The molecular evolution of amylase duplicates genes in D. melanogaster group

Abstract

Within D. melanogaster, the Amylase gene designation is made up of duplicate genes that can be designated as proximal or distal genes. Studies have divided up Drosophila coding regions by these classifications. And within these subgroups, there are significant evolutionary similarities and differences. These overlaps have caused for there to be an ambiguity when it comes to the relationship between these two subgroups. Do proximal and distal genes evolve independently (if distal genes cluster with their orthologs) or not (if the paralogs cluster)? By looking at the DNA material of nine species from the D. melanogaster Group, this paper conducts advanced genetic and statistical analyses using specialized software. By creating dependent and independent phylogenies for the nine species and conducting advanced bootstrap parameters on them, the paper demonstrates that it is more likely that the relationship between the proximal and distal genes is one of dependent evolution. This supports the notion that there exists a form of positive selection acting upon a coding region. For future research, The same analysis could be done using the amylase regulatory region. This analysis could provide insight into the relationship between upstream sequences of proximal and distal genes, multiple alignments and putative regulatory elements. And in future follow-ups, it could focus on the nucleotide divergence that appears to be taking place.

Introduction

Drosophila melanogaster is a species of fly in the family Drosophilidae. The alpha-amylase system of Drosophila is one of the most extensively examined systems of genes in the field of evolutionary research. Alpha-amylase (EC 3.2.1.1, alpha-1,4-glucan–4-glucanohydrolase) is a digestive enzyme. Its role in the digestive process is to break down starch into glucose and maltose to produce energy. There are six major and minor isozymes of amylase that have been recorded in natural populations of D. melanogaster (Inomata & Yamazaki T).1 Amylase activity is repressed by its produce, glucose, and maltose. It is induced by the substrate starch. Population genetic surveys provide information about the molecular characterization of fitness-related genes upon which natural selection acts. Differences in activity levels and the inducibility has been recorded within and between species. This variation is caused because of a mRNA abundance. However, the difference in the catalytic efficiency of an individual isozyme contributes to the recorded differences in activity between and within species. In the introduction to the differences in inducibility, differences in developmental and organ-specific expression have been recorded Popadić et al.2

The Amylase (Amy) gene of Drosophila is a member of a multigene family. The two main subdivisions of the Amy gene are proximal and distal genes. The closely-linked Amy genes, which are referred to as duplicated Amy genes, demonstrate different evolutionary patterns. The patterns involve concentrated evolution within coding regions and diverging evolution in flanking regions.1 The diverging evolution in flanking regions suggests that there is some differential selection at play within each flanking region. Evidence of differential selection comes from the evolutionary patterns of duplicate groups in Drosophila subgroups. An example of this is the Amyrel gene in the Sophophora subgroup Shapiro et al.1 It codes divergent proteins and has demonstrated different expression patterns to Amylase genes. Another subgroup that displays this pattern is in Drosophila kikkawai and its sister species. When these species exist two duplication groups of the Amy gene. In the first group, there are head–to–head duplicated genes. These are Amylase–1 and Amylase–2. The second group has tail–to–tail duplicated genes. These genes are Amylase–3 and Amylase–4. The Amylase–1 and Amylase–2 genes are shown to cluster with the Amylase genes of D.melanogaster, rather than clustering with the Amylase–3 and Amylase–4 genes of D. kikkawai Sella et al.4 The coding and flanking regions have extremely different in the duplication groups. Amylase–1 and Amylase–2 genes have higher GC (guanine–cytosine) content at the third positions of codons and more influenced codon usage compared to the Amylase–4 and Amylase–3 genes. In previous studies, Amylase–1 and Amylase–2 genes show contrasting evolutionary patterns. These patterns involve concentrating their diverging evolution in flanking regions. Another study looked at the nucleotide sequences of the 5’–flanking regions of duplicated Amylase genes in melanogaster species Good et al.5 The proximal and distal genes had sequence similarities. However, they found that coding regions are clustering into proximal and distal groups. This subdivision was causing increased divergence between the coding regions.1 This paper considers the existence of positive selection acting on one of the subregions, causing a divergence of the genes. However, shifting our attention to the relationship between the two regions, there is limited information on this. Even though the proximal and distal genes are diverging and becoming different, they still share similar characteristics that have evolved similar throughout time.

