Protein Residue Contact Network Analysis to understand function of Single Residue Mutations

Dhanasekhar V.K, Dr. Shashi Bhushan Pandit, Prof. Somdatta Sinha

Department of Biological Science, IISER Mohali

Introduction

- Insights into mechanistic details of protein function are crucial for protein design, rational drug discovery and understanding evolution of proteins.
- In general, Single Residue Mutations (SRM) or single site mutations have contributed immensely in identifying important residues and deciphering their roles in protein function.
- Recently, single site variation data from human 1000 genome sequencing projects and Deep Mutational Scanning (DMS) methods have generated plethora of SRMs, however, their effect on protein structure/function remains elusive.
- Previously, we have used Protein Contact Networks (PCN) [1] approach to understand functional defects in SRMs. We observed that global or local network properties between Wild type (WT) and mutant (MT) proteins are similar. There were some protein WT-MT pairs, which showed changes in global/local network properties.
- In continuation of the efforts to understand affect of SRM on protein structure/function, we make following two main objectives:
  - ✓ To study a subset of WT-MT protein pairs that have large change in amino acid properties.
  - ✓ To associate SRM to functional phenotypic effect we developed a score Amino acid Physiochemical Score Matrix (APSM).

Dataset and Methodology

Dataset A: In-house dataset consists of a 3600 wild-type mutant pairs having single residue mutations used for the network and physico-chemical analysis.

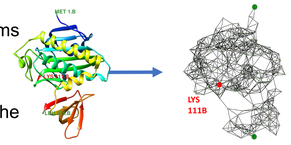
Dataset B: Non-redundant single domain created from pdb database (PISCES) [2] having resolution ≤ 2.5 Å.

Constuction of Protein residue Contact Network

- **Nodes**: C-α residues
- **Edges**: An edge between two nodes are drawn, if any two Cα atoms is within Euclidian distance ≤ 7 Å [3]
- **Degree (k)**: Number of of connections of a node.
- **Shortest path** (L_{ij}): Number of links that must be traversed by the shortest route, from one to node to another.
- **Clustering coefficien**t (CC $C_i = 2 * n/k_i(k_i - 1)$,
- **Betweeness centrality:** measures the extent to which a vertex lies on paths between other vertices.

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where, $\sigma_{st}(v)$ total number shortest path that passes through v
σ_{st} is total number shortest path between nodes s and t

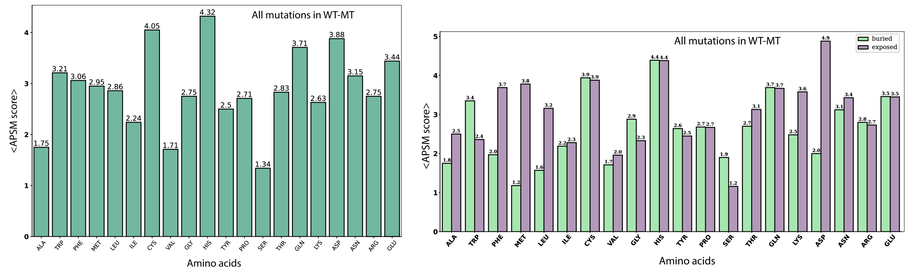


Classification of contacts based on the linear residue separation

To study distribution of contacts for each residue in single domain, we classified contacts based on linear separation [4] between contacting residues as:

Shorter Range: less than 6 residue apart
Medium range: 12-23 residue apart
Short range: 6 – 11 residue apart
Long range : > 23 residue apart

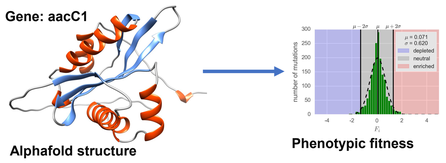
Case study of APSM score and phenotypic function



Distribution of average APSM score among all mutations in dataset A and score depednence on residue burial status

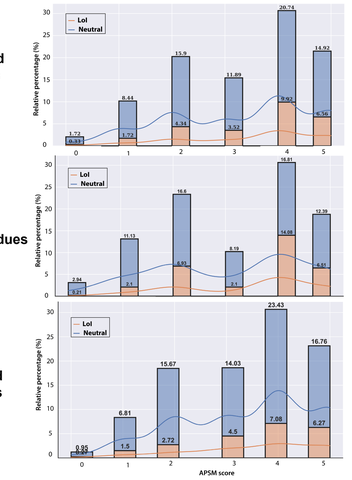
D. Coorespondence between APSM and phenotypic function

DMS studies on Gentamicin 3-N-acetyltransferase provided series of 2004 SRMs with their phenotypic effect at 37℃ in *E. coli.* We score these single residue mutations using APSM score and obtain correspondence with phenotypic fitness provided in the experimental studies[6]



LoF: defined when fitness score ≤ 1.311
Neutral: defined when fitness score is between and

a. Association between APSM score and protein function



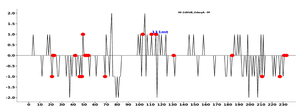
b. Residue level distribution of APSM and fitness

Results and Discussion

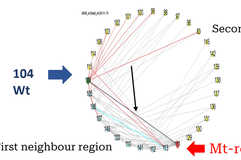
A. PCN Analysis of Selected WT-MT pairs

We selected 15 WT-MT pairs (having no significant structural changes) to construct their protein contact networks and analysed several network parameters to find residue contact perturbation due to mutation. Below is an representative example.

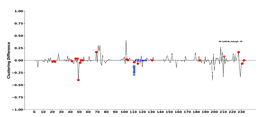
a. Degree difference



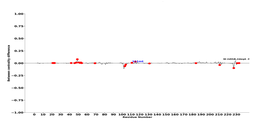
Ring network showing effect on first and second



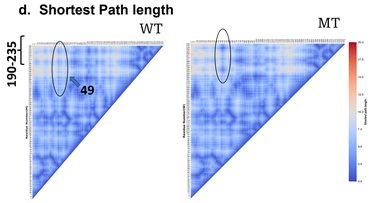
b. Clustering Coefficient difference



c. Betweenness Centrality difference



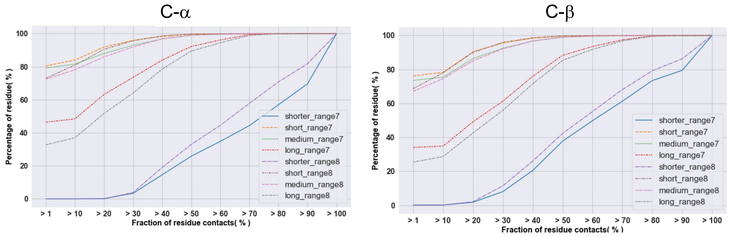
d. Shortest Path length



WT – 47,48,50,51,229 **MT – 47,48,50,51**

B. Distribution of contact types per residue

To understand whether short to long range contacts dominated in PCN, we classified types of contacts and examined their distribution in single domain proteins

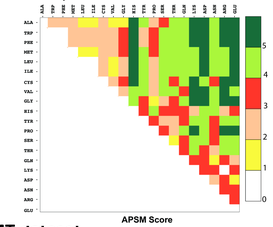


Most fraction of contacts per residues are medium/long range.

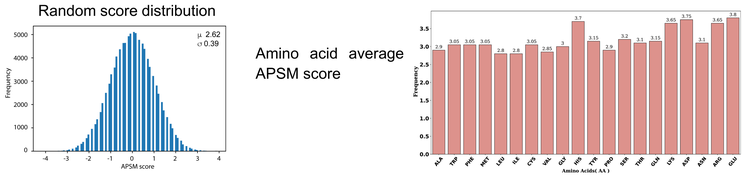
C. Construction of Physicochemical score matrix (APSM)

To quantitatively measure physiochemical change in SRMs, we developed an Amino acid Physiochemical Score matrix (APSM) [5], which can provides an interpretable metric.

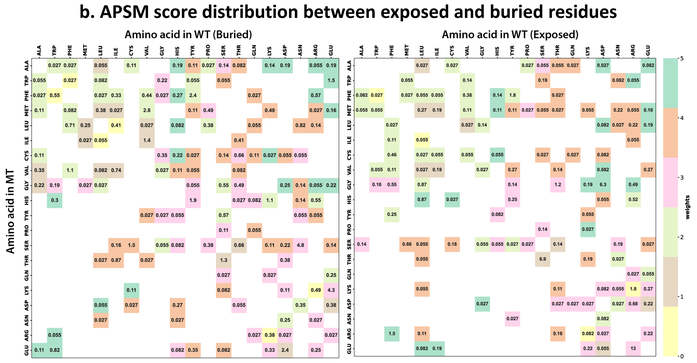
AMINO ACIDS	code	HYDRO	SIZE	CHEMICAl	CHARGE	POLARITY	CODE
Alanine	A	phobic	verysmall	aliphatic	neutral	neutral	ALA
Arginine	R	philic	large	basic	positive	polar	ARG
Asparagine	N	philic	small	amide	neutral	polar	ASN
Aspartic acid	D	philic	small	acidic	negative	polar	ASP
Cysteine	C	phobic	small	sulfur	neutral	neutral	CYS
Glutamic acid	E	philic	medium	acidic	negative	polar	GLU
Glutamine	Q	philic	medium	amide	neutral	polar	GLN
Glycine	G	neutral	verysmall	aliphatic	neutral	neutral	GLY
Histidine	H	neutral	medium	basic	positive	polar	HIS
Isoleucine	I	phobic	large	aliphatic	neutral	neutral	ILE
Leucine	L	phobic	large	aliphatic	neutral	neutral	LEU
Lysine	K	philic	large	basic	positive	polar	LYS
Methionine	M	phobic	large	sulfur	neutral	neutral	MET
Phenylalanine	F	phobic	verylarge	aromatic	neutral	neutral	PHE
Proline	P	neutral	small	aliphatic	neutral	neutral	PRO
Serine	S	neutral	verysmall	hydroxyl	neutral	polar	SER
Threonine	T	neutral	small	hydroxyl	neutral	polar	THR
Tryptophan	W	phobic	verylarge	aromatic	neutral	neutral	TRP
Tyrosine	Y	neutral	verylarge	aromatic	neutral	polar	TYR
Valine	V	phobic	medium	aliphatic	neutral	neutral	VAL



a. APSM score distribution in WT-MT dataset



b. APSM score distribution between exposed and buried residues



Occurrence of mutants in buried/exposed residues in WT-MT pairs and colored with APSM score

Conclusions

- In general, protein contact networks have limited ability to decipher affect of SRMs on protein structure of function. However, it can provide insights if the residue in the first/second shell shows perturbation on mutation.
- APSM score by itself is not able to discern effects of SRM on phenotypic function. We need to explore its combination with other features to predict phenotypic function.

References

1) (PDF) PROTEIN STRUCTURE: INSIGHTS FROM GRAPH THEORY (researchgate.net), http://14.139.227.205:8080/jspui/handle/123456789/1971

2) http://dunbrack.fccc.edu/lab/pisces

3) https://www.ncbi.nlm.nih.gov/pmc/articles/Phttps://www.academia.edu/en/18925120/Protein_cutoff_scanning_A_comparative_analysis_of_cutoff_dependent_and_cutoff_free_methods_for_prospecting_contacts_in_proteins

4) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4894841/#:~:text=Short%2Drange%20contacts%20are%20those,by%20at%20least%2024%20residues

5) https://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/IMGTclasses.html#:~:text=There%20are%205%20IMGT%20Volume,%2C%20C%2C%20P%2C%20T

6) https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007419#sec016