HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 3 (2023)

Computational Elucidation of Protein-Protein Interactions in the Minimal Proteome of JCVI-syn3A

Arnav Meduri[1], Abhinav Meduri[1], and Keith Robison[#]

[1]North Caroline School of Science and Mathematics
[#]Advisor

ABSTRACT

Proteins play a vital role in the regulation of biological processes, facilitating the transfer of intermediates and coordinating the sequential steps of biochemical pathways. Protein-protein interactions (PPIs) are crucial molecular events in which two or more proteins bind together, enabling the formation of protein complexes that govern various cellular activities, including signal transduction, gene expression, and enzymatic reactions. Evolutionary correlations arise due to the close proximity of amino acid side chains within these interactions, where amino acids on one side of an interaction surface may restrict which amino acids fit on the other side or encourage mutations that modify the surface. In this study, our aim is to investigate the correlation between protein sequence and structure in *mycoplasma mycoides* JCVI-syn3A, a minimal cell consisting of 493 genes. We utilize the EVCouplings framework, a coevolution-based approach with probabilistic scores for residue interactions, to predict protein-protein interactions and the specific surfaces that govern them. Our study demonstrates that coevolution-based computational methods can predict protein-protein interactions and their interaction surfaces. After analyzing multiple sequence alignment (MSA) data across 110 protein families, we identify a total of 33 inter-protein interactions. Our analysis of the protein-protein interactions in JCVI-syn3A provides valuable insights into the genetic architecture of *Mycoplasma*, one of the simplest cellular life forms known, and enhances our understanding of how the earliest cellular life forms might have functioned.

Introduction

Genome minimization is a synthetic biology approach that involves reducing the genome of an organism to its minimal gene set. The primary goal of genome minimization is to understand the fundamental physiological processes that give rise to life and to develop minimal organisms that can serve as useful tools in biotechnology and synthetic biology. In the pursuit of creating the first artificial cell with a fully synthesized minimal genome, a bottom-up approach was initiated by the J. Craig Venter Institute in 2016, utilizing JCVI-syn3.0 derived from the natural genome of *M. mycoides* (Hutchison et al., 2016). The J. Craig Venter Institute utilized global transposon mutagenesis, a technique used to introduce random mutations throughout the genome of an organism using transposons, to identify the essential genes to be retained in JCVI-syn3.0 (Hutchison et al., 1999). More recently, the concept of genome minimization has been driven by advancements in DNA synthesis, sequencing technologies, and computational tools for genome analysis. By systematically removing nonessential genes, researchers can identify core genetic elements necessary for cellular function and explore the functional dependencies within a genome.

Why did we choose JCVI-syn3A?

The reductionist approach provided by genome minimization has proven to be highly effective in scientific analysis. Scientists use reductionist methods to simplify complex phenomena that are not well understood. This approach not only enhances our understanding of scientific phenomena but also facilitates their reconstruction, thereby yielding

successful outcomes. In chemistry, this approach involves deconstructing a substance into its pure components and understanding the function of each. For instance, the periodic table provides a predictive framework for all elements, where the rows and columns make specific predictions about their properties and behaviors.

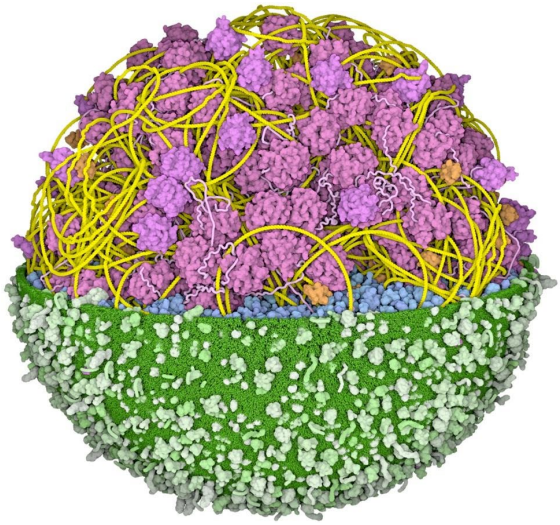

Figure 1. Whole-Cell Model of JCVI-Syn3A
Credit: Ludovic Autin

Despite recent advancements in genome minimization, understanding of cells at the molecular level remains limited. The JCVI-syn3.0 project successfully created a functional bacterium with only 473 genes, making it one of the smallest self-replicating cells ever produced. JCVI-syn3A, a mutant of JCVI-syn3.0, has only 19 additional genes and exhibits nearly normal morphology (Breuer et al., 2019). Moreover, it is the simplest cell that can be cultivated in the laboratory using media composed of pure components. Although we have identified all the genes in its genome, the precise function of many of these genes remains unknown. Further research into the functional organization of these genes and their products is essential to understanding the fundamental architecture of a simple cell.

Protein Folding and Multiple Sequence Alignment

At the molecular level, proteins are essential components that are involved in a wide range of biological processes, such as catalyzing chemical reactions, transmitting signals, and providing structural support. The unique three-dimensional structure of proteins plays a critical role in their ability to execute specific functions. By predicting a protein's three-dimensional structure, scientists can infer its function and how it interacts with other molecules, including drugs. Designing and testing new proteins with specific properties, such as increased stability or improved binding affinity, can have far-reaching applications in biotechnology and medicine.

