Project Proposal
BIOINF527
Mohamad Bairakdar, Nora Dobos, Matthew Junkin

Introduction
For our project we will be replicating the results of a paper that builds machine learning models and uses features derived from the Gene Ontology (GO) and multiple Protein-Protein Interaction (PPI) databases to classify DNA repair genes into “related to aging” and “not related to aging”. The end goal of building these models is to do feature selection and extraction in order to extract the most important features that contributed to the classification task. This would help find biologically meaningful information about what properties of DNA repair genes make them related to aging. This would help us better understand the aging process, as research has found that DNA repair genes play a non trivial role in aging.

Problem Statement
Current understanding of the roles individual DNA repair genes play in relation to one another and to aging is overall lacking if we wish to understand and study the aging process. Determining which features of DNA repair genes are predictive of their role as aging-related genes is an enormous task given the large number of both DNA repair genes and their features.

Goal of Project
We will be attempting to replicate the methods used by Fang et al. (2013) [1] using more recently updated genomics and proteomics data. We will compare our results with those found by Fang et al., which may lead to a better characterization of DNA repair genes related to aging.

Resources (software tools, datasets, etc)
- **Scikit learn Python Library**
  - Use library implementations of different ML algorithms.
- **GenAge database**
  - Extract list of aging related genes.
- **bioGRID database**
  - Gather PPI features for genes extracted from GenAge.
- **Gene Ontology (GO) Resource**
  - Gather GO features for genes extracted from GenAge.
- **UniProt (in place of Human Protein Ref. Database)**
  - Gather PPI features for genes extracted from GenAge.

- **We might also use Ingenuity Pathway Analysis (IPA) or a similar piece of software**
    - May be used to investigate the biological pathways of the attributes that were chosen by feature selection/extraction

Approach

We will be training four machine learning models from the Scikit-learn python library: Naive Bayes (NB), Decision Tree, Support Vector Machine (SVM), and Random Forest (RF). These models will be trained on data containing GO and PPI attributes of DNA repair genes derived from the bioGRID, GO, and UniProt databases. We will evaluate the performance of these models using cross-validation and will optimize for the following performance measures: AUC, MCC, and accuracy. Additionally, we will extract which GO and PPI attributes are the greatest predictors of a DNA repair gene being related to aging. We will then train whichever models lead to the best performance using only these selected features, and we will compare performance with the previous full models using the same performance measures. If we are able to access the IPA software tool, we will use it to analyze the biological functions related to our extracted attributes.

Resource Links/Keys

[1] Fang, Yaping, et al. "Classifying Aging Genes into DNA Repair or Non-DNA Repair-Related Categories." *Intelligent Computing Theories and Technology*, vol. 7996, Springer Berlin Heidelberg, 2013, pp. 20–29, doi:10.1007/978-3-642-39482-9_3.

[2] **BioGRID Key:** facf9868e3e0b115d8bf8e673a4c3c2f

Classifying human aging genes into DNA repair and non DNA repair

Mohamad Bairakdar, Nora Dobos, Matthew Junkin

Abstract

This work aims to reproduce the workflow outlined by Fang et al. to classify aging genes as DNA repair versus non DNA repair genes using updated data. The end goal was not to perform classification on new instances, but instead to perform feature selection in order to better understand what attributes of aging genes make them DNA repair vs non DNA repair related. We employ multiple machine learning algorithms to this end, and use GO terms and PPI attributes as features to characterize genes, which act as input to the algorithms. We obtain similar selected features to those obtained by Fang et al.

1 Introduction

Our understanding of the process of human aging is predicted to be much better understood through computational approaches like machine learning (De Magalhães et al.). Increasingly more studies are suggesting that DNA repair genes play an important role in the aging process (Fang et al.). We reworked the approach used by Fang et al. in using aging related genes to create machine learning models to distinguish between DNA repair and non DNA repair aging genes. The models were trained on features derived from the Gene Ontology (GO) and GenAge databases. Feature selection from these models was used to determine which features contribute most significantly to the task of classification.

