BIAPSS - BioInformatic Analysis of liquid-liquid Phase-Separating protein Sequences

Aleksandra E. Badaczewska-Dawid[1,✉] **and Davit A. Potoyan**[1,2,3✉]

[1]Department of Chemistry, Iowa State University, Ames IA 50011 USA
[2]Department of Biochemistry Biophysics and Molecular Biology, Iowa State University, Ames IA 50011 USA
[3]Bioninformatics and Computational Biology program, Iowa State University, Ames IA 50011 USA

Liquid-liquid phase separation (LLPS) has recently emerged as a foundational mechanism for order and regulation in biology. However, a quantitative molecular grammar of protein sequences underlying LLPS remains unclear. The comprehensive databases and associated computational infrastructure for biophysical and statistical analysis can enable rapid progress in the field. Therefore, we have created a novel open-source web platform named BIAPSS (BioInformatic Analysis of liquid-liquid Phase-Separating protein Sequences) which offers the users interactive data analytic tools for facilitating the discovery of statistically significant sequence signals for proteins with LLPS behavior. Availability: BIAPSS is freely available online at https://biapss.chem.iastate.edu/. Website is implemented within the Python framework using HTML, CSS, and Plotly-Dash graphing libraries, with all the major browsers supported including the mobile device accessibility.

LLPS | BIAPSS | Plotly-Dash

Correspondence: ***abadacz@iastate.edu, potoyan@iastate.edu***

Introduction

In the past few years, LLPS of biomolecules has become a universal language for interpreting intracellular signaling, compartmentalization, and regulation (1–5). The ability to phase separate appears to be encoded primarily in the protein sequences, frequently containing disordered and low complexity domains, which are enriched in charged and multivalent interaction centers (6–8). Nevertheless, the quantitative aspects of how amino acids encode and decode the phase separation remain largely unknown (9–11). This is because many different combinations of relevant interactions seem to be contributing to phase separation without anyone being universally necessary (12). So far, however, with a few exceptions (13–16) mostly case by case studies of different sequences are performed, with the broader context of many findings, including their statistical significance remaining unknown. To this end, we have developed a web framework BIAPSS: BioInformatic Analysis of liquid-liquid Phase-Separating protein Sequences. The objective of BIAPSS is to enable a rapid and on-the-fly deep statistical analysis of LLPS-driver proteins using the pool of sequences with empirically confirmed phase behavior.

Implementation

The back-end processing pipeline of BIAPSS is implemented in a Python framework, where in-house developed algorithms parse pre-computed data and perform on-the-fly analysis. The basic front-end user interface of the BIAPSS web platform is implemented with HTML5, CSS, JavaScript, and Bootstrap components which support the responsiveness and mobile-accessibility of the website. Specifically, our cross-platform framework is adjusted to be run on multiple operating systems and popular browsers. Modern display-layer solutions improve user experience by enabling smooth loading of contents, page transitions, and accompanying an in-depth presentation of the results. For instance, we included a lightbox slideshow with a brief overview of the features, collapsed menu, and modal images of quick guide within individual applications, side navigation, and more. Interactive graph plotting and data visualization accessible through web applications in SingleSEQ and MultiSEQ tabs were developed with the Plotly-Dash (17) browser-based graphing libraries for Python which create a user-responsive environment and follow remote, customized instructions. Thanks to the interactive interface users can go directly from exploratory analytics to the creation of publication-ready high-quality images.

Results

BIAPSS is designed as a user-friendly web platform that is billing itself as a central resource for systematic and standardized statistical analysis of biophysical characteristics of known LLPS sequences. The web service provides users with (i) a database of the superset of experimentally evidenced LLPS-driver protein sequences, (ii) a repository of pre-computed bioinformatics and statistics data, and (iii) two sets of web applications supporting the interactive analysis and visualization of physicochemical and biomolecular characteristics of LLPS proteins. The initial LLPS sequence set leverages the data from manually curated primary LLPS databases, namely PhaSePro (15) and LLPSDB (16). Given that the number of experimentally confirmed LLPS driver proteins is constantly growing, the BIAPSS pre-computed repository is updated annually and released to the public, which significantly saves the users time eliminating the need for exhaustive in-house calculations. The apps integrate the

DRAFT

results from our extensive studies, described in more detail elsewhere (). One of the aims of BIAPSS is to get an insight into the overall characteristics of the sufficient non-redundant set of LLPS-driver protein sequences. The comparison to benchmarks of various protein groups enables statistical inference of specific phase-separating affinities. Furthermore, the residue-resolution biophysical regularities inferred from BIAPSS will help not only to accurately identify regions prone to phase separation but also to design sequence modifications targeting various biomedical applications. The extended Cross-References section is designed as a central navigation hub for researchers for keeping track of the corresponding entries in the primary LLPS databases along with the other external resources relevant to the phase separation field. Since many users usually have specific single sequences of interest (natural or designed) our future efforts will be directed towards the creation of an upload section for parse user-defined cases and compare them with the benchmark of known LLPS-driver proteins.

The layout and main functionalities of BIAPSS services are summarized in the Figure 1. The general outline of the platform is designed to provide clarity and intuitive navigation by avoiding the excess of permanently visible information. Due to the multitude of analyses, available to meet the needs of a diverse audience of scientists, the extensive content of BIAPSS has been divided into 5 main tabs. The *Home* tab is a place where the user gets a high-level overview of the features of BIAPSS services. Next comes the *SingleSEQ* tab which is dedicated to the exploration of individual LLPS sequence characteristics. Besides a case summary and cross-reference section, there are multiple web applications dedicated to the in-depth analysis of biomolecular features, such as sequence conservation with multiple sequence alignment (MSA) (18), various sequence-based predictions by the state-of-the-art methods for secondary structure (18–23), solvent accessibility (22–24), structural disorder (22, 25–29), contact maps (22, 27, 29), and uniquely proposed detection of numerous short linear motifs (SliMs) (30–34) recently highlighted as key regions for driving the LLPS (35). The *MultiSEQ* tab provides the user with a set of web applications for a broad array of statistics on a superset of LLPS sequences. One may there investigate the regularities and trends specific only for disordered regions, such as amino acid (AA) composition, including AA diversity or regions rich in a given AA, general physicochemical patterns of polarity, hydrophobicity, the distribution of aromatic or charged residues, including not only the overall net charge but also charge decoration parameters that emerged as a relevant factor for electrostatic interactions of intrinsically disordered proteins (IDPs) (36), and more. Also, a deeper focus on the general frequency of particular short linear motifs, including LARKS (31), GARs (32), ELMs (30), and steric zippers (34), as well as pioneering identification of specific n-mers, can bring new perspectives in the field. The *Download* tab facilitates accessing the BIAPSS repository. The available data includes raw predictions pre-calculated using the well-established tools as well as the findings of our deep statistical analysis. For the

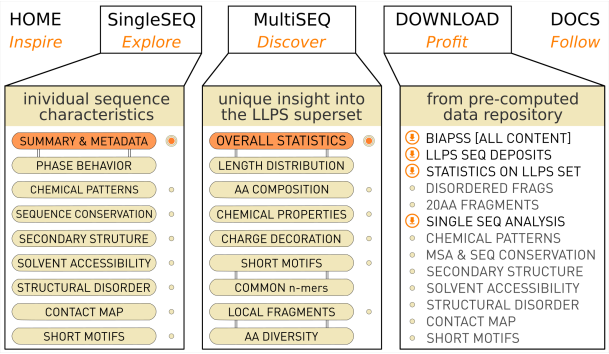

Fig. 1. The overall layout of BIAPSS web platform (https://biapss.chem.iastate.edu/) for comprehensive sequence-based analysis of LLPS proteins. The core of the implemented web applications and data repository is contained in the SingleSEQ, MultiSEQ, and Download tabs.

convenience of users, we have unified and integrated the pre-processed results into a standardized CSV format accompanied with intuitive descriptors to facilitate reuse and, specifically, allows the researcher to implement the pre-computed data directly or carry out further analysis. Finally, in the *Docs* tab, the user can follow the detailed data-analytic workflow and learn more about used tools with corresponding references to the original literature. The documentation also includes an easy-to-use tutorial dedicated to individual web applications, where all of the features are presented graphically with detailed descriptions (see also the user's manual attached in the Supplementary information).

Funding

A.E.B-D. acknowledges a generous financial support by Roy J. Carver Charitable Trust through Iowa State University Bioscience Innovation Postdoctoral Fellowship. This work was supported by the National Institute Of General Medical Sciences of the National Institutes of Health [R35GM138243 to D.A.P.]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
Conflict of Interest: none declared.

Author Contribution

Conceptualization, A.E.B-D.; Software development, A.E.B-D.; Writing an original draft, A.E.B-D. and D.A.P.

1. Clifford P Brangwynne, Christian R Eckmann, David S Courson, Agata Rybarska, Carsten Hoege, Jöbin Gharakhani, Frank Jülicher, and Anthony A Hyman. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science*, 324(5935): 1729–1732, June 2009.
2. Clifford P Brangwynne, Timothy J Mitchison, and Anthony A Hyman. Active liquid-like behavior of nucleoli determines their size and shape in xenopus laevis oocytes. *Proc. Natl. Acad. Sci. U. S. A.*, 108(11):4334–4339, March 2011.
3. Iain A Sawyer, Jiri Bartek, and Miroslav Dundr. Phase separated microenvironments inside the cell nucleus are linked to disease and regulate epigenetic state, transcription and RNA processing. *Semin. Cell Dev. Biol.*, July 2018.
4. Sudeep Banjade, Qiong Wu, Anuradha Mittal, William B Peeples, Rohit V Pappu, and Michael K Rosen. Conserved interdomain linker promotes phase separation of the multivalent adaptor protein nck. *Proc. Natl. Acad. Sci. U. S. A.*, 112(47):E6426–35, November 2015.
5. Sudeep Banjade and Michael K Rosen. Phase transitions of multivalent proteins can promote clustering of membrane receptors. *Elife*, 3, October 2014.
6. Jeong-Mo Choi, Alex S Holehouse, and Rohit V Pappu. Physical principles underlying the complex biology of intracellular phase transitions. *Annu. Rev. Biophys.*, January 2020.

DRAFT

7. Jie Wang, Jeong-Mo Choi, Alex S Holehouse, Hyun O Lee, Xiaojie Zhang, Marcus Jahnel, Shovamayee Maharana, Régis Lemaitre, Andrei Pozniakovsky, David Drechsel, Ina Poser, Rohit V Pappu, Simon Alberti, and Anthony A Hyman. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell*, 174(3): 688–699.e16, July 2018.
8. Gregory L Dignon, Robert B Best, and Jeetain Mittal. Biomolecular phase separation: From molecular driving forces to macroscopic properties. *Annu. Rev. Phys. Chem.*, 71:53–75, April 2020.
9. Castrense Savojardo, Pier Luigi Martelli, and Rita Casadio. Protein–Protein interaction methods and protein phase separation. *Annu. Rev. Biomed. Data Sci.*, 3(1):89–112, July 2020.
10. Wade Borcherds, Anne Bremer, Madeleine B Borgia, and Tanja Mittag. How do intrinsically disordered protein regions encode a driving force for liquid-liquid phase separation? *Curr. Opin. Struct. Biol.*, 67:41–50, October 2020.
11. Boris Y Zaslavsky, Luisa A Ferreira, and Vladimir N Uversky. Driving forces of Liquid-Liquid phase separation in biological systems. *Biomolecules*, 9(9), September 2019.
12. Brian Tsang, Iva Pritišanac, Stephen W Scherer, Alan M Moses, and Julie D Forman-Kay. Phase separation as a missing mechanism for interpretation of disease mutations. *Cell*, 183 (7):1742–1756, December 2020.
13. Kadi L Saar, Alexey S Morgunov, Runzhang Qi, William E Arter, Georg Krainer, Alpha Albert Lee, and Tuomas Knowles. Machine learning models for predicting protein condensate formation from sequence determinants and embeddings. October 2020.
14. Kaiqiang You, Qi Huang, Chunyu Yu, Boyan Shen, Cristoffer Sevilla, Minglei Shi, Henning Hermjakob, Yang Chen, and Tingting Li. PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic Acids Res.*, 48(D1):D354–D359, January 2020.
15. Bálint Mészáros, Gábor Erdős, Beáta Szabó, Éva Schád, Ágnes Tantos, Rawan Abukhairan, Tamás Horváth, Nikoletta Murvai, Orsolya P Kovács, Márton Kovács, Silvio C E Tosatto, Péter Tompa, Zsuzsanna Dosztányi, and Rita Pancsa. PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res.*, 48(D1):D360–D367, January 2020.
16. Qian Li, Xiaojun Peng, Yuanqing Li, Wenqin Tang, Jia'an Zhu, Jing Huang, Yifei Qi, and Zhuqing Zhang. LLPSDB: a database of proteins undergoing liquid–liquid phase separation in vitro. *Nucleic Acids Res.*, September 2019.
17. Plotly Technologies Inc. Collaborative data science, 2015.
18. Jaina Mistry, Robert D Finn, Sean R Eddy, Alex Bateman, and Marco Punta. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.*, 41(12):e121, July 2013.
19. Damiano Piovesan, Ian Walsh, Giovanni Minervini, and Silvio C E Tosatto. FELLS: fast estimator of latent local structure. *Bioinformatics*, 33(12):1889–1891, June 2017.
20. Rhys Heffernan, Kuldip Paliwal, James Lyons, Jaswinder Singh, Yuedong Yang, and Yaoqi Zhou. Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *J. Comput. Chem.*, 39(26):2210–2216, October 2018.
21. Mirko Torrisi, Manaz Kaleel, and Gianluca Pollastri. Porter 5: fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. October 2018.
22. Zhiyong Wang, Feng Zhao, Jian Peng, and Jinbo Xu. Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics*, 11(19):3786–3792, October 2011.
23. Daniel W A Buchan and David T Jones. The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res.*, 47(W1):W402–W407, July 2019.
24. Jack Hanson, Kuldip Paliwal, Thomas Litfin, Yuedong Yang, and Yaoqi Zhou. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, 35(14):2403–2410, July 2019.
25. Bin Xue, Roland L Dunbrack, Robert W Williams, A Keith Dunker, and Vladimir N Uversky. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta*, 1804(4):996–1010, April 2010.
26. Kang Peng, Predrag Radivojac, Slobodan Vucetic, A Keith Dunker, and Zoran Obradovic. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, 7:208, April 2006.
27. Jack Hanson, Kuldip K Paliwal, Thomas Litfin, and Yaoqi Zhou. SPOT-Disorder2: Improved protein intrinsic disorder prediction by ensembled deep learning. *Genomics Proteomics Bioinformatics*, 17(6):645–656, December 2019.
28. David T Jones and Domenico Cozzetto. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, 31(6):857–863, March 2015.
29. Yang Li, Jun Hu, Chengxin Zhang, Dong-Jun Yu, and Yang Zhang. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics*, 35(22):4647–4655, November 2019.
30. Manjeet Kumar, Marc Gouw, Sushama Michael, Hugo Sámano-Sánchez, Rita Pancsa, Juliana Glavina, Athina Diakogianni, Jesús Alvarado Valverde, Dayana Bukirova, Jelena Čalyševa, et al. Elm—the eukaryotic linear motif resource in 2020. *Nucleic Acids Research*, 48(D1):D296–D306, 2020.
31. Michael P Hughes, Michael R Sawaya, David R Boyer, Lukasz Goldschmidt, Jose A Rodriguez, Duilio Cascio, Lisa Chong, Tamir Gonen, and David S Eisenberg. Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks. *Science*, 359(6376):698–701, 2018.
32. P Andrew Chong, Robert M Vernon, and Julie D Forman-Kay. Rgg/rg motif regions in rna binding and phase separation. *Journal of molecular biology*, 430(23):4650–4665, 2018.
33. Izzy Owen and Frank Shewmaker. The role of Post-Translational modifications in the phase transitions of intrinsically disordered proteins. *Int. J. Mol. Sci.*, 20(21), November 2019.
34. Roland Riek. The Three-Dimensional structures of amyloids. *Cold Spring Harb. Perspect. Biol.*, 9(2), February 2017.
35. Simon Alberti, Amy Gladfelter, and Tanja Mittag. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell*, 176(3):419–434, 2019.
36. Greta Bianchi, Sonia Longhi, Rita Grandori, and Stefania Brocca. Relevance of electrostatic charges in compactness, aggregation, and phase separation of intrinsically disordered proteins. *International Journal of Molecular Sciences*, 21(17):6208, 2020.

DRAFT