Editorial Overview

Computation Resources for Molecular Biology: Special Issue 2019

Molecular biology has now firmly established itself as a data-intensive science. We are witnessing an explosive growth of genome sequence, gene expression and chromatin-associated experimental data, driven by the new, powerful genome-wide assays and the advances in sequencing technology. In addition, there is expansion of data on protein sequences, structures and interactions. The published data have an enormous potential to provide biological insights that were not reported in the original publications, but its rate of accumulation often outpaces the ability of biologists to extract those insights. It is therefore essential to make these data as easy as possible to access, manipulate and interrogate to answer biological questions.

This year's Special Issue on *Computational Resources for Molecular Biology* reflects these developments: in addition to traditionally strong representation of resources for studying protein structure and function, it brings several papers describing the resources that deal with various aspects of data produced by high-throughput sequencing.

The paper by Hait *et al*. [1] describes the updated and expanded release of EXPANDER (EXPression ANalyzer and DisplayER), a Java-based standalone software package for the processing, exploration and biological interrogation of transcriptomic data. Out of the box, it supports the analysis of gene expression data for 18 species, including all most commonly used model organism. Among the latest additions to the software is the functionality to integrate gene expression and ChIP-seq data for the inference and exploration of gene regulatory networks and the associated analysis of enriched DNA motifs.

Cap Analysis Gene Expression (CAGE) is a leading method for the determination of transcription initiation sites and promoters at nucleotide resolution. Its superiority to competing protocols has been demonstrated in a recent systematic comparison [2]. To date, the FANTOM consortium, ENCODE and several independent efforts have accumulated a large collection of high-quality datasets across several species and thousands of tissues and cell lines. CAGE-derived promoter definitions are more precise and more informative than conventional ones based on transcript models and should always be preferred where available. However, due to the non-straightforward relationship between promoters and transcription start locations, these data are still underutilized. In this issue, Abugessaisa *et al*. [3] offer to amend this situation for human and mouse promoters by introducing refTSS, a reference data set of transcription start sites that integrates the CAGE-derived promoter definitions with other resources commonly used to define promoter locations. The authors also provide a web interface to query the collection and download custom subsets thereof.

Whereas differential codon usage between different organisms and between different functional classes of genes is well studied and organism-specific codon usage tables are available, it is less well known that, in many genomes, the frequency of pairs of adjacent codons also deviates from random assortment. In this issue, Alexaki *et al*. [4] describe CoCoPUTs, a resource that compiles *Co*don and *Co*don *P*air *U*sage *T*ables for all genomes from Genbank with available open reading frame (CDS) annotation. The knowledge of codon and codon pair sequences is key to optimising translation in genetic engineering applications and holds the potential to explain the effect of a subset of synonymous mutations in disease.

A major development over the last decade in the area of modeling protein structure from sequence is the use of predicted residue/residue contacts identified from the correlated evolutionary conservation of residues in a multiple sequence alignment. This enables models to be generated without using a known template. Lamb *et al*. [5] report the development of the online PconsFam database, which provides models for structures based on predicted contact maps. The authors considered 6379 Pfam families [6] of unknown structure and were able to generate models for 558 families.

Two papers report the use of protein structure for the interpretation of genetic variants. Quan *et al*. [7] describe a new algorithm DAMpred, available both as an online server and a package, to identify whether a missense variants is disease associated. The prediction is based on a range of features including evolutionary conservation, the physiochemical properties at a variant position, tertiary models from I-TASSER, and residue–residue

contacts in both tertiary structures and binary complexes. A Bayes-guided artificial neural network is trained on these features to yield the prediction. Benchmarking shows higher predictive accuracy than SIFT and PolyPhen2, two widely used resources. However, it is important to note that comparison of predictive accuracy of variant predictors is notoriously difficult as it is highly influenced by the nature of testing set [8].

Ofoegbu *et al*. [9] present PhyreRisk, which is dynamic database to map human genetic variants onto experimental and protein structures predicted by Phyre. A linear display of structural coverage enables a user rapidly to identify whether a missense variant can be located on a structure. A valuable aspect of PhyreRisk is that the user can input genetic variants from either their protein or their genomic location. PhyreRisk links to Missense3D [10] that provides a stereochemical explanation of the effect of a missense variant and was developed to be applicable to both experimental and predicted structures.

Pearce *et al*. [11] present the second generation of EvoDesign, an online server to guide the design of both proteins and protein/protein interfaces. EvoDesign will propose sequences that increase the binding affinity between the two molecules. The approach integrates the identification of sequence evolutionary profiles from related structural complexes, and the evaluation of an optimised interaction energy. The user inputs either a single protein or a protein complex, with an option to load two chains separately and then predict the docked structure. For complexes, EvoDesign optimises not only the interface but the entire protein sequence. The server returns to the user the top 10 designed sequences.

Typically complex diseases such as diabetes and cardiovascular disease are the result of several genetic alterations within a related network. Aguirre-Plans *et al*. [12] report the second version of their tool GUILDify, which is accessible both as a web server and via an R package that provides programmatic access. GUILDify includes information about gene/disease and drug/gene-targets interactions. This is integrated with tissue and species-specific information about protein/protein interactions. Network prioritisation algorithms are included so novel associations can be identified. A new feature in this version is to facilitate two searches and identify overlaps. Applications of this resource include identifying novel genes associated with a disease and the suggestion of novel drug targets.

Smith *et al*. [13] report updates to Binding MOAD (Mother of all Databases), a well-established database of annotated high quality x-ray structures of proteins in complexes with their biologically relevant ligands. For many of the complexes, the database supplies binding affinity data reported in the literature. The update features a new viewer and almost 7000 new protein–ligand structures. In addition, it supplies pre-computed structural similarity scores of proteins, binding folds and ligands, making it a useful resource for polypharmacology—design of drugs that act on multiple protein targets and cellular pathways.

