Norwegian University
of Life Sciences

Master's Thesis 2021   30 ECTS
Faculty of Chemistry, Biotechnology and Food Science

Exploring Regulatory Evolution After Whole Genome Duplication Using Machine Learning

Tobias Bjørn
Master's Degree Programme in Chemistry and Biotechnology

Table of Contents

Abstract

The Atlantic salmon underwent a whole-genome duplication 80 million years ago and has kept around half of the duplicated genes. Over time, some genes have become more active, while others have become less active, due to regulatory changes. This thesis explores if it is possible to separate these genes by the number of nearby transcription factor binding sites.

With previously obtained information about the binding sites for different transcription factors for each gene and the direction of the expression level shift for this gene, a matrix was constructed containing the difference in bound transcription factor binding sites between the gene duplicates. One of the gene duplicates has a significant change in gene expression level, while the other is conserved. The duplicate pairs with increased expression in one copy are called upcons, and the pairs with decreased expression in one copy are called downcons.

Multiple machine learning algorithms were tested to classify upcons vs downcons. Overall, support vector machines performed best, achieving an accuracy of 67%.

In conclusion, the results are indicative that classification of the evolutionary direction of genes based on nearby transcription factor binding sites can be done.

Sammendrag

Atlanterhavslaken gjennomgikk en helgenomduplisering for 80 millioner år siden og har beholdt rundt halvparten av de dupliserte genene. Noen gener har blitt mer avlest, mens andre har blitt avlest sjeldnere, dette grunnet endringer i reguleringen. Denne oppgaven vil undersøke om det er mulig å skille slike gener etter antallet bindingsseter transkripsjonsfaktorer har i nærheten av genene.

Med tidligere innhentet informasjon om bundne bindingsseter for forskjellige transkripsjonsfaktorer for hvert gen og retningen på endringen for genuttrykksnivået for genet, ble en matrise laget som inneholdt forskjellen i bindingsseter mellom duplikatgenene. Det ene genet i duplikatparet har en signifikant endring i genuttrykksnivå, men ikke det andre. De parene hvor endringen er positiv, kalles «upcons», og de negative kalles «downcons».

Flere maskinlæringsmetoder var testet i klassifikasjonen av «upcons» og «downcons». SVM var den metoden som gjorde det best. Den klarte å velge riktig i 67% av tilfellene.

Konklusjonen er at det er gjennomførbart å klassifisere geners evolusjonære retning basert på transkripsjonsfaktorers bindingsseter.

# Acknowledgement

This thesis was written as a part of my master's degree at the Norwegian University of Life Sciences.

I want to thank my supervisor Torgeir R Hvidsten, and the BIAS group, for valuable help and discussions, and my co-supervisor Simen Sandve for the data.

I want to thank my parents for their support.

Tobias Bjørn
Ås, 2020

Introduction

The salmonids underwent a whole-genome duplication around 80 million years ago. After a whole-genome duplication every gene exist as two copies, along with their regulatory elements, these copies are referred to as gene duplicates. Today, Atlantic salmon (*Salmo salar*) has retained around half of the duplicated genes. It is interesting to know how the genes evolved to solve the challenges posed by a doubled genome. How did these genes evolve as to overcome the fitness costs and become a successful polyploid species?

This thesis is a continuation of the work done by Gillard *et al.*[1], where they looked at the evolutionary shifts of gene duplicates. They compared the gene expression level of the gene duplicates to the gene expression level of orthologs in species without the salmonid-specific whole-genome duplication. While most genes had a conserved gene expression level between salmonids and non-salmonids, some genes had a significantly different expression level between the gene duplicates.

The transcription of genes is regulated by the presence of transcription factors (TF). TFs can bind both in the promoter region and in enhancers far away from the gene. Bound transcription factors recruit the transcriptional machinery. Transcription factors bind to specific binding sites in the DNA sequence which are unique to each TF. TFs can bind at sites that deviate slightly from their preferred binding site. This behavior can be characterized by a position weight matrix, which is often called the motif. These transcription factor binding sites (TFBS) can be identified by immunoprecipitation techniques, and the motifs are found by aligning the results.

The gene expression level is dependent on the number of bound TFBS. A higher number leads to higher gene activity.

