Project ID: 201

SR - Computational Biology and Bioinformatics

Krish Pai
Grade 10
Canyon Crest Academy
Advisor: Ed Gerstin

Deep Learning Diagnosis and Reconstruction of Gastrointestinal Disorders from Capsule Endoscopy Videos

Gastrointestinal (GI) disorders are the diseases, and cancers relating to the GI tract (digestive tract), prevalent disorders such as colorectal cancer have rising mortality rates if not detected early. Traditional clinical diagnosis of GI disorders are done through screening procedures to locate biomarkers of disorders in the GI tract, while effective, patients are resistant to some procedures while other procedures are incomprehensive. This study provides a solution using deep learning to enable the novel capsule endoscopy medical device. The capsule endoscopy is a non-invasive screening procedure that uses a swallowable camera to visualize the GI tract, however it is a time-intensive and error prone process for radiologists to analyze over 50,000 frames the capsule endoscopy records. Deep neural networks that were optimized in image classification were used through transfer learning and data augmentation on a dataset of over 47,000 capsule endoscopy recorded lesions, from 14 different categories of biomarkers, to produce a model with 95% validation accuracy in diagnosis. Then deep learning was applied to the entire capsule endoscopy video for a real time diagnosis, and through techniques like class activation mapping the region of each video frame showing a biomarker was highlighted. Furthermore, using monocular depth estimation, biomarkers were segmented from the capsule endoscopy video and reconstructed into 3D models for proper size estimation before any biopsies are conducted. Overall this novel project automates the GI disorder diagnosis process on par with radiologists, can be used on a wide scale to identify new biomarkers of GI disorders and can be extrapolated to other medical imaging and video data.

Project ID: 202

SR - Computational Biology and Bioinformatics

Saanvi Dogra
Grade 9
Del Norte High School
Advisor: Sandeep Dhanda

Investigating the Underlying Molecular Patterns in Aggressive Brain Tumor ETMR

Embryonal tumors with multilayered rosettes (ETMR) is a rare and aggressive brain tumor primarily impacting infants 2-4 years old. This research studies the specific probes that result in gene amplifications, potentially causing ETMR. It was hypothesized that the methylation array analysis of copy-number variations in ETMR will reveal certain probes that cause gene amplification at C19MC in chromosome 19. Methylation arrays of ETMR samples wer gathered from the Capper Dataset and Illumina and compared to control samples using Conumee and Minfi in R. The significantly amplified probes in each sample were compared, and, from 480 thousand probes divided into 16 thousand bins, 1 detail region specified 149 overlapping probes. The range of the probes in chromosome 19 stretches from a starting point at 54168850 to an ending point at 54265500. The ETMR samples studied consistently demonstrated an amplification at C19MC and produced a gain ratio of 0.407, demonstrating a positive, large increase in number of those probes. This research makes important implications about the potential causes of ETMR and possible treatment options that result from targeting the amplified probes. In the future, more experiments can be conducted with more data to replicate the results. This amplification can also be used for the early diagnosis of ETMR, which is especially important for an increased likelihood of survival and ETMR's history of being frequently misdiagnosed.

Project ID: 203

SR - Computational Biology and Bioinformatics

Maanaskumar Gantla
Grade 12
Canyon Crest Academy
Advisor: Alex Siegel

Screening of FDA Approved Drugs for Potential Alzheimer's Therapeutics Using Machine Learning

Multiple factors contribute to the development of Alzheimer's disease making it difficult to treat, and there is no effective treatment available. Traditional drug discovery is a very complex and expensive process that takes several years to bring drugs to clinic. This project addresses these shortcomings, by developing machine learning (ML) models to screen FDA approved drugs, targeting key kinase CDK5 in Tauopathy pathway, to inhibit the formation of neurofibrillary tangles (NFT). I hypothesized that if computational ML models can screen existing FDA-approved drugs that can target Cyclin-dependent Kinase (CDK5) in Tauopathy in Alzheimer's disease, then effective therapeutics for Alzheimer's can be identified quickly.

I developed ML models for CDK5 using known active inhibitors, obtained from PubChem, and converted to the SMILES format. Afterwards, molecular descriptor data is processed to develop ML models using the internally built algorithms of WEKA software. For model development, a control set of random compounds from PubChem are also used in addition to known active inhibitors. Then, FDA-approved drugs were screened using this model to find novel therapeutics for Alzheimer's. ML models developed exhibited accuracies ranging from 95%-99.5%, with AUROC ranging from 0.97-1.00, considered as excellent predictive performance of the models. From the model prediction results, I identified 43 FDA-approved drugs that are active with prediction scores greater than 0.7, whose efficacies were further confirmed by protein-ligand docking experiments. This project not only provides drug candidates that could treat Alzheimer's, but also demonstrates the capability of ML models to find novel uses for existing drugs.

