COVERAGE

- Definition : The percentage of query sequence that overlaps the subject sequence.
- Coverage = (Length of Alignment / Length of Query sequence).
- Ideal coverage should be > 70%.

MULTIPLE SEQUENCE ALIGNMENT BY CLUSTALO

- Input fasta file include sequences of query, sequence list and structure list.
- Gives a better estimate of conservation.
- Clustal omega is an accurate alignment program for proteins of divergent organisms.
- It outperforms other programs in terms of runtime and quality.
- Important feature include adding newer sequences and using precomputed alignment information.


ervices/rest/clustalo/result/clustalo-I20151111-154248-0309-85263012-oy/aln-clustal
En
(99%)
16:52
File
Edit
Select
View
Format
Colour
Calculate
Help
10
20
30
40
50
60
70
80
90
100
110
120
4KY0:A|PDBID|CHAIN|SEQUENCE/1-431
3KBC:A|PDBID|CHAIN|SEQUENCE/1-425
4P19:A|PDBID|CHAIN|SEQUENCE/1-422
3V8F:C|PDBID|CHAIN|SEQUENCE/1-422
1XFH:A|PDBID|CHAIN|SEQUENCE/1-422
2NWL:A|PDBID|CHAIN|SEQUENCE/1-422
Conservation
Quality
Consensus

Source: Clustal Omega Jalview in EMBL-EBI

RATE4SITE SCORING

- Rate4Site evaluates the evolutionary conservation of positions in a protein sequence.
- If < 20 sequences, Maximum – likelihood algorithm, else > 20 sequences, Neighbor – joining algorithm .
- Maximum Likelihood rate is calculated for each position.
- The number of replacements expected at each site = l * (r[j])
- The rate at a specific site is estimated as :

$$P(\{X,Y\},data|r) = pi_X * P_{(X,Y)}(r * t)$$

- At every position the rate is been calculated and Lower scores mean higher conservation.

REFERENCE : Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Tal Pupko et.al.

NORMALIZATION

- The rate scores obtained are normalized by “Feature Scaling”.
- Its also known as unity-based normalization.
- All the scores which are generated are scaled to the range [0,1]. [fig 1]
- Feature Scaling ,

$X' = (X – X_{min}) / (X_{max} – X_{min})$

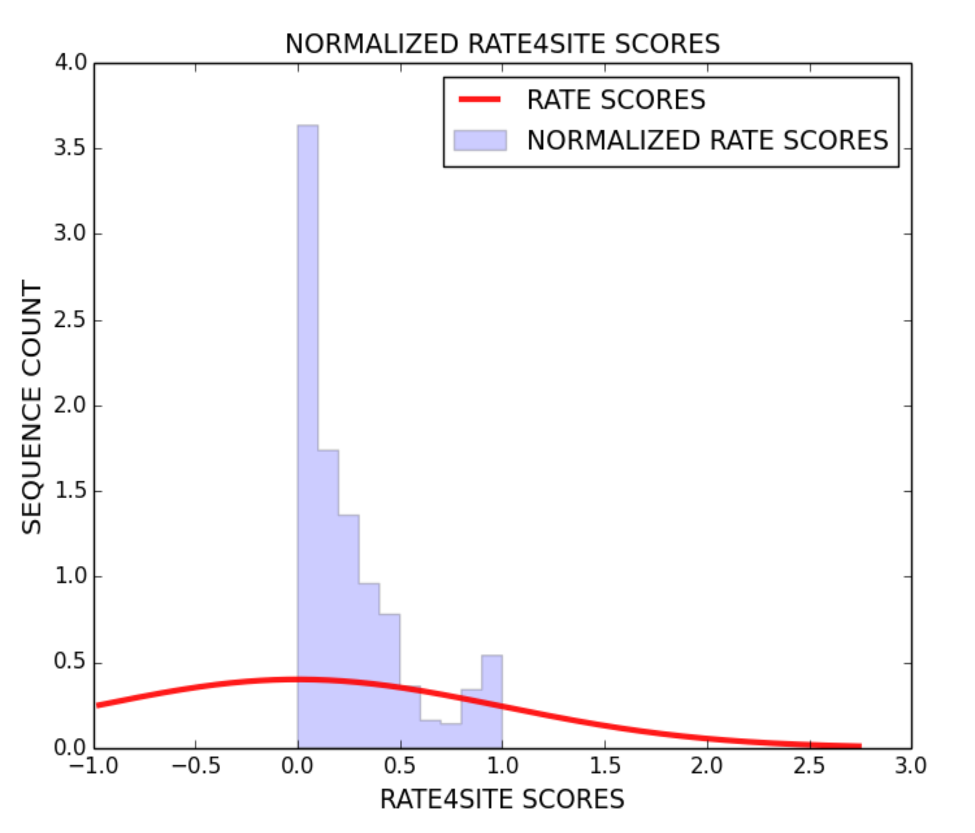


Fig 1 :   NORMALIZING THE RATE SCORES OF DAACS FAMILY

- The higher and lower values are then assigned to a color grade according to the color scheme 1 – 9. [fig 2]
- The highly conserved residues fall into the first three bins.
- If the interval in the specific position falls in 4 or more color bins, the score is considered unreliable.

