GENETIC-BASED DISEASE IDENTIFICATION WITH DEEP LEARNING ON NEURAL NETWORKS

HTTPS://GITHUB.COM/KALLEMPRATHUSHA/GROUP-PROJECT

Prathusha Kallem
700745065
Dept Computer Science
University of Central Missouri
Github link: https://github.com/KallemPrathusha/Group-Project

Abstract:

The difficult task of extracting disease genes from the human genome has drawn considerable interest from the biomedical research community. This is because there are a limited number of known disease genes in the entire genome, yet genetic disorders are brought on by a wide range of conditions. A third difficulty for researchers is the genetic variability of illnesses. The issue of finding new disease genes continues to be difficult despite multiple attempts to apply machine-learning techniques.

The idea of "guilt by association" has been used in recent methods to look into the relationship between the phenotype of an illness and the genes that cause it. The strategy is predicated on the notion that potential disease-related candidate genes have similar properties

diseases are more likely to be correlated with genes. However, utilizing unknown genes as training, semi-supervised methods such as label propagation techniques and positive-unlabeled learning are utilized to identify putative disease genes. The imbalance issues with disease gene identification are the cause of this. When there are a small number of known disease genes (labelled data) and a substantial amount of unknown genomes (unlabeled data), this is often done.

The usefulness of disease gene prediction models is constrained by the potential bias of a single learning model as well as the sparsity and noise of a single biological data source. Ensemble learning models, which incorporate a range of biological inputs and learning models, are used to increase predictive performance. Finding potential has been accomplished through the use of ensemble learning models. disease genes.

This thesis proposes three computer methods for identifying probable disease genes. These models attempt to address the issues surrounding the identification of disease genes by utilizing a variety of biological data sources and ensemble learning techniques. The suggested models have been tested on several benchmark datasets, and they perform noticeably better than the methods currently in use. The proposed models may facilitate the discovery of novel disease genes and enhance our knowledge of the genetic basis of diseases.

Keywords: Diseases, Deep learning, Neural network, Genetic algorithms, Accuracy, human genome, python

1. INTRODUCTION

Recognizing gene-disease joins is of extraordinary worth in human sickness conclusion and therapy. The known illness-related qualities answered to public data sets, like the Web-based and the Hereditary Affiliation Data set, address a little part of genuine connections. Consequently, finding infection qualities stays significant. Conventional quality planning approaches include linkage examination and a far-reaching affiliation study. Because of the set number of hybrids in tested families, linkage examinations for the most part distinguish just chromosomal spans that might contain up to many applicant qualities. Broad affiliation studies may likewise return numerous locales which still need to be analyzed in ongoing works. Exploratory approvals of so many up-and-comer qualities are tedious and costly. Since coordinating numerous helper wellsprings of information is fundamental for quality sickness characteristics, a progression of network-based computational options has been proposed in the previous ten years. The normal inspiration of these strategies is that qualities causing something similar or comparative sicknesses will generally intently relate with each other in the organic organizations.

The run-of-the-mill proof that these models can be arranged is as per the following: text-mining of biomedical writing, practical explanations, pathways and ontologies, aggregate connections, inherent quality properties, grouping information, protein collaborations, administrative information, orthologous relationships, and gene expression data. For instance, utilizing a text-mining way to deal with characterize the huge scope of human aggregates contained in the database. To define the similarities in protein-protein collaboration organizations, deduced quality illness associations by utilizing a worldwide organization distance measure called irregular walk examination. Plus, all the more as on the heterogeneous organization, which fosters an augmentation of there and walk utilizing walk build up to registers likenesses between two articles and consolidate significant data from different species, for example, natural product fly and mouse. The principal disadvantage of those organization-based techniques is that they miss the mark regarding summing up clever illnesses, for which there are no quality linkage concentrates yet. In such a manner, developed a subsequent strategy, Inductive Grid Fruition, in light of various organic sources, which can be applied to illnesses not seen at preparing time. Of the above techniques for focusing on qualities pertinent to a given illness, standard IMC plays out the best, despite the fact that it might bring about a shallow comprehension of the highlights.

