Indian J. L. Sci. 12 (1): 01-04, 2022 Perspective Review ISSN: 2277-1743 (Print)
ISSN: 2278-7879 (Online)

A Synopsis on the Yin Yang of Biology and Evolving Computational Science

Shaurya Jauhari[1*]

[1]Education, Training, and Assessment Division, Infosys Limited, India.
E-mail: shauryajauhari@yahoo.com

ABSTRACT

With the continuous evolution and adaptation of analytical approaches, the digital pro forma that Biology now presents has become a compelling domain of work. As a discipline, Biology has undergone multifaceted growth, effectively leveraging the benefits of significant methodological advancements. This transformation has shifted the field from its historical reliance on qualitative observations and foundational discoveries to one increasingly dominated by high-throughput data generation and sophisticated computational analysis, setting the stage for an even more integrated and predictive future. This review aims to briefly document this profound journey, exploring the field's past, its current state, and my perspective on its promising future.

KEYWORDS: Artificial intelligence, High-throughput data, Deep learning. Biology, Disease progression

According to Coursera, "Biological and Biomedical Sciences" is the fourth most popular major, accounting for approximately 6.5% of all bachelor's degrees conferred. From the aesthetic anatomical sketches of Leonardo da Vinci and the foundational insights of Gregor Mendel and Charles Darwin to the landmark discovery of the DNA double-helix structure by James Watson, Francis Crick, Rosalind Franklin, and Maurice Wilkins in 1953, and culminating in the monumental Human Genome Project initiated in 1989 and completed in 2001, Biology has consistently inspired intellectual excitement and scholarly exploration (Watson & Crick, 1953; International Human Genome Sequencing Consortium, 2001).

At the same time, Mathematics and Statistics, although traditionally outside the boundaries of Biology, have evolved with equal enthusiasm. These disciplines now form the foundation of Data and Computational Sciences, with artificial intelligence (AI) emerging as a defining achievement in the field.

As biological data becomes increasingly digitized, data science algorithms offer promising solutions to complex biological questions that have long puzzled wet-lab researchers. Although the cost of genome sequencing has significantly declined and associated technologies have improved, the task of deciphering the biological meaning behind the data remains challenging. This highlights the immense potential of digitally driven Biology in revealing insights once hidden in genetic codes.

Consequently, Computational Biology and Bioinformatics have emerged as pivotal disciplines. They represent the critical intersection where biology and data science converge to translate raw genomic information into meaningful scientific knowledge.

Could the Intelligence be *artificial*?

The evolving intersection of biology and computation has sparked a profound question in the scientific community: can intelligence be artificially created? The answer is a resounding yes. The rapid progress in robotics and automation serves as tangible proof. Interestingly, the foundation of AI partly lies in the fundamental principles of neuronal biology. The architecture of Artificial Neural Networks (ANNs) is modeled on how information is transmitted and processed through networks of neurons in the human brain (Wang 2003). Deep learning represents a more advanced extension of this concept, allowing for the management of complex data and dynamic refinements through internal feedback mechanisms (LeCun et al., 2015).

In the realm of data science, the objective is often dictated by the nature of the data itself. As Chris Anderson articulated in Wired magazine, the key innovation today is identifying the right question to ask given the data at hand. To uncover hidden patterns that define a specific cohort or category, the clustering or unsupervised learning approach becomes essential. In this paradigm, outcomes or class labels are not known beforehand. For example, when presented with an assortment of fruits, one might identify common features such as acidity or shape. A citrus fruit like an orange would naturally be grouped separately from a *Musa* species like a banana (D'Hont et al., 2012).

On the other hand, classification tasks involve mapping a new sample to an existing category using its attributes. Consider a scenario in which an excavation in the Arctic reveals a partially fossilized fruit embedded in ancient meltwater layers. By analyzing its physical traits, one could infer its biological origin. This process closely

***Corresponding author**

resembles a retrieval exercise from an internalized encyclopedia of prior knowledge.