Project ID: 204

SR - Computational Biology and Bioinformatics

Daniel Lu
Grade 11
Westview High School
Advisor: Scott Halander

Early Prediction of Diabetes Using Machine Learning Techniques on Multiple Datasets

The purposes of this project are to develop an advanced machine learning model that successfully predicts whether a patient is diabetic based on hospital data and to develop a classification framework to predict the patient's stage of diabetes mellitus, since current prediction methods are not ideal. Three online diabetes datasets (PIDD, UCIMLR, LMCH) were obtained and preprocessed. Five machine learning models (Random Forest, Logistic Regression, Decision Tree, XGBoost, and Ensemble Soft Voting Classifiers) were developed on each dataset. A random forest regression model was developed on the LMCH dataset in order to classify patients as nondiabetic, prediabetic, or diabetic based on their laboratory results. The results showed a new finding that the XGBoost classifier was the most successful and achieved an accuracy of 99.1% and a recall of 99.4% on the LMCH dataset. The most significant predictors of diabetes were determined to be hemoglobin A1C concentration, age, and body mass index (BMI). Additionally, a novel random forest regressor was created, and it accurately predicted the stage of diabetes 97.85% of the time, overestimated the severity of diabetes 1.37% of the time, and underestimated the severity of diabetes 0.78% of the time. This work was not reported in previous research. First, when provided with the necessary laboratory data, this random forest categorization model may play a significant role in health clinics to accurately diagnose the onset of diabetes before symptoms appear and worsen. Second, recommendations were proposed for the development of future diabetes datasets through feature analysis.

Project ID: 205

SR - Computational Biology and Bioinformatics

David Samy
Grade 11
Canyon Crest Academy
Advisor: Kevin Hare

Accelerating Cancer Drug Discovery through an Algorithmic Approach to Identifying Leads Against Target Protein Kinases

The goal of the project was to identify FDA-approved drugs that could be used to jumpstart drug discovery efforts on a different protein kinase, than for which it had been approved. The protein data bank was mined to generate a library of proteinligand complexes of FDA-approved drugs against protein kinases. Using a three dimensional structure of the target protein kinase as a query, the library was searched to identify the protein kinases with FDA-approved drugs that had the most similar binding pocket to our target of interest. Following structure superposition, the approved drug was placed in the structure of our target protein and analyzed to identify favorable interactions and potential modifications that could jumpstart a drug discovery effort. Since proteins are flexible molecules, multiple structures of the target protein in different conformations were used in the search query. The results of this project are a series of potential ligands that need to be tested against the target kinase to initiate the drug discovery effort.

Project ID: 206

SR - Computational Biology and Bioinformatics

Chloe Wang
Grade 10
Canyon Crest Academy
Advisor: Ed Gerstin

How Long Will It Take to Exhaust All Possible Mutations of SARS-CoV-2?

Background:
SARS-CoV-2 is the pathogen causing coronavirus disease 2019 (COVID-19). When the virus was first discovered in December 2019, it did not respond well to standard treatments and it was deadly. Starting in 2020, it began spreading rapidly around the world, infecting more than 750,000,000 people and killing almost 7 million people. During this period of time, multiple new variants emerged under evolutionary pressure from new vaccines and treatments. This project will study if it is possible to estimate how long it will take to go through all probable mutations of SARS-CoV-2 through a mathematical model, and if so, how?

Procedure:
Research and find average number of cell infection events in a patient per contraction of the COVID-19 virus, the average number of patients in a year, the mutation rate of RNA viruses, the number of nucleotides the virus has. Using these numbers, calculate the time it would take to exhaust all possible mutations.

Results:
At this stage, from the research and preliminary calculations, it was found that the time needed to go through all mutations of SARS-CoV-2 is 3.73 x 102292 years.

Discussion & Conclusion:
Although this estimation is done with a very simplified model and has a number of caveats, it is probably true that humans have to live with a continuously changing virus for a very long time unless an effective medicine outspeeding the virus evolution rate can be invented.

Project ID: 207

SR - Computational Biology and Bioinformatics

Andrew Yu
Grade 11
Canyon Crest Academy
Advisor: Ed Gerstin

Investigation of Potential Gene Pathways Associated with Resistance to Targeted Therapy in FLT3-ITD Acute Myeloid Leukemia

FLT3 is the most frequently mutated gene in AML patients and FLT3-ITD mutation accounts for poor prognosis, reduced survival and increased relapse. Drug resistance after FLT3 inhibitor therapy remains a significant challenge to the outcome of the patients. The goal of this research is to find potential pathways associated with the therapeutic resistance that could help the treatment of this disease. The hypothesis is that genes and pathways that play a role in cell regulation and growth will be affected by the FLT3 inhibitors. Bioinformatic analysis on differentially expressed genes (DEGs) on patient samples before and after the treatment of FLT3 inhibitors can generate meaningful gene pathways that contribute to the resistance of targeted therapy.

Two datasets GSE61804 and GSE74666 were selected from NCBI GEO database and analyzed with GEO2R in order to identify DEGs. The cut-off criteria were Padj<0.05. Data from GEO2R analysis were downloaded in SOFT format and imported into Microsoft Office Excel 2016 for screening for common DEGs across datasets. 67 common DEGs were identified across the datasets. GO and KEGG analysis together with literature review identified 4 DEGs that might be associated with the resistance to Sorafenib treatment in FLT3-ITD+ AML.

Conclusion: Resistance to FLT3 inhibitors involves multiple mechanisms. Four genes having specific expression profiles in FLT3-ITD+ AML patients and drug resistant cells were identified through bioinformatic analysis on DEGs . Combinatory treatments of CDH1/ABCB6/CCNK/HDGFRP3 pathways with the FLT-3 inhibitors would improve patients' outcome for this devastating disease.