This Special Issue reports software resources covering a wide range of areas. We have three contributions at the genomic level reporting codon usage, transcription binding sites and transcriptome analysis. At the protein level, we report a database of structures predicted from contacts, two resources to interpret missense variants, and a method to optimise the stability of protein structures including complexes. Drug discovery is aided by a resource that integrates genes, diseases, drug targets and the interactome and by an extensive database of protein/ligand interactions. We would like to thank all our contributors to this Special Issue.

References

[1] T.A. Hait, A. Maron-Katz, D. Sagir, D. Amar, I. Ulitsky, C. Linhart, R. Sharan, Y. Shiloh, R. Elkon, R. Shamir, The EXPANDER integrated platform for transcriptome analysis, J. Mol. Biol. (2019) 2398–2406, https://doi.org/10.1016/j.jmb.2019.05.013.

[2] X. Adiconis, A.L. Haber, S.K. Simmons, A.L. Moonshine, Z. Ji, M.A. Busby, X. Shi, J. Jacques, M.A. Lancaster, J.Q. Pan, & A.R., J.Z. Levin, Comprehensive comparative analysis of 5′-end RNA-sequencing methods, Nat. Methods 15 (2018) 505–511.

[3] I. Abugessaisa, S. Noguchi, A. Hasegawa, A. Kondo, H. Kawaji, P. Carninci, T. Kasukawa, refTSS: a reference data set for human and mouse transcription start sites, J. Mol. Biol. (2019) 2407–2422, https://doi.org/10.1016/j.jmb.2019.04.045.

[4] A. Alexaki, J. Kames, D.D. Holcomb, J. Athey, L.V. Santana-Quintero, P.V.N. Lam, N. Hamasaki-Katagiri, E. Osipova, V. Simonyan, H. Bar, A.A. Komar, C. Kimchi-Sarfaty, Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant Gene Design, J. Mol. Biol. (2019) 2434–2441, https://doi.org/10.1016/j.jmb.2019.04.021.

[5] J. Lamb, A.I. Jarmolinska, M. Michel, D. Menéndez-Hurtado, J.I. Sulkowska, A. Elofsson, PconsFam: an interactive database of structure predictions of Pfam families, J. Mol. Biol. (2019) 2442–2448, https://doi.org/10.1016/j.jmb.2019.01.047.

[6] S. El-gebali, J. Mistry, A. Bateman, S.R. Eddy, S.C. Potter, M. Qureshi, L.J. Richardson, G.A. Salazar, A. Smart, E.L.L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S.C.E. Tosatto, R.D. Finn, The Pfam protein families database in 2019, Nucleic Acids Res. 47 (2019) 427–432, https://doi.org/10.1093/nar/gky995.

[7] L. Quan, H. Wu, Q. Lyu, Y. Zhang, DAMpred: recognizing disease-associated nsSNPs through Bayes-guided neural-network model built on low-resolution structure prediction of proteins and protein–protein interactions, J. Mol. Biol. (2019) 2449–2459, https://doi.org/10.1016/j.jmb.2019.02.017.

[8] H. Zhou, M. Gao, J. Skolnick, ENTPRISE: an algorithm for predicting human disease-associated amino acid substitutions from sequence entropy and predicted protein structures, PLoS One 11 (2016), e0150965. https://doi.org/10.1371/journal.pone.0150965.
[9] T.C. Ofoegbu, A. David, L.A. Kelley, S. Mezulis, S.A. Islam, S.F. Mersmann, L. Strömich, I.A. Vakser, R.S. Houlston, M.J. E. Sternberg, PhyreRisk: a dynamic web application to bridge genomics, proteomics and 3D structural data to guide interpretation of human genetic variants, J. Mol. Biol. (2019) 2460–2466 https://doi.org/10.1016/j.jmb.2019.04.043.
[10] S. Ittisoponpisan, S.A. Islam, T. Khanna, E. Alhuzimi, A. David, M.J.E. Sternberg, Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated? J. Mol. Biol. 431 (2019) 2197–2212, https://doi.org/10.1016/j.jmb.2019.04.009.
[11] R. Pearce, X. Huang, D. Setiawan, Y. Zhang, EvoDesign: designing protein – protein binding interactions using evolutionary interface profiles in conjunction with an optimized physical energy function, J. Mol. Biol. (2019) 2467–2476, https://doi.org/10.1016/j.jmb.2019.02.028.
[12] J. Aguirre-Plans, J. Piñero, F. Sanz, L.I. Furlong, N. Fernandez-Fuentes, B. Oliva, E. Guney, GUILDify v2.0: a tool to identify molecular networks underlying human diseases, their comorbidities and their druggable targets, J. Mol. Biol. (2019) 2477–2484, https://doi.org/10.1016/j.jmb.2019.02.027.
[13] R.D. Smith, J.J. Clark, A. Ahmed, Z.J. Orban, J.B. Dunbar, H. A. Carlson, Updates to binding MOAD (Mother of all Databases): polypharmacology tools and their utility in drug repurposing, J. Mol. Biol. (2019) 2423–2433, https://doi.org/10.1016/j.jmb.2019.05.024.

Boris Lenhard
Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London SW7 2AZ, UK
Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London, W12 0NN, UK
E-mail address: b.lenhard@imperial.ac.uk.

Michael J.E. Sternberg
Structural Bioinformatics Group, Centre for Integrative systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London SW7 2AZ, UK
Corresponding author.
E-mail address: m.sternberg@imperial.ac.uk.