LLNL-TR-846348

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Rapid Computational Identification of Therapeutic Targets for Pathogens

J. E. Allen

March 15, 2023

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Rapid Computational Identification of Therapeutic Targets for Pathogens

Jonathan E. Allen (20-ERD-062)

Abstract

Biological threats continue to persist and evolve as an important challenge to national security. There are multiple ways in which novel viral pathogens could emerge to pose a serious threat to human health. This project developed a pathogen target identification tool that can rapidly respond to a novel or emerging viral biological threat. A set of computational tools were developed that provide detailed information on the newly sequenced genes, their protein products and the drug target sites for the proteins that are best suited for biological countermeasure development. Three key innovations were developed in the project. 1) Development of a new extensive database of protein pocket structures with structure-based search algorithms to rapidly link novel protein targets with the complete collection of previously experimentally solved protein structures. 2) A novel clustering pipeline was introduced to group matching structures and associated small-molecule binding ligands into a consensus protein pocket with the associated small-molecule chemotypes predicted to fit in the pocket site. The matching experimentally solved structures were used to inform the value of different target sites. 3) Where there are viral protein targets with pockets structurally matched to similar human proteins, a biological knowledge graph, which links molecular interactions with human disease, was used to further assess the potential negative impact of a viral protein target with similarities to human proteins that could have important off target side effects. In total, the project produced a new resource for rapid and detailed assessment of promising targets for countermeasures, reflecting the ongoing wet lab, clinical, and computational data being collected. These capabilities will improve the ability to respond to a biological threat in multiple domains.

Background and Research Objectives

There is a growing collection of new experimentally derived data being deposited in public databases, which can be used to help identify and characterize novel biological threats. The growth for publicly available protein structures has more than doubled over the last ten years from 95,509 to 202,292 structures currently (https://www.rcsb.org/stats/growth/growth-released-structures) . In addition to experimentally solved protein structures, there is a steady accumulation of biological data measuring molecular interactions and their links to human disease, drug treatments, and potential side effects of therapeutics. A database of molecular and disease data can be compiled into a graph database of relationships to link relationships between proteins, genes, drugs, and disease. An example of the schema for the graph database, called the Scalable Precision Medicine Open Knoweldge Engine (SPOKE), used in this project is shown in Figure 1 and shows how the disruption of protein function can be linked to disease and adverse health outcomes (Himmelstein & Baranzini, 2015).

The research objectives of this project were to develop new methods and insights into assessing a novel viral genome for druggable targets and to provide tools to help assess which protein regions to target and which regions to avoid.

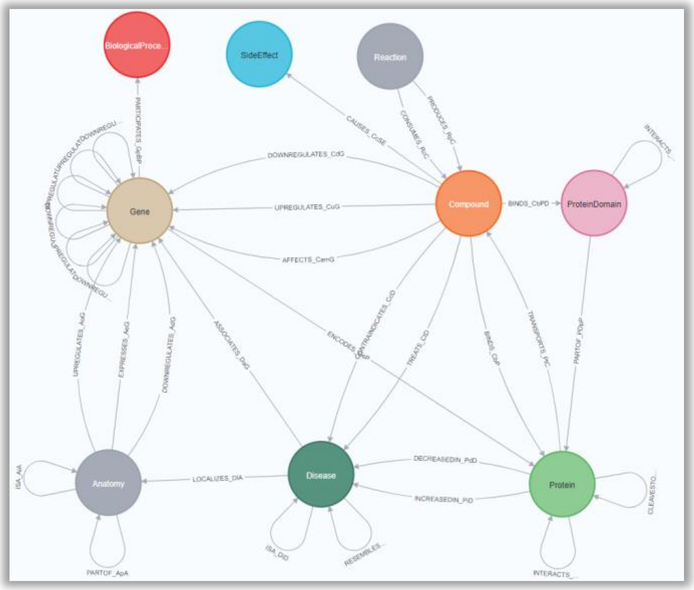

Figure 1. Schema for graph database showing experimentally derived relationships between molecular entities and human disease.

