Universitas Esa Unggul

Smart, Creative and Entrepreneurial

IBT 432 Aplikasi Bioinformatika

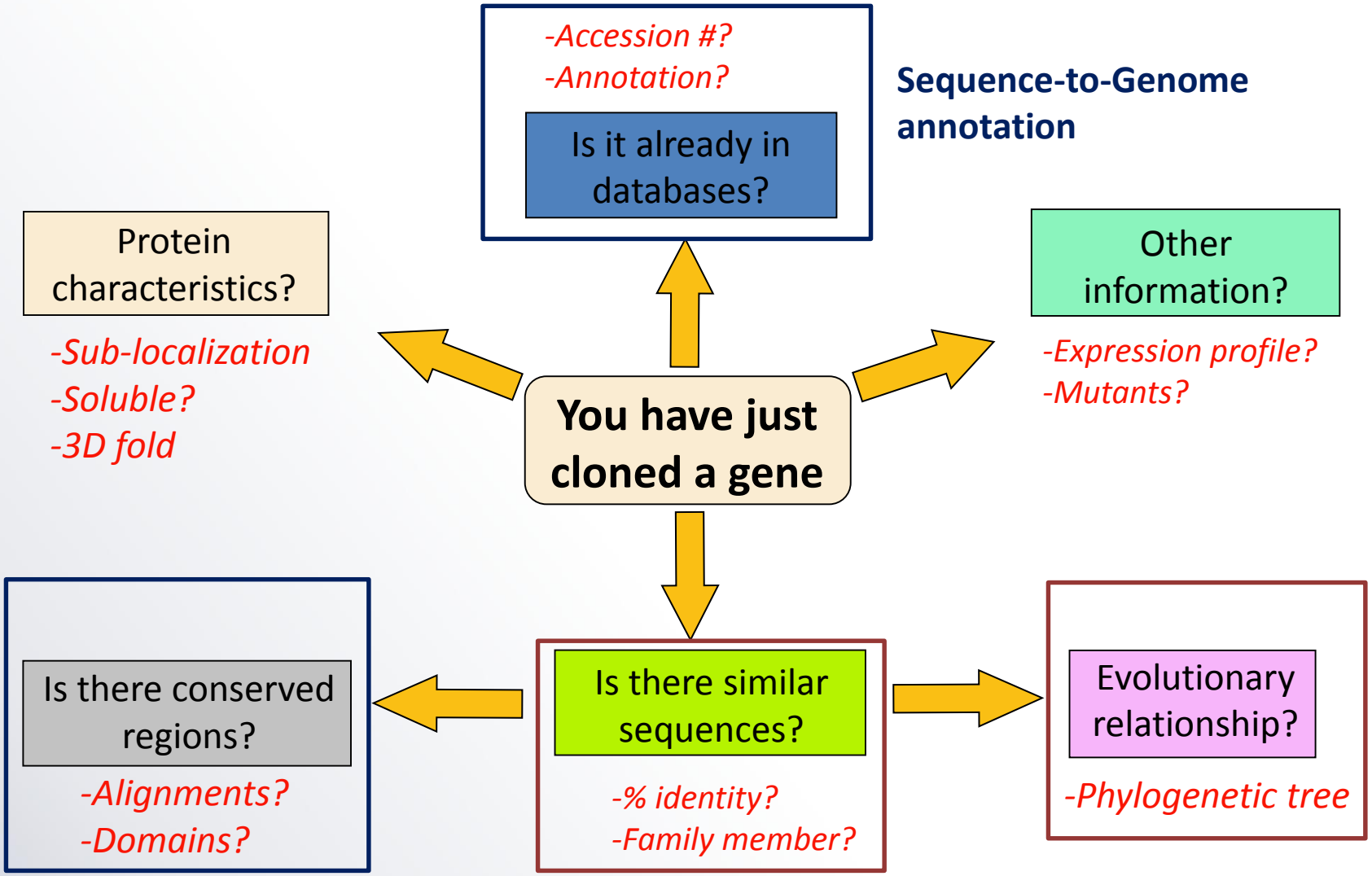
Analisis komparasi genomika III: Metodologi genomika komparatif dan contoh kasus

Riza Arief Putranto

Rencana Perkuliahan

1. ~~Kontrak belajar dan pengenalan bioinformatika aplikatif~~
2. ~~Database sekuen dan analisis genomika~~
3. ~~Anotasi sekuen ke genom - Praktik~~
4. ~~Analisis komparasi genomika I~~
5. ~~Analisis komparasi genomika II~~
6. Analisis komparasi genomika III
7. Analisis komparasi genomika – Praktik
8. Protein modelling I
9. Protein modelling II
10. Protein modelling III
11. Protein modelling - Praktik
12. Visualisasi protein modelling
13. Visualisasi protein modelling - Praktik
14. Presentasi mahasiswa

Remember this: analysis of genomics data



Homology searches

- Search **databases** of **DNA sequences**
- Use computer algorithms to **align sequences**
  - **Don't require perfect matches** between sequences
    - Allow for insertions, deletions, and base changes
- Most commonly used algorithms:
  - **BLAST**
  - FASTA

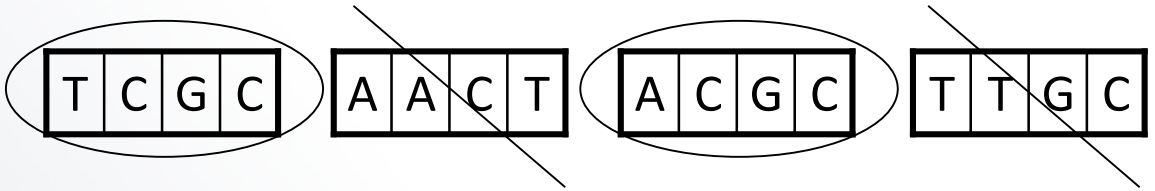
Methods – BLAST algorithm

- **Basic Local Alignment Search Tool**
- For comparing biological sequences (to find Homology – Paralogy or Orthology)

Example: Proteins, DNA sequences

Query

Library of sequences



(In the library – sequences of different lengths)

BLAST – Step 1

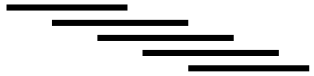
- Separate query to k-letter words

Example:

Proteins – Letters are Amino acids (L=Leucine)

Query sequence: RPPQGLF

(k=3)



3-letter words: RPP PPQ PQG QGL GLF

BLAST – Step 2

- Take one k-letter word – PQG
- **Search library** for similar words – LGMCPQA, DPPEGVV
- Define similarity: Use **scoring matrix** for two k-letter words

  High score for 2 words ⟺ Have common ancestor

  PQG – PQA : 12 PQG – PEG : 15
- Save similar words above a threshold T (save positions)
- Repeat for all k-letter words in query

BLAST – Step 3

- Align at saved positions:

- - - R P P Q G L F - - -

- - - D P P E G V V - - -

Scores: -2 7 7 2 6 1 -1

- Extend match right and left for positive score
- New pairs are called **High-scoring Segment Pairs (HSP)**
- Save significant HSPs (above a threshold S)

BLAST – Step 4

- Align saved HSPs (with gaps)

Example: 2 Sequences with 2 HSPs

.	.	.	R	P	P	Q	G	L	F	T	S	A	G	M	K	K	H	F	Y	Y	.	.	.	.
.	.	.	D	P	P	E	G	V	V	-	-	-	G	M	K	K	S	F	Y	D	N	C	D	.
.	.	.	D	P	P	E	G	V	V	G	M	K	K	S	F	Y	D	N	C	D	.	.	.	.

Insert gap ↗

- **Compute total score** (involves gap penalties)
- Report all matches above a threshold E

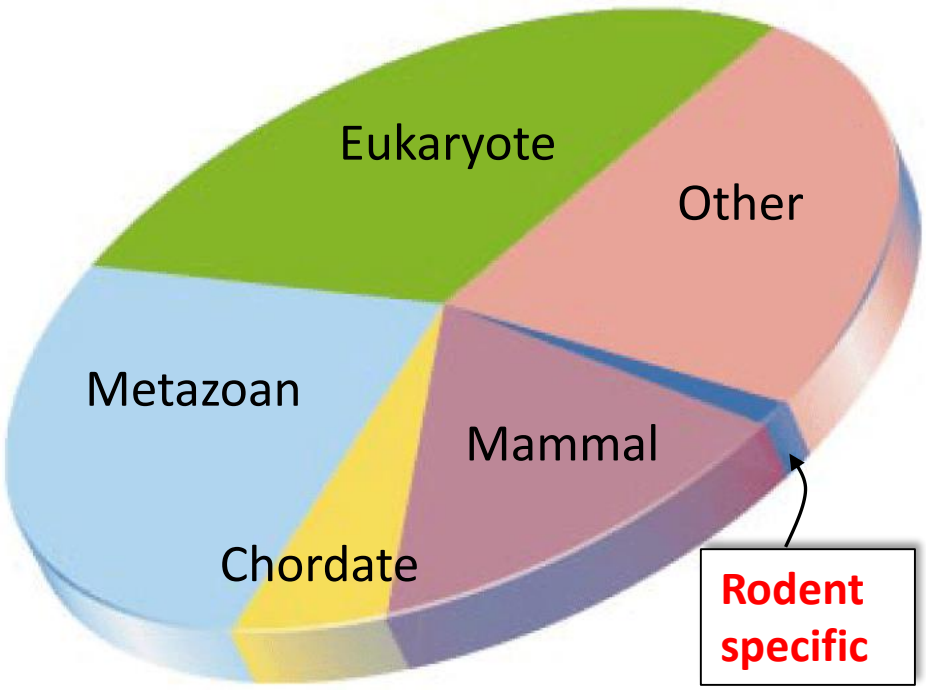
BLAST – Whole process

- Separate query to **k-letter words**
- **Search library** for similar k-letter words and save
- **Extend to HSPs** and save
- Align whole sequences and **compute total score**
- Return sequences with **score above E**

These are homologous to query

Homology search for the mouse genome

- Homology search of **all genes in the mouse genome**
- 27% in other metazoans
- 29% in other eukaryotes
- 6% in other chordates
- 14% in other mammals
- Less than 1% rodent specific

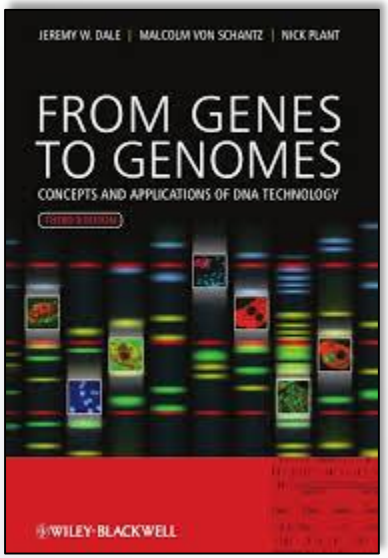


Comparative genomics on a higher scale – genome to genome

But on a smaller scale…

Genes-to-Genome

Genes-to-Genes



Case study of *Hevea* COBRA gene family

PROC. INTERNAT. CONF. SCI. ENGIN.
Volume 1, October 2017 | Pages: 39-47

ISSN 2597-5250 | EISSN 2598-232X

In silico Identification and Comparative Analysis of *Hevea brasiliensis* COBRA Gene Family

Riza Arief Putranto[1*], Irfan Martiansyah[1], Rizka Tamania Saptari[1,2]
[1]Indonesian Research Institute for Biotechnology and Bioindustry,
Jl. Taman Kencana No.1 Bogor 16128, Indonesia. Tel. +62-251- 8324048, Fax. +62-251- 8328516.
[2]School of Life Sciences and Technology – ITB,
Jl. Ganesa 10, Lebak Siliwangi, Coblong, Bandung, Indonesia 40132. Tel. +62-22- 2511575.
*Email: rizaputranto@iribb.org

Abstract. *Putranto R A, Martiansyah I, Saptari R T. 2017. In silico Identification and Comparative Analysis of Hevea brasiliensis COBRA Gene Family. Proc Internat Conf Sci Engin 1: 39-47.* In this paper, the *H. brasiliensis* COBRA gene family, alleged to be involved in laticifer differentiation, was identified from the public rubber tree genome of Reyan 7-33-97 clone. A comparative analysis was carried out against *A. thaliana* genomic database. This analysis has resulted to the in silico validation of thirteen putative genes encoding glycophosphatidylinositol anchors (GPI) proteins harbored by nine *Hevea* genomic scaffolds. The sequence's similarity of *HbCOBL* against *AtCOBL* genes were ranged from the threshold 50 to 81.58% covering 151 to 458 amino acid residues, respectively. Three partial and ten full-length protein sequences of *HbCOBL* genes were annotated. The partial protein sequences ranged from 89 to 184 amino acid residues as opposed to the full-length proteins ranging from 160 to 471 amino acid residues. Two types of COBRA domains (pfam04833 and cl04787) were found among *HbCOBL* genes. Phylogenetic analysis has clustered two subfamilies. Nine *HbCOBL* genes *(HbCOBL-B, HbCOBL-J, HbCOBL-C, HbCOBL-H, HbCOBL-F, HbCOBL-I, HbCOBL-M, HBCOBL-A,* and *HbCOBL-N*) were clustered as COBRA gene subfamily-I. By contrast, four genes (*HbCOBL-O, HbCOBL-P, HbCOBL-E,* and *HbCOBL-L*) were clustered as COBRA gene subfamily-II. The *HbCOB* subfamily-II was marked by the addition of 203 residues in C-terminal which is different with *Arabidopsis*. The gene *HbCOBL-C* was the putative ortholog to *AtCOB* carrying the unique COBRA domain cl04787 with 74 amino acid residues. Taken together, these results showed that *Hevea* and *Arabidopsis COBRA* genes might share similar functions while differ in gene structure.

Keywords: bioinformatics, COBRA, differentiation, laticifer, rubber tree

Abbreviations: glycophosphatidylinositol anchors (GPI); APETALA2/ETHYLENE RESPONSE FACTORS (AP2/ERF); Calcium-Dependent Protein Kinase (CDPK); Abscisic acid (ABA); Glucose-6-Phospate Dehydrogenase (G6PDH); Small Rubber Particle (SRPP); COBRA-like (COBL); European Nucleotide Archive (ENA); coding sequence (CDS); Conserve Domain Database Search (CDD); John-Taylor-Thornton (JTT); reads per kilobase per million mapped reads (RPKM)

Background

- Para rubber tree (*Hevea brasiliensis* Müll.Arg.) as the sole commercial source of natural rubber
- High NR production - related to **the storage of latex in the laticifer** (specialized tissue)
- **Secondary laticifer differentiation**, related to cell expansion, is an important limiting factor
- COBRA family has been **intensively studied in Arabidopsis** – role in the cell expansion/wood biosynthesis **but not in Hevea**
- **The transcript of putative COBRA was found in the laticifers** (Tang et al., 2016) – potential involvement in secondary laticifer differenciation

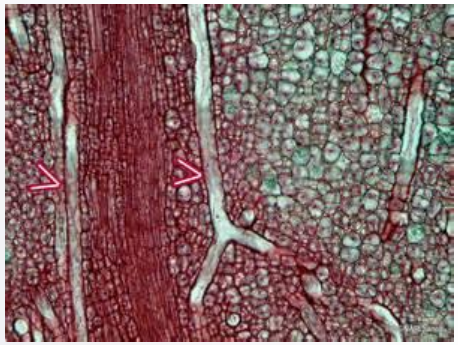
Red. Laticifer

Tan *et al.* 2017 *Scientific Reports* **7** 3126

Objective of the research

The objective of this work was to carry out *in silico* identification of *H. brasiliensis* COBRA gene family related to potential pivotal role in secondary laticifer differentiation

Comparative genome to genome analysis using Tang's rubber tree genome and annotated COBRA gene family in *Arabidopsis thaliana*



Workflow of the comparative analysis

QUERY

Rubber tree genome database
Clone Reyan-7-33-97
(Tang *et al* 2016)
https:// www.ebi.ac.uk/ena

REFERENCE

Arabidopsis genome database
TAIR
https://www.arabidopsis.org/browse/genefamily/COB.jsp

12 *AtCOBLs*

Putative CDS
9 Scaffolds

Sequence annotation using Geneious

tblastn of 9 scaffolds on *AtCOBLs* genes using Galaxy (www.galaxy.southgreen.fr/galaxy/)

Annotated putative CDS
9 Scaffolds

Conserved domain identification using NCBI CDD (www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)

17 amino acid residues for scaffolds encoding GPIs

13 annotated domains

Phylogenetic analysis of 13 amino acid residues *HbCOBLs* vs 11 *AtCOBLs* using MEGA 7.0

Phylogenetic tree of *HbCOBLs* vs *AtCOBLs* genes

Comparative analysis of Tang's genome vs TAIR database identifying *HbCOBLs*

- **tblastn in Galaxy**
- **NCBI Conserved Domain Database Search (CDD)**
- **Sequence annotation in Geneious**

Gene name	Tang's Scaffold		*Arabidopsis* ID	Galaxy tblastn*			Protein sequence		COBRA domain			Reference
	Scaffold ID	Length (Mb)		Similarity (%)	Length (bp)	E-value	Annotated	Length (aa)	CDD Search	Accession	Length (aa)	
HbCOBL-A	LVXX01000065	2,714	AT1G09790.1	53.89	167	7.00E-59	Partial	165	YES	pfam04833	151	(Lalanne *et al.* 2004)
HbCOBL-B			AT3G29810.1	61.84	228	2.00E-91	Full-length	164	YES	pfam04833	115	(Ben-Tov *et al.* 2015)
HbCOBL-C			AT5G60920.1	50.33	153	7.00E-36	Partial	89	YES	cl04787	74	(Schindelman *et al.* 2001)
HbCOBL-D	LVXX01000183	1,783	AT3G16860.1	53.07	179	3.00E-51	-	-	NO	-	-	-
HbCOBL-E			AT4G16120.1	70.93	454	0	Full-length	471	YES	pfam04833	180	(Borner *et al.* 2002)
HbCOBL-F	LVXX01000195	1,700	AT3G02210.1	81.58	152	5.00E-81	Full-length	160	YES	pfam04833	152	(Roudier *et al.* 2002)
HbCOBL-G			AT5G60920.1	66.51	209	5.00E-98	-	-	NO	-	-	-
HbCOBL-H			AT5G60920.1	77.63	152	1.00E-71	Full-length	159	YES	pfam04833	152	(Roudier *et al.* 2002)
HbCOBL-I	LVXX01000520	893	AT3G02210.1	81.58	152	3.00E-80	Full-length	172	YES	pfam04833	152	(Roudier *et al.* 2002)
HbCOBL-J			AT5G15630.1	76.43	157	2.00E-79	Full-length	171	YES	pfam04833	115	(Taylor-Teeples *et al.* 2015)
HbCOBL-K	LVXX01000625	733	AT3G16860.1	54.44	180	1.00E-52	-	-	NO	-	-	-
HbCOBL-L			AT4G16120.1	71.52	453	0	Full-length	471	YES	pfam04833	180	(Roudier *et al.* 2002)
HbCOBL-M	LVXX01000656	689	AT3G29810.1	50.00	152	1.00E-44	Full-length	181	YES	pfam04833	150	(Roudier *et al.* 2002)
HbCOBL-N	LVXX01001114	267	AT3G29810.1	52.32	151	2.00E-58	Partial	184	YES	pfam04833	148	(Roudier *et al.* 2002)
HbCOBL-O	LVXX01001412	126	AT4G16120.1	63.54	458	0	Full-length	451	YES	pfam04833	180	(Roudier *et al.* 2002)
HbCOBL-P	LVXX01003569	12	AT4G16120.1	63.54	458	0	Full-length	451	YES	pfam04833	180	(Roudier *et al.* 2002)
HbCOBL-Q			AT5G49270.1	54.37	160	1.00E-47	-	-	NO	-	-	-

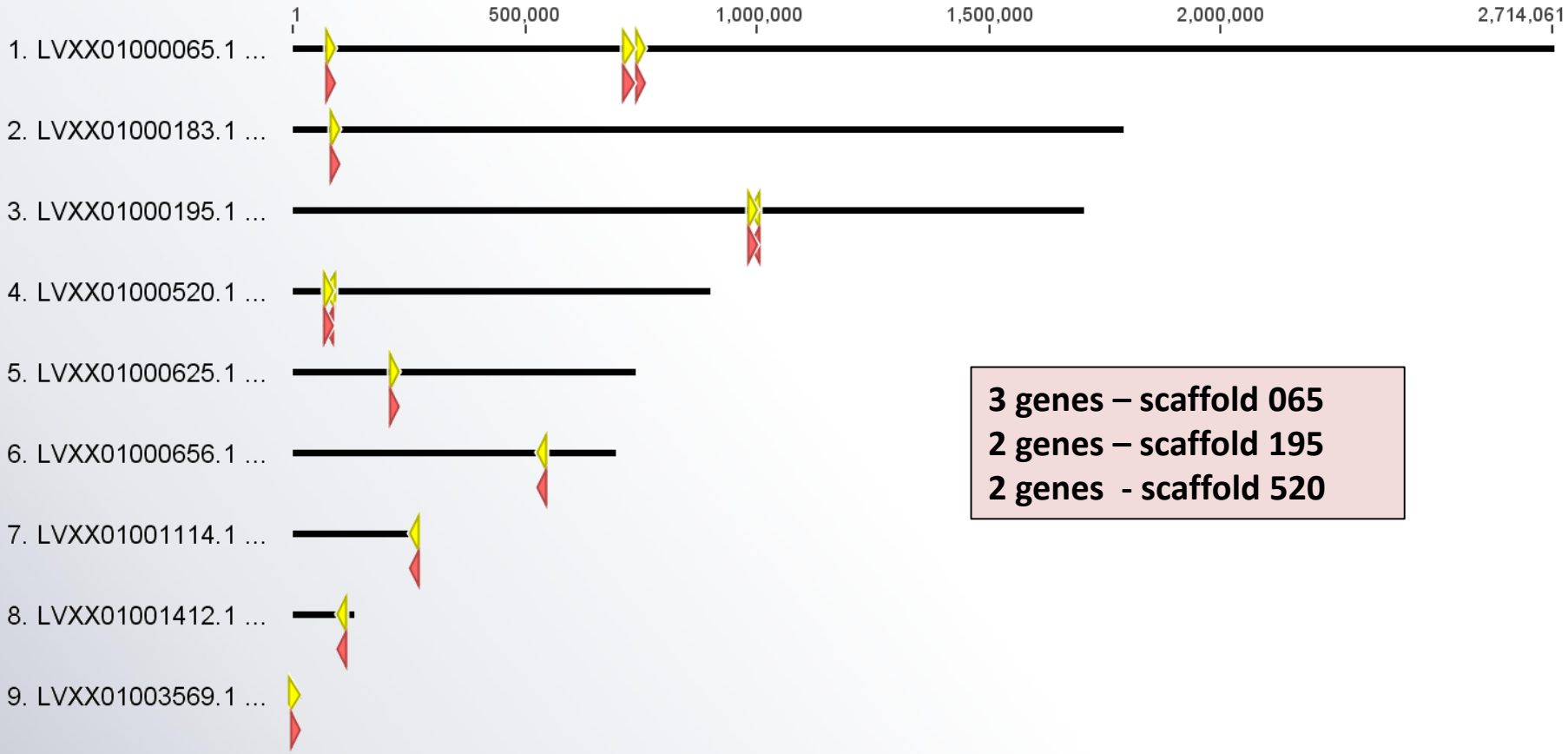
*NCBI BLAST+ tblastn: Search translated nucleotide database with protein query sequence(s) (Galaxy Version 0.1.04)

Identification of 17 sequences putative *HbCOBLs* (13 confirmed in 9 scaffolds)

Prediction of putative location HbCOBLs in the rubber tree genome

geneious

Putative location of each *HbCOBL* genes in *H. brasiliensis* genomic scaffolds.
Yellow triangle locates the CDS of the gene. Red triangle locates the COBRA domain of the gene.

