Bio-Informatics: Computer Science For Biological Modeling And Computation

Mario Panuco

Abstract—Bio-Informatics is a relatively young scientific field, that in many ways, is still searching for it's identity. It's applications can change the way medicine and epidemiology affects everyone's daily life. There are also applications of the field to things that don't directly affect most lives, but instead are meant to analyze ecosystems for the purpose of conservation and research.

COMPUTER science has found a way into almost every modern industry and science. Bio-informatics is one of the fields that has blended two sciences in a powerful way, where both biologists and computer scientists have the opportunity to collaborate and contribute through a mutual scientific language. This new scientific language has a promising future as both fields innovate and collaborate on applications of computer science to biological systems.

1 INTRODUCTION

1.1 Purpose

The purpose of this paper is to present a cursory overview of bio-informatics and its growing importance.

The purpose of the project is to educate other UCSC STEM students on the basics of bio-informatics and the importance of bio-informatics in the modern-world. I hope the students learn to expand their knowledge base on the limits of science and get inspired to apply their skills to other fields outside their major.

1.2 Intended Audience

The intended audience for this project is an undergraduate student with rudimentary knowledge of computer science and/or biology.

1.3 Style Guide

I will be styling my in LaTeX document according to the IEEE standards.

CONTENTS

0000–0000/00$1738.00 © 2007 IEEE

2 Origins of Bio-Informatics

The first baby-steps for bio-informatics encompassed creating databases for large data-sets and libraries of sequenced data. The first major data-set breakthrough was Dr. Margret Oakley's protein sequence database, which was created using computational models, entirely programmed on punch cards in FORTRAN (see Figure 1 for example).

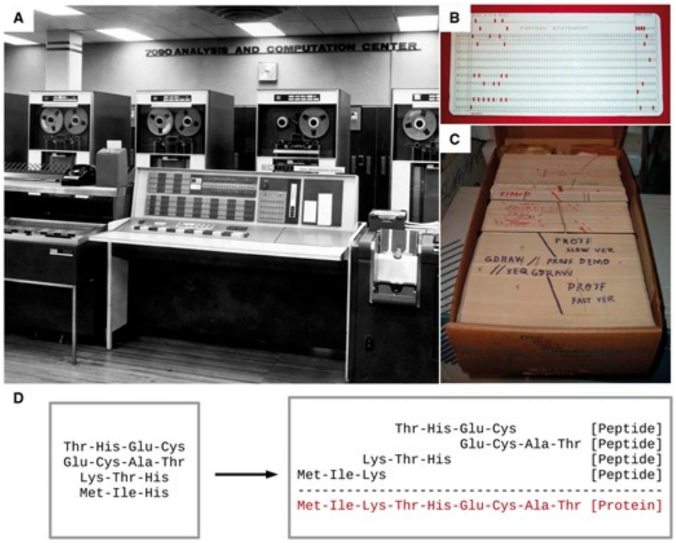

Fig. 1. Picture of IBM computer running FORTRAN code from a punch card for protein sequencing

After the release of Oakley's book on protein sequencing, *The Atlas Of Protein Sequence and Structure*, the infant bio-informatics community continued with improving on protein computation models.

In the 1970s the field started to focus more on DNA sequencing because of a new idea, now referenced as the ”Central Dogma” in molecular biology. The effect was that scientists focused on developing cheap and efficient methods for computing DNA sequencing as a means to protein sequencing. For example in 1970, the first dynamic programming algorithm for pairwise alignment was developed. The algorithm was developed for pairwise protein alignment because the focus of the field during the time was on protein sequences. [6]. From protein and DNA sequencing, the field started to research and develop solutions to more complex problems presented by its sister field, molecular biology, such as analyzing complete genomes. These methods of DNA and protein sequencing formed the foundation of bio-informatics, until the 2000s, where Bio-Informatic scientists were presented with a new issue pertaining to the management and structuring of big data being collected by Bio-Informatic research and computations. This is still an ongoing issue and is anticipated to remain as an issue as long as data remains integral to Bio-Informatic. Modern research and methods focus less on sequencing and more on modeling whole organisms and biological ecosystems.