This study explores the transcription factor binding sites surrounding genes and whether the number of TFBS for different TFs can explain the evolutionary shift in gene expression level between gene duplicates. The scope of the study is limited to looking at upregulation versus downregulation in gene duplicates where one copy is conserved.

Background

ATAC-footprinting

Assay of Transposase Accessible Chromatin Sequencing (ATAC-seq) is a technique which can identify the regions of the chromosomes that are accessible for transcription, also called open chromatin. To do so, ATAC-seq cuts DNA using Tn5 transposase, which only cut protein-free DNA. Each fragment is sequenced and aligned back to the genome. The accessible regions are then found by peak calling, finding the peaks in the read count from alignment. [2]

Interactions between transcription factors and genes can found by mapping motifs to the open chromatin of the genome, with the assumption that TFs regulate the closest gene. It is therefore necessary to impose a limit on the maximum distance between TFBS and gene.

Another popular approach to identify open chromatin is DNase-seq (DNase I hypersensitive sites sequencing), which is far costlier than ATAC-seq. Most computational tools are designed for DNase-seq data, and don't necessarily work as expected for ATAC-seq data. [2]

However, recently a computational framework designed for ATAC-seq data was developed, called TOBIAS. [3]

TOBIAS Method

TOBIAS (**T**ranscription factor **O**ccupancy prediction **b**y **I**nterference of **A**TAC-seq **S**ignal) is a comprehensive framework for footprinting analysis. It takes as input TF motifs, ATAC-seq data and annotated sequences, and outputs the positions of the TFBS, whether they are bound or not and their distance to the transcription start site of genes. [3]

The Tn5 enzyme is blocked by proteins bound to DNA like histones and TFs, which results in depletion in the ATAC-seq signal. An area of low signal in a larger region of high signal is called a footprint. Since Tn5 has a sequence preference, some motifs which disfavors Tn5 interaction will give false positives. The Tn5 signal might also hide some bound TFBS, causing false negatives. So, the first thing TOBIAS does is to correct the ATAC-seq data for Tn5 bias by removing the expected Tn5 cut sites from the signal. [3]

While the same TFBS will be present in every cell, the TF itself might not be expressed, leaving the TFBS unbound in that cell. TOBIAS determines if a TFBS is bound by looking at the footprint depth. If Tn5 is blocked by a bound TF, there won't be a read peak there. [3]

Expression Variance and Evolution (EVE) Model

The EVE model describes both phylogenetic evolution and expression variance in populations. It uses an Ornstein-Uhlenbeck (OU) process to model stabilizing selection. In contrast to Brownian motion, which is usual for modelling genetic drift, OU processes are constrained around an optimal value θ, parameterized over the strength of drift, σ^2, and the strength of the pull, α, towards that optimal value. [4]

The EVE model uses the parameter β to represent the ratio of between- and within-species variation. For each gene, there should be a linear relationship between the evolutionary expression level variance, defined by $\frac{\sigma_i^2}{2\alpha_i}$, and the population expression level variance, defined by $\beta_i \frac{\sigma_i^2}{2\alpha_i}$. Without selection, the value of β should be approximately the same for all genes. [4]

The EVE model can also be used to test for branch-specific expression level shifts. That is if the OU process for gene i has a different optimal value in one lineage than in the others. This way a hypothesis can be formulated by comparing the likelihood under $H_0: \theta_i^a = \theta_i^o$ versus $H_a: \theta_i^a \neq \theta_i^o$, where θ_i^a is the optimal value for gene i in the lineage of interest and θ_i^o is the optimal value for gene i in the other lineages. The resulting likelihood ratio test statistic is chi-squared distributed with one degree of freedom. The sign of $\theta_i^a - \theta_i^o$ gives the direction of the shift. [4]

Machine Learning Methods

Machine learning is about finding patterns in the data, it is a broad term that is usually divided into supervised and unsupervised learning. Supervised learning is about feeding the algorithm with many examples of how the data is distributed. This is usually for solving classification and regression tasks. With unsupervised learning, the algorithm tries to find the underlying structure of the data, usually meaning clustering methods, also including principal component analysis. Other machine learning methods include neural networks. [5]

Different machine learning methods have different strengths and weaknesses, here the following methods are tried.

