Stockholms universitet

Missing Edges: Unraveling Exon Duplication History

Saknade kanter: Att undersöka historiken för exon-dupliceringar

Maria Damerji

Handledare: Lars Arvestad
Examinator: Marc Hellmuth
Inlämningsdatum: 20-05-2024

Kandidatuppsats i datalogi
Bachelor Thesis in Computer Science

Contents

Abstract

Mutations are essential for evolution, helping to change and shape the genetic makeup of species. Gene duplication, especially exon duplication, can play a role in this by increasing the variety of proteins that genes can create. This study examines the identification and connection of missing links in exon expansion graphs of the human genome, which are graphs where each node represents an exon and each edge indicates similarities between exons. However, these graphs can sometimes be incomplete, meaning there are node pairs that are not neighbours (connected by an edge) in the expansion graphs.
In this study, different posterior probability trees are generated using Markov Chain Monte Carlo (MCMC) methods in MrBayes. The probability that two nodes which are not neighbours in the expansion graphs are neighbours in the posterior probability trees is then examined.
The result of this study is that the likelihood for exon pairs not neighbours in the expansion graph to be neighbours in the posterior tree was close to zero in the majority of cases.
This study the complexity in discovering exon duplication history, showing how posterior trees can be used on exon duplication events.

Sammanfattning

Mutationer är avgörande för evolutionen, och hjälper till att förändra och forma arternas genetiska sammansättning. Gen-duplicering, särskilt exon-duplicering, kan spela roll i detta genom att öka mängden proteiner som gener kan skapa. Den här studien tittar på hur man kan identifiera och koppla ihop saknade länkar i exonexpansionsgrafer av det mänskliga genomet, som är grafer där varje nod representerar en exon och varje kant visar starka likheter mellan exoner. Emellertid kan dessa grafer ibland vara ofullständiga, vilket innebär att det finns nodpar som inte är grannar (anslutna med kant) i expansionsgraferna.
I denna studie genererades olika posteriora sannolikhetsträd med metoden Markov Chain Monte Carlo (MCMC) i MrBayes. Sannolikheten att två noder som inte är sammankopplade i expansionsgraferna är grannar i de posteriora sannolikhetsträden undersöktes sedan.
Studien visade att sannolikheten för att exon-par som inte är kopplade i expansionsgrafen skulle vara grannar i de posteriora träden var nära noll i de flesta fall.
Den här studien belyser komplexiteten i att upptäcka exon- duplicerings-historik, och visar hur posteriora träd kan användas på exon-duplicerings-händelser.

1 Introduction

The dynamics of mutations are essential for driving evolutionary progress and reshaping the genetic landscape among species. Gene duplication is a significant mechanism that drives this process and facilitates the formation of different genetic transcripts [16]. Exon duplication is a term for a specific event that occurs when gene-coding segments are duplicated, producing several copies of a particular exon. [9]

Duplicated exons can greatly increase the range of proteins that a gene can encode through different combinations of exons being included in the transcripts that encode the proteins. Exon duplication can consequently potentially promote the creation of unique features that help species adapt to their surroundings [9]. Investigating exon duplications can provide important information on the dynamics of mutations within species and illuminate the evolutionary paths taken by those species.

1.1 Purpose

In genetics and genomics, researchers often construct evolutionary trees [14]. For this study, we aim to focus on a certain type of graph, known as an exon expansion graph. Expansion graphs illustrate how exons are duplicated and rearranged within a genome. However, these graphs may not always capture all possible connections between exons. Some edge connections may be missing due to limitations in data or analysis.

The purpose of the study is to investigate the feasibility of identifying and establishing connections among these missing edges in exon expansion graphs of the human genome. The goal is to answer the following question: *What is the likelihood for a node pair not neighbours in a expansion graph to be neighbours in posterior probability trees?* In essence, the study seeks to verify the precision of exon expansion graphs regarding exon duplication patterns by exploring the potential for connecting nodes that are not neighbours.

2 Background

2.1 Biology concepts

DNA, the blueprint of life, is comprised of twisted ladder-shaped sequence of nucleotides containing adenine (A), thymine (T), guanine (G), and cytosine (C) acids, dictating the synthesis of proteins through a genetic code embedded in its gene sequence, pivotal for biological growth, inheritance, and evolution.[1]

The **genome** refers to all the genetic material (DNA) present in an organism, which is usually found in chromosomes. It represents the entire set of instructions necessary for the development, functioning, and reproduction of an organism. [6]

Genes are sequences of nucleotides, which are molecular units that link together to form DNA and RNA, encompassing regions that are both coding and non-coding, where coding regions (exons) contains instruction for protein synthesis and non-coding (introns) regions play roles in gene regulation and expression[5]

Transcription is the process of copying a particular part of a DNA sequence specifying the amino acid order, by untangling the twisted ladder-shaped DNA and creating a copy called RNA, complementary to one side of the DNA. The RNA copy produces a molecule called **messenger RNA (mRNA)**, which serves as a template for protein production, proteins are the building blocks of the muscles and organs, providing the necessary instructions encoded within the gene. [2]