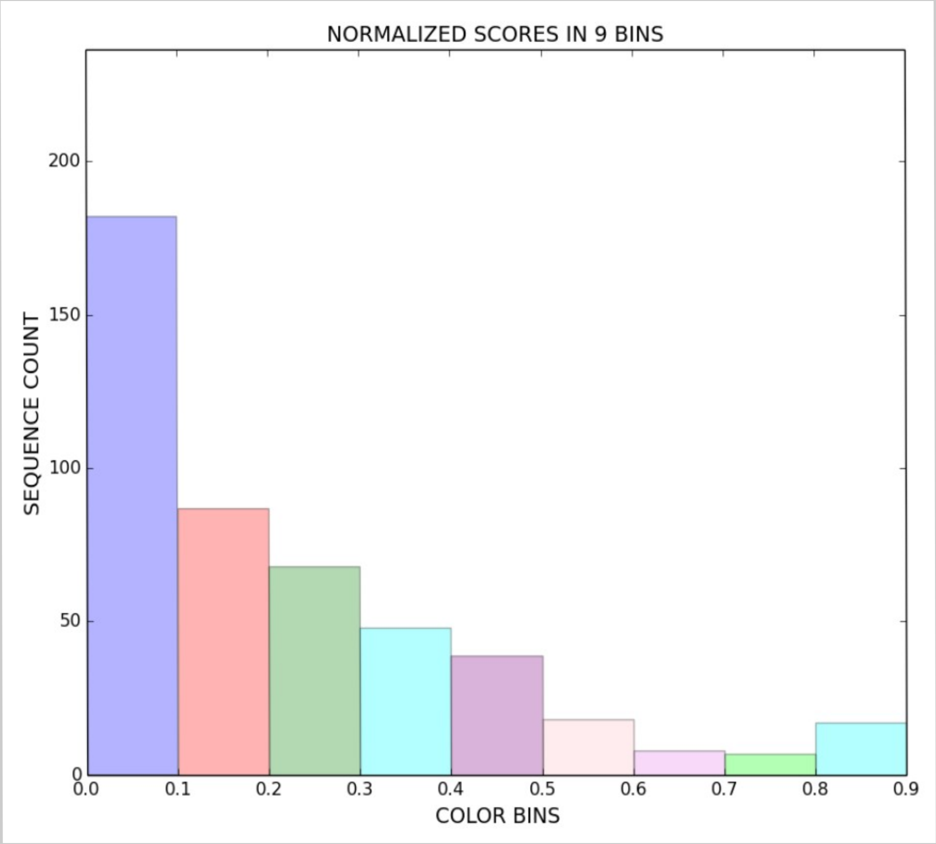


Fig 2 : ESTIMATING THE CONSERVED RESIDUES

TRANSPORTER PROTEIN FAMILIES

	TRANSPORTER FAMILIES	FUNCTION
1	ABC -family of transporters	Homodimer/Heterodimer
2	MFS Superfamily	Monomer
3	DMT Superfamily	- - No Hits --
4	MOP family	- - No Hits - -
5	BCAA (LIVCS) family	- - No Hits - -
6	LysE family	- - No Hits - -
7	POT family	Homodimer
8	RND family	- - No Hits - -
9	Solute/Sodium Symporter	- - No Hits - -
10	PTS System	Enzyme IIA complex
11	P- ATPase family	Enzyme complex
12	K+/Trk Superfamily	Homotetramer
13	F-ATPase Superfamily	Homodimer
14	Dicarboxylate/Amino Acid symporter	Homotrimer

OPM DATABASE

- Predicted spatial arrangement of membrane proteins in the lipid bilayer.
- This database includes features of membrane proteins
  - structural classification, species, destination membrane.
  - numbers of transmembrane segments and subunits.
  - numbers of secondary structures and the calculated hydrophobic thickness.
- The hydrophobic thickness is in the range of 21.1 - 46 A**.**

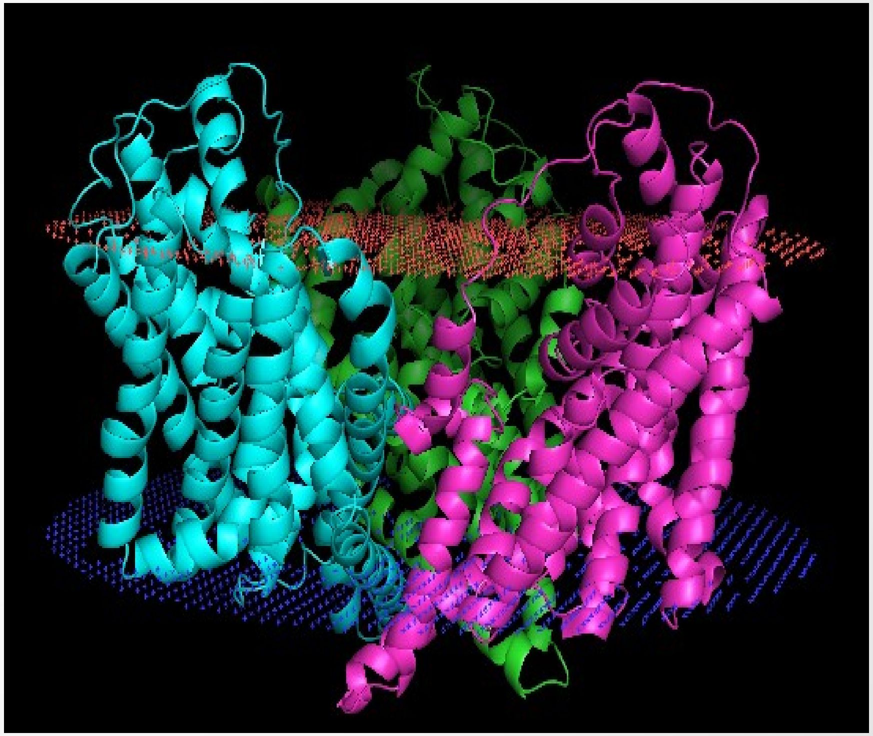


Fig 3 : 1xfhout.pdb

REFERENCE : OPM: Orientations of Proteins in Membranes database. Mikhail A et.al.

PROPORES

- Toolkit for identifying pockets, cavities and channels in a protein structure.
- Consists of 3 Perl programs namely
  - Pore_ID.pl for identifying the pores.
  - Pore_Trace.pl for pore axes determination.
  - Gate_Open.pl for opening the gate between neighboring pores.
- We use Pore_ID on the coordinate file of the output ".pdb" from OPM database.
- Ouput files consists of 3 text files namely: ".list", ".pdb" and ".PTin" indexed from 0.
- The list file contains the PLRs.

REFERENCE: Identifying continuous pores in protein structures with PROPORES by computational repositioning of gating residues. Lee PH1, Helms V.

POREWALKER

- Unfortunately PROPORES could not handle larger proteins (>= 15000 atoms).
- We used PoreWalker to identify PLRs on the larger proteins.
- Here we found that PoreWalker actually extracts the beta carbon atoms of the residues embedded on the helix that faces the pore.
- This gave us the idea to extract only the beta carbon ATOMS in OPM pdb file.
- By doing so, the size of the pdb file is reduced so much and made it easy for processing using PROPORES.
- However PROPORES remained unable to processes some families like Trk, SSPTS.

