Computational Resources for Molecular Biology 2024

Rita Casadio (a), David H. Mathews (b), and Michael J.E. Sternberg (c,*)

(a) Biocomputing Group, FABIT-University of Bologna, Bologna I-40126, Italy
rita.casadio@unibo.it

(b) Department of Biochemistry & Biophysics and Center for RNA Biology
University of Rochester
Rochester, NY 14642, USA
David_Mathews@urmc.rochester.edu

(c) Centre for Integrative Systems Biology and Bioinformatics,
Department of Life Sciences
Imperial College London
London SW7 2AZ, UK
m.sternberg@imperial.ac.uk

(*) To whom correspondence should be addressed

We are delighted to have assembled papers for this Special Issue on Computational Resources for Molecular Biology 2024. The range of topics highlights the reach of computational methodology across bioscience and biomedical research. Several of the reported resources exploit the added value of data integration while others focus on the application of a range of machine learning approaches. As in last year's issue, we report several resources that exploit the vast expansion in the number of available protein structures resulting from predictions using advanced machine learning.

Enricherator [1] addresses the question as to when a signal of interest arises genome-wide with "a Bayesan method for inferring regularised genome-wide enrichment from sequencing count data". **NucMap 2.0** [2]is a webserver that provide nucleosome position data across animal, plant, fungal, and protist species. This updated version provides pre-processed human data and information is summarized in a genome browser based on JBrowse. Analysis of genomic data often employs pathway enrichment analysis and consideration of the network structure of the interacting genes. **XGRm** [3]is a web server to enable users to perform such analyses on mouse genomic data thereby extending the developers' resource XGR, which considers human genomic data. **MVAR** [4] is a registry of mouse single nucleotide variants, indels, and their annotations comprising some 99 million canonical single nucleotide variants for 581 strains of mice, and this should also prove useful for biomedical human variant annotation.

Two papers report web servers that provide transcriptomic data. **LogoMotif** [5] provides predicted transcription factor binding sites from Actinobacteria. Genome-wide predictions and visualizations can be browsed for *Streptomyces*. **BeetleAtlas** [6] is a transcription atlas of tissues from the red flour beetle *Tribolium castaneum*. When homologs are known, expression information is also provided for *Drosophila* transcripts.

We present two resources that focus on RNA sequence data. **scRNA-Explorer** [7] provides a web-based tool for processing and analyzing single-cell RNA (scRNA) data. Users upload their own datasets for processing, and the server automates tasks such as clustering cell types and identifying gene correlations. **EVPsort** [8] provides a database of small non-coding RNAs found in extracellular vesicles. The data encompass over 3,000 sequence datasets and provides extensive analysis the small RNAs, including abundance and RNA family assignment.

Several contributions report methodologies to model the folding and stability of RNA. **RNA3DB** [9] is a browsable dataset of structured RNAs derived from the Protein Data Bank (PDB), designed for training and benchmarking deep learning models. **NNDB** [10] is a Nearest Neighbor Database to assist in modeling thermodynamic ensembles of secondary structures with direct application to the prediction of RNA and DNA secondary structure. The database provides the 1999 and 2004 sets of RNA folding nearest neighbor parameters and recent parameters that include the folding stability of sequences with N6-methyadenonosine. **LinearAlifold** [11] predicts the consensus structure of a set of aligned RNA homologs with an efficient algorithm which allows folding even of long nucleotide sequences, scaling linearly with the sequence length and the number of sequences. In this, it successfully complements the previous RNAalifold. **Ribovision2** [12] provides an interactive web server that demonstrates the evolutionary relationships between ribosomal RNA sequences. Users can interactively browse a sequence alignment, a secondary structure, or a tertiary structure; a chosen residue in one is highlighted in the others. Comprehensive alignments of bacterial, eukaryotic, and archaeal sequences are available.

As in previous years, many resources focus on protein analysis. Two papers report web servers for the enhanced prediction of structural features of proteins. The **E-pRSA** server[13] employs embedding to improve the prediction of residue relative solvent accessibility from protein sequence while **flDPnn2** [14] provides an accurate and fast predictor of intrinsic disorder in proteins. We also report the **mACPpred 2.0** server [15] which considers the primary sequence of a naturally-occurring peptide to evaluate if it is predicted to have anticancer properties. The approach uses deep-stacked learning on a set of features together with natural language processing embeddings.