Support Vector Machines

Support Vector Machines (SVM) tries to find the optimal hyperplane that separates the classes in the data. If the problem is not linearly solvable, the SVM can utilize a kernel function to transform the data into higher dimensional space where a hyperplane differentiates the classes neatly.

The optimal hyperplane is the one with the maximal distance to the support vectors, also called the margin. Support vectors are the data points closest to the decision boundary. Thus, the support vectors heavily influence the position and orientation of the hyperplane. [6]

SVM is parameterized by the cost function. The cost is a regularization parameter which allows for misclassifications in the training data, so that the hyperplane is less suspectable to outliers. With higher cost, the greater the hurdle to allow misclassification gets. [6]

The radial basis kernel also has the gamma parameter. The gamma parameter determines the influence of the training examples, with a low gamma, data points far apart can be considered similar, while a large gamma needs the points to be closer. In technical terms this refers to the width of the peaks of the hypersurface in feature space. [6]

In the R implementation, the default value for the cost is 1 and the gamma is the inverse number of features.

Random Forests

Random Forests (RF) is an ensemble of decision trees. Each tree is grown from a random subset of the training data, and all have one vote, with the most popular class being the prediction. Decision trees work by recursively partitioning the data such that the leaf nodes contain a single class. This is achieved by maximizing the Gini index, the purity of class labels at the node, at each split. [7]

Multiple trees are grown from random subsets of the data, a process called bagging. Since only some features are used for each tree, the forest can estimate the importance of each feature. [7]

Random Forests also reports the out-of-bag error. Each tree is tested on the training examples not used to construct the tree, and the average error of every tree is the estimated error of the forest.

Random Forests is parameterized by the number of trees to grow and the number of features to try at each split. In the R implementation, the default values are 500 trees and the square root of the number of features, rounded down.

K-Nearest Neighbor

K-nearest neighbor (KNN) is a technique that classifies new data by its distance to all the training data points, deciding on the most popular class among the neighbors, k representing the number of neighbors to consider. [8]

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) calculates the probability of an observation belonging to class A using Bayes theorem. It assumes that the data is normally distributed and estimates the mean and variance from the training data. When more than one explanatory variable is used, the distribution is multivariate normal, and the covariance matrix is computed instead. The class distribution is usually used as the prior. Thus, it classifies new data as the most likely class. [8]

Hierarchical Clustering

Clustering uses a distance matrix to cluster features that are similar. The distance between clusters can be computed in different ways. Single linkage is the shortest distance between two clusters, complete linkage is the longest distance between two clusters and average linkage is the average distance between all points, one from each cluster, in the two clusters.

Clustering is deterministic, meaning that the same dendrogram will be produced each time, given the same data. By cutting the dendrogram at a given height, one can discover groups in the data.

Model Evaluation

The data set is divided into test and training sets. The training set is the data used to fit the model and therefore also have class labels, and the test set, which is used to validate the model, also has class labels but is treated as unlabeled new data, which the model has not seen before. The true labels can then be compared to the predicted labels, for instance in the form of a confusion matrix.

Overfitting

Overfitting occurs when a model has more features than observations. The model will learn random patterns in training data that do not generalize well to unseen data.

Model A is overfitted if there is a model B which does better on test data but worse on training data than model A.

Cross Validation

Cross validation (CV) is used to check performance of a model on unseen test data. The dataset is split into multiple smaller sets, one is used as the test set and the rest as the training set. Each observation is in the test set once and in the training set the remaining times. This is called k-fold cross validation. Alternatively, a single observation can be used for validation each time, called leave-one-out cross validation.

Evaluation Statistics

Table 1 The confusion matrix

	Actual Positives, AP	Actual Negatives, AN
Predicted Positives, PP	True Positives, TP	False Positives, FP
Predicted Negatives, PN	False Negatives, FN	True Negatives, TN

Many performance metrics are based on the confusion matrix presented in Table 1.