To gain insights into protein structure and function, multiple sequence alignment (MSA) has become a vital bioinformatics tool that enables researchers to compare and analyze large volumes of genetic information. By aligning protein sequences derived from genome sequences, scientists can identify conserved regions, motifs, and domains that reveal evolutionary relationships, structural features, and functional implications. MSA can help researchers to identify drug-binding sites in proteins, enabling the design of novel drugs with high binding affinity. With this technique,

scientists can better understand how proteins interact with each other, design new drugs, and create novel biotechnological applications.

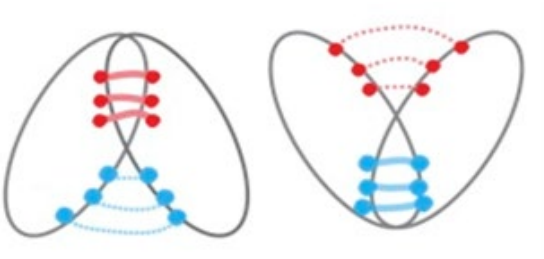

Figure 2. Schematic of Protein-Protein Interactions
Credit: Debbie Marks Lab

The EVCouplings Framework

Evolutionary Coupling (EC) is a computational method developed in the 2010s. The Evolutionary Couplings framework (EVCouplings) is an open-source bioinformatics tool that uses MSAs to predict co-evolving residues in protein families (https://github.com/debbiemarkslab/EVcouplings). The EC method aims to identify functional residues that are coevolving across different organisms, suggesting that they play a critical role in the protein's structure and function. The EVCouplings framework utilizes a statistical model that compares the evolutionary patterns of amino acids in the MSA to identify co-evolving residue pairs. The statistical model utilizes an algorithm that computes a score for each residue pair, which reflects the strength of the co-evolutionary signal between them. The resulting score matrix can be used to predict the 3D structure of a protein and identify residues that are critical for its function. By analyzing patterns of amino acid co-variation in an alignment of putatively interacting proteins, evolutionary couplings between co-evolving inter-protein residue pairs can be identified and the interaction surface can be mapped.

Research Questions

1. Can sequence coevolution be used to determine probable protein-protein interactions in JCVI-syn3A?
2. Can we identify inter-protein interaction surfaces between a subset of proteins in JCVI-syn3A?
3. Can we map the interaction surfaces deduced through statistical methods onto known 3D protein structures?

Methodology

Computational methods offer a cost-effective and scalable means of analyzing complex data, which can provide insights into experimental design and optimization. To this end, we utilized the EVCouplings framework in our analysis, which depends on the existence of diverse protein sequences in the proteome. The EVCouplings method is implemented using Python scripting and is capable of generating high-quality three-dimensional renderings, which were further analyzed using PYMOL.

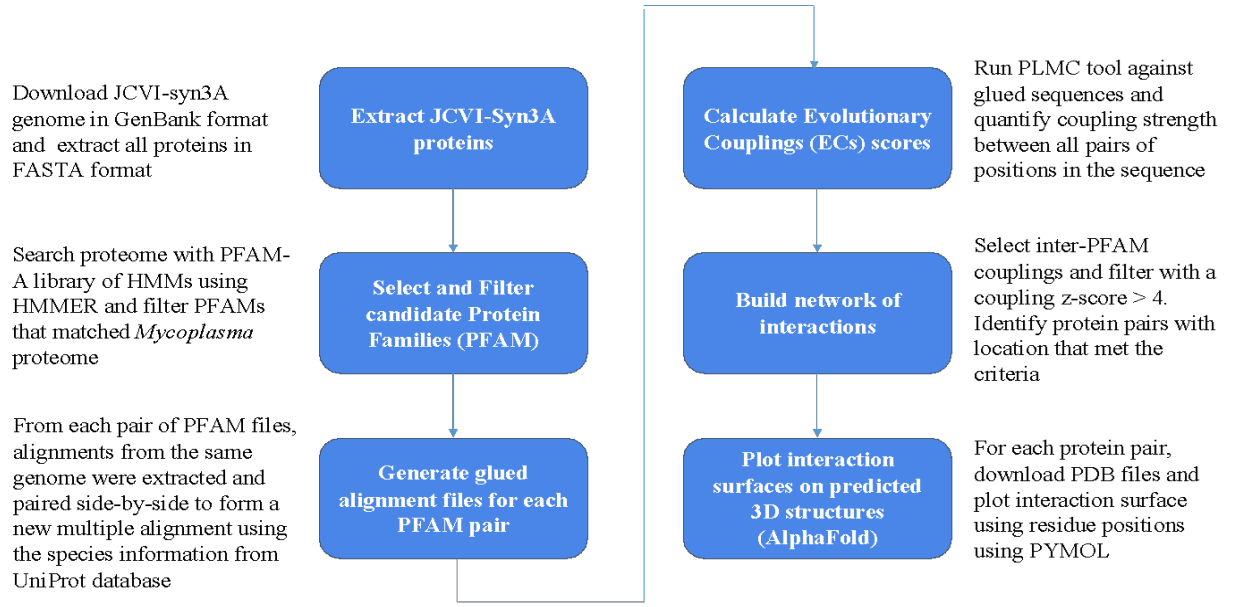

Figure 3. Methodology

By leveraging computational tools such as EC, we can conduct large-scale experiments that provide a wealth of data, which can then be used to inform and refine experimental approaches. This enables us to gain a deeper understanding of the underlying biology and develop novel solutions that can have a significant impact in biotechnology and medicine.