3 Hidden Markov Models

3.1 Background on Hidden Markov Models

The discovery of Hidden Markov Models (HMMs) as an efficient tool for problems such as gene prediction, sequence classification, and alignment of sequences propelled Bio-Informatics. HMMs are considered a Markov process, similar to a Markov-Decision-Process, which are used in fields like artificial intelligence as means of making reasonable predictions. HMMs are found to be an intuitive mathematical representation of biological systems. [8]

HMMs in isolation don't perform Bio-Informatic computation, the model simply serves as the mathematical model used to represent and manage bio-informatic problems such as multiple sequence alignment, genome mapping, and genome prediction.

3.2 Model of a Hidden Markov Model

Hidden Markov Models assume that the observations used for observation probability come from underlying ”hidden” states, rather than the state of our model.

The model can be defined by 3 components for N states:

Set of state transition probabilities T: Probability P_ij for state S_i and state S_j

Set of observation emission probabilities O - Probability of seeing observation O_k while in state S_i

State initialization probability $P(\lambda|S_i)$: The probability of our HMM λ occurring in state S_i

Figure 1 presents an example of a HMM that relates the number of ice cream eaten with the weather of the day. In the example, B_1 and B_2 are the observation emission probabilities, the state transition probabilities are the connections between the hot and cold observations, and the state initialization probability is the octagon at the bottom. The state initialization probability in the example shows that more ice cream is eaten on hot days [4].

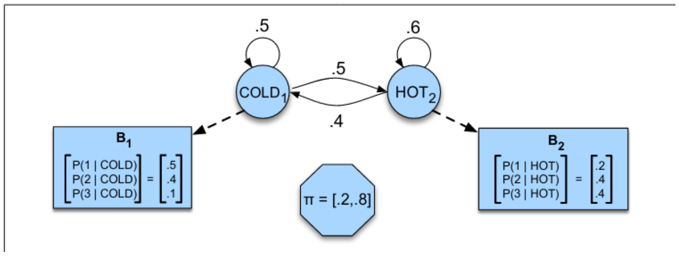

Fig. 2. Hidden Markov Model

3.3 HMM Algorithms

Within this field, there are numerous applicable algorithms that utilize HMMs, each with their own purpose and efficiency. To help conceptualize the application of these models to Bio-Informatics, it's important to understand the formal math behind HMM algorithms. More specifically, I wanted to introduce the evaluation problem for HMMs.

The evaluation problem for HMMs means we are trying to find the probability of a set of observations O given an HMM λ

$$P(O|\lambda) = \sum_S P(O|S;\lambda).(S|\lambda)$$

Now that we have defined the evaluation problem, we can present the "Forward Algorithm" for the evaluation problem. The algorithm begins its computation at the start of a sequence and finds the probability of a certain observation/genome sequence at each step.

Forward Algorithm for the state where we see those observations

$$\delta_K(u) = P(e_1, \ldots.., e_K; S_K = u|\lambda)$$

4 Dynamic Programming In Pairwise Alignment

A fundamental computer science concept for Bio-Informatics is dynamic programming. Most computer science students are initially introduced to this problem-solving technique in their algorithms classes (at UCSC this class is CSE 102). It's a fundamental tool, and is often evaluated during computer science interviews. This technique has found its way into numerous computer science subfields like artificial intelligence and Bio-Informatics.

For Bio-Informatics it's used as a way to efficiently compute pairwise alignment. Dynamic programming works by utilizing recursion and memoization to break down the problem into smaller chunks, until there are no more dependencies to other cells in the memoization table.

4.1 Needleman-Wunsh Algorithm

The Needleman-Wunsh algorithm is the basic algorithm for pairwise alignment using dynamic programming. The algorithm utilizes an NXN memoization table to store the currently misaligned sequence. Each cell in will represent a value for match, mismatch, or gap. The chain of dependencies start at the cell $i = 1, j = 1$, which is the value we want if we wish to know if the sequences are aligned, and the computation of a cell's value is recursively called the cell that requires its dependency. This continues until we reach $i = N, j = N$, where the cell no longer depends on another cell's value. The algorithm then backtracks back to $i = 1, j = 1$, filling the memoization table along the way and clearing dependencies.

4.2 Issue With Dynamic Programming

The main issue with dynamic programming for large data-sets is the bottleneck that comes with how interacting with large memory works on modern computers. To combat this Bio-Informatic scientists are designing frameworks that adjust the current use of dynamic programming to more efficiently interact with the memory on computer systems, specifically by introducing cache-oblivious algorithms for pairwise

alignment. [3]
Dynamic programming allows for efficient computation over large data-sets and models. The key idea of dynamic programming is that a state depends on the values of other adjacent or related states, so you start the algorithm at a state that has no dependencies and from there build the model of states until you have successfully computed every necessary state.
As a result of continuous research into algorithms, such as dynamic programming, for genome sequencing, the cost of whole genome sequencing has decreased past Moore's law, as illustrated by figure 3 [1].

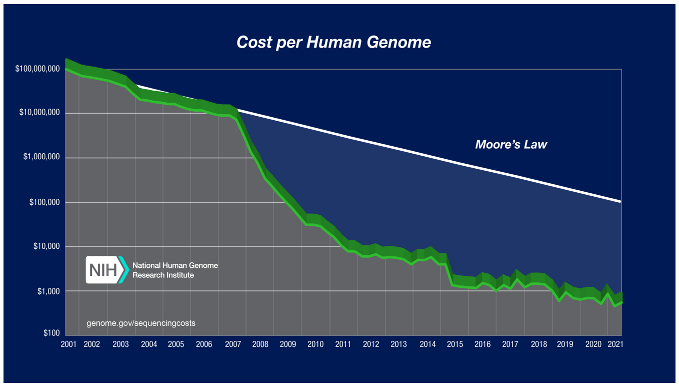

Fig. 3. Cost Human Genome

5 Genomic Epidemiology Applications and Research

Protein and DNA sequencing has brought us to modern genome sequencing, the basis for genomic epidemiology. Genomic epidemiology works by using whole genome sequencing to sorting whole genome sequences, by targeting specific sections of the sequence for similarity. For infectious diseases, those with similar sequences are cases originating from the same outbreak. This can be used to find the outbreak of new strains, as well as which the main strain's stake in new cases. Genomic epidemiology is also applicable to cancer prediction and prevention [9].

5.1 COVID-19

According to the World Health Organization as of May 29th 2022, the COVID-19 pandemic has affected 213 countries, over 526 million cases, and over six million deaths. As a result, there's been an effort to utilize genomic epidemiology for COVID-19 drug development, early detection, and prediction. With these new computation methods, there's hope to be able to more accurately predict the pandemic and mitigate vulnerabilities when they are observed.

5.2 Case Study: BeCaked

Deep learning and genomic epidemiology can be combined to more efficiently predict COVID statistics. The motivation for this is to notify the public of possible massive outbreaks in their area. This also helps researchers determine how new COVID-19 strains are spreading in counties.

The BeCaked project was created with the goal of creating an explainable AI that accurately forecasts COVID-19 statistics[10]. The training data that's fed into the machine learning model comes from John Hopkins University. The data set contains data for susceptible, infectious, recovered, and dead cases. It's important to include susceptible because if we are predicting the COVID-19 rate in a county with a higher average age, we need to associate a susceptible statistic because the model should behave differently for this county compared to a county with different demographics and age.
This formal representation of a model in the project was based on the data, the researchers call this the SIRD model (see Figure 4. There are edge cases to this model, as seen in figure , where more recovered cases affect your model's susceptible value.

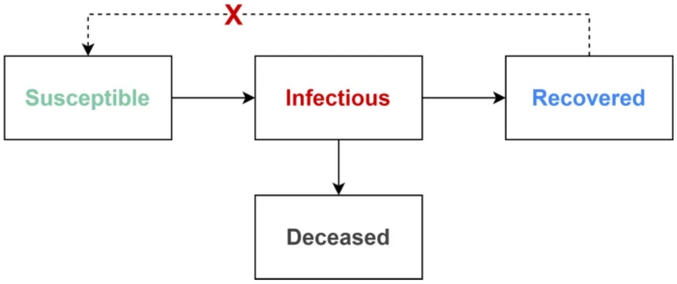

Fig. 4. SIRD Model

The deep learning model comes in the form of a $N * 4$ matrix, where N is the number of days we are training our model on. The j row of the matrix is length 4 because that is our model's data for each day i. The results of Be-Caked showed that it was accurate up to a 0.012 mean absolute percentage error at the global prediction level. The accuracy of the model was lower for modeling at the country level. This is a result of countries being open systems, as in they are affected by outside factors that can't be observed in a country's model. The global model is completely closed and isolated by external factors that affect COVID-19 statistics.

5.3 Importance Of This Research

I think it's important to establish importance on the fact that COVID prediction, biodiversity prediction, and other prediction-based Bio-Informatic problems can and should be tools used by legislators and those with power over these systems when they make decisions.

6 Applications of Bio-Informatics

Bio-Informatics is a very new and modern field, whose parent fields are either just as new, in the case of computer science, or constantly evolving new scientific methods and applications, in the case of biology. The uncertainty of this new field means there's no prediction on how applicable bio-informatics will be to our daily lives in 50 years. The field could reach a bottleneck or deadlock of what it's limits or it could maintain its growing rate and its applications could drastically change the way humans live. The current possible applications for bio-informatics are: genome sequencing, medical diagnosis, modeling of biological systems for the the purpose of analyzing bio-diversity and other aspects of a biological system, and genome data-collection for the purpose of analyzing variation and accuracy of models.

6.1 Medical Diagnosing

Medical diagnosing works very similarly to other prediction models for Bio-Informatics such as genomic epidemiology prediction. Using a person's whole genome sequence, we can search for genomes that are precursors or indicators of diseases. Along with diagnosing, this same concept is applicable to finding personalized cures and medicine based on a person's whole genome sequences. The figure below demonstrates a basic model for applying machine learning for the purpose of cancer prediction/susceptibility in a person's genome sequence.

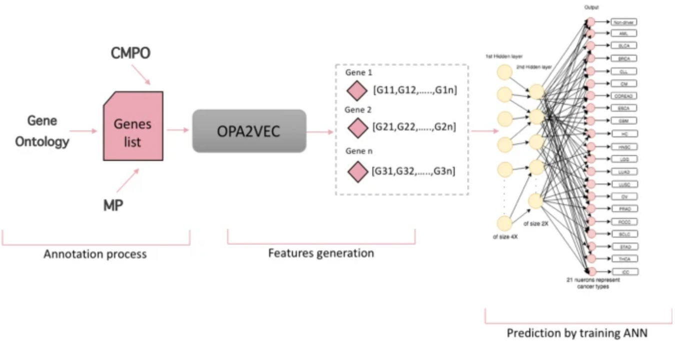

Fig. 5. Workflow Model For Cancer Prediction

In this specific model, the gene list is fed into OPA2VEC as a way to extract features from the genomic data[2]. A large portion of the application of machine learning for the purpose of cancer diagnosing comes from the ability to extract useful features that can be replicated across different genome data from multiple patients with the same form of cancer or similar genetic encoding. These features are necessary in order to train, analyze, and fine tune the cancer prediction model for accuracy.

6.2 Drug Discovery And Testing

Bio-Informatics for drug discovery works by using machine learning to create accurate alignments of phenotypic gene expressions. This process significantly reduces the amount of possible chemical combinations that can be used to create drugs that then go into clinical trials.

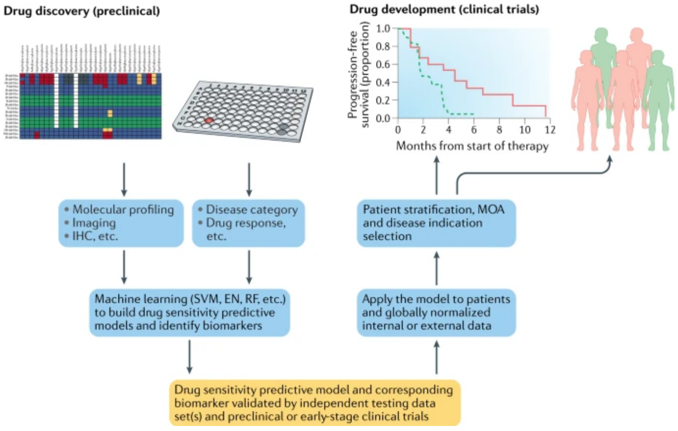
Drug discovery (preclinical)
Drug development (clinical trials)
Progression-free survival (proportion)
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
Months from start of therapy
• Molecular profiling
• Imaging
• IHC, etc.
• Disease category
• Drug response, etc.
Patient stratification, MOA and disease indication selection
Machine learning (SVM, EN, RF, etc.) to build drug sensitivity predictive models and identify biomarkers
Apply the model to patients and globally normalized internal or external data
Drug sensitivity predictive model and corresponding biomarker validated by independent testing data set(s) and preclinical or early-stage clinical trials

Fig. 6. Machine Learning For Drug Discovery

The above figure shows the workflow of drug discovery. It's important to not that drug discovery doesn't do all the computation needed to find useful and side-effect free drugs.

7 Privacy Concerns

As this technology improves, it's important to be conscious of the possible exploitative avenues this science can bring. Bio-Informatics is in the same case as other new technologies, like image recognition; this technology can be invasive and instinctually serve as data-collection mines. As illustrated by figure 6, the ownership of genome data can be transferred or owned by multiple entities, such as a genome research center and local authorities.

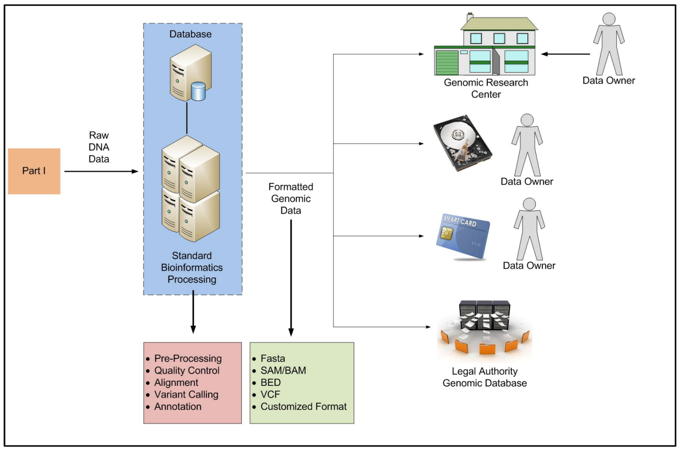

Fig. 7. Flow Of Bio-Informatic Data

The biggest concern for privacy comes with governments keeping collections of personal genome data, for the purpose of policing and surveillance. For example in 2018, genome data used in a genetic ancestry test was given to the police. This data was used to convict Joseph James DeAngelo of crimes committed by a serial killer of the name ”Golden State Killer” that committed at least 13 murders between 1974 and 1986.

7.1 Solutions

Solutions to this issue mostly come in the form of creating and introducing more security-enhancing technology to data collections at the infrastructure level. This solution is not enough to ensure privacy of your biological data, which is why there's been a growing number of legislation of data-privacy. As this field continues to grow and as genome data collections expand their data set, there's going to be a need for standardized human rights laws to protect data-privacy across countries.

8 Biodiversity Informatics

Biological systems in nature are rich in data, which means the data is impossible to process by just humans. This is where Bio-Informatics is stepping in. With the help of computers, biologists are able to see the true picture of their data and analyze the health of biological systems.

The introduction of these new analytical tools help, not only biologists, but those with direct authority over conservation and wildlife. The ability to have open and clear indicators of an ecosystem's health and condition in specific areas like: taxonomy, invasive species, and effects of the global environment system on the ecosystem.

8.1 Ecological Niche Modeling

Ecological Niche Modeling (ENM) is a model used to predict the distribution of species over a geographical area. This is important when it comes to surveying the health of an area because it helps research pin down on vulnerable areas for species or on areas that have a newly introduced invasive species that is disrupting the natural organization of the ecosystem. This following figure shows the different features that come into play when analyzing biodiversity and the status of a regions biodiversity health.

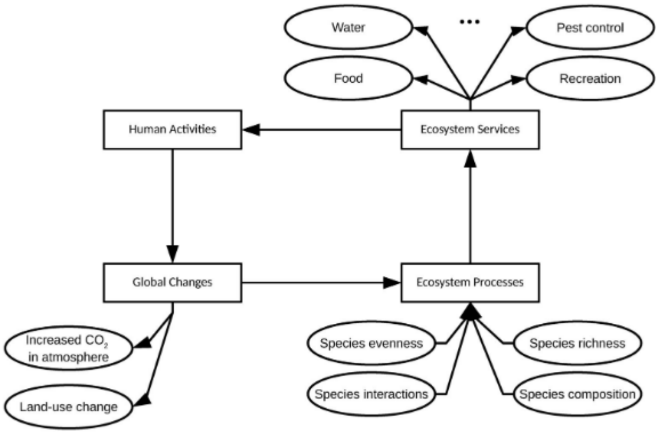

Fig. 8. Ecosystem connection to outside features and conditions

It's important to keep these external features in mind when creating layers and models for ENM because neglecting the fact that an ecosystem is not a closed system will introduce chaos into the model and ruin predictions for a species' distribution and health within the ecosystem.

The main factors that affect a species probability of existence in a region are abiotic conditions, biotic conditions, and dispersal capacity. Abiotic conditions are environmental conditions like rainfall measurements and temperature. Biotic conditions are interactions between the species and other biology in the environment, e.g. the ratio between our target species and one of its predator species. Bio-Informatic researchers currently focus more on abiotic conditions because they have produced the best accuracy for ENMs. [5] An example of layers that could go into an ENM model is illustrated in figure 8 [7]. To make the ENM in the figure more complex, a designer could introduce a new biotic layer for regions with trees taller than a certain height, which is a suitable habitat for a predator bird species to our subject species.

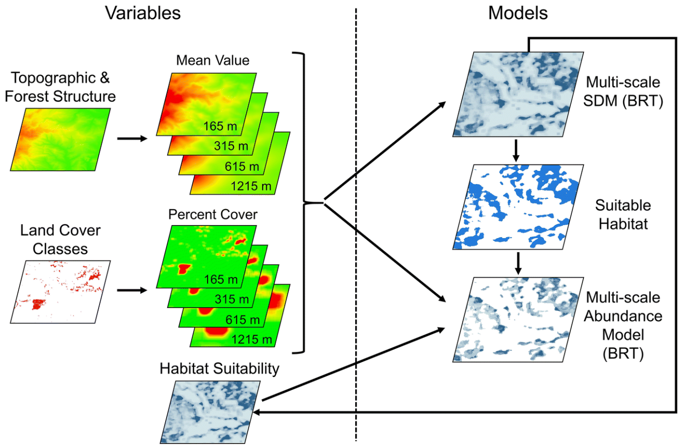

Fig. 9. Example of Layers in an ENM

This process of using ENMs for biodiversity informatics is split into three steps: preprocessing, modeling, and postprocessing. The preprocessing stage deals with selecting which data is important for the ENM, retrieving said data, and then filtering the data for noise. The modeling stage is where computational models are applied to the previously curated data. There are many different computational methods that can be applied during the modeling phase such as machine learning algorithms like Maxent. The post-processing phase consists of retrieving the actual model projection from our machine learning and evaluating such projection.

The more complex and biodiverse our subject environment is, the more difficult it is to create an accurate model of the system for use of ENMs.

9 Conclusion

Bio-Informatics is a rich scientific field that continues to progress and bloom. Computer science has proven to be invaluable for processing and modeling biological data, a shortcoming of biology before bio-informatics, for the purposes of biodiversity, epidemiological research, medical diagnosing, and cancer prediction. As we move into the future, we will see parallel progression between computer science, Bio-Informatics, as well as fields like biomolecular engineering.

ACKNOWLEDGEMENTS

I would like to thank Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison for their work on the textbook Biological sequence analysis [1]. The book is what initially sparked my interest in bio-informatics. I would also like to thank overleaf for their LaTeX software, which was used to edit and compile this document.

REFERENCES

[1] Dna sequencing costs: Data, 2021.

[2] Sara Althubaiti, Andreas Karwath, Ashraf Dallol, Adeeb Noor, Shadi Salem Alkhayyat, Rolina Alwassia, Katsuhiko Mineta, Takashi Gojobori, Andrew D. Beggs, Paul N. Schofield, Georgios V. Gkoutos, and Robert Hoehndorf. Ontology-based prediction of cancer driver genes. *Scientific Reports*, 9(1), November 2019.

[3] Rezaul Alan Chowdhury, Hai-Son Le, and Vijaya Ramachandran. Cache-oblivious dynamic programming for bioinformatics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):495–510, 2010.

[4] James H. Martin Daniel Jurafsky. *Hidden Markov Models*. December 29, 2021.

[5] Luiz M R Gadelha, Pedro C Siracusa, Eduardo Couto Dalcin, Luís Alexandre Estevão Silva, Douglas A Augusto, Eduardo Krempser, Helen Michelle Affe, Raquel Lopes Costa, Maria Luiza Mondelli, Pedro Milet Meirelles, Fabiano Thompson, Marcia Chame, Artur Ziviani, and Marinez Ferreira Siqueira. A survey of biodiversity informatics: Concepts, practices, and challenges. *WIREs Data Mining and Knowledge Discovery*, 11(1), 2021-01.

[6] Jeff Gauthier, Antony T Vincent, Steve J Charette, and Nicolas Derome. A brief history of bioinformatics. *Briefings in Bioinformatics*, 20(6):1981–1996, August 2018.

[7] Tyler A. Hallman and W. Douglas Robinson. Comparing multi- and single-scale species distribution and abundance models built with the boosted regression tree algorithm. *Landscape Ecology*, 35(5):1161–1174, 2020.

[8] Shary Jose, Prashant Nair, Vinai George Biju, Blessy Baby Mathew, and C. M. Prashanth. Hidden markov model: Application towards genomic analysis. In *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pages 1–7, 2016.

[9] Shiro Kitano, Jamie Myers, Junko Nakamura, Akio Yamane, Mami Yamashita, Masato Nakayama, Yusuke Tsukahara, Hiroshi Ushida, Wanqing Liu, Mark J. Ratain, and Masahiko Amano. A novel fully automated molecular diagnostic system (AMDS) for colorectal cancer mutation detection. *PLoS ONE*, 8(5):e62989, May 2013.

[10] Duc Q. Nguyen, Nghia Q. Vo, Thinh T. Nguyen, Khuong Nguyen-An, Quang H. Nguyen, Dang N. Tran, and Tho T. Quan. BeCaked: An explainable artificial intelligence model for COVID-19 forecasting. *Scientific Reports*, 12(1), May 2022.

Mario Panuco I'm a Junior computer science student at the University of California, Santa Cruz. My favorite fields within computer science are assistive technology, machine learning, bio-informatics.