Project ID: 208

SR - Computational Biology and Bioinformatics

Bhadra Rupesh
Grade 11
The Bishop's School
Advisor: Lani Keller

Determining Genetic Biomarkers as Predictors of Response to Anti-PD 1 Cancer Immunotherapy

Immunotherapy, and in particular anti-PD1 therapy, has been a revolution in the approach to treating cancer, particularly for non-small cell lung cancer. However, not all patients respond to this treatment, so effective stratification of patient pre-treatment would improve the clinical management of the disease. Being able to identify patients that will respond to a therapy early greatly improves the opportunity for successful therapeutic outcomes.

To investigate whether genetic factors can predict if anti-PD-1 therapy would be effective, I wrote a program in R using the Bioconductor library's DESeq2 package to restructure and statistically analyze data on gene expression in several human cancer patient tumor transcriptomes and correlate it to their response to anti-PD-1 therapy.

Based on the statistical significance threshold of p-value < 0.05, the results showed that nine genes (NOS2, MAGEA4, KREMEN1, S100A9, S100A8, NOTCH3, CD44, TFRC, VTCN1) were expressed the highest in patients who did not respond to anti-PD-1 therapy (with log fold change > 2), while five other genes (MS4A1, CD22, CD3G, TNFRSF9, CD37) were expressed the highest in patients who did respond (with log fold change < 0.5).

With normalized Transcript Per Million values greater than 0 showing upregulation, genes that were the best predictors were S100A9 (upregulated in seven of twelve non-responding patients) and TNFRSF9 (upregulated in seven of eight in responding patients). Based on these results, immune-related genes S100A9 & TNFRSF9 were found to act as good predictors of anti-PD 1 response and could be used to stratify patients for treatment with anti-PD1 therapy.

Project ID: 209

SR - Computational Biology and Bioinformatics

Alex Kumar

Vishnu Aravind

Grade 11

Del Norte High School

Advisor: Sanjay Kumar

Using Machine Learning to Repurpose FDA Drugs for the Treatment of Diabetic Cardiomyopathy

PARP-1 (Poly ADP Ribose polymerase) functions to repair damage to DNA and is implicated in a variety of malignant and inflammatory diseases including Diabetic Cardiomyopathy. The development of inhibitors for this protein, however, is a tedious process, involving a great deal of time and money, resulting in a very low number of inhibitors available for PARP-1. With this project, we developed a multilayer perceptron model that was able to distinguish between inhibitors and non-inhibitors of PARP-1 based on their chemical structure. We collected the most effective confirmed inhibitors of PARP-1 from PubChem and clustered them using RDKit to get groups of compounds with similar structure. The largest cluster was then isolated and molecular descriptors were calculated. Attribute selection was then performed with a feature-selection algorithm from WEKA. This data was used to develop the multilayer perceptron model, whose hyperparameters were tuned, enabling the model to predict inhibitors of PARP-1 with an accuracy of 90%, an MCC of 0.78, and an AUROC of 0.93. The model was then run on all FDA drugs, and the top 50 predictions were taken. After performing molecular docking on these predicted inhibitors, and a set of random FDA drugs, the predicted inhibitors had a significantly better binding affinity for PARP-1 than the control group. Our findings not only uncover potential treatment options for Diabetic Cardiomyopathy caused by PARP-1 but also showcases the use of machine-learning techniques to predict new uses for existing pharmaceuticals.

Project ID: 210

SR - Computational Biology and Bioinformatics

Shivani Ramesh
Grade 11
Scripps Ranch High School
Advisor: Patty Fowler

Predicting Right and Left Hemisphere Damage in Aphasia Patients Using NLP and Machine Learning

Purpose: Develop a Natural Language Processing (NLP) and Machine Learning (ML) based assessment tool that can predict the hemisphere (left or right) of brain damage or aphasia type based on speech patterns. This can significantly aid Speech-Language Pathologists(SLPs) in their patient evaluation and treatment, resulting in a more convenient and efficient therapy process for aphasia patients.

Procedure: This experiment utilizes recordings from aphasia patients and analyzes them using NLP and ML techniques. They were classified into Left Hemisphere Damage (LHD) and Right Hemisphere Damage (RHD). Data set was cleaned, annotated and only patients' speech was used. NLP techniques was applied for parts of speech (POS) tagging and other feature like gestures and pauses were extracted. Statistical analysis method of "Two samples assuming unequal variance" (T-Test) was used to find distinguishable features with P value of less than 5%.

Results: The study saw the separation between LHD/RHD and Broca/non-Broca for specific features, such as the ratio of "uhhh" and verbs used. These distinctive features were then used in a K-Nearest Neighbor(KNN) ML model. The study found the optimal value of K(5) for LHD/RHD based on â€œPauseâ€ and K(5) for Broca/Non-Broca based on verb count, which resulted in an optimal model.

Conclusion: The study revealed that those with LHD experienced difficulties in finding words leading to pauses while speaking. Broca patients had difficulty to find action words compared to non-Broca patients. Using these features, ML model was can distinguish between LHD/RHD and Broca/non-Broca. Developing an easily accessible automated tool using these could ease and enhance the ongoing assessment process and therapy by SLPs, allowing for a more transparent evaluation and leading to faster recovery in a home setting.