Step 1: To assemble the broadest possible data sets to make predictions, we first extracted all the protein sequences of the JCVI-syn3A genome in FASTA format.

Step 2: We searched the PFAM-A library of HMMs using HMMER (Hidden Markov Modeler), a bioinformatics tool widely used to find similarities between different biological sequences (*A New Generation of Homology Search Tools Based on Probabilistic Inference*, 2009). The output of HMMER is an expected value (e-value) score that represents the statistical significance of the match between the query sequence and the sequence models in the database. Using HMMER, we extracted a list of protein families (PFAMs) along with their e-value scores. For the exact commands executed, please see Steps.md in our GitHub project (https://github.com/arnav-meduri/JCVI-Syn3A-analysis). To prune the PFAM list and come up with a subset of protein targets for co-evolution analysis, we utilized the following strategies:

- Limit the potential combinatorial explosion of PFAMs. To compute co-evolution across proteins, individual protein sequences in each PFAM were paired with protein sequences with every other PFAM entry in the subset.
- In the interest of time and computing power available, we limited the amount of computation that was performed at the expense of possibly detecting fewer interactions.
- We selected PFAM entries that have only a single Mycoplasma hit so hits are less ambiguous. (Suppose PFAM family A has Mycoplasma hit X and PFAM family B has Y and Z, and if we get a positive, it is not easy to interpret whether the correct pairing is X-Y or X-Z or both, for example).
- By choosing an equal number of Mycoplasma hits (count = 1), we hypothesized that the statistical correlation signal would also be stronger.

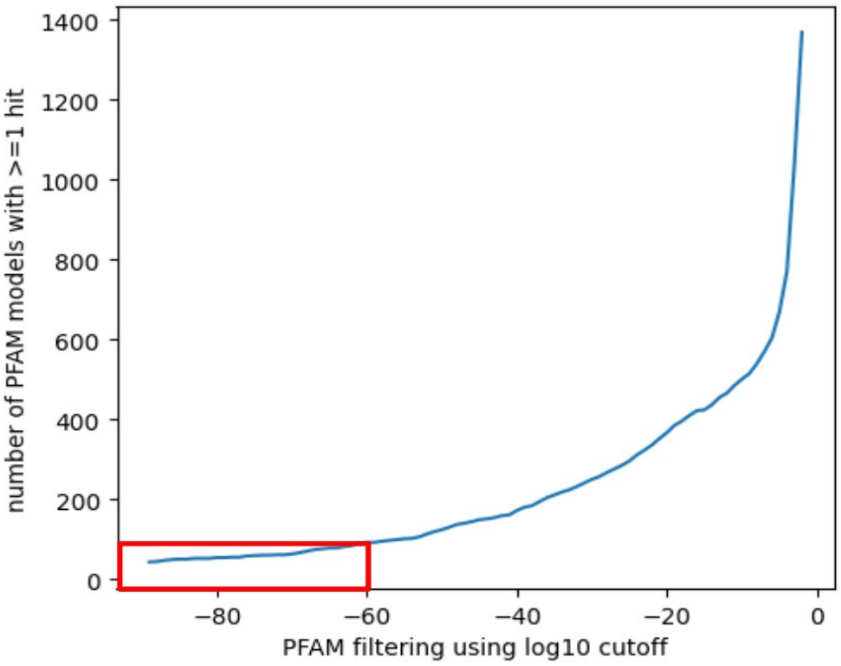

Figure 4. PFAM E-Value Cutoff

- We chose an E-value cutoff of 1E-60 while selecting a subset PFAMs and this resulted in 110 PFAM entries for pair analysis. A lower cutoff value implies a high confidence in the match, and that the match is not a coincidence. A lower cutoff has the potential to miss crucial matches. On the other hand, a higher cutoff will generate more noise in our analysis as it will include intra-protein interactions which are not the focus of our study. While the selection of a cutoff was somewhat arbitrary, it was carefully considered after evaluating the number of PFAMs that can be processed and analyzed with our available resources.

Step 3: Next, for each pair of PFAMs within the subset of 110 PFAMs, we created a new multiple sequence alignment file by concatenating paired sequences. We ignored glue alignments that did not have any proteins belonging to *Mycoplasma*.

Step 4: We performed statistical coevolution analysis using the PLMC tool in the EVcouplings framework (Ekeberg et al., 2014). The PLMC tool applies a pseudolikelihood maximization (PLM) approximation to determine the interaction parameters in the underlying maximum entropy probability model, generating both intra- and inter-EC scores of all pairs of residues within and across protein pairs. We used the protein sequence common across the Mycoplasma genus as the focus sequence with the PLMC tool and limited the number of iterations to 100. This optimization drastically reduced the amount of time it took for pairwise analysis. We used our personal laptops with M1 processors and 16 GB memory to complete each pair analysis. The computation required turned out to be both memory and CPU intensive. All the computations were completed over a 4-week period.