2. MOTIVATION

The Deep learning will be the python based application which contributes to find out the Genetic disease early stage . It will be helpful for the human to detect at early and to take necessary treaments in the correct time. The progression of profound learning influences is generally applied to classification assignments and portrayals learning. These profound frameworks with numerous layers have been displayed to yield promising execution in removing serious areas of strength for more of information. The streamlining of the goal capability becomes curved in the event that we adjust one variable and fix the others.

The finding of the application, includes the 'Clinical Elements' and 'Clinical Administration' segments of the website pages that report the side effects, prescription and reactions by patients, and related investigations of impacts of various courses of treatments.

Since cross-approval on review information presumably prompts overoptimistic results, cross-approval is improper for this issue. To assess the capacity of the models to anticipate newfound affiliations, we train and test the dataset.

Thus, via consistently consolidating the model for helper side data and the cooperative filter for the quality sickness affiliations grid, our model learns a significantly more significant portrayal for every quality and illness and gives more exact expectations.

3.MAIN CONTRIBUTIONS & OBJECTIVE

- The objective of Genetic disease identification with deep learning is to detect the Genetic disease in the early stage itself with the available attributes.
- In this work, the dataset containing the Genetic disease will be taken into consideration
- The pre-processing will be applied to the dataset, and the noisy and null value data will be removed
- After the data will be analyzed and visualized for further processing. The Deep learning algorithm will be chosen to make the prediction
- The dataset will be divided into train, test, and validity and it has genetic disease images of the brain.

4.RELATED WORK

The number of atomic descriptors was decreased by performing a highlight determination strategy. The strategy was led in two stages, for example, measurable investigation and Genetic Algorithm. In the initial step, the descriptors with low standard deviation or containing comparable qualities over half were taken out. Then, at that point, an investigation of the Pearson connection was completed to decide the relationship between the descriptor and between the descriptor and the target. This step was performed to lessen predisposition and eliminate the descriptors with comparative data. The descriptors that have a feeble connection with the target (relationship < 0.1) or a solid connection with another objective (connection > 0.9) were taken out.

On account of areas of strength for the two descriptors, the descriptor with a more fragile connection with the objective was eliminated. In the subsequent step, a mix of descriptors was chosen by utilizing the hereditary calculation Genetic Algorithm strategy.

This strategy observes Darwin's exemplary guidelines of normal advancement and utilizations irregular techniques to get ideal non-arbitrary arrangements.

The descriptor choice by GA was performed by characterizing the arrangement as an assortment of a whole number worth in a chromosome. For this situation, the quantity of the number worth is equivalent to the quantity of the chosen descriptor, in which the worth address the descriptor list. We involved the cross entropy misfortune as a goal capability during the element determination.

We fostered a forecast model by utilizing a Genetic Algorithm technique. This technique is a numerical model that looks like the construction and capability of the natural sensory system. The essential rule of the Genetic Algorithm is the execution of fake neurons, which are straightforward numerical models. Such a model has three straightforward arrangements of rules, for example, duplication, expansion, and actuation. Then, at that point, play out the y-scrambling examination to ensure that the exhibition of the model didn't relate to an unintentional relationship. This investigation was led by rearranging the class focus while saving the descriptors multiple times. The aftereffects of the y-scrambling present by giving the values to rearranged and unshuffled information.

5.PROPOSED FRAMEWORK

The proposed framework involves applying the convolutional neural network to the dataset in the identification of genetic diseases.

The testing and training variables are split and passed into the algorithm for the Genetic disease prediction.

