BIO2 CW:
James Wilsench - s1666320

March 28, 2017

Introduction

We were provided with two sequences from an unknown plant species *contig_28799* (Sequence 1) and *contig_326* (Sequence 2) with the objective of isolating and characterising potential protein coding genes within them. We also investigated the possible cladistics of the species of origin. These gene models were constructed using specialised software that use Hidden Markov Models (HMMs) to predict the location of important gene features such as transcriptional start and stop sites as well as intronic and exonic boundaries.

HMMs are a class of graphical model which uses a supervised learning approach. The HMM gene finding approach generally requires a large body of known sequence data already annotated with appropriate exons, introns and other features in order to make predictions. It is therefore necessary to identify related species with well annotated gene sets which we found through general and sub-sequence specific blast queries. This process also helps in identifying possible candidates for the species to which these sequences belong.

This report is broken up into two main sections followed by a discussion, with the primary objective of characterising and identifying the origins of the two sequences which share the same organismic source. In Sec. 1 we examined the origin of the first sequence using direct Gene finding, sequence analysis and regulatory motif finding tools including miRNA analysis. In Sec. 1.5 we examined a potential non-nuclear sequence and attempt to characterise its origin and function. In Sec. 3 we address interesting combined findings and possible methodological faults.

1 Gene Finding & Analysis: Sequence 1

The gene-finding procedure covered here involves the integration of multiple gene-finding resources and soft-ware packages. These include sequence alignment searches using NCBI Blast and Ensembl plant in combination with gene finding software GeneMark, GENSCAN and FGENESH

1.1 Procedural Overview

In order to determine the origin of the sequence provided we performed an initial NCBI BLASTN using NCBI plant sequence database. From this search we obtained an impression of the cladistics of the species in question and used this to inform our gene searches using the three gene finding tools.

Using these tools we recursively performed NCBI and Ensemnl BLAST alignments on predicted gene elements (exons, introns and inter-genic regions), returning to the gene prediction step when a more accurate species to use as a gene model was located. We then used these updated gene models to search new regions for possible alignment. The end result was an overview of several hot-spots containing strong domain signals, indicating possible exonic regions of the gene.

Finally, in order to identify possible regulatory elements we looked for short, highly conserved regions in intronic regions in both the gene model for *contig_28799* and aligned sequences. Although we did find such potential elements in our sequence, this method may lead to some inaccuracies since regulatory sequences and the transcription factors that bind to them, diversify rapidly, particularly in plants [18, 24].

1.2 Initial Sequence Analysis & Cladistics

Figure 1 shows the results of constructing a phylogenetic distance tree based on the 'Build Tree' option following an BLASTN query using NCBI Plant [1, 4]. All of the significant hits shown in the tree are from members of the eudicot clade, more specifically rosids, a large clade of flowering plants. The tree includes chromosomal and mRNA hits from a number of model organisms including (in order of closest to furthest alignment): *Medicago truncatula*, *Arabidopsis thaliana* and *Lotus japonica*. The mRNA hits suggest some sub-sequences may be involved or be the targets of transcriptional regulation (C2B2 type Zinc Finger protein family) and/or involved in metabolic pro-

cesses (phosphodiesterase superfamily - GDPD2).

This search suggests a number of possible candidates on with which to

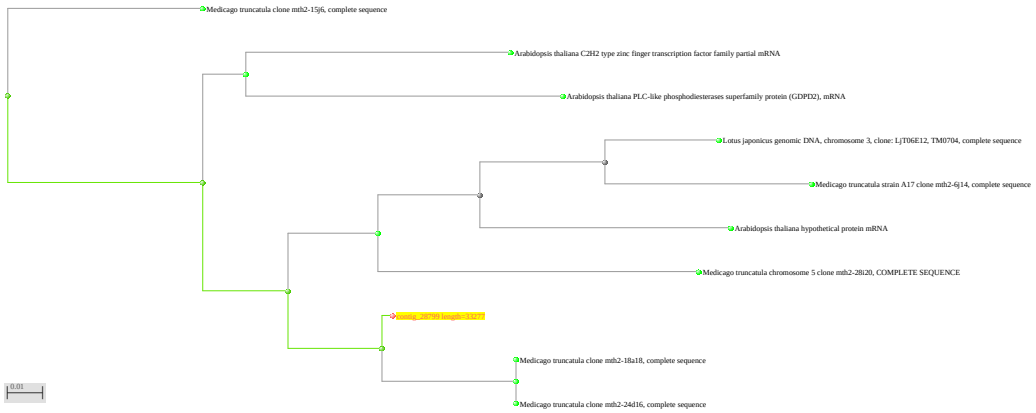

Figure 1: Phylogeny constructed from initial BLASTN query shows results in the eudicot

parametrise gene finding tools. The most generally well understood species in the list which we posit should have the best annotated genome is *A. thaliana* which is the species we used in our initial gene finding efforts below.