Scientific Approach and Accomplishments

Three technical areas were explored for demonstrating improved assessment of novel protein targets.   The first area was to investigate the use of protein and chemical structure similarity search between the new viral proteins and previously solved structures.   To this end, a novel structure based, updatable data resource was created called PDBspheres (Zemla et al., 2022). PDBspheres is a computational system and data resource for enumerating through the complete collection of experimentally solved structures available in the Protein Data Bank (PDB) (consortium, 2018). The system identifies each small-molecule (ligand) protein interface and extracts the structural region for a 12-angstrom radius from the center of the ligand. This generates a set of localized exemplars for ligand-protein interfaces which can support searching against novel protein structures. Rather than attempting to structurally align the whole protein, the goal is to focus on the key functional interface which can allow the searching algorithms to ignore other parts of the protein that less directly impact the binding. Since a single protein structure can have multiple ligand binding sites, there are currently 2 million searchable structural exemplars included in the system.

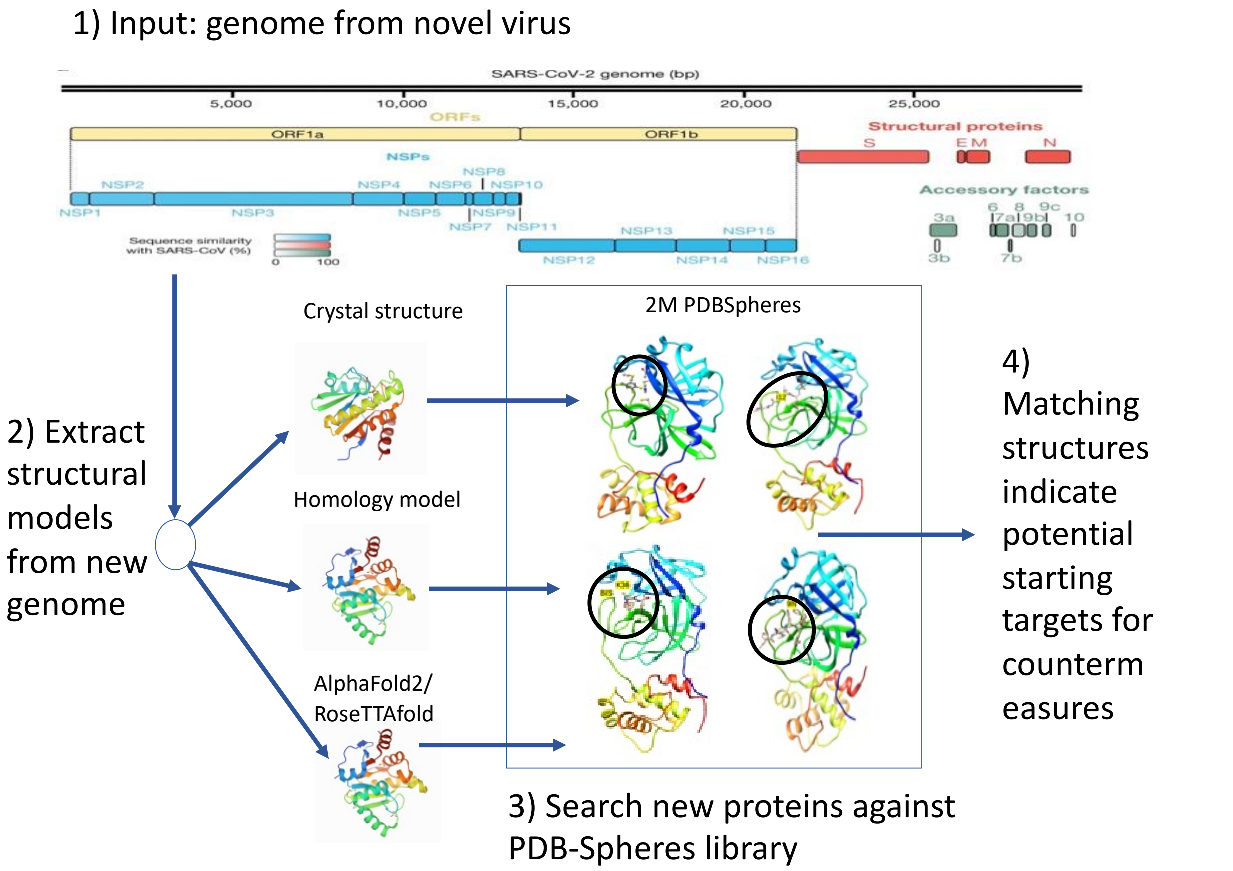

Figure 2. Part one of the pipeline described in four steps. 1) viral genome is sequenced, 2) protein sequences are extracted and 3D structures are predicted from different sources (crystal structures, homology models, AI/ML predictions), 3) new viral protein structures are searched against the PDBspheres library and 4) matches are reported as potential sites for counter measure targeting.