2 Methods

2.1 Gene List Collection

A list of 307 aging-related genes were downloaded from the GenAge database. Of these genes, 40 were identified as related to DNA repair by Wood et al., and were labeled accordingly.

2.2 GO Attributes

For each gene, its set of GO attributes was collected. Only biological process GO terms were considered. We then merged all these attributes, totalling 3556 GO terms, and considered them as binary features for the genes based on whether or not a gene was associated with a given GO term.

2.3 PPI Attributes

Protein-Protein Interaction pair data was downloaded from the Biological General Repository for Interaction Datasets (BIOGRID, http://thebiogrid.org/). A PPI was considered if at least one protein in the pair was a product of a human aging-related gene. A total of 13857 PPI attributes were collected, and were considered as binary features for the genes based on whether or not a gene interacted with a given gene product.

2.4 Data Mining Algorithms

Four machine learning algorithms were used from the SciKit Learn Python library: Decision Tree (DT), Multinomial Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM). For the RF and DT, Gini importance measure was used to evaluate the importance of features.

2.5 Performance Evaluation

Standard 5-fold cross validation was used to test performance of the different classifiers. Performance was measured using several metrics: AUROC (Area under the receiver operating curve), MCC (Matthew's correlation coefficient), and ACC (Accuracy).

3 Results

3.1 Improving Performance of Different Classification Algorithms through GO and PPI Term Cutoffs

Given that the feature matrix was sparse, with most GO and PPI attributes being associated with only a few genes, we evaluated different minimum threshold cutoff values to filter out features. For GO terms, we tried cutoff values between 1 and 20 terms. For example, in the case of a cutoff of 1, if a GO term was associated with only one gene, it was filtered out. For PPI attributes, we tried values between 1 and 200 in increments of 10. Employing these different filters, we obtained the AUC for the four algorithms, with the results displayed in the graph below. As seen in Figure 1, a cutoff of 1 through 4 was the best-performing cutoff for GO terms, achieved by the SVM, and a cutoff of 90 was the ideal cutoff for PPI features, achieved by NB. Surprisingly, the best performing algorithms in Fang et al were actually the RF and DT classifiers. In order to compare the results of the remaining work done by Fang et al with our work, we decided to use the RF and DT, despite the lower performance. Hence, we chose cutoffs based on the performances of these models. For the GO terms, we found that a cutoff of 9 GO terms resulted in a reasonable AUC value for both the RF and DT models and for PPI terms, the cutoff was 130. However, after examining the number of PPI terms that satisfied this cutoff, we found that none did. A vast majority of PPI terms satisfied a cutoff of 50, so we limited our cutoff value to be 50 or less. Hence, we chose a cutoff of 10, since it offered a reasonable AUC for both the RF and DT.

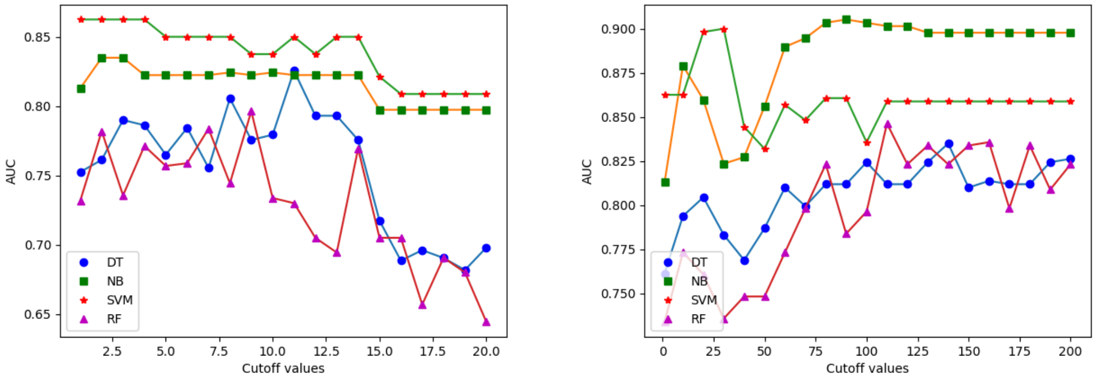

Figure 1: influence of cutoff values on AUC for the different algorithms. Cutoff applied to GO terms (left) and PPI terms (right).

3.2 Feature Selection And Performance Using Selected Features Only

We then performed feature selection since the above graphs indicated that not necessarily all, but perhaps only subsets of the features were important for classification. Thus, for the RT, we obtained the top 25 most predictive features ranked by Gini index. For the DT, we obtained all attributes that had non zero feature importance, also according to Gini index. There were 14 of these attributes (both sets of top attributes are listed in Appendix B). We then performed 5-fold cross validation to obtain AUC and ACC for both the DT and RF trained only on selected features. Indeed, this confirmed that only a subset of features is necessary for the task, as indicated by the high values for the performance metrics.

Table 1: Performance of selected features using DT and RF models

		Decision Tree		Random Forest	
Feature set	# Attributes	ACC	AUC	ACC	AUC
Top25	25	0.9344	0.8349	0.9508	0.8655
Top14	14	0.9770	0.9655	0.9771	0.9549

3.3 Performance Using Only PPI Attributes

We looked up the GO terms among the selected features, and we noticed that a lot of them are related to DNA repair. As done in Fang et al., to make sure that these were not the only important factor in the classification task (which is relevant to check since these terms describe exactly the function that we are trying to predict), we trained RF models on subsets of only PPI attributes, ranked using the Gini index. The performance measures for these models are shown in Fig. 2. Our performance measures were quite low relative to the measures obtained by Fang et al. For 15 PPI attributes, which we chose to examine because it was the number of attributes that led to the best performance in Fang et al., our ACC was 0.879, AUC was 0.665, and MCC was 0.373.

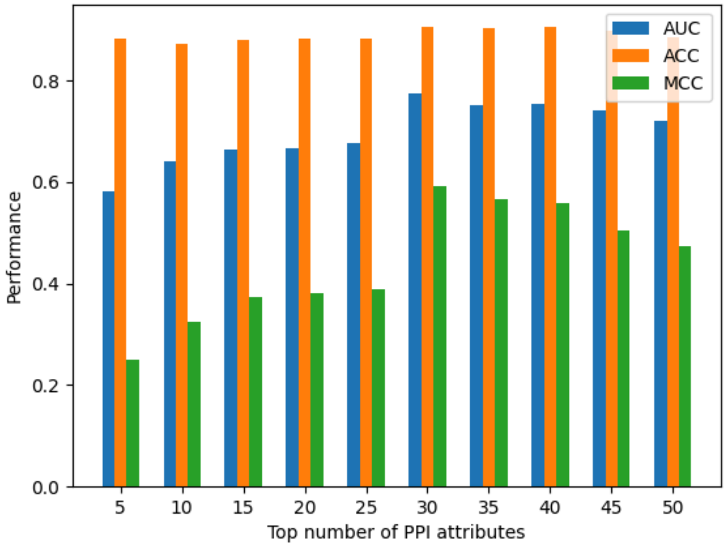

Figure 2: Performance of RF models using subsets of PPI attributes ranked by Gini index

Table 2 shows the top 15 selected attributes, along with their subcellular location. Similar to the PPIs identified by Fang et al., most of our proteins are located in the nucleus, which the authors argue makes sense because of their probable role as products of DNA-repair genes.

Table 2: Top 15 PPI attributes and their respective locations

Entrez ID	Symbol	Entrez Protein Name	Location(s)
1019	CDK4	Cyclin dependent kinase 4	Nucleus, cytoplasm
6884	TAF13	TATA-box binding protein associated factor 13	Nucleus
6118	RPA2	Replication protein A2	Nucleus
55215	FANCI	FA complementation group I	Nucleus
23293	SMG6	SMG6 nonsense mediated mRNA decay factor	Nucleus, Cytosol, Telomere
7532	YWHAG	Tryptophan 5-monooxygenase activation protein gamma	Cytoplasm
26993	AKAP8L	A-kinase anchoring protein 8 like	Nucleus, Cytoplasm
3732	CD82	CD82 molecule	Plasma membrane
5781	PTPN11	Protein tyrosine phosphatase non-receptor type 11	Nucleus, Cytoplasm
6449	SGTA	Small glutamine rich tetratricopeptide repeat containing alpha	Nucleus, Cytoplasm
26517	TIMM13	translocase of inner mitochondrial membrane 13	Mitochondrion
1523	CUX1	Cut like homeobox 1	Nucleus
1877	E4F1	E4F transcription factor 1	Nucleus

81704	DOCK8	Dedicator of cytokinesis 8	Cell membrane, cytoplasm
898	CCNE1	Cyclin E1	Nucleus

4 Discussion

Comparing the top 28 proteins selected by our DT and RF models combined, and comparing those to the top 18 proteins selected by the models in Fang et al., we noticed that most proteins were different between the two. The common genes were HSPA4 and XRCC1. We also found that RPA2 and RPA3 were selected by our models, while RPA1 was selected by the models in Fang et al. The fact that only a few genes overlapped made us somewhat question how sound our model was. After examining the genes that were not in common, we noticed that actually many of them had very similar functions, which made us more confident in our models. For example, Fang et al.'s models selected BLM, ATM, WRN, MSH2, MRE11A, ERCC6, PCNA, and ATR, which are all related to cell cycle regulation. Our models selected CDK4, TAF13, AKAP8L, CUX1, CDK9, CDK2, CDC42, E4F1 all of which are related to cell cycle regulation. Similarly, Fang et al. had a class of proteins related to DNA repair: BLM, WRN, MSH2, XRCC1, RPA1, MRE11A, ATM, ATR, PCNA, ERCC6, ERCC2 and XAB2. Our models also selected a class of proteins related to DNA repair: RPA2, FANCI, SMG6, NBN, XRCC1, PRKDC, RAD50. Fang et al. had a class of proteins related to apoptosis of tumor cell lines: BLM, ATM, RPA1, PCNA and HSPA4. Our models selected a class of genes related to tumor suppression: CD82, HSPA4. We present a full list of the genes selected by our models in Appendix A, each provided with a brief description of the gene and its major role indicated. As noted previously, the GO terms selected by our models and Fang et al.'s models are mostly related to DNA repair. The 4 GO terms with the highest importance from our models were: GO:0033683 (nucleotide-excision repair, DNA incision), GO:0006284 (base-excision repair), GO:0006289 (nucleotide-excision repair), and GO:0006302 (double-strand break repair).

5 Conclusions

Overall, our models had lower performance than Fang et al.'s, despite being trained on more data, but the features selected by both of our models were similar in nature. Our model may have performed worse due to multiple reasons. First, the implementations of the machine learning models used were different from Fang et al.'s. Moreover, although we had a few more training examples, our feature set expanded considerably compared to Fang et al.'s, which might have made our models more prone to overfitting. When the feature set was restricted, we noticed that our performance in some cases was very close to the performance obtained by Fang et al.; the similarities can be emphasized by comparing table 1 in our report to table 1 in Fang et al.

References

Cock *et al*, Biopython: freely available Python tools for computational molecular biology and bioinformatics, *Bioinformatics*, Volume 25, Issue 11, 1 June 2009, Pages 1422–1423, https://doi.org/10.1093/bioinformatics/btp163

De Magalhães, Joao Pedro, and Olivier Toussaint. "How Bioinformatics Can Help Reverse Engineer Human Aging." *Ageing Research Reviews*, U.S. National Library of Medicine, 2004, pubmed.ncbi.nlm.nih.gov/15177050/.

Fang, Yaping, et al. "Classifying Aging Genes into DNA Repair or Non-DNA Repair-Related Categories." *Intelligent Computing Theories and Technology Lecture Notes in Computer Science*, 2013, pp. 20–29., doi:10.1007/978-3-642-39482-9_3.

Pedregosa *et al*, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830 (2011) (publisher link)

Tacutu *et al*. "Human Ageing Genomic Resources: new and updated databases." *Nucleic Acids Research* 2018, 46(D1):D1083-D1090.

The UniProt Consortium, UniProt: a worldwide hub of protein knowledge, *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D506–D515, https://doi.org/10.1093/nar/gky1049

Wood, RD., et al. "Human DNA Repair Genes." . , edited by R. Wood and M. Lowery, 10 June 2020, www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-genes.html.

Appendix A

Selected Attributes:

CDK4	Phosphorylate and inhibit members of the retinoblastoma (RB) protein family including RB1 and regulate the cell-cycle during G_1/S transition
TAF13	Plays a critical role in the regulation of gene transcription in eukaryotic cells
RPA2+	Binds and stabilizes single-stranded DNA intermediates, that form during DNA replication or upon DNA stress, it plays an essential role both in DNA replication and the cellular response to DNA damage. (alt.)This gene encodes a subunit of the heterotrimeric Replication Protein A (RPA) complex, which binds to single-stranded DNA (ssDNA), forming a nucleoprotein complex that plays an important role in DNA metabolism, being involved in DNA replication, repair, recombination, telomere maintenance, and co-ordinating the cellular response to DNA damage through activation of the ataxia telangiectasia and Rad3-related protein (ATR) kinase
FANCI	Plays an essential role in the repair of DNA double-strand breaks by homologous recombination and in the repair of interstrand DNA cross-links (ICLs) by promoting FANCD2 monoubiquitination by FANCL and participating in recruitment to DNA repair sites.
SMG6	Component of the telomerase ribonucleoprotein (RNP) complex that is essential for the replication of chromosome termini. May have a general role in telomere regulation. Promotes in vitro the ability of TERT to elongate telomeres
YWHAG	Adapter protein implicated in the regulation of a large spectrum of both general and specialized signaling pathways. Binds to a large number of partners, usually by recognition of a phosphoserine or phosphothreonine motif. Binding generally results in the **modulation** of the activity of the binding partner.
AKAP8L	Could play a role in constitutive transport element (CTE)-mediated gene expression by association with DHX9. Increases CTE-dependent nuclear unspliced mRNA export May be involved in anchoring nuclear membranes to chromatin in interphase and in releasing membranes from chromating at mitosis. May regulate the initiation phase of DNA replication when associated with TMPO isoform Beta . Required for cell cycle G2/M transition and histone deacetylation during mitosis. In mitotic cells recruits HDAC3 to the vicinity of chromatin leading to deacetylation and subsequent phosphorylation at 'Ser-10' of histone H3; in this function seems to act redundantly with AKAP8.
CD82	Associates with CD4 or CD8 and delivers costimulatory signals for the TCR/CD3 pathway.
PTPN11	Acts downstream of various receptor and cytoplasmic protein tyrosine kinases to participate in the signal transduction from the cell surface to the nucleus
SGTA	Co-chaperone that binds misfolded and hydrophobic patches-containing client proteins in the cytosol. Mediates their targeting to the endoplasmic reticulum but also regulates their sorting to the proteasome when targeting fails
TIMM13	Mitochondrial intermembrane chaperone that participates in the import and insertion of some multi-pass transmembrane proteins into the mitochondrial inner membrane
CUX1	Plays a role in cell cycle progression, in particular at the G1/S transition
E4F1	May function as a transcriptional repressor. May also function as a ubiquitin ligase mediating ubiquitination of chromatin-associated TP53. Functions in cell survival and proliferation through control of the cell cycle. Functions in the p53 and pRB tumor suppressor pathways and regulates the cyclin CCNA2 transcription.
DOCK8	Guanine nucleotide exchange factor (GEF) which specifically activates small GTPase CDC42 by exchanging bound GDP for free GTP

NBN The encoded protein is a member of the MRE11/RAD50 double-strand break repair complex which consists of 5 proteins. This gene product is thought to be involved in DNA double-strand break repair and DNA damage-induced checkpoint activation.

XRCC1* The protein encoded by this gene is involved in the efficient repair of DNA single-strand breaks formed by exposure to ionizing radiation and alkylating agents

CDK9 Protein kinase involved in the regulation of transcription, works to avoid DNA damage by limiting certain parts of the cellular cycle.

CUL2 Ubiquitous expression in testis (RPKM 12.7), brain (RPKM 7.3) and 25 other tissues IDK

AP1B1 Adaptor protein complex 1 is found at the cytoplasmic face of coated vesicles located at the Golgi complex, where it mediates both the recruitment of clathrin to the membrane and the recognition of sorting signals within the cytosolic tails of transmembrane receptors.

HSPA4* Ubiquitous expression in testis (RPKM 45.7), esophagus (RPKM 33.4) and 25 other tissues “Stress response protein released when exposed to heat” (might protect something, but not clear)

NKX2-1 This gene encodes a protein initially identified as a thyroid-specific transcription factor. The encoded protein binds to the thyroglobulin promoter and regulates the expression of thyroid-specific genes but has also been shown to regulate the expression of genes involved in morphogenesis.

PRKDC This gene encodes the catalytic subunit of the DNA-dependent protein kinase (DNA-PK). It functions with the Ku70/Ku80 heterodimer protein in DNA double strand break repair and recombination.

RAD50 This protein, cooperating with its partners, is important for DNA double-strand break repair, cell cycle checkpoint activation, telomere maintenance, and meiotic recombination.

CDK2 This gene encodes a member of a family of serine/threonine protein kinases that participate in cell cycle regulation. The encoded protein is the catalytic subunit of the cyclin-dependent protein kinase complex, which regulates progression through the cell cycle. Activity of this protein is especially critical during the G1 to S phase transition.

CSK The protein encoded by this gene is involved in multiple pathways, including the regulation of Src family kinases.

CDC42 The protein encoded by this gene is a small GTPase of the Rho-subfamily, which regulates signaling pathways that control diverse cellular functions including cell morphology, migration, endocytosis and cell cycle progression

CREBBP This gene is ubiquitously expressed and is involved in the transcriptional coactivation of many different transcription factors.

RPA3 As part of the heterotrimeric replication protein A complex (RPA/RP-A), binds and stabilizes single-stranded DNA intermediates that form during DNA replication or upon DNA stress. It prevents their reannealing and in parallel, recruits and activates different proteins and complexes involved in DNA metabolism. Thereby, it plays an essential role both in DNA replication and the cellular response to DNA damage

----- Blatantly related to DNA/Telomere repair

----- Related to the cellular cycle

----- Modulation/expression of other proteins

----- Cell cycle checkpoints

----- Tumor suppressor

Appendix B

Top 25:
'GO:0033683', '4683', '7515', 'GO:0006284', '6118', '1025', '8453', 'GO:0006289', '162', '3308', '7080', 'GO:0006302', '5591', 'GO:0006283', '10111', 'GO:0006281', '1017', 'GO:0006295', '1445', 'GO:0006293', '998', '1387', 'GO:0000723', 'GO:0006294', '6119'

TopDT:
'GO:0046686', 'GO:0000724', 'GO:0006302', 'GO:0006303', 'GO:0010332', 'GO:0016579', 'GO:0006283', '6672', '5424', '5499', '10980', '1080', '3163', '55218'