MAGPIE: an interactive tool for visualizing and analyzing many simultaneous protein interactions with a binding partner

Daniel C. Pineda Rodriguez[1], Kyle C. Weber[1], and Anum Glasgow[1]*
[1] Department of Biochemistry and Molecular Biophysics, Columbia University Irving Medical Center, New York, NY, United States
* Corresponding author. Department of Biochemistry and Molecular Biophysics, Columbia University Medical Center, New York, NY 10032, United States. Email: ag4522@cumc.columbia.edu (A.G.)

Abstract
Summary: We present MAGPIE, a tool for analyzing thousands of protein-ligand interactions at once. MAGPIE generates an interactive 3D visualization from a set of protein complex structures that share a target ligand, as well as sequence logo-style frequency graphs that show all the amino acids from the set of protein binders that interact with user-selected target ligand positions. MAGPIE is compatible with small molecule or protein ligands.

Availability and implementation: MAGPIE is implemented in Python 3 and is freely available with two dataset/usage examples and a helper script to prepare MAGPIE inputs at https://github.com/glasgowlab/MAGPIE. Additional documentation is in the Supplementary data online.

1 Introduction

In the study of protein-protein or protein-small molecule interactions, it is challenging to compare many interactions at once. For example, analyzing how a library of computationally designed proteins interact with a binding partner usually requires ranking the design models according to several quality metrics, and then inspecting top-ranked models in detail individually. Similarly, although assembling a multiple sequence alignment (MSA) of evolutionarily related proteins is expedited by bioinformatic tools, it remains difficult to visualize how sequence variability at one or more positions in a ligand binding pocket manifests in three dimensions (3D), especially when they are non-contiguous in sequence space. Thus, there is a need for a versatile protein complex visualization tool that also depicts the sequence diversity inherent in molecular interactions on the fly. Such a tool would offer valuable insights into sequence conservation and variability among proteins binding to common ligands, and would be particularly beneficial for identifying the biochemical requirements for key interactions or protein functions. However, existing methods only partially address this need. For example, using sequence logos to study protein interactions requires MSAs with similar sequence lengths, and relies on the assumption that similar interactions occur at the same sequence positions (Crooks *et al.*, 2004). In fact, MSA generation algorithms present significant limitations for investigating protein interactions in 3D as they are designed for aligning sequences with similar lengths and high sequence homology, neglecting the 3D protein structure (Larkin *et al.*, 2007; Sievers *et al.*, 2011; Katoh and Standley, 2013; Wallace *et al.*, 2006; Brown *et al.*, 1998; Hagopian *et al.*, 2010). An MSA-based approach is inadequate for finding conserved interactions across structurally diverse proteins with one binding partner, such as a small molecule ligand that binds multiple proteins via different binding pocket geometries, or a single antigen that binds a library of engineered antibodies at non-overlapping epitopes. It remains difficult in such cases to translate trends at the 2D sequence level to the 3D space in order to identify molecular patterns that drive or tune biochemical interactions.

To address this problem, we introduce MAGPIE: Mapping Areas of Genetic Parsimony In Epitopes. MAGPIE a protein complex visualization and analysis tool to facilitate the simultaneous comparison of many protein-partner interactions in 3D by structure and sequence. MAGPIE identifies amino acids (AAs) in a set of proteins that make molecular interactions with a user-defined region on a single target ligand, which can be a protein or a small molecule. MAGPIE also generates AA frequency graphs on the fly that show how often different AAs in the set of proteins interact with user-defined target ligand residues or heavy atoms. Furthermore, MAGPIE generates an interactive 3D representation of the target ligand and visually illustrates the positions of all alpha carbons (Cα) that are near it, coloring these by AA biochemical characteristics and highlighting patches of conserved or similar interactions. We anticipate that MAGPIE will be useful in applications ranging from protein design, where it can identify biases in computational methods for engineering protein-ligand interactions, to biology, where it may illuminate functional adaptation in proteins by revealing how allosteric binding pockets have emerged over evolutionary history.