An example of the procedure for matching new viral proteins to the PDBspheres library is shown in Figure 2. While the reporting of structural matches of previously seen structures to the new viral protein can provide interesting information, it remains challenging to summarize the significance of the possibly hundreds of different matches to each protein sequence and in turn assess the relative importance of evidence supporting the targeting of one protein over another. For this reason, we embarked on exploring a second advance, which focused on clustering the results from the initial search to come up with consensus target sites and their associated chemotypes.   A schematic for the pipeline for clustering the resulting structure matches is shown in Figure 3. The challenge here was to devise an effective clustering scheme for the protein pocket, highlighting the regions of the pocket that are observed to be in physical proximity to previously solved ligand-protein structures and separately, clustering the ligands associated with these sites to summarize both the chemical type associated with the binding site and the consensus region of the binding site. This method is applied to the viral proteome to rapidly generate a list of targetable protein sites and associated chemotypes for initiating a drug discovery campaign. An example is shown in Figure 4 for the SARS-CoV-2 main protease target site, where 15 distinct chemotypes are identified and listed by their decreasing molecular weight. The middle plot in the figure shows the fraction of active compounds for a specific chemotype that have been shown experimentally to be active inhibitors. The results highlight potentially novel chemotypes to consider for the target site including the chemotypes labeled 1, 2, 4 and 5. In addition, the identified consensus pocket sites and respective chemotypes can be compared against observed mutations in the binding site to identify potentially significant viral mutations that may induce antiviral resistance. An example of this analysis was explored for clinical SARS-CoV-2 samples collected in California to find potential evidence for antiviral resistance potential (Kimbrel et al., 2022).

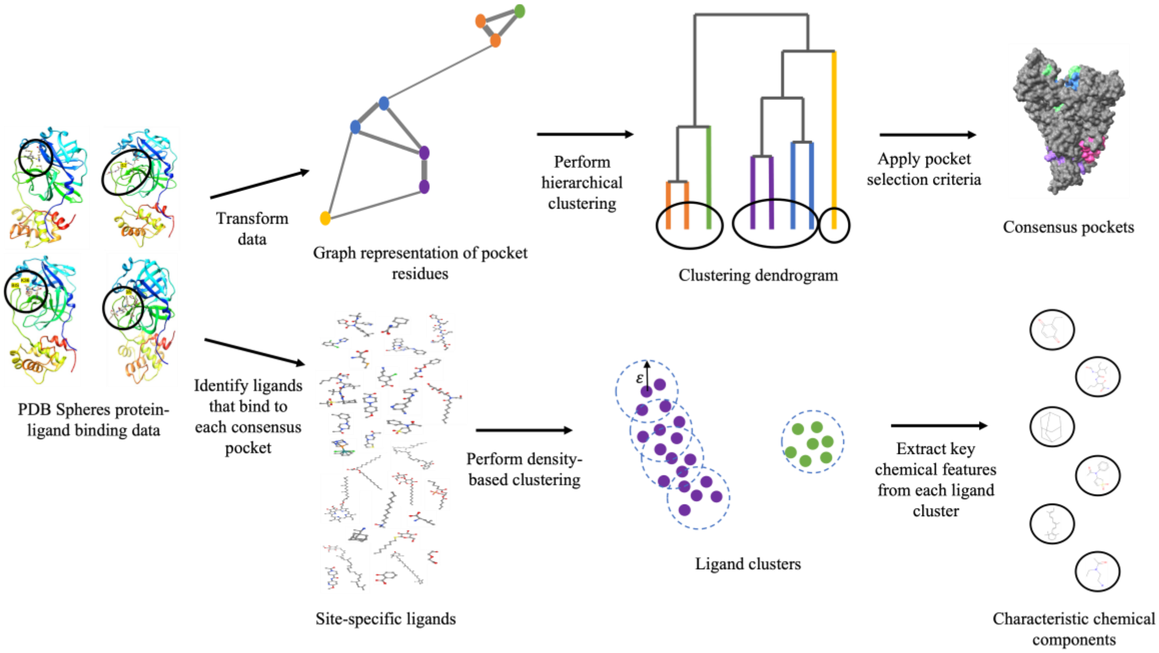
PDB Spheres protein-ligand binding data
Transform data
Graph representation of pocket residues
Perform hierarchical clustering
Clustering dendrogram
Apply pocket selection criteria
Consensus pockets
Identify ligands that bind to each consensus pocket
Site-specific ligands
Perform density-based clustering
ε
Ligand clusters
Extract key chemical features from each ligand cluster
Characteristic chemical components