The vast expansion in the number of accurately predicted protein structures resulting from using machine learning has stimulated the development of the updated version of **CATH 2024** [16] which classifies domain structures from experimental PDB structures and predicted structures in the DeepMind's AlphaFold Database thereby increasing the number of CATH relatives in the protein structure landscape. Comparison of different predictions of a protein structure provides an important guide to the accuracies of the models. Accordingly, the server **Alpha&ESMhFolds** [17] enables a user to compare models for human proteins provided by AlphaFold2 and ESM, two recent prediction approaches based on machine learning. A complementary method for model accuracy assessment is provided by the **ModFOLD9** server [18] . Experimentally-based evaluation of protein stability is provided by the **TPPU_DSF** server[19] which calculates thermodynamic parameters using differential scanning fluorimetry (DSF) data.

Increasing protein structures of complexes are determined by a mixture of experimental and computational methods. This issue includes a paper reporting the development from the PDB team and their collaborators of **IHMCIF** [20]**,** which is a data information framework for macromolecular structures determined by Integrative Hybrid Modelling (IHM). The framework is an extension of the mmCIF (macromolecular Crystallographic Information Framework) data representation. The aim is to provide a unified framework to archive and disseminate complexes solved by an integrative approach through the PDB. We also report two specialist databases documenting structural features of protein interacting with ligands. **AHoJ-DB** [21] provides a comprehensive resource for apo-holo pairs of PDB structures for 468,293 biologically relevant protein–ligand interactions across 27,983 proteins. This resource addresses the problem that because proteins undergo conformational rearrangement on ligand docking, a single protein structure is often insufficient to capture conformational diversity. One application of this database would be to provide alternate templates for homology modelling. The **PPInterface** [22] provides details of 680,442 protein/protein interfaces from over 190,000 structures in the PDB. Users can visualise the interface and download details. **CAPRI-Q** [23] provides

a resource to enable users to evaluate the quality of predicted protein complexes using the same metrics as employed in the international blind assessment of protein docking known as CAPRI (Critical Assessment of PRedicted Interactions).

Additional information about protein/protein interactions can be obtained from crosslinking mass spectroscopy (MS) and the **xiVIEW** [24] resource provides users with an interface to access a range of data from the MS crosslinking community including spectra, 2D and 3D information. Focussing on the implications for the role of protein interactions in disease, **AmyloComp** [25] analyses amyloid co-aggregation with the aim of helping the identification of microbial proteins capable of triggering amyloidosis in humans.

The docking of low molecular weight molecules into proteins is a central strategy in computer-aided drug discovery. **DockThor-VS** [26] uses high performance computing to assist virtual screening for drug development. The **@TOME 3.0** [27] server provides an extensive pipeline where a user can input a sequence, generate a predicted protein structure by homology modeling, then identify probable binding sites, and finally perform virtual docking. The web server **GalaxySagittarius-AF** [28] tackles the inverse modelling task and enables users to identify possible target protein structures for a given drug-like small molecule searching both experimentally-determined structures and those predicted by AlphaFold. The server has applications for drug repurposing and the identification of possible drug side effects.

Molecular dynamics is a widely-used technique to study protein dynamics and interactions. We report an extension to the **CHARMM-GUI** [29] toolkit that facilitates input and modelling of a range of protein-ligand linkages and has been benchmarked on over 1,000 covalent ligand structures in the PDB. The resource should prove valuable in studies of covalently bound drugs.

Some resources are directly focussed on plant biology**. Fungtion** [30] is a web server for predicting fungal effectors of plant fungal pathogens. Zhao et al. report the **Leaf senescence database v5.0** [31], which provides a repository of senescence-related genes and associated data. It includes over 31,000 genes from 148 species. The information is searchable with a web interface and includes phenotype information, homologies between genes, and high-throughput datasets such a proteomic and transcriptomic data.

Finally we would like to thank all the authors and the reviewers for their contributions that have enabled us to assemble this Special Issue.

Acknowledgements

RC was supported by: FABIT-ECOI2020 and RFO-UNIBO. DM was supported by National Institutes of Health grant R35GM145283. MJES was supported by Imperial College London.

[1] J.W. Schroeder, P.L. Freddolino, Enricherator: A Bayesian Method for Inferring Regularized Genome-wide Enrichments from Sequencing Count Data, J Mol Biol (2024) 168567. https://doi.org/10.1016/j.jmb.2024.168567.

[2] Z. Nie, Y. Zhao, S. Yu, J. Mai, H. Gao, Z. Fan, Y. Bao, R. Li, J. Xiao, NucMap 2.0: an updated database of genome-wide nucleosome positioning map across species, J Mol Biol (2024) 168655. https://doi.org/10.1016/j.jmb.2024.168655.

[3] C. Eickhorst, R. Babic, J. Rush-kittle, L. Lucya, C. Meisinger, D. Slade, XGRm: A Web Server for Interpreting Mouse Summary-level Genomic Data, J Mol Biol (2024) 168705. https://doi.org/10.1016/j.jmb.2024.168705.