This brings to light the question of the relationship between the proximal and distal genes? Do they evolve independently? Or do they evolve dependently? By looking at levels of polymorphism and at specific patterns of substitution, it is possible to infer what evolutionary patterns have acted upon these genes. It is also possible to determine the relationship between these two sub-groups of genes. To go about this, the paper firstly gathers the fly DNA sequences from the NCBI genome database. After which, the MEGA software was used in order to find i.) The substitution model for the bootstrap analysis and ii.) Develop a separate, dependent phylogeny for all the proximal and distal gene samples (Supplementary Material II). After this, a tree was made by combining these two trees. This would be the tree used in the
bootstrap analysis. The code for the independent tree was written in the Newick format and was used in the bootstrap analysis (Table 1). Another tree was made by the Phylogeny.fr software (Figure 4). This was used in order to compare the two versions of the phylogeny. In the bootstrap, the MEGA phylogeny was used. In all of the software used, the settings were on normal and default. Using the HyPhy software, a bootstrap analysis was conducted. The HyPhy program did the analysis with the HKY85, as it was chosen by the MEGA software. The possibility of proximal and distal genes evolving dependently as the alternative hypothesis. The possibility of proximal and distal genes evolving independently as the null hypothesis.

Materials and methods

Fly gene sequences

DNA sequences from nine species of flies from the *D. melanogaster* group were used in the paper. These species are:

1. Drosophila Orena
2. Drosophila Teisseire
3. Drosophila Yakuba
4. Drosophila Erecta
5. Drosophila Melanogaster
6. Drosophila Simulans
7. Drosophila Mauritania
8. Drosophila Sechellia
9. Drosophila Ananassae

All of the DNA sequences for these species were used from NCBI Genomic Database. The data was used in a constant FASTA format for the paper. The sequences of the species will be included in the supplementary data (Supplementary Data 1).

MEGA

MEGA (Molecular Evolutionary Genetics Analysis) is a software that specializes in analyzing FASTA DNA sequences. The software emphasizes the integration of sequence acquisition with evolutionary analysis. It contains an array of input data and multiple results explorers for visual representation; the handling and editing of sequence data, sequence alignments, inferred phylogenetic trees; and estimated evolutionary distances Gascuel et al. The software allows the user the ability to browse, edit, summarize, export, and generate publication–quality captions for their results. MEGA also includes distance matrix and phylogeny explorers as well as advanced graphical modules for the visual representation of input data and output results. The main features of this software used in this paper are the phylogeny construction software and the substitution software. The substitution software will analyze the DNA sequences that are uploaded onto the program, after which, it will calculate the AiC value for each substitution model. The model with the lowest AiC value will be the model that will be used to analyze that sequence. See the supplementary information for more specifics on the AiC calculation. After the substitution model was determined, a phylogeny would be created with the gathered information. All of the phylogenies created using the Neighbor–Joining method.

Hyphy

HyPhy (Hypothesis Testing using Phylogenies) is an open–source software package for the analysis of genetic sequences (in particular the inference of natural selection) using techniques in phylogenetics, molecular evolution, and machine learning Tamura et al. The paper uses this software to compare the independent and dependent phylogenies through the use of a bootstrap analysis. The bootstrap was run within normal and default parameters. In the bootstrap, the minimum number of simulation recommended was 100. The program allows 100–1000 simulations. For this analysis, 550 simulations were run. The simulations calculated the LR value. The LR value is the likelihood ratio, defined as 2 (log L− log L0), where LA is the likelihood for the alternative hypothesis, L0 is the likelihood for the null hypothesis (refer to the documentation for HyPhy). A simulation in the bootstrap is to pick random sites from the original sequence with replacement, rebuild the phylogenetic tree for two hypotheses, calculate the log likelihood and generate one likelihood ratio. The goal is to see the likelihood ratio from the data fall into the empirical distribution. In the program, the null hypothesis was entered and the alternative hypothesis. A null hypothesis supports the hypothesis that there is no significant difference between the specified population; that any observed difference being due to sampling or experimental error. The alternative hypothesis supports the hypothesis that there is a significant difference between the specified populations and that these differences share a cause. If the p–value is really small, ~ p<0.0005, the null hypothesis is rejected. The possibility of proximal and distal genes evolving dependently as the alternative hypothesis. And the possibility of proximal and distal genes evolving independently as the null hypothesis. The MEGA software was also used to align and organize the DNA before the phylogenies were created.

