SPRING 2023 FINAL

You're Cute Genes

WHICH GENES ARE DISCOVERED AND WHEN

ANNA VICTORIA LAVELLE
KAILLA MADERA
CECI NGUYEN

Table of Contents

Introduction

HUMAN GENE NOMENCLATURE REFERS TO THE SYSTEM OF NAMING HUMAN GENES IN A STANDARDIZED AND CONSISTENT WAY. THE NAMING OF GENES IS IMPORTANT BECAUSE IT ENABLES RESEARCHERS TO COMMUNICATE THEIR FINDINGS EFFECTIVELY AND EFFICIENTLY. HOWEVER, GENE NAMING CAN BE A COMPLEX PROCESS THAT IS SUBJECT TO ERRORS AND INCONSISTENCIES, WHICH CAN CAUSE CONFUSION AND HINDER SCIENTIFIC PROGRESS.

ONE OF THE PRIMARY ISSUES WITH HUMAN GENE NOMENCLATURE IS THAT MANY GENES HAVE MULTIPLE NAMES OR ALIASES, WHICH CAN LEAD TO CONFUSION WHEN DIFFERENT RESEARCHERS USE DIFFERENT NAMES FOR THE SAME GENE. ADDITIONALLY, GENES CAN BE NAMED BASED ON A VARIETY OF FACTORS, INCLUDING THEIR FUNCTION, LOCATION, OR SEQUENCE, WHICH CAN LEAD TO INCONSISTENT AND CONFUSING NAMES. TO ADDRESS THESE ISSUES, THE HUMAN GENOME ORGANIZATION (HUGO) HAS ESTABLISHED THE HUGO GENE NOMENCLATURE COMMITTEE (HGNC) TO OVERSEE THE NAMING OF HUMAN GENES. THE HGNC IS RESPONSIBLE FOR ASSIGNING UNIQUE AND CONSISTENT NAMES TO HUMAN GENES BASED ON ESTABLISHED GUIDELINES.

THE HGNC PLAYS A CRITICAL ROLE IN ENABLING COMMUNICATION AND COLLABORATION AMONG RESEARCHERS STUDYING HUMAN GENES. BY ENSURING THAT EACH GENE HAS A UNIQUE AND CONSISTENT NAME, THE HGNC HELPS TO PREVENT CONFUSION AND ERRORS THAT CAN HINDER SCIENTIFIC PROGRESS. IN ADDITION TO ITS IMPACT ON SCIENTIFIC RESEARCH, HUMAN GENE NOMENCLATURE ALSO HAS IMPORTANT APPLICATIONS IN MEDICINE AND BIOTECHNOLOGY. FOR EXAMPLE, THE ACCURATE AND CONSISTENT NAMING OF GENES IS ESSENTIAL FOR THE DEVELOPMENT OF GENE-BASED THERAPIES AND PERSONALIZED MEDICINE.

IN SUMMARY, THE ISSUES WITH HUMAN GENE NOMENCLATURE STEM FROM THE COMPLEXITY OF GENE NAMING AND THE POTENTIAL FOR ERRORS AND INCONSISTENCIES. THE HGNC PLAYS A CRITICAL ROLE IN ENSURING THAT GENE NAMES ARE STANDARDIZED AND CONSISTENT, WHICH IS ESSENTIAL FOR ADVANCING SCIENTIFIC RESEARCH AND ENABLING THE DEVELOPMENT OF NEW MEDICAL TREATMENTS AND TECHNOLOGIES.

Background, Data Description & Problem

THE HGNC HAS COMPILED A DATA FILE CALLED THE "HGNC_COMPLETE_SET" WHICH INCLUDES ALL APPROVED GENE SYMBOL REPORTS FOUND ON THE GENOME REFERENCE CONSORTIUM HUMAN BUILD 38 (GRCH38) AND THE ALTERNATIVE REFERENCE LOCI. THE GRCH38 IS THE MOST RECENT VERSION OF THE HUMAN GENOME ASSEMBLY, ESSENTIALLY THE COMPLETE SET OF GENES FOR A HUMAN BEING. THE ALTERNATIVE REFERENCE LOCI ARE ADDITIONAL GENETIC SEQUENCES THAT ARE NOT INCLUDED IN THE GRCH38 BECAUSE THEY ARE MORE COMPLEX AND VARIABLE. HOWEVER, THEY ARE STILL IMPORTANT AND RELEVANT IN ADVANCED GENETIC STUDIES. OVERALL, THE "HGNC_COMPLETE_SET" IS INCREDIBLY IMPORTANT FOR RESEARCHERS BECAUSE IT PROVIDES UNIQUE SYMBOLS THAT CAN BE USED TO REPRESENT GENES IN STUDIES, PUBLICATIONS, AND OTHER DATABASES AND ENSURES THAT RESEARCHERS ARE USING THE SAME NAME FOR RELEVANT GENES.

