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

Introduction

Microtubules (MTs) are filaments of the eukaryotic cytoskeleton assembled from the protein tubulin and highly involved in mitosis, motility and intracellular transport. MT are assembled from protein tubulin is a heterodimer of alpha- and beta-subunits and the sequence, and structure of tubulin subunits is highly conserved among species. Many eukaryotic organisms carry multiple genomic copies of functional α or β tubulin, commonly referred to as isoforms (or isotypes if they are confined to a single organism). In humans for example, 9 genes encode for the α-subunit whereas 10 genes encode the β-subunit which are assembled into functional microtubule polymers. Mutations in different mammalian tubulin proteins have been linked to a wide range of disorders many of which affect brain development (Mark I. Rees et al, 2014).

Beta tubulin is of critical function because most of the available data in the literature deals with this protein as a target for drug action and (Huzil et al, 2007). For instance, all drugs targeting microtubules for cancer chemotherapy are designed to bind to beta tubulin. The antitumor drugs stabilize microtubules and reduces their dynamicity, promoting mitotic arrest and eventually apoptosis. Mutations in specific beta-tubulin isotypes cause severe neuropathies that disrupt axonal transport and cancer (Huzil et al, 2007). Genetic variations affecting all beta-tubulin genes expressed at high levels in the brain (TUBB2B, TUBB3, TUBB, TUBB4A, and TUBB2A) have been linked with malformations of cortical development and disease phenotypes that arise from their disruption, include microcephaly, lissencephaly and polymicrogyria (Thomas D. Cushion, et al. 2014)

To date, mammalian beta tubulin has been grouped into 9 classes, class I, IIa, IIb, III, IVa, IVb, V, VI, VIII (HUGO Gene Nomenclature Commitee). However, searching through literature, no phylogenic study has been done/found categorizing these genes in their respective classes. Thus, the basis of beta tubulin classification is not yet clear. Despite lack of the consensus phylogeny, it has been shown that class I β-tubulin is the most commonly expressed isotype in humans and as such is also the most common isotype found in cancer cells. Alternatively, both classes II and III β-tubulin have been observed at increased levels in human tumors (Ferguson et al. 2005; Mozzetti et al. 2005). There is a degree of tissue specificity in the expression of some β- tubulins as described and some degree of gene redundancy where the loss of one gene can be compensated by over expression of another as has been shown in yeast (Nsamba et.al unpublished). Here I employ molecular phylogenetics to validate the nomenclature of mammalian tubulin and at the same time determine the selection pressure and molecular evolutionary relationship between the two major classes of mammalian tubulin.

Methods

Assembling the data

Databases; UniProtKB. and NCBI.

Protein sequences; The mammalian alpha tubulin protein sequences were similar to those used in V.K. Khodiyar et al. 2007, (A revised nomenclature for the human and rodent α-tubulin gene family) whereas the beta tubulin sequences were extracted from UniProtKB database (http://www.uniprot.org/). Nucleotide sequence; Both α or β-tubulin gene sequences were extracted from NCBI database using their corresponding gene names as published in literature. Only the protein-coding portion of each cDNA was used, to prevent differences in length of the UTRs biasing the alignments. These were extracted from (https://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi) a sub-project of NCBI for annotating coding regions. For phylogenic analysis, C. elegans α-2 tubulin was used as an outgroup.

Generating the Multiple Sequence Alignment

I employed a MAFFT module available on HPC class to align the sequences. Depending on the type of analysis, I converted the fasta file in different formats that suit a specific phylogenetic analysis for instance, phylip format for Maximum likelihood ad distance analysis.

Computing and Visualizing Phylogenetic Trees

Phylogenetic trees were generated for all data sets using the neighbor joining (NJ) and the maximum likelihood (ML) method. Maximum likelihood was inferred using RAxML77 v.8.0.22 with the GAMMA-LG model of evolution for protein sequenceand GTRGAMMA for nucleotide sequences. For Neighbor joining, pairwise distances among taxa was used as input for phylogenetic reconstruction estimated under five different models of amino acid substitutions available in protdist and 4 models of nucleotide substitutions available in dnadist. In all cases, a strict consensus and majority rue trees were built by neighbor joining. For each data set, bootstrapping with 1,000 replicates was performed.

Hypothesis Testing and Detecting Selection with codeml

To determine the rates of evolution in the beta tubulin genes and ascertain the underlying selection pressure, I used codeml, a PAML (Phylogenetic Analysis by Maximum Likelihood) package, by setting seqtype to 1 and carried out ML analysis using codon substitution models (e.g., Goldman and Yang 1994). Both alpha and beta tubulin gene sequences were analyzed separately and dN/dS ratios compared between the two.

Phylogenetic trees were visualized with FigTree, newick tree viewer and dendrocope for strict and majority rule consensus trees.

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 (ω □= 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

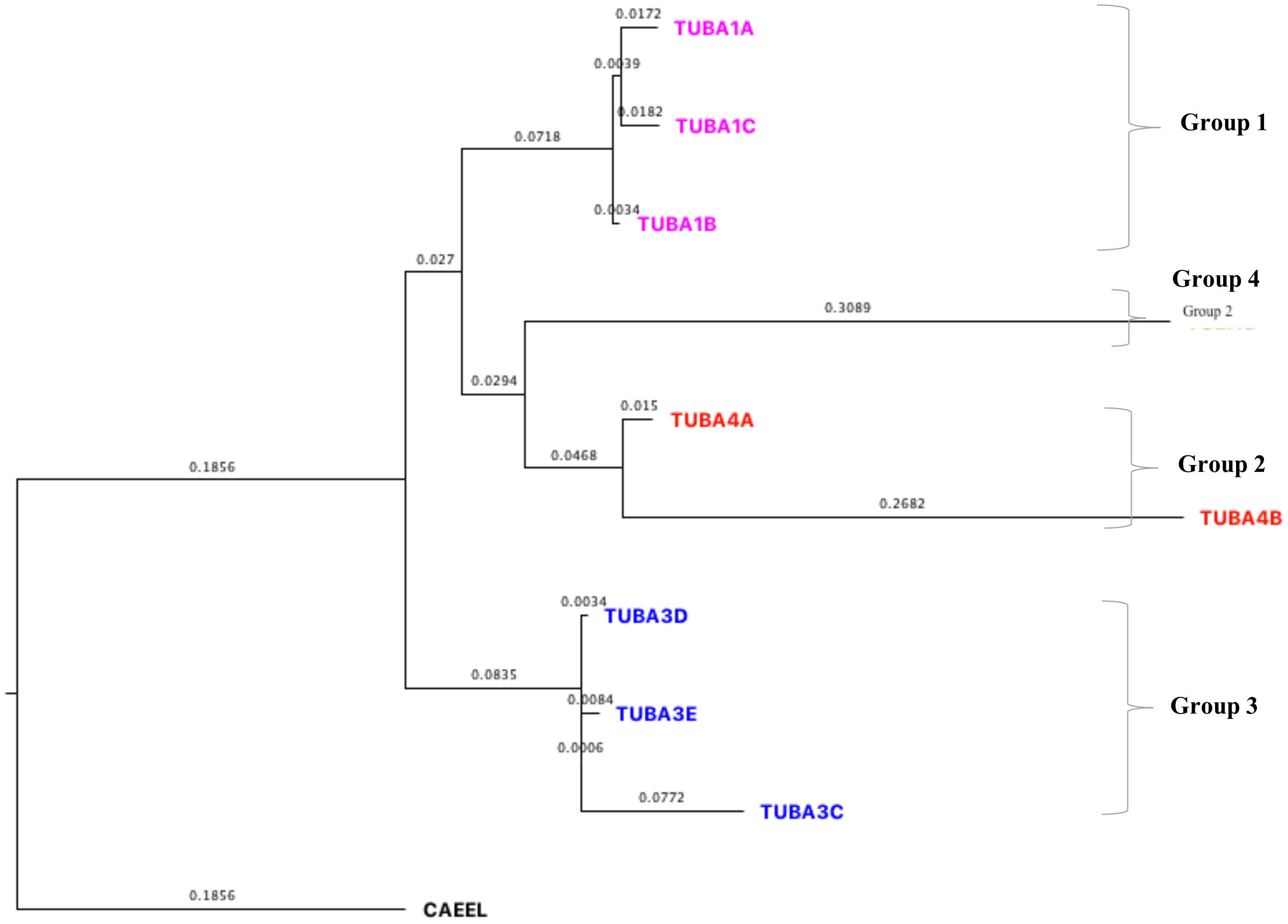
Table 3: Classification of alpha tubulin genes

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) and thus not belonging to group 2. This is because the human TUBA4A protein is 448 amino acids in length, whereas the human TUBA4B protein is only 241 amino acids. The alignment of the *TUBA4A* cDNA to the *TUBA4B* genomic sequence provides evidence that another tubulin-related exon containing multiple frameshifts and stop codons is present in this region as a pseudo gene and is not transcribed. 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).



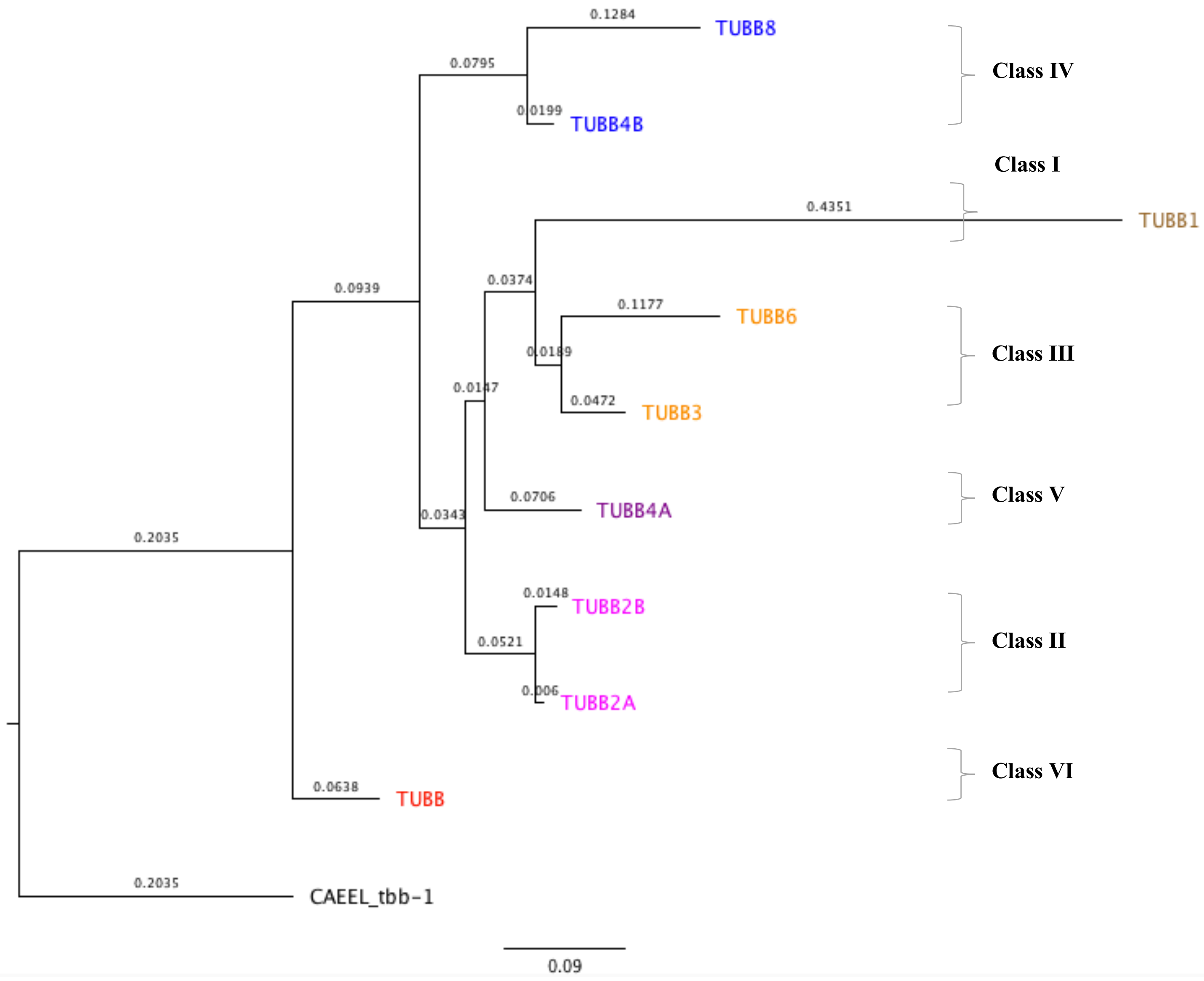


Figure 1. (A) Phylogenetic analysis of human α-tubulin genes by Maximum likelihood using the following settings. (1) Method; Maximum likelihood using GTRGAMMA model of nucleotide substitution. (b) substitutions to include: transitions + transversions; (c) rates among sites: uniform rates. (2) Include 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 protein sub-groups that correspond to branch specific gene duplication events gene sequences clustered 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 as a result of branch specific duplication events are represented as classes as it was found in the literature. Human beta tubulin protein names/classes used in this analysis are listed in table 2. Numbers on the branches refer to the bootstrap values after 1000 replicates.

2A

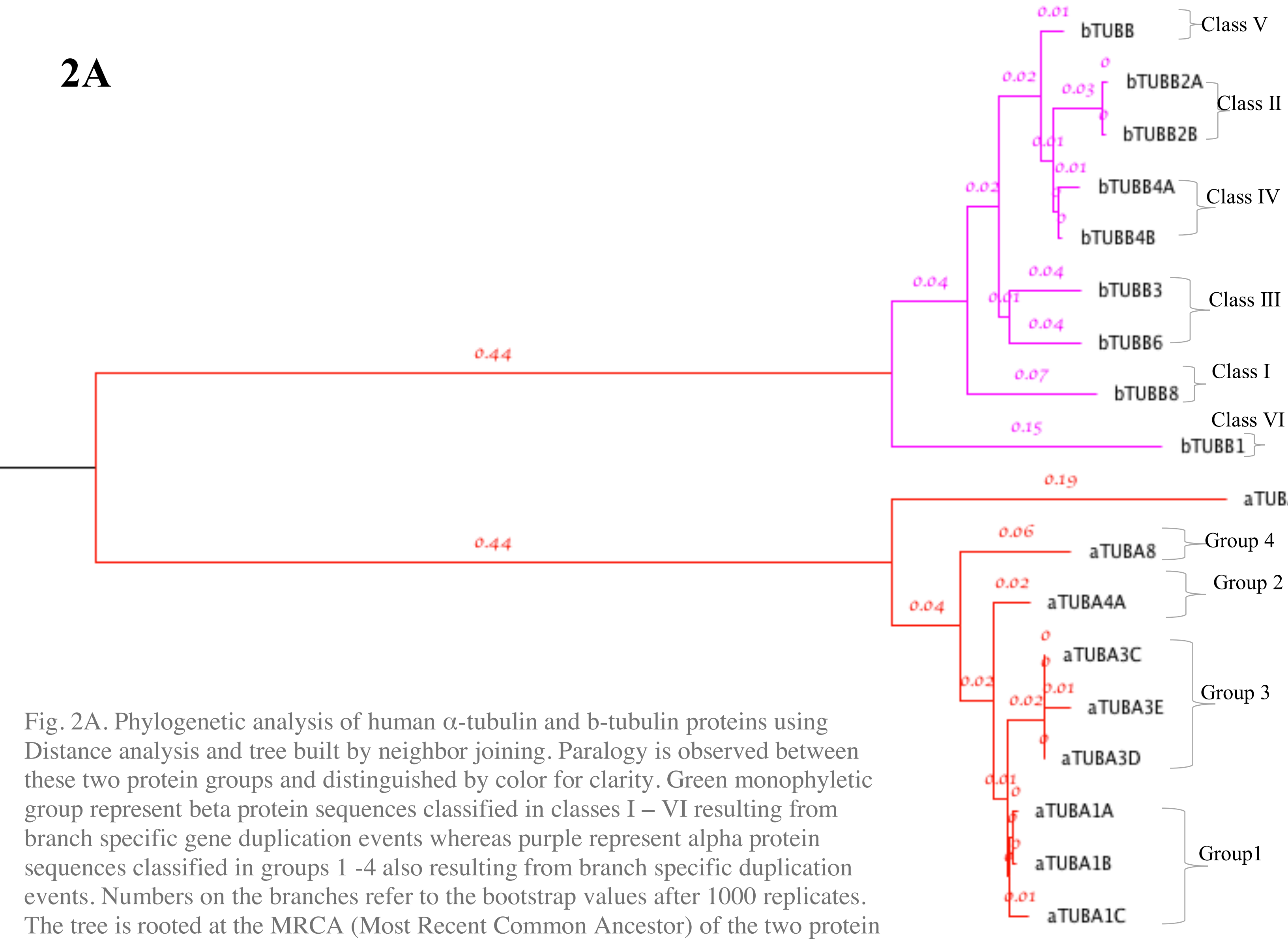


Fig. 2A. Phylogenetic analysis of human α-tubulin and b-tubulin proteins using Distance analysis and tree built by neighbor joining. Paralogy is observed between these two protein groups and distinguished by color for clarity. Green monophyletic group represent beta protein sequences classified in classes I – VI resulting from branch specific gene duplication events whereas purple represent alpha protein sequences classified in groups 1 -4 also resulting from branch specific duplication events. 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.

Both the nucleotide and protein based phylogenetic analyses are not a representative of the tissue-specific expression of the major Human b-tubulin isotypes (Cristina Rodr ´ıguez-Antona et al 2009) in humans. Through the use of RT-PCR, Cristina Rodr ´ıguez-Antona et al 2009 discovered that tissues with the highest beta- tubulin expression were as follows; thymus for TUBB; peripheral blood leukocytes for TUBB1; brain for TUBB2A, TUBB2B, TUBB3, and TUBB4 and heart for TUBB2C and TUBB6. All these expression patterns are not representative clades on protein based phylogenetic analysis of human beta tubulin (Fig. 2). The phylogenetic grouping represents the specific gene loci for example TUBB2A, TUBB2B are all positioned on chromosome 6 and their proteins only differ by 2 amino acids. However, TUBB2A although highly expressed in most tissues that TUBB2B, it was not found after examination of the **Pan troglodytes** (Common chimpanzee) genome which share a common recent ancestor with humans implying that TUBB2A was recently acquired in humans by duplication of TUBB2B.

Conclusion

There are multiple copies of tubulin genes across the eukaryotic kingdom. Cell biologists working with different species started isolating genes of mammalian tubulin and giving them randomized names but when they became too many, the naming became very complicated resulting in orthologous genes having different nomenclatures. The other discrepancy is due to the fact that these genes are highly similar, co-expressed and sometimes interchangeable. For example, on the protein level, human alpha tubulin proteins are 90% identical and 72% identical at the nucleotide level. The HUGO Gene Nomenclature Committee is a web database whose goal is to assign a unique and meaningful name to every human gene. The previous nomenclature of these proteins was not right due to the reasons cited above however, with the advancement in molecular phylogenetics, discrepancy in naming was resolved. Evolutionary relationships were defined after aligning the sequences first and then have a tree built by employing maximum likelihood and distance analysis. I have validated the nomenclature of mammalian beta and alpha tubulin protein coding nucleotide sequences into 4 groups for alpha tubulin and 6 classes for beta tubulin than it was earlier reported. These groups correspond to the prefixes 1, 2, 3 and 4 with the sister members (orthologus) assigned with letters A, B, C depending on how many they are and classes range from I to V1. In addition, 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 thus conclude that phylogenetic is a powerful tool to understanding evolutionary relationships among species and on the genome level.

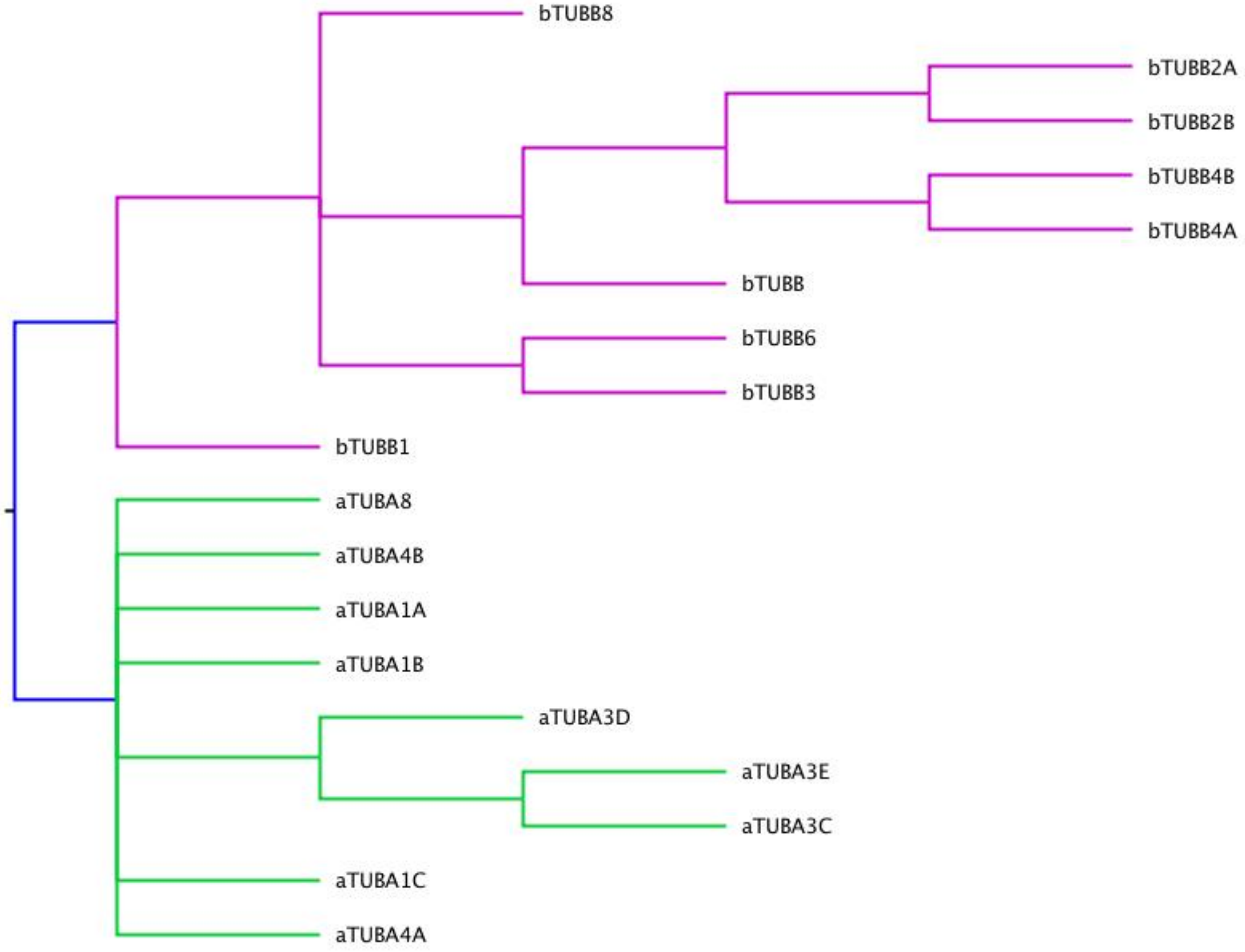
References

1. 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.
2. 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**]
3. Markova et al, 2015; Genetic Disorders Affecting Tubulin Cytoskeleton. Journal of Biomedical and Clinical Research
4. http://envgen.nox.ac.uk/bioinformatics/docs/codeml.html
5. http://abacus.gene.ucl.ac.uk/software/pamlFAQs.pdf
6. A. Villasante, *et al..1986*. **Six mouse alpha-tubulin mRNAs encode five distinct isotypes: testis-specific expression of two sister genes.** Mol. Cell. Biol., 6 (1986), pp. 2409-2419
7. Cristina Rodr ´ıguez-Antona et al 2010. Tumoral and Tissue-Specific Expression of the Major Human b-Tubulin Isotypes. Cytoskeleton, April 2010 67:214–223 (doi: 10.1002/cm.20436)

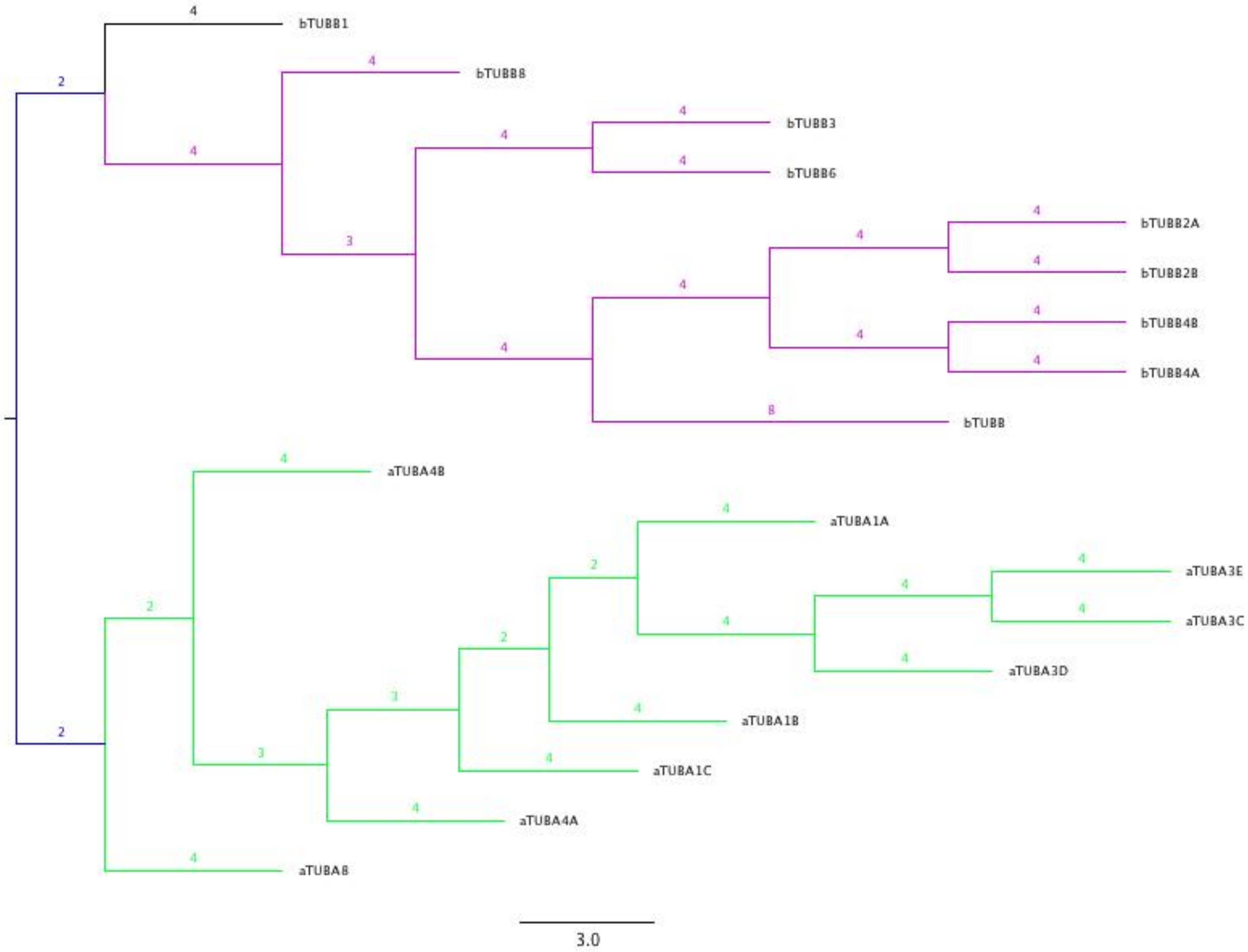
Supplemental data

S1

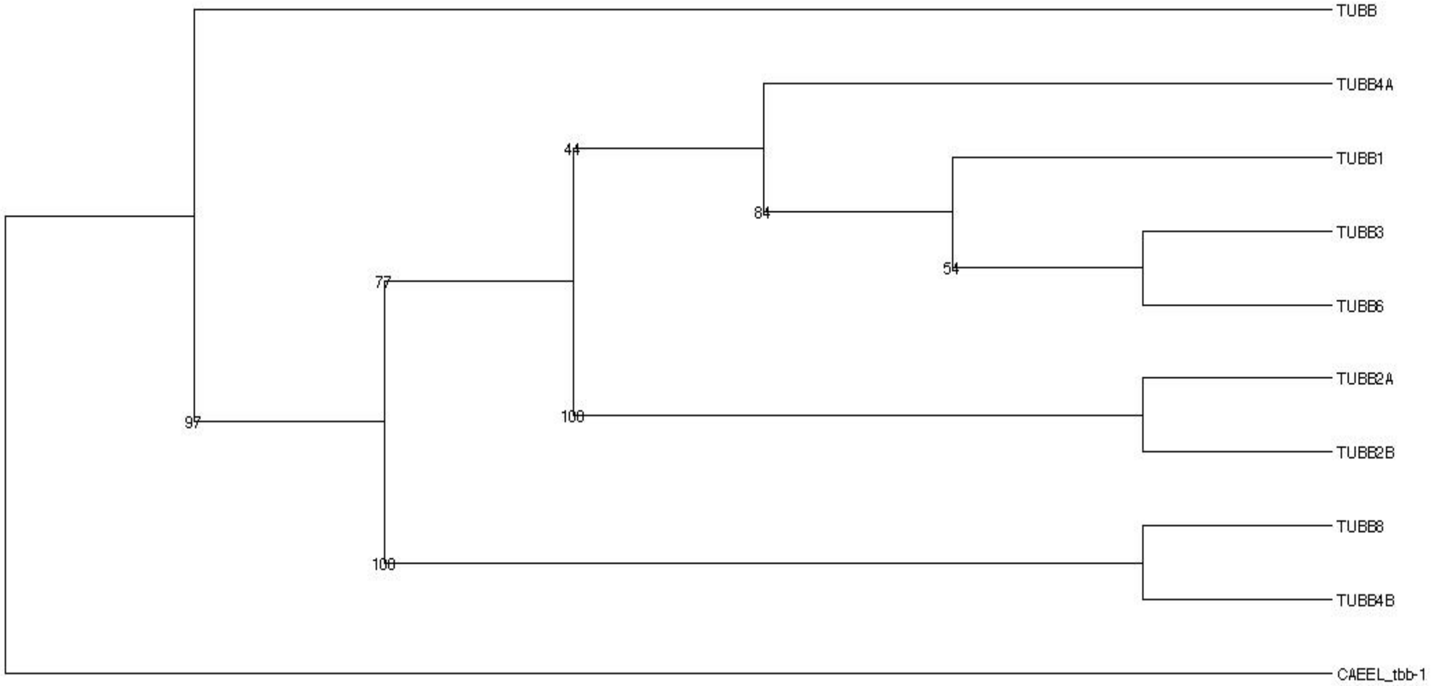
Strict consensus tree constructed from neighbor joining tree of 4 models of amino acid substitution, JTT, PMB, PAM and Kimura