Convolutional Neural Networks model:

In profound learning, a convolutional neural network (CNN) is a class of profound brain organizations, generally regularly applied to dissect visual symbolism. Presently when we consider a brain network we ponder framework increases yet that isn't true with ConvNet. It utilizes a unique strategy called Convolution. Presently in science convolution is a numerical procedure on two capabilities that creates a third capability that communicates how the state of one is changed by the other.

Keras model of Deep Learning:

Keras is a brain network Application Programming Connection point for Python that is firmly coordinated with Tensor Stream, which is utilized to construct AI models. Keras' models offer a straightforward, easy-to-understand method for characterizing a brain organization, which will then, at that point, be fabricated.

Keras is a strong and simple to-involve free open-source Python library for creating and assessing profound learning models. It is important for the Tensor Flow library and permits you to characterize and prepare brain network models in only a couple of lines of code.

Keras Deep Learning working process:

1. Dataset load

2. Keras Model Definition

3. Keras Model Compilation

4. Keras Model Fit and evaluation

5. Predictions

Keras is a Python library including a Programming interface for working with brain organizations and profound learning systems. Keras incorporates Python-based techniques and parts for working with different Profound Learning applications.

By adding and erasing layers, the Models Programming interface permits you to construct muddled brain organizations. The model could be consecutive, suggesting that the layers are heaped one on top of the other with solitary information and result. The model can likewise be functional, implying that it tends to be different.

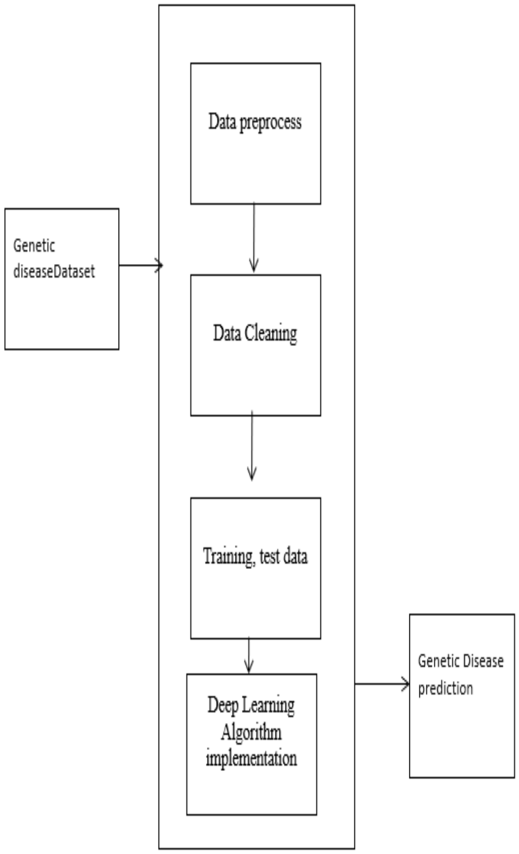
A preparation module is likewise remembered for the Programming interface, with techniques for producing the model, as well as the enhancer and misfortune capability, fitting the model, and assessing and gauging input messages. It likewise gives strategies to bunch information preparing, testing, and gauging. The models Programming interface likewise permits you to save and preprocess your models.

Keras models are partitioned into two classes:

- Keras Utilitarian Programming Interface

- Keras Consecutive Model

SYSTEM ARCHITECTURE DIAGRAM :



Figure[1].architecture diagram

6. DATA DESCRIPTION

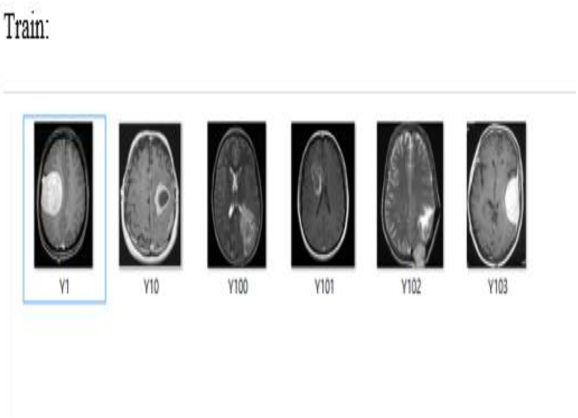
Dataset: The majority of the dataset's photos are cluttered with data. However, feature engineering will produce more successful outcomes. Importing libraries and loading data comes first. The next step is to gain a fundamental understanding of the data, including its shape, sample, and the presence of any NULL values in the collection. Understanding the data is a crucial stage in any deep learning research or prediction. That there are no NULL values is a good thing from fig(1).

The Kaggle website is used to download brain X-ray picture data.

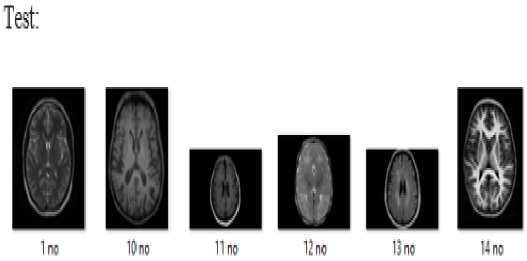
It contains brain X-ray images of patients with genetically impacted and unaffected conditions in three folders: train, test, and validate.

The fields required for the analysis of this dataset's detailed design of features of the dataset on genetic disorders. To produce highlight designing and Deep

Learning exhibiting steps smoothly and in line with anticipation, the exploratory examination is a cycle to investigate and comprehend the information and information connected in entire depth. The exploratory analysis helps us determine whether our assumptions are accurate or misleading**.**



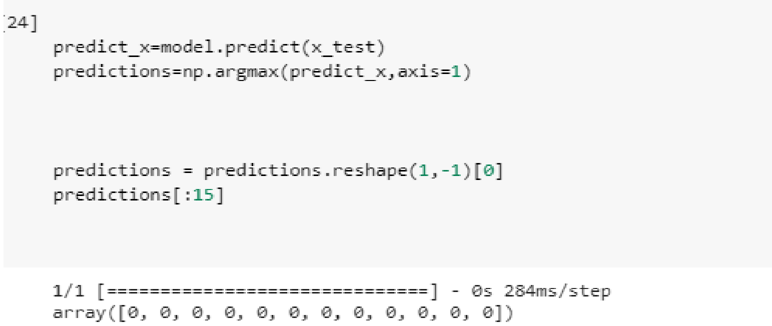
Figure[2].Train data



Figure[3].Test data

7.**RESULTS&ANALYSIS**

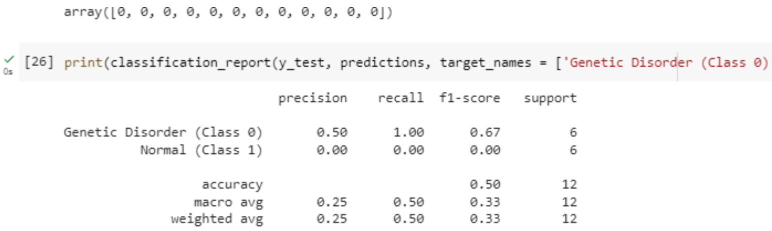
The accuracy, confusion matrix of the neural network is given below:



```
[24]
    predict_x=model.predict(x_test)
    predictions=np.argmax(predict_x,axis=1)



    predictions = predictions.reshape(1,-1)[0]
    predictions[:15]


    1/1 [==============================] - 0s 284ms/step
    array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

Figure[4].Prediction

The accuracy results:

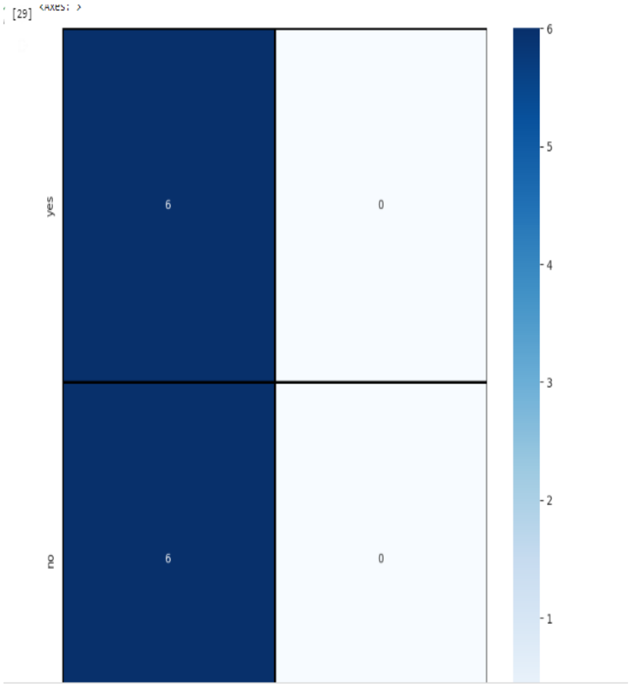


```
    array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])

[26] print(classification_report(y_test, predictions, target_names = ['Genetic Disorder (Class 0)

                            precision    recall  f1-score   support

Genetic Disorder (Class 0)       0.50      1.00      0.67         6
         Normal (Class 1)       0.00      0.00      0.00         6

                 accuracy                           0.50        12
                macro avg       0.25      0.50      0.33        12
             weighted avg       0.25      0.50      0.33        12
```

Figure[5].Accuracy

The final results are compared with a different type of algorithm accuracy levels.

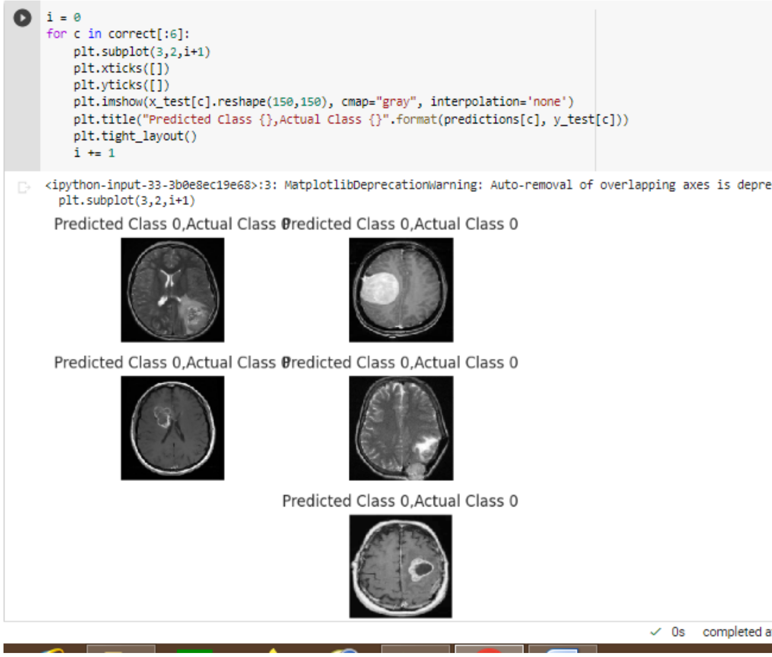
Heat map:



Figure[6].Heat Map

The heat map matches the genetic diseases is present(yes) and not-present(no) values in the given dataset.

Results of Prediction:



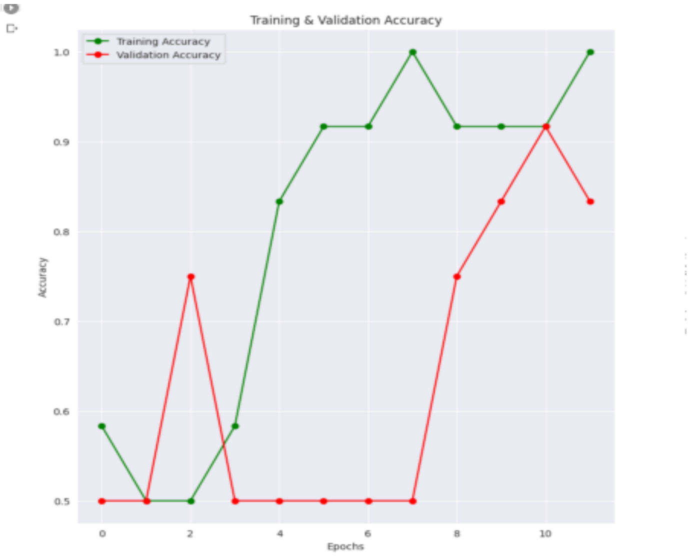
```
i = 0
for c in correct[:6]:
    plt.subplot(3,2,i+1)
    plt.xticks([])
    plt.yticks([])
    plt.imshow(x_test[c].reshape(150,150), cmap="gray", interpolation='none')
    plt.title("Predicted Class {},Actual Class {}".format(predictions[c], y_test[c]))
    plt.tight_layout()
    i += 1
```

```
<ipython-input-33-3b0e8ec19e68>:3: MatplotlibDeprecationWarning: Auto-removal of overlapping axes is depre
  plt.subplot(3,2,i+1)
```

```
Predicted Class 0,Actual Class 0Predicted Class 0,Actual Class 0
Predicted Class 0,Actual Class 0Predicted Class 0,Actual Class 0
                                Predicted Class 0,Actual Class 0
```

Figure[7].Prediction results

The prediction results of the genetic disorder are shown.



Figure[8].Training and Validation Accuracy

The training and the validation accuracy graph shows the results with a graphical format.

References

[1]. Nour eldeen m. khalifa 1, mohamed hamed n. taha 1, dalia ezzat ali 1, adam slowik 2, (senior member, ieee), and aboul ella hassanien “Artificial Intelligence Technique for Gene Expression by Tumor RNA-Seq Data: A Novel Optimized Deep Learning Approach”. February 6, 2020.Digital Object Identifier 10.1109/IEEE ACCESS.2020.2970210

[2]. Xiangxiang Zeng, Senior Member, IEEE, Yinglai Lin, Yuying He, Linyuan L¨u, Xiaoping Min∗, and Alfonso Rodr´ıguez-Pat´on” Deep collaborative filtering for prediction of disease genes”. DOI 10.1109/TCBB.2019.2907536, IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[3] W. R. J. Taylor and N. J. White, “Antimalarial drug toxicity: a review,” Drug Saf., vol. 27, no. 1, pp. 25–61, 2004, doi: 10.2165/00002018200427010-00003.

[4] E. A. Ashley et al., “Spread of artemisinin resistance in Plasmodium falciparum malaria,” N. Engl. J. Med., vol. 371, no. 5, pp. 411–423, Jul. 2014, doi: 10.1056/NEJMoa1314981.

[5] E. Tjitra et al., “Multidrug-resistant Plasmodium vivax associated with severe and fatal malaria: a prospective study in Papua, Indonesia,”. PLoS Med., vol. 5, no. 6, p. e128, Jun. 2008, doi: 10.1371/journal.pmed.0050128.

[6] A. M. Dondorp et al., “Artemisinin Resistance in Plasmodium falciparum Malaria,” N. Engl. J. Med., vol. 361, no. 5, pp. 455–467, Jul. 2009, doi: 10.1056/NEJMoa0808859.

[7] W. O. Godtfredsen, W. von Daehne, L. Tybring, and S. Vangedal, “Fusidic Acid Derivatives. I. Relationship between Structure and Antibacterial Activity,” J. Med.

Chem., vol. 9, no. 1, pp. 15–22, Jan. 1966, doi: 10.1021/jm00319a004.

[8] G. Kaur et al., “Synthesis of fusidic acid bioisosteres as antiplasmodial agents and molecular docking studies in the binding site of elongation factor-G,” MedChemComm, vol. 6, no. 11, pp. 2023–2028, 2015, doi: 10.1039/C5MD00343A.

[9] S. Tonmunphean, V. Parasuk, and S. Kokpol, “QSAR Study of Antimalarial Activities and Artemisinin-Heme Binding Properties Obtained from Docking Calculations,” Quant. Struct.-Act. Relatsh., vol. 19, no. 5, pp. 475–483, 2000, doi: 10.1002/15213838(200012)19:5 <475::AID-QSAR475>3.0.CO;2-3.