THE "HGNC_COMPLETE_SET" CAN BE FOUND AT 'HTTPS://WWW.GENENAMES.ORG/DOWNLOAD/ARCHIVE/', AND THE TSV FILE CAN BE EASILY OPENED IN EXCEL TO SEE THE 43,000+ ROWS AND 50+ COLUMNS. EACH ROW SIGNIFIES A UNIQUE GENE, AND THE ACCOMPANYING COLUMNS PROVIDE ADDITIONAL INFORMATION ABOUT EACH GENE. SOME OF THE MOST INTERESTING COLUMNS ARE "HGNC_ID," "LOCUS_TYPE," "DATE_APPROVED_RESERVED", AND "DATE_MODIFIED." "HGNC_ID" PROVIDES A UNIQUE ID FOR THE GENE WHICH IS AN IMPORTANT IDENTIFIER. "LOCUS_TYPE" PROVIDES INFORMATION ABOUT THE LOCUS TYPE WHICH IS THE TYPE OF GENE LOCATED AT THE PARTICULAR LOCUS. "DATE_APPROVED_RESERVED" AND "DATE_MODIFIED" PROVIDE TIME STAMPS FOR WHEN THE GENE ENTRY WAS FIRST APPROVED AND LAST MODIFIED.

THE "HGNC_COMPLETE_SET" HAS OVER 2 MILLION PIECES OF INFORMATION, MAKING IT NEARLY IMPOSSIBLE TO SORT THROUGH TO FIND THE MEANINGFUL INFORMATION. OUR APPLICATION SOLVES THIS PROBLEM BY CREATING ROUTES THAT, WHEN PROMPTED BY THE USER, CAN RETURN THE IMPORTANT INFORMATION THAT SCIENTISTS, RESEARCHERS, AND OTHER USERS WANT.

Strategies & Method

THE BEST STRATEGY THAT WE FOUND TO SOLVE THE PROBLEM WAS TO CREATE AN APPLICATION WITH DIFFERENT ROUTES TO INTERFACE WITH THE DATA SET.

ROUTES:

THERE ARE FIFTEEN DIFFERENT ROUTES FOR THE USERS TO QUERY. (1-6)

TABLE 1. A LIST OF ALL THE ROUTES IN THE APPLICATION WITH THEIR DESCRIPTIONS

Route	Method	Function
/data	POST	Post the data to the database
/data	GET	Return all the data in the database
/data	DELETE	Delete the data from the database
/genes	GET	Return a list of all HGNC IDs
/genes/<hgnc_id>	GET	Return all of the information for a specified HGNC ID
/image	POST	Generate a plot of the number of genes approved each year and post it to the database

Strategies & Method

ROUTES:

THERE ARE FIFTEEN DIFFERENT ROUTES FOR THE USERS TO QUERY. (7-15)

/image?start=int&end=int	POST	Generate a plot of the number of genes approved each year and post it to the database witha given range
/image	GET, DELETE	Return the plot to the user Delete the plot from the database
/imagedata	GET	Return the data used for generating the plot
/locusdata	GET	Return the number of entries in each locus group
/locus/<hgnc_id>	GET	Return the locus group of a specified HGNC ID
/when/<hgnc_id>	GET	Return dates of approval or modification for a specified HGNC ID
/help	GET	Return help text for the user
/jobs	POST	Creates a new job to do some analysis of the data
/jobs/<id>	GET	Gets the status perameter for the id
/jobs/<id>/results	GET	Retrieves the plot

Strategies & Method

KEY TECHNOLOGIES:
THE APPLICATION USES SEVERAL TECHNOLOGIES TO MAKE IT EASY TO USE AND ACCESSIBLE TO USERS ANYWHERE.

USING FLASK, THE PYTHON WEB FRAMEWORK USED FOR DEVELOPING WEB APPLICATIONS, WE CREATE AN APPLICATION THAT ACCESSES THE DATA AND RETURNS THE DESIRED INFORMATION DEPENDING ON THE USER'S QUERY. OUR FLASK APPLICATION ALSO DEPENDS ON REDIS, WHICH FUNCTIONS AS A DATABASE AND PROVIDES PERSISTENCE OF THE DATA BETWEEN RESTARTS OF OUR APPLICATION AND REDIS ITSELF. USING A DATABASE ALLOWS FOR DATA AND IMAGES TO BE RESTORED IF A USER EXITS REDIS AND THE APPLICATION.

THE APPLICATION IS THEN CONTAINERIZED BY WRITING A DOCKERFILE, A "RECIPE" FOR INSTALLING AND CONFIGURING OUR APPLICATION. THIS "RECIPE" BECOMES ACCESSIBLE ON DOCKER HUB, A CONTAINERIZATION PLATFORM THAT ALLOWS US TO PACKAGE UP OUR SOFTWARE. THIS WAY, ANY USER WHO HAS CLONED OUR APPLICATION CAN PULL THE "RECIPE" FROM DOCKER HUB AND RUN AN INSTANCE OF IT INTO A CONTAINER WITHOUT HAVING TO WORRY ABOUT LIBRARIES AND DEPENDENCIES THAT THEY MIGHT NOT HAVE INSTALLED.

IN ORDER TO AUTOMATE THE APPLICATION'S DEPLOYMENT, WE MUST USE DOCKER COMPOSE WHICH HELPS MANAGE APPLICATIONS THAT HAVE MULTIPLE OPTIONS FOR STARTING THE CONTAINER. WITH DOCKER COMPOSE, THE DESIRED START OPTIONS ARE ESSENTIALLY PREPARED FOR THE USER, MAKING STARTING THE CONTAINER AND USING THE APPLICATION ON A USER'S MACHINE EASY AND SUCCINCT.

KUBERNETES, A CONTAINER ORCHESTRATION SYSTEM, IS ALSO USED TO MANAGE THE CONTAINERIZED APPLICATION BEING DEPLOYED ON ONE OR MORE COMPUTERS. USERS CAN DEPLOY THE APPLICATION TO A KUBERNETES CLUSTER AND QUERY ROUTES FROM THAT VERY LOCATION, MAKING IT EVEN EASIER TO USE THE APPLICATION.

Results & Usage

WITHIN OUR APPLICATION, WE ARE ABLE TO PLOT AND PRODUCE A GRAPH THAT THE USER MAY USE TO BETTER ANALYZE THE DATA SET. BELOW SHOWS A GRAPH THAT WE PLOTTED:

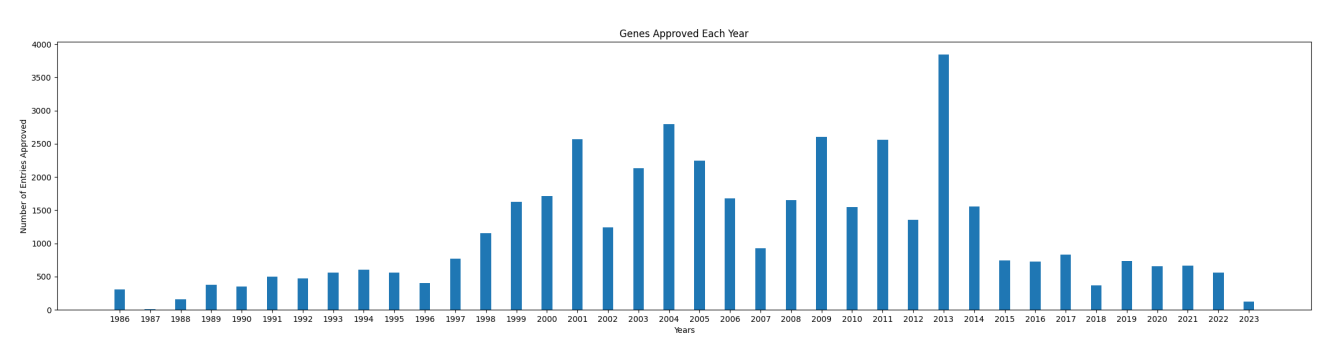

FIGURE 1. A PLOT FROM THE YOU'RE CUTE GENES APPLICATION

AS SHOWN IN THE GRAPH, THE PLOTS SHOW THE NUMBER OF GENES APPROVED EACH YEAR. THE USER CAN USE THIS INFORMATION TO SEE THE SATURATION OF GENES FOUND AT A GIVEN TIME/ YEAR. THIS COULD ALSO SUPPORT CLAIMS IN DIFFERENT ASPECTS OF BIOLOGICAL RESEARCH LIKE FINDING THE CORRELATION OF THESE DISCOVERIES AND THE TECHNOLOGICAL ADVANCEMENTS WITHIN THE BIOLOGY COMMUNITY. THE USER COULD EVEN REFERENCE THESE DIFFERENT PIECES OF ANALYSIS TO HELP GAIN A BETTER UNDERSTANDING WITHIN THEIR RESEARCH OR ACADEMIC PROJECTS. THE MAIN USAGE OF OUR RESULTS IS TO CREATE A VISUAL REPRESENTATION OF THE DATA TO ALLOW THE USERS TO GAIN A BETTER UNDERSTANDING OF THE DATA SET ITSELF.

NOT ONLY DOES THE APPLICATION CREATE AN IMAGE THAT CAN VISUALLY REPRESENT OUR DATA, THE APPLICATION CAN SUMMARIZE DIFFERENT PARTS OF THE HGNC DATA.

```
ubuntu@avlav-vm:~/FinalProj$ curl localhost:5000/locusdata
{
  "Locus Group": "Number of Entries",
  "non-coding RNA": 9056,
  "other": 992,
  "protein-coding gene": 19267,
  "pseudogene": 14351
}
```

FIGURE 2. AN EXAMPLE OF A SUMMARY OF THE DATA SET FOR LOCUS KEYS

AS SHOWN ABOVE, THE CONFIRMATION THAT THE COMPUTER SENDS TO THE USER INDICATES HOW MANY GENES ARE IN EACH TYPE OF LOCUS GROUP.

Software Design Principles

OUR APPLICATION REQUIRED SEVERAL OF THE SOFTWARE DESIGN PRINCIPLES THAT WE REVIEWED AT THE BEGINNING OF THE SEMESTER. BEGINNING WITH MODULARITY, WE ATTEMPTED TO INCREASE INTRA-MODULE COHESION WHILE REDUCING INTER-MODULE COUPLING. INTRA-MODULE COHESION REFERS TO GROUPING SIMILAR FUNCTIONS TOGETHER, WHICH WE DID BY ORGANIZING OUR FUNCTIONS INTO WORKER.PY, JOBS.PY, AND GENE_API.PY SCRIPTS.

EACH SCRIPT CONTAINS FUNCTIONS THAT ARE SIMILAR TO ONE ANOTHER, AND THE THREE SCRIPTS WORK TOGETHER TO FORM A COMPLETE SYSTEM. WE ALSO REDUCED INTER-MODULE COUPLING, WHICH IS THE DEPENDENCE OF COMPONENTS ON ONE ANOTHER, BY INCLUDING ENVIRONMENT VARIABLES THROUGHOUT THE APPLICATION, MAKING IT EASIER TO MODIFY DIFFERENT ELEMENTS INDEPENDENTLY WITHOUT BEING WORRIED ABOUT CHANGING IP ADDRESSES AND CAUSING COMPONENTS TO FAIL.

OUR APPLICATION ALSO FOCUSES ON PORTABILITY AND REPRODUCIBILITY SO THAT THE SAME RESULTS CAN BE OBTAINED AT ANY TIME ON ANY COMPUTER. USING CONTAINERIZATION, OUR APPLICATION BECAME PORTABLE BECAUSE ANY INDIVIDUAL CAN PULL THE IMAGE FROM DOCKER HUB, CLONE THE REPOSITORY, AND USE THE APPLICATION ON THEIR PERSONAL MACHINE. THEN, USING A DATABASE ALLOWED FOR REPRODUCIBILITY AS THE RESULTS GENERATED AT ONE TIME CAN BE OBTAINED LATER, EVEN AFTER RESTARTING THE APPLICATION.

Ethics

WHEN USING DATA SETS FROM OTHER SOURCES AND CODING AN APPLICATION FOR USERS TO USE, THERE ARE SEVERAL ETHICAL AND PROFESSIONAL RESPONSIBILITIES TO CONSIDER. SOME OF THESE RESPONSIBILITIES INCLUDE:

1. RESPECT FOR DATA PRIVACY: IT IS IMPORTANT TO ENSURE THAT THE DATA USED IN THE APPLICATION IS OBTAINED LEGALLY AND WITH PROPER CONSENT FROM INDIVIDUALS OR ENTITIES THAT OWN OR PROVIDE THE DATA. ADDITIONALLY, SENSITIVE DATA MUST BE KEPT CONFIDENTIAL AND SECURED TO PREVENT UNAUTHORIZED ACCESS OR MISUSE.
2. TRANSPARENCY: IT IS ESSENTIAL TO BE TRANSPARENT ABOUT THE DATA SOURCES USED AND HOW THE DATA IS BEING USED IN THE APPLICATION. USERS SHOULD BE INFORMED ABOUT THE TYPES OF DATA THAT ARE BEING COLLECTED, HOW THEY ARE BEING ANALYZED, AND WHAT THE APPLICATION INTENDS TO DO WITH THE DATA.
3. ACCURACY AND RELIABILITY: THE APPLICATION MUST ENSURE THAT THE DATA USED IS ACCURATE, RELIABLE, AND UP-TO-DATE. THE ALGORITHMS AND MODELS USED IN THE APPLICATION SHOULD BE DESIGNED TO MINIMIZE BIASES, ERRORS, OR INACCURACIES THAT MAY AFFECT THE RESULTS.
4. COMPLIANCE WITH LAWS AND REGULATIONS: THE APPLICATION MUST COMPLY WITH RELEVANT LAWS AND REGULATIONS GOVERNING DATA PROTECTION, PRIVACY, AND ETHICAL USE OF DATA.

USING DATA SETS FROM OTHER SOURCES AND CODING AN APPLICATION FOR USERS TO USE COMES WITH A SIGNIFICANT ETHICAL AND PROFESSIONAL RESPONSIBILITY. DEVELOPERS MUST ENSURE THAT THE DATA IS OBTAINED LEGALLY AND ETHICALLY, AND THAT THE APPLICATION IS DESIGNED TO BE ACCURATE, RELIABLE, FAIR, AND RESPECTFUL OF USER PRIVACY AND DATA PROTECTION LAWS.

Conclusion

IN CONCLUSION, OUR GROUP GAINED VALUABLE EXPERIENCE IN CODING FLASK APIS, SETTING UP KUBERNETES CLUSTERS, AND UTILIZING REDIS TO INTERFACE WITH THE HGNC DATA SET. WE LEARNED ABOUT THE EFFICIENCY OF QUEUES IN IMPROVING THE PERFORMANCE OF OUR API AND HOW TO USE DATA TO DEVELOP CORRELATIONS AND RELATIONSHIPS BETWEEN DIFFERENT GENES. WE ALSO GAINED INSIGHTS INTO THE REAL-WORLD APPLICATIONS OF THE DATA AND HOW ORGANIZING THE HGNC DATA SET COULD HELP RESEARCHERS IN THEIR WORK.

GIVEN UNLIMITED TIME, WE WOULD LIKE TO ADD MORE VISUALIZATIONS SUCH AS ADDITIONAL GRAPHS REPRESENTING LOCUS GROUPS, AS WELL AS ENHANCE THE HTML CONTENT OF OUR INTERFACE. WE WOULD ALSO HAVE LIKED TO DEBUG OUR JOB/WORKER SCRIPTS MORE TO GET THEM FULLY FUNCTIONING. WE RECOGNIZE THE POTENTIAL FOR FURTHER ANALYSIS OF CORRELATIONS AND RELATIONSHIPS BETWEEN GENES TO HELP SUPPORT RESEARCHERS IN THEIR OWN DATA AND RESEARCH ENDEAVORS. OVERALL, THIS PROJECT PROVIDED US WITH VALUABLE KNOWLEDGE AND SKILLS THAT WE CAN APPLY IN FUTURE PROJECTS AND IN OUR CAREERS.

References and Bibliography

BRUFORD, ELSPETH A, BRYONY BRASCHI, PAUL DENNY, TAMSIN E M JONES, RUTH L SEAL, AND SUSAN TWEEDIE. “GUIDELINES FOR HUMAN GENE NOMENCLATURE.” NATURE GENETICS. U.S. NATIONAL LIBRARY OF MEDICINE, AUGUST 2020. HTTPS://WWW.NCBI.NLM.NIH.GOV/PMC/ARTICLES/PMC7494048/.

“HOME: HUGO GENE NOMENCLATURE COMMITTEE.” HOME | HUGO GENE NOMENCLATURE COMMITTEE. ACCESSED APRIL 26, 2023. HTTPS://WWW.GENENAMES.ORG/.

“HOW ARE GENETIC CONDITIONS AND GENES NAMED?: MEDLINEPLUS GENETICS.” MEDLINEPLUS. U.S. NATIONAL LIBRARY OF MEDICINE. ACCESSED APRIL 26, 2023. HTTPS://MEDLINEPLUS.GOV/GENETICS/UNDERSTANDING/MUTATIONSANDDISORDERS/NAMING/#:~:TEXT=DURING%20THE%20RESEARCH%20PROCESS%2C%20GENES,AIDING%20THE%20ADVANCEMENT%20OF%20RESEARCH.