1.3 Investigating First Gene Model Based on *A. thaliana*

Table 1: HMM Gene Finding Results (*A. thaliana* model)

Software	Genes	Exons	Start	End	HS
FGENESH	5	2,1,5,4,8	65	33277	28462-30447
GENSCAN	4	1,1,6,8	800	33215	28474-30469
GeneMark	6	2,2,2,5,10,2	55	32458	N/A

Table 1 shows the results of running the three different gene finding programs FGENESH [13], GENSCAN [5] and GeneMark [12] which all use HMM-based methods to predict possible gene features. The Table shows how even methods using a similar model structure can give different results with differing numbers of exons (Exons), start (Start) and stop (Stop) sites and even number of predicted genes (Genes). However there is some agreement regarding the direction of transcription (reverse) . There is also consensus between the two methods that provide log-odds scores as to the exonic region with the highest score (HS) in that both suggest that it lies in a similar range

towards the end of the contiguous sequence extract.

The Highest Scoring region codes for a long hypothetical exonic sequence in both models. BLASTX searches using this sub-sequence on Ensembl Plant reveals that it may be a possible orthologue for the gene MEE44 a nucleotidyl polymerase-β gene in *A. thaliana* which is orthologous to other coding sequences with low search E-values (<1E-100) including PRUPE .

Table 2: Top 10 NCBI BLASTX Results in Dicots for 28462-30469

Species	Accession	ID	Pos.	Fr.	Ann.
Juglans regia	XP_018820264.1	70	77	-2	N/A
Jatropha curcas	XP_012083850.1	62	75	-2	NTT
Cephalotus follicularis	GAV62042.1	63	75	-2	N/A
Manihot esculenta	OAY49354.1	63	74	-2	PRUPE
Prunus persica	ONI22581.1	64	76	-2	N/A
Malus domestica	XP_008341655.1	62	74	-2	polβ
Citrus sinensis	XP_006490855.1	61	71	-2	N/A
Citrus clementina	XP_006445325.1	61	71	-2	polσ
Vitis vinifera	CAN65347.1	61	71	-2	RT
Theobroma cacao	EOX96314.1	60	70	-2	PAP

PAP - PolyA Polymerase, RT - Reverse Transcriptase, NTT - Nucleotidyl Transferase, pol - Polymerase

BLASTX Search for HS Exonic Sequence

Table 2 provides a number of hits with E-values very near to zero, ordered in descending match confidence which resulted from a NCBI BLASTX search based on the union of the two high scoring exonic regions in Table 1. The results show proteins with interrelated annotations (Ann.) most of which relate to DNA repair (DNA Polymerase β and Nucleotidyltransferase 2) but also sequences that may be involved in mRNA poly-adenalation, one of the genes associated with the *Vitis vinifera* also shows some evidence of reverse transcriptase (RT) activity, but there was no overlap with this domain in the actual alignment (though this could result from splicing).

BLASTN Search using related Species in Ensembl Planr

We also performed a whole sequence BLASTN alignment search using Ensembl Plant against all species from Table 2 that were available in the Ensembl database as well as *Populus trichocarpa*, another model organism with

significant sequence identity (60%) and coverage (99%) for the HS region in the NCBI BLASTN results. The results, while broadly similar to the narrower NCBI search show interesting areas of possible orthology elsewhere in the sequence. This includes significant (<1E-3, > 65% ID) but short matches with YCF-1 (*V. vifera*) and YCF2-B (*A. thaliana*), two chloroplastic genes [14] with predicted Nucleoside triphosphate hydrolase activity [3]. It also includes a significant match (E-value<1E-6) with an F-box family protein in *A. thaliana* .

Summary of Initial Model Results

Table 3 gives a summary of suspected domains and regions of possible gene activity in the sequence of innterest following the initial investigation. These include nucleotidyl transferase and possible chloroplastic genes for Nucleoside triphosphate hydrolase. In the next section we refined the search by looking for more specific parameter species on which to apply our HMM gene finding tools.