[10] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, and V. Prachayasittikul, “QSAR study of amidino bis-benzimidazole derivatives as potent anti-malarial agents against Plasmodium falciparum,” Chem. Pap., vol. 67, no. 11, pp. 1462–1473, Nov. 2013, doi: 10.2478/s11696-013-0398-5.

[11] M. C. Sharma, S. Sharma, P. Sharma, and A. Kumar, “Pharmacophore and QSAR modeling of some structurally diverse azaaurones derivatives as anti-malarial activity,” Med. Chem. Res., vol. 23, no. 1, pp. 181–198, Jan. 2014, doi: 10.1007/s00044-013-0609-1.

[12] M. Fernandez, J. Caballero, L. Fernandez, and A. Sarai, “Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM),” Mol. Divers., vol. 15, no. 1, pp. 269–289, Feb. 2011, doi: 10.1007/s11030-010-9234-9.

[16] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, and J. E. Richardson, “The mouse genome database (mgd): new features facilitating a model system,” Nucleic Acids Research, vol. 35,

no. Database issue, pp. 630–7, 2007.

[13] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, and G. Sherlock, “Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go),” Nucleic Acids Research, vol. 30, no. 1, pp. 69–72, 2002.

[14] T. L. Saito, M. Ohtani, H. Sawai, F. Sano, A. Saka, D. Watanabe, M. Yukawa, Y. Ohya, and S. Morishita, “Scmd: Saccharomyces cerevisiaemorphologicaldatabase,”NucleicAcidsResearch,vol.32, no. 1, pp. 319–22, 2004.

[15] K. L. Mcgary, I. Lee, and E. M. Marcotte, “Broad network-based predictability of saccharomyces cerevisiae gene loss-of-function phenotypes.” Genome Biology, vol. 8, no. 12, p. R258, 2007.

[16] M. E. Hillenmeyer, E. Fung, J. Wildenhain, S. E. Pierce, S. Hoon, W. Lee, M. Proctor, R. P. St Onge, M. Tyers, and D. Koller, “The chemical genomic portrait of yeast: uncovering a phenotype for all genes.” Science, vol. 320, no. 5874, pp. 362–365, 2008.

[17] R. J. Nichols, S. Sen, Y. J. Choo, P. Beltrao, M. Zietek, R. Chaba, S. Lee, K. M. Kazmierczak, K. J. Lee, and A. Wong, “Phenotypic landscape of a bacterial cell,” Cell, vol. 144, no. 1, pp. 143–156, 2011.

[18] J. Sprague, D. Clements, T. Conlin, P. Edwards, K. Frazer, K. Schaper, E. Segerdell, P. Song, B. Sprunger, and M. Westerfield, “The zebrafish information network (zfin): the zebrafish model organism database,” Nucleic Acids Research, vol. 34, no. 1, pp. 241– 243, 2003.

[19] G. W. Bell, T. A. Yatskievych, and P. B. Antin, “Geisha, a wholemount in situ hybridization gene expression screen in chicken embryos,”

Developmental Dynamics, vol. 229, no. 3, pp. 677–687, 2010.

[20] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork et al., “The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible,” Nucleic acids research, p. gkw937, 2016.