Statistical Methods on Incorporating Protein Structure Information with Rare Variant Association Studies

Abstract

Whole exome and whole genome sequencing association studies have long been used as a way to identify genes and variants that influence complex diseases and traits. However, despite their popularity, their statistical power is limited by a high background rate of neutral variants. In order to isolate rare variants more to improve the statistical power of analysis, many novel statistical methods have been developed. Three of these are POINT, PSCAN, and POKEMON, each of which use a unique method to identify the likelihood that a rare variant influences a trait or disease. With these three methods, I used a relatively well-studied set of genes to determine the statistical power of each of these three methods in identifying known disease-associated genes. Then, instead of using genes with experimentally mapped structures in the Protein Data Bank (PDB), I used AlphaFold2's revolutionary protein structure prediction algorithm, which allowed for more genes to have mapped structures, while some accuracy. My project analyzed the impact that AlphaFold2 had on the statistical power and accuracy of the three methods' analyses, seeing if the impact of AlphaFold 2 on rare variant analysis is as significant as the impact it had on the biological community as a whole.

Background

Rare genetic variants play an important role in complex diseases. Individual rare variants can be difficult to detect due to low frequencies of the mutant alleles though the genome-wide association study. There is a growing demand for an endeavor to effectively annotate the biological functions of genome-wide sequencing generated variants, since the variant of a gene will eventually impact the structure of a protein (Fig. 1) and can have varying impacts. Therefore, using protein structure as a function to predict the potential biological significance of genetic rare variants is a powerful method. There have been a number of statistical approaches developed to improve analysis using protein Structures from the Protein Data Bank. Some pioneer tests that utilized these protein structures were POINT, PSCAN, and POKEMON. All of them conceptually advance the field and shed light on its future. The introduction of AlphaFold 2, a revolutionized machine-learning protein folding algorithm, enormously boosts the biological field, including the application for these associationtests. My project thus aims to test if AlphaFold2 outperforms the limited empirical crystal structures. By evaluating this idea, a well-known tumor suppressor, *PTEN*, is used. *PTEN* gatekeeps the vital functions of the cell, and mutations can affect cell fate – survival or death, quiescent or proliferate (Fig. 2).

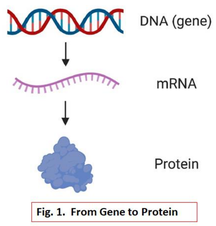


Fig. 1. From Gene to Protein

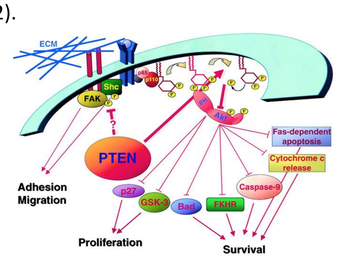


Fig. 2. Role of PTEN in the Control of Cell Death and Proliferation

Procedure

I chose to analyze the rare variants of the gene *PTEN.* As a widely studied tumor suppression gene, the experimental effects of the variants are more well known than a randomly selected gene from the human genome. I then made sure that the macromolecules formed from *PTEN* were fairly well mapped in the Protein Data Bank, and had corresponding structures predicted in the AlphaFold Database.

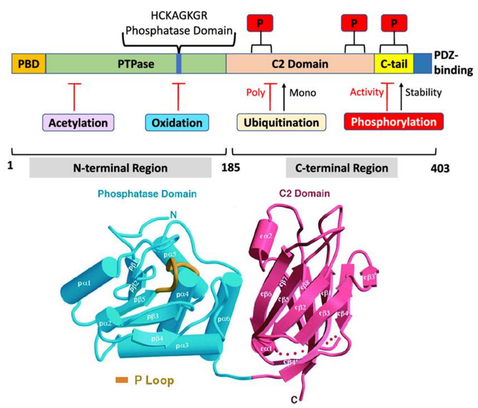


Fig. 3. PTEN Domains and 3D structure

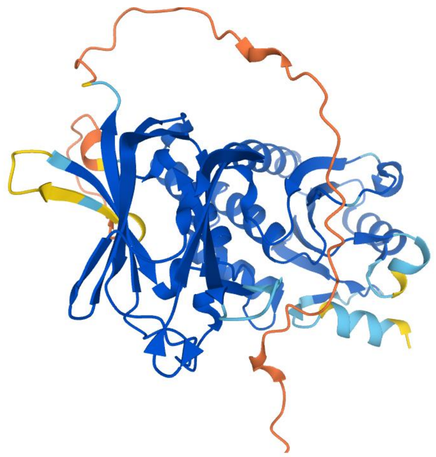


Fig. 4. PTEN 3D structure predicted by AlphaFold2

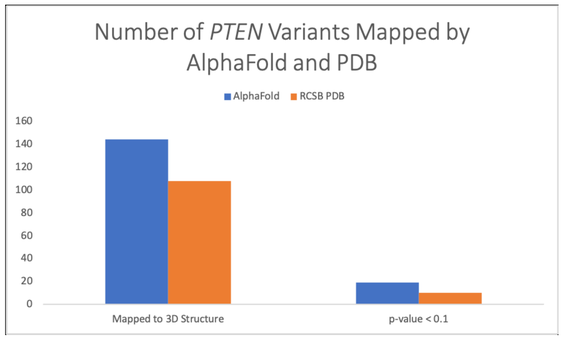
Once I knew that *PTEN* was suitable for analyzation, I used gnomAD to locate variants in *PTEN* that were within 75 base pairs of a coding exon. Then, I ran the list of variants through ENSEMBL's Variant Effect Predictor (VEP) to analyze the predicted consequences, as well as locate the PDB structure and location the variant mapped to. After this process, I manually filtered out common variants (MAF <1%), variants outside of coding regions, as well as synonymous variants. The reasoning for this was that although these variants still have a chance to be consequential to protein function, the rate of that occurring is much lower than with rare, non-synonymous variants.