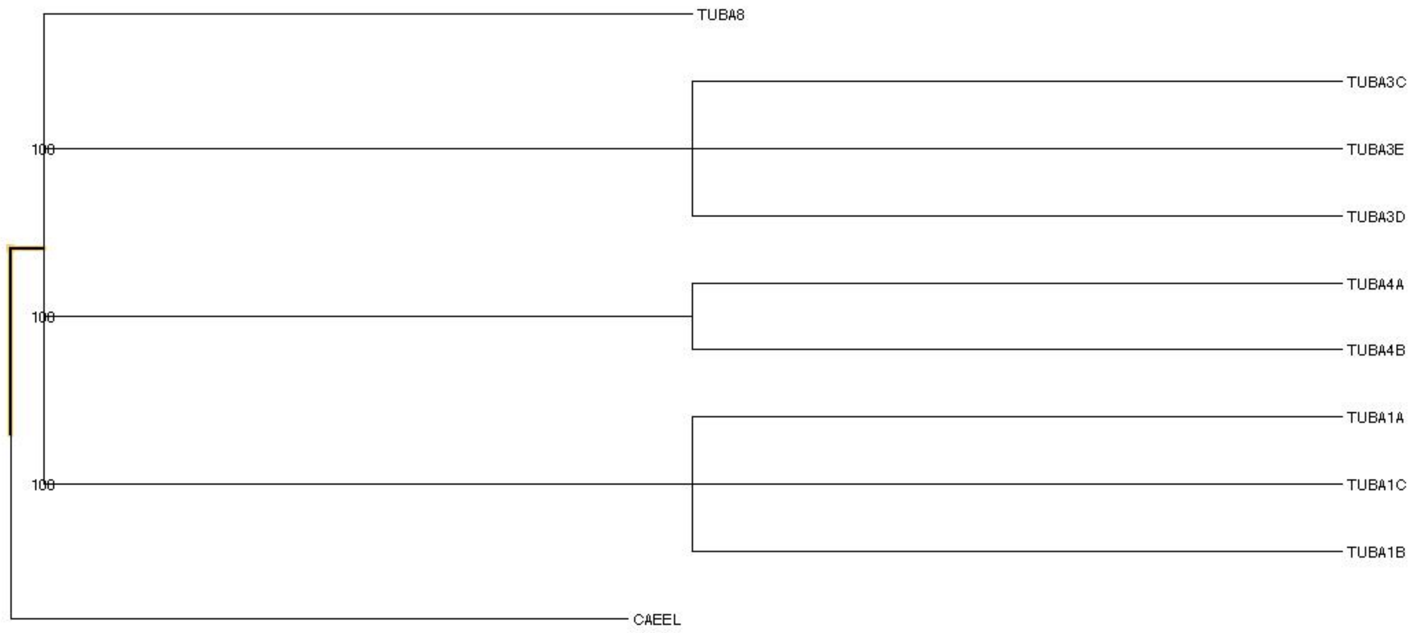
S2: Majority rule tree constructed from neighbor joining tree of 4 models of amino acid substitution employed in phylip JTT, PMB, PAM and Kimura



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.


TUBA8
TUBA3C
100
TUBA3E
TUBA3D
TUBA4A
100
TUBA4B
TUBA1A
100
TUBA1C
TUBA1B
CAEEL