[4] B. El Kassaby, F. Castellanos, M. Gerring, G. Kunde-Ramamoorthy, C.J. Bult, MVAR: A Mouse Variation Registry, J Mol Biol (2024) 168518. https://doi.org/10.1016/j.jmb.2024.168518.

[5] H.E. Augustijn, D. Karapliafis, K.M.M. Joosten, S. Rigali, G.P. van Wezel, M.H. Medema, LogoMotif: A Comprehensive Database of Transcription Factor Binding Site Profiles in Actinobacteria, J Mol Biol (2024) 168558. https://doi.org/10.1016/j.jmb.2024.168558.

[6] D.P. Leader, M.T. Naseem, K. V. Halberg, BeetleAtlas: An Ontogenetic and Tissue-specific Transcriptomic Atlas of the Red Flour Beetle Tribolium castaneum, J Mol Biol (2024) 168520. https://doi.org/10.1016/j.jmb.2024.168520.

[7] I. Baltsavia, A. Oulas, T. Theodosiou, M.D. Lavigne, E. Andreakos, G. Mavrothalassitis, I. Iliopoulos, scRNA-Explorer: An end-user online tool for single cell RNA-seq data analysis featuring gene correlation and data filtering, J Mol Biol (2024) 168654. https://doi.org/10.1016/j.jmb.2024.168654.

[8] H. chang Chen, J. Wang, R.J. Coffey, J.G. Patton, A.M. Weaver, Y. Shyr, Q. Liu, EVPsort: An Atlas of Small ncRNA Profiling and Sorting in Extracellular Vesicles and Particles, J Mol Biol (2024) 168571. https://doi.org/10.1016/j.jmb.2024.168571.

[9] M. Szikszai, M. Magnus, S. Sanghi, S. Kadyan, N. Bouatta, E. Rivas, RNA3DB: A structurally-dissimilar dataset split for training and benchmarking deep learning models for RNA structure prediction, J Mol Biol (2024) 168552. https://doi.org/10.1016/j.jmb.2024.168552.

[10] A. Mittal, D.H. Turner, D.H. Mathews, NNDB: An Expanded Database of Nearest Neighbor Parameters for Predicting Stability of Nucleic Acid Secondary Structures, J Mol Biol (2024) 168549. https://doi.org/10.1016/j.jmb.2024.168549.

[11] L. Zhang, S. Li, H. Zhang, D.H. Mathews, L. Huang, LinearAlifold: Linear-Time Consensus Structure Prediction for RNA Alignments., J Mol Biol (2024) 168694. https://doi.org/10.1016/j.jmb.2024.168694.

[12] H.M. McCann, C.D. Meade, B. Banerjee, P.I. Penev, L. Dean Williams, A.S. Petrov, RiboVision2: A Web Server for Advanced Visualization of Ribosomal RNAs, J Mol Biol (2024) 168556. https://doi.org/10.1016/j.jmb.2024.168556.

[13] M. Manfredi, C. Savojardo, P.L. Martelli, R. Casadio, E-pRSA: Embeddings Improve the Prediction of Residue Relative Solvent Accessibility in Protein Sequence, J Mol Biol (2024) 168494. https://doi.org/10.1016/j.jmb.2024.168494.

[14] K. Wang, G. Hu, S. Basu, L. Kurgan, flDPnn2: Accurate and Fast Predictor of Intrinsic Disorder in Proteins, J Mol Biol (2024) 168605. https://doi.org/10.1016/j.jmb.2024.168605.

[15] V.K. Sangaraju, N.T. Pham, L. Wei, X. Yu, B. Manavalan, mACPpred 2.0: Stacked Deep Learning for Anticancer Peptide Prediction with Integrated Spatial and Probabilistic Feature Representations, J Mol Biol (2024) 168687. https://doi.org/10.1016/j.jmb.2024.168687.

[16] V.P. Waman, N. Bordin, R. Alcraft, R. Vickerstaff, C. Rauer, Q. Chan, I. Sillitoe, H. Yamamori, C. Orengo, CATH 2024: CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds, J Mol Biol (2024) 168551. https://doi.org/10.1016/j.jmb.2024.168551.

[17] M. Manfredi, C. Savojardo, G. Iardukhin, D. Salomoni, A. Costantini, P.L. Martelli, R. Casadio, Alpha&ESMhFolds: A Web Server for Comparing AlphaFold2 and ESMFold Models of the Human Reference Proteome, J Mol Biol (2024) 168593. https://doi.org/10.1016/j.jmb.2024.168593.

[18] L.J. McGuffin, S.M.A. Alharbi, ModFOLD9: A Web Server for Independent Estimates of 3D Protein Model Quality, J Mol Biol (2024) 168531. https://doi.org/10.1016/j.jmb.2024.168531.

[19] P. Martin-Malpartida, C. Torner, A. Martinez, M.J. Macias, TPPU_DSF: A Web Application to Calculate Thermodynamic Parameters Using DSF Data, J Mol Biol (2024) 168519. https://doi.org/10.1016/j.jmb.2024.168519.

[20] B. Vallat, B.M. Webb, J.D. Westbrook, T.D. Goddard, C.A. Hanke, A. Graziadei, E. Peisach, A. Zalevsky, J. Sagendorf, H. Tangmunarunkit, S. Voinea, M. Sekharan, J. Yu, A.A.M.J.J. Bonvin, F. DiMaio, G. Hummer, J. Meiler, E. Tajkhorshid, T.E. Ferrin, C.L. Lawson, A. Leitner, J. Rappsilber, C.A.M. Seidel, C.M. Jeffries, S.K. Burley, J.C. Hoch, G. Kurisu, K. Morris, A. Patwardhan, S. Velankar, T. Schwede, J. Trewhella, C. Kesselman, H.M. Berman, A. Sali, IHMCIF: An Extension of the PDBx/mmCIF Data Standard for Integrative Structure Determination Methods: IHMCIF Data Standard for Integrative Structures, J Mol Biol (2024) 168546. https://doi.org/10.1016/j.jmb.2024.168546.

[21] C.P. Feidakis, R. Krivak, D. Hoksza, M. Novotny, AHoJ-DB: A PDB-wide Assignment of apo & holo Relationships Based on Individual Protein–Ligand Interactions: A database of apo-holo structure pairs, J Mol Biol (2024) 168545. https://doi.org/10.1016/j.jmb.2024.168545.

[22] Z. Abali, Z. Aydin, M. Khokhar, Y.C. Ates, A. Gursoy, O. Keskin, PPInterface: A Comprehensive Dataset of 3D Protein-Protein Interface Structures, J Mol Biol (2024) 168686. https://doi.org/10.1016/j.jmb.2024.168686.

[23] K.W. Collins, M.M. Copeland, G. Brysbaert, S.J. Wodak, A.M.J.J. Bonvin, P.J. Kundrotas, I.A. Vakser, M.F. Lensink, CAPRI-Q: The CAPRI resource evaluating the quality of predicted structures of protein complexes, J Mol Biol (2024) 168540. https://doi.org/10.1016/j.jmb.2024.168540.

[24] C.W. Combe, M. Graham, L. Kolbowski, L. Fischer, J. Rappsilber, xiVIEW: Visualisation of Crosslinking Mass Spectrometry Data, J Mol Biol (2024) 168656. https://doi.org/10.1016/j.jmb.2024.168656.

[25] S.A. Bondarev, M. V. Uspenskaya, J. Leclercq, T. Falgarone, G.A. Zhouravleva, A. V. Kajava, AmyloComp: A Bioinformatic Tool for Prediction of Amyloid Co-aggregation, J Mol Biol (2024) 168437. https://doi.org/10.1016/j.jmb.2024.168437.

[26] I.A. Guedes, M.M. Pereira da Silva, M. Galheigo, E. Krempser, C.S. de Magalhães, H.J. Correa Barbosa, L.E. Dardenne, DockThor-VS: A Free Platform for Receptor-Ligand Virtual Screening, J Mol Biol (2024) 168548. https://doi.org/10.1016/j.jmb.2024.168548.

[27] V. Moreau, T.E. Exner, G. Labesse, @ TOME 3 . 0 : Interfacing Protein Structure Modeling and Ligand Docking, (2024) 168704. https://doi.org/10.1016/j.jmb.2024.168704.

[28] S. Kwon, N. Jung, J. Yang, C. Seok, GalaxySagittarius-AF: Predicting Targets for Drug-Like Compounds in the Extended Human 3D Proteome, J Mol Biol (2024) 168617. https://doi.org/10.1016/j.jmb.2024.168617.

[29] L. Kong, S.J. Park, W. Im, CHARMM-GUI PDB Reader and Manipulator: Covalent Ligand Modeling and Simulation, J Mol Biol (2024) 168554. https://doi.org/10.1016/j.jmb.2024.168554.

[30] J. Li, J. Ren, W. Dai, C. Stubenrauch, R.D. Finn, J. Wang, Fungtion: A Server for Predicting and Visualizing Fungal Effector Proteins, J Mol Biol (2024) 168613. https://doi.org/10.1016/j.jmb.2024.168613.

[31] Y. Zhao, Y. Zhang, S. Li, S. Tan, J. Cao, H.L. Wang, J. Luo, H. Guo, Z. Zhang, Z. Li, Leaf Senescence Database v5.0: A Comprehensive Repository for Facilitating Plant Senescence Research, J Mol Biol (2024) 168530. https://doi.org/10.1016/j.jmb.2024.168530.