Figure 3. Pipeline for clustering both the structural matches using both the bound ligands and protein pockets to give a summary of consensus pocket and chemotypes.

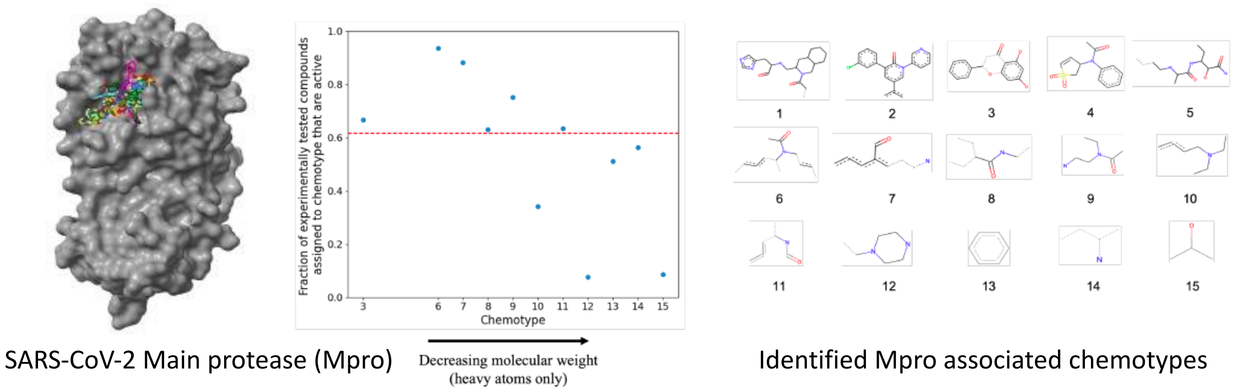

Figure 4. Summary of SARS-CoV-2 Main protease (Mpro) target site and associated chemotypes.

Having established a computational system for identifying the potential target sites in the viral proteome, the final step was to investigate use of a biological knowledge network such as the one shown in Figure 1 to make predictions on potential drugs or targets using orthogonal non-structure based biological data. We explored a dataset of compounds that were carefully labeled for their observed impact on causing Drug Induced Liver Injury (DILI). A collection of 1,139 compounds were evaluated, 757 “safe” drugs and 382 DILI positive drugs (Posada et al., 2022). Nodes in the graph included compounds connected to proteins they interact with as well as disease types, molecular pathways, genes, and side effects. For each drug, a random walk is taken on the graph to create a graph encoded feature vector using an algorithm called node2vec (cite). The graph embedding is meant to create a biological descriptor for each drug, which can be used to train a machine learning classifier. The results from the analysis showed that prediction model accuracy could be as high as 84% (citation).