Therefore, AI goes beyond a mere technological tool, it reflects a deep integration of biological principles with computational logic. As the distinction between human and machine intelligence fades, we are entering an era of unprecedented growth in our ability to create and understand intelligent systems.

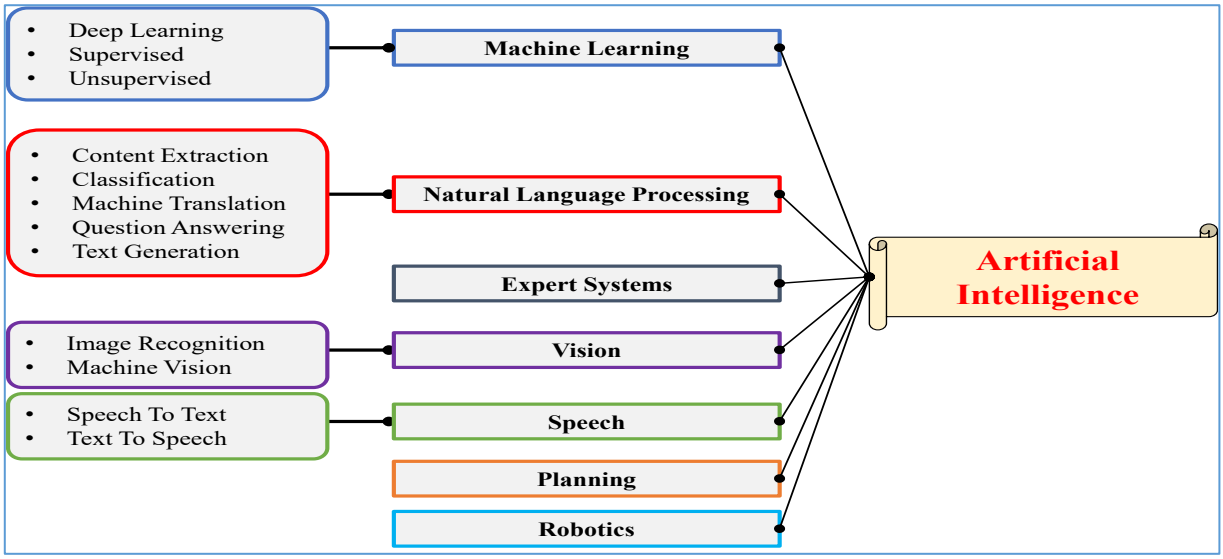

Figure 1: The overarching field of Artificial Intelligence incubates majority of applications in Data Science. Ultimately, it is all about the computational optimality to corroborate a precise, analytical pipeline, and henceforth, meaningful insights. [*adapted from* Ye D (2020)].

Generative AI

Given the scope of available data, the application of algorithms related to Machine Learning, Neural Networks, and Deep Learning has, to some extent, become commonplace. A multitude of research submissions now routinely employ these methodologies. While the concept of AI was first introduced by John McCarthy, a professor at Stanford University in 1955, with the aim of emulating human cognition, these disciplines originally had a long and relatively dormant existence rooted in mathematics. They were later revitalized in response to the growing demand for advanced data analysis.

In the current technological landscape, however, the focus has increasingly shifted toward the prominence of Generative AI. As the name implies, Generative AI systems learn from existing data artifacts to create new, realistic outputs (Jovanovic and Campbell, 2022). By leveraging aggregated knowledge, implicit correlations, and internalized reasoning, these models deliver remarkably sophisticated results. Large Language Models (LLMs), in particular, have emerged as highly scalable tools capable of generating coherent and contextually rich responses, despite the potential limitations caused by repetitive patterns in their training data.

Built upon deep learning architectures, LLMs are typically pre-trained using supervised learning techniques. A well-crafted prompt, consisting of carefully chosen input keywords, can guide the model's attention mechanism to produce high-quality, context-aware outputs. This evolution marks a significant advancement in our ability to generate, interpret, and interact with language-driven AI systems, demonstrating the practical realization of AI in everyday life.