FPRAT : Functional Pore Residue Annotation Tool

Tool works in as follows:

- Extract the coordinates to a text file for each OPM output, eg. “1xfhATOM”.
- Calculate the Center of mass (COM) of the coordinate file of the z-oriented OPM file.
- Create a box with the COM as its center and spanning

  -15.0 < = z < = 15.0, -1.0 < = x < = 1.0 and -1.0 < = y < = 1.0.
- Output the coordinates of the box as a text file, eg. “1xfhout_box.txt”.
- Filter out the pores identified using PROPORES.
- Parse out the PLRs and convert them to single letter codes.

CENTER OF MASS (COM)

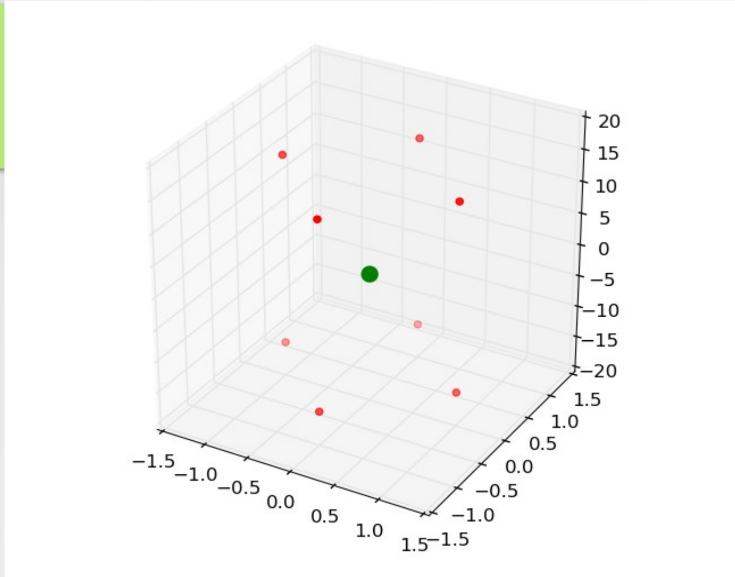


Fig 4 :   COM OF 1xfhATOM

- *Fig 4* depicts the box generated for 1xfhout.pdb.
- The green dot represents the COM.
- The red dots represent the box coordinates.

PORE FILTER

- The red dots in *Fig 5* depict the filtering out of the coordinate points of the pores for 1xfhout.pdb.
- The enclosed pore of these points is then identified.
- The corresponding list files are then identified.

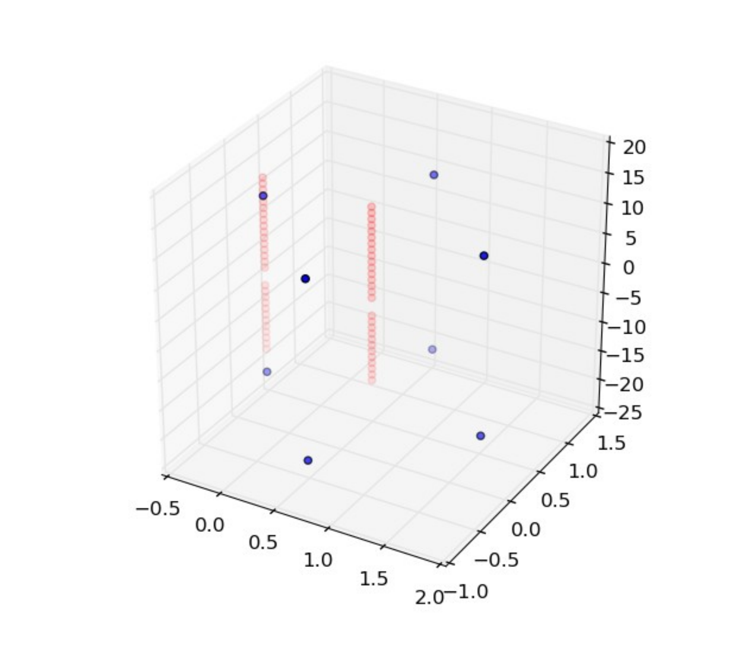
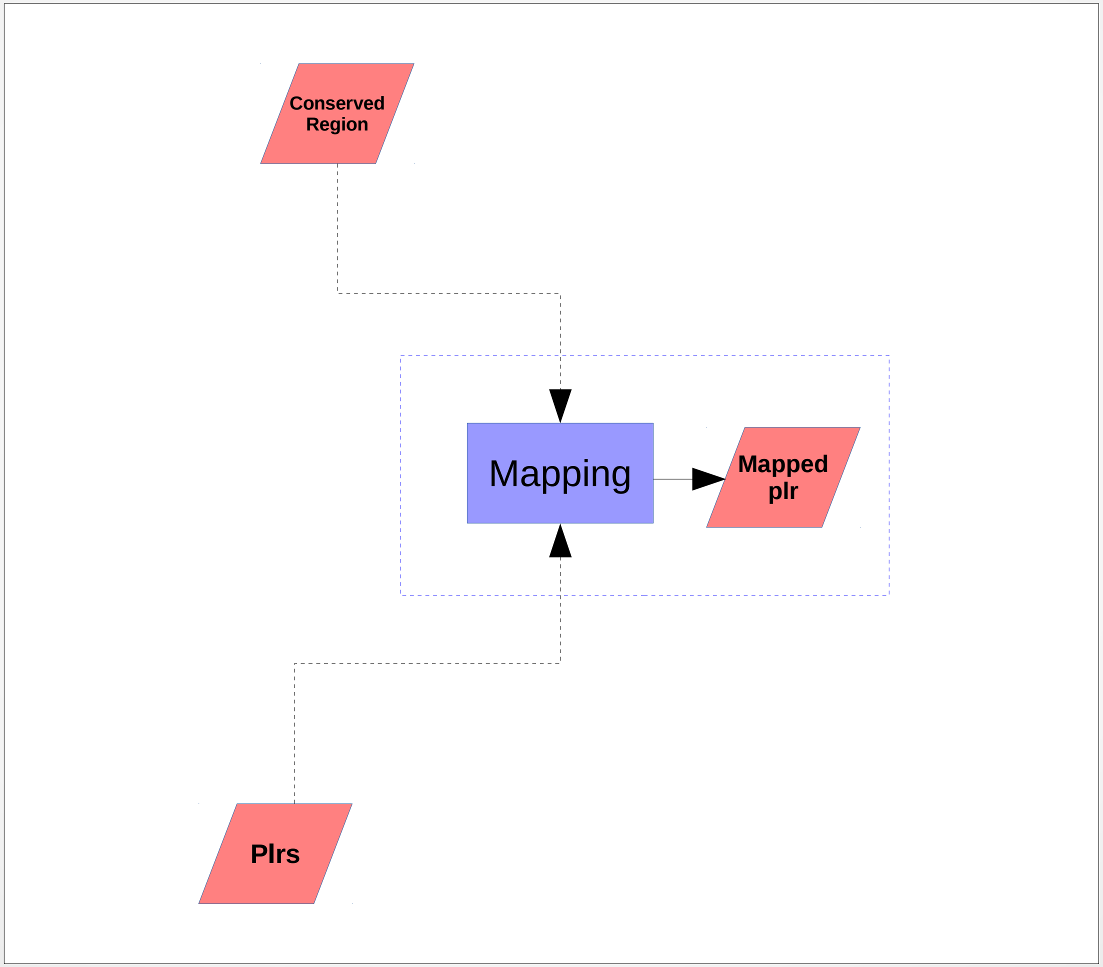


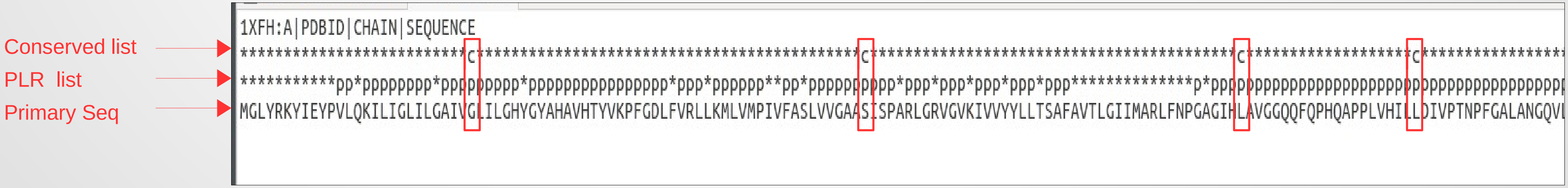
Fig 5 :   FILTER OUT THE PORES OF 1xfhATOM

RESULTS



MAPPING

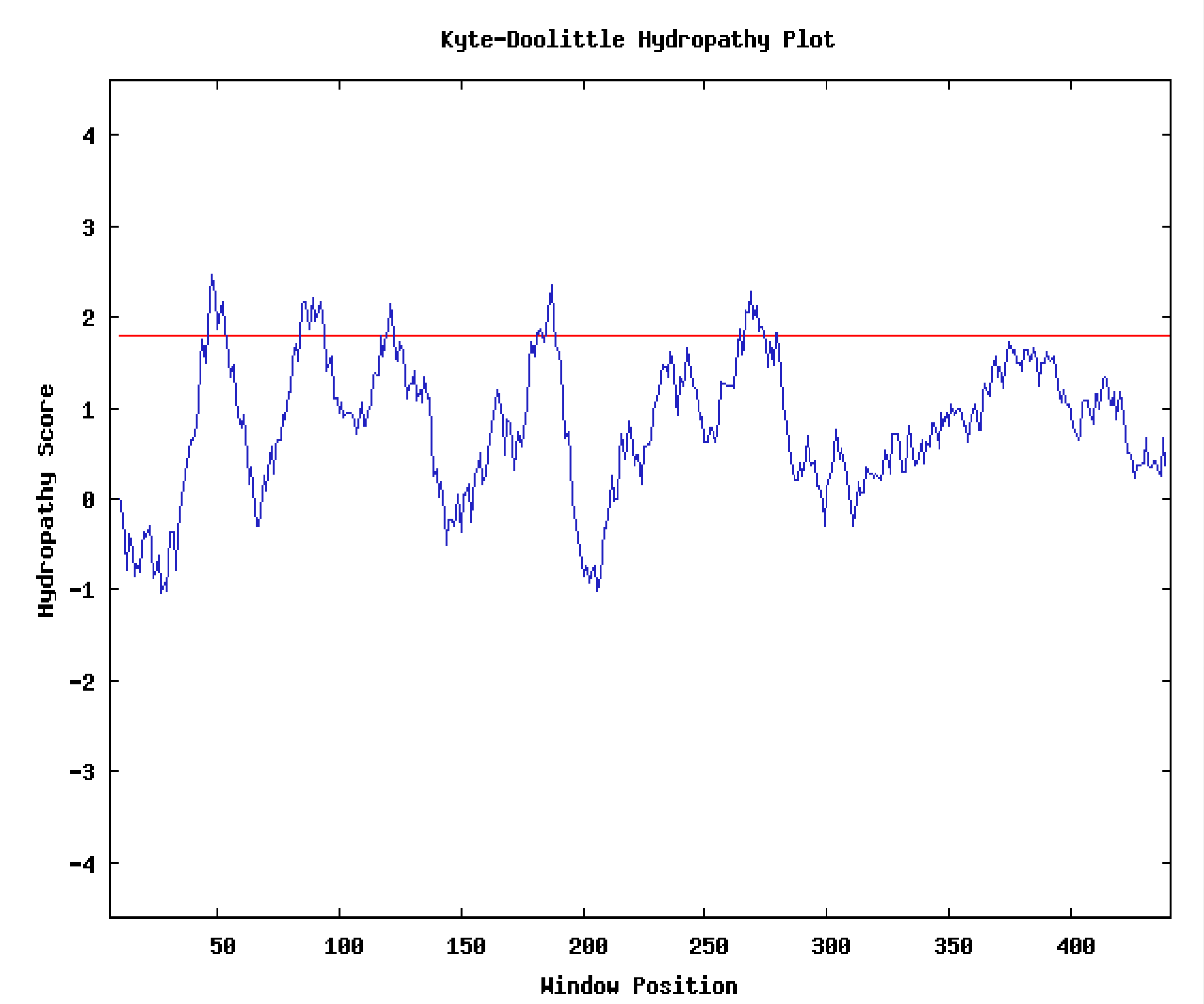
- The conserved list (represented in fig as “c”) obtained by module2 along with the PLR list (represented in the fig as “p”) from module3 is mapped to the primary sequence of the protein.
- This helps us to identify the set of conserved PLRs along with their positions.
- The degree of conservation is calculated for the conserved PLR and Non-PLR sets.
- For most of the transporter families the DOC >50%.



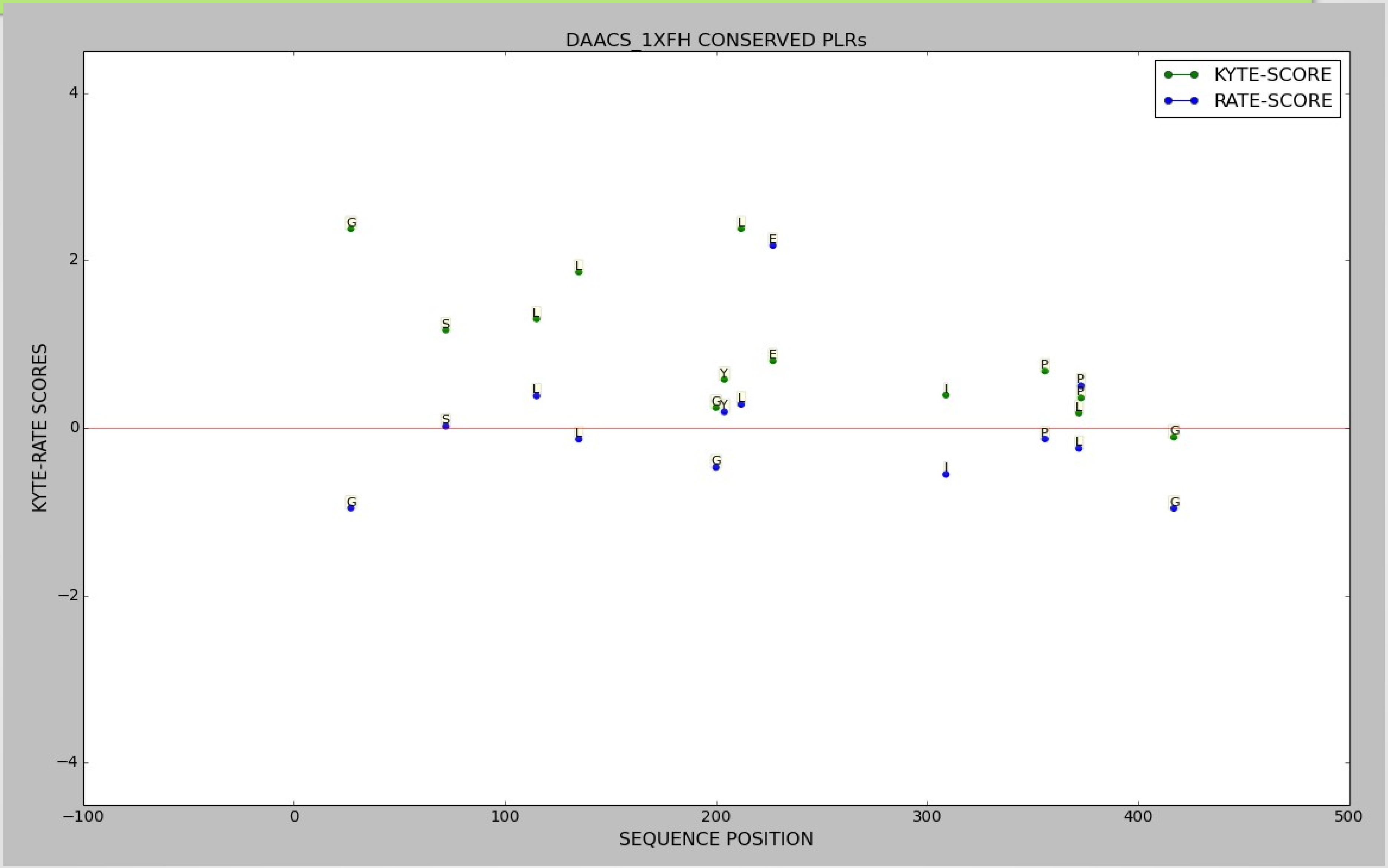
```
1XFH:A|PDBID|CHAIN|SEQUENCE
**************************c********************************************c******************************************c*******************c*****************
***********pp*pppppppp*ppppppppp*ppppppppppppppppp*ppp*pppppp**pp*ppppppppp*ppp*ppp*ppp*ppp*ppp**************p*pppppppppppppppppppppppppppppppppppppppppp
MGLYRKYIEYPVLQKILIGLILGAIVGLILGHYGYAHAVHTYVKPFGDLFVRLLKMLVMPIVFASLVVGAASISPARLGRVGVKIVVYYLLTSAFAVTLGIIMARLFNPGAGIHLAVGGQQFQPHQAPPLVHILLDIVPTNPFGALANGQVL
```

KYTE-DOOLITTLE SCORING

- Quantitative measure of the degree of hydrophobicity.
- It is useful to identify the possible domains of a protein structure.
- The peaks above the red line indicate the possible hydrophobic regions.
- A stretch of 20 amino acid residues in the plot with positive score indicate that they are a part of the alpha-helix spanning the lipid bilayer.