Project ID: 211

SR - Computational Biology and Bioinformatics

James Sun
Grade 10
Canyon Crest Academy
Advisor: Ed Gerstin

Type 2 Diabetes Prediction and Risk Factors Using Machine Learning Models

Type 2 Diabetes is a chronic disease where a person's cells stop reacting to insulin, or their pancreas stops producing insulin. It affects over 10% of the adult population in the US, and can come with severe complications like risk of heart disease, nerve damage, and even sleep apnea. Early diagnosis of diabetes is crucial to stopping these complications from happening, which is why it is important to develop ways to detect diabetes without a blood test, as it can be easily applied to the general population with low cost and high efficiency. This study aims to create several machine learning models that can detect if a person has diabetes based on general physical and demographical data. Classification models like Logistic Regression, Support Vector Machine, Decision Tree, etc. are trained and tested on batches of data taken from the 2021 BRFSS dataset. Their accuracy, precision, f1 score, precision-recall curve, and receiver operating characteristic curve are then recorded and analyzed on how well they perform. Overall, models were shown to have an average accuracy around 79%, and performed well as early indicators of diabetes.

Project ID: 212

SR - Computational Biology and Bioinformatics

Miles Keiser
Grade 10
Del Lago Academy
Advisor: Igor Tsigelny

Predicting Radiosensitivity of Tumors Using Genetic Biomarkers and Machine Learning

Radiation therapy is a common and cost effective cancer treatment. Radiation therapy is carried out via Ionizing Radiation (IR), which slows growth or even destroys the tumor. However, different cancers can have very different responses to radiation therapy. Some are very sensitive, while others require much larger doses of radiation. Not knowing how radiosensitive a tumor will be severely reduces the precision and personalization of a patientâ€™s radiation therapy. We believed that radiosensitivity could be predicted using mutation and pathway expression data.

We studied the CCLE (Cancer Cell Line Encyclopedia) Database which contains information on mutations in a large number of cell lines. Each gene was then correlated to a specific pathway that it affects. This data was then combined into one file that contained the mutated genes and pathways for 133 cell lines. Finally, we performed attribute selection in WEKA to filter out the most important genes.

We created several models, the most successful of which were Logistic, SMO, BayesNet, HoeffdingTree, and NaiveBayes. The highest accuracy achieved was 94.7% using the BayesNet architecture.

In conclusion, using machine learning and gene mutation data, it is possible to predict the radiosensitivity of tumors accurately. We were able to develop a model that could successfully elucidate the radiosensitivity of cancer cell lines based on mutations with a high degree of accuracy. With more development, this method could be used to greatly increase the degree of personalization in cancer treatment and radiation therapy.

Project ID: 213

SR - Computational Biology and Bioinformatics

Trevor Chen
Grade 10
Westview High School
Advisor: Scott Halander

Use of Deep-Learning to Find Inhibitors for Aldose Reductase as Treatment of Diabetic Cataracts

Diabetes is a disaster in the modern world, and it leads to many complications. One of the most frequently occurring complications is cataract: millions of people estimated to have a cataract recently. High glucose condition activates the polyol pathway and enzyme aldose reductase (AR), leading to reactive oxygen stress and accumulation of sorbitol in the lens, which attracts water, making disruption of fibrils and cataract formation. So, the inhibition of AR is the main point to treat diabetic cataract. To explore the use of machine learning in finding possible compounds to inhibit AR, blocking the reduction process of glucose to sorbitol in the lens, possibly mitigating the onset of diabetic cataracts. Known AR inhibitors have failed to be used for patients, but new predicted AR inhibitors have the potential to possess more favorable characteristics to be successfully implemented after clinical testing. Exploring new inhibitors can improve patient well-being and lower surgical complications while decreasing long-term medical expenses. Elucidate a set of molecular descriptors related to known inhibitors of AR based on their activity in inhibiting AR. In turn, using the important descriptors to train machine-learning classifiers to sort inhibitors of AR based on their binding affinity. The trained machine-learning classifiers predict the FDA-approved drugs binding affinities to AR. The compounds with the highest predicted scores are docked upon AR, in order to see how the machine-learning prediction correlates with the possible binding affinity. The compounds, which were predicted using machine learning, when docked on AR had a higher binding affinity compared to random compounds. The best compounds had affinities of -9.7 and -8.7, while the prediction and binding affinity closely followed one another. Machine learning is a viable method to find possible inhibitors for AR. With future testing, the compounds could become medically viable. The machine-learning algorithms can also be applied to find other inhibitors for proteins, which follow similar pathways such as AR. Using known AR inhibitors and PaDel chemical descriptors, we prepared a machine-learning model for prediction of new AR inhibitors. The cross-validation of the model demonstrated more than 90% accuracy of prediction of new compounds.