2 Features and implementation

MAGPIE operates in several steps (Fig. 1A). First, the user inputs a set of protein complex structures in Protein Data Bank (PDB) format that are structurally aligned on the target ligand. The target ligand must have the same chain index in all input structures, and the protein binding partners must also share a chain index, which must be different from the target ligands. MAGPIE stores the AA position information of the user-provided chain indices for the target ligand and the protein binding partners. Specifically, it records the positions of all Cα within the target chain, or all the heavy atoms (HAs) if the target is a small molecule. Subsequently, it also stores the positions of all Cα from the binder chain in the user-provided PDB dataset. Second, MAGPIE employs a kd-tree to efficiently query which AAs from the binding partners fall within a user-defined distance of any AAs/HAs from the target ligand (Bentley, 1975). This step is crucial for identifying potential interacting residues within a specified proximity of regions of the target ligand. Third, MAGPIE generates plots to display the frequency of AAs among the binding partner residues relative to user-provided target positions. By default, it includes all AAs, but users have the option to customize the target ligand AAs (for protein ligands) or HAs (for small molecule ligands) for their specific analysis needs. Next, MAGPIE employs the Plotly library to create color-coded interactive 3D graphs, displaying the 3D locations of all binding partner residue Cα and the target ligand structure main chain (Sievert, 2020). This allows users to explore the spatial arrangement of the protein-ligand interactions. The target structure is represented in black (for protein targets) or CPK coloring (for small molecules), while the binding residues can be colored using the RasMol-based "Shapely" or "Amino" colors (Sayle, 1995). Finally, users can select which AAs/HAs from the target ligand to include in the AA frequency graphs. Once the positions are selected, MAGPIE identifies binding partner residue Cα atoms using the kd-tree from Step 2, calculates the frequency of each AA, and generates a frequency map using the LogoMaker (Crooks *et al.*, 2004) Python package with a customizable color scheme (Najafabadi *et al.*, 2017). This process helps in visualizing and understanding the distribution and conservation of AAs at specific positions involved in the binding interactions. In the resulting MAGPIE-generated figure, the *x*-axis denotes the index position in the target ligand, while the *y*-axis indicates the frequency of AAs among binding partner residues from the user-input PDB dataset found within the specified distance, implying molecular interactions with the target ligand index position. Each one-letter AA code is represented by a height that corresponds to its frequency in the set of binding partners. MAGPIE also generates a combined graph that contains all the user-specified target ligand positions**.** This graphical representation illustrates the prevalence of specific AA interactions for distinct positions in the target.

Performance and scalability: MAGPIE is implemented in Python and uses multiple CPU cores to optimize speed. If a GPU is available, MAGPIE can use CUDA acceleration to further improve performance. We recommend using MAGPIE on the cloud-based Google CoLab server if the input PDB dataset includes fewer than 1000 structures. For private and/or larger datasets where limited computing resource allocations are a concern, we recommend instead downloading the MAGPIE source code and running it locally, or using the multithreaded implementation.

Recommendations for preparing datasets for MAGPIE, a list of additional features, troubleshooting guide, documentation for our helper script for preparing input structures, information about multithreaded MAGPIE, and runtime benchmarks are available in Sections 1-6 of the Supplementary information.

3 Results

We present two case studies to demonstrate the utility of MAGPIE. The first case study (Fig. 1B-E) explores the interactions that a set of 63 nonredundant antibodies make with the SARS-CoV-2 spike protein receptor binding domain (RBD) (Fig. 1B) (Gowthaman *et al.*, 2021). MAGPIE aids in the visualization of the antibody binding epitopes (Fig. 1C). We generated AA frequency graphs for one region of the RBD (Fig. 1D, E). The graph showcases the diversity of interactions with antibodies from the dataset for four residues in the RBD belonging to that region. The second case study (Fig. 1F-H) highlights similarities and differences among interactions with coenzyme A (COA) in 23 binding pockets/interfaces from bacterial enzymes, including ligases, acetyltransferases, synthases, and epimerases (Fig. 1F). MAGPIE conglomerated the local chemical environments for each of the heavy atoms in the adenosine moiety of COA (Fig. 1G, H) and calculated the frequency of different types of interactions that occur in the enzymes with each COA atom. For example, in Fig. 1I, we observed that whereas the 3' phosphorus atom in the COA adenosine group is located near polar and charged AAs in most of the enzymes, the 5' P1 atom environment instead includes many hydrophobic AAs, likely due to its proximity to the phosphopantetheine group of COA. We also compared the chemical environments of the nitrogen atoms to quantify their hydrogen bonding propensities. Additional details about these case studies are in Section 7 of the Supplementary information.