Step 5: The PLMC (Pseudo-Likelihood Maximization Couplings) coupling score is a measure to evaluate the coevolution between pairs of amino acid residues of pair proteins in a glued sequence (Ekeberg et al., 2014).. It represents the strength of the inferred interaction between two amino acid residues, with a higher score indicating a stronger interaction. In a multiple alignment, the rows are individual sequences, and the columns are amino acid positions. In the process of aligning protein sequences, each position represents a putative hypothesis that assumes all the amino

acids at that location occupy the same spatial arrangement in the protein's three-dimensional structure. We filtered out PFAM coupling scores using the following strategy:

- We filtered out scores that belonged to intra-protein matches, as we were interested in protein-protein interactions only.
- We selected scores with a z-score of > 4. A z-score is a statistical measure that indicates how many standard deviations a particular score is from the mean score. Higher counts in figure 6 left of the red circle generally imply intra-protein couplings. A z-score of greater than 4 implies that the score is 4 standard deviations higher than mean score and is a rare coupling. Such rare couplings of higher statistical significance are worth further investigation in our study. The filtering resulted in a total of 33 possible interactions of pair proteins.

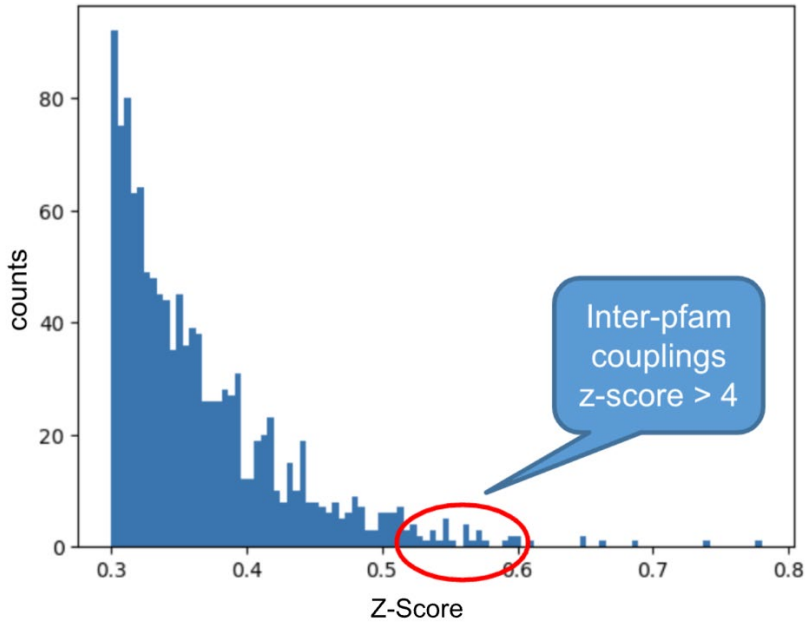

Figure 5. Histogram of PLMC Scores for PFAM Pairs

- For all of the identified interactions, we mapped the residue (amino acid) positions in the original sequence from the positions identified in the glued sequence by the EV couplings analysis.

Step 6: We downloaded the 3D images of the proteins (Alpha fold PDB files), loaded them in PYMOL to examine and identify the interaction surfaces of these protein pairs based on the residue pair and position information. Furthermore, for each of the interactions that provided a clear surface, we classified the interactions as either hydrophilic or hydrophobic.

Results

Figure 7 illustrates interaction between proteins P47345 and P47346 (Glutamyl amidotransferase subunits A & B), one of the 33 interaction surfaces identified in our analysis. For background, in order to function in translation a tRNA first requires a specific enzyme known as a tRNA synthetase, which is responsible for attaching the appropriate amino acid to the tRNA. *Mycoplasma*, however, doesn't have glutamine (Gln) tRNA synthetase. Instead, it charges the Gln-tRNA with glutamate (Perona, 2013). The enzyme glutamyl amidotransferase (with subunits A & B shown below) adds an amino acid group to glutamate to form glutamine. Analysis using BLASTP confirmed that these proteins are

homologous to each other and therefore will have the same fold. Since this is bacterial specific, this interaction surface could be a good target for further study of antibiotic discovery and development.

