AUTHOR:
Chenjerayi Kashangura[1]

AFFILIATION:
[1]Biological Sciences Department, University of Zimbabwe, Harare, Zimbabwe

CORRESPONDENCE TO:
Chenjerayi Kashangura

EMAIL:
ckashangura@kutsaga.co.zw

HOW TO CITE:
Kashangura C. Artificial intelligence enhanced molecular databases can enable improved user-friendly bioinformatics and pave the way for novel applications. S Afr J Sci. 2021;117(1/2), Art. #8151. https://doi.org/10.17159/sajs.2021/8151

ARTICLE INCLUDES:
☐ Peer review
☐ Supplementary material

KEYWORDS:
deep learning, deep reasoning, algorithm

PUBLISHED:
29 January 2021

Artificial intelligence enhanced molecular databases can enable improved user-friendly bioinformatics and pave the way for novel applications

Molecular databases have enabled scientists across the globe to collaborate and contribute to the growth of the databases. The current form of the databases involves researcher input which is acted upon by algorithms developed by bioinformaticians leading to outputs for researchers. Experimental data analysis using the molecular databases normally results in a reduction in cost and time for in vitro experiments preceded by in silico stages. Molecular biology technologies are applied in multiple disciplines, generating enormous amounts of data every day, which, when deposited, requires professional staff to annotate, verify submissions and generally maintain the database. The rapid rise of artificial intelligence (AI) can be used to enhance molecular databases through incorporation of deep learning and deep reasoning to enable the molecular databases to partially self-maintain, bringing novel applications and the potential for an improved user-friendly interface for researchers who are not trained in bioinformatics to generate data that require bioinformatics-related analysis.

Bioinformatics has been around since the 1960s, whilst online molecular databases that handle data generated by disciplines in the life sciences have existed since the 1990s[1], with researchers contributing by submitting data generated through experiments conducted in vitro and in vivo or by utilising the molecular databases for in silico analysis. The backbone of this analysis is bioinformatics presented in various algorithms tailor-made for different molecular data, such as DNA (genomics), RNA (transcriptomics) and protein (proteomics), to produce particular outcomes. This analysis relies on the researcher interrogating the database through the algorithms.

A plethora of databases such as GenBank has been published in the *Nucleic Acids Research* Database issues for the past 26 years, highlighting a range of different molecular data and a synchronous range of bioinformatics capabilities within the databases.[2-4] The tools to use on data contained in the molecular databases is determined by researchers. Often a researcher may opt to use the bioinformatics tools they are well acquainted with and opt not to use other tools which might require arduous training. However, the data submitted to providers of molecular databases require skilled professionals to update and secure systems and verify the accuracy of the data submitted[1] and funds are required for such staff.

The growing availability of AI brings the possibility of self-learning, self-reasoning and self-improving molecular databases that can be considered to be 'next-generation enhanced molecular databases' which can assist in lowering the cost of maintenance, ultimately reducing databases becoming defunct and introducing novel applications and a new in-depth analysis of molecular data. These next-generation databases can enable bioinformatics to become more user-friendly to non-bioinformatics trained researchers who, owing to the multi- and interdisciplinary nature of molecular life sciences data, sometimes face a daunting task in analysing data.

This next generation may take the form of AI led or incorporated databases. These can have various forms of deep learning, deep reasoning and reading algorithms that enable the databases to learn experientially. These databases will not be static but will be capable of increasing the database routine tasks that can be accomplished through experiential learning so that activities such as curation, annotation and archiving can be partly accomplished, with human verification required, together with the introduction of new potential applications.

Molecular databases enhanced with AI may be useful for a new level of deep analysis that involves the AI selecting the parameters to be used through experiential learning in areas that include three-dimensional modelling[5], binding predictions and domain calling[6], epigenome, especially differential DNA methylation patterns[7], deep analysis of multi-omic data[8,9], sequence-based taxonomy[3], precision medicine and drug discovery[10,11]. The experiential learning capability suggests that the database can, for example, interrogate a submitted set of molecular data and determine the nature of the data and carry out basic analysis to assign identity, annotation and generation of output from the AI selected parameters, like a phylogenetic tree or identified potential antimicrobial agents. A user interface can then pop up with suggestions for further data analysis, for example, if the AI has identified a potential therapeutic agent against existing or emerging pathogens, then prediction AI simulations[12] are selected and run using deep learning algorithms that would run in silico trained parameter algorithms for predicting pathogen 'growth inhibition', or 'neutralise pathogen receptor access', or 'prevent multiplication'. This will greatly assist those researchers who are not bioinformatics trained to generate molecular life sciences data that require bioinformatics analysis. These next-generation databases can potentially provide improved user-friendly interaction with bioinformatics by providing single-click buttons for running particular bioinformatics tools whilst the parameters are selected by the AI after analysis of the submitted data set.

The current challenge of big data analysis may be ameliorated by development of AI that performs a comparative analysis of recently submitted big data sets against existing similar data sets and possibly suggests areas that need modification in terms of analysis for bioinformaticians to develop.

Competing interests

I declare that there are no competing interests.

References

1. Imker HJ. 25 Years of molecular biology databases: A study of proliferation, impact and maintenance. Front Res Metr Anal. 2018;3, Art. #18, 13 pages. https://doi.org/10.3389/frma.2018.00018
2. Johnson G, Wu TT. Kabat Database and its applications: 30 Years after the first variability plot. Nucleic Acids Res. 2000;28(1):214–218. https://doi.org/10.1093/nar/28.1.214
3. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2013;41(D1):D36–D42. https://doi.org/10.1093/nar/gks1195
4. Rigden DJ, Fernández XM. The 26th annual Nucleic Acids Research database issue and molecular biology database collection. Nucleic Acids Res. 2019;47(D1):D1–D7. https://doi.org/10.1093/nar/gky1267
5. Ay F, Noble WS. Analysis methods for studying the 3D architecture of the genome. Genome Biol. 2015;16(1), Art. #183, 15 pages. https://doi.org/10.1186/s13059-015-0745-7
6. Joyce AP, Zhang C, Bradley P, Havranek, JJ. Structure-based modelling of protein: DNA specificity. Brief Funct Genomics. 2014;14(1):39–49. https://doi.org/10.1093/bfgp/elu044
7. Huh I, Wu X, Park T, Yi SV. Detecting differential DNA methylation from sequencing of bisulfite converted DNA of diverse species. Brief Bioinform. 2019;20(1):33–46. https://doi.org/10.1093/bib/bbx077
8. Sangalaringam A, Ullah AZD, Marzek J, Gadaleta E, Nagano A, Ross-Adams H, et al. 'Multi-omic' data analysis using O-miner. Brief Bioinform. 2019;20(1):130–143. https://doi.org/10.1093/bib/bbx080
9. Spies D, Renz PF, Beyer TA, Ciaudo C. Comparative analysis of differential gene expression tools for RNA sequencing time course data. Brief Bioinform. 2019;20(1):288–289. https://doi.org/10.1093/bib/bbx115
10. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688–702. https://doi.org/10.1016/j.cell.2020.01.021
11. Hoffman M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-Cov-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically-proven protease inhibitor. Cell. 2020;181(2):271–280. https://doi.org/10.1016/j.cell.2020.02.052
12. Rodriguez AC. Simulation of genes and genomes forward in time. Curr Genomics. 2010;11(1):58–61. https://doi.org/10.2174/138920210790218007