4 Availability and requirements

MAGPIE is implemented in Python 3 and freely available to download with documentation and two dataset/usage examples. Local installation requirements are listed in Section 8 of the Supplementary information. A Google CoLab server running MAGPIE, where users can run the case studies or their own uploaded datasets, is also available. See the Data availability section.

We also provide a helper script (MAGPIE_input_prep.py) to modify input PDB files for seamless usage in MAGPIE, with documentation in Section 4 of the Supplementary information. The helper script allows the user to input files or directories with options to define target ligands and protein binders by chain, protein sequence, and/or ligand name (for small molecule targets). Usage examples are also available in the Supplementary data section.

5 Conclusion

As highly accurate protein structure prediction methods spur the development of computational approaches to model and design protein-ligand binding interactions, the need for versatile tools to analyze the biochemical details of structural complexes *en masse* grows. Platforms to easily compare protein interactions will aid in biological discovery and therapeutics development.

MAGPIE provides a practical and beautiful way to explore the diverse interactions that natural and designed proteins alike can form with protein and small molecule ligands.

Supplementary data

Supplementary data are available online.

Conflict of interest

None declared.

Funding

This work has been supported by the NIH NIGMS grant R00 GM135529.

Data availability

MAGPIE is available open source from the Glasgow Lab GitHub: https://github.com/glasgowlab/MAGPIE/tree/local-version. Access the Google CoLab server at: https://colab.research.google.com/github/glasgowlab/MAGPIE/blob/GoogleColab/MAGPIE_COLAB.ipynb

Acknowledgements

The authors thank the Glasgow Lab and Dr. James Lucas for beta testing, feature requests, and bug reports.

References

Bentley,J.L. (1975) Multidimensional binary search trees used for associative searching. *Commun. ACM*, **18**, 509–517.

Brown,N.P. *et al.* (1998) MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.

Crooks,G.E. *et al.* (2004) WebLogo: A Sequence Logo Generator. *Genome Res.*, **14**, 1188–1190.

Gowthaman,R. *et al.* (2021) CoV3D: a database of high resolution coronavirus protein structures. *Nucleic Acids Res.*, **49**, D282–D287.

Hagopian,R. *et al.* (2010) SATCHMO-JS: a webserver for simultaneous protein multiple sequence alignment and phylogenetic tree construction. *Nucleic Acids Res.*, **38**, W29–W34.

Katoh,K. and Standley,D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.*, **30**, 772–780.

Larkin,M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Najafabadi,H.S. *et al.* (2017) Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding. *Genome Biol.*, **18**, 167.

Sayle,R. (1995) Rasmol.

Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence

alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
Sievert,C. (2020) Interactive Web-Based Data Visualization with R, plotly, and shiny.
Wallace,I.M. *et al.* (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.*, **34**, 1692–1699.