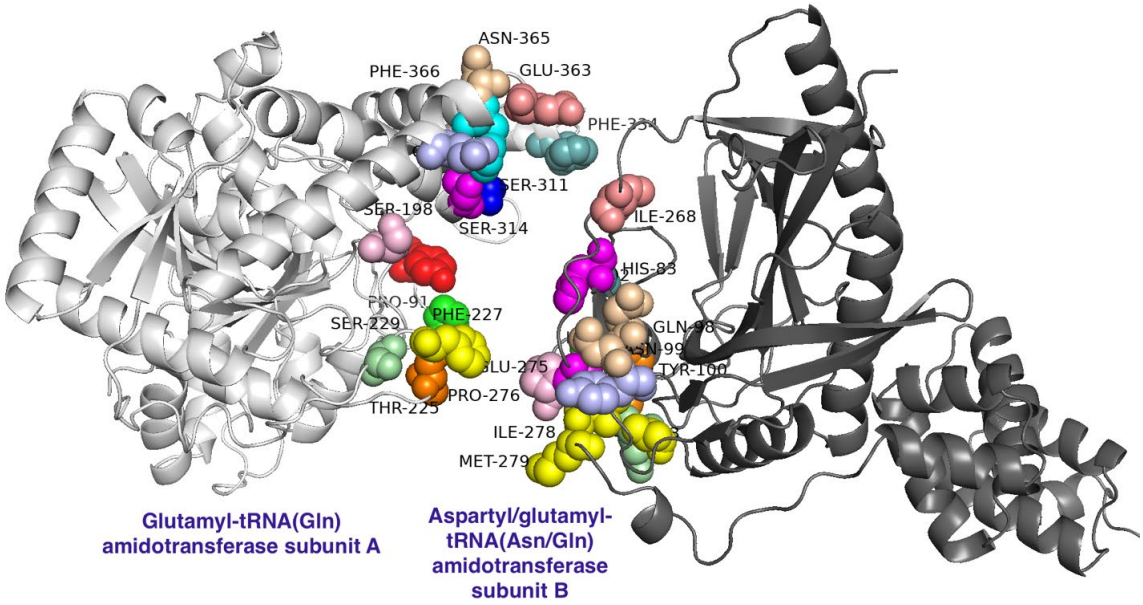

Figure 6. Interaction Surface between Proteins P47345 and P47346

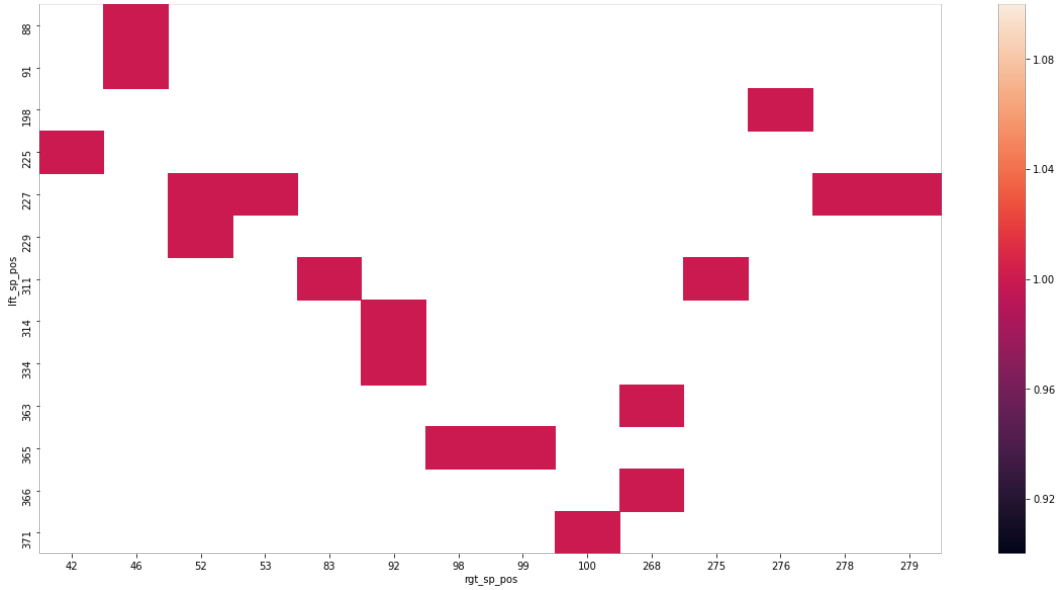

Figure 7. Heat map of amino acid interactions between Proteins P47345 and P47346

Figure 7, shown above, illustrates a large interaction surface between the protein pairs, where multiple residues on one protein are interacting with a single residue on the other protein in the pair.

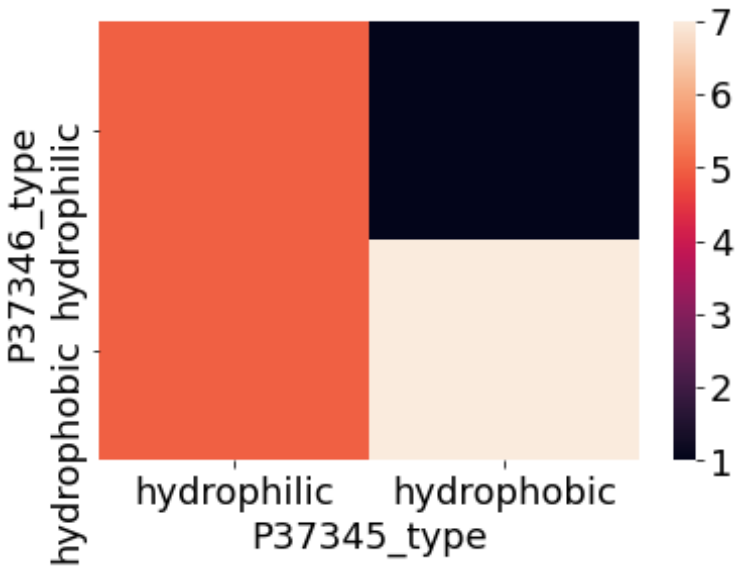

Figure 8. Hydrophilic vs Hydrophobic interactions between P47345 and P47346

Below are three additional interaction surfaces identified by our study. It is extremely likely that these interactions are real and can be confirmed further through experimentation. You can view all the interactions identified by the study in our GitHub repository.

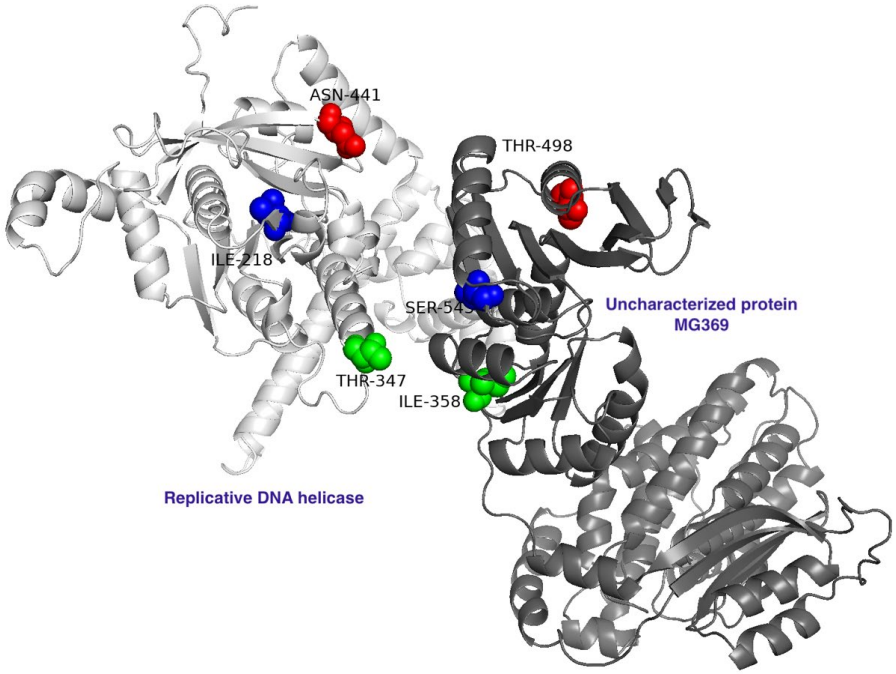

Figure 9. Interaction surface between replicative DNA helicase and uncharacterized protein MG369

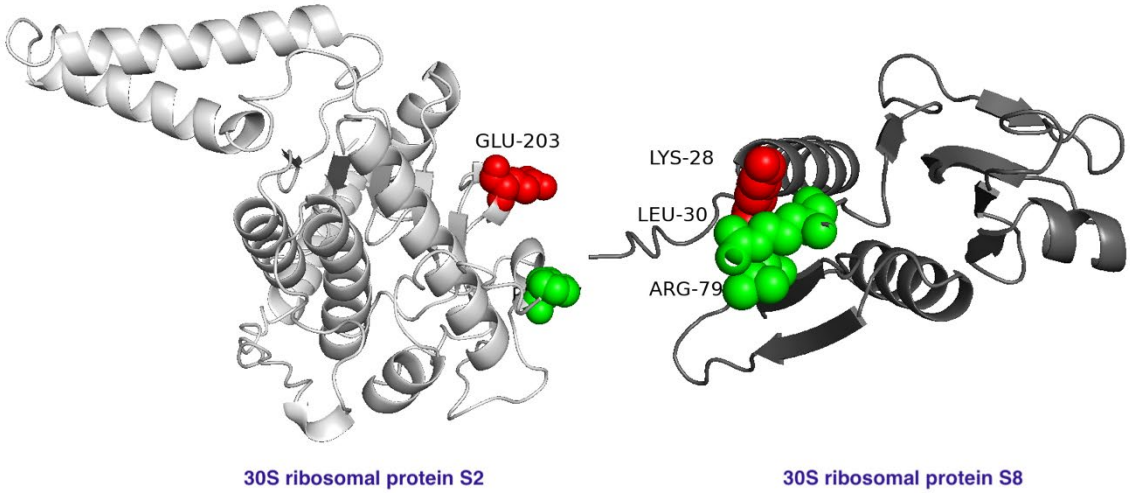

Figure 10. Interaction between 30S ribosomal protein S2 and 30S ribosomal protein S8

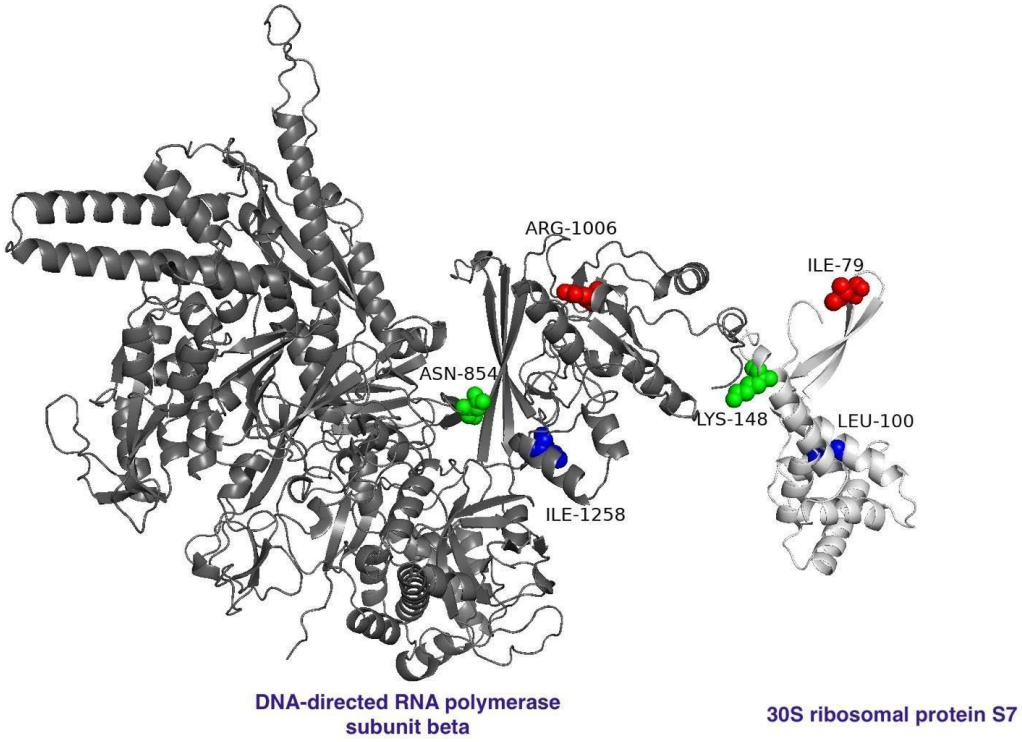

Figure 11. Interaction surface between DNA-directed RNA polymerase and 30S ribosomal protein S7