Table 3: Summary of Possible Orthologues and Domains

Orthologues	Domains	Start	Stop
ME44/PRUPE	polβ,polσ,RT,PAP	28491	30447
FBX8	FBOX	6867	6992
YCF1/2	NTH	12876	13081

NTH - P-loop Nucleoside Triphosphate Hydrolase, FBOX - F-Box Domain

1.4 Refining Search Based on *J. regia* Gene Model

Juglans regia also known as the Persian or English walnut tree was the highest scoring match in both the NCBI full sequence BLASTN search and the exonic subsequence BLASTX search in Sec 1.3. This suggests that use of *Juglans regia* as a parameter species for HMM training may provide better results than the original *A. thaliana* approach, however this option was not available for all HMM tools, notably GeneMark and GENSCAN, and thus results were limited to the remaining package, FGENESH.

The truncated FGENESH results in Table 4 present a slightly more limited list of possible gene features with only four genes predicted, however, all of the subsequences so far implicated are covered by one or more of the predicted exonic regions except for those overlapping with the chloroplastic YCF1/2 genes which exist between the transcriptional start site (TSS) and only exon

Table 4: FGENESH Gene Model Based on *J. regia*

Gene	Feature	Start	End	Score
1	PolA	211		0.01
1	CDSo	681	1046	17.49
1	TSS	1873		-2.7
2	PolA	5517		2.2
2	CDSo	5856	7040	53.91
2	TSS	7111		0.49
3	PolA	8026		2.2
3	CDSl	8469	8514	-1.43
3	CDSi	11360	11457	4.39
3	CDSi	13522	13601	-0.49
3	CDSi	18771	18879	4.45
3	CDSi	19559	19643	3.95
3	CDSi	19758	19835	2.79
3	CDSi	19986	20074	-0.43
3	CDSf	20713	20823	6.97
3	TSS	21187		0.4
4	PolA	21237		2.2
4	CDSl	21309	21374	-6.57
4	CDSi	23470	23529	2.43
4	CDSi	23740	23859	15.49
4	CDSi	25315	25578	4.22
4	CDSi	27758	27820	2.36
4	CDSi	27904	28086	1.54
4	CDSi	28260	28376	7.75
4	CDSi	28462	30447	131.85
4	CDSi	30559	30672	2.57
4	CDSi	30770	31312	11.34
4	CDSi	32458	33277	87.48

TSS - Transcriptional Start Site, CDSi - Internal CDS, CDSf - First CDS, CDSl - Last CDS, CDSo - Only CDS, PolA - PolyA tail

of predicted Gene 1. Note that a possible TSS is predicted for three of these genes which could be at or near the site of potential regulatory elements. Also note that for predicted Gene 4 only a partial CDS is recovered. We chose to investigate each predicted gene one at a time using a combination of NCBI and Ensembl tools and related databases. Although we do discuss the details of their individual function here, functional information is used to support arguments as to whether a predicted gene is likely to be accurate.

Investigation of Predicted Gene 1

An NCBI BLASTX search on the entire genomic sequence spanned by Gene 1 (TSS to PolA), limited to the more specific rosid clade (taxid: 71275), revealed a another reverse transcriptase domain (E-value<1E-47), this time in the forward direction (see Table 5). This strongly suggests some degree of genomic interference due to viral or transposon activity. A similar NCBI BLASTP search (see Table 6) with the same organismic filter gave a hypothetical protein with a reverse transcriptase zinc binding domain as its top hit (E-value<1E-19). The overall high level of consensus with known and suspected protein coding sequences with RT domains suggests that Gene 1 is or was recently a protein coding gene that may have some reverse transcriptase activity.

Investigation of Predicted Gene 2

An NCBI BLASTX search on the entire genomic sequence spanned by Gene 2 (TSS to PolA), limited to the more specific rosid clade (taxid: 71275), returned a similar result to the whole sequence Ensembl search which suggested a possible F-box family orthologue in *A. thailiana*. A similar result is obtained here but this time for another potential *J. regia* F-box orthologue (E-value<1E-55). This time the BLASTP search returned another F-box protein, also in *J. regia* (E-value<1E-58). In addition, the predicted homollogues also consist of a single exon. This strongly suggests that Gene 2 encodes an F-box protein making it a regulatory protein, possibly involved in protein degradation [15].