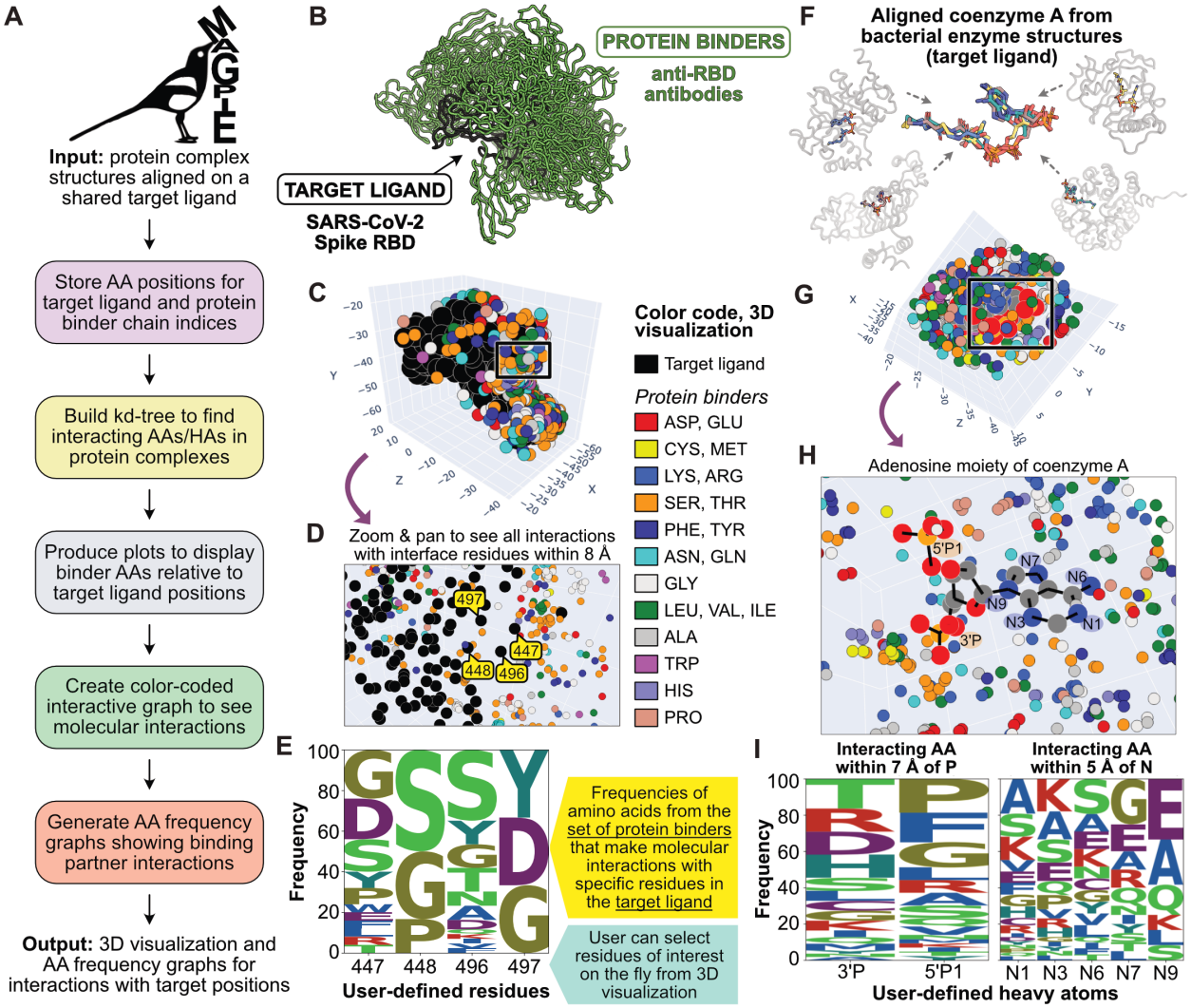

Figure 1. MAGPIE. (A) MAGPIE takes protein-protein or protein-small molecule complex structures as inputs, aligned on the target ligand. It produces an interactive graph of protein binder interactions with amino acids (AAs) and heavy atoms (HAs) on the target ligand, with AA frequency graphs showing binder interactions with user-defined target AAs or HAs. **(B)** Structural representation of the SARS-CoV-2 spike receptor binding domain (RBD, black) bound to 15 anti-RBD antibodies (green), from a set of 63 antibodies. **(C, D)** MAGPIE-generated 3D representation of the RBD-antibody complexes. The spike antigen Cα are shown as black circles and the anti-RBD antibody Cα are shown as colored circles. **(D)** is the zoomed-in region boxed in **(C)**, with individual residues shown. **(E)** AA frequency graph showing enrichment of polar interactions for several spike residues indicated in **(D)**. **(F)** Overlaid coenzyme A (COA) structures for four of a set of 23 enzymes. Originating structures in gray, with COA in color. **(G)** COA-protein interactions. **(H)** Zoomed-in view of COA adenosine moiety. **(I)** AA frequency graphs for the two COA P atoms and five N atoms shown in (H), highlighting the differences and similarities in their chemical environments from the set of 23 enzymes.