Exons are segments of the genes that contain coding information for protein synthesis in cells containing a nucleus, i.e., eukaryotes and are made of DNA segments of the gene. Also specific to eukaryotes, the introns, are any nucleotide sequences within a gene not expressed or operative in the final RNA product. [7]
The mRNA only consists of combined parts of the coding regions of the DNA (exons) and the introns are removed. [8, 15]

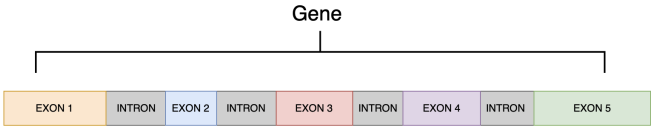

Figure 2.1: A gene is a segment of the DNA, consisting of coding regionse(Exons) and non-coding regions (Introns), and is part of the genome

Alternative splicing that allows different combinations of exons to be included or excluded from the final mRNA product, through which a single gene can generate multiple mRNA transcripts and, consequently, multiple variations of protein. Alternative splicing significantly increases the diversity of proteins that can be produced from a limited number of genes, contributing to the complexity and adaptability of organisms. [17]

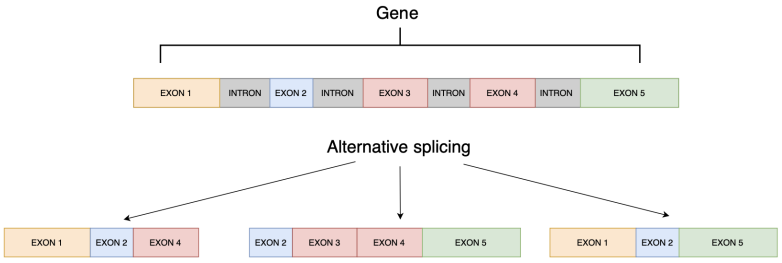

Figure 2.2: Illustration of the three alternative mRNA that have been transcribed from a gene in the genome. During the transcription, different exons are chosen and put together in segments. In the figure exon 4 is a copy of exon 3, which means there has been an exon duplication in the gene.

Exon duplications, are important for how genes change and work in living things. In rare instances, certain segments of a gene, known as exons, are alternately copied and spliced, deviating from the constitutive pattern. This alternative splicing process can result in the generation of diverse protein variants[12]. There is growing evidence

that exon duplication can influence the included pieces in the making of mRNA. The process, known as alternative splicing, is a way genes can create different versions of proteins from the same DNA code. Researchers have found examples of repeated exons using computer programs that analyze genetic information. They have discovered that repetitions are common and can affect how genes work and evolve. By looking at genetic data from different organisms, like humans and fruit flies, scientists are learning more about how exon duplications shape life. [9]

3 Method

3.1 Expansion graphs

Expansion graphs are graph representations of the relationship between different exons within the context of exon duplication. Each node (vertex) represents an exon and is labelled by the exon's genomics coordinates (starting reference, ending reference) within the gene. The edges between the exons represent nucleotide patterns in the exon sequences that are nearly identical (See figure 3.1). The expansion graphs can be complete graphs due to the transitivity of sequence similarity between the exons. A complete graph is a graph where all the nodes are connected by an edge to all other nodes in the graph.

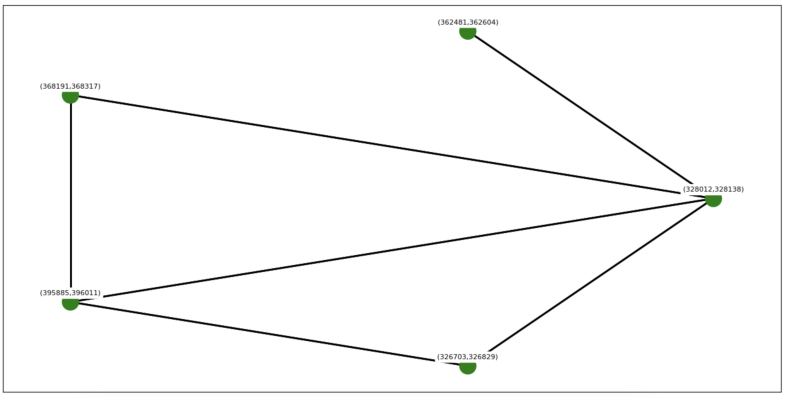

Figure 3.1: Expansion graph of an exon duplication event with 5 exons, where every node is the coordinates to the exon within the gene and the edge between nodes stands for large similarities in sequences.

Phylogenetic tree A phylogenetic tree visually represents the evolutionary relationships among entities, with nodes depicting common ancestors, leaves representing individual entities, and branches connecting them. Each node signifies a divergence event, where an ancestor node gives rise to child nodes. The structure of the tree, composed of nodes and branches, illustrates the evolutionary history of the entities being compared.
In the context of this project, the phylogenetic tree serves as a representation of the evolutionary history of exon sequences, illustrating how they have duplicated and diverged over time. Each branch node represents a common ancestor exon sequence,

and the branches leading from it depict the duplication events that have given rise to descendant exon sequences. The leaves of the tree, which are the terminal nodes, represent the current exon sequences resulting from these duplication events (See figure 3.2).

A tree representation can be useful for understanding the patterns of exon duplication and divergence within a genome or a set of related genomes. [14]

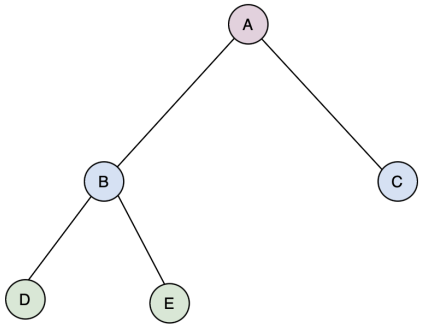

Figure 3.2: Illustration of a phylogenetic tree where A is the ancestor exon of B, C, D, E and B is the ancestor exon of D and E

3.2 Multiple Sequence Alignment

Due to the variable sequence lengths of exons within the same exon duplication event, multiple sequence alignment (MSA) aligns sequences by systematically arranging them to identify common patterns or regions of similarity, enabling easy comparison even if they differ in length. This MSA process inserts gaps when necessary to maintain coherence (Edgar, 2004). MUSCLE was used, due to being a practical MSA program with a good balance between speed and accuracy (Edgar, 2004), where default settings were used. See Figure 3.3 for an example of an exon duplication event alignment [4].

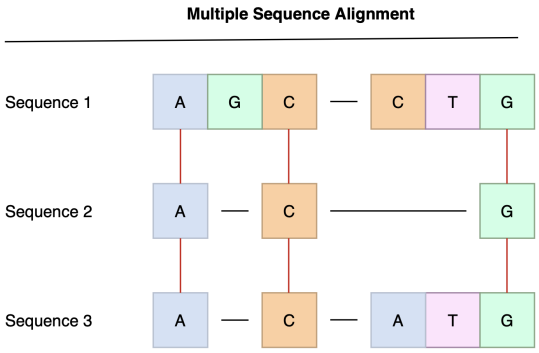

Figure 3.3: Illustration of a multiple sequence alignment where the sequences AGC-CTG, ACG and ACATG where aligned to be 7 nucleotides long and where the same nucleotides are aligned to be in same position in the different sequences. The lines are inserted to align the sequences.

3.3 MrBayes/posterior probability trees

In phylogenetics, Bayesian inference of phylogeny is a commonly used method. It produces a posterior distribution of trees, rather than a single best estimate. This distribution encapsulates the uncertainty inherent in phylogenetic reconstruction and provides a more comprehensive view of the evolutionary relationships. Instead of focusing solely on a single tree, Bayesian methods allow for the exploration of multiple possible evolutionary scenarios and the assessment of the likelihood based on the available data [3].

The posterior distribution of trees obtained from Bayesian phylogenetic analysis contains information about the relative probabilities of different tree topologies. Each tree in the distribution represents a possible evolutionary history of the groups under study, with its posterior probability indicating the degree of likelihood for the outcome of that particular topology given the data and the model assumptions. Rather than relying on a single 'best' tree, researchers can examine the entire distribution to understand the range of plausible evolutionary scenarios. [3]

Bayes' Theorem provides a mathematical framework for updating our beliefs about a hypothesis based on new evidence. In the context of phylogenetics, this theorem is crucial for understanding how Bayesian inference generates posterior probabilities. These probabilities represent the revised degree of confidence in a hypothesis (such as a specific tree topology) after analyzing the observed data in light of any prior information or assumptions.

Bayes' Theorem states:

$$P(\text{Hypothesis}|\text{Data}) = \frac{P(\text{Data})}{P(\text{Data}|\text{Hypothesis}) \times P(\text{Hypothesis})}$$

Here's what each component represents:

$P(\text{Hypothesis}|\text{Data})$ is the posterior probability of the hypothesis given the data.
$P(\text{Data}|\text{Hypothesis})$ is the likelihood of observing the data given the hypothesis.
$P(\text{Hypothesis})$ is the prior probability assigned to the hypothesis.
$P(\text{Data})$ is the overall probability of observing the data. [3, 13]

In the context of exon duplication research, the generated posterior probability trees serve as the hypotheses, while the exon sequences constitute the observed data. By applying Bayes' Theorem, it is possible to systematically evaluate the credibility of different evolutionary scenarios regarding exon duplication, refining their understanding based on both existing knowledge and newly acquired sequence data. [3]

Monte Carlo Markov Chains (MCMC) are utilized in MrBayes inference due to their efficiency in handling large datasets MCMC methods offer a scalable solution. Iteratively sampling from the posterior distribution, MrBayes navigates complex tree topologies and model parameters, providing reliable estimates of posterior probabilities even with extensive sizes of genomic data. [13]

3.4 Aligning data

The exon sequences of each exon duplication event are represented as nodes in the different expansion graphs. The exons sequences in each duplication event were aligned using the multiple sequence alignment tool MUSCLE, because it is necessary that the sequences are equal lengths for examining how similar the exons are that are not neighbours in the expansion graphs. The input and output files were in FASTA format, a text-based representation of nucleotide sequences[10]. After the sequences were aligned, the data was rewritten and transformed into matrix-formatted data blocks in the NEXUS-format[11], enabling inference with MrBayes.

3.5 Running MrBayes Inference

Each Nexus file was run through MrBayes, generating posterior probability trees for each exon duplication event for each gene.

A database was created to store posterior probability trees and their posterior probability for each gene's expansion, along with relevant tables from the provided database and a translation table for sequence reference indexing, for easier analysis of the data and less memory complexity.

3.6 Parsing through the trees

In the posterior probability tree, the term "Neighbours" has been employed to denote instances where the distance between two nodes, both emanating from a shared parent node, is precisely one unit.

The expansion graphs for each exon duplication event were generated. Out of any given dataset, only exons should be examined. The exons not connected in the expansion graphs for each different exon duplication event are identified. The cumulative posterior probability, obtained by summing the posterior probabilities of the trees where exons that are not connected in the expansion graph are considered neighbours, was computed to aggregate the individual posterior probabilities for the missing edges across all expansion graphs.

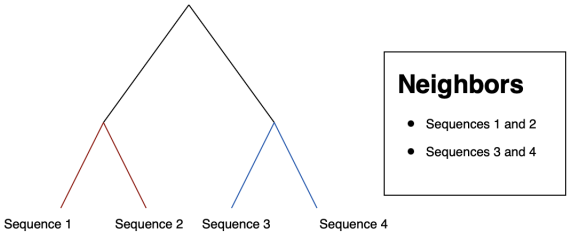

Figure 3.4: Illustration of what it means that two sequences are neighbours

When the total posterior probability for all missing edges in all expansion graphs is added to a database, they are then extracted into a list visualized by histograms with the tool MatPlotLib.

4 Results

4.1 Exon duplication in the human genome

4.1.1 Data

Two files were utilized: The first one was a database that included information on exon duplication events, with each gene and each exon duplication event for each gene assigned a unique identifier, along with details such as the starting and ending reference of each exon in each gene and the chromosome of the gene. The second one was a FASTA file of the human genome DNA sequences.

Data from the database was extracted and organized into a dictionary. Only exon duplication events including at least 4 exons of a minimum sequence length of 30 nucleotides were considered in the analysis (see: Chapter 3.0.7). Analyzing expansions with less than 4 exons would not be meaningful and exon sequences shorter than 30 nucleotides could be to short to have significant patterns required for meaningful comparison with other exon sequences.

4.1.2 Posterior trees

In figures 4.2 and 4.3 there are examples of posterior trees generated by MrBayes for an exon duplication event that has the expansion graph in figure 4.1.

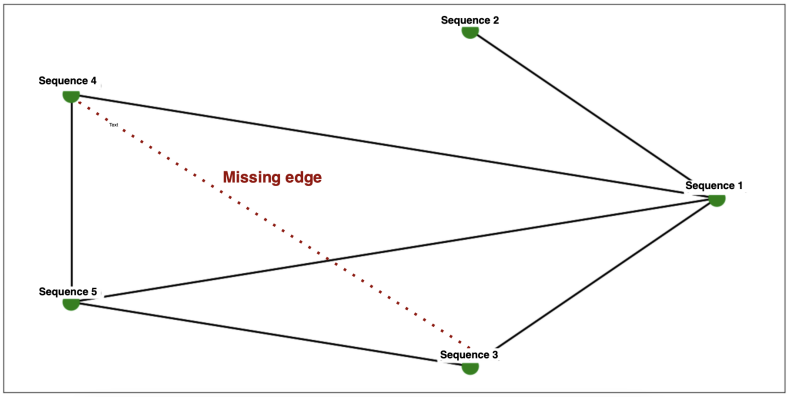

Figure 4.1: The expansion graph from figure 3.1 with sequences indexes instead of coordinates in gene as nodes. The missing edge between sequences 3 and 4 is marked with a red dotted line.

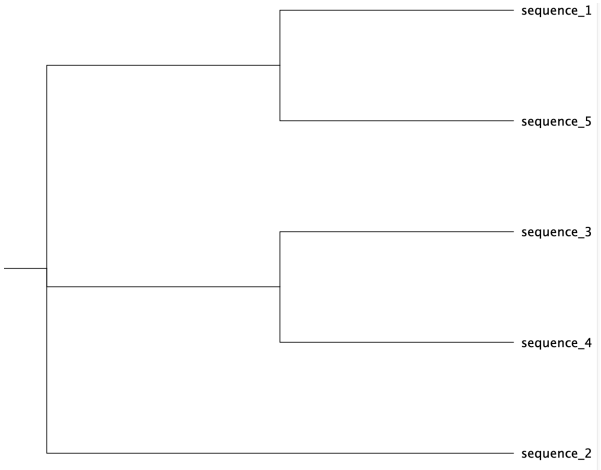

Figure 4.2: Posterior tree generated from an exon duplication event where sequences 3 and 4 represent two nodes that are missing an edge between them in the expansion graph (see figure 4.1). There was 15 different posterior trees where generates and this is tree had posterior probability p=0.33.

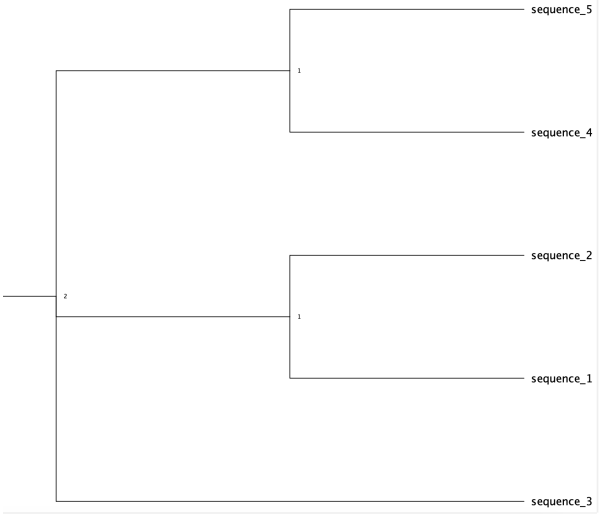

Figure 4.3: Posterior tree generated from an exon duplication event where sequences 3 and 4 represent two nodes that are missing an edge between them in the expansion graph (see figure 4.1). There was 15 different posterior trees where generates and this is tree had posterior probability p=0.16.

4.2 Probabilities for node pairs to be neighbours

4.2.1 Data characteristics

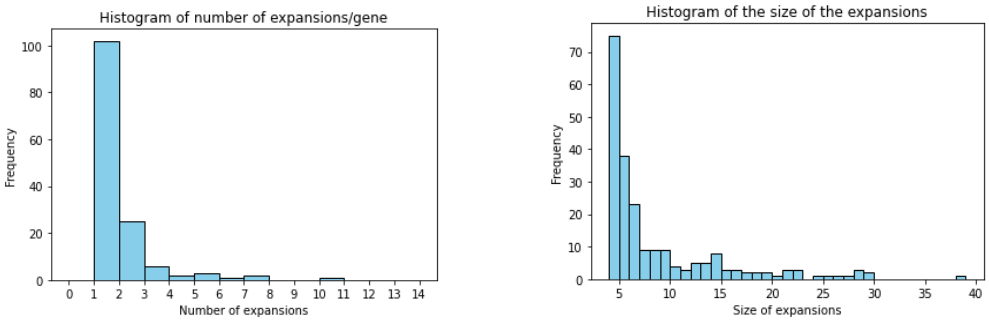

Figure 4.4: Histograms showing the number of exons in each expansion and number of expansions per gene.

The most common number of exon duplication events for a gene is 1 (See figure 4.4) and the most common number of exons per expansion is 4, decreasing according to negative exponential functions. The total number of exon duplication events that where studied after the filtering of the data was 223.

4.2.2 Missing edge frequency in expansion graphs

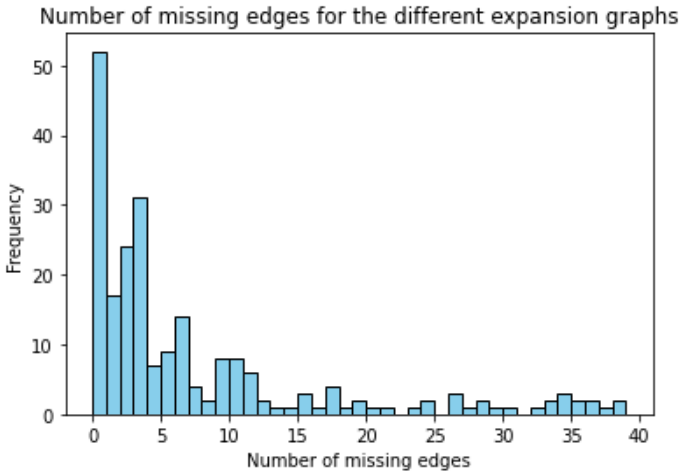

Figure 4.5: Histogram over the number of missing edges for each expansion graph.

The majority of expansion graph of exon expansion are complete graphs, meaning that all nodes of the graph are connected by an edge to every other node in the graph. The number of missing edges in each graph ranges between 0-38 (see figure 4.5).

4.2.3 Number of generated posterior trees per expansion

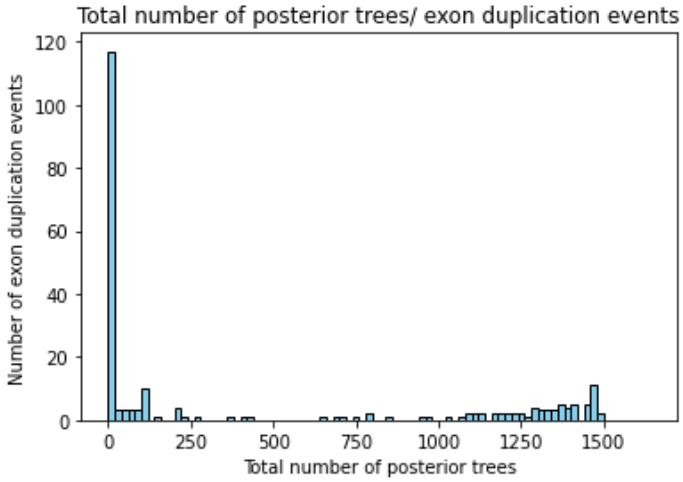

Figure 4.6: Histogram over the number of posterior trees generated for each exon duplication event.

For the majority of exon duplication event 0-20 posterior probability trees where generated. are complete graphs. However the number of posterior trees generated was ranging between 0 to around 1500. (see figure 4.6).

Total probability of neighbour relationship for node pairs of each missing edge

For each detected missing edge in an expansion graph, the closeness of the two non-neighbouring nodes was examined in all generated probability trees for the exon duplication event on which the expansion graph is based. The total posterior probability of the trees where the sequences of the node pair were neighbours was calculated. The histogram in Figure 4.7 shows the probabilities for the different nodes that were not neighbours in the expansion graphs to be neighbours in the posterior trees.

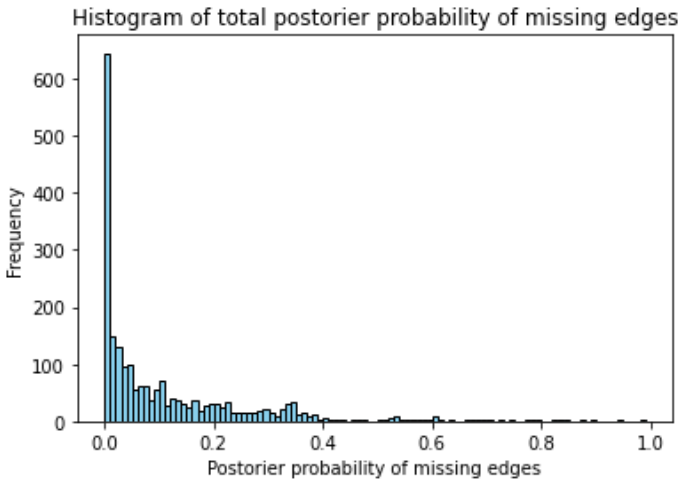

Figure 4.7: The total posterior probability for the node pair not neighbours in the expansion graphs to be neighbours in posterior trees.

Most of the exons that was missing an edge between them in the expansion graphs had a total posterior probability of less than 0.01 of being neighbours in the generated posterior probability trees (See figure 4.7). However the probabilities for a neighbour relationship in the posterior trees for the node pair with a missing edge in the expansion graphs ranges from 0 to 1.

Total probability of missing edges depending on number of generated posterior trees

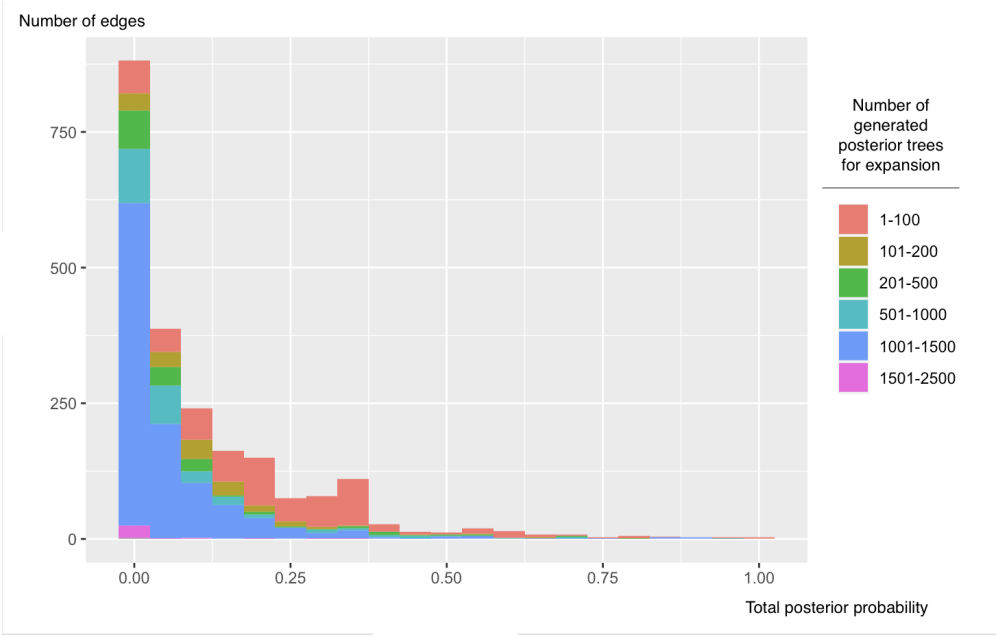

Figure 4.8: The total posterior probability for the node pair not neighbours in the expansion graphs to be neighbours in posterior trees. This is the same as figure 4.7. The colors indicate how many posterior trees where generated of the exon duplication events (see figure 4.6). The diagram is a combining the information from figure 4.6 and figure 4.7.

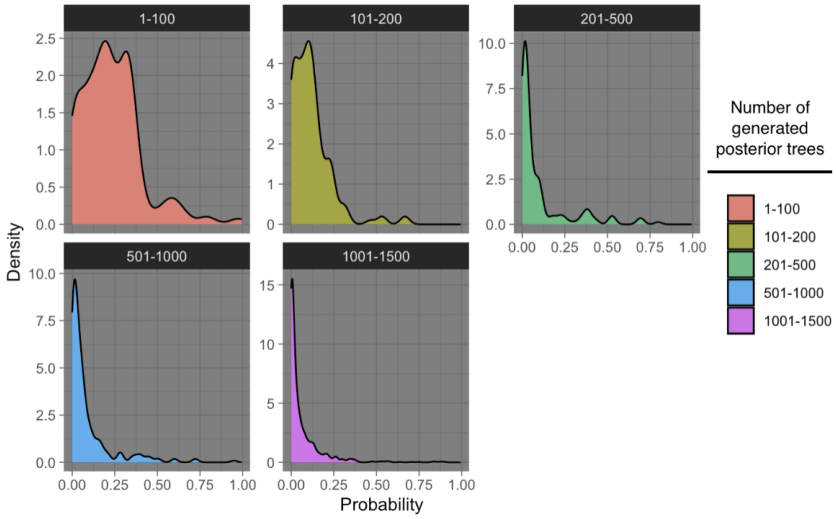

Figure 4.9: Density plot of each separate bracket of number of generated posterior trees with the total posterior probability of the different missing edges on the x-axis.

The combined likelihood of different missing edges spans from 0 to 1, as shown in figures 4.8 and 4.9. Yet, when there are more posterior trees generated for an expansion, the chances of a missing edge occurring between nodes not directly connected in the expansion graph decrease. The density plot reveals that expansions with 1-1500 posterior trees often exhibit instances where the likelihood of a missing edge surpasses 0.5. However, this trend is more pronounced when fewer posterior trees are generated, as depicted in figure 4.9.

4.3 Analysis

4.3.1 Improvement of results

Due to only 223 exon duplication events being analyzed in this study, more far-reaching conclusions could be drawn if the sample size was larger. Despite the small sample size, data is collected from different demographics and could be seen as a small representation of the total population.

This study investigates the probability of exons not being neighbours in the expansion graphs to be neighbours (distance of 1) in the posterior probability trees. A study with a larger magnitude could take into consideration how close the node pairs are in the posterior trees (see figure 4.8).

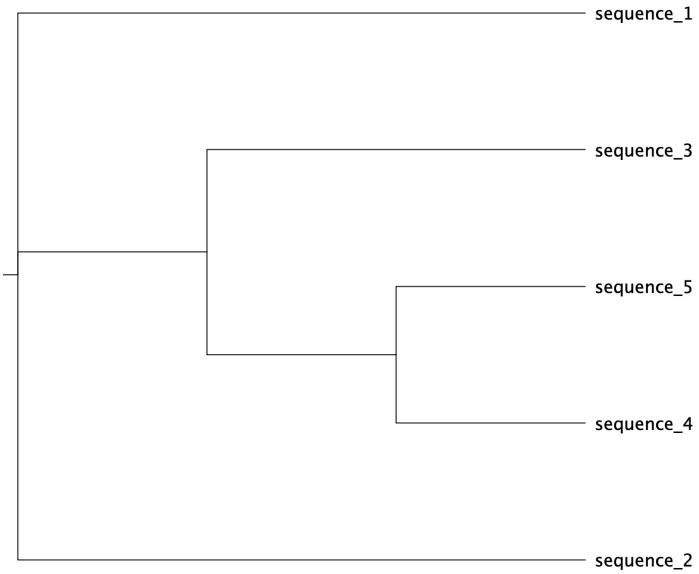

Figure 4.10: Posterior tree where sequence 4 and sequence 3 are close, but not neighbours. Sequence 3 and 4 are not connected in the expansion graph (see figure 4.1). The algorithm used in this study did not take into consideration how close the sequences are in the posterior tree, but instead only checked if they where neighbours or not.

Knowing how the distance between the exons are in the different posterior trees and taking that into consideration in the posterior probability of a node pair being neighbours in the expansion graph, could give more insight in their similarities.

4.3.2 Possible improvements

The methods used for the study had to take time and hardware limitations into account. MrBayes is only one of many programs that could be used, although recognized in phylogenetics.

One suggestion of a further study would be to investigate the proximity of the nodes more thoroughly. In this study, we only examined whether the two nodes not connected by an edge in the expansion graphs were neighbours in the generated posterior trees. However, investigating their average distance of the node pair in the posterior trees could provide a more nuanced view of the similarity between exons (nodes) that are not neighbours in the expansion graphs. Figure 4.8 the sequences 3 and 4 are close (two steps apart) but not neighbours, so investigating the average distance could be interesting.

Additionally, it would be interesting to analyze a different dataset and compare the conclusions. The parameters used can also be changed and give other results.

4.3.3 Discussion

When looking at histogram depicting the posterior probability for each node pair (representing exons) not neighbours in the expansion graphs to be neighbours in the generated posterior tree (see figure 4.7), it becomes apparent that in the majority of cases, the posterior probability $p < 0.01$ for each two exons to be neighbours. This suggests that there is a substantial likelihood of exon pair not neighbours in the expansion graphs to not be neighbours in posterior probability tree or to be neighbours in a posterior tree with posterior probability less than 0.01.

Nevertheless, it's crucial to acknowledge that despite the absence of a neighbour relationship, there remains a possibility for these exons to still maintain a degree of proximity in the posterior trees. Even though the exon pairs not neighbours in the expansion graphs are unlikely to be neighbours in the posterior trees, they could still exhibit a close relationship in the posterior trees and the exon sequences could still be very similar. The number of posterior trees generated for each exon duplication event was ranging from 0-1520 (see figure 4.6), indicating that in the exon duplication events where many posterior trees where generated the sequences are very similar making the MCMC algorithm predict for example 1000 trees with small posterior probabilities each, instead of fewer trees with larger probabilities. The exon pairs not neighbours in the expansion graphs with more than 500 posterior trees generated all had a likelihood of less than 0.1 of being neighbours in the posterior trees (see figure 4.8), but that doesn't necessarily mean that the sequences of the exon pair are very different.

In summary, a node pair not neighbours in expansion graphs has a low likelihood of being neighbours in the posterior trees, but the sequences of the exons might still be similar and the nodes might still have a proximity in the posterior trees.

Bibliography

[1] NE Nationalencyklopedin AB. DNA. https://www.ne.se/uppslagsverk/encyklopedi/lång/dna, 2023. (accessed: 06.05.2024).

[2] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J.D. Watson. *Molecular Biology of the Cell*. Garland, 4th edition, 2002.

[3] Karen A. Cranston and Bruce Rannala. Summarizing a posterior distribution of trees using agreement subtrees. *Society of Systematic Biologists*, 56(4):578—-590, 2007.

[4] Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. 2004.

[5] Ph.D. Eric Green, M.D. Gene. https://www.genome.gov/genetics-glossary/Gene. (accessed: 05.05.2024).

[6] Ph.D. Eric Green, M.D. Genome. https://www.genome.gov/genetics-glossary/Genome. (accessed: 05.05.2024).

[7] National Human Genome Research Institute. Exon. https://www.genome.gov/genetics-glossary/Exon. (accessed: 05.05.2024).

[8] National Cancer Intitute. Transcription. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/transcription, 2017. (accessed: 07.05.2024).

[9] T. M. Ivanov and D. D. Pervouchine. Tandem exon duplications expanding the alternative splicing repertoire. 2017.

[10] National library of medicine. Fasta format for nucleotide sequences. https://www.ncbi.nlm.nih.gov/genbank/fastaformat/. (accessed: 07.06.2024).

[11] Maddison WP Maddison DR, Swofford DL. Nexus: an extensible file format for systematic information. *Systematic biology*, 46(4):590–621, 1997.

[12] Pozo F. Walsh T. A. Abascal F. Martinez Gomez, L. and M. L. Tress. The clinical importance of tandem exon duplication-derived substitutions. *Nucleic acids research*, 49(14):8232–8246, 2021.

[13] Tom Britton och Sven Erick Alm. *Stokastik : Sannolikhetsteori och statistikteori med tillämpningar*. Liber AB, 1st edition, 2008.

[14] A.D. Scott and D.A. Baum. Phylogenetic tree. In Richard M. Kliman, editor, *Encyclopedia of Evolutionary Biology*, pages 270–276. Academic Press, Oxford, 2016.

[15] Ph.D. Shurjo K. Sen. Messenger RNA (MRNA). https://www.genome.gov/genetics-glossary/messenger-rna. (accessed: 06.05.2024).

[16] Claudine Landès Carène Rizzon Tanguy Lallemand, Martin Leduc and Emmanuelle Lerat. An overview of duplicated gene detection methods: Why the duplication mechanism has to be accounted for in their choice. *Genes*, 11(9):1046, 2020.

[17] Yan Wang, Jing Liu, BO Huang, Yan-Mei Xu, Jing Li, Lin-Feng Huang, Jin Lin, Jing Zhang, Qing-Hua Min, Wei-Ming Yang, et al. Mechanism of alternative splicing and its regulation. *Biomedical reports*, 3(2):152–158, 2015.

Datalogi
www.math.su.se

Beräkningsmatematik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm