Identifying Functional Protein Domains Positions Using Population Variation

Anat Etzion-Fuchs[1] and **Mona Singh**[1,2]

[1]Lewis-Sigler Institute for Integrative Genomics; [2]Princeton University, Department of Computer Science, Princeton University
{anatf,mona}@princeton.edu

Abstract

Identifying essential protein residues is a major challenge in computational biology. we developed a computational framework that incorporates variation from large-scale population sequencing data, to highlight functional positions of protein domains.

Background

Challenges in identifying proteins "key" residue positions

- Structure-based methods - limited to ~30% of human proteins with a solved structure.
- Sequence-based methods – often miss non-conserved specificity-determining positions (SDPs):

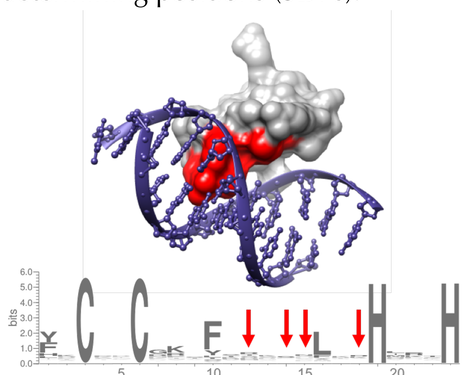


Fig 1: Non-conserved SDPs exemplified in the Cys_2-His_2 Zinc finger domain. Known[1] Functionally-important positions are colored in red on the domain surface[2], and pointed by red arrows on the domain sequence logo[3]. They are critical for the DNA-binding specificity, and are not conserved.

Results

Fig 3: Human genome coverage.
Most of the human genes have at least one protein domain (> 95%)[5].
Large fraction of the genes are covered by frequent domains $\left(\sim\frac{2}{3}\right)$.
Methods that rely on structure have Significantly lower coverage $\left(<\frac{1}{3}\right)$[5].

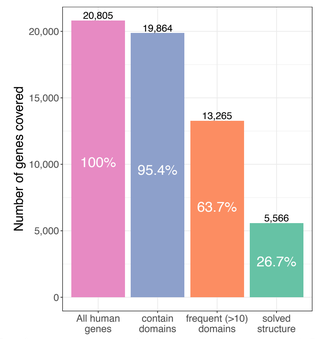


Fig 4: Population-based positional features.
Features calculated for protein domains (≥ 10 instances) positions: 409 domains, 5,842 binding and 44,692 neutral. Positional labels according to Kobren & Singh binding scores[6]. P-values are Wilcoxon one-tailed test.

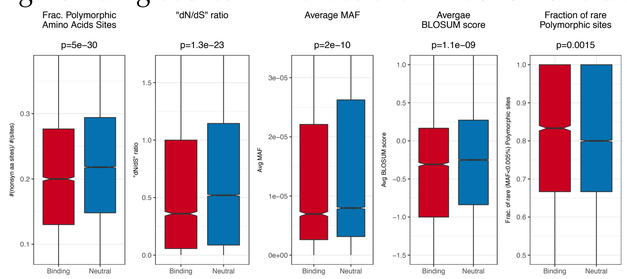
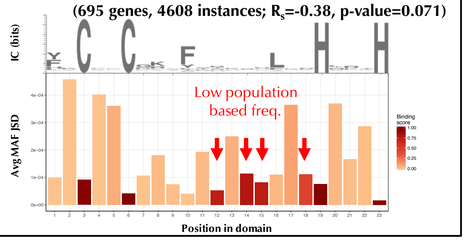


Fig 5: Predicting binding potential. Jensen-Shannon Divergence (JSD) per domain position, based on population allele frequency. The known SDPs in the C2H2-zinc finger domain have low JSD, as expected.



Significance

Systematic identification of "key" residue positions is critical for understanding protein **functions**, predicting **interactions**, and characterizing the **effect of mutations** in the context of **disease**.

Future plans

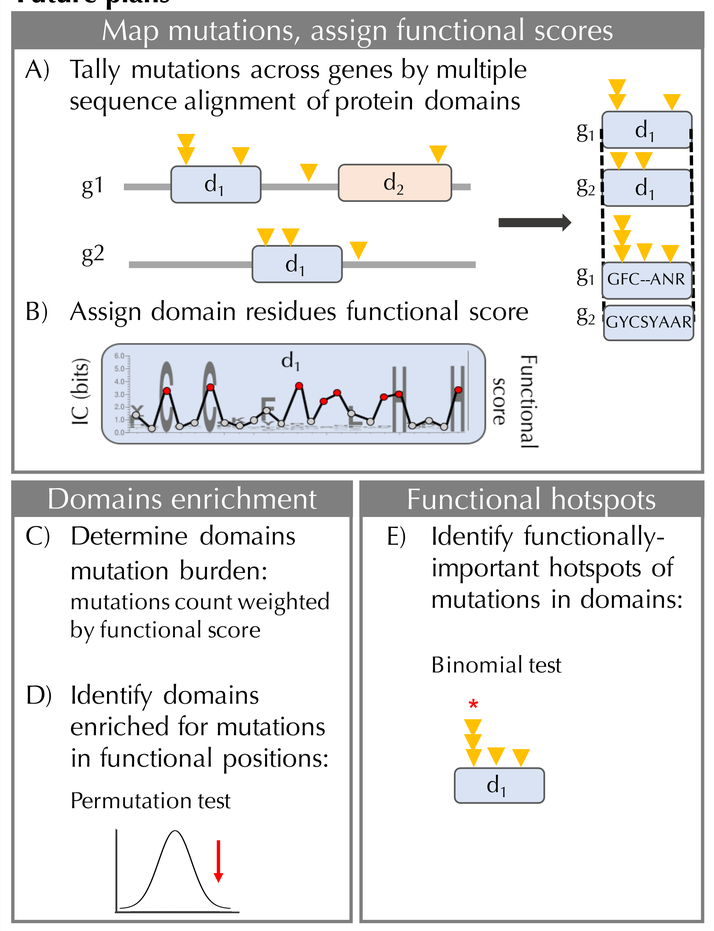
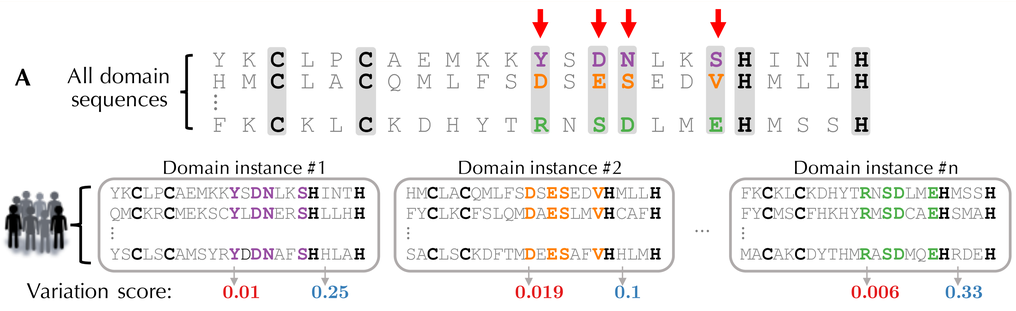


Fig 6: Proposed method for using domains positional functional scores in cancer mutations analysis.

Methods



Population-based domain-centered method
Predict functional importance from natural variations within sequencing data of 60,000 healthy individuals[4]:

- Single nucleotide polymorphisms (SNPs) in each proteins domain, from all protein sequences across all individuals are aggregated by domain position (Fig2A).
- Domain position distributions are analyzed to predict functional importance.

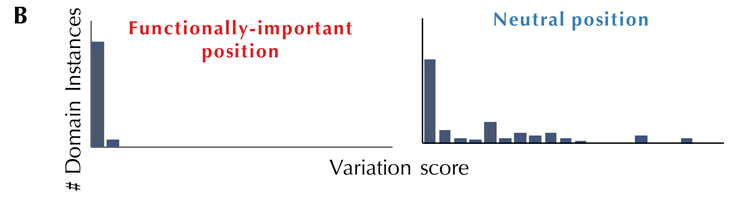


Fig 2: Outline of approach using the C2H2-zinc finger domain as an example. (A) Aggregation of SNPs population variation score across domain instances per domain position. (B) The distributions for two such positions are represented here as histograms.

***crucial protein domain positions are more conserved across the healthy population, even if they vary across proteins (Fig2B).**

References

[1] Wolfe et al. Ann. Rev. Biochem. (2001) [2] Pettersen et al. J. Comput. Chem. (2004) [3] Crooks et al. Genome Res. (2004) [4] Lek et al. Nature (2016) [5] Yates et al. Nucleic Acids Res. (2016) [6] Kobren and Singh (Unpublished).

Acknowledgements

We would like to thank Shilpa N. Kobren for helping with the computational framework, and the rest of the Singh lab for their helpful insights. Funded by T32 HG003284, NIH CA208148, and the Forese Family Fund for Innovation.