Scientific Project

Master 1 GENIOMHE

2023–2024

Further development on FTAG Finder, a pipeline to identify Gene Families and Tandemly Arrayed Genes

Samuel Ortion

Advisors:
Carène Rizzon
Franck Samson
Laboratoire de Mathématiques
et Modélisation d'Évry
carene.rizzon@univ-evry.fr
franck.samson@inrae.fr
+33 (0) 1 64 85 35 40
IBGBI
23 Bd. de France
91037 Évry Cedex

keywords: duplicate genes, tandemly arrayed genes, pipeline

Contents

List of Figures

List of Tables

Glossary

allopolyploidisation Polyploidisation with genetic material coming from a diverged species 9

autopolyploidisation Polyploidisation within the same species 9

neofunctionalization Acquisition of a new function by the duplicate gene 13

orthologues Homologous genes whose divergence started at a speciation event 13

polyploidisation Mechanism leading to the acquisition of at least three versions of the same original genome in a species 9

polyspermy Fertilization of an egg by more than one sperm 9

pseudogene A gene-like sequence that lost its capacity to transcribe 13

retroduplication Duplication of a gene through retro-transcription of its RNA transcript 11

segment duplication DNA sequences present in multiple locations within a genome that share high level of sequence identity 11

subfunctionalization Fate of a duplicate gene which gets a part of the original gene function, the function being shared among multiple duplicates 13

Acronyms

GO Gene Ontology 20

TAG Tandemly Arrayed Genes 11, 15, 16

WGD Whole Genome Duplication 9

1 Scientific context

Duplicate genes represent an important fraction of Eukaryotic genes: It is estimated that between 46% and 65.5% of human genes could be considered as duplicate[1] (Correa et al., 2021). Duplicate genes offers a pool of genetic material available for further experimentation during species evolution.

1.1 Gene duplication mechanisms

Multiple mechanisms may lead to a gene duplication. Their effect ranges from the duplication of the whole genome to the duplication of a fragment of a gene.

1.1.1 Whole genome duplication and polyploidisation

During an event of Whole Genome Duplication (WGD), the entire set of genes present on the chromosomes is duplicated (figure 1.1 (A)). WGD can occur thanks to polyspermy or in case of a non-reduced gamete. Polyploidisation is a mechanism leading to a species with at least three copies of an initial genome. A striking example is probably *Triticum aestivum* (wheat) which is hexaploid due to hybridisation events (Golovnina et al., 2007).

We distinguish two kinds of polyploidisations, based on the origin of the duplicate genome: (i) Allopolyploidisation occurs when the supplementary chromosomes come from a divergent species. This is the case for the *Triticum aestivum* hybridisation, which consisted in the union of the chromosome set of a *Triticum* species with that of an *Aegilops* species. (ii) Autopolyploidisation consists in the hybridisation or duplication of the whole genome within the same species.

1.1.2 Unequal crossing-over

Another source of gene duplication relies on unequal crossing-over. During cell division, a crossing-over occurs when two chromatids exchange fragments of chromosome. If the cleavage of the two chromatids occurs at different positions, the shared fragments may have different

[1] The estimate vary strongly depending on the criteria in use, because ancient duplication event may be hard to detect.

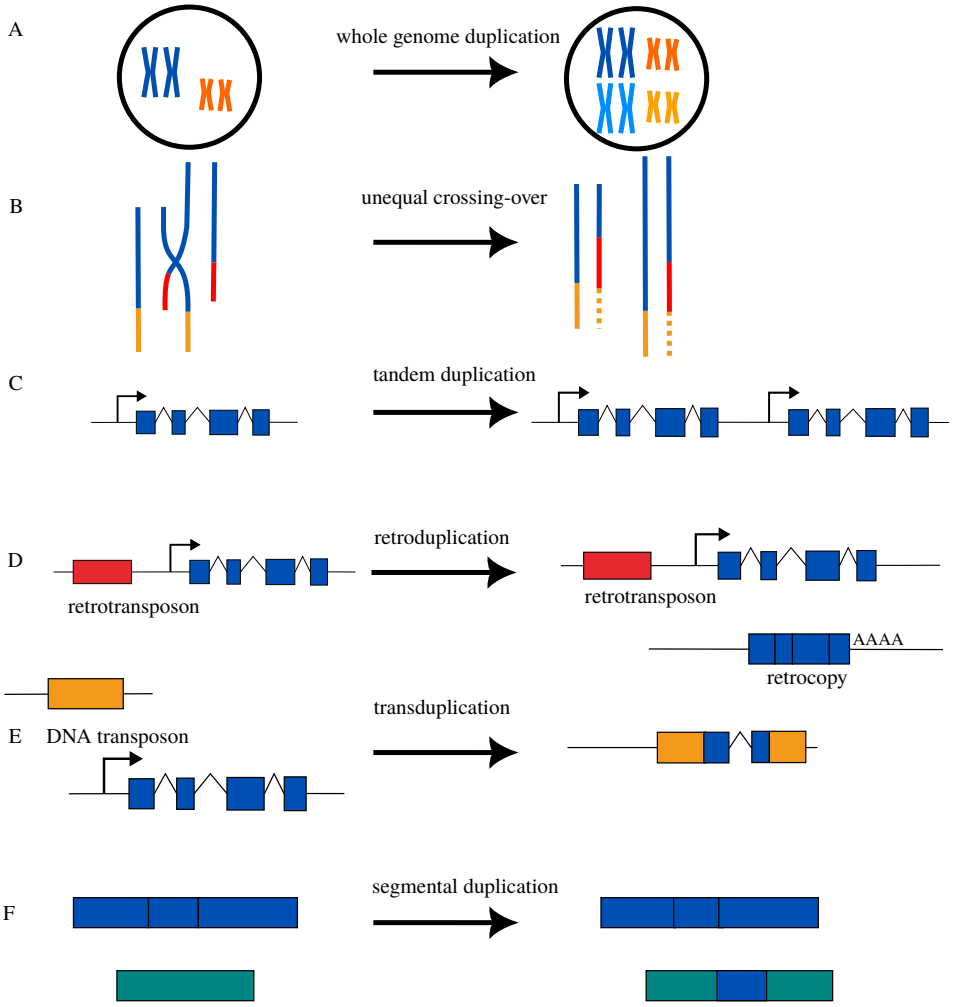

Figure 1.1. Different types of duplication. (A) Whole genome duplication. (B) An unequal crossing-over leads to a duplication of a fragment of a chromosome. (C) In tandem duplication, two (set of) genes are duplicated one after the other. (D) Retrotransposon enables retroduplication: a RNA transcript is reverse transcribed and inserted back without introns and with a polyA tail in the genome. (E) A DNA transposon can acquire a fragment of a gene. (F) Segmental duplication corresponds to long stretches of duplicated sequences with high identity. Adapted from (LALLEMAND et al., 2020) (fig. 1).

lengths. Homologous recombination of such uneven crossing-over leads to the incorporation of a duplicate region, as depicted in figure 1.1 (B, C). This mechanism leads to the duplication of the whole set of genes present in the fragment. These duplicate genes locate one set after the other: we call them Tandemly Arrayed Genes (TAG). TAG are the kind of gene duplication we will be particularly interested in during this internship.

1.1.3 Retroduplication

Transposable elements play a major role in genome plasticity, and enable gene duplication too. Retrotransposons, or RNA transposons are one type of transposable elements. They share similar structure and replication mechanisms with retroviruses. Retrotransposons replicate in the genome through a mechanism known as "copy-and-paste". These transposons typically contain a reverse transcriptase gene. This enzyme proceeds in the reverse transcription of an mRNA transcript into its reverse complementary DNA sequence which can then insert elsewhere in the genome. More generally, retroduplication refers to the duplication of a sequence through reverse transcription of a RNA transcript. Genes duplicated through retroduplication lose their intronic sequences and bring a polyA tail with them in their new locus (figure 1.1 (D)).

1.1.4 Transduplication

DNA transposons are another kind of transposable elements whose transposition mechanism can also lead to gene duplication. This type of transposable element moves in the genome through a mechanism known as "cut-and-paste". A typical DNA transposon contains a transposase gene. This enzyme recognizes two sites surrounding the donnor transposon sequence in the chromosome resulting in a DNA cleavage and an excision of the transposon. The transposase can then insert the transposon at a new genome locus. A transposon may bring a fragment of a gene during its transposition in the new locus (figure 1.1 (E)), leading to the duplication of this fragment.

1.1.5 Segment duplication

Finally, segment duplications, also called *low copy repeats* are long stretches of DNA with high identity score (figure 1.1 (F)). Their exact duplication mechanism remains unclear (Lallemand et al., 2020). They may come from an accidental replication, distinct from an uneven cross-over or a double stranded breakage. Transposable elements may well be involved in the mechanism, as a high enrichment of transposable elements is found next to duplicate segment extremities, in *Drosophila* (Lallemand et al., 2020).

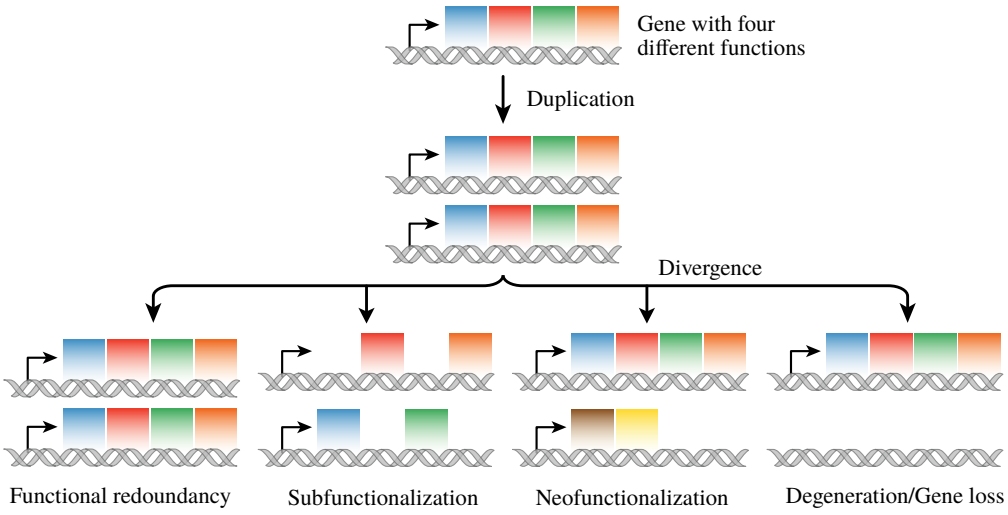

Figure 1.2. Fate of duplicate genes. An original gene with four functions is duplicated. Its two copies may both keep the original functions (functional redoundancy). The original functions may split between the different copies (subfunctionalization). One of the copy may acquire a new function (neofunctionalization). It may also degenerate and lose its original functions (pseudogenization).
Adapted from Smedlib, CC BY-SA 4.0, via Wikimedia Commons

1.2 Fate of duplicate genes in genome evolution

In his book *Evolution by Gene Duplication*, Susumu Ohno proposed that gene duplication plays a major role in species evolution (Ohno, 1970), because it provides new genetic materials to build on new phenotypes while keeping a backup gene for the previous function. Indeed, duplicate genes evolve after duplication: they may be inactivated, and become pseudogenes; they may be deleted or conserved, and if conserved, the may or may not acquire a new function. Figure 1.2 depicts the different possible fates of a duplicate gene.

As genome evolves, duplicate genes may be inactivated and become pseudogenes. These pseudogenes keep a gene-like structure which degrades as and when further genome modifications occur but they are no longer expressed.

After duplication, the new gene copy may gain a new function. We call this possible outcome neofunctionalization. For instance, the current set of olfactory receptor genes result from several duplication and deletion events (for *Drosophila*, see: Nozawa and Nei (2007)), after which each duplicate olfactory gene specialized in the detection of a particular chemical compound.

Two duplicate genes with the same original function may encounter a subfunctionalization: each gene conserves only one part of the function.

Another possibility is that the two gene copies keep the ancestral function, resulting in a functional redoundancy. In this case the quantity of gene product may increase.

1.3 Methods to identify duplicate genes

Different methods exists to detect duplicate genes. These methods depend on the type of duplicate genes they target and vary on computation burden as well as in the ease of use (for a review, see Lallemand et al. (2020)).

1.3.1 Paralog detection

Paralogs are homologous genes derived from a duplication event. We can identify them as homologous genes coming from the same genome, or as homologous genes between different species once we filtered out orthologues (homologous genes derived from a speciation event).

We can use two gene characteristics to assess the homology between two genes: gene structure or sequence similarity. The sequence similarity can be tested with a sequence alignment tool, such as `BLAST` (Altschul et al., 1990), `Psi-BLAST`, and `HMMER3` (Johnson et al., 2010), or `diamond` (Buchfink et al., 2021), which are heuristic algorithms, which means they may not provide the best results, but do so way faster than exact algorithms, such as the classical Smith and Waterman algorithm (Smith and Waterman, 1981) or its optimized versions `PARALIGN`

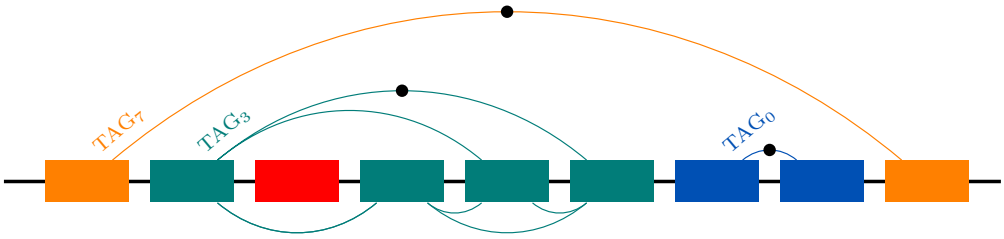

Figure 1.3. Schematic representation of TAG definitions. Several genes are represented on a linear chromosome. The red box represent a singleton gene. Orange boxes represent a TAG with two duplicate genes seperated by 7 other genes (TAG_7). Four green boxes constitute a TAG, the gene at the extremities are seperated by three genes (TAG_3). The two blue boxes represents a TAG with two genes next to each other (TAG_0). The bended edges represents the homology links between each pair of genes within a TAG.

(Rognes, 2001) or SWIMM. This is the case for Triticum aestivum hybridisation, which consisted in the union of the chromosome set of a Triticum species with those of an Aegilops species

1.3.2 FTAG Finder

Developed in the LaMME laboratory, the FTAG Finder (Families and Tandemly Arrayed Genes Finder) pipeline is a simple pipeline targeting the detection of TAG from the proteome of single species (Bouillon et al., 2016).

The pipeline proceeds in three steps. First, it estimates the homology links between each pair of genes. Then, it deduces the gene families. Finally, it searches for TAG.

Estimation of homology links between genes

This step consists in establishing a homology relationship between each genes in the proteome. In this step, the typical tool involved is BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990) run "all against all" on the proteome.

Several BLAST metrics can be used as an homology measure, such as bitscore, identity percentage, E-value or variations of these. The choice of metrics can affect the results of graph clustering in the following step, and we should therefore chose them carefully (Gibbons et al., 2015).

Identification of gene families

Based on the homology links between each pair of genes, we construct an undirected weighted graph whose vertices correspond to genes and edges to homology links between them. We apply a graph clustering algorithm on the graph in order to infer the gene families corresponding to densely connected communities of vertices. FTAG Finder proposes three clustering algorithm alternatives: single linkage, Markov Clustering (van Dongen, 1998) or Walktrap (Pons and Latapy, 2005).

Detection of TAGs

The final step of FTAG Finder consists in the identification of TAG from the gene families and the positions of genes. For a given chromosome, the tool seeks genes belonging to the same family and located close to each other. The tool allows a maximal number of genes between the homologous genes, with a parameter set by the user. Figure 1.3 is a schematic representation of some possible TAG positioning on a genome associated with their definition in FTAG Finder *Find Tags* step.

2 Objectives for the internship

2.1 Scientific questions

The underlying question of FTAG Finder is the study of the evolutionary fate of duplicate genes in Eukaryotes. Duplicate genes are

2.2 Extend the existing FTAG Finder Galaxy pipeline

Galaxy is a web-based platform for running accessible data analysis pipelines, first designed for use in genomics data analysis (Goecks et al., 2010). Last year, Séanna Charles worked on the Galaxy version of the FTAG Finder pipeline during her M1 internship (Charles, 2023). I will continue this work. FTAG Finder is currently deployed on the server of the *Laboratoire d'Analyse et Modélisation d'Évry*[fn: http://stat.genopole.cnrs.fr/galaxy].

2.3 Port FTAG Finder pipeline on a workflow manager

Another objective of my internship will be to port FTAG Finder on a workflow manager better suited to larger and more reproducible analysis.

We will have to make a choice for the tool we will use. The two main options being Snakemake and Nextflow. Snakemake is a python powered workflow manager based on rules *à la* GNU Make (Köster and Rahmann, 2012). Nextflow is a groovy powered workflow manager, which rely on the data flows paradigm (Di Tommaso et al., 2017). Both are widely used in the bioinformatics community. Their use have been on the rise since they came out in 2012 and 2013 respectively (Djaffardjy et al., 2023).

	List ref	List L
related to go	a	b
unrelated	c	d

Table 2.1. Contingency table for a Fisher exact test on gene lists

3 Methodological approaches

Based on the output of the FTAG Finder pipeline, which consist in lists of genes, researchers could perform bespoke subsequent analyses on TAGs.

3.1 Analysis of over-represented gene functions among TAGs

The Gene Ontology (GO) describes biological concepts across three main classes: Cellular Component, Molecular Function and Biological Process. It describe a controlled vocabulary of concepts and the relationship between them. The genes with known functions can be associated with a particular GO term. We can perform an GO enrichment analysis to assess whether a particular GO term is over-represented in a particular gene list, compared to an other. We can use a Fisher exact test, using the FDR (False Discovery Rate) control procedure of Benjamini and Hocheberg to do so.

Let *go* be a GO term. We construct a contingency matrix based on the count of genes associated with this GO term (or associated with one of its brother GO term) for the reference gene list and the list of interest (here, the list of genes in a TAG) (see table 2.1).

3.2 Are TAG located preferentially on specific chromosome region?

3.3 Are there chromosomes enriched or depleted in TAG?

3.4 Do genes located next to each other in a TAG share the same orientation?

The concordance of two genes of a TAG falls in three possible cases: either both genes are on the same strand ($\rightarrow\rightarrow$), either they have a divergent orientation ($\leftarrow\rightarrow$), or a convergent one ($\rightarrow\leftarrow$). Graham conjectured that genes of a TAG that are close to each other would be more likely to share the same orientation, and it seems to be effectively the case (Shoja and Zhang, 2006).

3.4.1 TODO write down the hypotheses

3.5 What is the robustness and accuracy of the detection method?

Lê-Hoang (2017) started analyses of the impact of parameter choice on FTAG Finder output lists. A more detailed benchmark of FTAG Finder in comparison with other methods on some known test dataset might be of particular interest.

References

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (Oct. 5, 1990). "Basic Local Alignment Search Tool". In: *Journal of Molecular Biology* 215.3, pp. 403–410. issn: 0022-2836. doi: 10.1016/S0022-2836(05)80360-2. url: https://www.sciencedirect.com/science/article/pii/S0022283605803602 (visited on 04/30/2023).

Bouillon, Bérengère, Franck Samson, Etienne Birmelé, Loïc Ponger, and Carène Rizzon (2016). *FTAG Finder: Un Outil Simple Pour Déterminer Les Familles de Gènes et Les Gènes Dupliqués En Tandem Sous Galaxy*.

Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost (Apr. 2021). "Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND". In: *Nature Methods* 18.4, pp. 366–368. issn: 1548-7105. doi: 10.1038/s41592-021-01101-x. url: https://www.nature.com/articles/s41592-021-01101-x (visited on 03/28/2024).

Charles, Séanna (2023). *Finalisation du pipeline FTAG (Families and TAG) Finder, un outil de détection des gènes dupliqués sous Galaxy*. Internship Report. Laboratoire de Mathématiques et Modélisation d'Évry.

Correa, Margot, Emmanuelle Lerat, Etienne Birmelé, Franck Samson, Bérengère Bouillon, Kévin Normand, and Carène Rizzon (May 1, 2021). "The Transposable Element Environment of Human Genes Differs According to Their Duplication Status and Essentiality". In: *Genome Biology and Evolution* 13.5, evab062. issn: 1759-6653. doi: 10.1093/gbe/evab062. url: https://doi.org/10.1093/gbe/evab062 (visited on 09/15/2023).

Di Tommaso, Paolo, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame (Apr. 2017). "Nextflow Enables Reproducible Computational Workflows". In: *Nature Biotechnology* 35.4, pp. 316–319. issn: 1087-0156, 1546-1696. doi: 10.1038/nbt.3820. url: https://www.nature.com/articles/nbt.3820 (visited on 03/27/2024).

Djaffardjy, Marine, George Marchment, Clémence Sebe, Raphael Blanchet, Khalid Bellajhame, Alban Gaignard, Frédéric Lemoine, and Sarah Cohen-Boulakia (2023). "Developing and Reusing Bioinformatics Data Analysis Pipelines Using Scientific Workflow Systems". In: *Computational and Structural Biotechnology Journal* 21, p. 2075. doi: 10.1016/j.csbj

.2023.03.003. pmid: 36968012. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10030817/ (visited on 03/26/2024).

Gibbons, Theodore R., Stephen M. Mount, Endymion D. Cooper, and Charles F. Delwiche (Dec. 2015). "Evaluation of BLAST-based Edge-Weighting Metrics Used for Homology Inference with the Markov Clustering Algorithm". In: *BMC Bioinformatics* 16.1, p. 218. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0625-x. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0625-x (visited on 03/19/2024).

Goecks, Jeremy, Anton Nekrutenko, James Taylor, and Galaxy Team (2010). "Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences". In: *Genome Biology* 11.8, R86. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-8-r86. pmid: 20738864.

Golovnina, K. A., S. A. Glushkov, A. G. Blinov, V. I. Mayorov, L. R. Adkison, and N. P. Goncharov (Apr. 1, 2007). "Molecular Phylogeny of the Genus Triticum L". In: *Plant Systematics and Evolution* 264.3, pp. 195–216. ISSN: 1615-6110. DOI: 10.1007/s00606-006-0478-x. URL: https://doi.org/10.1007/s00606-006-0478-x (visited on 03/27/2024).

Johnson, L. Steven, Sean R. Eddy, and Elon Portugaly (Aug. 18, 2010). "Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure". In: *BMC Bioinformatics* 11.1, p. 431. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-431. URL: https://doi.org/10.1186/1471-2105-11-431 (visited on 04/09/2024).

Köster, Johannes and Sven Rahmann (Oct. 1, 2012). "Snakemake–a Scalable Bioinformatics Workflow Engine". In: *Bioinformatics (Oxford, England)* 28.19, pp. 2520–2522. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts480. pmid: 22908215.

Lallemand, Tanguy, Martin Leduc, Claudine Landès, Carène Rizzon, and Emmanuelle Lerat (Sept. 4, 2020). "An Overview of Duplicated Gene Detection Methods: Why the Duplication Mechanism Has to Be Accounted for in Their Choice". In: *Genes* 11.9, p. 1046. ISSN: 2073-4425. DOI: 10.3390/genes11091046. URL: https://www.mdpi.com/2073-4425/11/9/1046 (visited on 03/19/2024).