Project ID: 214

SR - Computational Biology and Bioinformatics

Akshaj Bharadwaj
Grade 11
Canyon Crest Academy
Advisor: Ed Gerstin

Predicting Gene Expression In Cancer Tissues Using Machine Learning on Histopathological Images

Objective: The objective of this study is to examine the potential for utilizing machine learning algorithms in determining the precise gene expression levels within cancerous tissu specimens. This aims to provide a more comprehensive understanding of the gene expression dynamic present in these tissues and contribute to the advancement of computational biology.
Purpose: The purpose of this project is to employ high-resolution tissue images and corresponding gene expression data to train and evaluate a machine learning model that will predict the expression levels of targeted genes. The predicted expression levels will then be plotted on a boolean analysis chart to analyze the relationship between the levels.
Procedure: The procedure involved the organization and processing of the tissue images and gene expression data to be used as input for the machine learning model. The model was trained using a subset of the data, and its accuracy was evaluated using the remaining data. The predicted expression levels were then plotted on a boolean analysis chart to visualize the relationship between the levels.
Results: The machine learning model produced highly accurate predictions of gene expression levels, with an average absolute difference of 0.05 between the predicted and actual expression levels. The predicted gene expression values were very close to the actual values, as evidenced by the low mean squared error (MSE) and high correlation coefficient (R) values. Specifically, the MSE was 0.03 and the R value was 0.97, indicating that the model had a strong predictive performance. The boolean analysis chart revealed that the majority of the gene pairs had either a high-low or high-high relationship, with few instances of low-low relationships. This suggests that there is a strong positive correlation between many of the genes, which could have implications for further research in computational biology.
Conclusion: This study demonstrates the potential of machine learning algorithms in accurately predicting gene expression levels in cancerous tissue specimens. The close alignment between predicted and actual values indicates that the model was effective in learning to predict gene expression levels. The boolean analysis chart provides insights into the relationship between gene pairs, which can have implications for understanding the biology of cancer. The study highlights the importance of using machine learning algorithms to analyze large amounts of gene expression data, which can lead to new discoveries and advancements in computational biology. In conclusion, this study provides a framework for future research in this area and contributes to a better understanding of gene expression dynamics in cancer.

Project ID: 216

SR - Computational Biology and Bioinformatics

Grace Wang
Grade 11
Del Norte High School
Advisor: John Mortensen

Generating a Database of Experimental Data for Evaluating Algorithms for Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) is a computational method that identifies differentially expressed sets of genes between two phenotypes and is used for bioinformatics research. Given a set of genes, GSEA can determine whether there is statistically significant difference between the two biological states. Currently, there are continuous efforts to modify GSEA algorithms and create a more robust technique for GSEA analysis. However, no database that stores GSEA data currently exists. To support continuous efforts to enhance and accelerate GSEA, a gold standard benchmarking data set collection is required to be used to compare algorithmic changes. Using hundreds of real-world datasets, my goal is to produce a repository of GSEA results that can be used to design and test enrichment algorithms. Users can query Gene Expression Omnibus databases and find statistical scores of the GSEA analysis to compare to new developments. This database will be made available to the public to aid in the widespread development of reliable gene set enrichment techniques.

Project ID: 217

SR - Computational Biology and Bioinformatics

Kevin Ma
Grade 12
Canyon Crest Academy
Advisor: Ed Gerstin

Variant Annotation Through a Highly-Efficient Bioinformatics Tool

Currently, there are many publicly available Next Generation Sequencing tools developed for variant annotation and classification. However, as modern sequencing technology produces more and more sequencing data, a more efficient analysis program is desired, especially for variant analysis. In this study, I updated SNPAAMapper, a variant annotation pipeline by converting perl codes to python for generating annotation output with an improved computational efficiency and updated information for broader applicability. The new pipeline written in Python can classify variants by region (Coding Sequence, Untranslated Regions, upstream, downstream, intron), predict amino acid change type (missense, nonsense, etc.), and prioritize mutation effects (e.g., synonymous > non-synonymous) while being faster and more efficient. The new pipeline works in five steps. First, exon annotation files are generated. Next, the exon annotation files are processed, and gene mapping and feature information files are produced. Afterward, the python scrips classify the variants based on genomic regions and predict the amino acid change category. Lastly, another python script prioritizes and ranks the mutation effects of variants to output the result file. The Python version of SNPAAMapper accomplished the overall speed by running most annotation steps in a substantially shorter time. SNPAAMapper-Python was developed and tested on the ClinVar database, a NCBI database of information on genomic variation and its relationship to human health. I believe the developed Python version of SNPAAMapper variant annotation pipeline will benefit the community by elucidating the variant consequence and speed up the discovery of causative genetic variants through whole genome/exome sequencing.

Project ID: 218

SR - Computational Biology and Bioinformatics

Eshika Pallapotu
Nithika Vivek
Grade 9
Del Norte High School
Advisor: Juli Cheskaty

What Factors are Most Influential in the Spike in Influenza Levels

This project analyzed the most influential factors affecting influenza cases in the United States. It was presumed beforehand that the most positively correlated factor affecting influenza is population density, and the most negatively correlated factor was age. The influenza data utilized was gathered by CBSAs, in the time span of 2019 through 2023 from CDC.gov. Using Python, we first standardized the data and merged all the factors with the influenza data, and then used various regression models to identify the highest correlation coefficient. The regression models we experimented with were Linear Regression, Lasso, and RandomForest. We calculated the error rate with these models, and used the model with the maximum R squared value and lowest Mean Absolute Error. This turned out to be Linear Regression, with an error rate of 1.87 and R2 value of 0.14. This model was used to identify the most important factors (with both a negative and positive correlation) influencing influenza. It was found that areas with people who travel more by flight have lower flu activity (correlation coefficient of 0.98), while weeks with higher temperatures have lower flu activity (correlation coefficient of -0.74). Our study and findings can be applied to general public health in the future. By identifying the most important factors affecting influenza, we can try to be more prepared for an increase in influenza cases and better inform the population on how to take care of themselves.

Project ID: 219

SR - Computational Biology and Bioinformatics

Vivian Nguyen
Giselle Geering
Grade 12
Bonita Vista High School
Advisor: Michelle Mardahl

Identifying Novel Inhibitors of Cholera Toxin Subunit B

This project screened over 47,800 compounds to identify novel inhibitors for the cholera toxin subunit B (CTB). We hypothesized that our screenings would allow us to identify small drug-like molecules that would make similar interactions with the CTB as galactose-based inhibitors. After collecting data on publicly available ligands known to inhibit the canonical CTB site, we created pharmacophore models for each ligand in order to pinpoint critical “features” areas of the ligand that formed hydrogen bond interactions with the CTB. After running a pharmacophore model search across three compound libraries, we isolated 137 compounds that best matched with the canonical ligands’ features. We then performed template-docking trials for all of the 137 compounds (using the ganglioside (GM1) receptor as the template. After, we clustered results using a Tandimodo coefficient of 50 in order to determine common features of each compound cluster. After careful analysis and deliberation, from these 137 compounds, we identified 15 top compound candidates for in vitro experimentation, separated into three groups based on docking S-value, atomic weight, and search frequency. Our research is intended to provide a starting place for in vitro studies which will further confirm the efficacy of the potential CTB inhibitors.

Project ID: 220

SR - Computational Biology and Bioinformatics

Nandana Madhukara
Grade 10
Canyon Crest Academy
Advisor: Ed Gerstin

Lipid Bilayer Bending Due to Membrane-Protein Interactions: A Computational Investigation

Every single cell in an organism has a membrane that in addition to being compartmental barriers play a role in maintaining cell shape and function. For example, sickle cell anemia is a very deadly disease that is caused when the cell membrane deforms in an irregular manner. Therefore, maintaining the desired cell shape is vital for living organisms.

To investigate the role of how different mechanical forces can change the shape of these cellular membranes, we can use tools from mechanics to understand the force balance relationships. One way to do this is with Discrete Differential Geometry which is what Mem3dg does. This tool mathematically models a membrane based on different external factors like the surrounding osmotic pressure and spontaneous curvature-inducing proteins. However, this model accounts for the curvature caused by one protein, so the ultimate goal is for the model to support any number of proteins. In this project, we improve this existing model and allow it to simulate 2 proteins.

In the beginning of this project, derived the necessary equations for the model that described the forces acting on the membrane. In this phase of the project, we created parallel equations for the new model that involved more than one protein. Finally, we coded up the new model and ran simulations. In order to test the implementation of a second protein, we ran a series of tests including testing whether the model can accurately model endocytosis and the model was able to pass all these tests. One way I would like to continue this project is by making the model support any number of proteins the user inputs which would be valuable for researchers studying membrane shape.

Project ID: 221

SR - Computational Biology and Bioinformatics

Arunraj Jeyaprakash
Grade 10
Canyon Crest Academy
Advisor: Alex Siegel

Artificial Intelligence Techniques for Automated Detection of Autism Spectrum Disorder Based on Mobile Camera Imaging

People with Autism Spectrum Disorder (ASD) face ongoing struggles with communication, socializing, and behavior. An early and accurate diagnosis of ASD is essential for successful treatment and better outcomes for those with the disorder. However, the process of diagnosing ASD can be difficult, costly, and lengthy, requiring the participation of multiple experts. I have observed many people, including those close to my age and younger, who are affected by ASD. This disorder not only alters the lives of those with it, but also has an impact on the lives of their families.

I propose using the Mobile camera imaging based scanpath of the human eye to diagnose ASD using artificial intelligence. The scanpath is the sequence of eye movements that people make while they look at an object or scene. Research has shown that individuals with ASD have distinct eye-scanpath patterns, which are different from those of typical individuals. By analyzing these patterns, I aim to develop a deep/machine learning model that can accurately detect the symptoms of ASD. I hypothesized that my proposed model would have an accuracy rate of at least 80%, which would be a significant improvement over current diagnostic methods.

To test this hypothesis, I used eye-scanpath data collected from a large number of individuals, both with and without ASD. I augmented this data to improve the accuracy of the analysis. I then processed this data and trained multiple machine learning models, including a deep and convolutional neural network. The results of my analysis showed that the average accuracy of the machine learning models was 72%, which is relatively low compared to the desired accuracy. However, the neural networks showed an average accuracy of 85%, which was significantly higher. This indicates that the accuracy of the model can be improved and that the prototype with a neural network has the potential to be implemented on mobile cameras, making early, inexpensive, and accurate diagnosis of ASD more accessible to individuals and families.

Project ID: 223

SR - Computational Biology and Bioinformatics

Arshia Nayebnazar
Grade 11
Canyon Crest Academy
Advisor: Tony Mauro

Precision-Recall Adversarial Network (PRAN): A Novel Deep Learning Framework for Improvement of Clinical Binary Classification

The progression of Machine Learning (ML) techniques coupled with the rise of Electronic Health Records (EHR) has taken the medical field by storm, demonstrating the capability of predicting vital clinical outcomes. However, practical use of ML is uncommon due to the lack of sufficiently high results, preventing ML in various health applications from being more than a proof-of-concept. Part of the challenge is the frequency of imbalanced data in clinical datasets. This causes ML models to have misleadingly high accuracies despite ignoring the minority class. More appropriate metrics to evaluate a model's performance are precision and recall. However, the rise of one metric's value typically drops the other metric's value. Although there are many approaches to tackling imbalanced data, such as data resampling or cost-sensitive learning, many of them have a lack of control over balancing both precision and recall and have limited overall performance. I propose the Precision-Recall Adversarial Network (PRAN) framework, which leverages competition between a network that aims to increase recall with another that aims to increase precision. I hypothesized that this would allow both metrics to balance and be more flexible to control, allowing for stronger overall performance in turn. This framework was tested on four clinical imbalanced datasets from UCI's Machine Learning Repository, including on the classification of hepatitis survival, hyperthyroid, breast cancer prognosis, and Parkinson's disease. PRAN had among the highest median F1-Scores with 10 stratified K-Fold cross validation splits for the first three respective datasets in comparison with the most common clinical ML classifiers, including SMOTE-enhanced and cost-sensitive classifiers. This indicates promise for the PRAN framework to improve the progression of clinical ML and may also be used in any imbalanced classification problem.

Project ID: 224

SR - Computational Biology and Bioinformatics

Vidha Yadav Ganji
Grade 9
Del Norte High School
Advisor: Nagamalleswara Rao Ganji

Machine-Learning Based Early Detection of Breast Cancer Using Blood-Based Biomarkers

Recent research has indicated that an astute, dependable, early diagnosis of breast cancer is difficult to achieve. Current methods are expensive, detrimental to general human health, or lack sensitivity to certain types of tumors. This project serves to highlight the importance in the use of certain mRNA blood biomarkers to diagnose breast cancer, taking into consideration that the laboratory and clinical work to isolate such blood biomarkers is cost-effective compared to existing test designs. Research was done to isolate certain genes, the expression levels of which were differentially expressed between control and cancer patients. We hypothesized that certain mRNA based blood biomarkers involved in crucial cellular pathways could effectively be used to predict breast cancer prior to symptomatic progression of the disease. The procedure required to evaluate this hypothesis involved the use of a control and diseased dataset from PltDB (a platelet-based cancer database) which was preprocessed then reordered to accommodate for genes that held the lowest to highest p-values. The thousand most differentially expressed genes were fitted into a machine learning algorithm to determine the top two hundred and fifty feature importance's. The genes that were extracted from the model were then analyzed using gene ontology and correlations to crucial cellular pathways were found. Therefore, our project concluded that expression levels for these two-fifty genes could be used to at least partially determine an early diagnosis for breast cancer, alleviating some of the technological and fiscal burden off of clinical professionals and patients, both.

Project ID: 225

SR - Computational Biology and Bioinformatics

Aarav Arora
Grade 11
Del Norte High School
Advisor: Andrea Callicott

Novel Laryngeal Cancer Diagnostic Approach Using MiRNA and Metabolite Biomarkers

Laryngeal cancer is the most common head and neck cancer, which often goes undiagnosed due to the expensiveness and inaccessible nature of current diagnosis methods.
In this study, a neural network model is created for the diagnosis of laryngeal cancer using a created series of microRNA (miRNA) attributes. Unique features are extracted from each miRNA such as sequence-based information, known gene targets, and miRNA pathways. The model was trained with a combination of known laryngeal cancer-associated miRNA and random non-associated miRNA. The results showed that the created model can identify miRNA associated with laryngeal cancer with 86% accuracy using a multilayer perceptron model. Furthermore, the model was validated with existing cancer datasets and can accurately classify publicly available cancer datasets at >80% accuracy. The model reached 0.864 precision, 0.860 TP rate, 0.868 ROC AUC, and 0.859 F-Measure for miRNA classification. Our study demonstrates that the proposed model and an inexpensive miRNA testing device have the potential to serve as a cost-effective and accessible method for diagnosing laryngeal cancer.

Project ID: 226

SR - Computational Biology and Bioinformatics

Ritvik Irigireddy
Grade 11
Canyon Crest Academy
Advisor: Ed Gerstin

Machine-Learning Based Rapid and Efficient Diagnosis of Tuberculosis

In 2021, tuberculosis, a bacterial disease affecting the lungs, resulted in 1.6 million deaths and spread to 10.6 million people. The World Health Organization (WHO) made it a goal to have 0% tuberculosis patients by 2020 but failed due to not enough funding. Tuberculosis is a completely preventable disease yet without a quick enough diagnosis, severe issues may arise. Low and middle income countries make up the vast majority of tuberculosis cases but due to a lack of funding from the WHO and an insufficiency of current diagnosis methods, 40% of cases ended up not diagnosed. Most of these low and middle income countries have X-Ray machines but no dedicated health-care professional to analyze the results readily. A promising solution to allow for rapid diagnosis of tuberculosis is the use of a neural network that can classify x-ray images of the lung as having tuberculosis or not. The data was obtained from Kaggle and split into 80% being allocated for training and 20% being allocated for testing. After Tensorflow and Keras were used and optimized for the creation of the model, an accuracy rate of approximately 95% resulted. These results from the model are extremely promising and current work is being done by the author to enhance the diagnosis process.

Project ID: 227

SR - Computational Biology and Bioinformatics

Zhijing Wang
Grade 11
Canyon Crest Academy
Advisor: Ed Gerstin

System Biology Analysis of Rheumatoid Arthritis Samples for Personalized Treatment

Rheumatoid Arthritis (RA) is an autoimmune disease that inflames the joints, affecting approximately one percent of the United States population. It is triggered by different mechanisms in different people and effective treatments are often found through “trial and error.” During this lengthy and costly process, the patients’ symptoms may get worse and they may suffer for a longer time. My research aims to use a systems biology approach to discover new drug targets and develop personalized treatment for RA patients. In this study, I integrated RNA-Seq and ATAC-Seq data from 10 RA and 10 Osteoarthritis (OA) patients to identify genes and pathways important for RA pathogenesis in individuals, which would help find the most effective drugs for each RA patient. Furthermore, I discovered new potential drug targets such as a kinase protein BST1 to develop new potent therapeutics. The next step is to further improve this approach and ultimately apply it to clinical practices.

Project ID: 228

SR - Computational Biology and Bioinformatics

NS

Eleanor Jung
Grade 12
Mt. Carmel High Shool
Advisor: Amy Klingborg

Novel Method to Image-based Intracellular Force Estimation with Discrete Differential Geometry

The study of how force is generated and translated to exert output functions within cells has become a topic of growing interest within recent decades. Not only do these processes drive fundamental cell functions including division, migration, and differentiation, they are also key to understanding widespread diseases like cancer. Defects in mechanobiological pathways have been traced to several detrimental conditions that, if treated, could prevent a number of common disorders. However, studies that aim to measure intracellular forces directly are limited by current techniques. Many existing methods, including the use of probes and other instruments that attempt to take physical measurements pose risks of damaging the cell, thus ruining experimental data, and accuracy is often difficult to obtain, as measurements can be easily influenced by slight changes in the environment. Conversely, image-based techniques offer greater potential to combine multiple known variables (e.g., material properties) and experimental data for estimating forces with less invasiveness and more accuracy. In this study, we develop a novel technique which marries advances in automatic differentiation and biological image analysis to predict the forces on membranes within cells. To test our method, we apply our work to ActuAtor, a molecular tool engineered from ActA proteins (actin nucleation factors) for controlled deformation of intracellular structures. Using EM images of deformed nuclei, we trace the shapes of the nuclear membranes and use the Helfrich-Canham-Evans free energy model of biological membranes to estimate quantitative measures of the force generated by ActuAtor. This technique has wide-ranging applications, from allowing future experiments to accurately study cell dynamics from a mechanical standpoint to providing a basis for future advancements in clinical medicine.

Project ID: 229

SR - Computational Biology and Bioinformatics

Vidhi Kulkarni
Grade 11
Del Norte High School
Advisor: Courtney Craig

Implementation of Machine Learning-Based System for Early Diagnosis of Feline Mammary Carcinoma through Blood Metabolite Profiling

Feline mammary carcinoma (FMC) is a prevalent and fatal carcinoma that predominantly affects unspayed female cats. FMC is the third most common carcinoma in cats, but is still underrepresented in research. Current diagnosis methods include physical examinations, imaging tests, and fine-needle aspiration. The diagnosis through these methods is sometimes delayed and unreliable, leading to increased chances of mortality.

The objective of this study was to identify the biomarkers, including blood metabolites and genes, related to feline mammary carcinoma, study their relationships, and develop a machine-learning (ML) model for the early diagnosis of the disease.

I analyzed metabolites of felines with mammary carcinoma using the pathway analysis and gene-metabolite features in the MetaboAnalyst software. The metabolic pathways that were explored include alanine, aspartate and glutamate metabolism, D-glutamine and D-glutamate metabolism, arginine biosynthesis, and glycerophospholipid metabolism. Furthermore, I identified various genes that play a significant role in the development of FMC. I utilized machine-learning methods in the development of a â€œRandom Forestâ€ classifier for blood metabolites of FMC. The best-performing model was able to predict metabolite class with an accuracy of 85%.

I also developed an app that allows pet owners to verify the preliminary symptoms of their feline for FMC. A Raspberry Pi device was also designed to process blood metabolite data, and output whether the feline has mammary cancer. In addition, the device takes in visual and auditory input to further verify its diagnosis. The app and device provide a non-invasive and accurate means of diagnosing FMC, improving the chances of early detection and effective treatment.

In conclusion, my findings demonstrate that the identification of the biomarkers associated with FMC and the affected metabolic pathways for carcinoma can aid in the early diagnosis of feline mammary carcinoma.

Project ID: 230

SR - Computational Biology and Bioinformatics

Katherine Ge
Grade 11
The Bishop's School
Advisor: Lani Keller

Docking Peptides into HIV/FIV Protease for Structure-based Drug Design

Molecular docking is a staple in structure-based drug design. However, the process to accurately compute each prediction is immensely complex. Different docking platforms have been investigated. They differ not only in sampling /scoring algorithms, but also vary in the flexibility of receptors and ligands.

Traditional methods, such as AutoDock CrankPep (ADCP) developed by the Sanner Lab at TSRI, rely on a rigid representation of the macromolecule and a Monte Carlo search. AlphaFold2, a DeepMind deep learning system, predicts interactions between peptides and their binding partners using a fully flexible receptor and ligand.

In this project, my goal is to evaluate the performance of four docking methods (AlphaFold2 Monomer, AlphaFold2 Multimer, ADCP, and OmegaFold). Furthermore, I will assess the ability of each method to discriminate between native and non-native ligands, by docking sets of native and designed peptides (constructed with opposite physicochemical properties) into HIV/FIV protease.