MAPPING OF CONSERVED PLRs TO THE PROTEIN SEQUENCE



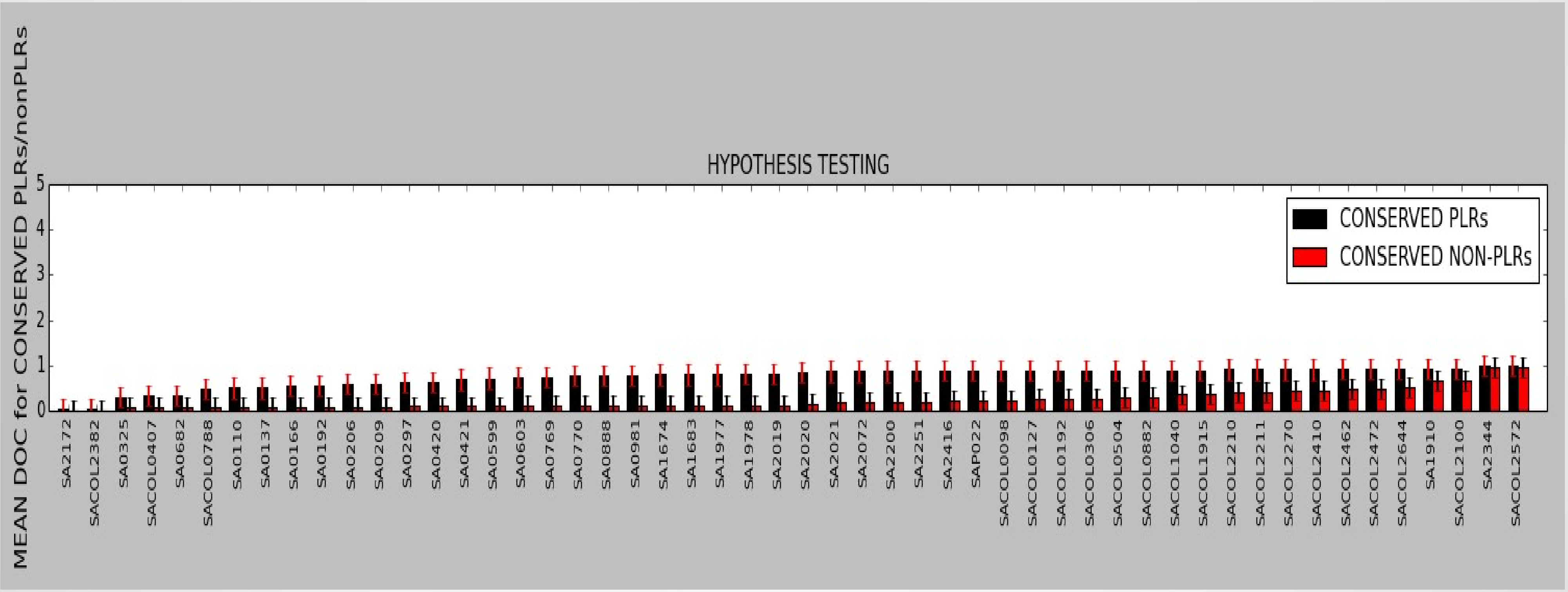
DEGREE OF CONSERVATION

- The degree of conservation of each position is the inverse of the site's evolutionary rate.
- Rapidly evolving positions are variable while slowly evolving positions are conserved.
- Structural and functional conservation gives us a measure of the evolutionary relationships between subjects.
- Relatively fewer conserved residues are sufficient for understanding the molecular architecture of the TM fold.
- The PLRs in the active site tend to be highly conserved.
- These residues can occur in the conserved pocket region which represents a useful druggable site.