Table 11. PPIs Identified in Our Analysis

Protein Name	Protein ID	Protein ID	Protein Name
Enolase (EC 4.2.1.11) (2-phosphoglycerate dehydratase)	**P47647**	**P47609**	Hypothetical protein MG369
Phosphoglycerate kinase (EC 2.7.2.3)	**P47542**	**P47609**	Hypothetical protein MG369
30S ribosomal protein S7	**P47334**	**P47583**	DNA-directed RNA polymerase beta chain (EC 2.7.7.6) (Transcriptase beta
30S ribosomal protein S7	**P47334**	**P47619**	Glucose inhibited division protein A
50S ribosomal protein L16	**P47404**	**P47583**	DNA-directed RNA polymerase beta chain (EC 2.7.7.6) (Transcriptase beta
Ribonucleoside-diphosphate reductase beta chain (EC 1.17.4.1)	**P47471**	**P47473**	Ribonucleoside-diphosphate reductase alpha chain (EC 1.17.4.1)
30S ribosomal protein S2	**P47316**	**P47411**	30S ribosomal protein S8
30S ribosomal protein S2	**P47316**	**P47583**	DNA-directed RNA polymerase beta chain (EC 2.7.7.6) (Transcriptase beta
30S ribosomal protein S2	**P47316**	**P47609**	Hypothetical protein MG369
30S ribosomal protein S2	**P47316**	**P47619**	Glucose inhibited division protein A
Preprotein translocase secY subunit	**P47416**	**P47583**	DNA-directed RNA polymerase beta chain (EC 2.7.7.6) (Transcriptase beta)
Preprotein translocase secY subunit	**P47416**	**P47331**	Probable HPr(Ser) kinase/phosphatase (EC 2.7.1.-) (EC 3.1.3.-)
DNA-directed RNA polymerase beta chain (EC 2.7.7.6) (Transcriptase beta	**P47583**	**P47582**	DNA-directed RNA polymerase beta' chain (EC 2.7.7.6) (Transcriptase beta)
DNA-directed RNA polymerase beta chain (EC 2.7.7.6) (Transcriptase beta	**P47583**	**P47609**	Hypothetical protein MG369
DNA-directed RNA polymerase beta chain (EC 2.7.7.6) (Transcriptase beta	**P47583**	**P47619**	Glucose inhibited division protein A
Ribonuclease R (EC 3.1.-.-) (RNase R) (VacB protein homolog)	**P47350**	**P47609**	Hypothetical protein MG369
Probable cytosol aminopeptidase (EC 3.4.11.1) (Leucine aminopeptidase)	**P47631**	**P47609**	Hypothetical protein MG369
DNA topoisomerase I (EC 5.99.1.2) (Omega-protein) (Relaxing enzyme)	**P47368**	**P47609**	Hypothetical protein MG369
Cysteinyl-tRNA synthetase (EC 6.1.1.16) (Cysteine--tRNA ligase) (CysRS)	**P47495**	**P47609**	Hypothetical protein MG369
Phenylalanyl-tRNA synthetase alpha chain (EC 6.1.1.20) (Phenylalanine--tRNA	**P47436**	**P47609**	Hypothetical protein MG369

Alanyl-tRNA synthetase (EC 6.1.1.7) (Alanine--tRNA ligase) (AlaRS)	**P47534**	**P47609**	Hypothetical protein MG369
Glutamyl-tRNA(Gln) amidotransferase subunit A (EC 6.3.5.-) (Glu-ADT subunit	**P47345**	**P47346**	Aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B (EC 6.3.5.-)
Glutamyl-tRNA(Gln) amidotransferase subunit A (EC 6.3.5.-) (Glu-ADT subunit	**P47345**	**P47609**	Hypothetical protein MG369
Cytidylate kinase (EC 2.7.4.14) (CK) (Cytidine monophosphate kinase) (CMP	**P47572**	**P47609**	Hypothetical protein MG369
Probable thiamine biosynthesis protein thiI	**P47612**	**P47609**	Hypothetical protein MG369
Hypothetical protein MG110	**P47356**	**P47609**	Hypothetical protein MG369
Replicative DNA helicase (EC 3.6.1.-)	**P47340**	**P47609**	Hypothetical protein MG369
DNA-directed RNA polymerase beta' chain (EC 2.7.7.6) (Transcriptase beta'	**P47582**	**P47619**	Glucose inhibited division protein A
Probable HPr(Ser) kinase/phosphatase (EC 2.7.1.-) (EC 3.1.3.-)	**P47331**	**P47609**	Hypothetical protein MG369
Probable HPr(Ser) kinase/phosphatase (EC 2.7.1.-) (EC 3.1.3.-)	**P47331**	**P47619**	Glucose inhibited division protein A
Preprotein translocase secA subunit	**P47318**	**P47609**	Hypothetical protein MG369
Preprotein translocase secA subunit	**P47318**	**P47619**	Glucose inhibited division protein A
Hypothetical protein MG369	**P47609**	**P47619**	Glucose inhibited division protein A

Discussion

Despite significant advances in predicting the three-dimensional structures of proteins, such as the AlphaFold deep learning software, predicting protein-protein interactions remains an active research frontier. To date, a comprehensive molecular machine learning algorithm for predicting inter-protein interactions has yet to be developed.