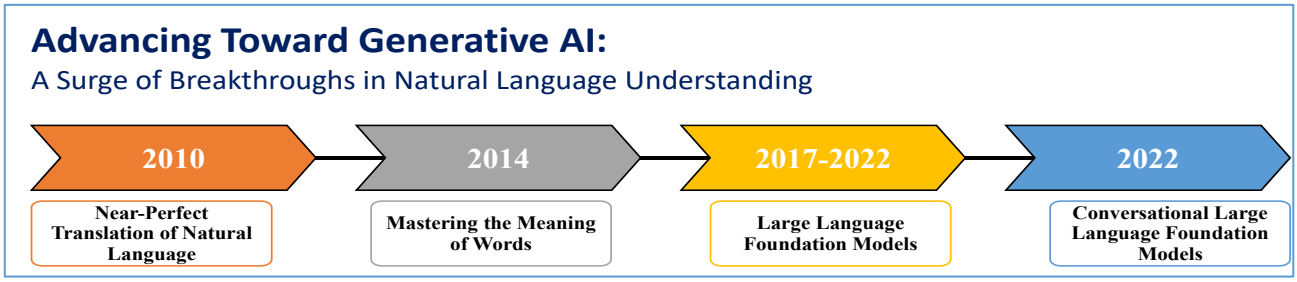

Figure 2: The recent timeline of tangible Generative AI development. (Adapted from: Gartner; https://www.gartner.com/en/topics/generative-ai**).**

Applications in Biology

As suggested earlier, there is now a vast array of aspects in Biology that can be computationally explored. These include, but are not limited to, the prediction of biomarkers, protein structure elucidation, and the shortlisting of druggable gene candidates in relation to diseased physiological states. A narrow set of examples from neuroscience are presented below.

Ocular and MRI Scans

Ophthalmic imaging modalities, such as ocular and magnetic resonance imaging (MRI) scans, represent a crucial use case in medical image analysis. Brain disorders account for a substantial number of mortalities worldwide. Neuroimages often present issues related to morphology and scale, and conclusions about the underlying anatomy are further complicated by noise in the data. This increases the complexity of interpretation, particularly when medical practitioners are expected to perform this task manually. Additionally, the subjectivity inherent in the diverse expertise levels of physicians makes it challenging to reach a consensus, and the adequacy of clinical sample sizes remains debatable.

Alzheimer's and Parkinson's Diseases

As previously mentioned, AI algorithms often outperform traditional physical examinations in determining the extent of disease progression in a patient. More significantly, these algorithms can predict the likely onset of a disease and monitor its course using pathology and diagnostic reports. This capability is crucial, as early detection can allow for timely medical intervention and potentially spare the individual from the full impact of the illness (Nussbaum and Ellis, 2003).

Beyond image analysis, the development of fluid biomarkers for the detection of various diseases, including cancers, can reinforce computational findings through wet-lab validation. Furthermore, the analysis of biopsy samples can help confirm computational results. Collectively, *in vitro*, *in vivo*, and *in silico* approaches work synergistically to improve predictive outcomes and lend greater confidence to scientific hypotheses.

A Prelude to Proteins and the Core Challenge

Shifting focus to proteins, undoubtedly the central players in all physiological processes, their structure fundamentally dictates their function. Within a living organism, a diverse array of cells exists, each specialized in function, based on the quantity and quality of proteins synthesized from messenger RNAs (mRNAs) through the cellular process of translation (Jauhari and Rizvi, 2017). Notably, not all transcripts are translated into proteins, and these translation patterns vary among cell types, each with distinct nuclear genomic content. Therefore, gene expression profiles and resulting protein structures have a direct influence on an organism's physiological state, whether *normal* or *diseased*. This framework presents a compelling opportunity for differential analysis, which can help illuminate key distinctions between health and pathology.

While gene and protein expression levels can be measured, understanding protein structure and function is even more critical, given that proteins are the ultimate functional products of gene expression. Protein structure prediction has a long history filled with both breakthroughs and setbacks. Traditional experimental methods such as Nuclear Magnetic Resonance (NMR) spectroscopy, Electron Microscopy, and X-ray Crystallography have long been used for structure elucidation. However, recent advances in computational approaches have reshaped the landscape. In a de novo protein structure prediction task, the amino acid sequence of an unknown protein is compared with sequences of previously studied proteins, using alignment strategies. This comparative approach is foundational, as biological relevance often lies in similarity. However, the scarcity of sequence homologs poses a significant challenge. At this juncture, AI offers a promising solution by generating proxy data that can provide meaningful insights.

Additionally, accurate prediction of secondary structural features, such as helices, turns, sheets, and strands, is crucial to faithfully model a protein's structure. At higher organizational levels, predicting ligand–receptor interactions and protein–protein interactions provide a more holistic understanding from a systems biology perspective.

The Methods

Among contemporary tools such as RaptorX and OmegaFold, AlphaFold stands out as a remarkable AI model, perhaps due in part to its strong institutional backing, for its ability to predict the three-dimensional structure of proteins from their amino acid sequences. AlphaFold, developed by DeepMind (a subsidiary of Alphabet), utilizes a network-based architecture (Jumper et al., 2021). This AI model has demonstrated high accuracy and efficiency in predicting protein folding, a notoriously difficult NP-hard problem and a long-standing grand challenge in computational biology.

While each tool has its strengths, AlphaFold's performance has set a new benchmark in the field, prompting comparisons with other structure prediction models. Ultimately, the type and complexity of the target

protein remain key factors influencing which model performs best in a given context.

Conflicts of Interest: The author declares no conflicts of interest.

REFERENCES

D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, Da Silva C, Jabbari K, Cardi C, Poulain J, Souquet M, Labadie K, Jourda C, Lengellé J, Rodier-Goud M, Alberti A, Bernard M, Correa M, Ayyampalayam S, McKain MR, Leebens-Mack J, Burgess D, Freeling M, Mbéguié-A-Mbéguié D, Chabannes M, Wicker T, Panaud O, Barbosa J, Hribova E, Heslop-Harrison P, Habas R, Rivallan R, Francois P, Poiron C, Kilian A, Burthia D, Jenny C, Bakry F, Brown S, Guignon V, Kema G, Dita M, Waalwijk C, Joseph S, Dievart A, Jaillon O, Leclercq J, Argout X, Lyons E, Almeida A, Jeridi M, Dolezel J, Roux N, Risterucci A-M, Weissenbach J, Ruiz M, Glaszmann J-C, Quétier F, Yahiaoui N, Wincker P (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature **488**: 213-217.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature, **409**: 860–921.

Jauhari S, Rizvi SAM (2017) A priori, de novo mathematical exploration of gene expression mechanism via regression viewpoint with briefly cataloged modeling antiquity. International Journal of Biomathematics **10:** 1750006.

Jovanovic M, Campbell M (2022) Generative Artificial Intelligence: Trends and Prospects. Computer **55**: 107-112

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature **596**: 583-589

LeCun, Y., Bengio, Y. & Hinton, G. Deep learning (2015). *Nature*, **521**: 436-444.

Nussbaum RL, Ellis CE (2003) Alzheimer's disease and Parkinson's disease. N Engl J Med **348**: 1356-1364

Wang Sun-Chong, Artificial neural network. (2003). In *Interdisciplinary computing in java programming*, 81–100. Springer.

Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature, **171**: 737–738

Ye D (2020) Artificial Intelligence and Deep Learning Application in Evaluating the Descendants of Tubo Mgar Stong Btsan and Social Development. In. Springer Singapore, Singapore, pp 1869-1876