With the filtered dataset, 145 rare variants of the *PTEN* gene remained. The next step was to map each variant to its 3D position on its protein tertiary structure, either retrieved from RCSB PDB or AlphaFold DB. In this process, the final variants were filtered out, based on whether or not they occurred on a mapped structure of the macromolecule.

Finally, enough data was collected to run simulation trials of POINT. 100 samples were ran under standard simulation parameters, using a randomly generated SNP matrix based on 1000 subjects. After running POINT, the minimum p-value across different distances was recorded. Then, for the purpose of comparison, the p-values and variants were mapped back to the original, annotated variant data set to analyze predicted and experimental results.

Results

When comparing the difference in the results of experimentally determined protein structures (RCSB PDB) and computationally predicted protein structures (AlphaFold DB), differences started occurring after mapping the variants to 3D coordinates. Out of the set of 145 variants, 108 of them were mapped to RCSB PDB structures. On the other hand, a substantial 144 of 145 variants were able to be mapped on AlphaFold's predicted structures (Table below).



Then, after running the 100 trials of POINT, variants with a minimum p-value of less than 0.1 were considered, as these were the ones that POINT determined were most likely to be associated with complex diseases. In the association test conducted using AlphaFold DB, there were 19 variants that fulfilled this requirement, while there were only 10 variants that had a p-value under 0.1 when the RCSB PDB was used.

Of the 10 variants determined using RCSB PDB, all were predicted, via a plethora of other association tests and experimental determination, to have at least moderate consequences, while 10% (1 of 10) were predicted to have severe consequences. Of the 19 variants determined using AlphaFold DB, once again, all were determined to have at least moderate consequences, while 10.5% (2 of 19) were determined to have severe consequences.

Conclusion

Based on the results of my *PTEN* simulation, AlphaFold appears to improve the results of association studies. During the procedure, AlphaFold was able to map 3D coordinates to significantly more variants than RCSB PDB. As a result, when using AlphaFold, the number of variants was almost double that when using RCSB PDB. When evaluating the accuracy of these results, all were predicted to have at least moderate consequences, and the percentage that had severe consequences hovered at 10%.

The results indicate that although AlphaFold may not increase the accuracy of association studies, it certainly improves the number of variants that can be mapped to a protein tertiary structure, increasing the number of variants that pass the p-value requirement. Therefore, my simulation indicates that AlphaFold 2 does in fact improve the results association tests over experimentally determined PDB structures.

Future Research

My research was mainly focused on the gene *PTEN* and the association test POINT. In the future, I would like to expand this. I am currently working on simulations using *PTEN* on PSCAN and POKEMON, two other prominent rare variant association tests. After simulating *PTEN* on those tests, I would like to move onto testing different genes, more than one gene at once, and move out of simulation parameters to hopefully gain additional and more accurate insight on the effect of AlphaFold 2 on rare variant association studies.

Additionally, I would like to improve the code I currently have to make it more streamlined, organized, and accessible, hopefully creating a pipeline in the future. One of the main problems I ran into when writing my code was the brevity with which the relevant research papers covered the parameters necessary to run their test, which made the process more difficult. By writing a pipeline, I hope to make this process more accessible for others in the field of bioinformatics.

References

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank *Nucleic Acids Research*, 28: 235-242.

Jin, B., Capra, J.A., Benchek, P., Wheeler, N., Naj, A.C., Hamilton-Nelson, K.L., Farrell, J.J., Leung, Y.Y., Kunkle, B., Vadarajan, B., et al. (2021). An Association Test of the Spatial Distribution of Rare Missense Variants within Protein Structures Improves Statistical Power of Sequencing Studies. BioRxiv 2021.08.09.455695.

Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021).

Karczewski, K.J., Francioli, L.C., Tiao, G. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020).

Marceau West R, Lu W, Rotroff DM, Kuenemann MA, Chang SM, et al. (2019) Identifying individual risk rare variants using protein structure guided local tests (POINT). PLOS Computational Biology 15(2): e1006722.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biology Jun 6;17(1):122. (2016) doi:10.1186/s13059-016-0974-4

Tang, ZZ., Sliwoski, G.R., Chen, G. *et al.* PSCAN: Spatial scan tests guided by protein structures improve complex disease gene discovery and signal variant detection. *Genome Biol* **21,** 217 (2020).

Varadi, M et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research (2021).

Acknowledgements

I deeply appreciate Prof. Hongyu Zhao from Yale University, who offered me an internship opportunity and was a great mentor for my research. Additionally, special thanks to Yuhan Xie, Yale University, for her direct supervision and extreme patience while I was working on the project.