Our research has demonstrated the feasibility of predicting protein interactions using co-evolution-based computational methods, as applied to the JCVI-Syn3A proteome. The method we employed also predicted the interaction surfaces and a subset of the proteins. Our analysis of multiple sequence data across 110 protein families (PFAMs) revealed a total of thirty-three potential inter-protein interactions in JCVI-syn3A. Using computational methods to predict interactions in three dimensions from sequence data alone is a fascinating and compelling approach. Our study highlights the tremendous potential of computational methods in predicting protein-protein interactions, and their subsequent validation through experimental testing. By integrating computational methods with experimental validation, we have demonstrated a powerful and cost-effective approach for investigating protein interactions, particularly in situations where experimental methods may be limited. Our research underscores the value of combining sequence data with the EVCouplings framework to gain insights into the underlying molecular mechanisms of

biological processes. This approach has broad applications across various fields, including drug discovery and development.

Overall, our study highlights the significance of leveraging computational methods to advance our understanding of the complex interplay between protein sequences, structures, and functions. We hope that our findings will inspire further research and development of innovative computational tools and methodologies to deepen our understanding of biological systems.

Acknowledgements

We wish to express our heartfelt thanks to Dr. Keith Robison, whose encouragement and invaluable mentorship have been integral to our research journey. With his expert guidance, we were able to identify the most relevant problem to investigate, maintain our focus in overcoming challenges at every stage, and ultimately evaluate the significance of our work.

In addition, we are deeply grateful to the Debora Marks lab at Harvard University for their generous open-source contribution of the EVCouplings framework, which has empowered us and numerous others to advance our research pursuits. We especially appreciate their exceptional educational web sessions, which have played a vital role in enhancing our proficiency in utilizing this powerful tool.

References

Brown, G., Yakunin, A. F., Kurilyak, I., Folz, J., Fiehn, O., Glass, J. I., Hanson, A. D., … de Crécy-Lagard, V. (2022). Metabolite Damage and Damage Control in a Minimal Genome. *mBio*, *13*(4), e0163022. https://doi.org/10.1128/mbio.01630-22

Bianchi, D. M., Pelletier, J. F., Hutchison, C. A., 3rd, Glass, J. I., & Luthey-Schulten, Z. (2022). Toward the Complete Functional Characterization of a Minimal Bacterial Proteome. *The journal of physical chemistry. B*, *126*(36), 6820–6834. https://doi.org/10.1021/acs.jpcb.2c04188

Breuer, M., Earnest, T. M., Merryman, C., Wise, K. S., Sun, L., Lynott, M. R., Hutchison, C. A., Smith, H. O., Lapek, J. D., Gonzalez, D. J., de Crécy-Lagard, V., Haas, D., Hanson, A. D., Labhsetwar, P., Glass, J. I., & Luthey-Schulten, Z. (2019). Essential metabolism for a minimal cell. *eLife*, *8*, e36842. https://doi.org/10.7554/eLife.36842

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press. https://doi:10.1017/CBO9780511790492

Ekeberg, M., Hartonen, T., & Aurell, E. (2014). Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics*, *276*, 341–356. https://doi.org/10.1016/j.jcp.2014.07.024

Green, A. G., Elhabashy, H., Brock, K. P., Maddamsetti, R., Kohlbacher, O., & Marks, D. S. (2021). Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nature communications*, *12*(1), 1396. https://doi.org/10.1038/s41467-021-21636-z

Pelletier, J. F., Sun, L., Wise, K. S., Assad-Garcia, N., Karas, B. J., Deerinck, T. J., Ellisman, M. H., Mershin, A., Gershenfeld, N., Chuang, R. Y., Glass, J. I., & Strychalski, E. A. (2021). Genetic requirements for cell division in a genomically minimal cell. *Cell*, *184*(9), 2430–2440.e16. https://doi.org/10.1016/j.cell.2021.03.008

Hutchison, C. A., 3rd, Chuang, R. Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J., Ma, L., Pelletier, J. F., Qi, Z. Q., Richter, R. A., Strychalski, E. A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K. S., Smith, H. O., Glass, J. I., … Venter, J. C. (2016). Design and synthesis of a minimal bacterial genome. *Science (New York, N.Y.)*, *351*(6280), aad6253. https://doi.org/10.1126/science.aad6253

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(49), E1293–E1301. https://doi.org/10.1073/pnas.1111471108

Pfam: The protein families database in 2021 https://pubmed.ncbi.nlm.nih.gov/33125078/

A new generation of homology search tools based on probabilistic inference (HMMER) https://pubmed.ncbi.nlm.nih.gov/20180275/

Jumper, J. M., Evans, R., Pritzel, A., Green, T. J., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. a. A., Ballard, A. J., Cowie, A. M., Romera-Paredes, B., Nikolov, S., Jain, R. D., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Perona, J. J. (2013). *Glutaminyl-tRNA Synthetases*. Madame Curie Bioscience Database - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK6506/