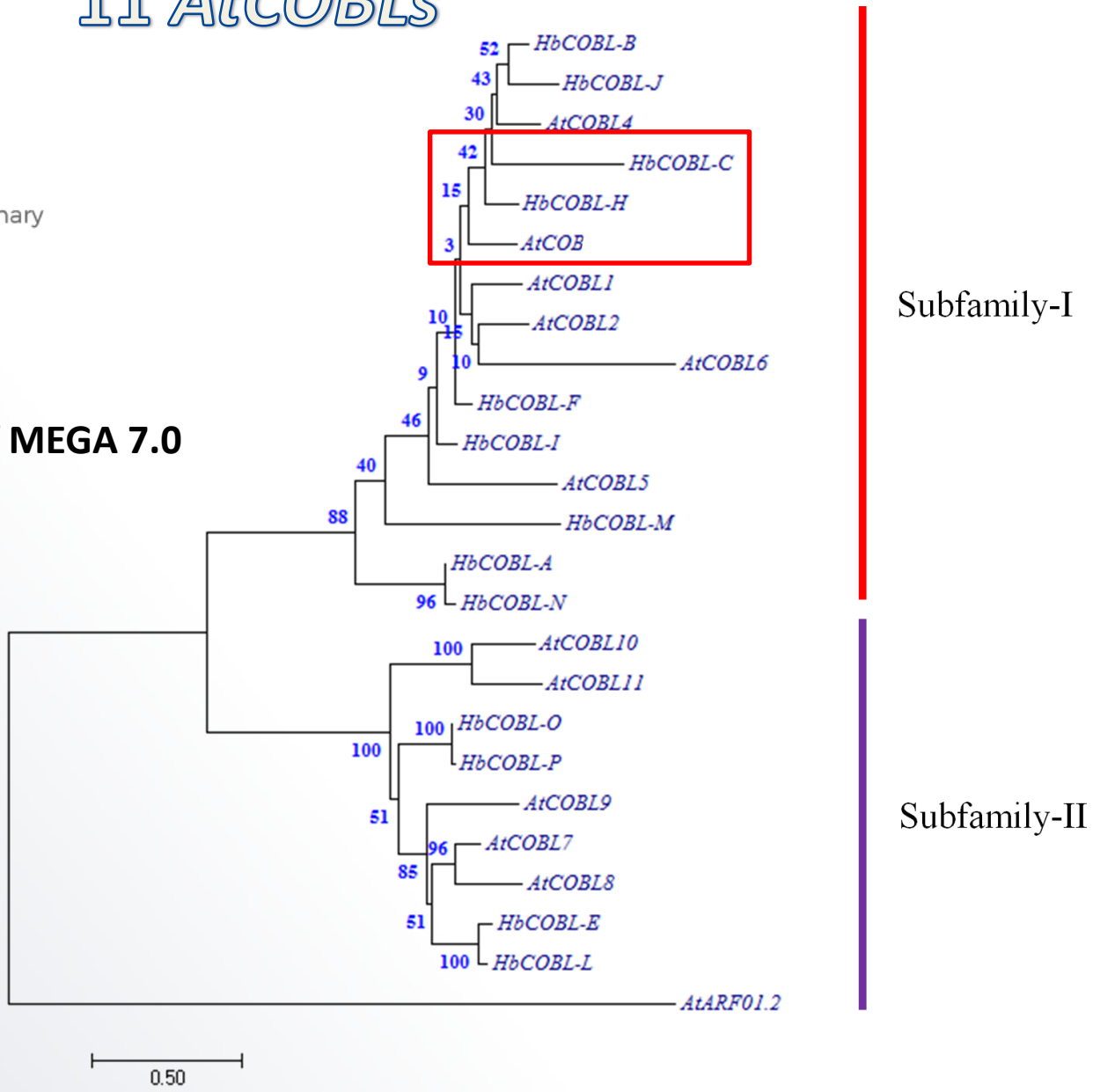


3 genes – scaffold 065
2 genes – scaffold 195
2 genes - scaffold 520

Phylogenetic analysis of 13 *HbCOLs* vs 11 *AtCOBLs*

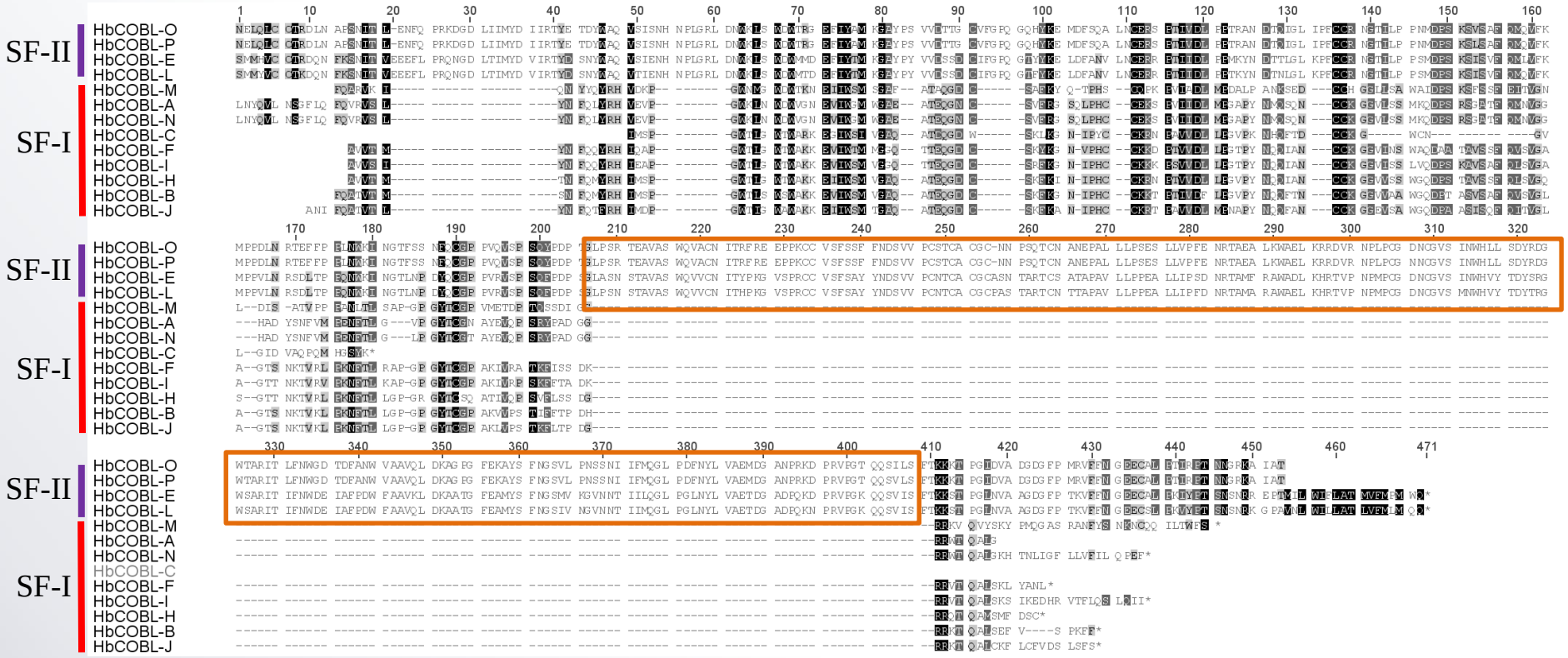
Molecular Evolutionary Genetics Analysis

- **Amino acid residues**
- **Neighbor-Joining method of MEGA 7.0**
- **1000 bootstrap replicates**



HbCOBL-C
HbCOBL-H | **AtCOB**

AA motif identification of 13 *HbCOBLs*



Two bars with red and blue colors refer to COBRA subfamily-I and –II.

- **Amino acid residues**
- **MUSCLE alignment**
- **Orange box: additional sequence on C-terminal**

Conclusion

- By implementing comparative analysis**, the putative *H. brasiliensis* COBRA gene family was identified**
- The gene family was clustered into **two subfamilies** with total **13 members**
- **Additional amino acid motifs** was found uniquely on subfamily-II

It was still the sixth course, don't get dizzy yet