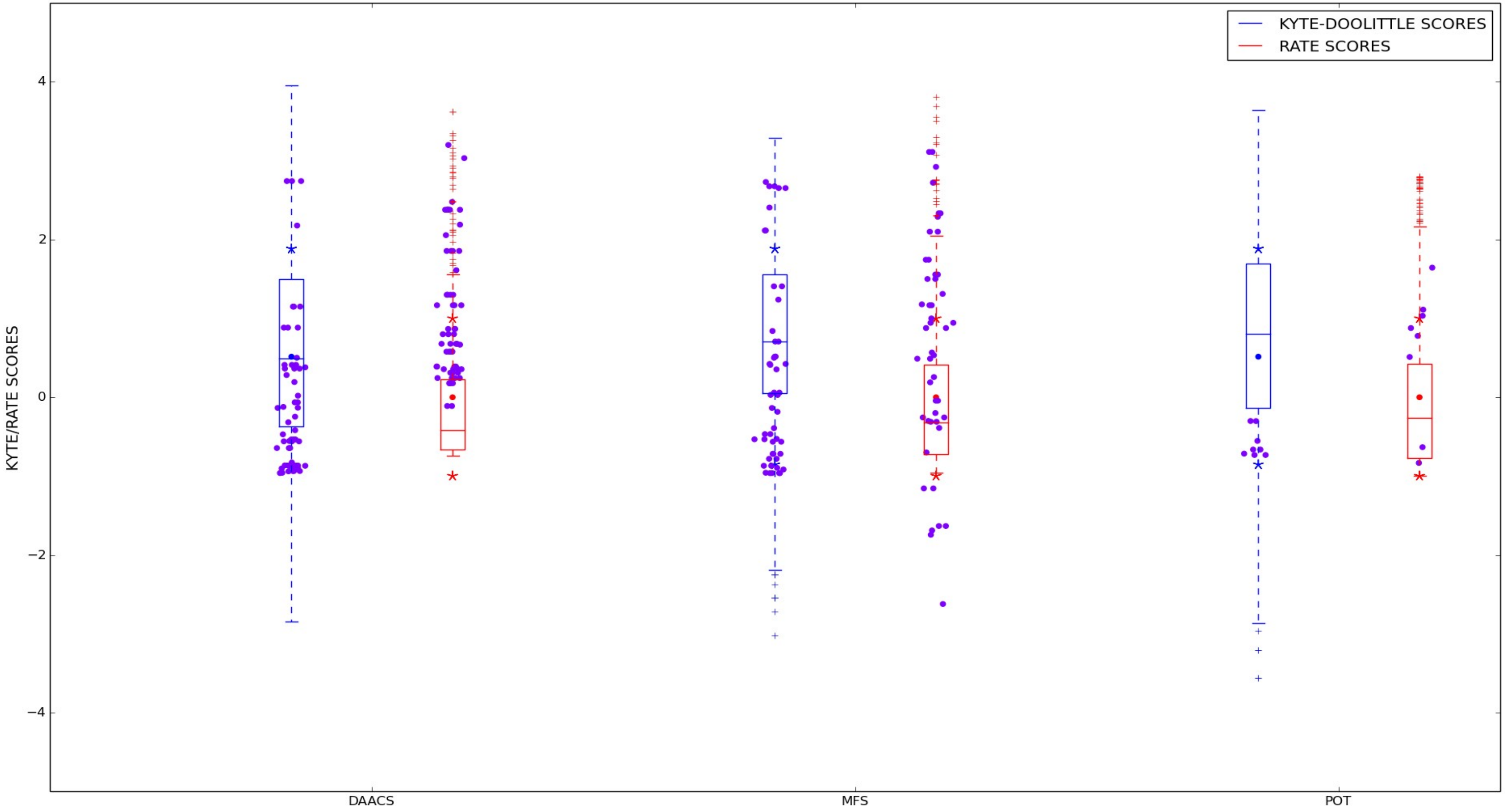
DEGREE OF CONSERVATION OF PLRs FOR PROTEIN FAMILIES

SR	TRANSPORT FAMILY	TOTAL No. OF PROTEINS SAN315+SACOL	FRACTION OF CONSERVED PLRs	FRACTION OF CONSERVED NON-PLRs
1.	DAACS	3+3	0.46,0.63	0.54,0.37
2.	MFS	1	0.34	0.66
3.	POT	1+1	0.57,0.44	0.43,0.56
4.	ABC	12+12	0.77,0.81	0.23,0.19
5.	F-ATPase	1+2	0.70,0.85	0.30,0.15
6.	P-ATPase	1+1	0.0,0.06	1.0,0.94

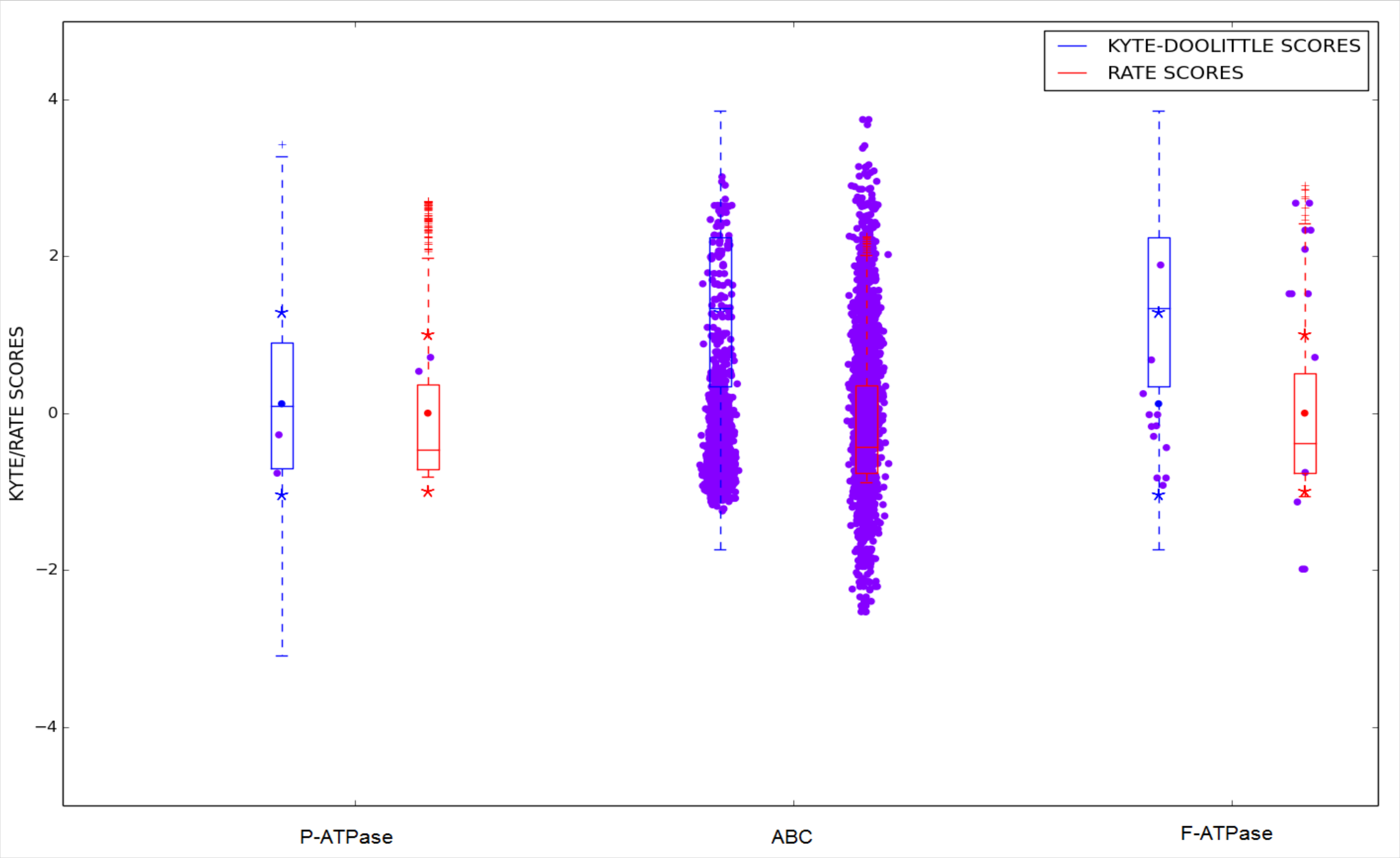
COMPARISON OF MEAN DOC BETWEEN FAMILIES



HYDROPHOBICITY Vs CONSERVATION FOR SMALLER PROTEIN FAMILIES



HYDROPHOBICITY Vs CONSERVATION FOR LARGER PROTEIN FAMILIES



HYPOTHESIS TESTING

- The hypothesis statement asserted is called $H_{0.}$
- The argument statement is H1 $\rightarrow \mu\ 0 \neq \mu\ 1$.
- P-value assumption(reductio ad absurdum).
  - Pr(X ≥ x|H) for right-tailed event.
  - Pr(X ≤ x|H) for left-tailed event.
  - 2 min{P r (X ≤ x|H)} , {P r (X ≥ x|H)} for double-tailed event.

WILCOXCON SIGNED – RANKED TEST

- Wilcoxon signed-ranked test is a hypothesis test used when we compare two related/matched samples.
- Also called a paired difference test.
- In order to use this test, we assume the following assumptions:

(1)The data is paired and come from the same population.

