Selection pressure and evolution rates of alpha and beta sequences; Beta is under negative selection and alpha positive selection

To determine the selection pressure exposed to each tubulin family proteins, I employed Hypothesis Testing and Selection with codeml. My goal was to determine the rates of evolution in the beta and alpha tubulin genes and to ascertain the underlying selection pressure exposed to these proteins. For this particular analysis since I was dealing with protein-coding DNA sequences, I used seqtype set to 1 and carried out ML analysis using codon substitution models (e.g., Goldman and Yang 1994). codeml is a part of the PAML package, which is a suite of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood (ML). For more information visit (http://envgen.nox.ac.uk/bioinformatics/docs/codeml.html)

Consistent with my hypothesis, the omega ($\omega = dN/dS$) between alpha and beta was different (Table 1). Omega (ω) ratio is a measure of natural selection acting on the protein and is very informative in understanding natural selection acting on genomes of species. Hypothesis testing using codeml revealed that beta tubulin genes are under a negative selection pressure ($\omega < 1$) whereas alpha tubulin genes are under positive selection ($\omega > 1$), Table 1. This was quite unexpected as the margin is very strong (X1000). This dramatic difference can somewhat be explained with the recent discovery of tubulin disorders (largely known as Tubulinopathies) where at least 60% of these disorders are associated with mutations in the beta tubulin genes (Markova et al, 2015). The fact that such mutations are deleterious, they are quickly removed from the gene pool and those that persist cause tubulin diseases. However, such cases are rare as carriers of these harmful mutations have fewer offspring each generation thus reducing the frequency of the mutation in the gene pool.

Furthermore, under Ohta's hypothesis of slightly deleterious mutations, purifying selection is more effective in large populations than in small populations, and so differences in population sizes along lineages provide another compatible hypothesis. If amino acid changes are slightly deleterious, we expect them to be removed from the population at a higher rate in a large population than in a small population. As a result, we expect to see a smaller dN/dS ratio in a large population than in a small one, even if there is no difference between the two lineages in selective pressure or gene function. In the context of population sizes, beta tubulin evolved with more protein coding genes (10) than alpha (09) which is consistent with a smaller dN/dS ratio observed. We thus conclude that beta tubulin is essential components of eukaryotic cytoskeleton function and accumulation of deleterious mutations increases a risk to tubulin diseases

Table 1. Omega values for alpha and beta tubulin sequences

Nucleotide sequence	dN	dS	w = dN/dS	Sites
Beta-genes	0.08056	37.02937	0.0022	208.3
Alpha-genes	0.5331	0.2293	2.3247	189.1

Gene family evolution of mammalian alpha and beta tubulin genes and the underlying selection pressure

I used Phylogenetic analysis to validate the nomenclature of both alpha and beta tubulin protein groups as represented in Tables 1 and 2 (https://www.genenames.org/. The basis of alpha tubulin classification has been reported in literature (Varsha K. Khodiyar et al, 2007) whereas the classification of beta tubulin hasn't been done yet.

Table 2: Classes of beta tubulin genes (https://www.genenames.org/)

Group (New Class)	Gene name (CDS) Approved symbol	Class (**Approved Name)**
I	TUBB	tubulin beta class I
V	TUBB1	tubulin beta 1 class VI
II	TUBB2A	tubulin beta 2A class IIa
II	TUBB2B	tubulin beta 2B class IIb
III	TUBB3	tubulin beta 3 class III
IV	TUBB4A	tubulin beta 4A class IVa
IV	TUBB4B	tubulin beta 4B class IVb
III	TUBB6	tubulin beta 6 class V
VI	TUBB8	tubulin beta 8 class VIII
Not classified	TUBB7P	tubulin beta 7 pseudogene

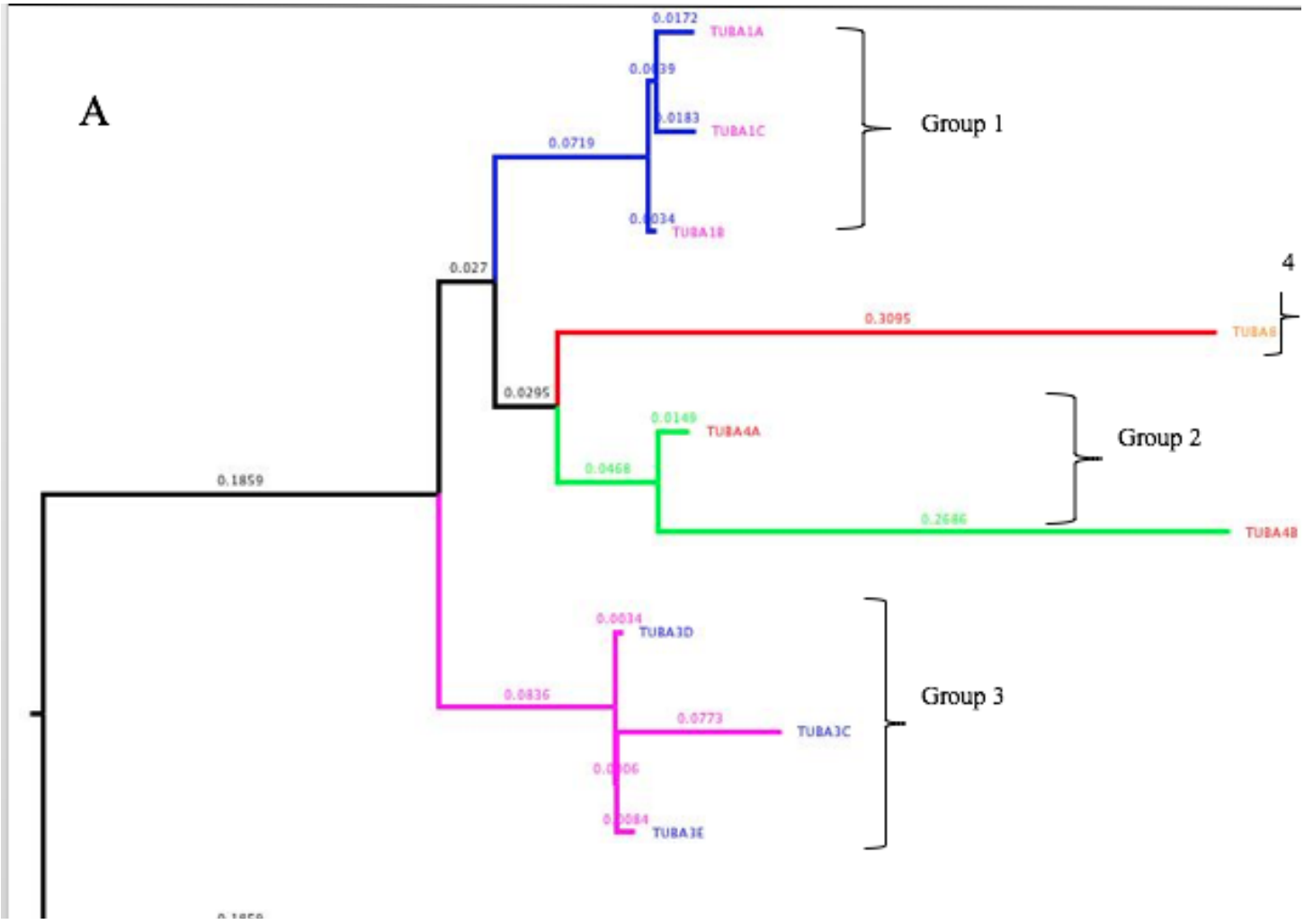
Table 3: Classes of alpha tubulin genes

Phylogenetic group	**Approved Symbol**	**Approved Name**
Group 1	TUBA1A	tubulin alpha 1a
	TUBA1B	tubulin alpha 1b
	TUBA1C	tubulin alpha 1c
Group 3	TUBA3C	tubulin alpha 3c
	TUBA3D	tubulin alpha 3d
	TUBA3E	tubulin alpha 3e
Group 2	TUBA4A	tubulin alpha 4a
	TUBA4B	tubulin alpha 4b
Group 4	TUBA8	tubulin alpha 8

From the phylogenetic analysis of protein coding sequences, of alpha tubulin genes, it can be observed that there are four α-tubulin subgroups (Fig.1A, and 2) which is consistent with the classification by HUGO Gene Nomenclature Committee, (Table 3 and Varsha K. Khodiyar et al, 2007).

In contrast, phylogenetic analysis of b-tubulin protein coding sequences, classified these genes into 6 subgroups (Fig 1B and 2) which was one less compared to the HUGO gene nomenclature classification. TUBB6 and TUBB3 form a monophyletic clade and thus belong to the same beta protein class (III) than belonging to classes V and III respectively (Table2) as previously classified.

When the two protein groups were combined together and analyzed by neighbor joining using amino acid sequences, the resulting tree shows us that the 2 protein groups are paralogs (Fig 2), implying that they arose as a result of gene duplication. Within each protein group, the proteins are orthologous to each other which suggests that these rose through gene speciation that gave rise to these proteins. I found out that the phylogenies constructed from the protein coding nucleotide sequences (Fig 1) and amino acid sequences (Fig 2) are conflicting. This could be due to the different methods which I used that is Maximum likelihood for protein coding nucleotide sequences and distance analysis (neighbor joining) for the amino acid sequences.
In the nucleotide-based analysis human TUBA4B falls within group 2 (Fig 1A, S4), whereas the amino acid-based analysis positions TUBA4B outside the outgroup chosen for the phylogenetic analysis (see Fig 2). The same applies to TUBB4A of the beta tubulin (Fig 1B, S3) where it forms a distinct paraphyletic group after the nucleotide-based analysis and thus given class V whereas after the amino acid-based analysis if forms a monophyletic clade with TUBA4A (Fig 2, S1). The notion behind this trend is yet clear to me.



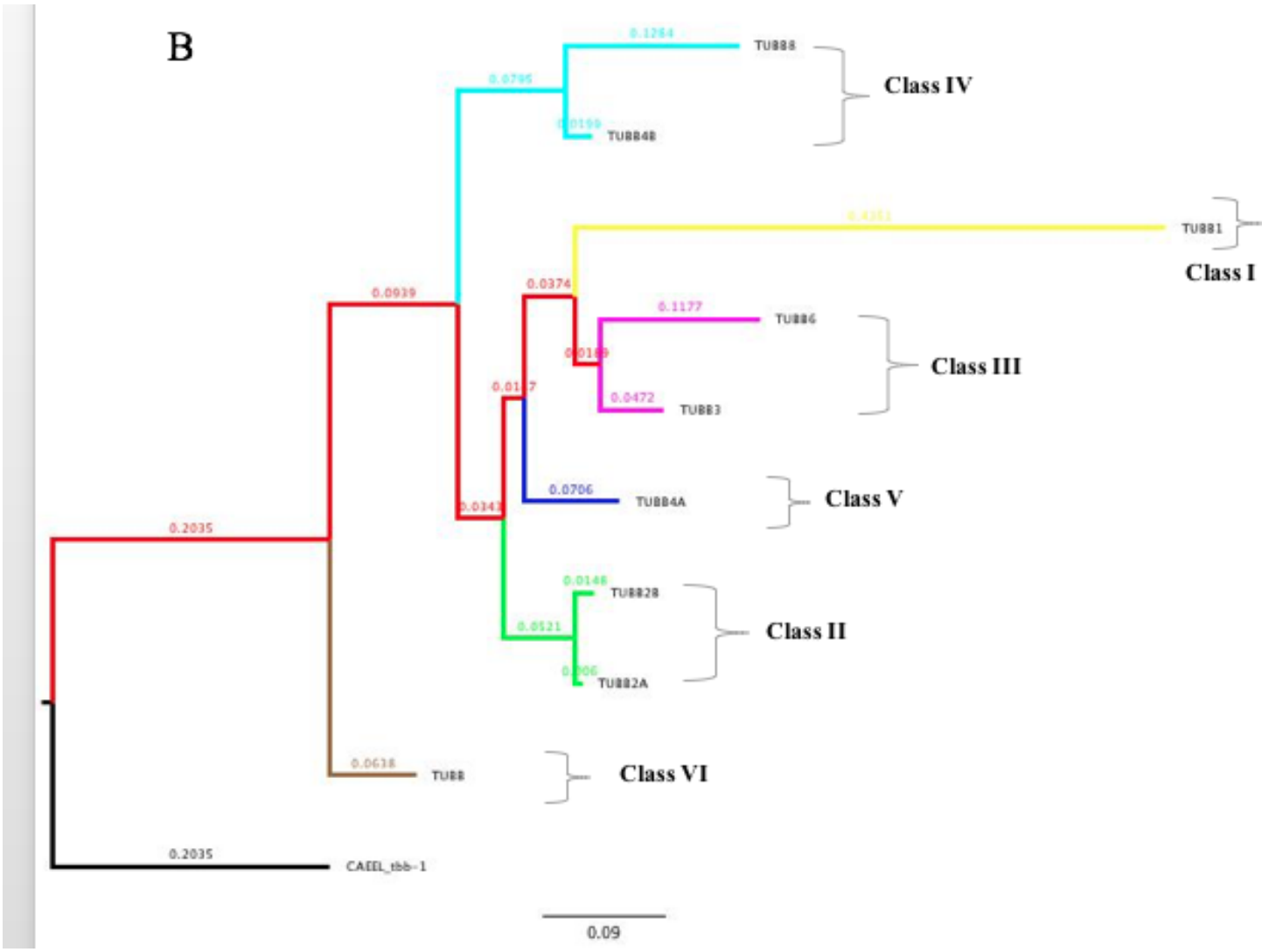


Figure 1. (A) Phylogenetic analysis of human α-tubulin genes by Maximum likelihood using the following settings. (1) Method; Maximum likelihood using GTRGAMMA model of nucleotide substitution. (b) substitutions to include: transitions + transversions; (c) rates among sites: uniform rates. (2) Include sites (a) gaps/missing data: complete deletion; (b) codon positions: 1st+2nd+3rd+noncoding. (b) Numbers on the branches refer to the bootstrap values after 1000 replicates.
Human sequences, used in this analysis are listed in Table 3 and rooted with *Caenorhabditis elegans* α-2 tubulin CAEEL.
Phylogenetic tree, visualized by fig tree represent groups that correspond to gene sequences clustered on the monophyletic clade and colored for clarity.

(B). Phylogenetic analysis of human b-tubulin genes by Maximum likelihood using similar settings as for a-tubulin. Monophyletic clades clustering beta-tubulin genes are represented as classes as it was found in the literature. Human sequences used in this analysis are listed in table 2. Numbers on the branches refer to the bootstrap values after 1000 replicates.

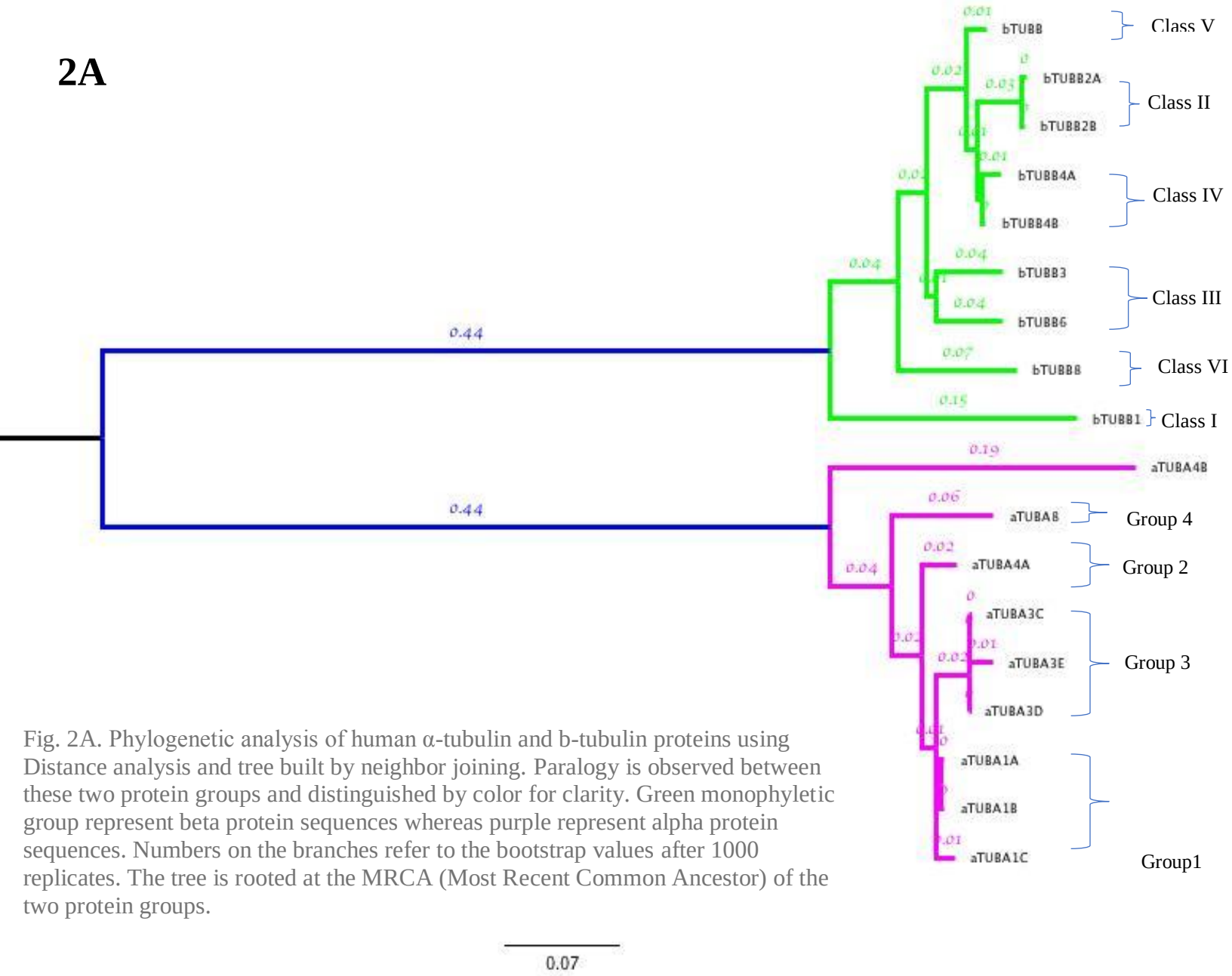


Fig. 2A. Phylogenetic analysis of human α-tubulin and b-tubulin proteins using Distance analysis and tree built by neighbor joining. Paralogy is observed between these two protein groups and distinguished by color for clarity. Green monophyletic group represent beta protein sequences whereas purple represent alpha protein sequences. Numbers on the branches refer to the bootstrap values after 1000 replicates. The tree is rooted at the MRCA (Most Recent Common Ancestor) of the two protein groups.

Look at the tissue specific localization and gene locations, it was quite interesting to find out that the subgroups are enriched is specific tissues and also located on different chromosomes (data not shown)

Conclusion

Hypothesis testing using codeml has revealed to us that alpha and beta tubulin sequences are exposed to different rates of selection alpha being under positive selection whereas beta negative selection. We have also validated the nomenclature of mammalian beta tubulin protein coding nucleotide sequences into 6 classes than it was earlier reported. We thus conclude that phylogenetic is a powerful tool to understanding evolutionary relationships among species and on the genome level

References

Adachi, J., and M. Hasegawa. 1996a. MOLPHY Version 2.3: Programs for molecular phylogenetics based oÂn maximum likelihood. Computer science monographs, 28:1-150. Institute of Statistical Mathematics, Tokyo.

Adachi, J., and M. Hasegawa. 1996b. Model of amino acid substitution in proteins encoded by mitochondrial DNA. Journal of Molecular Evolution 42:459-468.[**Entrez**]

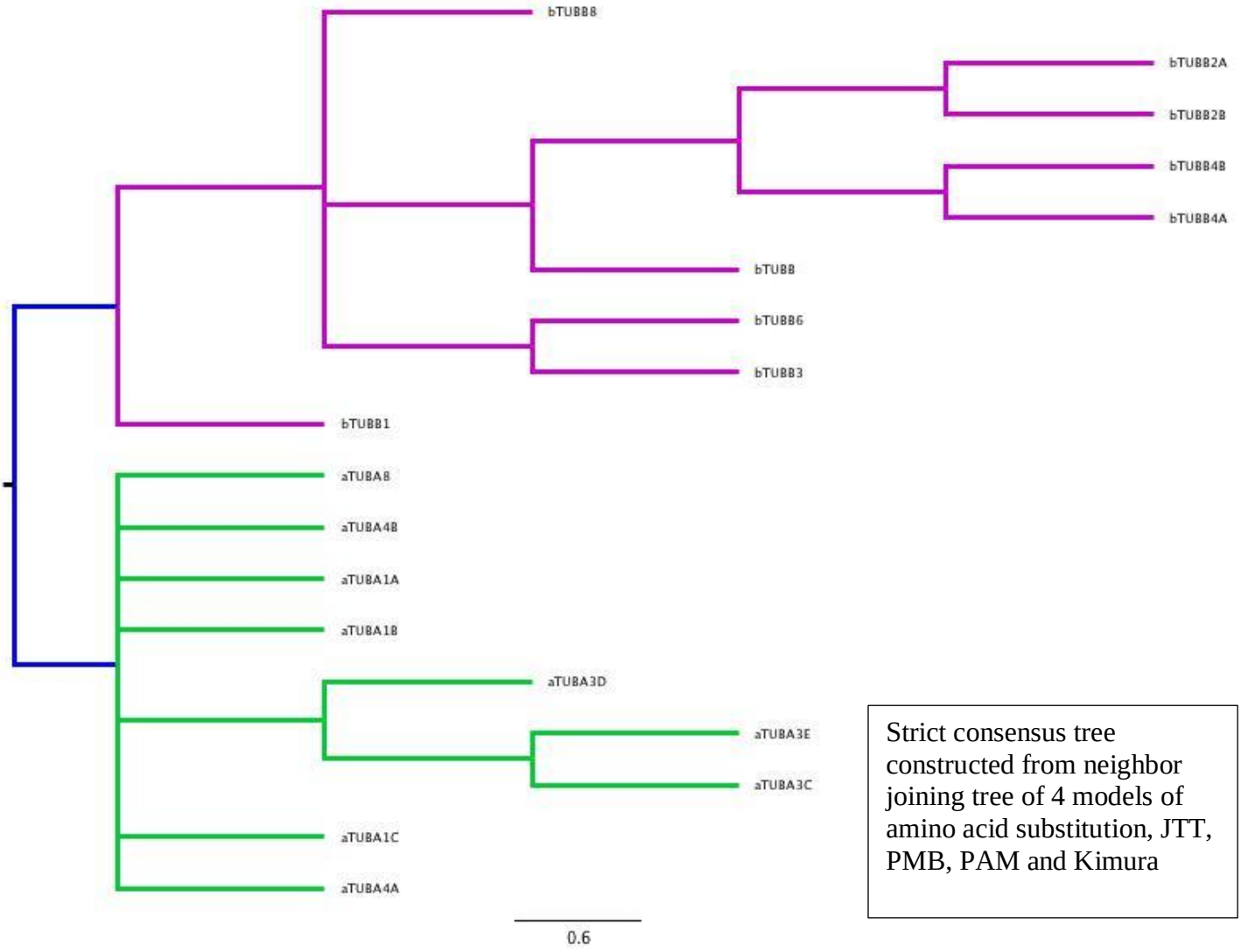
Markova et al, 2015; Genetic Disorders Affecting Tubulin Cytoskeleton. Journal of Biomedical and Clinical Research

http://envgen.nox.ac.uk/bioinformatics/docs/codeml.html

http://abacus.gene.ucl.ac.uk/software/pamlFAQs.pdf

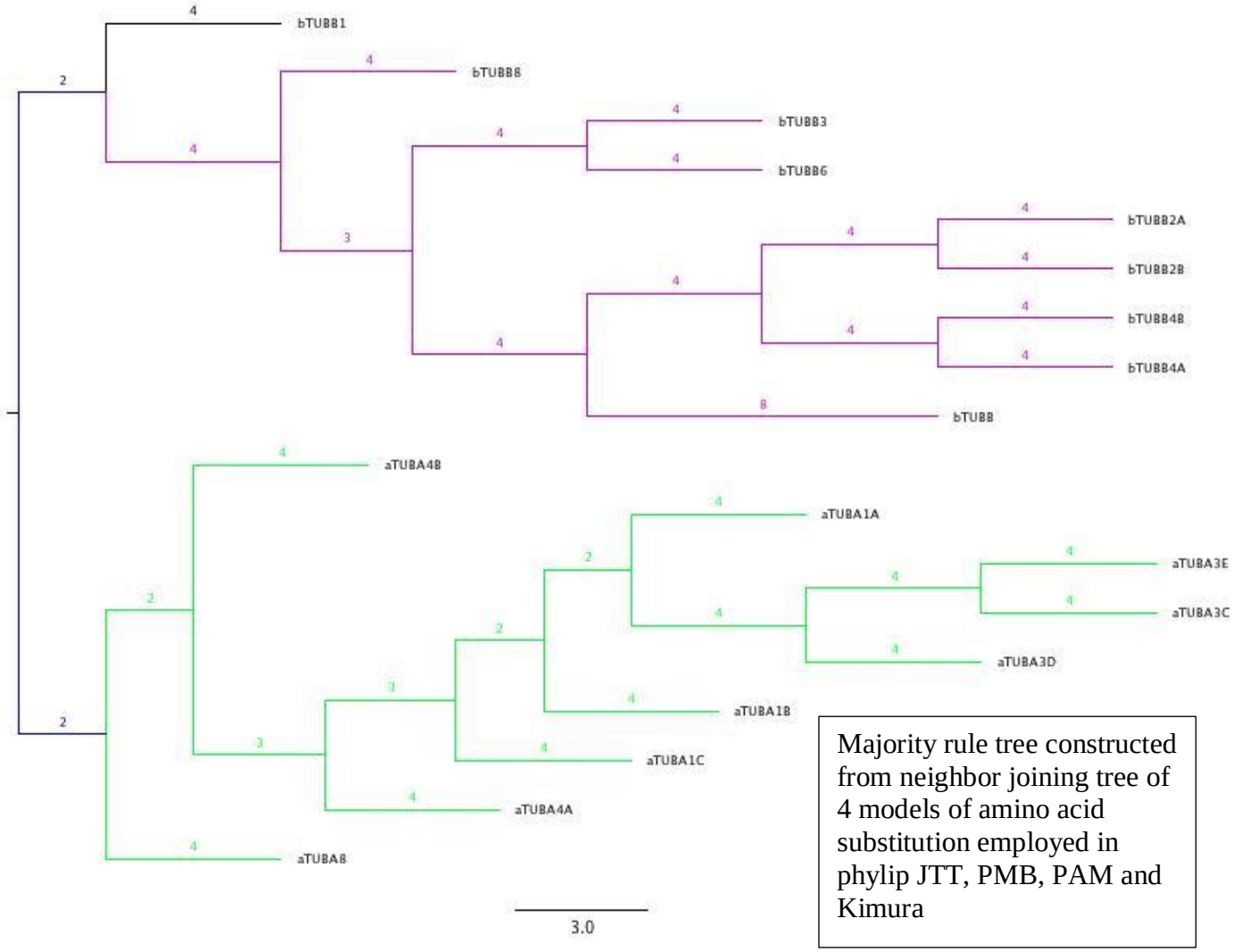
Supplemental data

S1



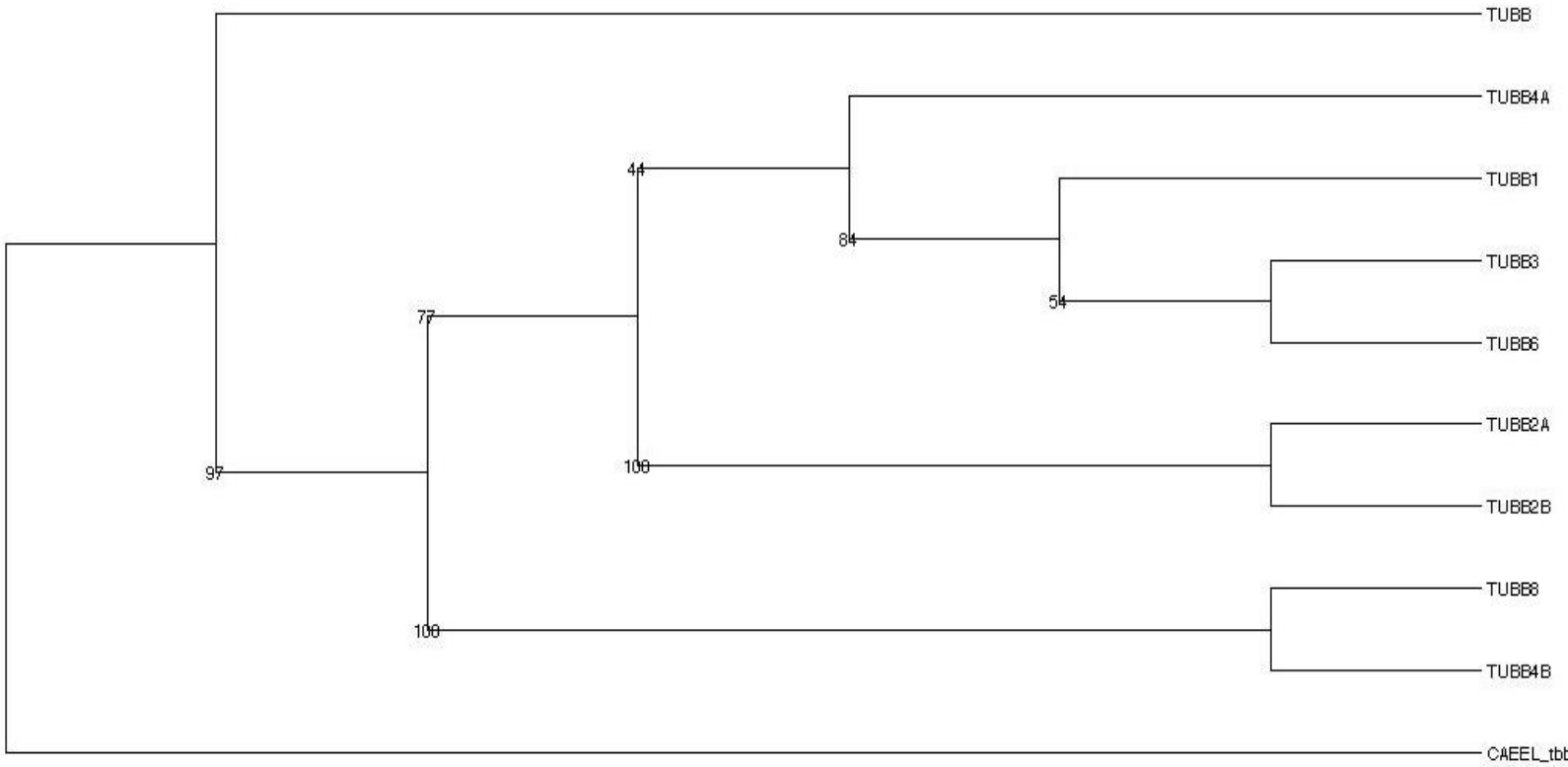
Strict consensus tree constructed from neighbor joining tree of 4 models of amino acid substitution, JTT, PMB, PAM and Kimura

S2



Majority rule tree constructed from neighbor joining tree of 4 models of amino acid substitution employed in phylip JTT, PMB, PAM and Kimura

S3 Majority rule extended tree of beta tubulin built by maximum likelihood



S4 Strict Consensus tree of alpha tubulin sequences built by maximum likelihood.
The clades are clearly distinguished and the numbers represent bootstrap support values after 1000 iterations.