Investigation of Predicted Gene 3

Following the same procedure for Gene 3 as for 1 and 2 we found in the BLASTX search that the predicted gene region is much larger than its closest hit resulting in very low coverage of only 4%. However, the aligned region did receive a low E-value (<1E-21) and high identity (43%) suggesting that the hit is genuine. The annotations for the aligned predicted protein (see

Table 5) suggest that it is yet another Nucleotidyl Transferase. The first BLASTP result is for an uncharacterised, predicted *J. regia* protein with 92% identity but only 45% coverage. These results suggest that while Gene 3 may encode a real polymerase protein, the overall gene model may require some fine tuning.

Investigation of Predicted Gene 4

Predicted Gene 4 is almost certainly a Nucleotidyl transferase protein with potential polyA polymerase capabilities since the BLASTX results indicate a strong match (E-value$\approx$ 0) with reasonable coverage (30%) and high sequence identity (62%). The initial BLASTP results are inconclusive but suggest a near complete alignment (99& coverage) with an uncharacterised *J. regia* protein. Although incomplete, Gene 4 is or was in the recent past likely to be a polymerase or other protein involved in the transfer of nucetides.

Table 5: Top Annotated Results of NCBI BLASTX for Predicted Genes

Gene	Species	Start	Stop	Coverage	ID	Ann.
1	*C. capsularis*	222	876	65	33	RT
2	*J. regia*	5865	6992	70	38	FBOX
3	*G. max*	18760	20029	4	43	NTT/Polyβ/σ/PAP
4	*J. regia*	25312	30468	30	62	NTT/polyβ

Table 6: Top Results of NCBI BLASTP for Predicted Gene Products

Gene	Species	Length	Start	Stop	Coverage	ID	Ann.
1	*C. olitorius*	121	5	114	90	37	RT
2	*J.regia*	394	17	392	95	38	FBOX
3	*J. regia*	231	53	157	45	92	-
4	*J. regia*	1444	2	1434	99	77	-

1.4.1 Consensus Species

Initial investigation in Sec. 1.3 showed that the sequence of interest likely bwlonged to a eudicot plant, more specifically a rosids. The rosids form a large family of flowering plants (angiosperms) that include many of the worlds crop plants and trees [16]. Judging from the number of very strong hits found for *J. regia*, with significant results for all 4 postulated genes, it seems most likely that the sequence belongs to J. regia or a closely related

species in a nearby clade. The Juglans genus is itself large containing all known true species of walnut tree with *J. regia* being the most well-known member. It may therefore be the case that a member of this or a related plant genus may include the species of interest

1.5 Structure and Function of Predicted Gene 2

In Sec. 1.4 we inferred that Gene 2 is most probably a member of the family of proteins containg F-box protein domains, specialised domains that are thought to be involved in the regulation of proteins through ubiquitin degradation pathways [8]. In plants, these proteins are themselves generally regulated by hybridising short micro RNAs (miRNAs) which work in concert with protein complexes to cleave target mRNAs [10]. Because of their supposed complementarity, with the their mRNA, such an miRNA might be detected as a short alignment sequence (as miRNAs need not be co-located with their targets) or using a miRNA prediction tool. We chose to use psRNATarget, a web based tool for miRNA prediction using an miRNA database [6].

The initial psRNATarget search using the single predicted exon of Gene 2 yielded no results, however, miRNAs are often well-conserved between closely related species. We thus chose to use the mRNA of a possible *A. thaliana* orthologue (the top result of a direct *A. thaliana versus Gene 2 BLASTX search*) as a target. Unfortunately, no matches could be found, however there is still some possibility that regulatory sequences exist at or near to the TSS or promoter region.

Next, we investigated regulation using the NSITE-PL tool for predicting promoter and regulatory elements based on the TSSW architecture [17, 20]. We used the entire TSS to PolyA sequence with 1000bp of padding on the TSS end to look for potential upstream regulatory regions and 2000bp on the PolA end in search of downstream elements. After reversing the strand to read in the direction of translation, the test provides a list of possible transcription factor binding sites that are over represented at a 5% (by default) level of significance.

Of those sites returned by the NSITE-PL search 7, only five were in non-coding regions, located either up or downstream of the CDSo. Of these, two of them belong to the MYB family of transcription factors, a large superfamily of known plant TFs which are known to be involved in cell-cycle regulation [21]. The GAGA family are a collection of zinc-finger TFs which have been implicated in processes unique to plants such as flowering time [22]. The WRKY superfamily of plant transcription factors are involved in many pro-

cesses such as heat stress response [7].The only potential upstream enhancer region is a possible bZIP binding site. bZIP proteins form a very large family of TFs that span all eukaryotes. In *A. thaliana* bZIP TFs are known to be involved in various pathways including light and stress signalling [9].

Table 7: Detected Upstream and Downstream TF Motifs

Species	TF Family	Start	End
Pisum sativum	MYB	7891	7879
Lycopersicon esculentum	MYB	7614	7595
Glycine max	GAGA	7508	7524
Catharanthus roseus	WRKY	7538	7523
Arabidopsis thaliana	bZIP	3870	3878

2 Analysis of Sequence 2

An initial BLASTN NCBI search (using organismic parameter taxid: rosids) suggests that Sequence 2 contains plant mitochondrial (MT) DNA (E-value$\approx$ 0) based on an alignment with *Cannabis sativa*. However, although identity is high (91%), it seems that sequence coverage (34%) is fairly low given that the two sequences are both contiguous genomic sequences. This could be due to the rapid divergence of intergenic regions in mitochondria [19], whilst the conserved regions are protein coding or regulatory.

To test this hypothesis we applied individual BLASTX searches (taxid: rosids) to the three aligned regions. This approach was adopted as an alternative to gene finding techniques used in Sec. 1 due to the difference in gene structure between MT and nuclear DNA, specifically the lack of exons. In order to apply such an HMM gene finding technique, new models would need to be trained using fully annotated (with gene boundaries) MT DNA.

2.1 Subsequence 1

The top BLASTX hit foor this sequence is from *M. truncatula* which encodes a MT cytochrome C protein (84% ID, 36% Coverage). Cytochrome C proteins are large transmembrane proteins found in MT and bacteria which are involved in electron transfer chain processes needed for ATP production [23].

2.2 Subsequence 2

The BLASTX search on this subsequence returned another MT protein belonging to *Malus domestica*, the apple tree (95% ID). The protein in question is a Sec-independent protein translocase (TATC) protein. These proteins are involved in protein transport across the cell or membrane and are found in MTs as well as most prokaryotes [2].

2.3 Subsequence 3

The hit returned shows that this subsequence may code (Evalue$\approx$ 0) for the S3 subunit of the ribosome in the MT with top hits from the *Hevea brasiliensis* MT S3 gene.

2.4 Consensus

From these three sources of evidence, it seems clear that Sequence 2 is MT in nature. Although some of the alignments have relatively low coverage for potential protein coding regions (roughly 20-40%) this could be explained by the rapid divergence of all non-functional protein subcomponents that occurs in MT as well as other prokaryotes [11].

3 Discussion

It is interesting to note both the present of reverse transcriptase and chloroplastic protein domains in Sequence 1. This perhaps suggests that retrotransposon or viral activity lead to the inclusion of chloroplastic gene fragments into this nuclear DNA region.

One potential caveat with the search and refine methodology we have employed is that pre-mature refinement can lead to exclusion of viable candidate protein and sequence alternatives that could lead to a lack of viable results. In order to alleviate the problem, previous gene models could be returned to when the newest model fails to yield results. Another problem with the refinement approach is that it could lead to bias towards a subset of species and sequences that happen to have achieved good initial alignment scores. One other problem is that by using *J. regia* as our final Gene Model we may actually have obtained worse results, since it is not as well understood as the more studied *A. thaliana*.

Overall it seems clear that the genomic regions explored above are dense with

possible protein coding and regulatory techniques and isolating the signal from a noisy evolutionary history can be extremely difficult. It is therefore important to conclude that our results may not turn out to be accurate as better, more subtle tools become available.

References

[1] Nucleotide blast: Search nucleotide databases using a nucleotide query. `https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=megaBlast&BLAST_SPEC=Plants_MV`. (Accessed on 03/27/2017).

[2] Tatc - sec-independent protein translocase protein tatc, chloroplastic precursor - arabidopsis thaliana (mouse-ear cress) - tatc gene & protein. `http://www.uniprot.org/uniprot/Q9SJV5`. (Accessed on 03/28/2017).

