Structure-based modeling of cysteine and serine disease variants of human proteome

D. Podgalo
Smolensk State Medical University

Introduction. During the early 1980s, the ability to rationally design drugs using protein structures was an unrealized goal for many structural biologists. The first projects were underway in the mid-80s, and by the early 1990s the first success stories were published. Today, even though there is still quite a bit of fine-tuning necessary to perfect the process, structure-based drug design is an integral part of most industrial drug discovery programs and is the major subject of research for many academic laboratories.

The completion of the human genome project, the start of both the proteomics and structural genomics revolutions, and developments in information technology are fueling an even greater opportunity for structure-based drug design to be part of the success story in the discovery of new drug leads. Excellent drug targets are identified at an increased pace using developments in bioinformatics. The genes for these targets can be cloned quickly, and the protein expressed and purified to homogeneity. Advances in high-throughput crystallography, such as automation at all stages, more intense synchrotron radiation, and new developments in phase determination, have shortened the timeline for determining structures. Faster computers and the availability of relatively inexpensive clusters of computers have increased the speed at which drug leads can be identified and evaluated in silico [1].

There are many disease-associated mutations that endow pharmacological target (typically a protein), drug resistance, e.g. G12C amino acid substitution in oncogenic target KRAS [10]. People carrying such mutations may need the development of personalized drugs that take into account structural peculiarities of the mutated protein. One of the promising strategies is to develop covalent drugs that are specific for a given mutation [9].

The goal of this project is to model structures of human proteins with disease-associated amino acid substitutions. Two types of amino acid substitutions are selected: X to cysteine or X to serine (X is any amino acid residue) – these residues are often used as the attachment points for covalent drugs.

Materials and methods.

We used the following software:

1. Conda [4]
2. Ensembl Variant Effect Predictor (VEP) to predict amino acid substitution based on single nucleotide polymorphism [11]
3. Rosseta ddg_monomer to model mutant proteins [2]
4. SCons to install Rosetta ddg_monomer [6]
5. UniProt Retrieve/ID mapping to map VEP-annotated proteins ID to UniProt ID [12]

We used the following databases:

1. gnomAD 2.1.1 ExAC (a database of single nucleotide polymorphism) with 9362318 variants (60705 exomes) for human genome assembly GRCh37/hg19 [7]
2. VEP human database GRCh37/hg19 for VEP-annotation
3. ClinVar database for disease detection [8]
4. AlphaFold2 database – normal sturcture-based models of human proteins [5]

Results.

We annotated all 9362318 gnomAD variants using VEP and then also using VEP filtered these variants to choose only missense mutations on coding sequence with amino acid substitution to cysteine or serine. Also we chose only pathogenic variants that are reliably known from the disease.

After this we got 1339 cysteine and serine variants associated with the disease. Then each variant we linked with the AlphaFold2 model and using Rosseta ddg_monomer modeled mutant protein.

We also carried out statistical processing of the results. Most diseases are associated with proteins: DYHC2 (8.36 %), USH2A (7.84 %), VWF (3.36 %). The most frequently

substituted amino acid is arginine (52.05 %), also commonly substituted amino acids are glycine (14.71 %) and tyrosine (10.9 %). In 66.69 %, the substitution was made with cysteine, in 33.31 % one was made with serine. Most frequent diseases associated with amino acid substitution in proteins are asphyxiating thoracic dystrophy (6.05 %), primary ciliary dyskinesia (4.03 %) and retinal dystrophy (4.03 %).
All our results and pipeline of this work can be found in the GitHub repository: DmitriiPodgalo/POP.

Discussion. The obtained structural models will be used as the starting conformations for the structure-based drug design pipelines [3].
Structure-based drug design is a powerful method, especially when used as a tool within an armamentarium, for discovering new drug leads against important targets. After a target and a structure of that target are chosen, new leads can be designed from chemical principles or chosen from a subset of small molecules that scored well when docked in silico against the target. After a preliminary assessment of bioavailability, the candidate leads continue in an iterative process of reentering structural determination and reevaluation for optimization. Focused libraries of synthesized compounds based on the structure-based lead can create a very promising lead which can continue to phase I clinical trials.
As structural genomics, bioinformatics, and computational power continue to explode with new advances, further successes in structure-based drug design are likely to follow. Each year, new targets are being identified, structures of those targets are being determined at an amazing rate, and our capability to capture a quantitative picture of the interactions between macromolecules and ligands is accelerating [1].

References.

1. Anderson, Amy C. "The process of structure-based drug design." *Chemistry & biology* 10.9 (2003): 787-797.
2. Barlow, Kyle A., et al. "Flex ddG: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation." The Journal of Physical Chemistry B 122.21 (2018): 5389-5399.
3. Cohen, Nissim Claude. "Structure-based drug design and the discovery of aliskiren (Tekturna): perseverance and creativity to overcome a R&D pipeline challenge." Chemical biology & drug design 70.6 (2007): 557-565.
4. Grüning, Björn, et al. "Bioconda: sustainable and comprehensive software distribution for the life sciences." Nature methods 15.7 (2018): 475-476.
5. Jumper, John, et al. "AlphaFold 2." (2020).
6. Knight, Steven. "Building software with SCons." Computing in Science & Engineering 7.1 (2005): 79-88.
7. Koch, Linda. "Exploring human genomic diversity with gnomAD." Nature Reviews Genetics 21.8 (2020): 448-448.
8. Landrum, Melissa J., et al. "ClinVar: public archive of interpretations of clinically relevant variants." Nucleic acids research 44.D1 (2016): D862-D868.
9. Li, Biao, et al. "Automated inference of molecular mechanisms of disease from amino acid substitutions." Bioinformatics25.21 (2009): 2744-2750.
10. Lievre, Astrid, et al. "KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer." Cancer research 66.8 (2006): 3992-3995.
11. McCarthy, Davis J., et al. "Choice of transcripts and software has a large effect on variant annotation." Genome medicine 6.3 (2014): 1-16.
12. Pundir, Sangya, et al. "UniProt tools." Current protocols in bioinformatics 53.1 (2016): 1-29.