Phylogeny.fr

The Phylogeny.fr platform transparently chains programs to automatically perform these tasks. It is run by the Reseau National de Genopoles. Phylogeny.fr offers three main modes. The ‘One Click’ mode targets non–specialists and provides a ready–to–use pipeline chaining programs with recognized accuracy and speed: MUSCLE for multiple alignments, PhyML for tree building, and TreeDyn for tree rendering. All parameters are set up to suit most studies, and users only have to provide their input sequences to obtain a ready–to–print tree. The ‘Advanced’ mode uses the same pipeline but allows the parameters of each program to be customized by users. The ‘A la Carte’ mode offers more flexibility and sophistication, as users can build their own pipeline by selecting and setting up the required steps from a large choice of tools to suit their specific needs. Prior to phylogenetic analysis, users can also collect neighbors of a query sequence by running BLAST on general or specialized databases. A guide tree then helps to select neighbor sequences to be used as input for the phylogeny pipeline. This paper uses the advanced mode in order to create a second version of the dependent phylogeny.

Results

In the distal phylogeny, the tree is divided up into two main clades. One main clade is made up of *D. melanogaster, D. simulans,* and *D. mauritania.* Within the clade, *D. melanogaster* is the outlier. *D. simulans* and *D. mauritania* form a sub–clade. In the second main clade, see that it diverges into one sub–clade and two nodes (Figure 1). The sub–clade is made up of *D. yakuba* and *D. teisseire.* The nodes are *D. orena* and *D. erecta.* In the entire phylogeny, there are low
consensus values (Figure 1). Most values range from 0–10%, meaning that there is little consensus and much variability when it comes to the development of the phylogeny. In the upper clade, there is a sub-clade with a consensus value of 0.0%, which means that it was placed there because the program had no other alternatives (Figure 1).

The evolutionary history was inferred using the Neighbor-Joining method. The optimal tree with the sum of branch length = 0.18459390 is shown. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The analysis involved 7 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 1724 positions in the final dataset. Evolutionary analyses were conducted in MEGA7.

In the proximal phylogeny, the tree is divided up into two main clades. On the top clade, there are two sub-clades. In the first sub-clade, there are *D. orena* and *D. erecta* (Figure 2). In the bottom sub-clade, there are *D. yakuba* and *D. teisseire*. On the bottom clade, there exists two nodes and one sub-clade. The sub-clade is made up of *D. mauritania* and *D. sechellia*. The two nodes are *D. simulans* and *D. melanogaster*. The consensus values along the phylogeny were extremely low. They range from 0–10%. In the bottom clade, there is a sub-clade with a consensus value of 0% (Figure 2). This means that this clade was created because of the computer program and not from genetic data. Overall, between the proximal and the distal phylogeny, there are massive similarities. Both of them share the same overall shape, as they have two main clades. And within these clades, one clade divided into two sub-clades and the other clade divided up into one sub-clade and two nodes. Also, the trees both have low consensus values and both share a sub-clade with a consensus value of 0% (Figure 1–2).

![Figure 1](image_url) Evolutionary relationships of taxa.

The evolutionary history was inferred using the Neighbor-Joining method. The optimal tree with the sum of branch length = 0.12814470 is shown. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The analysis involved 8 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 1678 positions in the final dataset. Evolutionary analyses were conducted in MEGA7.

In the MEGA consensus tree of the proximal and distal phylogenies, it’s shown that most of the nodes create one main clade. There is one sample, *D. ananassae*, that creates a separate node (Figure 3). This was expected as *D. ananassae* was picked to be the outlier sample.

Looking at the main clade, it further divides into two main sub-clades. Within the sub-clades, most of the proximal genes are in the same clade as their distal counterpart. For example, both *D. erecta* versions are within the same sub-clade. However, there are instances within the consensus phylogeny in which the distal and the proximal genes end up on different sub-clades (Figure 3). For example, *D. mauritania* and *D. sechellia* are both within the same sub-clade. However, they are not a proximal and distal pair. Also, *D. simulans* is in a sub-clade with *D. mauritania*. And for most of the phylogeny, the consensus values were low. Most range 2–20%. There is a trend in which proximal genes group with other proximal genes in a clade with little overlap, and vice versa. These consensus values are better than the consensus values that came from the distal and proximal phylogenies individually (Figure 1–3).

Citation: Marzouk SM. The molecular evolution of amylase duplicates genes in *D. melanogaster* group. *Biom Biostat Int J.* 2018;7(5):389–395. DOI: 10.15406/bbij.2018.07.00236
The molecular evolution of amylase duplicates genes in D. melanogaster group

Figure 2 Evolutionary relationships of taxa.

Figure 3 Evolutionary relationships of taxa.

The evolutionary history was inferred using the Neighbor-Joining method. The optimal tree with the sum of branch length = 0.42876639 is shown. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The analysis involved 16 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 1585 positions in the final dataset. Evolutionary analyses were conducted in MEGA7.

In order to have another tree to compare, Phylogeny.fr was used to develop another dependent tree. This tree was extremely similar to the MEGA phylogeny. The main shape of the phylogenies was identical. They both were divided into one big clade and one individual node. In both cases, D. ananassae was the outlier. And the phylogeny goes...
along with the pattern of subdividing into further sub-clades. A difference between the two phylogenies is that there is more overlap between different strains of genes (Figure 3). Whereas the previous phylogeny only had two cases of non-distal/proximal genes being in the same clade, this tree offers more. And the consensus values were much higher than the previous phylogeny. Most of the consensus values are between 66–100%. This phylogeny supports the previous as they share the same shape and overall trends. However, for the rest of the paper, phylogeny in Figure 3 will be the referred phylogeny, as it will be the one used for the HyPhy Bootstrap analysis (Figure 3). Tamura et al. The MEGA phylogeny was chosen as the consensus phylogeny (Figure 4).

Figure 4 Phylogeny from Phylogeny.fr.

The evolutionary history was inferred using the Neighbor-Joining method. The optimal tree with the sum of branch length = 0.8 is shown. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The analysis involved 16 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 1585 positions in the final dataset. Evolutionary analyses were conducted in Phylogeny.fr.

In order to develop an independent phylogeny, this paper built the phylogeny from scratch in Newick format. This tree is based on a previous consensus tree for Drosophila Seetharam & Stuart (Table 1). This is the consensus tree developed for the HyPhy Bootstrap analysis. It represents the independent phylogeny. It is written and involves all of the listed species. It does not include D. ananassae, as the bootstrap analysis does not need and does not consider outliers.

Table 1 Kenwick tree format - Independent tree

(((D_MELANOGASTER_PROXIMAL,D_SIMULANS_PROXIMAL,D_MAURITIANA_P_OXIMAL,
D_SECHELLIA_PROXIMAL),(D_YAKUBA_PROXIMAL,
D_TEISSIERI_PROXIMAL),(D_ORENA_PROXIMAL,
D_ERECTA_PROXIMAL)),(D_MELANOGASTER_DISTAL,
(D_SIMULANS_DISTAL,D_MAURITIANA_DISTAL),(D_YAKUBA_DISTAL,
D_TEISSIERI_DISTAL),(D_ORENA_DISTAL,D_ERECTA_DISTAL)));

After the development of the dependent and independent phylogenies, it was time to organize the data within Hyphy. After uploading the sequences of both the proximal and distal, they had to be partitioned and organized. After which, the program created a tree based on the data of the genetic profiles. This “Created Tree” is the one that was used for the tree morphology. And the substitution model was HKY85. After which, the bootstrap analysis began. In the bootstrap, the possibility of proximal and distal genes evolving dependently was the alternative hypothesis. The possibility of proximal and distal genes evolving independently was the null hypothesis. From the bootstrap, the program calculated a p-value of about 0. The mean was $-1053.34$ and the variance was $22056.4$ from the dataset. The mean here is the average difference between the different tree models. This value indicates that there is extreme similarity between the phylogeny datasets, as it is a extremely high, negative value. And the variation value shows the magnitude of a standard deviation away from the mean. Individual branches from the tree could have different values of significance and the scale of this difference is measured by the variation and standard deviation values. Standard deviation sigma is calculated through the formula $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$, After which, to calculate the p-value, the mean and the standard deviation are put into the z-level formula. $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$. The bootstrap analysis demonstrated
The molecular evolution of amylase duplicates genes in D. melanogaster group

a normal distribution curve. And the scale went increased by a factor 500 for each tick (Figure 5).

![Figure 5 Bootstrap Analysis.](image)

This is the bootstrap that was developed from the HyPhy software. In the bootstrap, the possibility of proximal and distal genes evolving dependently was the alternative hypothesis. The possibility of proximal and distal genes evolving independently was the null hypothesis. From the bootstrap, the program calculated a p–value of about 0. The mean was −1053.34 and the variance was 22056.4. The bootstrap analysis demonstrated a normal distribution curve. And the scale went increased by a factor 500 for each tick.

**Discussion**

From the bootstrap, the p–value calculated was ~0. Since the bootstrap value is less than 0.005, this means that the null hypothesis is rejected. The null hypothesis was that the proximal and distal genes evolved independently. Because it was rejected, this means that the alternative hypothesis is the dominant hypothesis Saito et al.," (Figure 2) (Figure 3). The alternative hypothesis is that the proximal and distal genes evolved dependently. Because the alternative hypothesis is the dominant hypothesis, this means that there is a relationship between the proximal and distal copies of amylase. This signifies that the overall trend is that proximal and distal copies influence each other when it comes to their evolution (Figure 3). From the data, it appears that there is regular transferring of genes between proximal and distal genes. This is because, in the dependent phylogeny, there are clear examples of the distal and proximal genes overlapping and being placed within the same clade (Figure 3). In addition, there appears to be evidence of enough recombination so that mutations can take place. This is demonstrated in the lower sub–clade with *D. sechellia* (proximal and distal) as well as *D. mauritiana* (Figure 2) (Figure 3).

A mutation might have taken place along these genes because of the fact that they have a much lower consensus number than all of the other clades, of about 2% (Figure 3). In addition, there is evidence that these genes have undergone drastic recombination, as they are all included in the same sub–clade. It would be expected that within the clade, it would just be *D. sechellia* (proximal and distal). However, because *D. mauritiana* is in the clade, this shows that a recombination event has happened (Figure 3). However, for the most part, most of the genetic samples do not have mutations. Many of them recombine, but many of them do not recombine (Figure 3). Overall, the trend is not recombining, with the occasional recombining event, and low mutations. From these results, it is possible to estimate phylogenetic trees that are compatible with the known Drosophila phylogeny. This is possible through the usage of the “Maximum Likelihood” statistical model Castresana1 (Figure 3). There are five total parameters to focus on when comparing different phylogenies. These are: mean, variance, rates, differential transformation costs, and the phylogeny shape. The parameters that would yield different results would be the mean and the shape of the phylogeny. By shape here, it means the physical and observable trends of the phylogeny (Figure 3). The shape of the phylogeny would yield different results because different shapes impact all four of the previously stated parameters. The parameters of shape can impact the mean that is calculated from a bootstrap analysis (Figure 3). And the inverse relationship exists, in which the mean can impact the shape of the phylogeny, as the shape reflects the mean (Figure 3). The rates and the differential transformation costs usually do not impact the yield results. And the variance is a byproduct of the mean and the shape (Figure 2) (Figure 3). The variance can be reflective of the shape and mean of a phylogeny. However, it is not usually the cause of these parameters. This placement suggests that there is significant gene commonality, which demonstrates a regular transferring of genes. Overall, this data suggests that dependent evolution of amylase copies is more likely. For future research, more work should be done on developing the relationship between the genes of species in the wider Drosophila family and to identify which coding region the positive selection is acting upon Good et al.1 (Figure 3). A future application of this research can be applied to future analyses done in the amylase regulatory region.1 The same analysis could be done using the amylase regulatory region, which is upstream from the coding region that was discussed in this paper.1 This provides the chance to gain further insights into amylase (Figure 2) (Figure 3). This analysis could provide insight into the relationship between upstream sequences of proximal and distal genes. It could shed more light on multiple alignments and putative regulatory elements. And in future follow–ups, it could focus on the nucleotide divergence that appears to be taking place.12–20

**Acknowledgements**

I would like to thank Dr. Kreitman for his assistance throughout the project. And I would like Dr. Wenhan Chang for his assistance in programming the HyPhy software.

**Conflict of interest**

Author declares that there is no conflict of interest.

**References**