Accuracy is the proportion of correct predictions. Sensitivity, also known as recall, is the proportion of true positives among actual positives. Specificity is the proportion of true negatives among actual negatives. Precision is the proportion of true positives among predicted positives. The F1 Score is the harmonic mean of precision and recall. Matthew's correlation coefficient is the difference between the product of true positives and true negatives and the product of false positives and false negatives, divided by the square root of the product of actual positives, actual negatives, predicted positives and predicted negatives. [9]

Gene Ontology

Gene ontology (GO) is a comprehensive network of gene annotations. It is divided in three ontologies: biological process, molecular function, and cellular component. Biological process refers to the goals of the cell and are processes that are accomplished by multiple molecular activities. Molecular function is for the actual enzymatic function and explains what happens on a molecular level. Cellular component refers to physical location in the cell. The network is a directed graph with each of the ontologies as the top node. The child nodes are specifications of the parent node. [10, 11]

Gene ontology can be used to check if a gene set of interest is enriched for some GO term. That is if the genes in the set, more often are annotated by the same GO term than what one would expect by chance.

GO Enrichment Analysis

The type of enrichment analysis is usually divided into Function Enrichment Analysis (FEA) and Gene Set Enrichment Analysis (GSEA).

GSEA uses a list L of all genes, ranked by some metric, and checks if the gene set S is enriched at the top or bottom of list L. It computes an Enrichment Score (ES) by walking down the list L, increasing ES whenever it encounters a gene in S or decreasing when not in S. By randomizing the classes, the p-value is the fraction of randomized lists with a higher ES than the real data.

FEA looks at whether a gene set of interest is enriched for a functional category. To test this, it uses the hypergeometric test, or Fisher's exact test, to test if there is a significant overlap between the gene set and the set of genes annotated with this functional category.

Evolutionary Fates

After whole-genome duplication, there are three different fates for gene duplicates that are retained. The first is that one copy retains the ancestral gene function while the other is allowed to accumulate mutations and achieve a novel gene function, called neofunctionalization. Second is that the ancestral gene function is split between the two, called subfunctionalization. Third is that both retains the ancestral gene function but that the absolute gene dosage is reduced as to maintain the ancestral dosage. Most gene duplicates are not retained and undergo a slow pseudogenization. Pseudogenization is when a gene accumulates so many deleterious mutations that it can no longer be transcribed and translated, or the protein is no longer functional.

Methods

Dataset Preparation

Gillard *et al*.[1] used the TOBIAS method to find bound transcription factors in the liver of Atlantic salmon.

Gillard *et al*.[1] also used the EVE model to study the evolution of gene expression between salmonids and non-salmonid fish.

Data Analysis

All data analyses were done in R, version 4.0.2[12], using the Orion compute cluster at NMBU.

The result table from TOBIAS was a huge dataset, where only a subset was of interest. It was filtered such that only motifs expressed in liver within a 20 kilobase window around the transcription start site were kept. Furthermore, only genes also used in the EVE analysis were used.

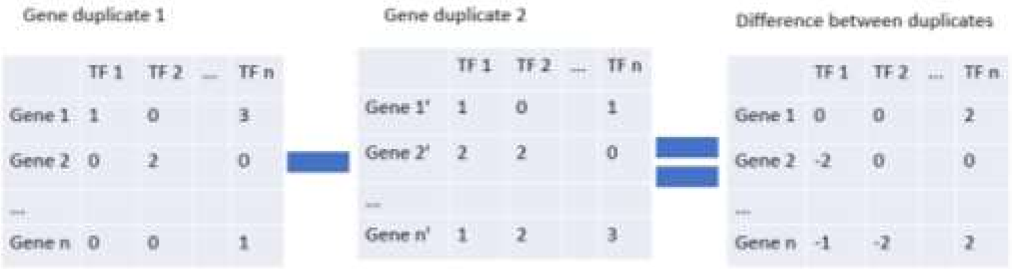
The resulting table were summarized as a list of tables. One table for each motif, with genes in one column and the corresponding number of times the motif was near this gene in the other. By iterating through the list, a data matrix was constructed. This matrix has genes as rows and TF motifs as columns. It was constructed in the following manner: for each TF, if the gene were an element of the set, return the number of bound motifs, else return zero. This matrix had dimensions 10360 x 746.

Information about motifs within transposable elements, was taken from Lien *et al.*[13]. This was the same type of table as the result from TOBIAS and was processed the same way, resulting in a matrix of the same dimensions, and was column-bound to the other matrix.

The result table from EVE was filtered such that only duplicate pairs, where both genes were present in the data matrix, were kept. Furthermore, two subsets of this table were made, 1) one with duplicate pairs where one gene was upregulated and the other was conserved, called upcons, and 2) one with duplicate pairs where one gene was downregulated and the other was conserved, called downcons. 378 pairs of upcons and 1100 pairs of downcons.

Now, the data matrix was subset into four smaller matrices. 1) One with upregulated genes, 2) one with conserved upcons, 3) one with downregulated genes, and 4) one with conserved downcons. Concurrently, transformed versions were made by applying the R function *as.logical* to each column. Resulting in a binary-valued matrix where zeros are treated like zeros and nonzero numbers like ones.

Then the contrasts were found, defined as the difference between the up- /downregulated copy and the conserved copy, as illustrated in Figure 1. The differences between the transformed matrices were also found. Positive numbers mean that there are motifs in the up/down copy that are not present in the conserved copy, while negative numbers mean there are motifs in the conserved copy that are not present in the up/down copy, and zeros mean no change.



	TF 1	TF 2	...	TF n
Gene 1	1	0		3
Gene 2	0	2		0
...				
Gene n	0	0		1

	TF 1	TF 2	...	TF n
Gene 1'	1	0		1
Gene 2'	2	2		0
...				
Gene n'	1	2		3

	TF 1	TF 2	...	TF n
Gene 1	0	0		2
Gene 2	-2	0		0
...				
Gene n	-1	-2		2

Figure 1 Illustration of how the contrast was computed.

To get an equal number of observations for both classes, random subsampling of the downcons was done.

Since there is twice as many features as observations, feature selection must be done before model fitting. Hierarchical clustering was chosen, using the R function *hclust*. The distance metric was 1 – correlation between the columns of the data matrix, and clustering was done with average linkage. The dendrogram was cut at height 0.7, and one TF was chosen at random from each cluster, resulting in 297 chosen features. The random seed was set to “123”.

Feature selection by appearance shift were also tried. By looking at the column sums of the transformed matrices, that is how often a motif has appeared or disappeared between duplicates, one can choose the motifs with the biggest shifts in appearance. Among upcons, the motifs with column sums above 40 were kept, while among downcons, the motifs with column sums below -100 were kept. The intersection of these were chosen, resulting in 26 selected TFs.

The machine learning methods used were Random Forest by the R function *randomForest* from the randomForest package[14], Support Vector Machines by the R function *svm* from the e1071 package[15], k-Nearest Neighbor by the R function *knn* from the class package[8], and Linear Discriminant Analysis by the R function *lda* from the MASS package[8].

Models using the four methods, RF, SVM, KNN and LDA, were trained using default parameters under 6-fold cross validation. The k of KNN was set to 5. The confusion matrices for each of the methods were aggregated at each iteration of the cross validation, resulting in an average CV accuracy score for each model.

Hyperparameter Tuning

The cost and gamma parameters of SVM were tuned by grid search, under cross validation. The cost was in the range 10^{-2} to 10^{3} and the gamma was in the range 10^{-6} to 10^{0}.

GO Enrichment Analysis

GO enrichment analysis was used to see whether some gene functions were easier to predict than others. The hypergeometric test was used to check if any GO terms were overrepresented among the correctly predicted genes, as opposed to the misclassified genes. The set of correctly predicted genes were the genes that were correctly predicted by both the SVM and RF model. Upcons and downcons were tested separately.

The R function *fish_GO* from the salmonfisher package[16] was used to retrieve all the GO terms associated with the genes. The hypergeometric test was performed by the R function *HyperGTest* from the GOstat package[17], this function takes a parameter object as input, which was delivered by the R function *GSEAGOHyperGParams* from the GSEABase package[18], with correctly predicted genes as the gene set of interest, all tested genes as the background to test against, and testing for overrepresentation of Biological Process terms, the p-value threshold was set to 0.01.

Randomization

To check if the results could have happened by chance, the class labels were randomized, and cross validation was run again 100 times.

Reproducibility

The code is available on GitLab, at https://www.gitlab.com/tobibjor/thesis.

Results

Multiple different approaches were attempted. Initially I tried to classify three classes: upcons, downcons and conscons (gene duplicates where both copies were conserved). Since it was difficult for the models to be able to discriminate the classes, this was dropped in favor of only classifying upcons vs downcons. The class imbalance problem was attempted solved by copying the training examples of the minority classes, but it did not work. The absolute value of the contrasts was tried, but this had a negative impact on performance.

Regression was also tried, with the shift in θ from EVE as response. This was dropped as the residuals were larger than the deviation from the mean, meaning that the mean was closer to the true value than the prediction was.

Machine Learning

Table 2 Results from model fitting under 6-fold CV. Feature selection by hierarchical clustering (297 features) and feature selection by appearance shift (26 features).

		RF	SVM	KNN	LDA
297 features	Accuracy	0.651	0.651	0.571	0.571
	F1 Score	0.657	0.662	0.357	0.573
26 features	Accuracy	0.642	0.669	0.614	0.660
	F1 Score	0.642	0.792	0.574	0.668

The dataset contained 1478 genes and 1492 TF motifs. 756 genes were used for model fitting, equally distributed between classes upcons and downcons. Two methods of feature selection were performed on the 1492 motifs, leading to models with 297 and 26 features, respectively.

Table 2 presents the results for models trained with default parameters and under 6-fold cross validation. SVM gave the best results, overall, and an increase in performance with fewer features was observed for all but one method.

Hyperparameter Tuning

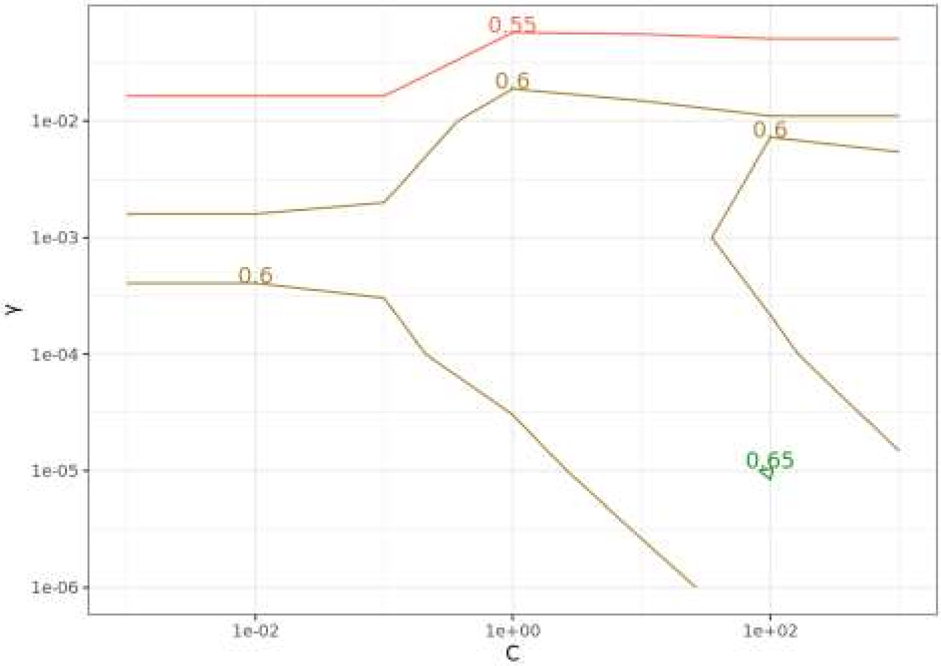


Figure 2 Contour plot of accuracies of SVM from grid search

The optimal parameters of the SVM model with 297 features found by grid search, were cost equal to 100 and gamma equal to 10^{-5}, as seen in Figure 2. Surprisingly, this did not lead to an improvement of either the accuracy or the F1 Score, compared to the untuned model.

GO Enrichment Analysis

Five GO terms were overrepresented among correctly predicted upcons. These were related to the ERAD pathway and regulation of signaling pathways.

49 GO terms were overrepresented among correctly predicted downcons. Many of these were related to cell death, however “digestive tract development” was the most significant.

Randomization

None of the accuracies from the randomized models exceeded the observed accuracy, but it is reasonable to assume these accuracies to be normally distributed. The randomized accuracies had a mean of 0.5 and standard deviation of 0.022. Thus, the p-value of achieving an accuracy of 65% by chance, is $1.5 \cdot 10^{-9}$.

Discussion

The results are indicative that the changes in gene regulation can be explained by changes in the number of TFBS and by which TFs that bind.