Lastly, we explored the potential to evaluate off target interactions by modeling compounds that structurally fit both the viral protein target and off target human proteins. As an example, the SARS-CoV-2 Mpro protein target is used and compared with the related MERS Mpro protein target and each protein in the human proteome for which a structure can be produced. Each matching human protein structure is evaluated with the biological knowledge graph using Yen’s shortest path algorithm starting from the overlapping human protein and traversing the graph to the closest nodes representing related side effects of concern. The expectation is that a drug targeting the viral protein could also bind to the human protein and the knowledge graph can assess the biological significance of disrupting the human protein’s function by measuring the protein’s close proximity in the graph to an adverse health outcome. For the protease example, the SARS-CoV-2 main protease protein pocket was found to share significant similarity with 124 human proteins across 377 distinct ligands, and each of the matching human proteins turns out to be a member of the protease family. As a positive control, we can look at the active component in the FDA approved SARS-CoV-2 main protease inhibitor Paxlovid and see significant structural similarities to binding pockets of 7 human proteases, but the structural similarity of the binding site is low (slightly above 50%) based on a structural similarity metric called the Global distance calculation (GDC), which measures the fraction of atoms that align between the two structures(GDC values range from 0-100 with 100 being a perfect match). Moreover, all of the human proteases show a path length greater than two to any adverse side effect. In contrast, while other human proteases directly connected to a ligand of interest were not found to have a direct association with adverse side effects, in some cases the structural pocket site similarity was much higher (GDC=89%), indicating a stronger potential for that ligand to interfere with a small subset of human proteases.

Impact on Mission

Our collection of software tools is designed to effectively search existing experimental biological data to rapidly report on the protein sites to target for novel viral pathogens. The software exploits the high-performance computing resources by employing parallelized protein structure search algorithms, which allow new target proteins to be searched against an extensive reference collection of protein-ligand structure interfaces. This work is supporting the high-performance computing, simulation and data science core competency. In addition, this is supporting a core capability in advancing national security concerns in support of bioassurance and biosecurity. It is expected that the newly developed tools will provide important insights in prioritizing protein targets as novel pathogens may emerge with little knowledge on therapeutic development.

Conclusion

The potential to rapidly produce an antiviral for any new viral pathogen remains an important challenge. The new tools developed in this project provide a novel data driven approach to assessing individual protein targets for therapeutic development. As more experimental data is collected, these data driven approaches will continue to improve and provide improved levels of detailed evidence for focusing countermeasure development on specific proteins. Future extensions of this work are to improve integration of more detailed biophysical modeling of protein-ligand interactions, recognizing new protein-interaction sites (e.g. protein- protein interactions and others), and develop additional knowledge network profiles for defining negative health outcomes. The basic work in this project has contributed to several proposals in response to calls from the Department of Defense and the Department of Energy.

Notes to the Editors

References

consortium, w. (2018). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, *47*(D1), D520-D528. https://doi.org/10.1093/nar/gky949

Himmelstein, D. S., & Baranzini, S. E. (2015). Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLoS Comput Biol*, *11*(7), e1004259. https://doi.org/10.1371/journal.pcbi.1004259

Kimbrel, J., Moon, J., Avila-Herrera, A., Martí, J. M., Thissen, J., Mulakken, N., Sandholtz, S. H., Ferrell, T., Daum, C., Hall, S., Segelke, B., Arrildt, K. T., Messenger, S., Wadford, D. A., Jaing, C., Allen, J. E., & Borucki, M. K. (2022). Multiple Mutations Associated with Emergent Variants Can Be Detected as Low-Frequency Mutations in Early SARS-CoV-2 Pandemic Clinical Samples. *Viruses*, *14*(12), 2775. https://www.mdpi.com/1999-4915/14/12/2775

Posada, R., Silva, M., Torres, M., Allen, J., Drocco, J., Sandholtz, S., & Zemla, A. (2022). Graph-based featurization methods for classifying small molecule compounds. *14*. https://doi.org/https://doi.org/10.5070/m414157338

Journal Name: UC Merced Undergraduate Research Journal; Journal Volume: 14; Journal Issue: 1

Zemla, A. T., Allen, J. E., Kirshner, D., & Lightstone, F. C. (2022). PDBspheres: a method for finding 3D similarities in local regions in proteins. *NAR Genomics and Bioinformatics*, *4*(4). https://doi.org/10.1093/nargab/lqac078

Publications and Presentations

- Accepted - "A Computational Pipeline to Identify Broad-Spectrum Drug Targets and Interacting Chemotypes in Viral Pathogens," Gordon Research Seminar on Chemical and Biological Defense, March 19, 2023
- "A Computational Pipeline to Identify and Characterize Binding Sites and Interacting Chemotypes in SARS-CoV-2," American Chemical Society Fall 2022 National Meeting, August 22, 2022
- "A Computational Pipeline to Identify Broad-Spectrum Drug Targets and Interacting Chemotypes in Viral Pathogens," DTRA CBD S&T Conference, December 8, 2022