[3] Ycf2 (chloroplast) [arabidopsis thaliana] - protein - ncbi. `https://www.ncbi.nlm.nih.gov/protein/NP_051101.1`. (Accessed on 03/27/2017).

[4] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[5] Chris Burge and Samuel Karlin. Prediction of complete gene structures in human genomic dna. *Journal of molecular biology*, 268(1):78–94, 1997.

[6] Xinbin Dai and Patrick Xuechun Zhao. psrnatarget: a plant small rna target analysis server. *Nucleic acids research*, 39(suppl 2):W155–W159, 2011.

[7] Thomas Eulgem, Paul J Rushton, Silke Robatzek, and Imre E Somssich. The wrky superfamily of plant transcription factors. *Trends in plant science*, 5(5):199–206, 2000.

[8] Margaret S Ho, Pei-I Tsai, and Cheng-Ting Chien. F-box proteins: the key to protein degradation. *Journal of biomedical science*, 13(2):181–191, 2006.

[9] Marc Jakoby, Bernd Weisshaar, Wolfgang Dröge-Laser, Jesus Vicente-Carbajosa, Jens Tiedemann, Thomas Kroj, and François Parcy. bzip

transcription factors in arabidopsis. *Trends in plant science*, 7(3):106–111, 2002.

[10] Matthew W Jones-Rhoades, David P Bartel, and Bonnie Bartel. Micrornas and their regulatory roles in plants. *Annu. Rev. Plant Biol.*, 57:19–53, 2006.

[11] B Franz Lang, Michael W Gray, and Gertraud Burger. Mitochondrial genome evolution and the origin of eukaryotes. *Annual review of genetics*, 33(1):351–397, 1999.

[12] Nathalie Pavy, Stephane Rombauts, Patrice Dehais, Catherine Mathe, Davuluri VV Ramana, Philippe Leroy, and Pierre Rouze. Evaluation of gene prediction software using a genomic data set: application to arabidopsis thalianasequences. *Bioinformatics*, 15(11):887–899, 1999.

[13] Asaf A Salamov and Victor V Solovyev. Ab initio gene finding in drosophila genomic dna. *Genome research*, 10(4):516–522, 2000.

[14] Shusei Sato, Yasukazu Nakamura, Takakazu Kaneko, Erika Asamizu, and Satoshi Tabata. Complete structure of the chloroplast genome of arabidopsis thaliana. *DNA Research*, 6(5):283–290, 1999.

[15] Brenda A Schulman, Andrea C Carrano, Philip D Jeffrey, Zachary Bowen, Elspeth RE Kinnucan, Michael S Finnin, Stephen J Elledge, J Wade Harper, Michele Pagano, and Nikola P Pavletich. Insights into scf ubiquitin ligases from the structure of the skp1–skp2 complex. *Nature*, 408(6810):381–386, 2000.

[16] Robert W Scotland and Alexandra H Wortley. How many species of seed plants are there? *Taxon*, 52(1):101–104, 2003.

[17] Ilham A Shahmuradov and Victor V Solovyev. Nsite, nsiteh and nsitem computer tools for studying transcription regulatory elements. *Bioinformatics*, 31(21):3544–3545, 2015.

[18] Shin-Han Shiu, Ming-Che Shih, and Wen-Hsiung Li. Transcription factor families have much higher expansion rates in plants than in animals. *Plant physiology*, 139(1):18–26, 2005.

[19] Daniel B Sloan, Andrew J Alverson, John P Chuckalovcak, Martin Wu, David E McCauley, Jeffrey D Palmer, and Douglas R Taylor. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol*, 10(1):e1001241, 2012.

[20] Victor V Solovyev, Ilham A Shahmuradov, and Asaf A Salamov. Identification of promoter regions and regulatory sites. *Computational Biology of Transcription Factor Binding*, pages 57–83, 2010.

[21] Ralf Stracke, Martin Werber, and Bernd Weisshaar. The r2r3-myb gene family in arabidopsis thaliana. *Current opinion in plant biology*, 4(5):447–456, 2001.

[22] H Takatsuji. Zinc-finger transcription factors in plants. *Cellular and Molecular Life Sciences CMLS*, 54(6):582–596, 1998.

[23] David C Wharton and Alexander Tzagoloff. [45] cytochrome oxidase from beef heart mitochondria. *Methods in enzymology*, 10:245–250, 1967.

[24] Patricia J Wittkopp and Gizem Kalay. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1):59–69, 2012.