On average downcons have much fewer TFBS than upcons. Looking at the sign of the contrasts, 67% of the downcons are negative, and 57% of upcons are positive. A naive classifier solely based on the sign, would still get an accuracy of 62%, given the confusion matrix in Equation 1, with upcons in the first position and the predictions as rows.

$$\begin{bmatrix} 378 \cdot 57\% & 378 \cdot 33\% \\ 378 \cdot 43\% & 378 \cdot 67\% \end{bmatrix} = \begin{bmatrix} 215 & 125 \\ 163 & 253 \end{bmatrix} \quad (1)$$

GO

The gene set were genes that were correctly predicted by both SVM and Random Forest, so the enriched GO terms more common among genes that were correctly predicted, rather than with the dataset as a whole. The results indicate that some gene functions are more prevalent in the dataset. Genes with similar function are probably regulated in the same way, and the models seem to pick up on that, which seem to be the case for the downcons in particular.

The endoplasmic reticulum-associated degradation (ERAD) pathway is responsible for marking misfolded proteins for degradation [19]. So, it is logical that many genes in this pathway are upregulated, as to offset the risk from increased pseudogenization of unneeded genes.

Biological Insights

The transformed matrices show that the up copy have primarily gained new motifs, meaning that different TFs are regulating than the conserved copy, while the down copy have primarily lost motifs. This likely indicate that upcons have probably begun the process of neofunctionalization while downcons have started the slow process of pseudogenization. Gillard *et al.*[1] noted that genes where both duplicates were downregulated were often involved with ribosomes and attributed this to the prevention of faulty proteins ruining protein complexes and keeping within the size constraints of the cell.

In conclusion, my results are indicative that patterns that allow for the prediction of the evolutionary direction of genes based on nearby TF motifs, do exist in the data.

References

1. Gillard GB, Grønvold L, Røsæg LL, Holen MM, Monsen Ø, Koop BF, Rondeau EB, Gundappa MK, Mendoza J, Macqueen DJ, et al: **Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication.** *Genome Biology* 2021, **22:**103.
2. Yan F, Powell DR, Curtis DJ, Wong NC: **From reads to insight: a hitchhiker's guide to ATAC-seq data analysis.** *Genome Biology* 2020, **21**.
3. Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, Fust A, Preussner J, Kuenne C, Braun T, et al: **ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation.** *Nature Communications* 2020, **11:**4267.
4. Rohlfs RV, Nielsen R: **Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution.** *Systematic Biology* 2015, **64:**695-708.
5. Dey A: **Machine Learning Algorithms: A Review.** *International Journal of Computer Science and Information Technologies* 2016, **7:**6.
6. Chang CC, Lin CJ: **LIBSVM: A Library for Support Vector Machines.** *Acm Transactions on Intelligent Systems and Technology* 2011, **2:**27.
7. Breiman L: **Random forests.** *Machine Learning* 2001, **45:**5-32.
8. Venables WN, Ripley BD: *Modern Applied Statistics with S.* Fourth edn. New York: Springer; 2002.
9. Chicco D, Jurman G: **The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation.** *BMC genomics* 2020, **21:**6-6.
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
11. Gene Ontology Consortium: **The Gene Ontology resource: enriching a GOld mine.** *Nucleic Acids Res* 2021, **49:**D325-d334.
12. R Core Team: **R: A Language and Environment for Statistical Computing.** 4.0.2 version. Vienna, Austria: R Foundation for Statistical Computing; 2020.
13. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, et al: **The Atlantic salmon genome provides insights into rediploidization.** *Nature* 2016, **533:**200-205.
14. Liaw A, Wiener M: **Classification and Regression by randomForest.** *R News* 2002, **2:**18-22.
15. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F: **e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.** 2020.
16. Gillard G: **salmonfisher: Retrieve salmonid genomic data.** *GitLab* 2021.
17. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association.** *Bioinformatics* 2007, **23:**257-258.
18. Morgan M, Falcon S, Gentleman R: **GSEABase: Gene set enrichment data structures and methods.** 2020.
19. Hoseki J, Ushioda R, Nagata K: **Mechanism and components of endoplasmic reticulum-associated degradation.** *The Journal of Biochemistry* 2009, **147:**19-25.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway