Initial Teaching plans

Bioinformatics for Functional Annotation and Phylogeny

My teaching philosophy emphasizes fostering curiosity and critical thinking through an engaging learning environment. I aim to make complex biological concepts, like enzyme function and evolution, accessible through practical bioinformatics examples. By incorporating real-world applications and promoting collaborative learning, I hope to inspire students to apply interdisciplinary approaches to solve biological challenges. I propose a course titled "Bioinformatics for Functional Annotation and Phylogenetics," covering essential bioinformatics concepts (Homology and Analogy, Galperin & Koonin, 2012), tools and databases (Mount, 2004; Lesk, 2020), with a focus on functional annotation of proteins, identification of homologous sequences, and phylogenetic analysis. The course will introduce bioinformatics basics, such as exploring databases like NCBI and UniProt, and will teach practical skills using tools like BLASTp (Altschul et al., 1990) and InterProScan (Jones et al., 2014) for protein function prediction. Students will explore sequence alignment algorithms like MAFFT (Katoh & Standley, 2013) and the Needleman-Wunsch (Needleman & Wunsch, 1970) and Smith-Waterman (Smith & Waterman, 1981) algorithms, with workshops on multiple sequence alignment and sequence conservation. The course will also cover phylogenetic tree construction using methods like maximum likelihood and Bayesian inference (Felsenstein, 1981; Yang, 2014), and introduce AI-driven tools such as AlphaFold2 (Jumper et al., 2021) for protein structure prediction. Students will apply these concepts to global challenges, including enzymes related to climate change and pollution. They will formulate hypotheses and mini-projects on enzyme functions (Yoshida et al., 2016; Kumar & Bharadvaja, 2019) that can be utilized through synthetic biology to address emerging issues such as global warming, environmental pollution from metals, plastics, textiles, etc. They will design *in silico* experiments and work collaboratively on projects. Teaching methods will include lectures and seminars to introduce theoretical concepts (30%), hands-on workshops for practical experience (40%), group projects to encourage teamwork (30%) with presentations and discussions to develop communication skills and critical thinking. Assessment will consist of in-class assignments (30%) involving short reports, a group project (40%) culminating in a detailed report and presentation, and examinations (30%) to assess understanding of theoretical concepts. By the end of the course, students will gain proficiency in bioinformatics tools, develop scientific hypotheses, enhance teamwork and communication skills, and understand biotechnology's role in solving environmental challenges. The course aims to bridge the gap between theory and practical application, preparing students for future careers in science and biotechnology.

Integration of Artificial Intelligence (AI) and Bioinformatics

My teaching philosophy centers on fostering curiosity and developing critical thinking skills in postgraduate students. I aim to create an engaging learning environment where complex concepts—such as the integration of Artificial Intelligence (AI) and Bioinformatics—become accessible through practical examples and hands-on experiences. By emphasizing real-world applications and collaborative learning, my goal is to inspire students to think creatively and apply interdisciplinary approaches to solve complex biological challenges. I propose to offer a foundational postgraduate course titled **"Integrating AI and Bioinformatics for Biotechnological Applications (Postgraduate Level)."** This course will cover recent advancements in AI and Bioinformatics integration, focusing on Natural Language Processing (NLP) models like Word2Vec (Mikolov et al., 2013) and neural networks such as the Perceptron (Rosenblatt, 1958), Multilayer Perceptron (Rumelhart et al., 1986), and transformers (Vaswani et al., 2017). Students will learn how these algorithms can extract protein sequence embeddings to identify hidden patterns in protein sequences and integrate them with traditional machine learning methods such as Support Vector Machines (SVM) and Random Forest for functional classification of proteins. The course will also explore advanced models based on transformer algorithms like ESM2 (Lin et al., 2023) for enzyme annotation and ProtGPT2 (Ferruz et al., 2022) for protein design. A significant component of the course will be practical workshops using Python notebooks exploring the methods presented. Students will engage in hands-on coding sessions to implement the algorithms and models discussed, allowing them to apply AI techniques to real bioinformatics problems. During the course, students will be divided into groups, with each group assigned to present one of the methods covered and lead a corresponding Python workshop showing the use of the methods presented during the class session. This collaborative approach will enhance peer learning and ensure a deep understanding of each technique. Throughout the course, we will demonstrate the integration of these methods with both supervised learning algorithms (e.g., SVM, Random Forest) used to train classifiers aimed at the functional annotation of sequences, and unsupervised learning algorithms (e.g., K-means clustering, Principal Component Analysis (PCA), Non-metric Multidimensional Scaling (NMDS)) for the analysis of embeddings generated from protein sequences. Topics will include the history of Natural Language Processing and large language models, foundational AI concepts like the Perceptron and Multilayer Perceptron, advanced AI models applied to functional annotation of proteins and protein prediction (3D structures), practical applications of AI in biomedical research—including drug discovery—and short discussions on ethical and security considerations. Assessment will consist of in-class assignments (20%), Python workshops (20%), and group presentations (30%) where students present on the assigned methods and propose how these methods can be used for other applications, participation (20%), and a final exam (10%). By the end of the course, students will have a solid understanding of foundational AI concepts and advanced AI models, equipping them with practical skills to apply AI in predicting protein functions, designing novel proteins, and addressing real-world challenges in molecular biology and biotechnology—all while considering ethical and security implications.

Outreach Vision

Outreach is a fundamental part of my academic vision, as I believe that scientific knowledge should extend beyond the classroom and laboratory. I plan to engage with different communities by organizing free workshops presented by students, showcasing the programs they developed during their training and internships in the lab, and hosting small talks organized with the students on topics involving bioinformatics and artificial intelligence to demystify complex subjects such as enzyme engineering and AI. These events will be tailored to various audiences, including school students, business leaders, and the general community. The goal of this outreach is to stimulate interest in science, especially among underrepresented groups, by demonstrating how AI can lead to environmentally friendly industrial solutions. Additionally, I see great potential in using online platforms and social media (Instagram, Facebook, and LinkedIn and Youtube) to share research findings in an accessible format, ensuring the general public understands the societal impact of biotechnological innovations. Through collaboration with schools and educational programs, I plan to introduce young students to biotechnology and bioinformatics concepts, aiming to foster future interest in STEM fields.

Technology Transfer (Valorization) Vision

My project on Non-Homologous Isofunctional Enzymes (NISEs) has significant potential for technology transfer, particularly in industries such as biofuels, pharmaceuticals, and environmental sustainability. I envision the development of a comprehensive database of NISEs, along with AI programs based on transformer algorithms like ProtGPT2, becoming a vital tool for both academic and industrial partners. This platform will facilitate the creation of custom-designed enzymes and genetic circuits with enhanced efficiency and specificity for industrial applications. My goal is to establish collaborations with industry leaders to explore the commercialization of synthetic enzymes generated through this project, and to better understand market needs and the existing gaps in tools that need to be developed to meet societal demands. This could lead to patents and licensing opportunities. Additionally, I plan to work closely with technology transfer offices to ensure that research outcomes are translated into tangible products and processes, ultimately contributing to sustainable industrial practices and economic growth.

Other potential disciplines/courses: Python Programming, Algorithms, Introduction to Bioinformatics, Microbiome and Metagenomic Analyses, Functional annotation of Genomes, RNAseq and Transcriptome Analyses—topics in which I have extensive experience in data analysis.

References:

- Galperin, M. Y., & Koonin, E. V. (2011). Divergence and convergence in enzyme evolution. *Journal of Biological Chemistry, 287*(1), 21–28. https://doi.org/10.1074/jbc.R111.241976
- Mount, D. (2004). *Bioinformatics: Sequence and Genome Analysis* (2nd ed.). Cold Spring Harbor Laboratory Press.
- Lesk, A. (2020). *Introduction to Bioinformatics* (5th ed.). Oxford University Press.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2
- Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., & Nuka, G. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics, 30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution, 30*(4), 772-780. https://doi.org/10.1093/molbev/mst010
- Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology, 48*(3), 443-453. https://doi.org/10.1016/0022-2836(70)90057-4
- Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology, 147*(1), 195-197. https://doi.org/10.1016/0022-2836(81)90087-5
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution, 17*(6), 368-376. https://doi.org/10.1007/BF01734359
- 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, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

- Yoshida, S., Hiraga, K., Takehana, T., Taniguchi, I., Yamaji, H., Maeda, Y., Toyohara, K., Miyamoto, K., Kimura, Y., & Oda, K. (2016). A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science, 351*(6278), 1196-1199. https://doi.org/10.1126/science.aad6359
- Kumar, L., & Bharadvaja, N. (2019). **Enzymatic bioremediation: A smart tool to fight environmental pollutants**. In *Smart Bioremediation Technologies*, *Microbial Enzymes* (pp. 99–118).
- Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Rosenblatt, F. (1958). **The perceptron: A probabilistic model for information storage and organization in the brain**. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533–536. https://doi.org/10.1038/323533a0.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). **Attention is all you need**. *arXiv preprint arXiv:1706.03762v7*. https://arxiv.org/abs/1706.03762
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). **Evolutionary-scale prediction of atomic-level protein structure with a language model**. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574