(2)Each of the pair is randomly taken and is independent to each other.

(3)The data are measured on an ordinal scale.

- The P-value for conserved PLRS and conserved Non-PLRs is 3.45001e-10.

HYPERGEOMETRIC TESTING

- Definition : Hypergeometric distribution is a discrete probability distribution which describes the probability k successes in n draws from a finite population of size N that contains K successes, wherein each draw is either a success or a failure.
- A random variable X follows the hypergeometric distribution if its probability mass function (pmf) is given by:

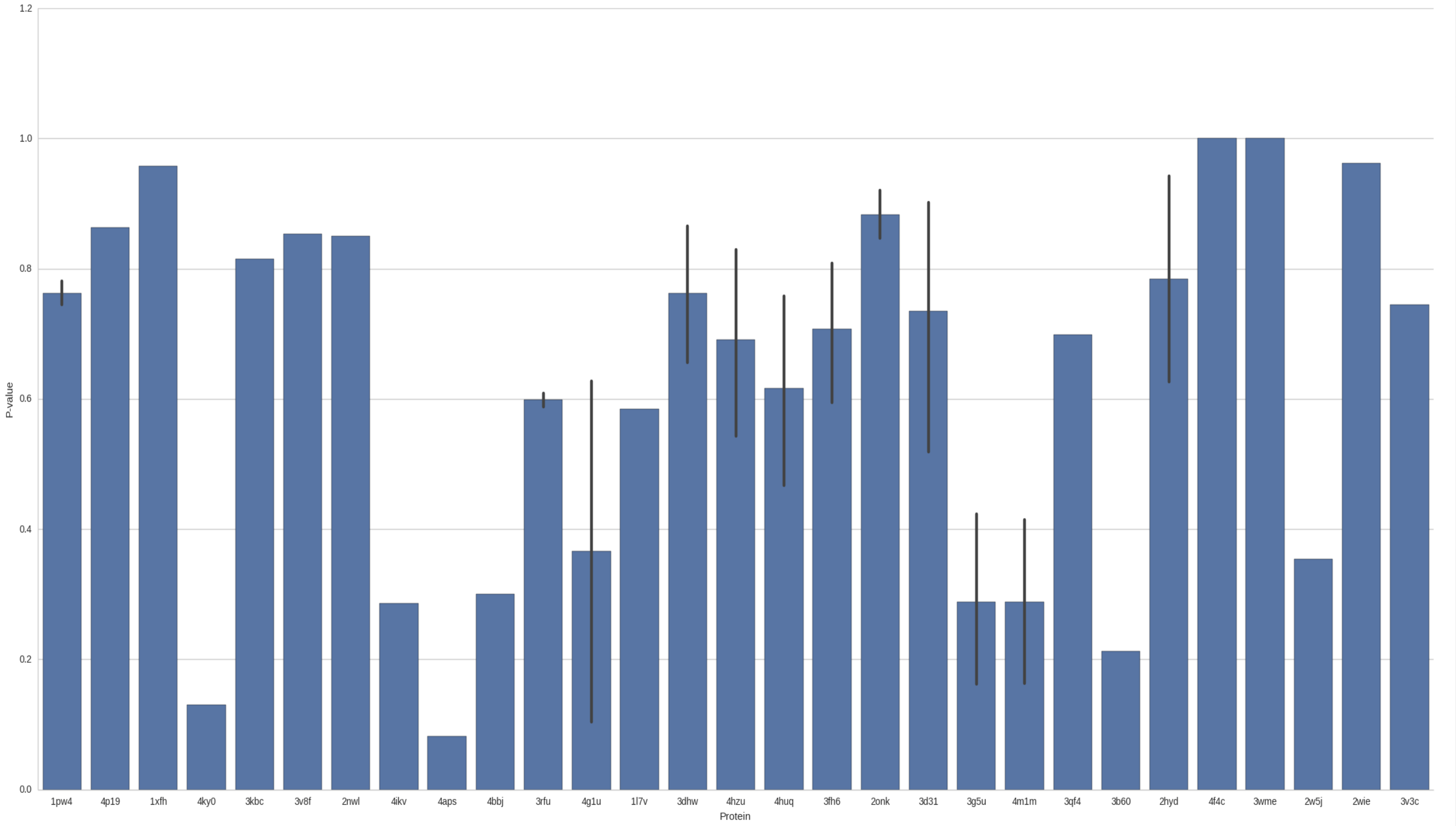
P(X = k) = $\binom{K}{k}\binom{N-K}{n-k}/\binom{N}{n}$

where N is the population size, K is the number of success states in population, n is the number of draws, k is the number of observed successes and (ab)is a binomial distribution.

- The pmf is positive when

max(0, n+K-N) ≤ k ≤ min(K,n)

P-VALUE BY HYPERGEOMETRIC TESTING



- Here we set N as the total length of the protein sequence, K as the number of PLRs, n as the number of conserved residues through the length of the protein sequence and k as the number of conserved PLRs through the length of the protein sequence.
- We see that the P-values for each of the proteins are very high proving less significance in the test as p-value need to be a low value.
- This is insufficient to support the hypothesis.
- This can mainly due to the following reasons:
  - very low balance between the conserved data set and PLR data set.
  - The initial setting on PROPORES for the PLR prediction. The grid settings have to be proper in order to predict the PLRs in a much effective way.
  - No difference in the protein sequence of the transporters of both the strains SN315 and SACOL obtained from the transportDB database.

REFERENCES

- *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Fabian Sievers et.al.*
- *Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Tal Pupko et.al.*
- *PRIMSIPLR: prediction of inner-membrane situated pore-lining residues for alpha-helical transmembrane proteins. Nguyen D, Helms V, Lee PH.*
- *Positioning of proteins in membranes: a computational approach. Lomize AL1 et.al.*
- *OPM: Orientations of Proteins in Membranes database. Mikhail A et.al.*
- *Identifying continuous pores in protein structures with PROPORES by computational repositioning of gating residues. Lee PH1, Helms V.*
- *A simple method for displaying the hydropathic character of a protein. Jack Kyte, Russell F. Doolittle.*
- *The Consurf Server : Server for the Identification of Functional Regions in Proteins.*

THANKYOU