Lê-Hoang, Julie (2017). *Etude transcriptomique des gènes dupliqués en tandem (TAG) chez Arabidopsis thaliana*. Internship Report. Laboratoire de Mathématiques et Modélisation d'Évry, p. 74.

Nozawa, Masafumi and Masatoshi Nei (Apr. 24, 2007). "Evolutionary Dynamics of Olfactory Receptor Genes in Drosophila Species". In: *Proceedings of the National Academy of Sciences* 104.17, pp. 7122–7127. DOI: 10.1073/pnas.0702133104. URL: https://www.pnas.org/doi/full/10.1073/pnas.0702133104 (visited on 04/02/2024).

Ohno, Susumu (1970). *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-86661-6. DOI: 10.1007/978-3-642-86659-3. URL: http://link.springer.com/10.1007/978-3-642-86659-3 (visited on 03/21/2024).

Pons, Pascal and Matthieu Latapy (Dec. 12, 2005). *Computing Communities in Large Networks Using Random Walks (Long Version)*. DOI: 10.48550/arXiv.physics/0512106. arXiv: physics/0512106. URL: http://arxiv.org/abs/physics/0512106 (visited on 03/30/2024). preprint.

Rognes, Torbjørn (Apr. 1, 2001). "ParAlign: A Parallel Sequence Alignment Algorithm for Rapid and Sensitive Database Searches". In: *Nucleic Acids Research* 29.7, pp. 1647–1652. ISSN: 0305-1048. DOI: 10.1093/nar/29.7.1647. URL: https://doi.org/10.1093/nar/29.7.1647 (visited on 04/09/2024).

Shoja, Valia and Liqing Zhang (Nov. 2006). "A Roadmap of Tandemly Arrayed Genes in the Genomes of Human, Mouse, and Rat". In: *Molecular Biology and Evolution* 23.11, pp. 2134–2141. ISSN: 0737-4038. DOI: 10.1093/molbev/msl085. pmid: 16901985.

Smith, T. F. and M. S. Waterman (Mar. 25, 1981). "Identification of Common Molecular Subsequences". In: *Journal of Molecular Biology* 147.1, pp. 195–197. ISSN: 0022-2836. DOI: 10.1016/0022-2836(81)90087-5. URL: https://www.sciencedirect.com/science/article/pii/0022283681900875 (visited on 04/29/2023).

Van Dongen, S. (Jan. 1, 1998). "A New Cluster Algorithm for Graphs". In: R 9814. URL: https://ir.cwi.nl/pub/4604 (visited on 03/22/2024).

Summary

Duplicate genes is an important feature of Eukaryotic genomes. They contribute to the plasticity of genome, hence to the capacity of species to evolve.

Several mechanisms may lead to gene duplication. Among them, an unequal crossing-over leads to the formation of Tandemly Arrayed Genes (TAG) corresponding to homologous genes located one set after the other on the same chromosome.

There are multiple methods for detecting duplicate genes from sequences. These methods vary in terms of the particular gene duplication mechanism they target, computational efficiency and ease of use.

FTAG Finder is a simple Galaxy pipeline aiming at the detection of families of duplicate genes and the identification of TAG based on the proteome of a single species. FTAG Finder is developed in the *Laboratoire de Mathématiques et Modélisation d'Évry*, where I will do my internship.

One the one hand, the aim of my internship is to extend the current Galaxy implementation of FTAG Finder with new export lists best suited to the analysis requirements of the laboratory. On the other hand, the objective of my internship will be to port the Galaxy pipeline on another scientific workflow manager better suited to reproducible analyses such as Snakemake and Nextflow.

Then, the updated version of the FTAG Finder pipeline will be used to perform an analysis on the TAG of a model species, to assess its proper behavior. A benchmark of the pipeline will probably be run to compare the FTAG Finder with alternative published methods targetting duplicate genes and TAG in particular.