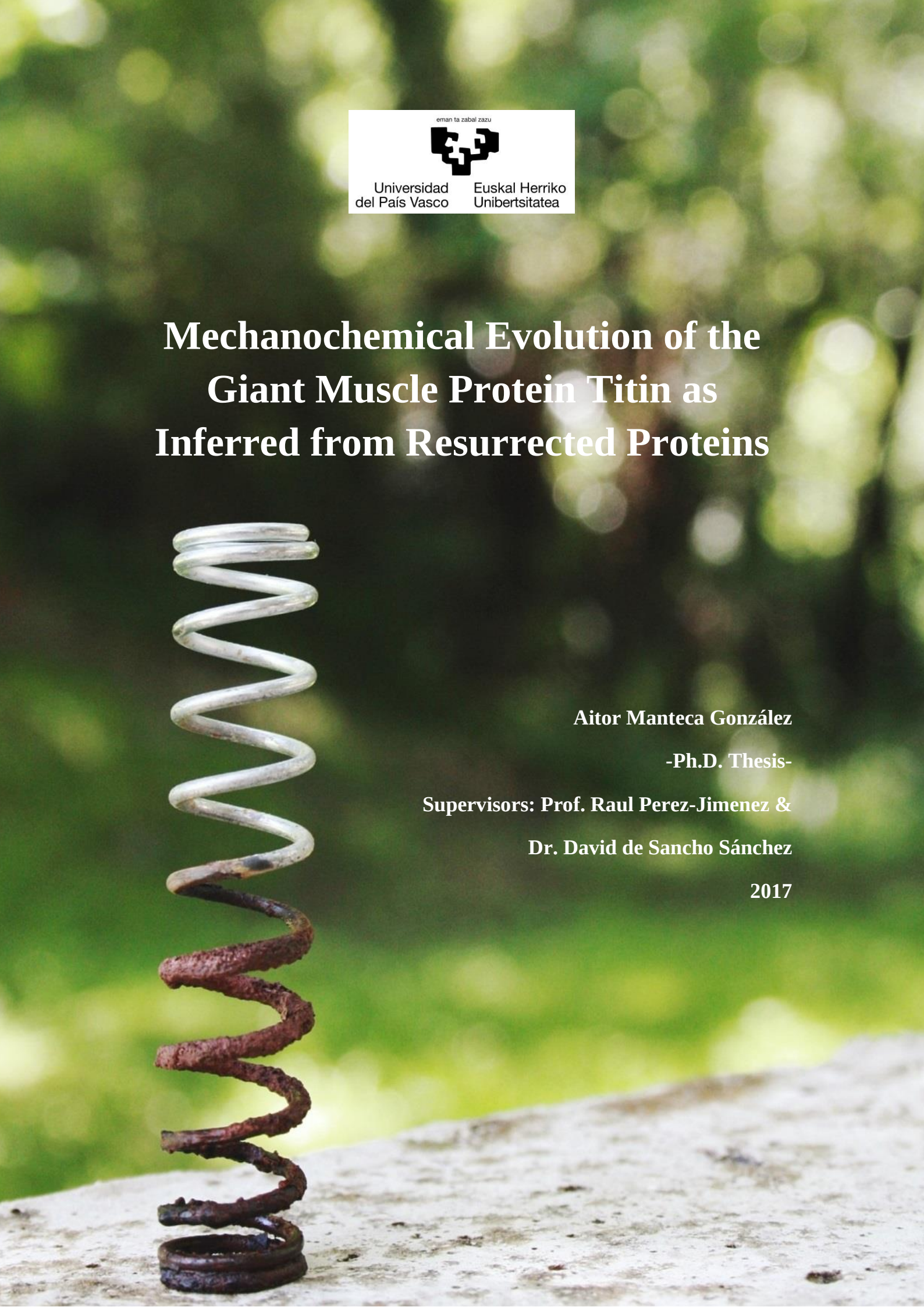


eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Mechanochemical Evolution of the Giant Muscle Protein Titin as Inferred from Resurrected Proteins

Aitor Manteca González

-Ph.D. Thesis-

Supervisors: Prof. Raul Perez-Jimenez &

Dr. David de Sancho Sánchez

2017


CIC
nanoGUNE
nanoscience cooperative research center

Lilurarik ez!
ez dago itzultzerik
eguna atean dago
haize hotza dakar
ez da izango beste goizerik.

Liluraren kontra
Bertolt Brecht / Mikel Laboa

Laburpena

Giharrak ornodun guztietan aurki ditzakegun ehun uzkurkor bigunak dira. Giharren funtzio nagusia indarra sortzea da. Hori dela eta, animalien lokomozioaren arduradunak dira, baita ere zenbait barne organoen mugimendu pasiboena (bihotzarena edo hesteena, esaterako). Giharrak zuntz edo filamentu ezberdinez osaturik daude eta sarkomeroa da oinarrizko unitatea (ikus **1a irudia**).

Gihar filamentuen aurkikuntzaren gurasotzat jotzen dituzte sarri Hugh Huxley eta Jean Hanson. Lagin biologikoen mikroskopia elektronikoaren alorrean 1950eko hamarkadako hobekuntzak probestuz, giharretako filamentu "lodi" eta "meheak" bereizi zituzten lehen aldiz. Huxleyk berak filamentu lodia miosinaz eta mehea aktinaz eginak zeudela frogatu zuen ondoren. Gertakari horien harira, Hugh Huxley eta Adolf Huxleyk lehen "Filamentu labainkorren teoria" (*sliding filament theory*, ingelesez) proposatu zuten aldi berean, nahiz eta familia bereko kideak ez izan, ezta lankideak ere. Hala ere, 1960ko hamarkadara arte, komunitate zientifikoak ez zuen teoria hori onartu. Halaber, teoria hori garatzearekin batera, Huxley eta Hanson ohartu ziren giharrak elastikoak izaten jarraitzen zutela, nahiz eta filamentu lodia eta mehea erauzi. Horrek, sarkomeroan, gihar ehunen oinarrizko egituran, hirugarren filamentu bati ateak ireki zizkion. Handik bi hamarkadara, teknika inmunohistokimiko eta elektroforesi gelen aurrerapauso nabarmenei esker, **titina** lehen aldiz isolatua izan zen. Horregatik, batzuetan, titina bi Huxleyen filamentu ezkutua bezala da ezaguna.

Titina bizkarmuina duten animalia guztiek komunean duten mikrometro bat luzeko proteina da. Giharretan aurki dezakegun hirugarren proteina ugariena da (aktina

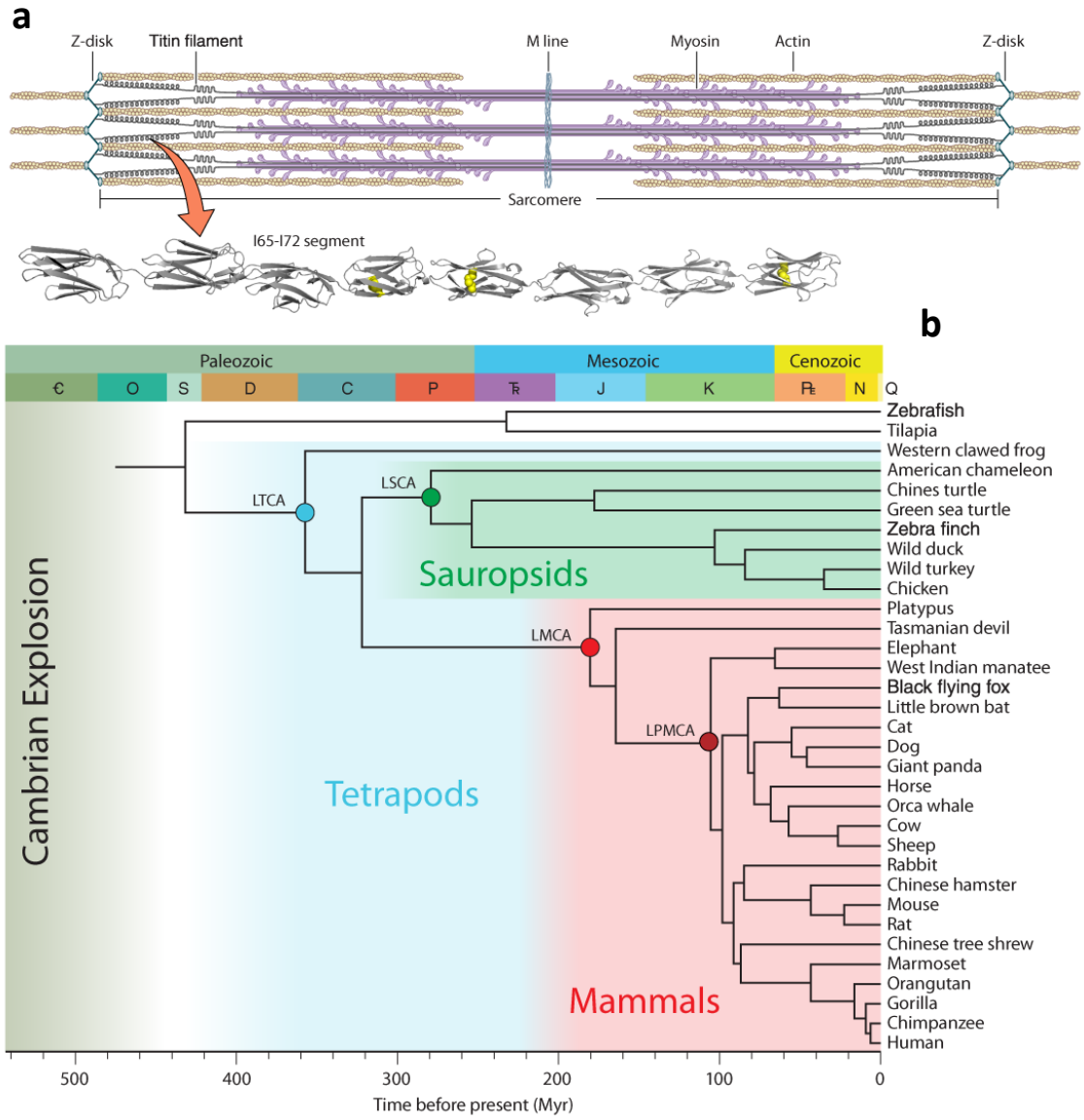
eta miosinaren ondoren) eta ezagutzen den handiena, batzuetan 35.000 aminoazidoko kopurua gainditzen baitu. Banan-banan tolesturiko ehunka domeinuz osaturik dago. Bere osagai nagusiak inmunoglobulina (Ig) domeinuak, III motako fibronektina (FnIII) domeinuak eta egitura tertziariorik gabeko PEVK (prolina-glutamatoa-lisina-balina) erregioa dira. Titina bi eskualdek osatzen dute: A-banda, gehien bat ez elastikoa; eta I-banda, titinari malguki propietateak ematen dizkiona. A-banda Ig-FnIII zorizko tandem errepikapenez osaturik dago. I-banda, berriz, oso anitza da isoformaren arabera. Hala eta guztiz ere, baditu ezaugarri komunak: Ig super-errepikapenak eta PEVK erregioa. Gihar motaren arabera ehunka isoforma izan ditzakeen arren, denak hiru taldetan bana ditzakegu: Hezurdurako N2A, Bihotzeko N2B eta Bihotzeko N2BA. Lehena gorputzeko edozein giharretan aurki dezakegu. Azken biak, aldiz, bihotzean soilik.

Ezaguna da titinak hainbat prozesu biologikotan parte hartzen duela; seinale transdukzioan edo sarkomeroaren egituraren egonkortasunean, esate baterako. Hala ere, bere funtzio esanguratsuena giharrei malgutasun pasiboa ematea da. Ezaugarri harrigarri horientzat gaur egun onarturiko mekanismoa honakoa da: Gihar bat luzatzen den bakoitzean titinaren Ig domeinuak lerrokatzen hasten dira, eta giharrari elastikotasuna ematen diote, tentsio pasibo ahul batekin. Ondoren PEVK erregioa luzatzen da. Azkenik, I-bandan kokaturik dauden Ig domeinuak destolesten dira, zurruntasuna erabat handituz. Horren harira, azken urtetan titinaren propietate mekanikoen ezagutza modu esponentzialean handitu da **molekula bakarreko indar espektroskopia**ren garapenari esker (ingelesez *single-molecule force spectroscopy*, smFS). smFSa indar atomikoko mikroskopioaren (ingelesez, *atomic force microscope*, AFM) aldakuntza bat da. Aldaketa horren bitartez, laginaren irudi topologikoa atera beharrean, haren propietate mekanikoak neur daitezke. Modu horretan, smFSak titina domeinu ezberdinetan indar mekaniko kalibratuak aplikatzea ahalbidetzen du, hots, domeinu bakoitzaren egonkortasun mekanikoa kalkula daiteke. smFSa lehen aldiz ADNaren hari osagarriak askatzeko behar zen indarra kalkulatzeko erabili zen 1994ean. Handik gutxira, titinaren lehen immunoglobulina domeinuak neurtu zituzten. Gaur egun, titinaren hamarnaka domeinuren egonkortasun mekanikoa jakina da. Saiakuntza mota hauetan proteinarik erabiliena da eta molekula bakarreko esperimentuen *prima donna* bezala da ezaguna.

Titinaren propietate mekanikoak eta funtzio biologiko ugariak kontutan hartzen baditugu, horrek ornodunek duten izugarrizko dibertsitate fisiologikoarekin eta giharrek

jasan duten eboluzioarekin erlazio zuzena izan duela pentsa dezakegu. Bale urdin kolosaletik kolibri nimiñora, ornodun guztiek dute amankomunean titina. Hala ere, dibertsitate honen ezaugarri molekularra oraindik ezezaguna dugu. Azken hamabost urteetako informazio genetikoaren eztandarekin batera, baina, analisi estatistikoen hobekuntzak etorbide berriak ireki ditu konparazio biologiaren alorrean. Horrela, metodo filogenetiko berriek ADN zein proteinen **eboluzio molekularra** azter ditzakete. Hau da, egungo organismo bizien biomolekulak alderatuz, erlazio ebolutiboak ezar daitezke haien artean eta, bide batez, arbaso komunen sekuentziak ondorioztatu. Beraz, **Antzinako sekuentzien berregituratzea**k (ingelesez, *ancestral sequence reconstruction*, ASR) duela milioika urte bizi izan ziren espezieen biomolekulak berpizteko aukera ematen du, eta, punta-puntako teknika biofisiko eta biokimikoekin bat eginik, haien hainbat ezaugarri azter daitezke. Lanabes horiekin guztiekin, eta biologikoki garrantzitsuak diren egiturak aztertzeko asmoz, ASRaren arloan makina bat ikerketa loratu dira 2010etik aurrera. Adibidez, biokimikan giltzarri diren zenbait entzimen (tiorredoxina edo hainbat zelulasa, esaterako) arbola filogenetikoak eraikiak izan dira eta, biologia molekularreko teknikei esker, duela milioika urte bizi izan ziren arbasoak laborategian berpiztuak. Horri esker, entzimen errendimendua tenperatura altutan hobetua izan da, arbasoak bizi izan ziren garaian lurreko tenperatura ere altuagoa baitzen. Prozesu horri termoadaptazioa deitzen zaio.

Titinaren kasuan, nahiz eta zenbait ikerketa filogenetiko burutu diren, gaur egun oraindik ez dago 10-15 espezie baino gehiago batzen dituen arbolarik. Arrazoia, bere tamaina erraldoia eta horrek atxikiak dakartzan koste konputazional izugarriak dira. Hortaz, tesi honetan, titinaren arbola filogenetikoa lehen aldiz berreraikia izan da 34 espezie kontutan hartuz (**1b irudia**), animalia hauen eta haien arbasoen ezaugarri nanomekanikoak aztertzeko asmoz. Hau gauzatzeko, biologikoki garrantzitsuak iruditzen zaizkigun lau arbaso hautatu ditugu: tetrapodoen azken arbaso komuna (LTCA, *last tetrapod common ancestor* ingelesez), sauropsidoen azken arbaso komuna (LSCA, *last sauropsid common ancestor* ingelesez), ugaztunen azken arbaso komuna (LMCA, last mammal common ancestor ingelesez) eta ugaztun karendunen azken arbaso komuna (LPMCA, *last placental mammal common ancestor*, ingelesez). Arbaso horiek duela 356, 278, 179 eta 105 milioi urte (Myr) bizi izan ziren, hurrenez hurren. Gaur egungo beste bost espezie ere aukeratu ditugu haien titinen propietate mekanikoak aztertzeko: gizakia, arratoi arrunta, orka, oiloa eta zebra txonta.

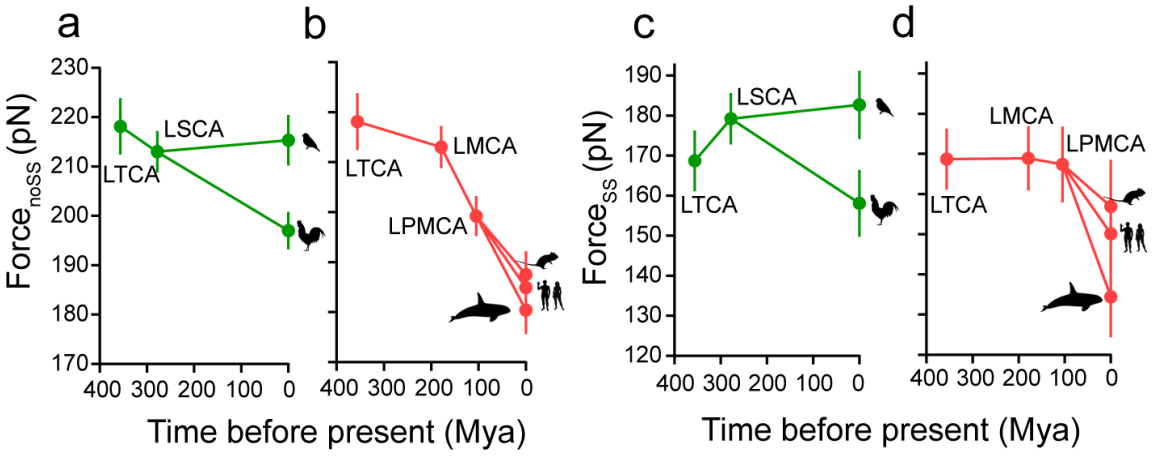


1 irudia: (a) *Sarkomeroaren egitura eskematikoa. Bertan filamentu lodia, mehea eta titina ikus ditzakegu. Bertan esi honetan ikertutako I65-I72 domeinuen zooma ere ikus daiteke.* ***(b)*** *Titinaren kronograma edo arbola filogenetikoa. Ugaztunak, sauropsidoak eta tetrapodoak kolore ezberdinetan banaturik daude.*

Titinaren kronogramarekin batera, miosinaren arbola ere eraikia izan da giharreko zuntzek jasan duten eboluzioa ikus ahal izateko (aktinarena egitea ezinezkoa da proteina honek duen espezieen arteko identitate altuagatik). Emaitzei begira, titina eta miosinaren mutazio tasak alderatzen baditugu, lehenak, era harrigarrian, bigarrenaren tasa bikoizten duela ikus dezakegu. Hori dela eta, titina eboluzioan zehar informazio genetikoaren

eramailea izan dela hipotetiza dezakegu eta, horrela, gaur egungo animalien dibertsitate harrigarrian erantzukizun zuzena edukiko luke.

Titina ezberdinen propietate mekanikoak aztertzeko, haren segmentu bat aukeratu dugu smFSarekin ikertzeko asmoz, bere tamainak teknikoki ezinezkoa egiten baitu osorik arakatzeak. Hau dela eta, I65etik I72ra doan 8 Ig domeinuko zatia hautatu dugu. Zati hau esanguratsua da, titinaren zati elastikoan aurkitzen baita. Emaitzak aztertuz, Ig domeinuen egonkortasun mekanikoa eta disulfuro zubien presentzia titinaren eboluzioaren muina direla ikus dezakegu. Orohar, antzinako titinen egonkortasun mekanikoa handia den bitartean, gaur egungo animalien Ig domeinuen propietate mekanikoak xumeagoak dira haien arbasoekin alderatzen baditugu. Disulfuro zubien ehunekoa ere handiagoa da antzinako titinetan. Bestalde, disulfuro zubidun eta zubigabeen domeinuen egonkortasun mekanikoaren konparaketa eginez gero, lehenek propietate mekaniko ahulagoak erakusten dituzte espezie guztietan. Gainera, emaitza horiek guztiak espezie bakoitzaren adinarekin alderatuz gero, joera paleomekaniko desberdinak aurkituko ditugu animalia klado desberdinen eta domeinu motaren arabera (**2 Irudia**). Hala, sauropsidoen joera egonkortasun mekanikoa mantentzean izan den bitartean, ugaztunen kasuan egonkortasun honek jaitsiera itzela jasan du, batez ere disulfuro zubirik ez duten domeinuetan. Gainera, elementu nanomekaniko horiek animalien hainbat ezaugarri fisiologikorekin estuki loturik daudela ikusi dugu.



2 irudia: *Titinaren joera paleomekanikoa domeinu motaren eta animalia kladoen arabera. Sauropsidoen joera paleomekanikoa disulfuro zubigabeko (**a**) eta zubidun (**c**) domeinuentzat. Ugaztunen joera paleomekanikoa disulfuro zubigabeko (**b**) eta zubidun (**d**) domeinuentzat.*

Antzinako titinen disulfuro zubien proportzio handiago hau baieztatzeko asmoz, beste bi teknika esperimental frogatu ditugu. Lehena, molekula bakarreko disulfuro erredukzio entseiu bat da. smFSak indar konstantean lan egiteko duen aukera probestuz, metodo honekin disulfuro kriptikoen erredukzioa lor daiteke inmunoglobulina domeinuetan zenbait oxidorreduktasa entzimen bitartez. Gure kasuan tiorredoxina entzima erabili dugu. Bigarrena, tiol askeak detektatzeko gai den teknika biokimiko bat dat. Bi teknika hauen emaitzak koherenteak dira aldez aurretik smFSak lortutako emaitzekin eta aurkitutako joera baieztatzen dute.

Azkenik, emaitza horiek guztiak aintzat hartuta, titinak denboran zehar jasan izan dituen egokitze mekanikoek gaur egungo animalien aldaketa fisiologikoetan eta dibertsitate morfologikoan eragin zuzena izan duela pentsa daiteke. Animalia txikiek inmunoglobulina domeinu zurrunak eta disulfuro zubi kopuru handia duten bitartean, handien egonkortasun mekanikoa ahulagoa da eta disulfuroen ehunekoa, txikiagoa. Hala, animalien ezaugarri fisiologikoen eta haien titinen propietate mekanikoen analisi konparatiboak antzinako espezieen hainbat ezaugarri (pisua, luzera edo bihotz taupadak, esaterako) aurreikusteko aukera ematzen digu. Oro har, estimazio horiek antzinako animaliak txikiak, arinak eta azkar mugitzen zirela diote eta garai hartako fosil aurkikuntzekin bat egiten dute.

Summary

Muscles are soft tissues that can be found in all vertebrates, being responsible of the locomotion and the motion of some internal organs. Muscles are mainly composed of three filaments: thin filaments compounded of actin, thick filaments made of myosin, and titin. Muscles are arranged in basic units called sarcomeres. The sarcomere-based structure of muscles is conserved among vertebrates; however, vertebrate muscle physiology is extremely diverse. A molecular explanation for this muscle diversity and its evolution has not been proposed. In this Thesis, we use phylogenetic analysis and single-molecule force spectroscopy (smFS) to investigate the mechanochemical evolution of titin, a giant protein responsible for the elasticity, integrity and signal transduction of muscle filaments.

We used phylogenetic methods based on maximum parsimony and Bayesian inference to infer the chronograms of titin and myosin. The phylogeny of actin could not be inferred due to its extremely high conservation among species. Results revealed that the mutation rate is double in titin that in myosin since the Cambrian radiation, suggesting that the first one may be the main evolutionary information carrier in muscles of modern animals. Encouraged by this phenomenon and using maximum likelihood methods, we brought back to life eight immunoglobulin fragments of titin (specifically, the I65-I72 segment) corresponding to ancestors to mammals, sauropsids, and tetrapods, that lived 105-356 million years ago. These species were compared with some of their modern descendants, such us human, brown rat, orca, chicken and zebra finch.

Using smFS we observe that the mechanical stability of titin domains and the presence of disulfide bonds are key elements in the evolution of titin. Our experiments

demonstrate that ancient titin molecules were rich in disulfide bonds and displayed high mechanical stability. Moreover, the unfolding force is always higher in domains that do not contain a disulfide bond for the nine-species studied. These mechanochemical elements seem to have changed over the course of evolution creating different paleomechanical trends for birds and mammals, which correlates with animal physiological properties such as heart rate, body size and body lenght.

Two additional studies were carried out to confirm the presence and percentages of disulfide bonds on these titin segments. The first one is a single-molecule disulfide reduction assay in the presence of thioredoxin that allows to detect the cryptic disulfides in immunoglobulin domains. The second one is a biochemical assay that detects cysteines in the form of free thiols by labeling them with the fluorophore monobromobimane (mBBr). In both experiments the results are consistent with our previous data, corroborating the trend that, in general, ancestral titins contain more disulfide bonds than the ones from extant species.

Thus, considering all these results, we hypothesize that mechanical adjustments in titin contributed to physiological changes that allowed the muscular development and morphological diversity of modern vertebrates. While small animals seem to have stiffer domains with more disulfide bonds, large animals show weaker domains with fewer disulfides. A comparative analysis between animal physiological properties and titin mechanical properties allows us estimating the heart rate, the body size and the body length of the ancestors of modern tetrapods, sauropsids and mammals. The estimations revealed that these ancestors were small, light and fast-moving animals. Moreover, these estimations fit surprisingly well with fossil records found on that geological periods.

Table of contents

Part I

Chapter 1: Introduction

Muscles are a type of soft tissue that can be found in most vertebrates. They are composed of protein filaments that slide past one another, generating extensions and contractions that modify the shape of the cell. Their main function is to create force and motion, being responsible of the locomotion and the passive movement of some internal organs, such as the heart or the intestines. They can be classified as skeletal (or striated), cardiac, or smooth. While skeletal muscles are voluntary and related to posture and locomotion, cardiac and smooth muscles are involuntary and, most of the times, related to the internal organs mentioned above [1].

Muscle filaments constitute the building blocks of muscle tissues in all vertebrates. Their discovery is often assigned to Hugh Huxley and Jean Hanson. Taking advantage of the novel improvements of the electron microscope for biological samples in the early fifties, they discovered for the first time the "thick" and the "thin" filaments in muscles [2, 3]. Hugh Huxley himself demonstrated that the thick filament was mainly composed of myosin and the thin filaments, of actin [3]. Right after these discoveries, the first "*sliding filament theory*" was proposed simultaneously by Hugh Huxley and Andrew Huxley [4, 5]. Nevertheless, it was not until 1960s when this theory was finally accepted by the scientific community. But developing this theory, Huxley and Hanson realized that muscle filaments were elastic even when removing the thick and thin filaments and hypothesized that a third filament may be present in sarcomeric structures, the basic unit of muscle tissues **(Fig. 1.1)**. Two decades later due to the advantages of gel electrophoresis [6] and

immunohistochemical techniques [7] titin was first discovered. Lindstedt and Nishikawa brilliantly refer to titin as “Huxleys’ missing filament” in a recent review [8].

This third filament, titin, is a micrometer-long muscle protein present in all vertebrates. It is the largest known protein [9] (~35000 amino acids) and the third most abundant in striated muscle [10]. It is composed of hundreds of individually folded domains and disordered regions [7]. The main constituents of titin are immunoglobulin (Ig) domains, fibronectin type III (FnIII) domains and the unstructured PEVK (proline-glutamate-valine-lysine) region [11]. In the sarcomere, titin connects the Z disc to the M line, where its carboxy terminus and a kinase domain are located [7]. Titin is composed of two regions; the A-band, which is predominantly inelastic, and the I-band, which confers its elastic properties to titin. The A-band is a tandem Ig-FnIII random super-repeat region. The position of these domains is associated to the surrounding myosin filaments [12, 13]. The I-band region structure strongly depends on the isoform of the titin [14, 15]. Even so, they share common features: the Ig super-repeats and the PEVK region. Although there are tens of isoforms depending on the muscle, all of them can be clustered in three supergroups **(Fig. 1.2)**: the skeletal N2A and the cardiac N2B and N2BA.

The skeletal N2A is the isoform that can be found in all skeletal muscles. It is composed of a proximal tandem Ig-domain region, the N2A sequence insertion, the PEVK region and the distal Ig-domain region. The length of the proximal Ig region is determined by type of muscle and, in general, its length rules the slack length of muscle sarcomere. The cardiac N2B is the shortest and the stiffest of titin isoforms. It is exclusive of cardiac muscles and it is composed of a short proximal tandem Ig-region, the N2B sequence insertion, the PEVK region and the distal tandem Ig-region. Finally, the cardiac N2BA is

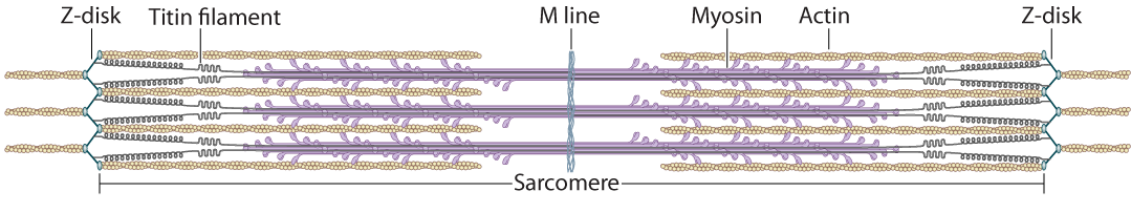


Figure 1.1. *Scheme of the sarcomere from Z disk to M line. The three main sarcomeric proteins actin, myosin and titin are shown.*

the other titin isoform that can be found only in cardiac muscles. Although it is much more elastic than N2B isoform, it does not reach the elasticity of the N2A skeletal isoform. It is composed of a short proximal Ig-tandem region, the N2B sequence insertion, a medium Ig-region with variable length, the N2A sequence insertion, the PEVK region, and the distal Ig tandem region.

Although titin has been shown to participate in several processes related to signal transduction [16, 17], the main and most studied function of titin is providing passive elasticity to the muscle [18, 19]. Thanks to the advances in immunohistology and single-molecules techniques, the currently accepted mechanism for this outstanding property is that when a muscle stretches, the Ig domains align themselves from the initial zig-zag orientation at first, providing elasticity to the muscle with a low passive tension. For longer stretches, the PEVK region elongates, and finally, the Ig domains unfold, increasing the stiffness drastically [20]. In the heart, both N2B and N2BA isoforms coexist in the same muscles. In this case, the ratio of N2BA to N2B regulates the stiffness and elasticity of the muscles. This ratio often changes in different animals depending on their size and heart rate. In addition, higher expression levels of the compliant N2BA titin isoform have been

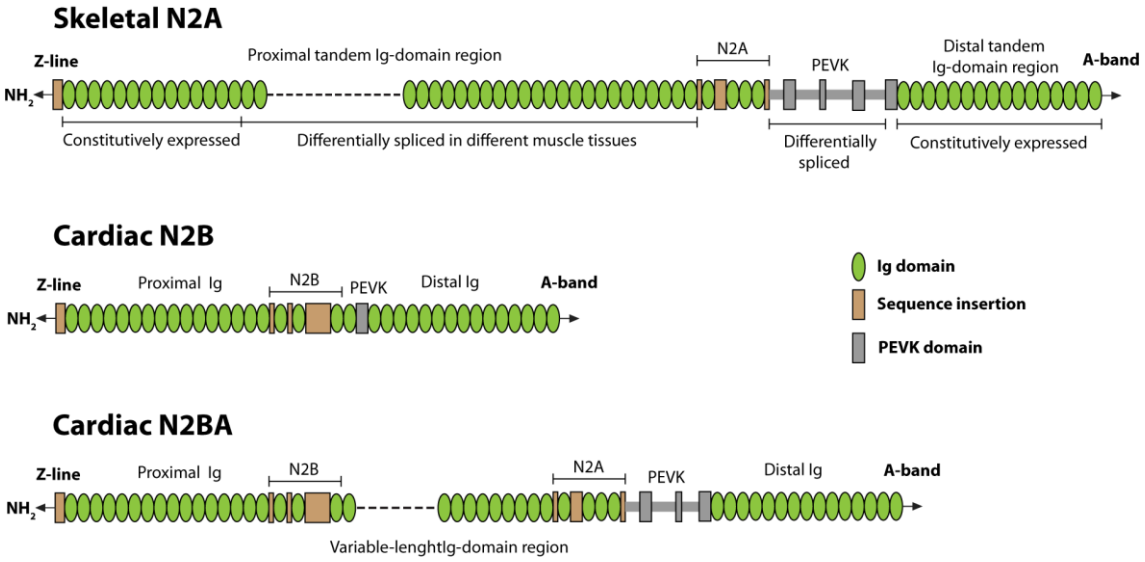


Figure 1.2. *Domain scheme of skeletal (N2A) and cardiac (N2B, N2BA) isoforms of titin in the I-band region. In each of the isoforms the location of Ig domains, the specific insertions and the PEVK region are shown.*

observed in human heart failure due to chronic ischemic cardiomyopathy [21], and an increased N2BA/N2B isoform expression ratio was accompanied by decreased myofibrillar passive force in dilated cardiomyopathy patients [22]. Moreover, recent studies support an important role of titin Ig domains in refolding during contraction, although these findings may be controversial [23].

In the past two decades, our knowledge of the mechanical properties of titin has increased dramatically due to the use of single-molecule force spectroscopy (smFS) techniques, which make it possible to apply calibrated mechanical forces to titin domains [24, 25]. The concept of employing the atomic force microscope (AFM) for measuring mechanical forces of biological samples was applied first in 1994 in the study of the forces that rely between the complementary strands of DNA [26]. Shortly after, this setup was used to calculate the unfolding forces of titin immunoglobin domains for the first time [24]. Afterwards, several Ig domains or consecutive Ig domain constructs of titin have been studied Such as I1 [27], I4 [19], I5 [19], I4-I11 [19], I27 (now termed I91) [24, 28-36], I28 [19, 30, 37], I27-I34 [38, 39], I32 [19] and I34 [19], I65-I70 [40] and I91-I98 [40], among others. The average unfolding forces for these domains in the constant velocity mode ranges between 127 pN for I1 and 330 pN for I34. Moreover, apart from Ig domains, other titin fragments have been also mechanically analyzed; various FnIII domains [41, 42], the N2B bus [19, 43] or the PEVK region [19, 43-45], showing a huge variety of mechanical stabilities (<20 pN for the PEVK region vs. 220 pN for ^{1}FNIII and ^{2}FNIII). In fact, titin is the most studied protein by smFS by far.

In the typical setup for smFS experiments the protein of interest is absorbed onto a gold surface and the cantilever is moved in z-direction towards the surface. Once the cantilever reaches the surface and thereupon retracts, if the protein gets attached to the tip of the cantilever it will be subjected to a mechanical force. The setup of the experiments carried out in this thesis has an upside-down design (**Fig. 1.3a**), where the cantilever, the laser and the sensor are fixed while the substrate with the attached protein is moved by a piezoelectric actuator. A laser is focused towards the cantilever and a photo detector (PD) will measure the deflection of the cantilever in terms of changes in the laser intensity. In force extension mode experiments the velocity of the piezo is constant, so the unfolding of the polyprotein will be expressed as a sawtooth pattern graph (**Fig. 1.3b**), where each of the peaks correspond to the unfolding of a single domain. On the other hand, in force clamp

mode, the AFM operates at constant force. The applied force can be controlled with a PID controller (proportional-integral-derivative) that generates a feedback loop. This time, the unfolding of a polyprotein will result in staircase pattern graph (**Fig. 1.3c**) where each step corresponds to the unfolding of a domain.

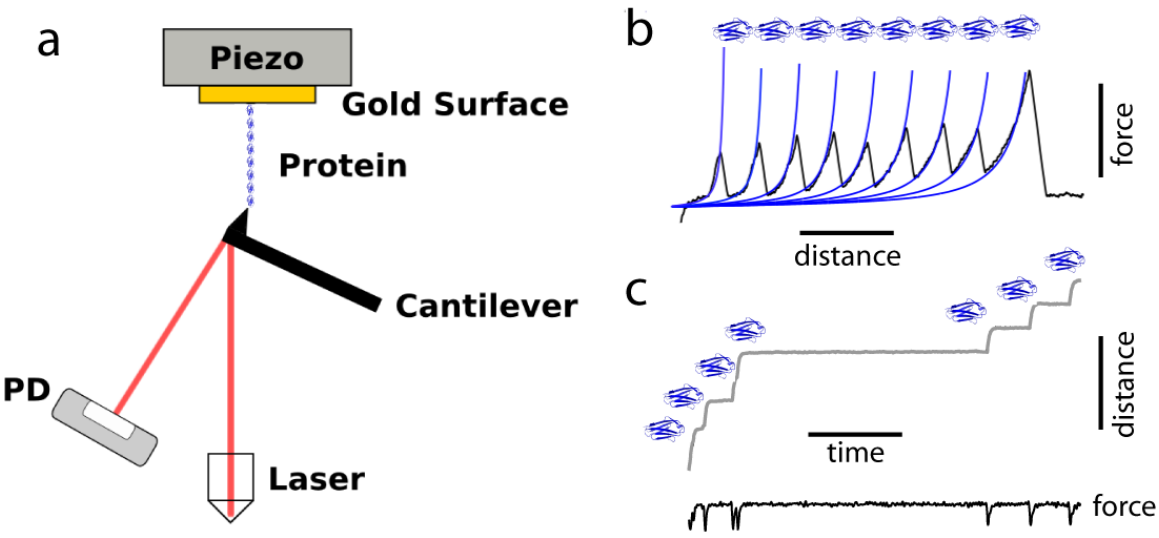


Figure 1.3. ***(a)*** *Schematic representation of a single-molecule force spectrometer.* ***(b)*** *Characteristic force extension trace for the* $(I27)_8$ *polyprotein.* ***(c)*** *Characteristic force clamp trace for the same homopolyprotein.*

In characteristic force-extension or force-clamp experiments, hundreds of curves are recorded, but only a small amount of these curves contain relevant information related to the behavior of a single-molecule. Hence, a large amount of curves are discarded (up to 99%) and the user has to search for a "needle in a haystack" [46] among a huge amount of unfit traces. Most of the times this happens because a lack of the protein of interest, but also to other factors such as interaction of the protein with the surface, the attachment of two proteins to the tip or the lack of detachment of the protein from the previous pull [47]. For a proper filtering of this data, there are few proteins that can be used as explicit identifiers for smFS traces. The best known one is the I91 domain (formerly I27) of titin. I91 has been used as a molecular fingerprint for the study of several proteins due to its well-known mechanical properties. I91 is composed of 89 amino acids that form a

characteristic β-sandwich structure consisting of two four-stranded sheets (**Fig. 1.4**). The mechanical barrier to unfold the domain completely comes from the three hydrogen bonds (H-bonds) between the β-stands A and B in a first phase, reaching an intermediate state, and the six H-bonds between the strands A'-G in a second phase [48], so called the "molecular clamp". After this, the rest of the domain is extended with very little resistance. Several proteins have been mechanically characterized by smFS taking the I91 domain as a molecular fingerprint, for instance the bacterial protein barnase [49], the HIV receptor CD4 [50] or the green fluorescent protein GFP [51].

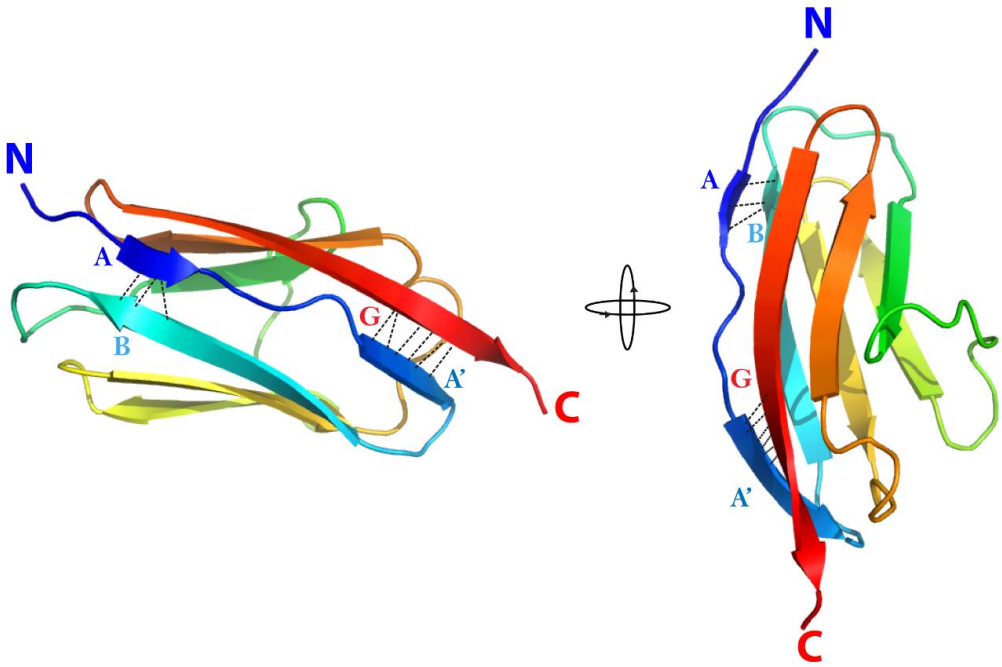


Figure 1.4. *Three-dimension structure of the I91 (formerly I27) domain. The 8 beta-strands can be visualized.*

The central hypothesis of this thesis in how the enormous physiological variety of the vertebrates could be related to titin, since the mechanical properties and functional features of muscles are directly connected to this gigantic protein. From the tiny hummingbird to the colossal blue whale, all the animals with a spinal cord share this protein or a similar one in terms of structure. The molecular component of this diversity remains unknown. In addition, the increasing amount of genetic data the last 20 years offers new

avenues for comparative biology to better understand biological systems. Thanks to the growing number of genome sequencing for different organisms, nowadays it is possible to compare and mechanically characterize titins from diverse animals and compare these results to their physiological features.

In this regard, phylogenetic methods applied to genomic information have made it possible to establish evolutionary relationships among different living organisms, including the possibility of inferring the putative sequences of the genes of their already extinct ancestors [52, 53]. Since Charles Robert Darwin sketched an evolutionary tree in 1837 for the first time (**Fig. 1.5a**) to the current *Time Tree of Life* [54] for all the known living organisms (**Fig. 1.5b**), the field of phylogenetics has raised constantly. Today, combined with biophysical and biochemical state-of-the-art techniques, ancestral sequence reconstruction allows to study and compare features of extinct biomacromolecules that are otherwise intractable. Thus, it is also possible to reconstruct the phylogenetic tree of titin and resurrect this protein from species that are biologically relevant in terms of the geological time scale.

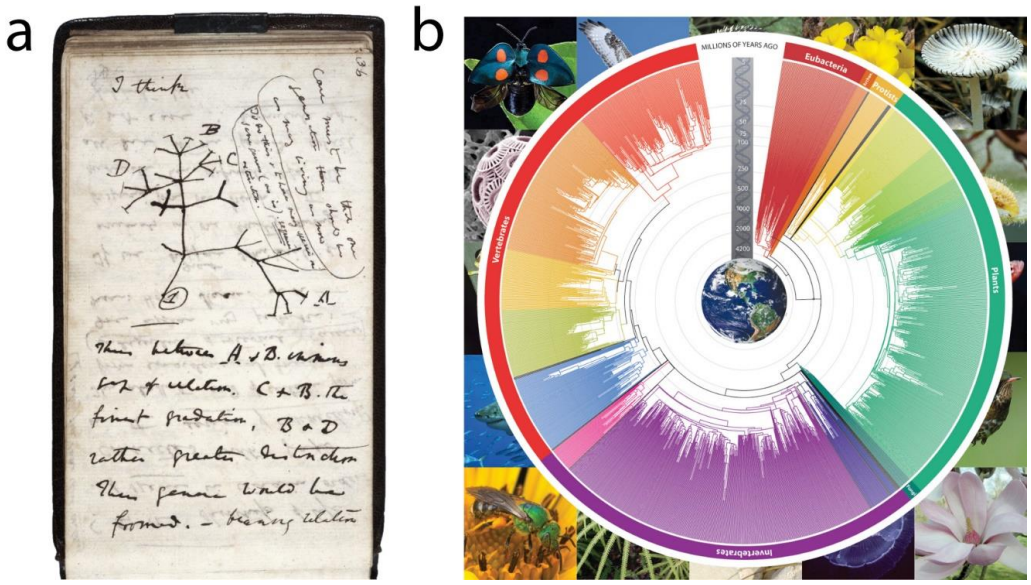


Figure 1.5. (a) *First sketch of an evolutionary tree in the notebook of Charles R. Darwin "Transmutation of Species B" with the sentence "I think" above.* ***(b)*** *Current Time Tree of Life, correlating all the known species (taken from http://www.timetree.org).*

The ancestral protein reconstruction has four well defined steps. First, a selection of homologous sequences from extant species is performed. After this, the sequences are organized by a creation of a multiple alignment. Later, using statistical tools such as maximum parsimony, maximum likelihood or Bayesian inference, it is possible to compute a phylogenetic tree that relates how species are correlated between each other by descent from common ancestors. Finally, one can infer the ancestral sequences and bring them to back to life by molecular biology techniques. The methodology for this process is illustrated in **Fig. 1.6**.

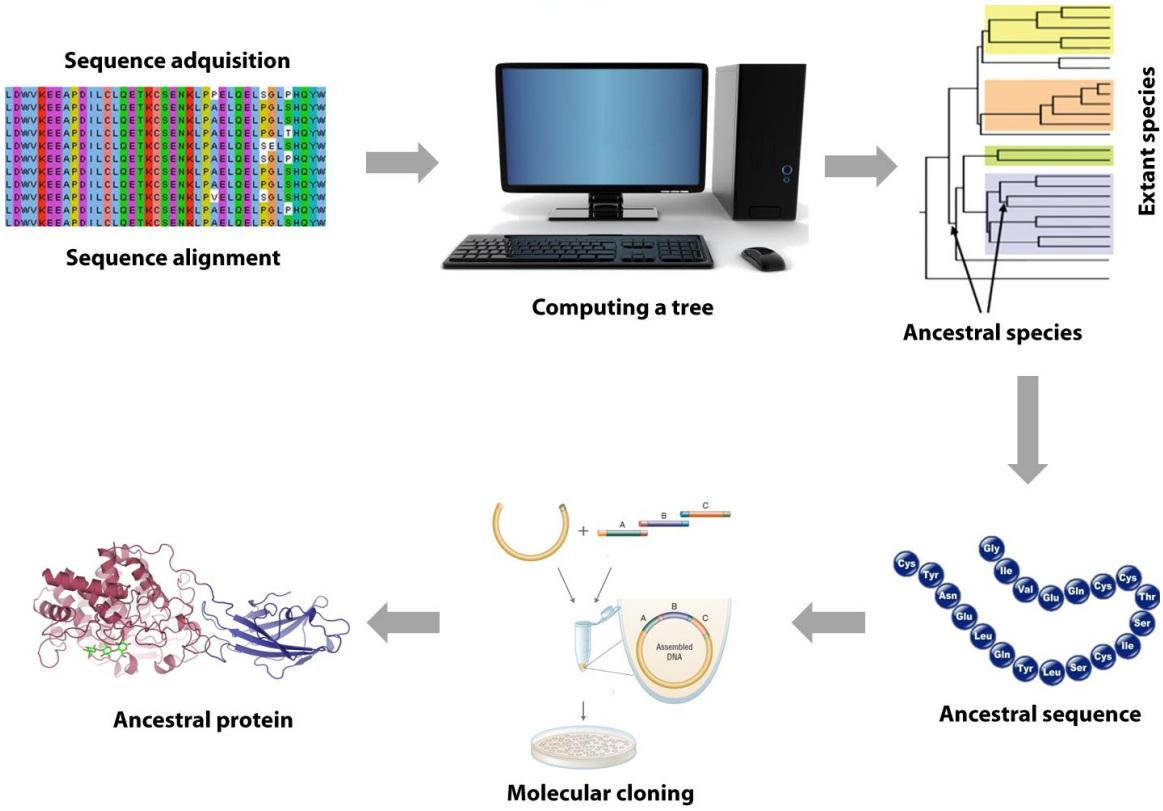


Figure 1.6. *Methodology for ancestral protein reconstruction.*

Taking advantage of this, several notable studies have been developed the last decade. This information relates to physiological and metabolic features [55, 56], but also to the environmental conditions that hosted ancestral organisms [57, 58]. One of the pioneering works of ancestral reconstruction in the field of molecular evolution is the prediction of the tertiary structure of some proteins by means of its residue contacts in 1994

[59]. One of the most relevant applications is the deduction of the environmental conditions of different geological eras. Several studies involving the reconstructions of billions-of-years old ancestral proteins have reported the conditions of the earth at the Precambrian era as a serendipitous but enormous valuable result [57, 58, 60-63]. Chronologically, the first study in this field was reported by Erik Gaucher and collaborators in 2008 [57]. In this work, they reconstructed the translation elongation factors of species that lived in the range of 3.5-0.5 Gyr ago and calculated their melting temperature. According to this study, the temperature on earth cooled down over 30 ºC during that period, matching the previous works related to the temperature of ancient oceans calculated from silicon isotopes [64]. Over the next years, several similar studies with different proteins were reported confirming this cooling trend. For example, Perez-Jimenez et al. studied the thermochemical evolution of thioredoxins from 4 to 1.4 Gyr with single-molecule force spectroscopy [58]. This thermoadaptation has also been recently studied by Kern and collaborators for adenylate kinase ancestral enzymes spanning 3 Gyr, although in this case the main focus of the work was to refute the activity/stability trade-off and to establish the catalytic speed of this enzyme as an evolutionary driver [65].

Despite titin has been studied for decades, there is yet much to be explored regarding the correlation between muscle physiological diversity in animals and the biochemistry and nanomechanics of titin. A series of phylogenetic studies have been carried out with giant titin-like proteins [66-69] (**Fig. 1.7**), but never related to its nanomechanical properties. This kind of proteins is present in bilaterian metazoans, deuterostome echinoderms, hemichordates, and chordates [70]. They have a common Ig-FnIII repeat structure and most of them share a kinase domain near the C terminus. However, their phylogenetic analysis is complicated due to a huge isoform variety, their low sequence identity and the non-standardized nomenclature [71]. Some studies [66, 69] hypothesize that duplication of Ig and FnIII domains took place with the appearance of striated muscle. They determine the age of this duplication in the common ancestor of nematodes and vertebrates (~800 Myr according to *Time Tree of Life* [54]).

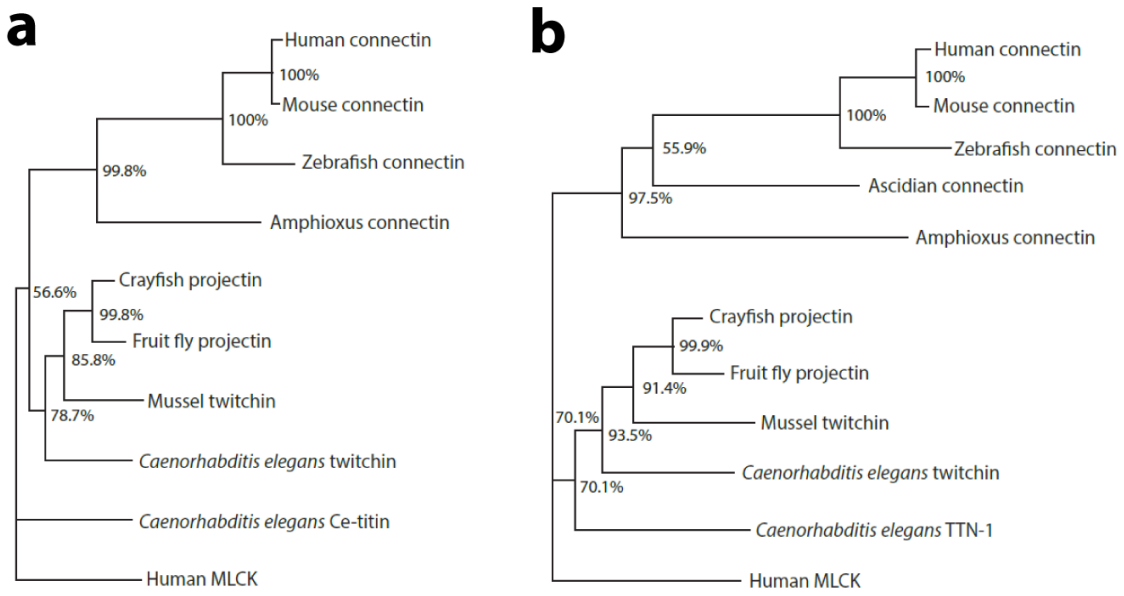


Figure 1.7. *Phylogenetic analyses of giant sarcomeric titin-like proteins carried out by (**a**) Ohtsuka et al. [67] and (**b**) Hanashima et al. [68]. Tree is not solved for early ancestors in both cases. More data is necessary to avoid polytomies.*

Thus, with regards to all these considerations, and given the morphological and locomotor diversity in vertebrates, titin may hold the key for some phenotypes displayed by animals in terms of muscle physiology. In this respect, it can be hypothesized that the evolution of titin has been central to the acquisition of muscle diversity in animals. For instance, the role of titin in evolution after the Cambrian explosion 542 Myr [72], remains unexplored. The so-called Cambrian radiation was a relatively brief period in terms of geological time (20-25 Myr), where most of the main animal phyla appeared, increasing the rate of diversification by one order of magnitude. Hence, it is reasonable to think that titin could have experienced major changes on its structure during this huge physiological outbreak.

Figure 1.8. *Schematic smFS representation for I65-I72 titin constructs. Force extension and force clamp experiments were carried to unravel the properties of the different titins.*

In this thesis, a combination of ancestral sequence reconstruction, smFS and biochemical techniques has been employed for the first time to investigate the evolution of the mechanical and biochemical properties of titin. Phylogeny has been used to reconstruct the I65-I72 titin fragment from different extinct species, including the last common ancestors of tetrapods, sauropsids, mammals, and placental mammals. This titin fragment is relevant because is a part of both the cardiac and the skeletal isoforms of titin. These four ancestral titins and five of their modern counterparts (human, zebra finch, orca, chicken and rat) were expressed in the laboratory and their mechanochemical properties measured and compared using smFS (**Fig. 1.8**) and biochemical assays. The different evolutionary lineages for birds and mammals were analyzed, showing very different behaviors. A special emphasis was given to the study and correlation of disulfide bonds in these domains, since it was recently proposed that they could regulate the unfolding forces in Ig domains. Results displayed diverse disulfide percentages depending on the geological age and the lineage of the specie. Altogether, this thesis aims to unravel the paleomechanical history of titin, since the first vertebrates to modern animals, in terms of mechanochemical stability.

Part II

Chapter 2: Phylogenetic methods

2.1. Background

Ancestral reconstruction is the generic name of the mathematical and statistical methods used to infer ancient information (in the form of strings of characters) from current data. Although it has some non-biological applications such as the phylogenies of the phonemes and vocabulary of ancient languages [73], oral traditions of extinct cultures [74] or ancestral marriage practices [75], it has been massively used in the field of phylogenetics. In this context, the strings of characters are either protein or nucleic acid sequences and the current data comes from the extant species that have been sequenced. Phylogenetics is the study and correlation of the evolutionary relationships between extant individuals, species and populations and their corresponding ancestors. Nowadays it is possible to reconstruct ancestral biological macromolecules; polynucleotide sequences of DNA and distinct types of RNA, or amino acid sequences of proteins. This is the so-called ancestral sequence reconstruction.

The first steps of reconstructing ancestors from measurable biological features of extant individuals or species are related to cladistics. In cladistics, the organisms are classified based on the common characteristics that they share. These common characteristics can be traced to a group's most recent common ancestor. Cladistics is known to be one of the precursors of modern phylogenetics. Cladistic methods appeared for the first time in the very beginning of 20th century. Specifically, Peter Chalmers Mitchell was

the first person who carried out a cladistic analysis for birds in 1901 [76, 77], followed by the works of Robert John Tillyard for insects (1921) [78] and Walter Max Zimmermann for plants (1943) [76].

The pioneering works of ancestral sequence reconstruction are credited to Emile Zuckerkandl and Linus Pauling in 1963. Under the historical context of the emergence of techniques for sequencing the primary structure of proteins started by Frederick Sanger in 1955 [17, 79], Zuckerkandl and Pauling proposed that, based on the amino acid sequence of extant proteins, it is possible to infer the phylogeny of that protein and the sequences of all the common ancestors, including the earliest point of the tree, the root [80, 81]. However, it wasn't until 1971 when Walter M. Fitch developed the first algorithm for ancestral sequence reconstruction using the principles of maximum parsimony [82]. Maximum parsimony is a non-parametric statistical method to infer a phylogenetics tree from a set of extant sequences minimizing the amount of mutations that are necessary to match the available data. David Sankoff optimized this algorithm in 1975 by adding a cost to the mutations [83]. This statistical work gave birth to the first phylogenetics program [84], called PAUP, in 1989. Developed by David L. Swofford, PAUP became very popular in the phylogenetics community.

Despite widespread usage, the limitations of maximum parsimony became evident. For instance, Fitch's approach overestimates the amount of rare changes [85]. Simultaneously, the exponential increase of the computing power (i.e. Moore's law) permitted the implementation of much more complex algorithms, such as maximum likelihood approaches [86-88] or Bayesian methods [89-93]. Briefly, maximum likelihood is a parametric approach where the algorithm looks for the most probable tree when the phylogenetics model and the extant sequences are introduced. In the Bayesian approach, the computer program searches for the highest posterior probability, which is determined by both the likelihood of the data under a certain evolutionary model and by a set of prior probabilities set for the trees. Nowadays almost all the procedures for ancestral sequence reconstruction are based in maximum parsimony, maximum likelihood and Bayesian inference. In this thesis, maximum parsimony and Bayesian inference have been used for computing the phylogenetic tree, whereas maximum likelihood has been chosen to infer the extant sequences.

The last two decades the range of applications of ancestral reconstruction has increased exponentially. Some of the most promising in the field of molecular evolution are the optimization of the fluorescence performance of opsins [94] and GFP proteins [95], novel anticancer drug's mechanism and design [96], the uric acid and evolution in mammals [97], the amino acid persistence in proteins [98] or mammalian diving capacity evolution [99]. Other relevant fields of application include calculating spatial migration traits in order to infer the location of the ancestors [100], inferring ancestral ranges of species from phylogenetic trees in order to obtain historical biogeographic ranges [101] and genome rearrangements [102].

One of the research lines of the Nanobiomechanics group in CIC nanoGUNE is focused on the ancestral reconstruction of proteins. Several proteins have been reconstructed since it was stablished in 2013. Specifically, we have resurrected not only structural proteins like titin, but also many enzymes such as different cellulases, thioredoxins and laccases for their further nanomechanical and biochemical analysis.

2.2. Theory

Every endeavor on reconstructing ancestors starts with a phylogeny, a hypothetical tree that encompasses the order in how species are correlated between each other by descent from common ancestors, starting with the last universal common ancestor (LUCA). In a phylogenetic tree, terminal nodes correspond to the extant species. These nodes are successively connected to their common ancestors by branches. The common ancestors are the inner nodes. At the end, all the species and, thus, all the evolutionary lines converge in the LUCA (**Fig. 2.1**).

.

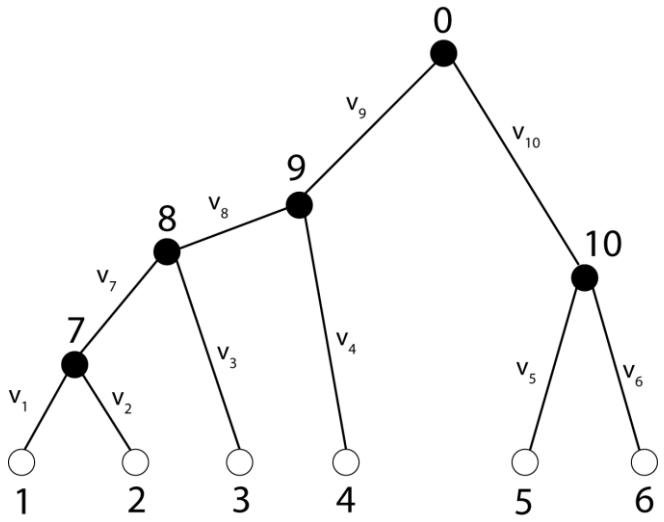


Figure 2.1. *Schematic example of a phylogeny. White (1-6) and black (6-10) circles refer to extant and ancestral species, respectively. The upper circle (0) is the last universal common ancestor (LUCA). Nodes are connected by branches, named* v_n.

2.2.1. Pioneering methods: Maximum parsimony

Parsimony, so-called *Occam's razor* [103], is the principle of choosing the simplest hypothesis and it was one of the first methods used in phylogenetics. Applied to phylogeny, it refers to looking for the distribution of ancestral states with the minimum number of mutations to explain the changes observed at the leaves of the tree. Therefore, the optimality criterion [84, 104] correspond to the total tree length L_{tree}:

$$L_{tree} = \sum_{k=1}^{B} \sum_{j=1}^{N} w_j \, diff\left(x_{k'j} x_{k''j}\right), \qquad (2.1)$$

where B is the number of branches, N is number of nucleotide or amino acid sites, k' and k'' are the two nodes connected by the branch k, and $x_{k'j}$ and $x_{k''j}$ correspond to the nucleotides or the amino acids of the observed in extant species or the inferred ones in the ancestral nodes. The cost of mutation between two sites is represented by the function *diff (y,z),* whereas w_j weights each of the sites.

Maximum parsimony is a very useful method due to its low computational costs and high efficiency for huge datasets and when *ab initio* phylogenies are needed [105] in

order to optimize more complex algorithms. However, it has a series of limitations [106, 107]. Some of these are:

- Fast evolution. The hypothesis of "minimum changes" that guides maximum parsimony methods implicitly assumes that mutations are rare. This assumption is not correct in cases of rapid evolution, such as some retroviruses [108-110].
- Change in rates of evolution. Maximum parsimony assumes that mutations between all sites have the same probabilities to take place. Therefore, any mutation has the same cost w_j for a certain tree. This hypothesis can limit the precision of the algorithm, but it can be partially fixed by weighting specific mutations, creating a weighted parsimony algorithm [83].
- Variation in time among evolutionary lines. Maximum parsimony algorithm takes the assumption that every branch in the tree has the same evolutionary time. Therefore, it doesn't take into account the different branch lengths of the tree, distorting the mutations rates of some nodes. [107, 111]. This limitation can be solved with more complex model-based methods, such as maximum likelihood or Bayesian inference [112]. This method applies the stochastic process of evolution for every branch of the tree.

2.2.2. State of the art methods

The following section describes briefly some of stochastic and phylogenetic concepts necessary to understand state-of-the-art phylogenetic methods:

Estimating mutations with Markov models

Phylogenies can be inferred with *cladistic* methods, such us the non-parametric maximum parsimony or the parametric maximum likelihood, but also with *phenetic* approaches, that build a tree considering a matrix of pairwise distances for the studied sequences. This last approach is often used to show phylogenetic relationships due to its low computational costs, but it can be only treated as an approximation because it lacks an

evolutionary model. Thus, precise models for ancestral sequence reconstruction require a proper evolutionary model.

Hence, for a stochastic DNA model (that is chosen for the sake of simplicity) the probability for each site is $p_{ij}(t)$, where the nucleotide $i\epsilon\{A, C, G, T\}$ will mutate to nucleotide j in time t. Consequently, a Markov chain with a state space $S_{DNA}\epsilon\{A, C, G, T\}$ with a random variable $X(t)\epsilon S_{DNA}$ defines the substitution process. For a homogeneous Markov process, we assume that $p(X(s + t) = j|X(s) = i)$, i.e. the probability for a nucleotide replacement i with j in the time t, is independent with the actual time point $s \geq 0$. If we define a constant rate of mutation as μ per unit time and a constant prior probability π_j for a mutation leading to nucleotide j, then we can assume that the probability for no mutations at the considered site after t is $(1-\mu)^t$.

$$p_{mut} = 1 - (1 - \mu)^t \approx 1 - e^{-\mu t} \qquad (2.2)$$

Moreover, the probability can be also displayed with equations for continuous time instead of using discrete generations:

$$p(t + dt)p(t) + p(t)Qdt = p(t)(I + Qdt), \qquad (2.3)$$

where Q is the rate matrix of transition probabilities including the terms for the individual transitions and I is the unit matrix. This equation can be rewritten as:

$$p(t) = e^{tQ} \qquad (2.4)$$

Substitution models

In the case of DNA, sixteen π_j are necessary to consider all the possible mutations and many Q matrices have been developed. In the case of proteins, it is also possible to study the mutational events in the level of amino acids. Early amino acid substitution models are assigned to PAM-matrices [113] developed by M. Dayhoff, but because of the few number of sequences available at that time, they are rough approximations. Thanks to the boom of the genomics in the last two decades, more complex matrices have been established more recently, such the JTT-matrix [114] or the WAG-matrix [115]. At the same time, homogeneous models were discarded for more complex ones with continuous

distributions that provide a specific rate for every site. Most of the times a gamma distribution representing a whole family of probability distributions is used [116]. The shape of this distribution depends of the parameters α and β. Nevertheless, it has been proved that a discrete gamma model performs well and is computationally efficient [86]. This model consists of a certain number of equally categories of rates (normally 4 to 8) that are selected to approximate the gamma distribution. So, the density of the gamma distribution *g(α, β)* is given by:

$$g(r;\ \alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \exp(-\beta r)\, r^{\alpha-1},\ \ 0 < r < \infty \qquad (2.5)$$

Here, *α* is given or calculated and the scale parameter *β* is redundant and can be fixed equal to *α*. Then, the range of *r (0, ∞)* is split into *k* categories by cutting points and every category is characterized by a rate r_i (the mean of the portion of the gamma distribution in that specific category). Hence, the probability *p(x)* for an observed symbol *x* at a site is related to the rate specific conditional probabilities:

$$p(x) = \int p(x|r)\, g(r) dr \approx \sum_{i=1}^{k} \frac{1}{k} p(x|r = r_i), \qquad (2.6)$$

where *g(r)* refers to the gamma density with *α*, which is selected so that $r_1, \ldots, r_k$ obtain the largest approximate likelihood and *p(x|r)* is the conditional probability of *x* at a rate *r* for a given site. Thus, using one of these models of evolution, it is possible to compute the likelihood of a tree.

2.2.2.1. Maximum likelihood

Phylogenetic methods based on maximum likelihood consider the sites of the internal nodes of the tree as parameters and try to find the values of these parameters that maximize the probability of the data (the extant sequences) for a given evolution model. In other words, a time-reversible Markov process models the evolution of the sequence, assuming that all the mutations are independent [117]. The likelihood of the phylogeny is calculated from a sum of intermediate probabilities of the nodes for the proposed tree. Thus, in every ancestral node, the likelihood of the descendants is computed to obtain a maximum posterior probability:

$$L_x = \sum_{S_x \in \Omega} P(S_x) \left(\sum_{S_y \in \Omega} P(S_y \,|S_x, t_{xy}) L_y \sum_{S_z \in \Omega} P(S_z \,|S_x, t_{xz}) L_z \right), \qquad (2.7)$$

where the node x is the ancestor of y and z. S_i represents the sequence of the i-th node, t_{ij} refers to the branch length from i to j. Ω is the set of all the possible combinations (for instance, the four nucleotides or the 20 basic amino acids). The ultimate objective of the reconstruction is to search for the best configuration in all the previous nodes to obtain the maximum likelihood for their ancestor in a given tree.

At the end of the process, each descendant obtains a likelihood value. In order to find the most probable evolutionary lineage to the common ancestor, two different conventions have been proposed. First, one can consider the probabilities of all the descendants for a certain ancestor and calculate the joint combination with the maximum likelihood. This approach is called joint reconstruction. And second, instead of calculating the global likelihood, one can successively select the most likely ancestor for every node. This procedure is referred to as marginal reconstruction (**Fig. 2.2**). The first software application for this approach was implemented in 2007 under the name PAML [93, 118].

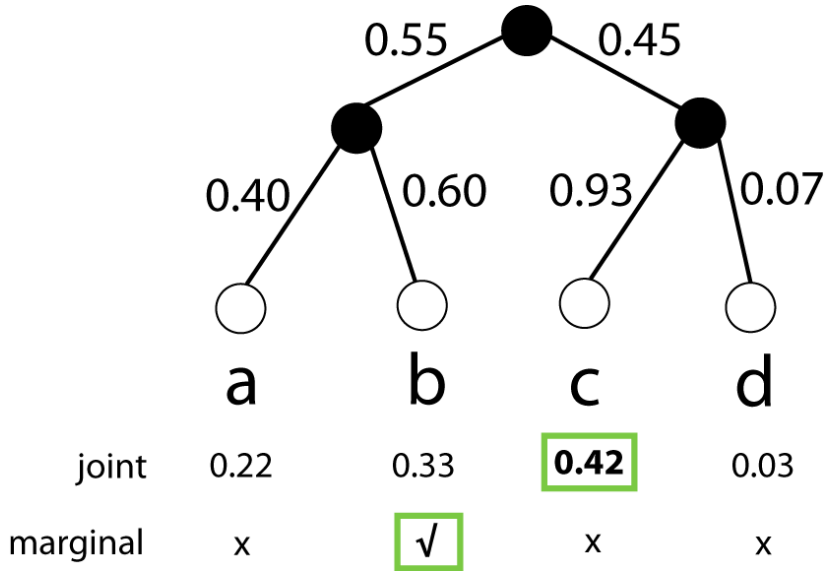


Figure 2.2. *Joint and marginal reconstructions for a given phylogeny.*

2.2.2.2. Bayesian inference

Bayesian inference employs both the likelihood of the experimental data, described before, and a prior knowledge about the possible solutions. Thus, the aim in ancestral sequence reconstruction is to obtain the posterior probabilities for every internal node of a known tree. On top of that, the posterior probabilities could be combined with the posterior distributions over the parameters for a given evolutionary model and the structure of all possible trees, giving the following applications of Bayes' theorem:

$$P(S \mid D, \theta) = \frac{P(D \mid S, \theta)\, P(S \mid \theta)}{P(D \mid \theta)} \tag{2.8}$$

$$\propto\ P(D \mid S, \theta) P(S \mid \theta)\, P(\theta), \tag{2.9}$$

where D is the experimental data, S corresponds to the ancestral states and θ represents the phylogenetic tree and the evolutionary model. In equation 2.8, $P(D \mid S, \theta)$ is the likelihood of the experimental data that could be computed, $P(S \mid \theta)$ refers to the prior probability of an ancestral node for a known tree and model and $P(D \mid \theta)$ corresponds to the probability of the data for a known tree and model, integrated for all possible ancestral states. Note that two different formulations have been given (2.8 and 2.9), one for each of the applications of Bayesian inference, the empirical and the hierarchical Bayes. Empirical Bayes approach estimates the probabilities of several ancestral nodes for a given tree and model of evolution. On the other hand, hierarchical Bayes approach calculates these probabilities over all possible trees and model of evolution, comparing how likely they are, with a given experimental data [119].

2.3. Methodology

Generally, the process to obtain an ancestral sequence consists of four well defined steps (**Fig. 2.3**): Select the extant species (**1**), create a multiple alignment (**2**), construct a phylogenetic tree (**3**), and finally, reconstruct the ancestral sequences (**4**). For the sake of simplicity and, as this thesis is related to protein reconstruction, we will focus on just ancestral protein reconstruction in this chapter.

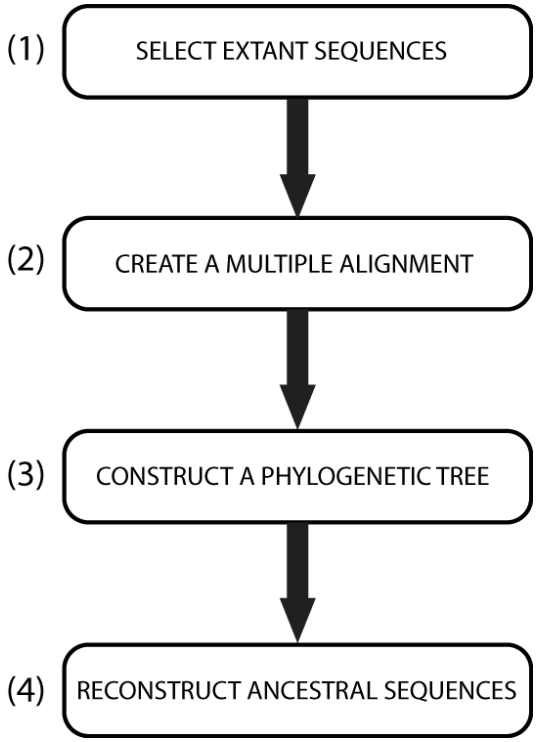


Figure 2.3. *Typical methodology used for the reconstruction of ancestral sequences consisting in four steps: **(1)** Selection of extant sequences, **(2)** creation of a multiple alignment, **(3)** construction of a phylogeny, and **(4)** reconstruction of ancestral sequences.*

2.3.1. Selection of extant sequences

The first step to reconstruct a phylogenetic tree is to find homologous sequences of the protein of interest. Homologous sequences mean that two sequences are descendant from the same common ancestor. Thus, the identical residues at a site are identical by character state in these sequences. In general, homologue sequences are retrieved from online databases using BLAST (Basic Local Alignment Search Tool) [120]. This tool searches for regions of local similarity between sequences that can be used later to infer evolutionary relationships. The most used databases are: **UniProtKB** of the EBI (European Bioinformatics Institute) and **GenomeNet** of the Kyoto University.

2.3.1.1. UniProtKB

UniProtKB (Universal Protein Resource Knowledge Base) [121] is a catalog of information on proteins. To find the protein of interest (for now on, **query**), one can either use the search tool (**Fig. 2.4**) or directly enter the protein sequence, or its UniProt identifier (i.e. Q8WZ42 for human titin).

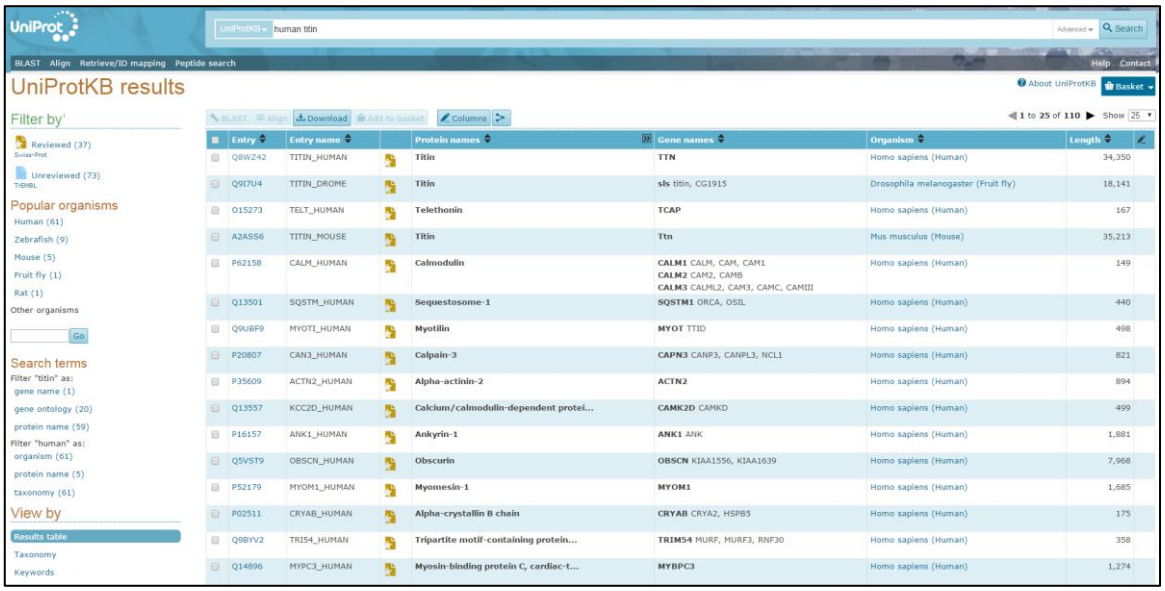


***Figure 2.4**. Search tool in UniProtKB database.*

Once the query is identified and selected the BLAST tool (**Fig. 2.5**) should be used to find homologous sequences. UniProtKB allows setting some of the parameters of this tool:

- Target database: is the database against which the search is performed. One can choose UniProtKB for different phylum or clusters of sequences with 100%, 90% or 50% identity.
- E-Threshold: The expectation value (E) threshold is a statistical measure of the number of expected matches in a random database. The smaller the e-value is, the more likely the match is to be significant.

- Matrix: The matrix assigns a probability score for each position in an alignment. The BLOSUM [122] matrix assigns a probability score for each position in an alignment that is based on the frequency with which that substitution is known to occur among consensus blocks within related proteins.
- Filtering: One can filter by lower complexity regions or by masking the lookup table only.
- Gapped: allows gaps to be introduced once the sequences are selected.
- Hits: allows to choose the number of hits in the search.

Figure 2.5. *BLAST tool in UniProtKB database.*

Once all the parameters are selected, "Run BLAST" should be pressed in order to obtain the homologous sequences of the query. This process takes up to several minutes depending on the complexity of the query sequence and the applied parameters. After the process is completed, homologous sequences will be shown in order of identity with the query (**Fig. 2.6**). One can select the proteins of interest and download them in different formats. It is convenient to download the sequences in FASTA format (a text based format for representing nucleotide or amino acid sequences, where each of them is represented by a single letter) to facilitate the use of the sequences during the following steps of the process.

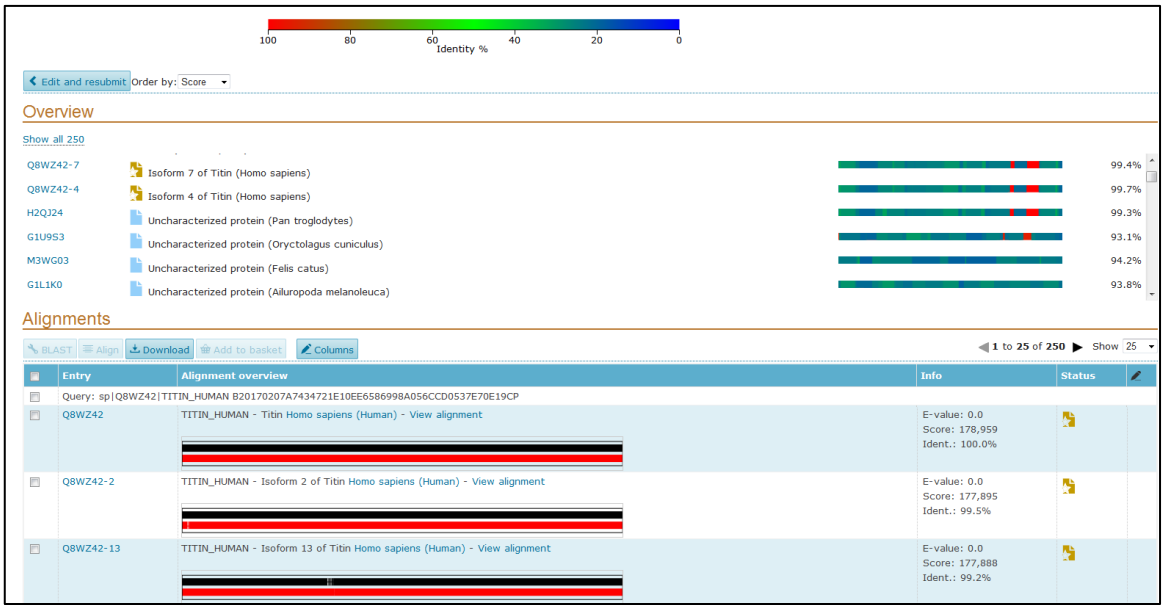


Figure 2.6. *List of homologous sequences after BLAST search.*

2.3.1.2. GenomeNet

GenomeNet [123] is a bioinformatics online platform that offers numerous services. To use the BLAST application for finding homologous sequences of the protein of interest, one should select "BLAST" in the homepage. The BLAST tool (**Fig. 2.7**) of this website offers search services for nucleotides and amino acids. To perform a protein query search against a protein database, BLASTP option should be selected. Here, the protein query could be found by introducing the sequence ID or the sequence itself, or by uploading a file with the sequence (normally in FASTA format). GenomeNet also provides of several options to modify the parameters of the search:

- KEGG GENES: is the option to delimit the search to eukaryotes, prokaryotes, viruses, or a certain organism.
- KEGG MGENES: is the option to delimit the search to environmental or organismal sequences. One can also choose a favorite group of samples.
- Microbial Reference Genes: searches only in microbes from ocean or the human gut.
- nr-aa: allow to discriminate the search against a certain database.
- Scoring matrix: One can choose different matrixes for the search.

- Maximum number of database sequences to be reported.
- Maximum number of alignments to be displayed

Again, once the process is completed the homologous sequences will appear in a new window ordered by identity. Sequences of interest can be selected and downloaded in several formats.

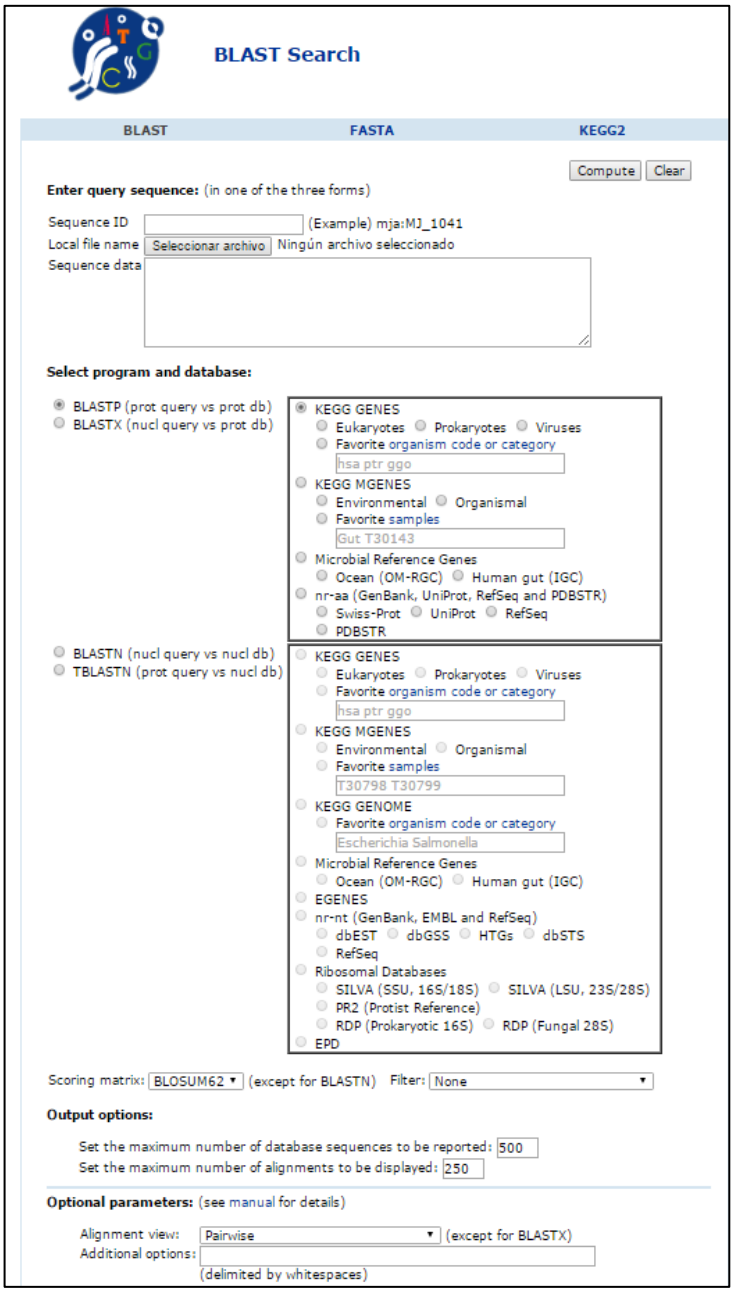


Figure 2.7. *BLAST tool in GenomeNet database.*

2.3.2. Creation of a multiple alignment

After the homologous sequences are obtained they must be aligned. Nowadays, heuristic processes are the only approaches to compute multiple alignments due to the complexity of the algorithms, where all the residues must be mapped to protein positions. Over the last years, many algorithms were developed for this this purpose, being Clustal (with Clustal X [53] being the version with a graphical interface and Clustal Ω [124, 125] the current standard method) and MUSCLE [126] the most popular in the phylogenetic community. In this thesis, MUSCLE algorithm (integrated in the MEGA software) has been used for all the multiple sequence alignments. MEGA6 [127, 128] is a user-friendly software for creating multiple sequence alignment, inferring phylogenetic trees, estimating rates of molecular evolution, inferring ancestral sequences, and several more phylogeny related applications. To perform the sequence alignment of the pool of selected sequences, first one must load the sequences (in general in FASTA format). The sequences will appear unaligned in the interface of the program (**Fig. 2.8**)

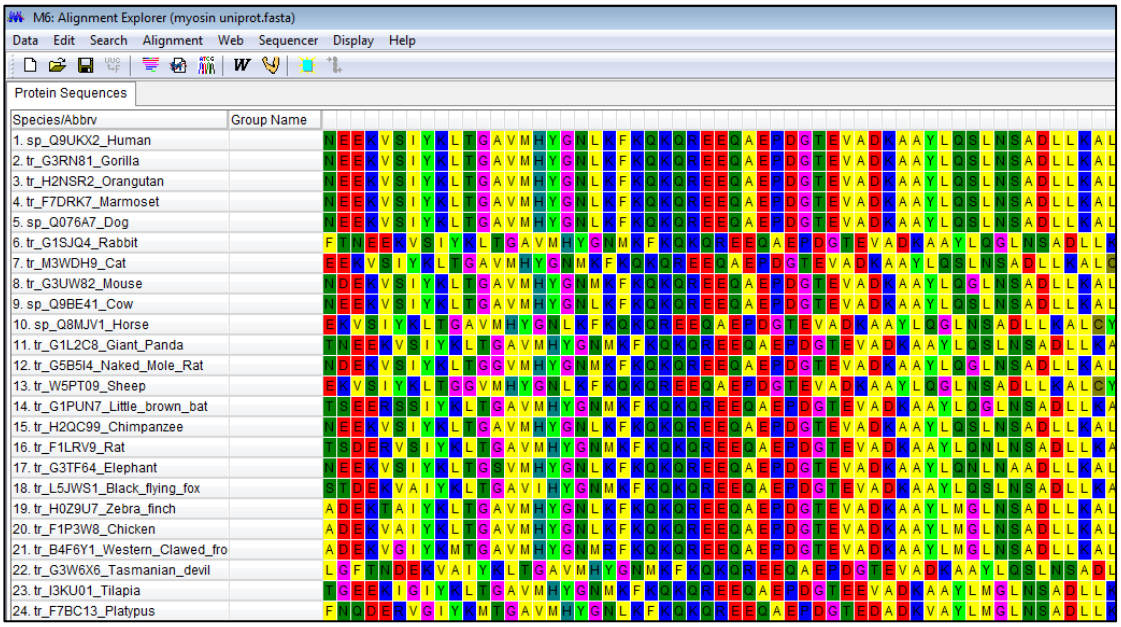


Figure 2.8. *Unaligned pool of sequences in MEGA 6.*

To run the MUSCLE algorithm, one should first select all the sequences and then click on the “align with MUSCLE” icon, which is an arm in the top-left of the interface. A

pop-up window will appear showing the different parameters that one can change before computing the algorithm (**Fig. 2.9**). The parameters are the following:

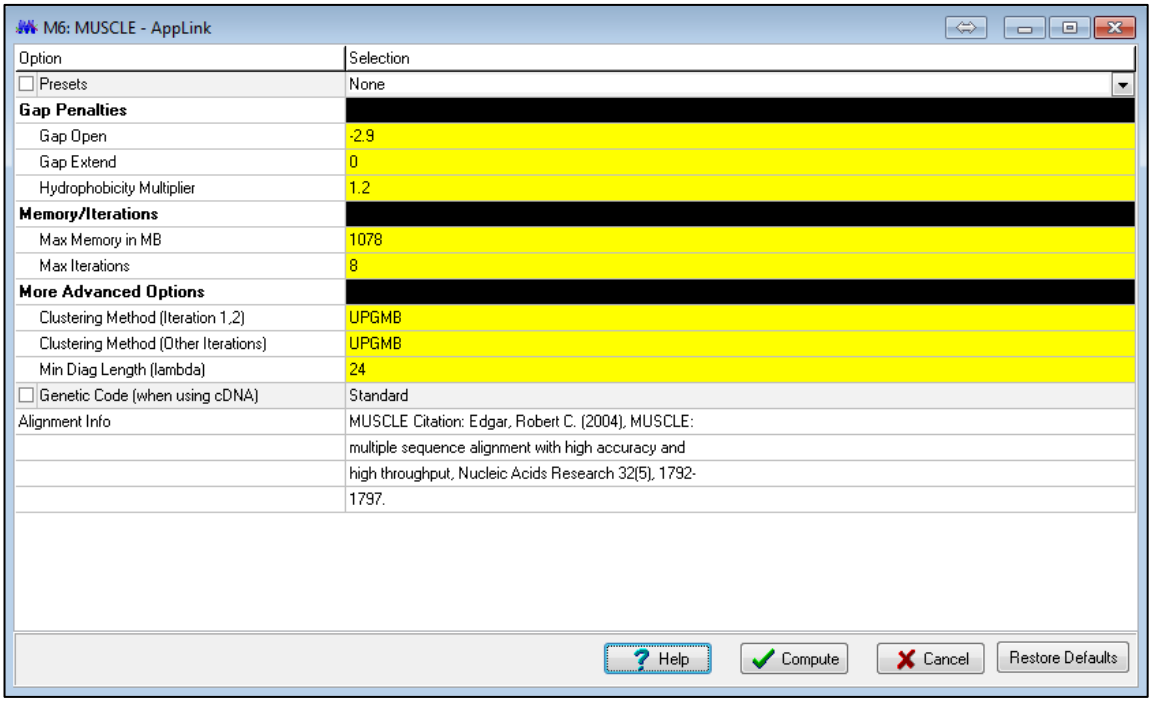


Figure 2.9. *MUSCLE algorithm parameters in MEGA 6 interface.*

- Gap Opening Penalty: By increasing this value the gaps are less frequent in the alignment.
- Gap Extension Penalty: Sets a penalty for extending a gap by one residue. By increasing this value, the gaps are shorter in the alignment. Terminal gaps don't penalize.
- Max Memory in MB: The algorithm sets a computational memory upper limit. It can use by default (in Megabytes) in order to avoid using all the computer resources.
- Max Iterations: Sets the maximum number of permitted iterations.
- Clustering Method (Iteration 1,2): Sets the clustering method used in the first two iterations.

- Cluster Method (Other Iterations): Sets the clustering method used in the following iterations
- Max Diagonal Length: Maximum length of the diagonal.

Once we set all the parameters, the "Compute" button must be selected and the program will start running the algorithm. Depending on how many iterations have been selected the process will last longer or shorter. Once all the iterations are finished, the interface will show the aligned pool of sequences (**Fig. 2.10**). Sometimes regions of ambiguous alignment or with gaps in several sequences can be removed manually or using GBLOCKS [129]. An asterisk in the top of the alignment means that the residue below the asterisk is conserved for all the sequences. The color of the residues is related to its biochemical properties.

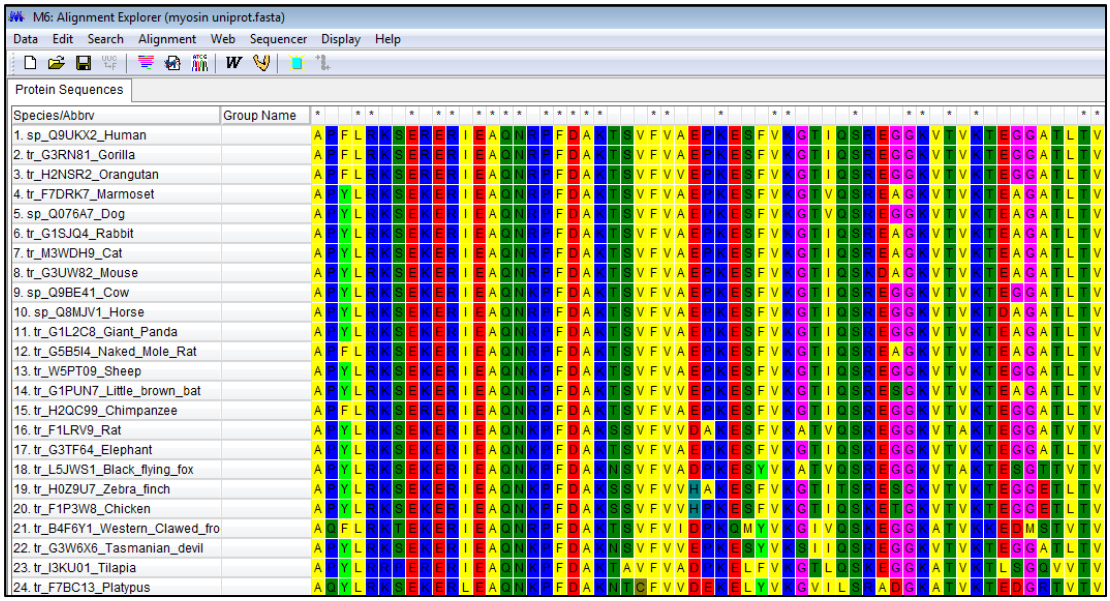


Figure 2.10. *Aligned pool of sequences in MEGA 6.*

2.3.3. Computing a phylogenetic tree

Once the sequences are properly aligned, the computing of the phylogenetic tree takes place. In order to achieve this, we used two different programs based on maximum parsimony (**PAUP**) and Bayesian inference (**BEAST**). The following sections explain how to run them.

2.3.3.1. PAUP

PAUP (Phylogenetic Analysis Using Parsimony) [130, 131] has been one of the most widely used software package for the inference of phylogenies over the last years. In this thesis PAUP 4.0 version has been used to compute phylogenetic trees. Although it can be run with commands, there is also a user-friendly interface (**Fig. 2.11**). Here, an alignment in NEXUS format (a file format widely used in bioinformatics encompassing taxa, data and trees) must be selected. The interface will show the number of taxa and the amount of amino acids in the alignment.

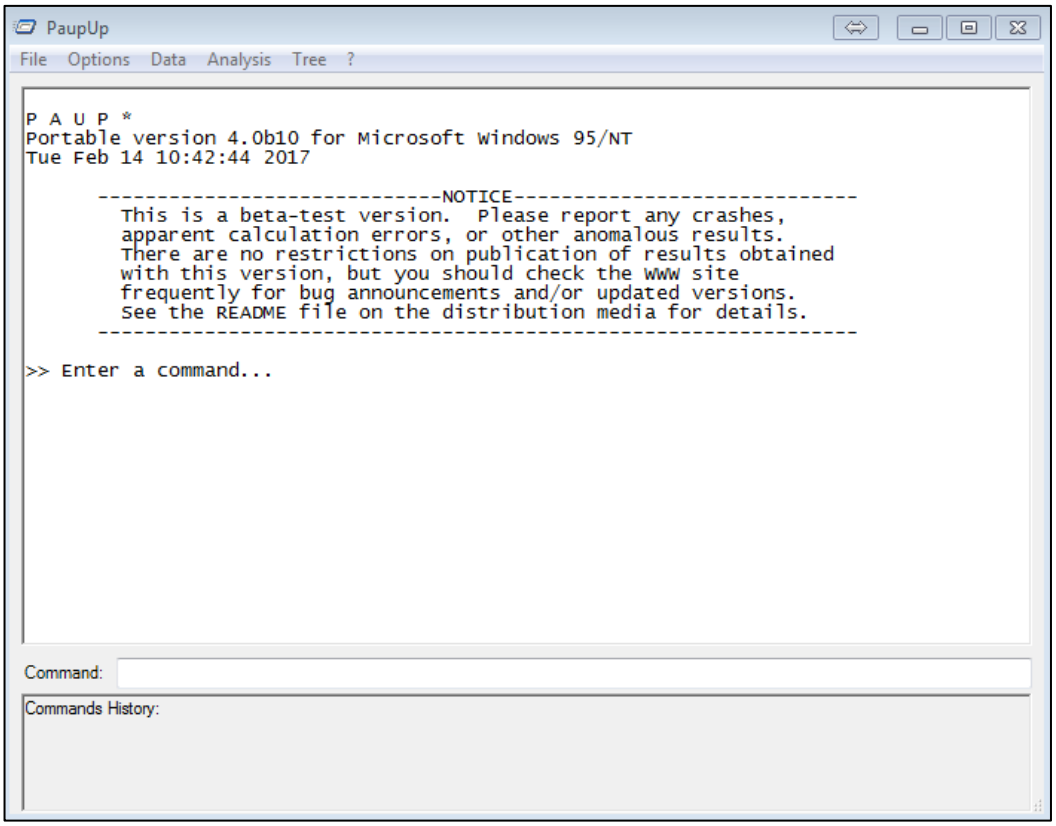


Figure 2.11. *PaupUp. PAUP 4.0 user-friendly interface.*

To compute the phylogeny with maximum parsimony algorithm, one should select the parsimony criterion in the "Analysis". In the same tab, there is an option called "Bootstrap/Jackknife" that allows to resample the data (**Fig. 2.12**). Once these parameters are selected, one can run the phylogeny.

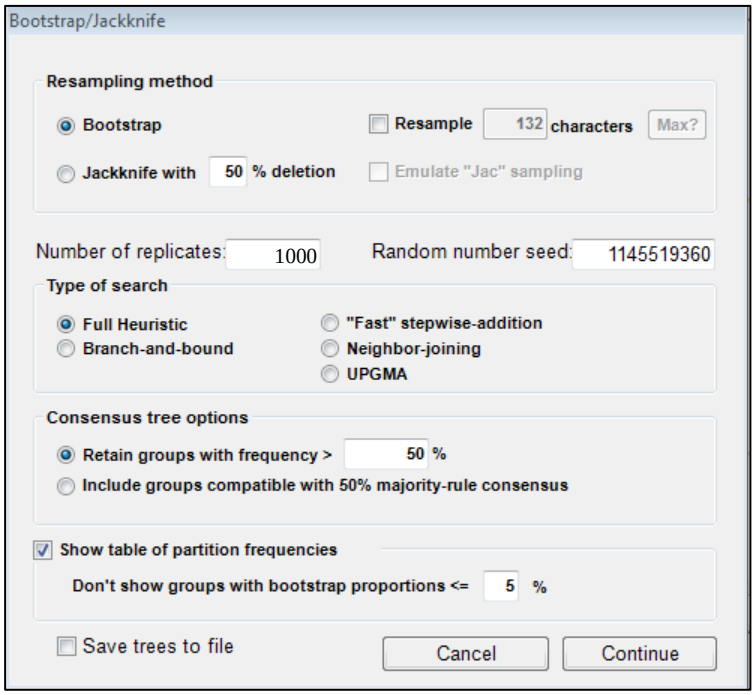


Figure 2.12. *Bootstrap/jackknife option in PAUP 4.0. The parameters of the algorithm can be selected here.*

Once the process is complete, the phylogenetic tree will appear in the interface (**Fig. 2.13**). The tree can be saved in NEXUS format.

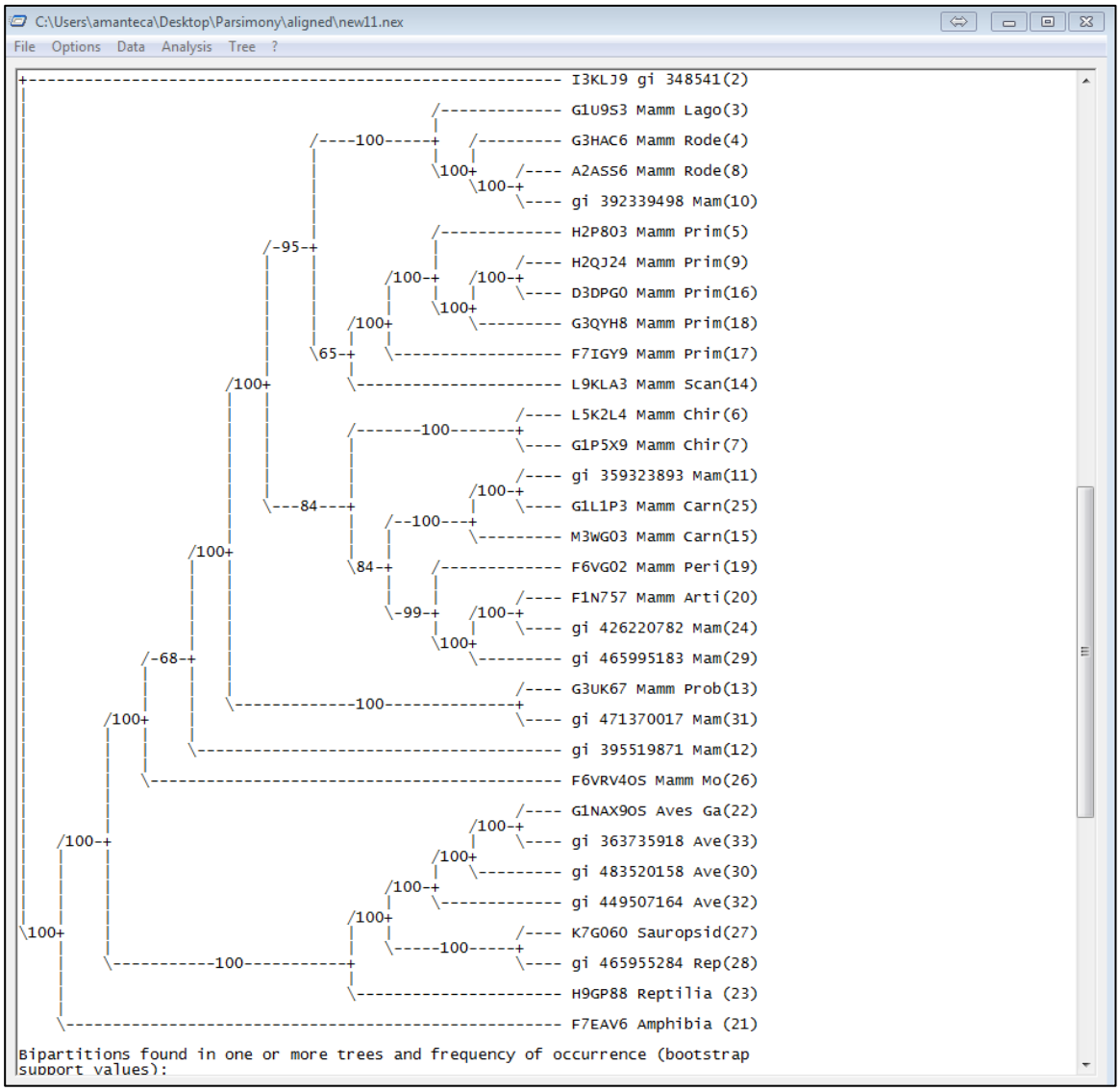


Figure 2.13. *Consensus tree of a phylogeny in PAUP 4.0 interface.*

2.3.2.2. BEAST

BEAST (Bayesian Evolutionary Analysis Sampling Trees) [132, 133] is a package of programs for Bayesian analysis of molecular sequences using Markov chain Monte Carlo (MCMC), a class of algorithm for sampling the probability distribution based on constructing a Markov chain. BEAST can be used for reconstructing phylogenies using MCMC to average over tree space, so that each tree is weighted proportional to its posterior probability. This section describes how to use the programs of the package used in this thesis: **BEAUti, BEAST, TreeAnotator, Tracer** and **FigTree**.

BEAUti

Bayesian Evolutionary Analysis Utility (BEAUti) is a graphical user interface that allows the creation of input files (.xml) to run BEAST. It permits to set the evolutionary model and options for the MCMC Once the aligned sequences are imported (through a NEXUS file), a user-friendly interface with several tabs allows modifying many parameters (**Fig. 2.14**):

- Partitions: allows to load sequences that were not in the initial pool and to make partitions within this pool.
- Taxa: clusters selected taxa into subgroups. There is the possibility to force these subgroups to be monophyletic.
- Tips: permits data sampling of individual taxa.
- Traits: Sets the phenotypic trait analysis.
- Sites: allows selecting the substitution model and the site heterogeneity model.
- Clocks: permits to choose the clock model. The different clock models use the mutation rate of biomolecules to estimate when they diverged.
- Trees: sets the tree prior
- States: allows reconstructing the states of all the ancestors or only certain subgroups.
- Priors: sets the prior distribution for the subgroups, the gamma shape parameter, the proportion of invariant sites parameter (a parameter that assumes which residues do not mutate) and the root height of the tree.
- Operators: Switches on or off some of the parameters set in previous tags.
- MCMC: sets the MCMC values for phylogeny computing.

Once all the parameters are selected, "Generate BEAST File…" button should be pressed to obtain the XML file for a further phylogeny computing using BEAST.

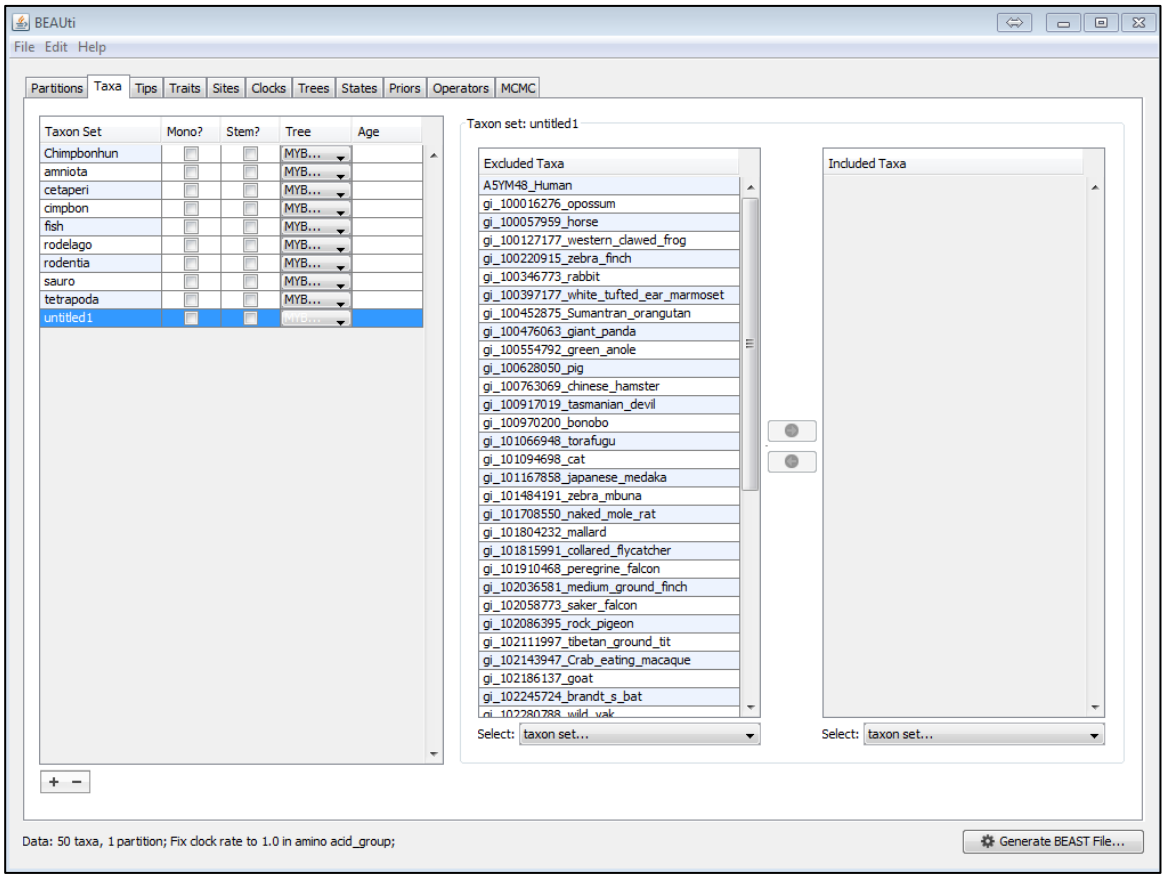


Figure 2.14. *Graphical user-interface (GUI) application of BEAUti.*

BEAST

BEAST uses as input an XML command file previously generated with BEAUti and returns a log file as output. The log file records a sample of the states that the Markov chain found. An examination of this output is needed to determine whether the Markov chain has been run for long enough to obtain accurate estimates of the parameters. This post-analysis is carried out by an application called **Tracer**.

To compute the phylogeny, one should open the BEAST graphical user-interface (**Fig. 2.15**). Here, one can open the XML file and run the program. The application also has the option to activate the **BEAGLE** library [134]. BEAGLE is a high-performance library that can take advantage of the parallel processors available in most of current PCs. Using this option is highly advisable to improve the performance of the program.

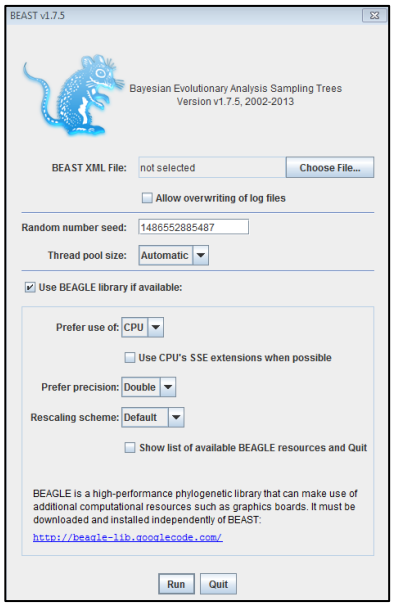


Figure 2.15. *Graphical user-interface (GUI) application of BEAST.*

Once the XML file of interest is selected and BEAGLE library option chosen, one can click on the "run" button and the phylogeny will start computing (**Fig. 2.16**). When the process is over, the program will return the log files that contain in the information of the process.

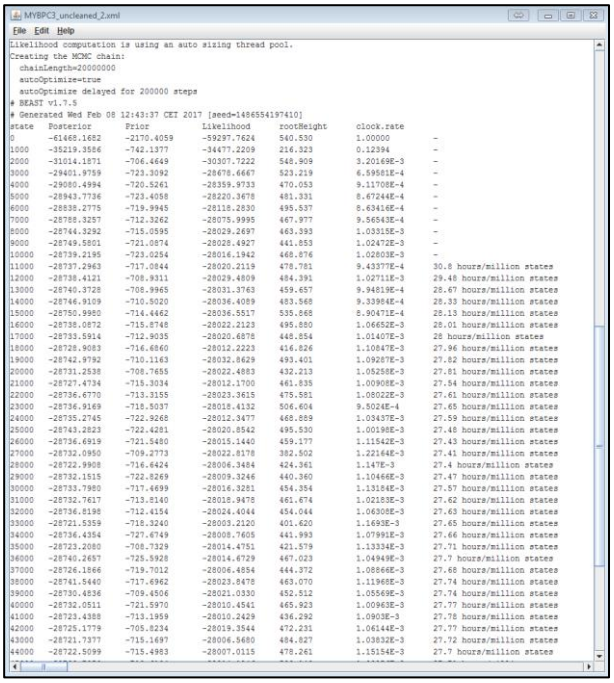


Figure 2.16. *BEAST application computing a phylogeny.*

We note that particularly in the case of very demanding trees like the ones computed in this thesis, one can also run the tree and set the parameters for the command-line.

Tracer

Tracer is graphical interface (**Fig. 2.17**) that allows to monitor and analyze the MCMC output carried out in BEAST. Once the log file that corresponds to the analysis of the phylogenetic computation is opened, the name of the log file lo the traces that it contains and many parameters related to MCMC analysis will appear on the left side of the interface. These parameters are pondered by their Effective Sample Sizes (ESSs). A low ESS means that the trace contained a lot of correlated samples and may not represent the posterior distribution well. It is advisable to run BEAST again until ESS reaches a value higher than 100.

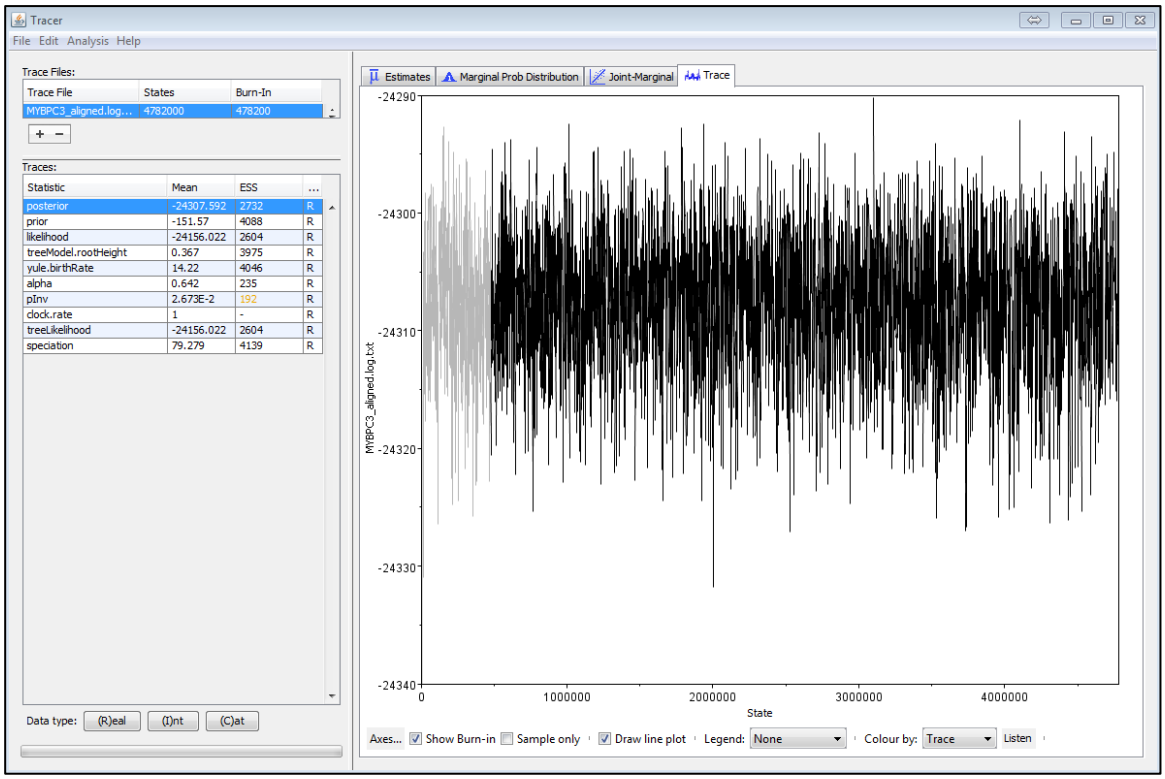


Figure 2.17. *Tracer graphical interface. The evolution of the posterior probability can be visualized in this picture. The data in grey is excluded for the calculation of averages.*

TreeAnotator

TreeAnotator summarizes the sample of trees generated by BEAST onto a single consensus tree. This tree contains information about the posterior probabilities of the nodes in the consensus tree, the posterior estimates and (in the case of a relaxed molecular clock model) the rates. The program (**Fig. 2.18**) includes a sort of options:

- Burnin: This option sets the percentage of samples that are discarded to analyze the part of the trace that is in equilibrium.
- Posterior probability limit: This option is the equivalent of setting a limit for the bootstrap in PAUP. Only the nodes in the target tree that have a posterior probability greater than the specified limit will be saved.
- Target tree type: two options can be chosen here. In "Maximum clade credibility" option the node height and rate statistics will be merged on the tree in the posterior sample with the maximum summation of posterior probabilities on its (n – 2) internal nodes. The "User target tree" option summarizes the tree statistics on a user-specified tree.
- Node heights: This option sets the way the node heights are summarized on the target tree.

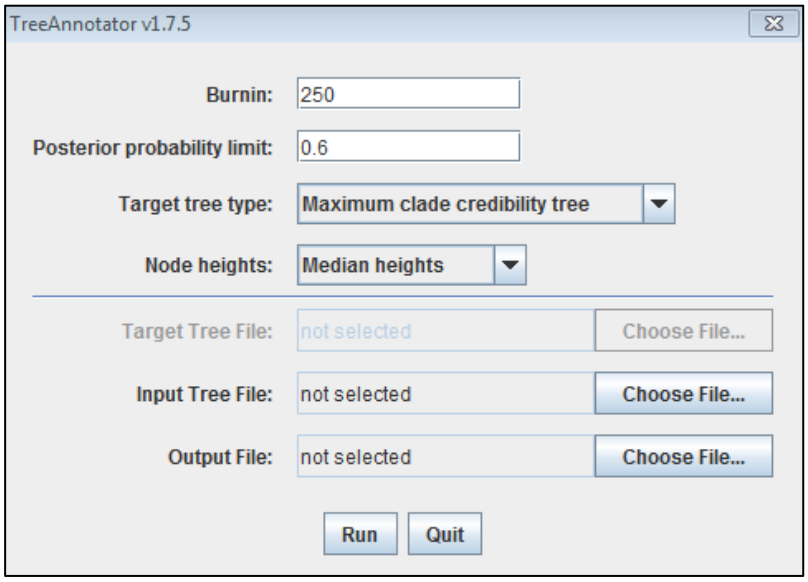


Figure 2.18. *TreeAnotator graphical interface and its options.*

FigTree

Finally, FigTree is a user-friendly interface to visualize and modify phylogenies. It permits to generate publication quality figures. Here, the tree file generated with TreeAnotator should be opened. A sort of different options will appear on the left side of the interface (**Fig. 2.19**) to change the appearance of the tree.

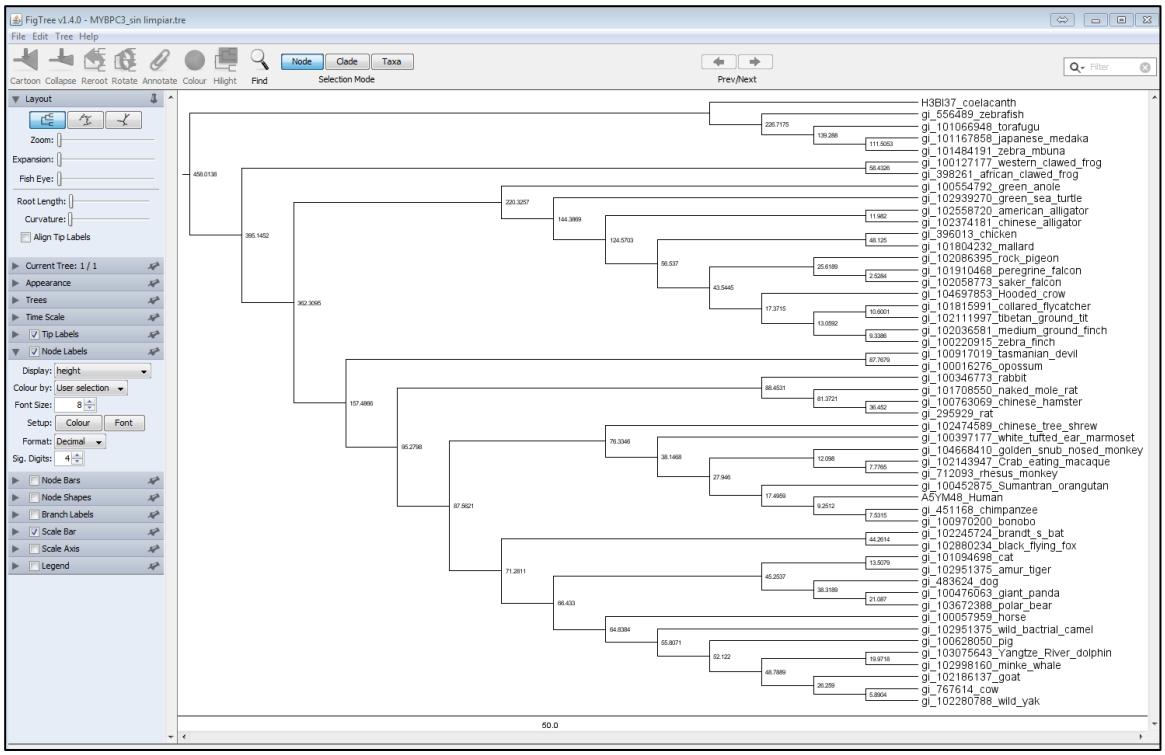


Figure 2.19. *FigTree graphical interface and its options. A consensus tree can be visualized in the figure.*

FigTree allows to change many parameters of the tree, such as the time scale, the node labels, the node bars, the branch labels, the scale bar, the scale axis or the legend. Once all these parameters are set, one can export the tree in a publication quality format (**Fig. 2.20**).

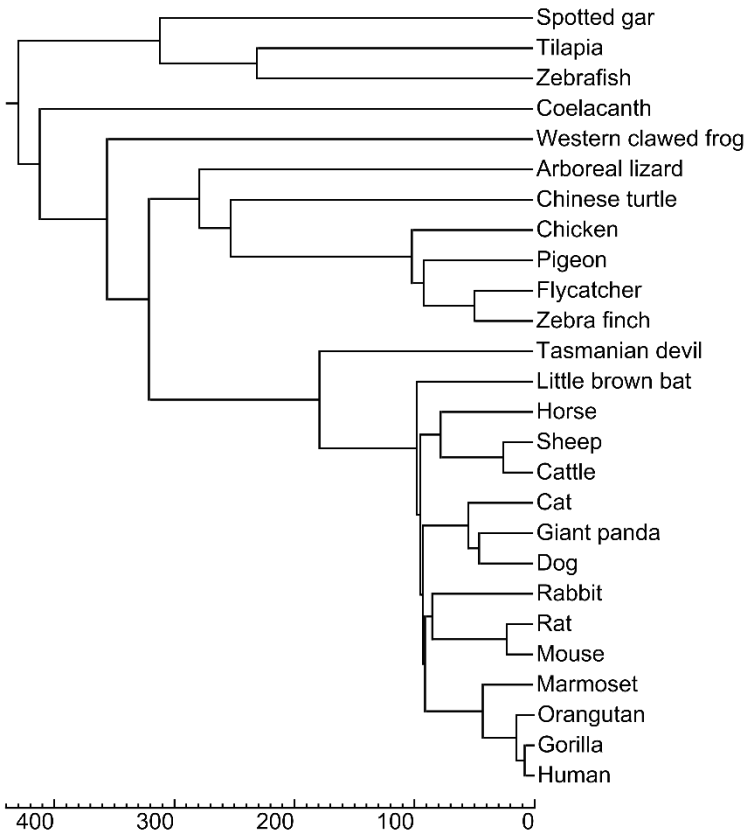


Figure 2.20. *Publication quality chronogram for myosin 2 muscle protein exported from FigTree.*

2.3.4. Reconstruction of ancestral sequences

Once the consensus tree is obtained and we are confident with the statistics, one can start inferring the ancestral sequences of interest. Although there are many algorithms and programs to achieve this purpose, in this thesis PAML has been used to infer all the ancestral sequences. PAML is based on the maximum likelihood algorithm mentioned in previous sections.

2.3.4.1. PAML

PAML (Phylogenetic Analysis Using Maximum Likelihood) [135, 136] is a package of programs for maximum likelihood analysis of protein and DNA sequences. The specific program used in this thesis is called **codeml** and it is used for the reconstruction of

ancestral codons and proteins). Running of codeml requires a sequence data file, a tree file (in Newick format), a matrix file (usually Jones matrix) and the control file before running the programs (**Fig. 2.21**).

codeml (2) - Notepad

File Edit Format View Help

```
      seqfile = act.aa * sequence data filename
     treefile = act.tre      * tree structure file name
      outfile = act_ef_jones           * main result file name

        noisy = 9  * 0,1,2,3,9: how much rubbish on the screen
      verbose = 0  * 0: concise; 1: detailed, 2: too much
      runmode = 0  * 0: user tree;  1: semi-automatic;  2: automatic
                   * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

      seqtype = 2  * 1:codons; 2:AAs; 3:codons-->AAs
    CodonFreq = 2  * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table

*       ndata = 10
        clock = 0  * 0:no clock, 1:clock; 2:local clock; 3:CombinedAnalysis
       aaDist = 0  * 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a
   aaRatefile = jones.dat  * only used for aa seqs with model=empirical(_F)
                   * dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or your own

        model = 3
                   * models for codons:
                       * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
                   * models for AAs or codon-translated AAs:
                       * 0:poisson, 1:proportional, 2:Empirical, 3:Empirical+F
                       * 6:FromCodon, 7:AAClasses, 8:REVaa_0, 9:REVaa(nr=189)

      NSsites = 0  * 0:one w;1:neutral;2:selection; 3:discrete;4:freqs;
                   * 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;
                   * 10:beta&gamma+1; 11:beta&normal>1; 12:0&2normal>1;
                   * 13:3normal>0

        icode = 0  * 0:universal code; 1:mammalian mt; 2-10:see below
        Mgene = 0
                   * codon: 0:rates, 1:separate; 2:diff pi, 3:diff kapa, 4:all diff
                   * AA: 0:rates, 1:separate

    fix_kappa = 0  * 1: kappa fixed, 0: kappa to be estimated
        kappa = 2  * initial or fixed kappa
    fix_omega = 0  * 1: omega or omega_1 fixed, 0: estimate
        omega = .4 * initial or fixed omega, for codons or codon-based AAs

    fix_alpha = 1  * 0: estimate gamma shape parameter; 1: fix it at alpha
        alpha = 0. * initial or fixed alpha, 0:infinity (constant rate)
       Malpha = 0  * different alphas for genes
        ncatG = 8  * # of categories in dG of NSsites models

        getSE = 0  * 0: don't want them, 1: want S.E.s of estimates
 RateAncestor = 1  * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)

   Small_Diff = .5e-6
    cleandata = 0  * remove sites with ambiguity data (1:yes, 0:no)?
*  fix_blength = -1  * 0: ignore, -1: random, 1: initial, 2: fixed
       method = 0  * Optimization method 0: simultaneous; 1: one branch a time

* Genetic codes: 0:universal, 1:mammalian mt., 2:yeast mt., 3:mold mt.,
* 4: invertebrate mt., 5: ciliate nuclear, 6: echinoderm mt.,
* 7: euplotid mt., 8: alternative yeast nu. 9: ascidian mt.,
* 10: blepharisma nu.
* These codes correspond to transl_table 1 to 11 of GENEBANK.
```

Figure 2.21. *Control file of codeml program in PAML.*

In the control file, one can set and control many parameters that are explained in the file itself. Once everything is set up, one can run the executable file and program will start computing the algorithm to calculate the ancestral sequences (**Fig. 2.22**).

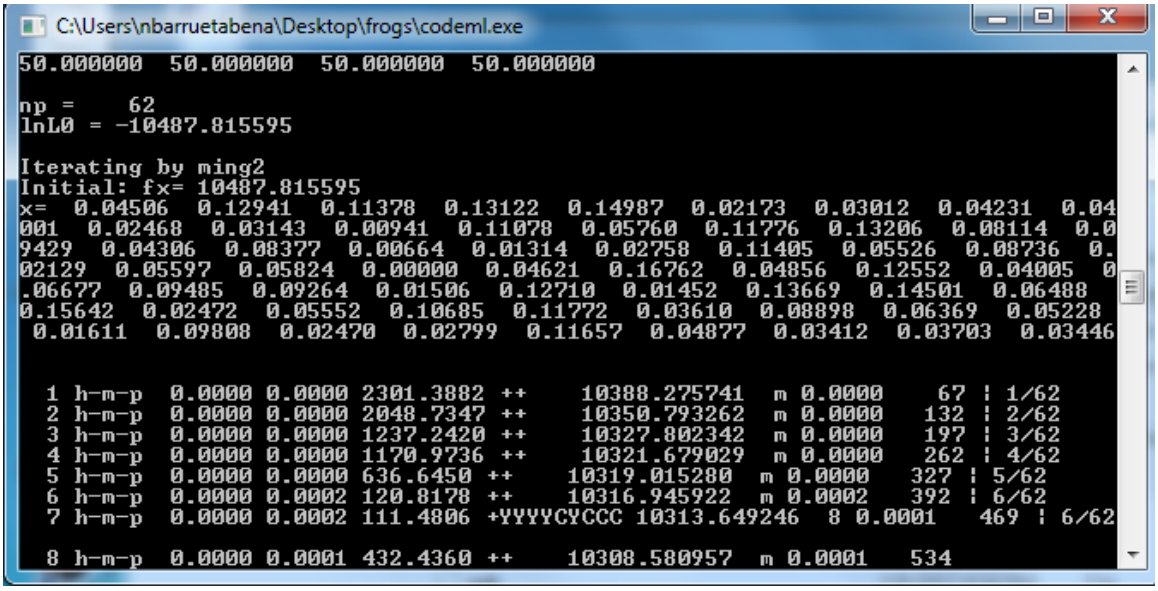


Figure 2.22. *PAML program running the calculations for the ancestral reconstruction of proteins.*

Once the process is over, the program will create an rst file will all the information concerning the process the posterior probabilities and the joint and marginal protein reconstructions. In this thesis, we have selected the marginal reconstructions for their further nanomechanical analysis

2.4. Available software

Over the last years, the software packages to implement ancestral reconstruction have raised very fast. To simplify the following section, the next table comprises the most used programs for ancestral sequence reconstruction in the field of molecular evolution, discarding the packages used in other applications.

Name	License	Supported Formats	Characters	Methods	[Ref.]
APE	General Public License	Nexus, Fasta, Clustal	N, P	ML	[137]
BEAST	General Public License	Nexus, Beast XML	N, P	Bayesian	[132, 133]
FastML	Copyright	Fasta	N, P	ML	[138]
HyPhy	Free Documentation License	Mega, Nexus, Fasta, Phylip	N, P	ML	[139]
MEGA	Proprietary	MEGA	N, P	MP, ML	[127, 128]
Mesquite	Creative Commons License	Fasta, NBRF, Genbank, Phylip, Clustal, TSV	N, P	Parsimony, ML	[140]
MrBayes	General Public License	Nexus	N, P	Bayesian	[141, 142]
PAUP	Proprietary	Nexus	N, P	Parsimony	[130, 131]
PAML	Proprietary	Phylip, Nexus, Fasta	N, P	ML	[135, 136]
PHAST	BSD License	Multiple alignment	N	ML	[143]

Table 4.1. *Some of the most popular software used for ancestral sequence reconstruction if the field of molecular evolution. In the column "Character" N refers to nucleotide while P is protein. In the column methods ML and MP correspond to maximum likelihood and maximum parsimony, respectively.*

Chapter 3: Experimental methods

3.1. Molecular biology techniques

The following sections explain in detail all the molecular biology techniques used in this thesis. To produce the pool of ancestral and extant titin sequences for their further nanomechanical analysis, a series of steps must be performed in the laboratory (**Fig. 3.1**). First, the DNA sequences that encode the protein of interest must be purchased. These genes should be inserted into a bacterial expression vector (called plasmid). This process is called ligation. The plasmids are later transformed into host bacteria in order to induce the protein expression. Once the protein of interest is expressed, the cells should be lysed to liberate the proteins. Finally, the resulting protein must be purified. Titin I65-I72 constructs for human, zebra finch, orca, rat, chicken, LTCA, LSCA, LMCA and LPMCA were produced using this protocol. The entire process is explained step-by-step in the following sections. For thioredoxin, the process starts in section **3.1.5.**

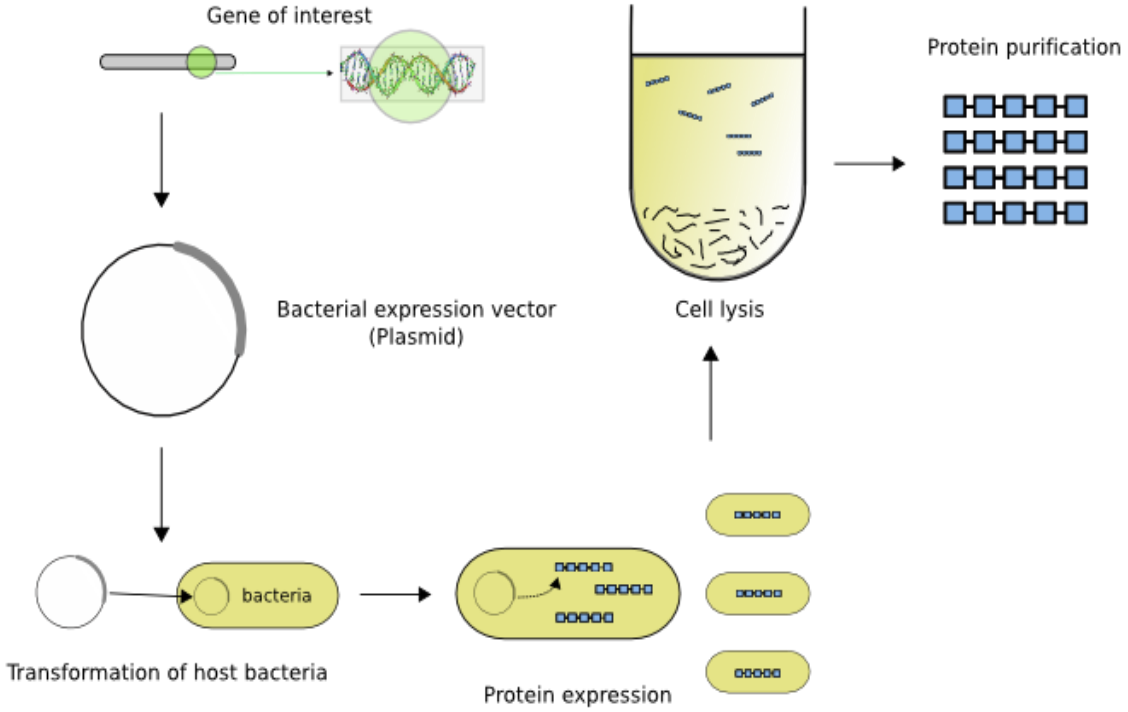


Figure 3.1. *Schematic representation of the molecular biology techniques used to produce proteins.*

3.1.1. Cloning of commercial plasmid

The genes encoding the I65-I72 titin constructs were codon optimized and purchased in a commercial plasmid (from *Life Technologies*). This plasmid also contains a kanamycin antibiotic resistance gene for the proper selection of them when they are cloned 1 µL of the commercial plasmid (50 ng/µL) was transformed into *E.coli-XL1Blue* competent cells (*Agilent Technologies*) following the manufacturer's protocol [144]. After transformation, the competent cells were grown in 450 µL of SOC medium (*Invitrogen*) for one hour, spread out onto LB-agar-kanamycin plates and incubated overnight at 37 ºC. Later, single colonies were isolated and grown in 10 mL of LB media + 1% 100 mg/mL kanamycin for 16 h at 37 ºC while gently agitating. Cells were harvested by centrifugation (14000 rpm, 10 min, 4 ºC, *Eppendorf Centrifuge 5810R*) and plasmids were extracted using a DNA-plasmid extraction kit (*Thermo Scientific*) following the manufacturer's protocol [145]. Purified plasmids were eluted with nuclease-free water and their concentration was calculated with the *Nanodrop 2000L* system.

3.1.2. Digestion of commercial plasmid

Once the commercial plasmid containing the titin gene is amplified, the enzymatic digestion of the gen is carried out. For that purpose, a triple digestion strategy with the cutting enzymes *BamHI* - *KpnI* - *SphI* is used. *BamHI* and *KpnI* restriction sites are flanking the borders of the titin gene, whereas *SphI* has two restriction sites over the rest of the plasmid (**Fig. 3.2**). A triple digestion is necessary because the products obtained with a standard double digestion with *BamHI* – *KpnI* are very similar in size (2269 bp for titin gene versus 2280 bp for the plasmid) and, therefore, very difficult to separate later on a DNA-agarose gel.

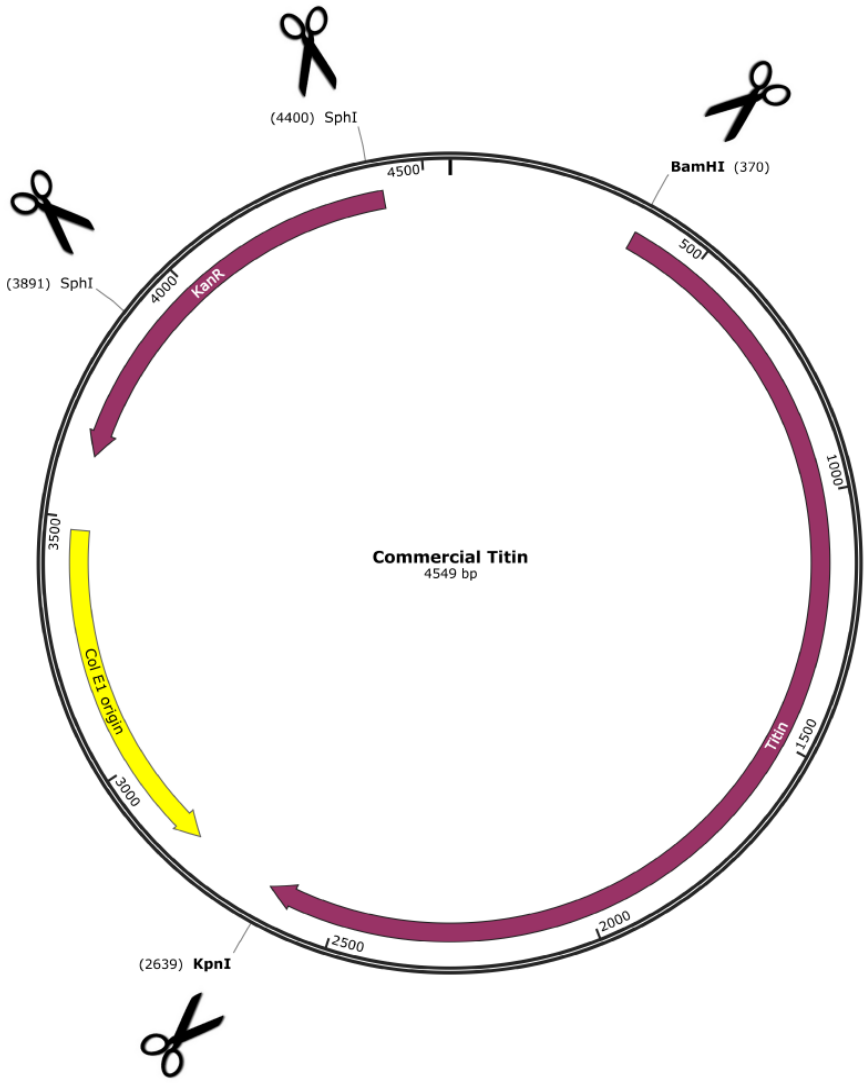


Figure 3.2. *Commercial plasmid encoding a titin gene. Restriction sites for BamHI, KpnI and SphI are illustrated with scissors.*

All the digestions done in this thesis are carried out with enzymes purchased from *Thermo Scientific* and following the manufacturer's *Fast Diges*t protocol. The final digestion volume is adjusted to 50 µL and incubated at 37 ºC for one hour. The digestion products are screened in a DNA-agarose gel (1%) in TAE buffer**.** All DNA-agarose gel is run using the *BioRad* agarose electrophoresis equipment for approximately 90 min. Once the gels are run, the band corresponding to the titin is extracted from the gel. The gene is purified with a DNA-extraction kit from *Thermo Scientific* following the usual protocol [146]. Concentrations are calculated using the *Nanodrop 2000L* system.

3.1.3. pQE80-titin construct ligation

After digestion, the genes encoding the titin must be ligated onto a high-efficiency bacterial expression vector with compatible cohesive ends. For that purpose, the previously digested *BamHI-pQE80-KpnI* open plasmid was used (**Fig. 3.3**). This plasmid was a kind gift from Professor Julio Fernandez's lab at Columbia University. It also contains an ampicillin resistance gene. *Invitrogen*'s T4-DNA ligase protocol [147] is used for the ligation process between pQE80 and the titin genes. The Mol ratio between the amount of plasmid vector and the titin gene insert is 3:1. All calculations for the needed amount of plasmid and DNA inserts are obtained following the formula (3.1). Ligations were incubated overnight at room temperature. Thereafter, to stop the process, ligations were diluted 5 times with deionized water.

$$DNA\ insert\ (ng) = 3 * \frac{Plasmid\ vector\ (ng) * DNA\ insert\ (bp)}{Plasmid\ vector\ (bp)} \qquad (3.1)$$

Figure 3.3. *pQE80-titin plasmid. Compatible cohesive ends for BamHI and KpnI are illustrated with a glue symbol.*

3.1.4. Cloning of pQE80-titin plasmid

For the amplification of pQE80-titin plasmid a similar methodology that the one described in section **3.1.1.** is used. In this case, 5-30 µL of the recombinant plasmid (depending on the concentration) is transformed into *E. coli-XL1Blue* competent cells following the same protocol. Competent cells are later spread out onto LB-agar-ampicillin plates and incubated overnight at 37 ºC. Again, single colonies are taken out and grown in 10 mL of LB media + 0.1% 100 mg/mL ampicillin for 16 h at 37 ºC. Cells are harvested by centrifugation (14000 rpm, 10 min, 4 ºC) and plasmids are extracted using a DNA-

plasmid extraction kit. Purified plasmids were eluted with nuclease-free water and their concentration is calculated with the *Nanodrop 2000L* system. The plasmids are screened and verified in a DNA-agarose (1%) gel and concentration is calculated using the *Nanodrop 2000L* system.

3.1.5. Screening of titin constructs

Once the pQE80-titin plasmids are amplified, an amount between 1-10 µL of the plasmid is transformed onto *E.coli-Origami2* competent cells following the vendor protocol [148]. After transformation, cells are grown in 450 µL of SOC medium for 1 hour at 37 ºC and spread out in LB-agar-ampicillin plates. Plates are incubated overnight at 37 ºC to grow the colonies. After this, 4 single colonies are isolated and grown in 10 mL LB medium + 0.1% 100 mg/mL ampicillin for 8 hours or until the optical density (OD) of the medium reaches 0.6. ODs are calculated with the *Nanodrop 2000L* system.

To induce the overexpression of titin by T7 promoter activation, 10 µL of IPTG (isopropyl-β-D-thiogalactopyranosid, *Sigma Aldrich*) 100 mg/mL is added to the medium and the solution is incubated overnight at 37 ºC. After this, 1 mL of each colony is taken to screen the overexpression. Bacteria are harvested by centrifugation (14000 rpm, 10 min, 20 ºC). Supernatant is discarded and bacteria are resuspended in 10 µL of extraction buffer. 20 µL of 2xSDS page Sample Buffer solution is added for the denaturation and charging of the protein in acrylamide electrophoresis gel separation. The bacteria are centrifuged again (14000 rpm, 30 min 20 ºC) and boiled at 95 ºC for 2 min.

To verify which colony overexpresses the titin construct better, 20 µL of each of the solutions are run in an 8% acrylamide gel for approximately 100 min in a *BioRad* acrylamide electrophoresis system. After the run, gels are rinsed in deionized water for 30 min. Proteins in the gel are dyed with Bradford solution (*Thermo Scientific*) for 20 min and rinsed with deionized water again. Negative controls without IPTG are also added to the gel to visualize the overexpression better (**Fig. 3.4**).

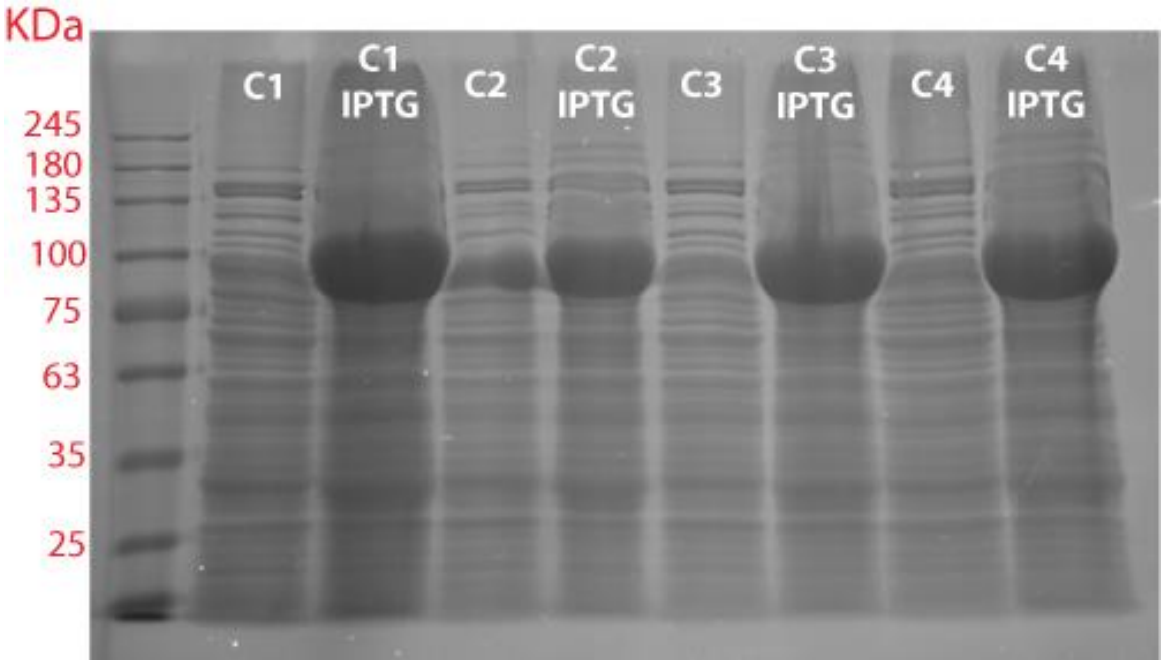


Figure 3.4. *8% Acrylamide electrophoresis gel for an I65-I72 titin screening (~88000 KDa). Protein ladder (Nippon Genetics) can be visualized in the left side of the picture. C1-C4 nomenclature refers to the 4 colonies isolated.*

3.1.6. Protein expression and purification

Once the colony with the best overexpression is selected, the remaining 8 mL of LB media with the desired bacteria are added to 800 mL more LB media + 0.1% 100 mg/mL ampicillin + 0.1% 50 mg/mL chloramphenicol. Chloramphenicol is added to maintain the ability of overexpression of the bacterial pLys system.

The media is incubated for 8 h or until OD > 0.6 at 37 ºC shaking (250 rpm) to facilitate the growth of the bacteria. After this, 0.1% 100 mg/mL IPTG is added to induce the overexpression of the protein. Media is again incubated for another 16 h at 37 ºC while shaking. Thereafter, bacteria are separated from the media by centrifugation (4000 rpm, 4 ºC, 20 min). Supernatant is discarded. The bacteria pellet is later resuspendend in 16 mL of extraction buffer and 160 µL of protease inhibitor (*Merck Millipore*) is added.

In order to achieve a proper cell lysis with French press, several reactives have to be added to the bacteria suspension. The first step is to add 160 µL of 100 mg/mL lysozyme (*Thermo Scientific*) solution for the enzymatic destabilization of the bacterial membrane

and incubate the suspension at 4 ºC with orbital shaking (5 rpm). After this, a series of reactives are added:

- 1.6 mL of 10% Triton X-100 (*Sigma Aldrich*) for the chemical destabilization of the bacterial membrane.
- 80 µL of 11 mg/mL DNAse I (*Invitrogen*) for the enzymatic degradation of DNA.
- 80 µL of 1 mg/mL RNAse A (*Ambion*) for the enzymatic degradation of RNA.
- 160 µL of 1M $MgCl_2$ (*Sigma Aldrich*) as a catalyzer to increase the enzymatic activity of DNAse and RNAse.

The suspension is incubated again for 10 min at 4 ºC with orbital shaking prior to the cell lysis. Cell lysis is carried out with a Frech press machine (*G. Heinemann HTU DIGI-F Press*). Suspension of cells is introduced in the press chamber. Cells are lysed at 18000 psi during 30 min. The lysate obtained is centrifuged in a high-speed centrifugation system (33000 rpm, 4 ºC, 90 min; *Beckman Coulter Avanti J-26 XPI*). After centrifugation, the pellet is discarded and the supernatant filtered with 0.8, 0.45 and 0.22 µm syringe filters (*Merck Millipore*) consecutively.

The first purification process is carried out by means of a HisTrap cobalt affinity resin (*Thermo Scientific*). All the I65-I72 titin constructs contain a HisTag composed of 6 consecutive histidines in the N terminus of the construct. The HisTag end poses the ability to specifically bind to the cobalt affinity column. This binding can be later eluted by adding imidazol in the buffer due to its higher affilnity to cobalt. After this first purification titin constructs are incubated overnight at room tempreature with 0.5% H_2O_2 (*Sigma Aldrich*) to enhance the formation of disulfide bonds.

A second purficiation process is performed with an *ÄKTA pure* fast protein liquid chromatography (FPLC) system (*GE Healthcare*) with a *Superdex 200* column of 30 cm in diameter (*GE Healthcare*). Fractions of interest are collected from the chromatogram (**Fig. 3.5**) and stored at -20 ºC in HEPES medium.

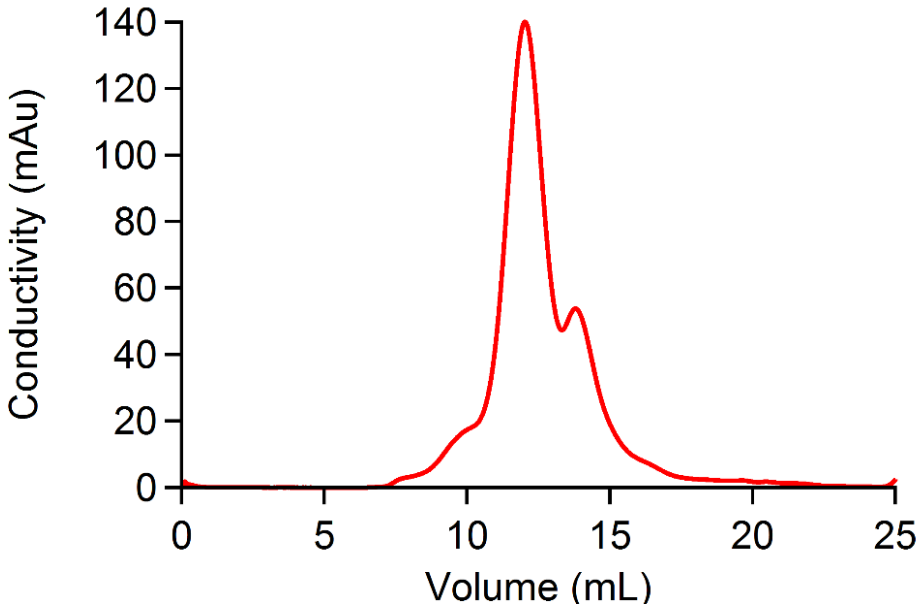


Figure 3.5. *Characteristic chromatogram for an I65-I72 titin construct. Fractions corresponding to the peak maximum (~12 mL) were collected and stored in the freezer.*

3.2. Single-molecule force spectroscopy

The following section will explain the setup used to carry out the single-molecule experiments. All the smFS experiments were conducted with a Luigs & Neumann commercial microscope in both force extension and force clamp modes. This AFM is specifically designed to perform single-molecule force spectroscopy experiments.

3.2.1. Initial setup

To run a smFS experiment, the initial setup is the following. First, a clean gold surface (previously cleaned by an isopropanol bath and rinsed with deionized water) is attached to the piezoelectric actuator with vacuum grease (*Dow Corning*). Later, a certain amount of the purified heteropolyprotein (normally, 20 µL at 10-20 µg/mL) is placed onto the gold substrate (**Fig. 3.6a**) until it is absorbed. In this way, we promote the attachment of the proteins to the gold substrate due to the two terminal cysteines added at the C terminus of their sequence. These cysteines create a strong thiol bond with the gold surface. Thereafter, the cantilever is mounted into a cantilever holder (**Fig. 3.6b**).

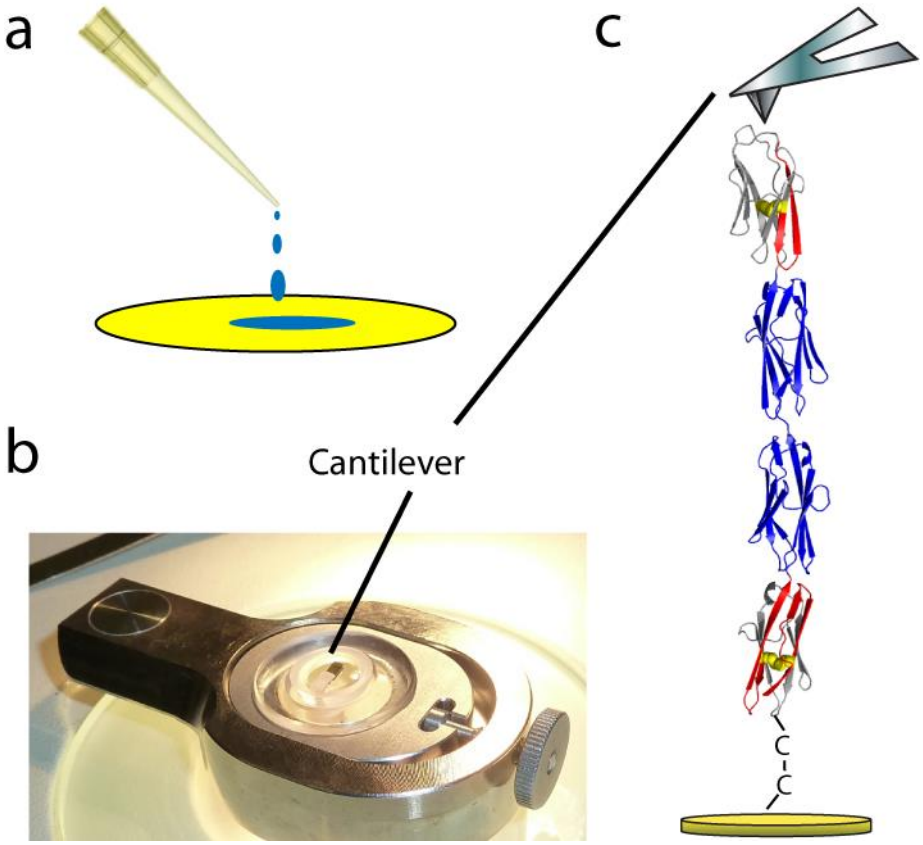


Figure 3.6. *Initial setup of a smFS experiment. (**a**) 10-20uL of protein is placed in a clean gold surface. (**b**) The cantilever is positioned on the holder. (**c**) Schematic representation of the experiment after the mounting process and stretching of the polyprotein.*

Once the cantilever is properly held and fixed, the fluid cell is filled with HEPES solution. After this, the cantilever holder is mounted in the AFM head and the gold surface is approached to the cantilever by moving the piezoelectric actuator. (**Fig 3.6c).** Once the cantilever is close to the gold surface, the force calibration of the cantilever is needed prior to start the experiment.

3.2.2. Force calibration

Once the cantilever holder is placed into the AFM head, the laser beam is focused on the backside of the cantilever and reflected onto a four-quadrant photodiode (PD) (**Fig. 3.7**). The backside of the cantilever is normally coated with a thin gold surface to improve

the reflection of the laser. The photodiode measures the electrical potential of the laser beam in volts (V). This voltage needs to be converted into a force (in Newton, N).

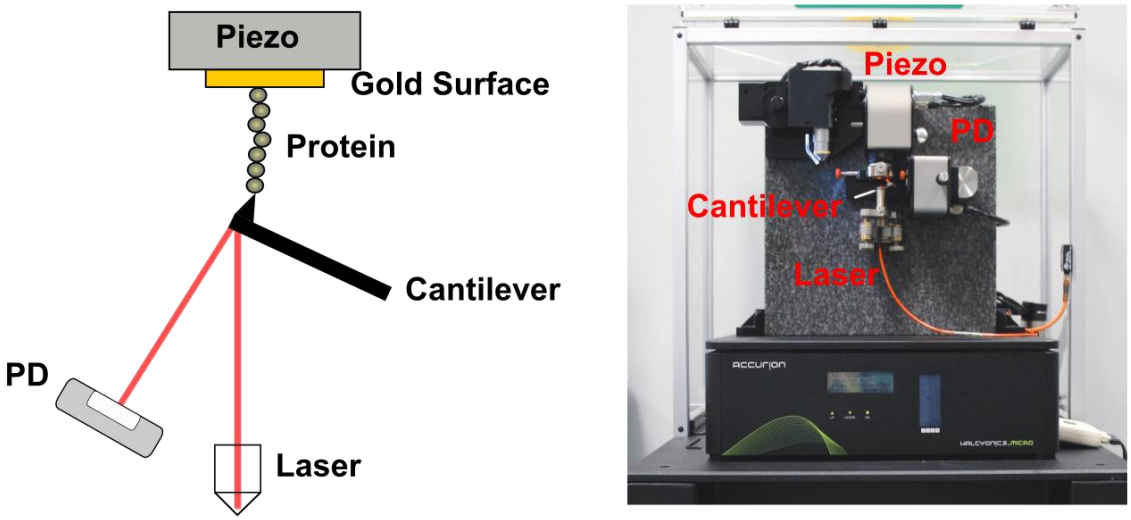


Figure 3.7. *Schematic representation and real picture of the AFM. The important parts of the microscope are named.*

To calibrate the cantilever, it is assumed that it behaves as a Hookean spring. Hence, the force (F) is calculated as the product of the cantilever deflection in the z axis (Δz_c) and its spring constant (k_c). Therefore, the force can be calculated with the following formula:

$$F = -k_c \, \Delta z_c \qquad (3.2)$$

The Luigs & Neumann AFM is controlled with the Igor Pro software (https://www.wavemetrics.com/) (**Fig. 3.8a**). Here, one can calibrate the deflection of the cantilever for protein unfolding by the thermal fluctuation method [149], which assumes that the cantilever tip behaves as a simple harmonic oscillator under equilibrium conditions. This assumption then models the cantilever tip as an ideal spring for small deflection angles. When the tip is far enough from the sample, the frequency of the motion near the resonant frequency permits an approximation for the k_c. Therefore, the cantilever deflects due to thermal motion with a harmonic resonant frequency ω_0 (**Fig. 3.8a and b**).

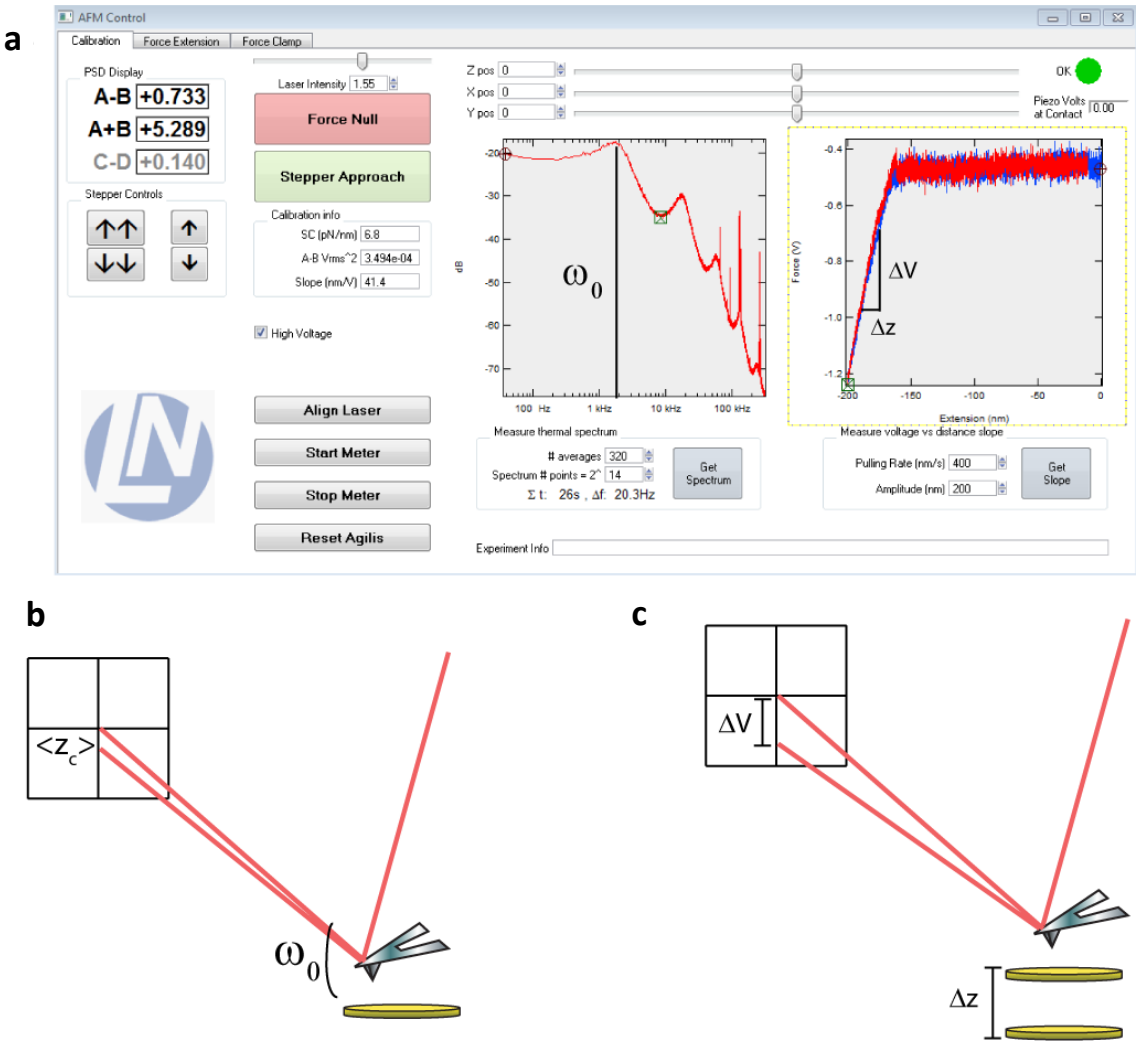


Figure 3.8. *(**a**) Igor Pro interface for force calibration. The harmonic resonant frequency and ratio between the photodiode output and the displacement of the piezoelectric actuator can be calculated here. (**b**) Schematic representation of the thermal fluctuation effect interactions with the laser and the photodiode. (**c**) Scheme representing the ratio photodiode output and the displacement of the piezoelectric actuator.*

Moreover, to calculate the force, it is also necessary to determine then ratio between the photodiode output and the displacement of the piezoelectric actuator in *z* direction ($\Delta V/\Delta z_p$) (**Fig. 3.8a and c**). This ratio can be calculated by approaching and retracting the gold surface to the cantilever and obtaining the slope $\Delta V/\Delta z_p$ when the surface and the cantilever are in contact. Under these conditions, it can be now assumed that:

$$\Delta z_c = \Delta z_p \qquad (3.3)$$

The displacement of the piezoelectric actuator and the deflection of the cantilever are the same. According to the equipartition theorem [149], the kinetic energy of each degree of freedom (vibrational modes, for instance) equals half of the thermal energy. Then:

$$\frac{1}{2}m\omega_0^2\langle z_C^2\rangle = \frac{1}{2}k_c\langle z_C^2\rangle, \qquad (3.4)$$

where:

$$k_c\langle z_C^2\rangle = k_B T, \qquad (3.5)$$

$\frac{1}{2}m\omega_0^2\langle z_C^2\rangle$ refers to the kinetic fraction of the Hamiltonian. Here, z_C is the displacement of the spring in the z direction. When $k = m\omega_0^2$ the constant is calcluted by obtaining the mean –square displacement of the harmonic frequency:

$$k = \frac{k_B T}{z_C^2} \qquad (3.6)$$

Moreover, to calibrate the cantilever, the laser beam signal detected in the photodiode and the displacement of the piezoelectric actuator should be correlated. Thus, *k* is finally calculated by:

$$k = \frac{k_B T}{\omega_0^2 s^2}, \qquad (3.7)$$

where $s = \Delta z/\Delta V$. Every cantilever tip has a characteristic spring constant range that should be provided by the manufacturer. In this thesis, the experiments were carried out with MLCT (*Bruker*) and Biolever (*Olympus*) silicon nitride cantilevers. For MLCT cantilevers tips C (spring constant k ~ 10-20 pN/nm; resonance frequency f_R ~ 1 kHz in buffer) and D (spring constant k ~ 30-40 pN/nm; resonance frequency f_R~ 3-4 kHz in buffer) were used (**Fig. 3.9a**). For Biolever cantilevers tip B was used (spring constant k ~ 3-6pN/nm; resonance frequency f_R ~ 1-2 kHz in buffer) (**Fig. 3.9b**).

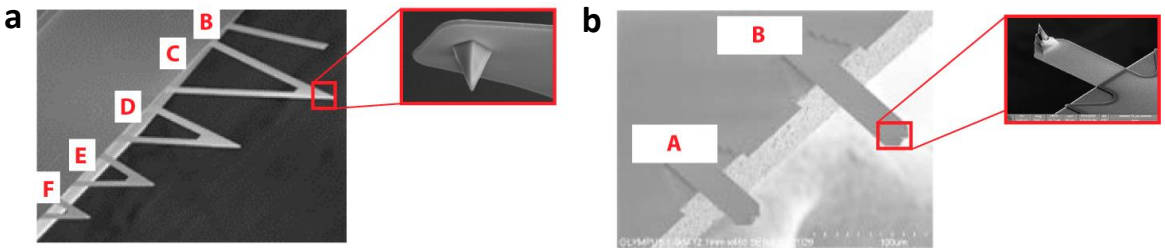


Figure 3.9. *(**a**) MLCT and (**b**) Biolever cantilevers From Bruker and Olympus respectively. The different tips of each cantilever are shown (images modified from www.brucker.com and www.olympus.es).*

3.2.3. Force extension mode

In force extension mode, the AFM completes the approach-retraction cycle at a constant velocity: Therefore, the applied force on the protein cannot be controlled. In a typical experiment, the polyprotein is attached to the gold surface by the terminal cysteines while the cantilever tip "fishes" the other terminal by a still unclear phenomenon, often assigned to physioadsorption, due to the its electrostatic nature [150]. The unfolding of the polyprotein will result in a force-extension graph with a sawtooth-like pattern (**Fig. 3.10**). Here, each of the peaks of the graph corresponds to the unfolding of a single domain. Finally, the last peak corresponds to the detachment of the polyprotein from the cantilever tip.

Fig. 3.10 depicts a characteristic unfolding trace of the $(I91)_4$ polyprotein. First, the piezoelectric actuator approaches the gold surface until it touches the cantilever tip and bends it (**I**). Thereafter, in a positive event, the cantilever picks a polyprotein (**II**) and the piezoelectric actuator starts to retract, increasing the distance between the gold surface and the tip (**III**). A further extension of the polyprotein causes the unravelling of a single domain (**IV**). This phenomenon releases the amino acids present in domain, which continue elongating until the domain is fully stretched (**V**). This process will take place for every domain until the polyprotein is fully extended and the detachment between the polyprotein and the cantilever tip takes place (**VI**), which can be most of the times visualized as a bigger peak. Finally, the cantilever returns to the non-bent state (**VII**).

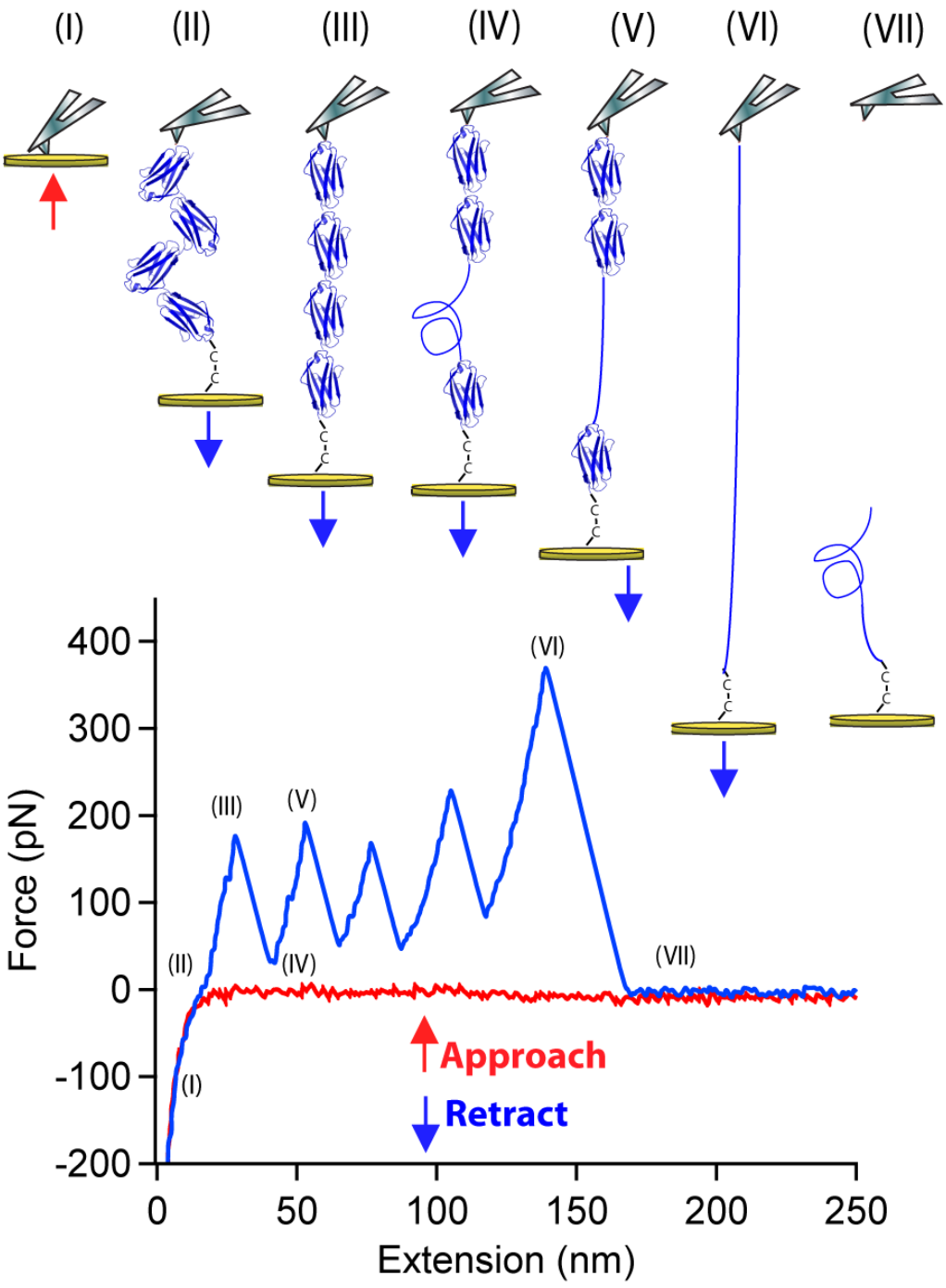


Figure 3.10. *Schematic representation of the unfolding forces and characteristic force extension trace for a $(I91)_4$ titin construct. The different steps of the process are represented in roman numbers (I-VII).*

In force extension mode, the Igor Pro interface allows to control various parameters (**Fig. 3.11**):

- Pulling rate: Controls the velocitiy of the approach-retract cycle in nm/s.
- Amplitude: Controls the length of the approach-retract cycle in nm.
- Sawtooth detector: permits to save the desired traces. It can be regulated by the stringency and by the minum forces of the peaks (in pN).
- Filter: Allows to filter the number of points of the trace by applying a Nyquist fraction. It applies a cut-off frequency after the complete adquisition of the trace.

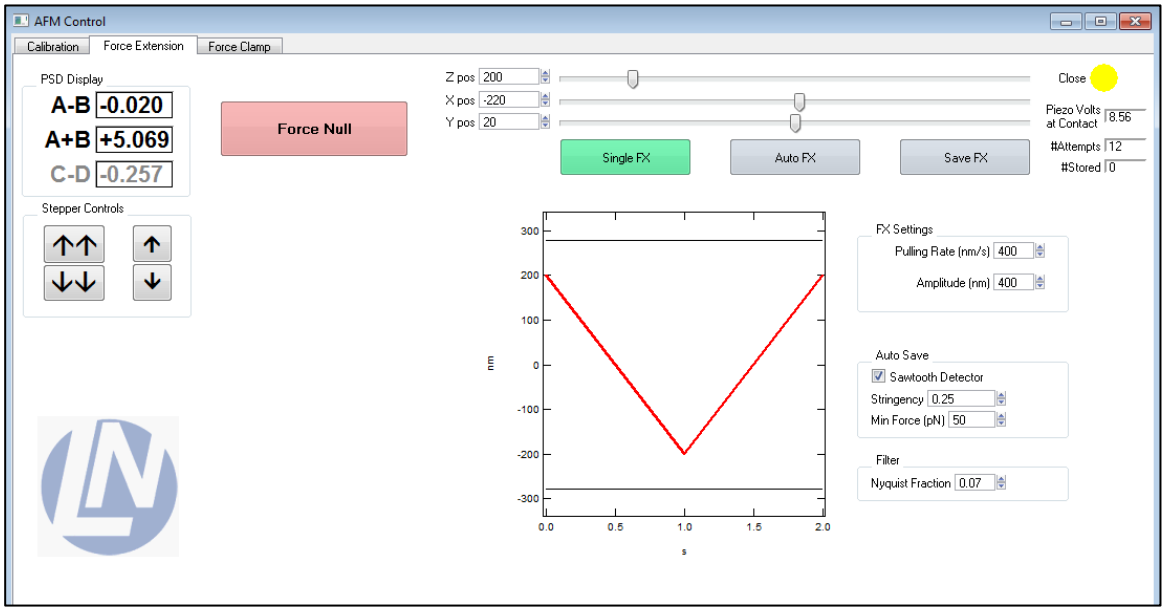


Figure 3.11. *Igor Pro interface for force extension experiments that permit to modify a series of parameters.*

All the I65-I72 heteropolyproteins analyzed in this thesis with the force extension mode of the AFM were pulled at 400 nm/s with amplitude of 400-450 nm. The sawtooth detector was set at a stringency of 0.5 and a minimum unfolding force of 50 pN. Finally, the Nyquist fraction ranged between 0.07 and 0.10 for all the experiments.

3.2.4. Force clamp mode

In the force clamp mode, the AFM works at a constant force. Here, the force applied to the polyprotein can be adjusted due to the implementation of a PID controller (**Fig. 3.12**). The PID controller creates a feedback loop that constantly corrects the extension of the protein to keep the desired force values. Therefore, in this mode, the unfolding of the polyprotein is detected as a staircase pattern in a length versus time trace, where each step corresponds to the unfolding of a single domain.

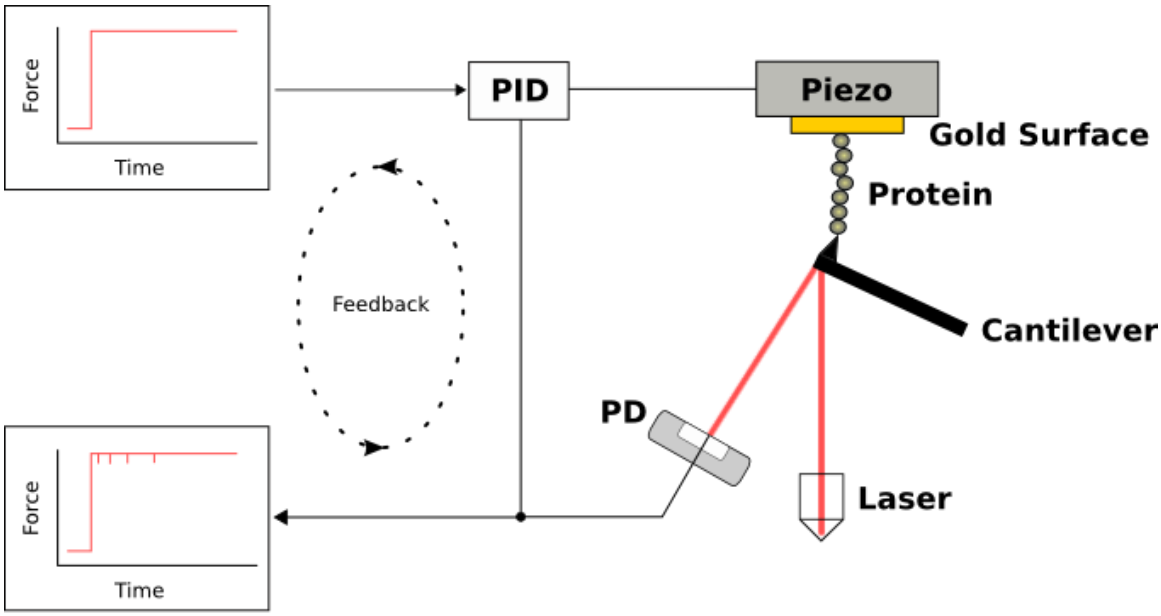


Figure 3.12. *Schematic representation of the AFM running in force clamp mode. The PID controller constantly adjusts the force applied to the protein with a feedback loop.*

Thus, every time that a domain unfolds the PID readjusts the force to the set point and moves the cantilever until it reaches the desired defection/force. This phenomenon can be visualized in the spikes of the force vs. length curve right after a staircase event in the length vs. time curve (**Fig. 3.13**).

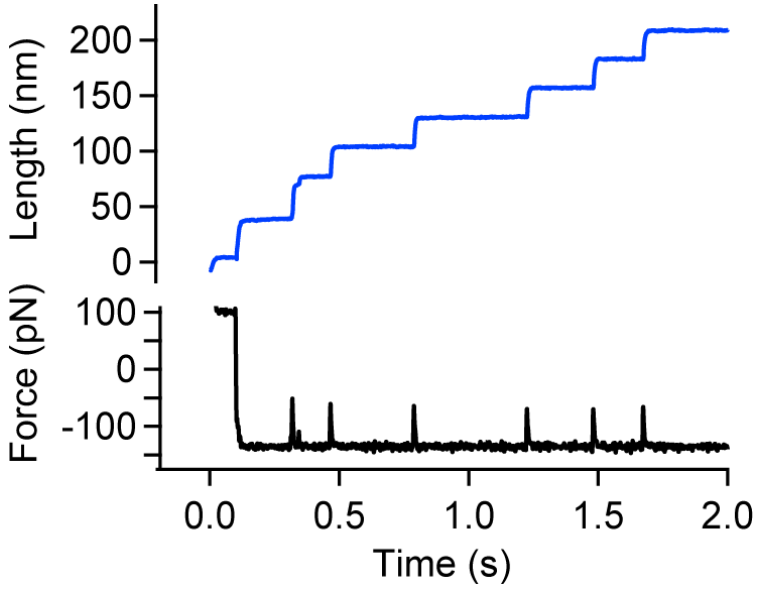


Figure 3.13. *Typical force clamp trace for an I65-I72 titin construct. Each step in the length curve corresponds to the unfolding of a single domain.*

In force clamp mode, one can develop a protocol with different forces and times. It is also possible to increase or decrease the force linearly with time (force ramp) [151]. Hence, this mode enhances greatly the flexibility of single-molecule experiments. In this sense, the Igor Pro interface allows to modify many parameters of the experiment and design a protocol in terms of force and time (**Fig. 3.14**).

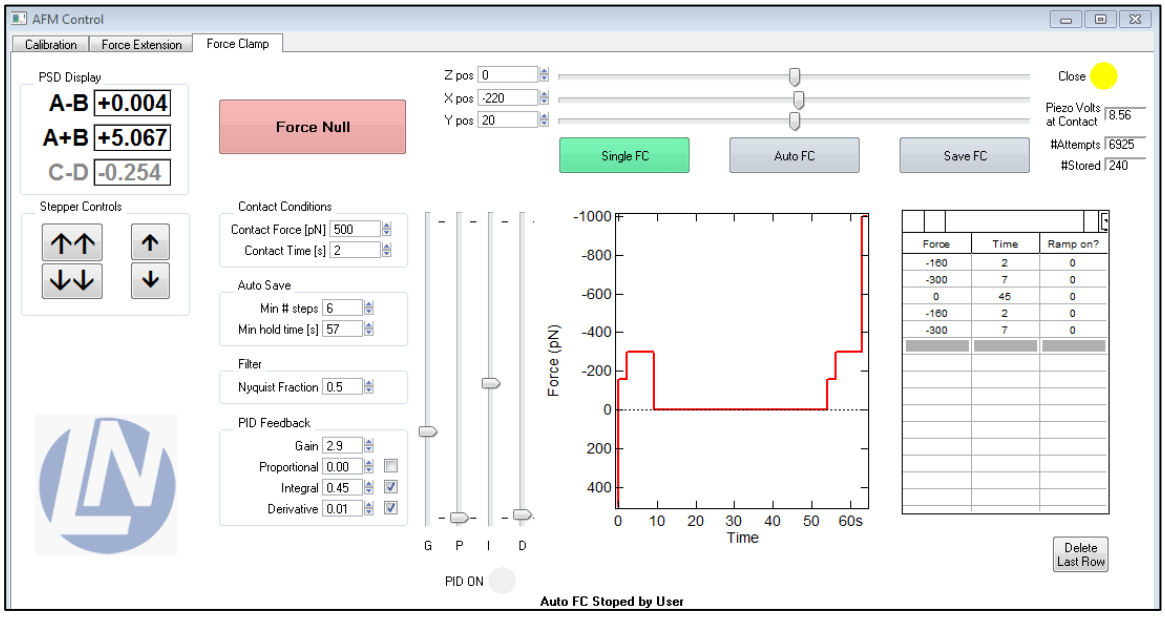


Figure 3.14. *Igor Pro interface for force clamp experiments.*

The initial parameters that can be modified are the following:

- Contact conditions: permits to set the contact force (pN) and the duration (s) of the contact phase before the experiment. High values imply higher probabilities to "fish" a protein, but also multiple tethers.
- Auto save: Controls the requirements to save a trace. The saved trace can be filtered by the number of well-defined steps or by a minimum hold time (s).
- Filter: Again, allows controlling the sampling rate by means of a Nyquist fraction after the complete acquisition of the trace.
- PID feedback: Sets the overall gain of the controller and permits to adjust individually the proportional (P), integral (I), and derivative (D) terms of the feedback. The feedback system should be adjusted to obtain a time resolution in the range of 2 ms (see **Fig. 3.15**).

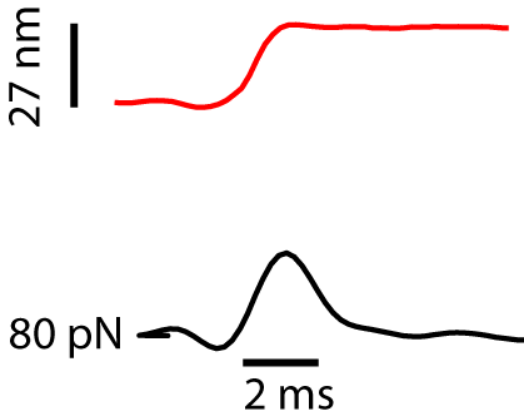


Figure 3.15. *Zoom of a force clamp experiment showing the feedback time of the force after an I27 unfolding event.*

Once the initial parameters are set, the force clamp protocol can be designed. Here, one can set the desired force and duration with many different force sequence steps if needed (**Fig. 3.16**). In this thesis force clamp was used for the detection of cryptic disulfide reduction events in I65-I72 constructs of LSCA and human titin in presence of thioredoxin (10-20 µg/mL in HEPES buffer). To achieve that, a first step at 135 pN during 2 s was applied to unfold the polyprotein and expose the possible disulfide bonds to the buffer with

the enzyme. Thereafter, the force was kept at 80 pN for 20 s to maintain the disulfides exposed and see their possible interactions with the thioredoxin (**Fig 3.16**).

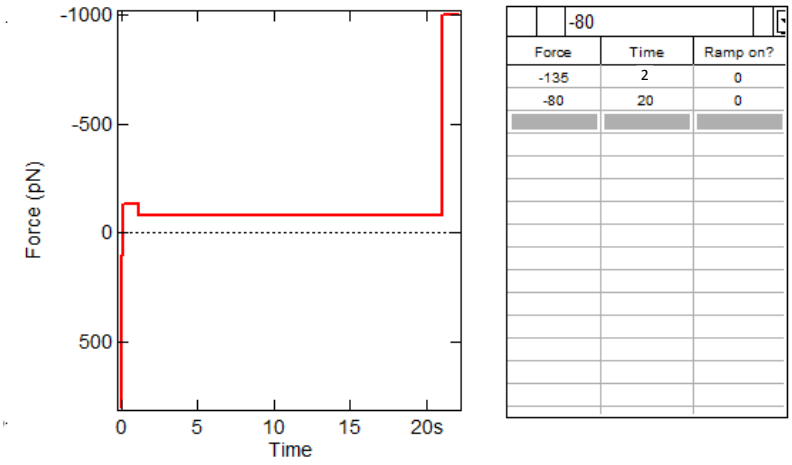


Figure 3.16. *Force clamp protocol for the detection of cryptic disulfides.*

3.2.5. Data analysis

All the smFS experiments were analyzed with Igor Pro software. The analysis file (AFM Analysis V2.40.ipf) is a kind present from Prof. Julio Fernandez's lab. In force extension mode experiments, each peak of the traces was fitted to the worm-like chain (WLC) model [152, 153] (**Fig. 3.17**):

$$F(x) = \frac{k_B T}{p}\left[\frac{1}{4}\left(1 - \frac{x}{L_C}\right)^2 - \frac{1}{4} + \frac{x}{L_C}\right], \qquad (3.8)$$

where k_B is the Boltzmann constant, T the temperature and p the persistence length. This model describes the entropic behavior and mechanic properties of an elastomer under an applied force with a specific contour length L_C. It has been proved also that WLC model fit well in an unfolding of a protein [152] when the persistence length is set at 0.4 nm (the average length of an amino acid). Thus, the difference of two consecutive unfolding peaks ΔL_C is the extension of a fully extended domain and can be calculated with this formula.

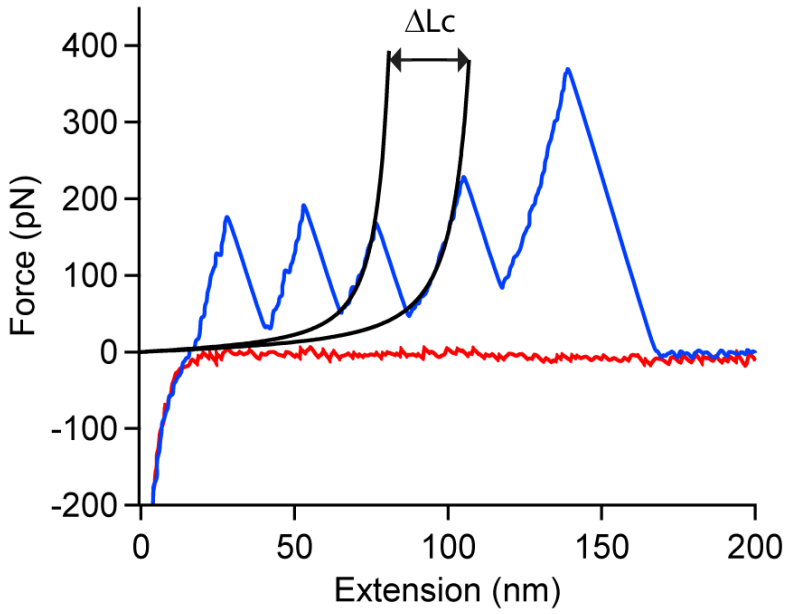


Figure 3.17. *Characteristic force extension trace of the $(I91)_4$ construct. Worm-like chain model fitting for an I91 domain is shown.*

3.3. Biochemical assays

A protocol for in-gel determination of oxidized thiols was adapted and optimized from previous reports [154]. 1 µg of each protein was incubated with 10 mM N-Ethylmaleimide in HEPES buffer in the presence of 3% w/v SDS for 30 min at 60 ºC to block all initially reduced thiols by irreversible alkylation. Samples were subsequently run on a 12% SDS-PAGE gel, and oxidized thiols were then reduced by incubation of the gel with 10 mM DTT for 1 hour at 60 ºC.

After three washes with 50 mM Tris-HCl, 10 mM EDTA, 3% w/v SDS, pH 6.8, the gel was incubated with a 5 mM mBBr solution in the same buffer for 2 h in the dark. Excess mBBr was removed by distaining the gel with 40% ethanol, 10% acetic acid overnight (3 changes). Fluorescent bands resulting from the reaction of the newly reduced thiols with mBBr were visualized on a Gel-Doc with UV excitation using standard filters for ethidium bromide emission.

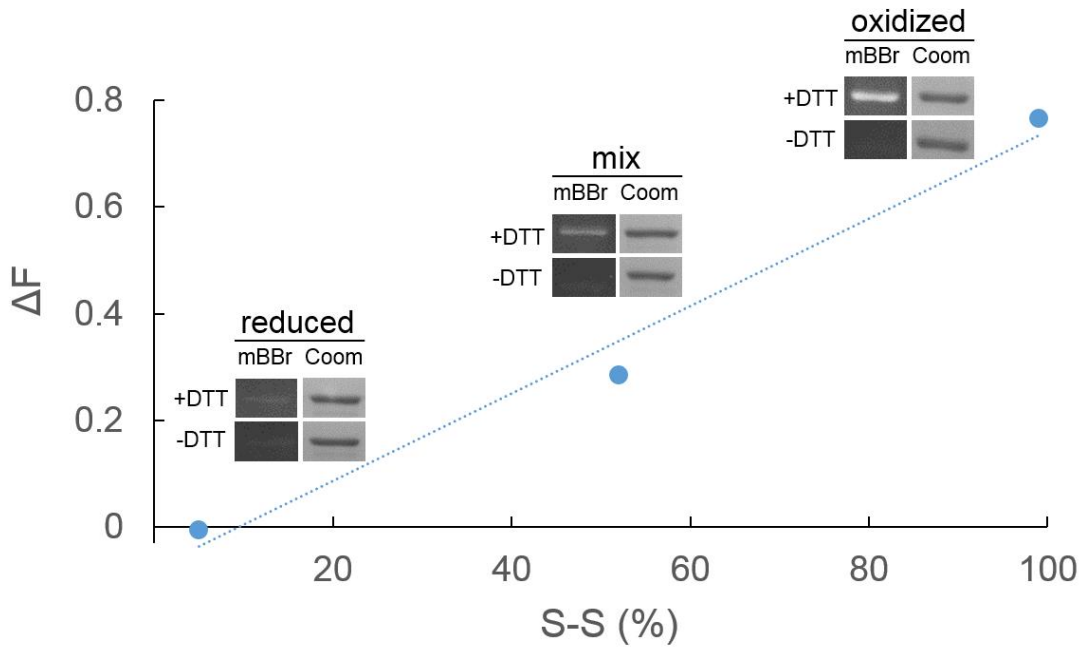


Figure 3.18. *Calibration experiment performed with (I9 $_{32-75}$)8 samples purified under reducing (5% S-S according to AFM and oxidizing conditions (99% S-S according to AFM).*

Quantification of the bands was performed by densitometry using the Quantity One software. The amount of protein in each well was later assessed by Coomasie staining and densitometry, and was used to normalize fluorescence signals. The fluorescent background of a replicate gel that was not reduced with DTT was subtracted. Oxidized and reduced ($I91_{32-75}$)$_8$ control proteins were also included in the experiment. Using force-clamp measurements [155], we estimated that 99% of I91-32/75 domains of the oxidized sample were oxidized, while 95% of the domains in the reduced sample were reduced (**Fig. 3.18**). These experimental values were used to estimate number of disulfides in the I65-I72 samples. These experiments were a collaboration with Dr. Jorge Alegre-Cebollada's group at National Center of Cardiovascular Diseases (CNIC).

Part III

Chapter 4: Phylogenetic results

This chapter focuses on the phylogenetic and mutational research carried out in this thesis. Specifically, we have reconstructed the chronograms of titin by parsimony and Bayesian inference. Phylogeny of myosin was also computed by this last method. After this, several internal nodes of both trees were inferred by maximum likelihood to compare them with relevant extant species in terms of identity and mutability. Finally, we have studied the role that cysteines have played during evolution in these sarcomeric proteins.

4.1. Ancestral reconstruction of titin

4.1.1. Ancestral reconstruction of titin using parsimony

To reconstruct the phylogeny of titin, a set of 33 titin sequences were used from which 28 corresponded to the full sequence of titin with over 30,000 residues. The sequences represent five different classes of vertebrate animals: mammals, amphibians, reptiles, birds and bony fishes, and were retrieved from UniProt [121] and GenomeNet [123] databases. All sequence ID numbers are listed in **Appendix I**. The sequences were aligned using the MUSCLE software [126] (**Fig. 4.1**) and further edited manually to remove unaligned regions.

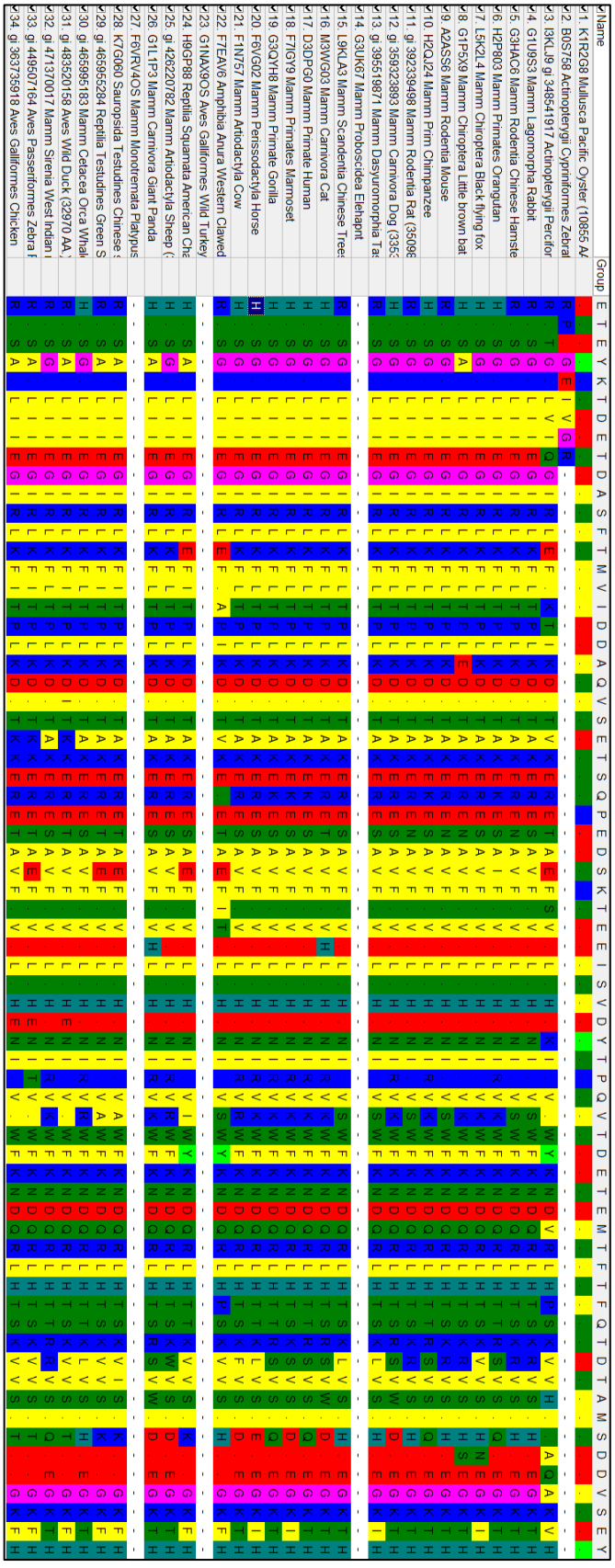


Figure 4.1 . Fragment of the alignment of the titins used for the ancestral reconstruction.

A phylogenetic tree was inferred with this alignment based on the maximum parsimony criterion (**Fig. 4.2**). We used PAUP* 4.0 software [130] with the heuristic search option and performing 2,000 bootstrap replicates. All bifurcations showed high bootstrap support, with most of them around 100% and a minimum of 67% for the ancestral node encompassing Chinese tree shrew and the hominids. Another ancestral node with relatively low values is the one that merges the Tasmanian devil and the placental mammals (69%). Nevertheless, these values are in accordance with the standards for the maximum parsimony criterion, which sets the minimum parsimony values in in 60%.

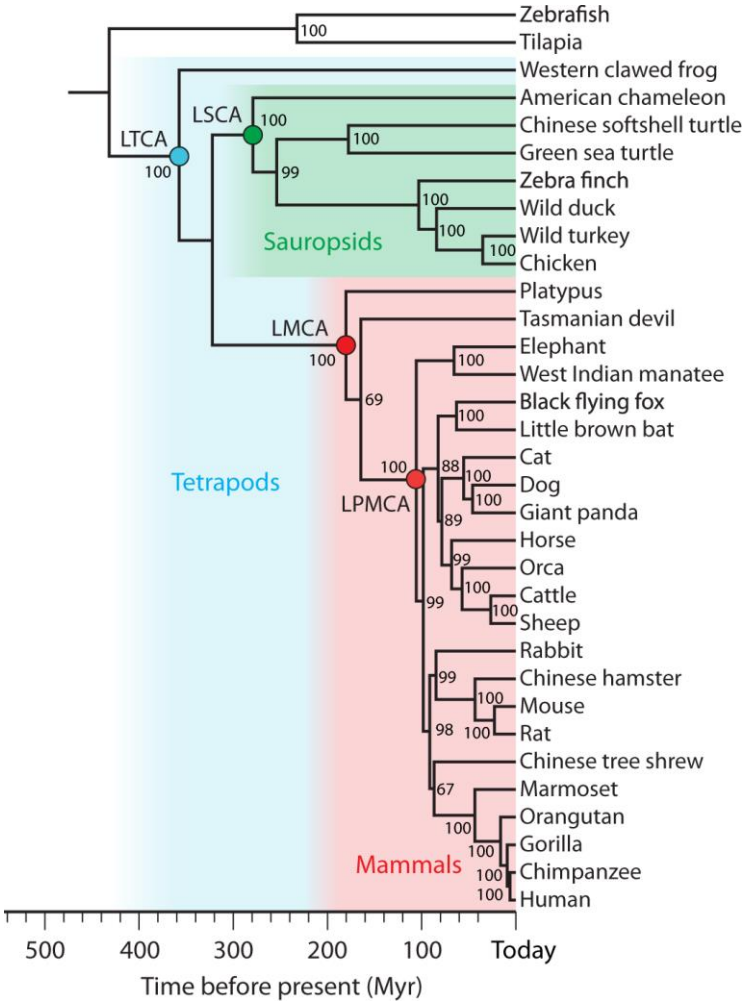


Figure 4.2. *Phylogenetic tree of titin using parsimony. A total of 33 titin molecules from different animals were used to compute the phylogenetic tree. Bootstrap support for each bifurcation is indicated. Diverse sources [54, 156] were considered for divergence times of ancestral nodes.*

The resulting tree splits properly the five clades representing the 33 animals present in the tree: mammals, birds, reptiles, amphibians and fishes. Each clade is also well separated on their corresponding subgroups according to the Time Tree of Life [54]. In the case of mammals, the non-placental mammals appear properly separated from placentals. Specifically, the separation includes two single-species monophyletic groups: one from the platypus (*Monotremata*) and another one for Tasmanian devil (*Marsupialia*). The groups encompassing rodents, primates, *Carnivora*, *Chiroptera* and *Proboscidea – Sirenia* are properly separated as well according to Time Tree of Life. The only possible discrepancy is in the group of *Laurasiatheria*. In this tree *Perissodactyla* (horse) is placed as a sister group of *Cetartiodactyla* (orca whale plus cattle and sheep). Despite the historical separation based on morphological features is contrary to this tree, recent studies based on genetic data are agreement with our phylogeny [157]. For birds and reptiles, the phylogeny is also consistent with the Time Tree of Life, being all the posterior probability values 1. The tree also groups all the animals of these two clades in a monophyletic group, which encompasses all the sauropsids. The only amphibian in this analysis (Western clawed frog) is well placed too, as it segregates before reptiles by 356 Myr, which agrees with data from Time Tree of Life. Altogether, they compose a monophyletic group for all the tetrapods in this phylogeny.

4.1.2. Ancestral reconstruction of titin using Bayesian inference

We also computed the phylogeny of with a more robust method to optimize the statistics, as the bootstrap support was not optimum in the previous tree. The same alignment was tested to find the best model of protein evolution using ProTest [158], resulting the Jones-Taylor-Thornton (JTT) with gamma distribution model as the best evolution model. With this alignment, we constructed a phylogenetic chronogram using Bayesian inference **(Fig. 4.3)** using BEAST v1.8.2 package software [159] incorporating the BEAGLE [134] library for parallel processing with the aim of improving the probability values obtained with the maximum parsimony method. Monophyletic groups for primates, *Rodentia, Carnivora, Chiroptera*, *Cetartiodactyla*, archosaurs, testudines, Squamata and fishes, being the latest the selected outgroup. Following ProTest, we selected the model JTT with 8 categories in gamma distribution and Yule model for speciation and length chain of 25 million generations, sampling every 1000 generations. Calculations were run

for 12 days in a single node of an HPC cluster of Intel Xeon 2680v2 processors, using 16-cores at 2.6 GHz and 64 GB of memory. We discarded the initial 30% of trees as burn-in. All nodes were supported by posterior probabilities above 0.99.

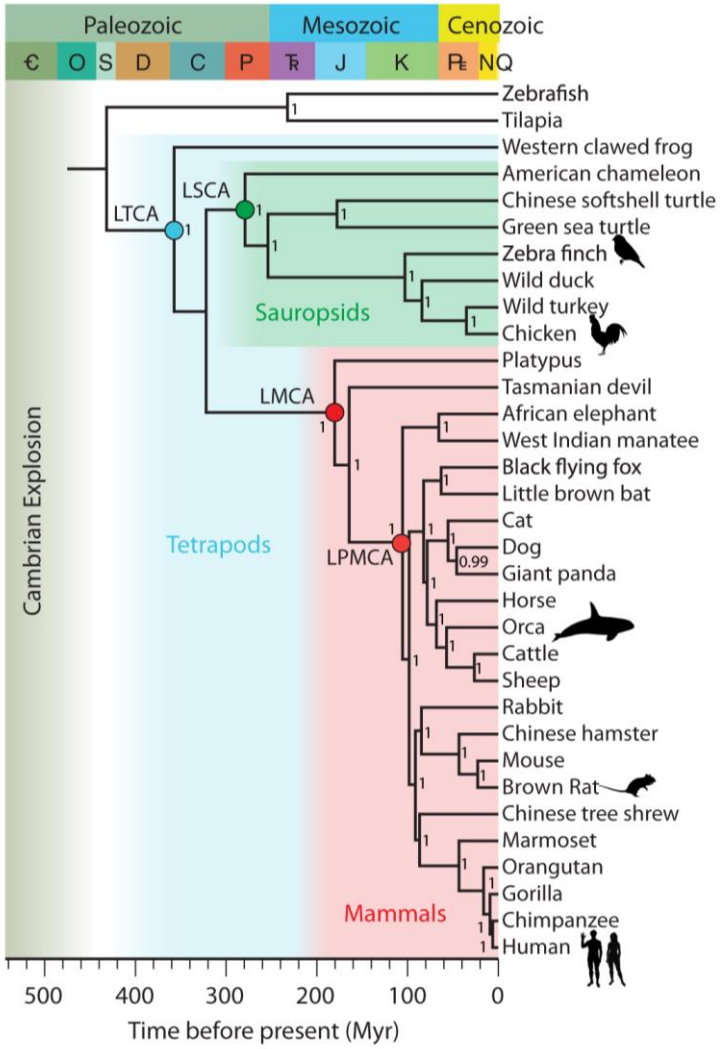


Figure 4.3. *Uncorrelated lognormal relaxed-clock chronogram of titin with geological time inferred with Bayesian inference. A total of 33 titin genes were used. The modern species studied are indicated by the animal outlines: zebra finch, chicken, orca, rat and human. The internal nodes LTCA, LSCA, LMCA and LPMCA were selected for resurrection and laboratory testing. Posterior probabilities for branch support are shown in the nodes. Geological times are shown in the upper bar. Outlines were retrieved from www.phylopic.org.*

A second set of sequences was used containing new entries that were added to databases at a later stage during this thesis to test the robustness of the tree. Most additions were in the *Sauropsida* supergroup and *Amphibia* clade that were less represented in our initial set. Phylogeny was constructed again with Bayesian inference **(Fig. 4.4)** with the same parameters mentioned above. Once again, the tree is consistent with the previous trees and with time tree of life. Most of the posterior probabilities are above 0.98, with the only exception of the node grouping the Tasmanian devil and the placental mammals (0.63) and the Western clawed frog with the rest of the amniotes (0.81), probably because the node is defined by only one species from one side.

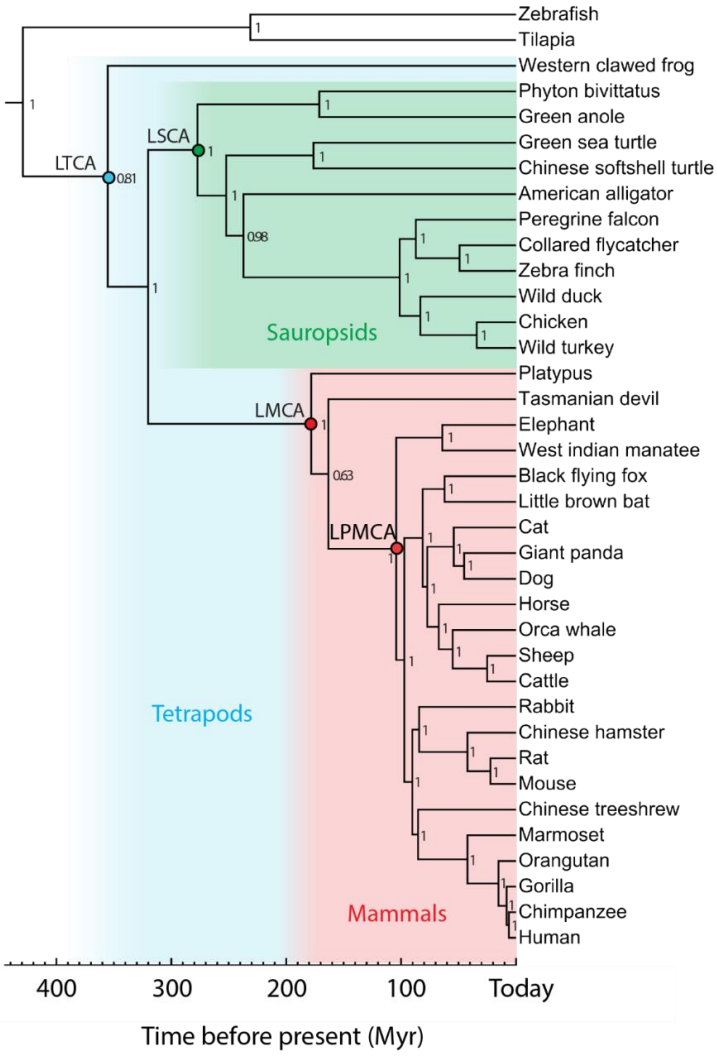


Figure 4.4. *Phylogenetic tree of titin using Bayesian inference. A total of 37 titin molecules from different animals were used to reconstruct the phylogenetic tree by Bayesian inference using MCMC. Posterior probabilities for each bifurcation are indicated.*

After inferring the phylogeny, we used Tracer to check the proper convergence of the posterior probability and the likelihood of the tree (**Fig. 4.5**). Despite in the first 10 million generations the posterior probability was fluctuating, from 15 million to the end it maintained stable. Ancestral sequence reconstruction was performed by maximum likelihood considering the tree of **Fig. 4.3** using PAML 4.8 [55, 57, 58] and incorporating a gamma distribution for variable replacement rates across sites and the JTT model.

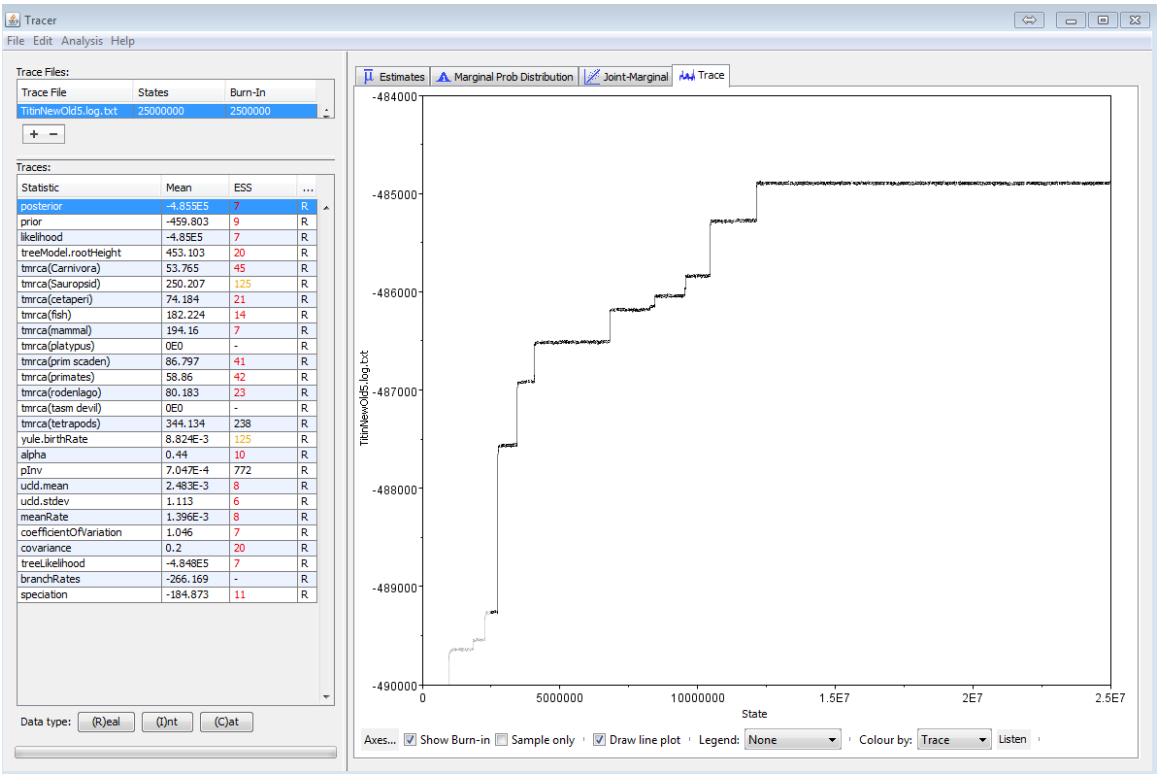


Figure 4.5. *Evolution of the posterior probability checked with Tracer.*

4.1.3. Dating and selection of ancestral nodes

Nodes were dated using multiple sources from the TTOL [54], as well as paleontological data [156]. Specifically, we have incorporated to calibrate the corresponding ancestral nodes in the tree the fossil records of the last *Carnivora* common ancestor (55 ± 7 Myr)[156], the last common ancestor of fishes (232 ± 61 Myr) [156], the last common ancestor of *Rodentia and Lagomorfa* clades (85 ± 19 Myr) [156] and the last

primate common ancestor (61 ± 5 Myr) [160-162]. The rest of the nodes were calibrated using TTOL and are available in **Appendix II**.

Several internal nodes for reconstruction of the most probable ancestral sequence were sampled. In particular, we have the chosen the following nodes for the further mutational and mechanochemical analysis: (**1**) the tree node corresponding to the last common ancestor of tetrapods (**LTCA**) that lived in the early Carboniferous Period (356 ± 11 Myr), (**2**) the node corresponding to the last common ancestor of sauropsids (**LSCA**) that is thought to have lived in the Permian Period (278 ± 14 Myr), (**3**) The last common ancestor of mammals (**LMCA**) that lived in our planet in the Jurassic (179 ± 38 Myr), and (**4**) the last common ancestor of placental mammals (**LPMCA**) from the mid-Cretaceous (105 ± 17 Myr).

The ancestral and extant titin sequences are listed in **Appendix III**. Posterior probability distributions across all sites for ancestral sequences are reported in **Fig. 4.6**. The overall posterior probability of the sites lies between 0.90 and 0.99. Logically, the higher overall posterior probability corresponds to LPMCA (the closest in time) whereas LTCA has the lowest probability values.

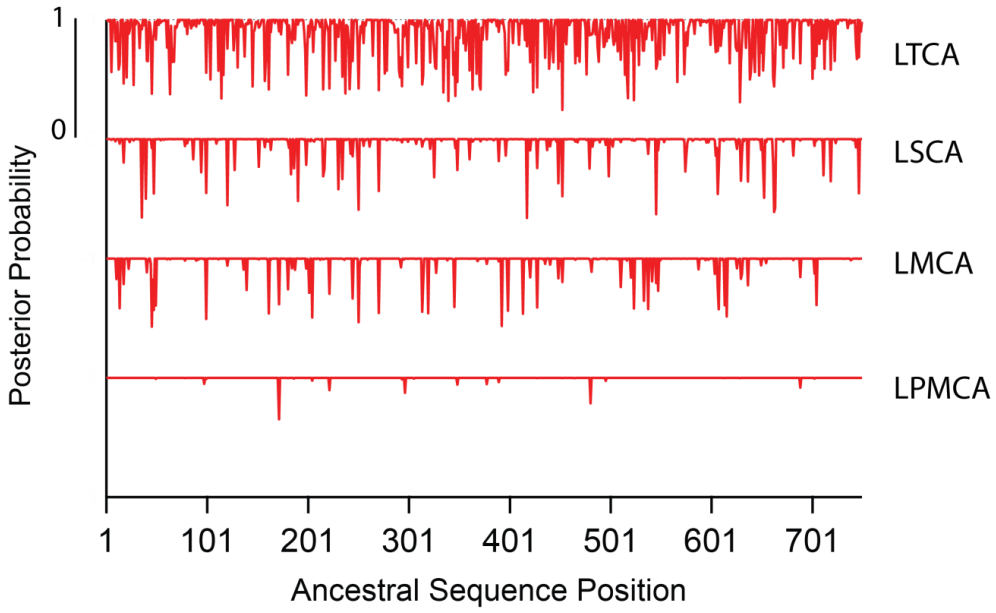


Figure 4.6. *Posterior probability distribution for each inferred residue of all ancestral titin fragments. The residue with the highest posterior probability is assigned at each position.*

An eight-domain fragment of titin encompassing domains I65 to I72 in the canonical human titin sequence (UniProt Q8W2Z42), and the homologous in other species were selected for resurrection and laboratory testing. This fragment is a good proxy of the elastic I-band region, located in the proximal tandem-Ig region of N2A skeletal titin and N2BA cardiac titin isoforms [163]. Up to six domains of this fragment (the I65-I70 segment, concretely) have been characterized in terms of structure [164] (**Fig. 4.7**) and mechanics [40]. Knowing the approximate structure and mechanical properties of these domains allows a direct comparison with the results obtained in this thesis. Also, the alignment of this fragment is well resolved suggesting that it is structurally conserved.

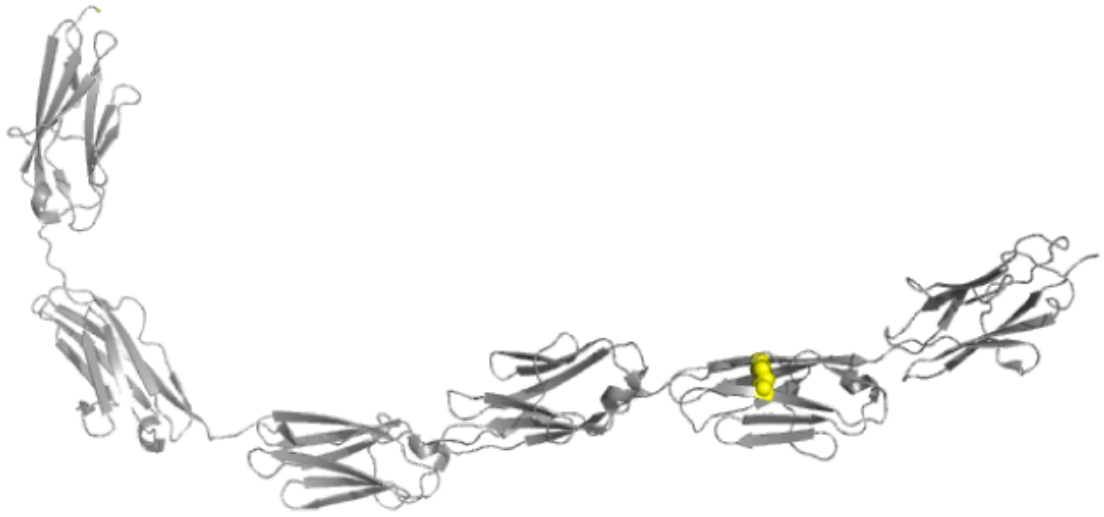


Figure 4.7. *Crystal structure of the I65-I70 fragment from the elastic I-band fraction of titin using small angle x-ray scattering [18].*

4.2. Ancestral reconstruction of myosin II

The main objective of this reconstruction was to establish a comparison between the main filaments of the sarcomere in terms of mutability and sequence conservations over the time. We brought back to life and analyzed the sequences from LTCA, LSCA, LMCA and LPMCA myosin molecules in terms of mutability and replacement of amino acids. The ancestral and extant myosin sequences to study are listed in **Appendix IV**.

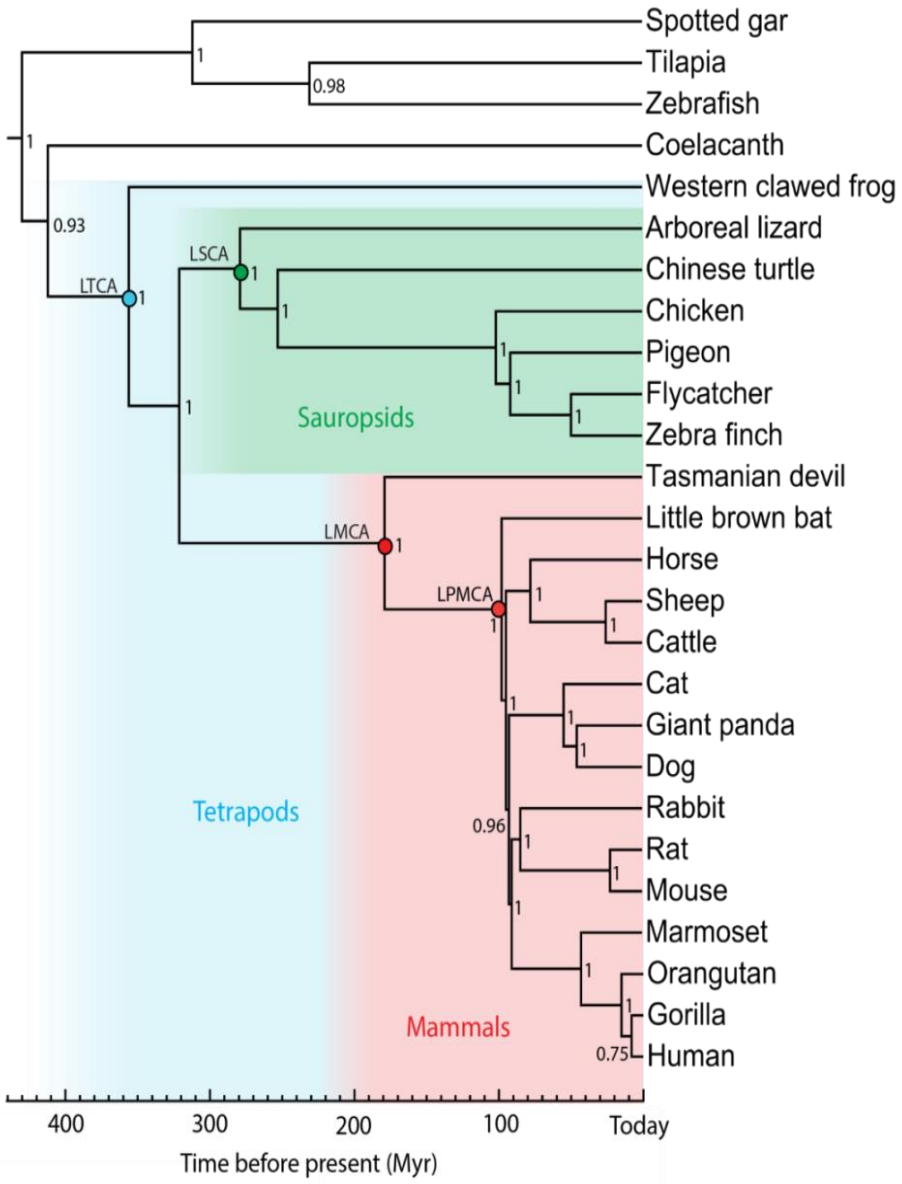


Figure 4.8. *Phylogenetic tree used for the reconstruction of the ancestral myosin. The tree was built by Bayesian inference using MCMC. The ancestors of interest are displayed with colored dots. Divergence times for ancestral nodes were collected from Time Tree of Life and fossil records again [54, 156]. Posterior probabilities for each bifurcation are indicated.*

Thus, we reconstructed the phylogenetic tree of myosin II from a similar pool of vertebrates as in the case of titin (**Fig. 4.8**) and inferred ancient myosin sequences. A total of 26 sequences were used (codes are listed in **Appendix I**). The myosin II tree was performed using only BEAST in a 12-core iMac computer. Tree Annotator was used to estimate a maximum clade credibility tree removing 30% of initial trees as burn-in. The

resulting tree is consistent with the trees generated for titin ancestral reconstruction and with Time Tree of Life. All posterior probabilities vary between 0.93 and 1, with the only exception of the ancestor between the human and the gorilla, which is 0.75.

4.3. Determination of mutation rate in sarcomeric proteins

To analyze the role that each of the main sarcomeric proteins (titin, myosin II, and actin) have played during evolution, we calculated the identity values of the ancestral proteins of interest compared with their descendants. Results clearly show that the mutations in titin are double than in myosin, the comparison of sequences of the inferred ancestral titin's I65-I72 with their modern counterparts yielded amino acid identities ranging from 72 to 92% (**Table 4.1a**). Concretely, the highest identity value corresponds to the pair human-orca whereas the lowest value is for the pair zebra finch-rat. In general, identities between species of the same clade are in the range of ~90% for both mammals and birds. Opposite to this, in animal pairs from distinct species, the identity decreases to values of around 70%. For identity values regarding ancestral species, the identity varies considerably depending on the closeness of the ancestral node with the extant species. These low identity values are significant considering that the structure and function of the domains is highly conserved between the species, suggesting that titin could have driven the muscle diversification of modern animals.

In the case of myosin, identities of modern forms reach over 90% for almost all the pairs of species (**Table 4.1b**). Here, the same phenomenon takes place for the identity values of intra and inter-clade pairs of species, but with different values. This time, identities for species of the same clade are approximately 95%, whereas identities for inter-clade species drop to ~90%. Again, the closeness phenomenon with the ancestral sequences is similar to that seen for titin. Finally, identity analysis for actin shows that is an extremely conserved protein, showing 99-100% identity across most living vertebrates. These values suggest that actin had very little influence on the molecular diversification of the sarcomere and, therefore, in the muscle diversification of modern animals. Due to its large conservation, ancestral reconstruction of actin could not be performed (**Table 4.1c**).

a. Titin

	Human	Orca	Rat	Zebra finch	Chicken	LTCA	LSCA	LMCA	LPMCA
Human		92	91	73	74	78	80	90	96
Orca	92		88	72	73	77	79	91	93
Rat	91	88		72	73	77	78	86	92
Zebra finch	73	72	72		91	76	86	77	74
Chicken	74	73	73	91		77	87	78	75
LTCA	78	77	77	76	77		86	84	80
LSCA	80	79	78	86	87	86		87	82
LMCA	90	91	86	77	78	84	87		93
LPMCA	96	93	92	74	75	80	92	93	

b. Myosin

	Human	Orca	Rat	Zebra finch	Chicken	LTCA	LSCA	LMCA	LPMCA
Human		-	94	91	92	91	93	95	97
Orca	-		-	-	-	-	-	-	-
Rat	94	-		92	92	92	94	96	96
Zebra finch	91	-	92		97	87	96	94	93
Chicken	92	-	92	97		92	95	94	93
LTCA	91	-	92	87	92		90	89	88
LSCA	93	-	94	96	95	90		96	95
LMCA	95	-	96	94	94	89	96		97
LPMCA	97	-	96	93	93	88	95	97	

c. Actin

	Human	Orca	Rat	Zebra finch	Chicken	LTCA	LSCA	LMCA	LPMCA
Human		-	100	99	100	*	*	*	*
Orca	-		-	-	-	*	*	*	*
Rat	100	-		99	100	*	*	*	*
Zebra finch	99	-	99		99	*	*	*	*
Chicken	100	-	100	99		*	*	*	*
LTCA	*	*	*	*	*		*	*	*
LSCA	*	*	*	*	*	*		*	*
LMCA	*	*	*	*	*	*	*		*
LPMCA	*	*	*	*	*	*	*	*	

Table 4.1. *Titin, myosin and actin sequence identities (%).* ***(a)*** *Titin has lowest identity with values that vary from 72% (zebra finch- rat) to 92% (human-orca).* ***(b)*** *Myosin identities fluctuate between 87 and 96% whereas actin* ***(c)*** *has the highest identity (99-100%). Orca myosin and actin sequences are not available at the time of this study. (*) Ancestral reconstruction of actin could not be performed due to the high identity of the sequences.*

Using ancestral and modern sequences, the mutation rates of titin I65-I72 and myosin II were also calculated for all the ancestral/modern pairs. This value is calculated as the number of mutations from the ancestral species to its extant descendants per 100

amino acids and per million years. Results reveal that mutation rates for titin's I65-I72 fragment (**Fig. 4.9a**) vary between 0.0381 (number of mutations per 100 aa/ Myr) for the pair LPMCA-human and 0.0763 for the pairs LPMCA-rat and LMCA-rat. In the case of myosin (**Fig. 4.9b**) the normalized mutation rate values differ between 0.0144 for LSCA-zebra finch pair and 0.0381 for LPMCA-rat. Remarkably, the highest mutation rate value corresponds to LPMCA-rat pair again. This occurrence could have significance, meaning that muscle physiology diversification of rodents, and mammals in general, could have happened faster in terms of evolutionary time. We should also take into consideration that the high number of hominids in both phylogenies could introduce a bias inferring the sequences of LMCA and LPMCA. This phenomenon would explain the differences between the mutation rates in different mammals. Comparing the average mutation rates on both proteins (0.0616 for titin and 0.0258 for myosin), titin has more than twice the mutation rate of myosin II. Again, we conclude that titin has contributed to the molecular diversification of the sarcomere more extensively than myosin.

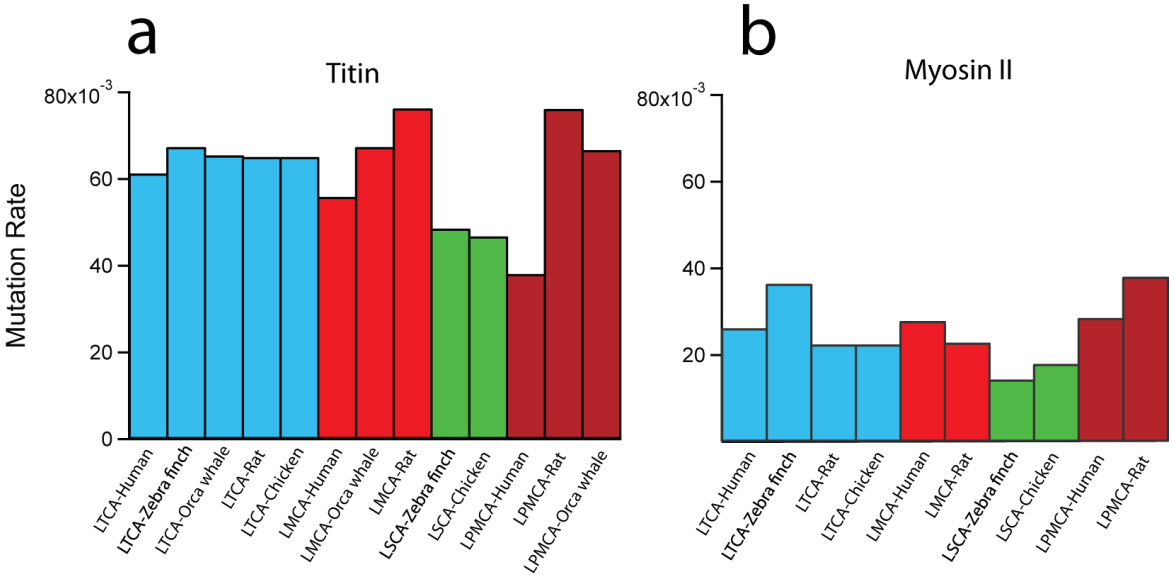


Figure 4.9. *Titin and myosin mutation rates between ancestral and extant species. Mutation rates are estimated as the number of mutations from the ancestral forms to their modern counterparts per 100 residues and per Myr.*

4.4. Role of cysteines in the evolution of muscles

With the aim of studying the behavior of each amino acid during the evolution of muscle filaments, we calculated the residue replacements in different zones of titin and in myosin II and weighted by their relative mutability. To achieve this, the occurrence of each residue type in titin fragments I65-I72 (proximal I-band), I88-I95 (distal I-band), I126-Fn90-94-I128 (A-band) as well as in myosin was estimated for human and LTCA. The replacement value for each residue refers to the increase or decrease in the number of a specific amino acid between LTCA and human for every construct. In order to weight these changes, we applied the concept of relative mutability [113], which estimates statistically the mutation probability of every amino acid, considering alanine as reference with a value of 100. Thus, the ratio between replacement and relative mutability was calculated for the

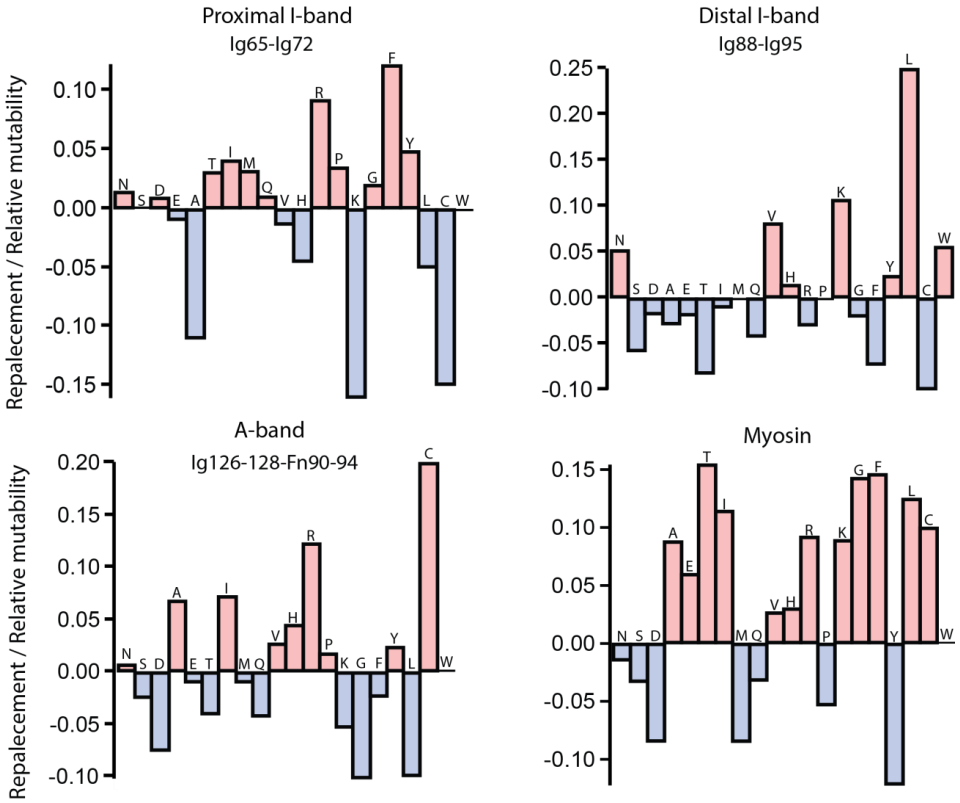


Figure 4.10. *Analysis of residue replacement and mutability for the transition from LTCA to human for different titin constructs and myosin. Cysteine residues have decreased their representation in the I-band of human titin much more prominently than any other residue. In the contrary, these residues have increased their number in the A-band of titin and myosin.*

4 constructs. In **Fig. 4.10** this ratio is represented for each residue. Positive values (red bars) indicate an increasing number of the specific amino acid in the transition from LTCA to human. Opposite to this, blue bars represent a decreasing number of residues.

Results show that cysteine (Cys) residues display a different mutability than expected during the evolution of titin I65-I72 and I88-I95 from LTCA to human, the fragments located in the I-band. Human titin fragments contain significantly fewer Cys residues than LTCA. Assignment of Cys residues in the ancestral sequences are supported by values of posterior probability close to 1 in the computed phylogenetic trees consistently. Contrary to this, a relatively prominent increase of Cys can be seen in the fragment located in the A-band rigid region (I126-Fn90-94-I128) and in myosin. This finding is significant given than Cys is one of the least mutable residues in proteins, second only to tryptophan [113]. Moreover, Wong J. H. and coworkers demonstrated that Cys residues are rarely lost once acquired [165-167], especially when they are disulfide bonded. This behavior of Cys conservation in the proteome is consistent in all the analyzed species [165].

4.5. Summary

In this chapter, we have reconstructed the phylogeny of the giant muscle protein titin with two different methods, maximum parsimony and Bayesian inference, being the results consistent for both cases. Moreover, a new set of sequences that was not available at the beginning of this study was added and the phylogeny computed again by Bayesian methods, being the tree correct too. The reconstruction of myosin was also carried out by the same method to compare the ancestors in terms of identity and mutation rate. We could not compute the phylogeny of actin due to its extremely high identity between species. The mutability results highlight that mutations in titin were double than in myosin. This finding is quite remarkable and it could mean that titin has been the evolutionary information carrier of the muscles since the Cambrian radiation. We also calculated the replacement of each amino acid in various parts of the A-band and I-band of titin and in myosin. Results reveal that Cys are only lost in the elastic region of titin, suggesting that these residues are directly related with the elastic properties and passive force of muscles.

Chapter 5: Experimental results

In this chapter, we show the nanomechanical behavior of nine I65-I72 titin constructs (4 ancestral plus 5 extant) analyzed with the smFS in in the force extension mode. To confirm the obtained results, two additional experiments were also performed: a single-molecule disulfide reduction assay to detect cryptic disulfides with force clamp mode of the smFS, and a biochemical experiment capable to detect free thiols.

5.1. smFS force extension experiments

5.1.1. Mechanical stability of titin constructs

Four ancestral (LTCA, LSCA, LMCA, and LPMCA) and five extant (human, brown rat, orca, chicken and zebra finch) I65-I72 fragments from representative modern vertebrate species covering different clades in the tree were purified under equal oxidative stress conditions to overcome limited disulfide bond formation in the host. To investigate the mechanical properties of all I65-I72 titin variants smFS was performed by mechanically stretching the proteins at a constant speed of 400 nm/s (**Fig. 5.1a)**. The stretching of titin domains leads to sawtooth patterns in force versus extension recordings, in which each peak represents the mechanical unfolding of an individual domain (**Fig. 5.1b to 5.1e top**).

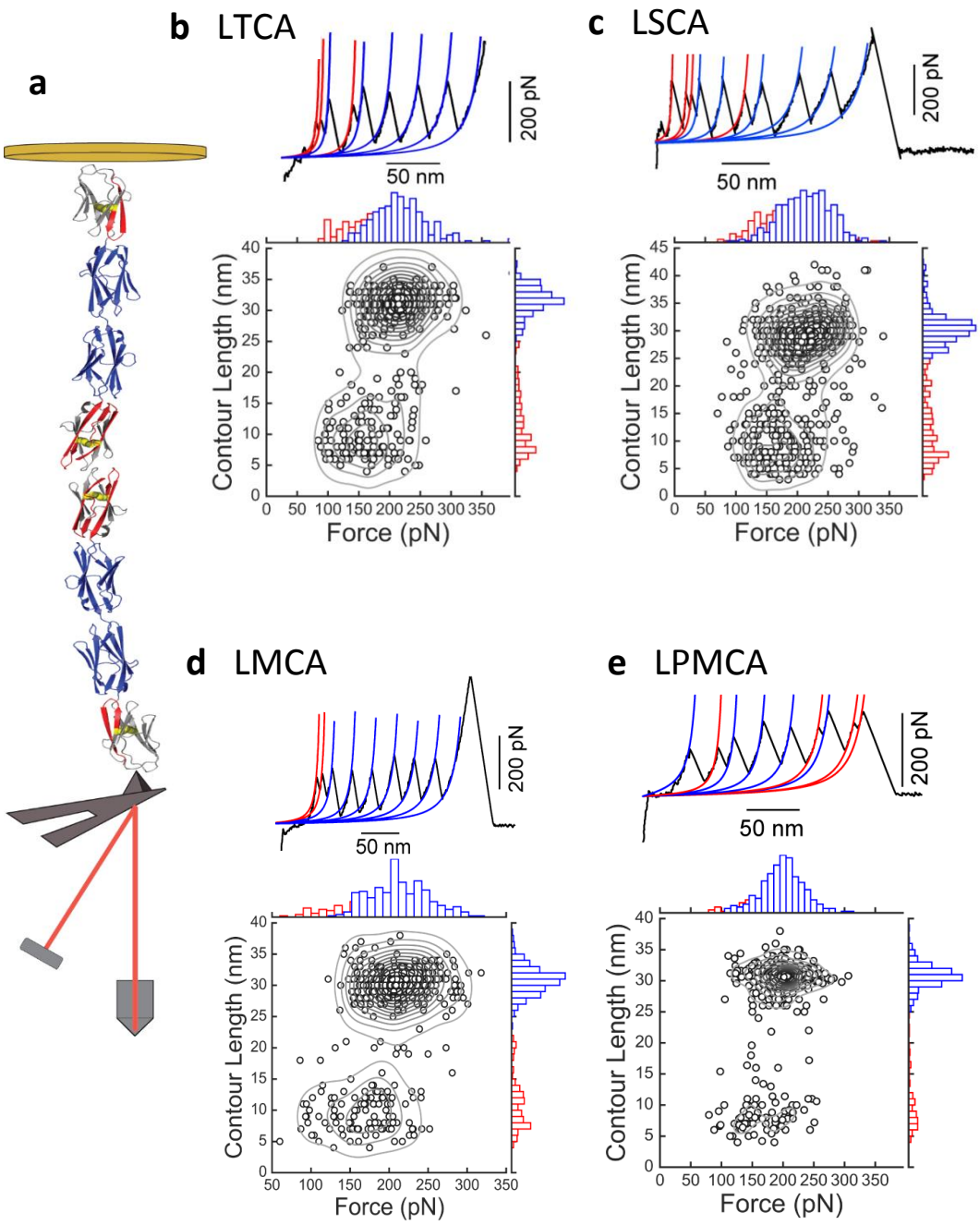


Figure 5.1. ***(a)*** *Schematic representation of a single-molecule experiment using the smFS (not to scale). Disulfide bonded domains are shown in red/grey with the cysteine highlighted in yellow. Non-disulfide bonded domains are shown in blue. The protein is mechanically stretched between a cantilever tip and a gold-coated surface.* ***(b to e - Top)*** *Experimental traces of LTCA, LSCA, LMCA and LPMCA I65-I72 titins. The unfolding of domains is monitored as a sawtooth pattern of force versus extension peaks. The worm-like chain model was used to fit the data. Fits to fully extended and disulfide bonded domains are shown in blue and red, respectively.* ***(b to e - Down)*** *Scatter plot of contour lengths and unfolding forces for LTCA (n=374), LSCA (n=614), LMCA (n=407) and LPMCA (n=366) I65-I72 titins, with kernel density estimates shown as lines. The histograms are shown on each axis.*

The analysis of the traces using well established procedures [168] (see **Chapter 3: Experimental methods**, section **3.2.5**) allows to determine the mechanical stability and contour length of the constituent domains.

For all the ancestral (**Fig. 5.1b to 5.1e down**) an extant (**Fig. 5.2a to 5.2e down**) variants tested two distinct populations of peaks are observed. The first one is a compact population that has contour lengths of ~30 nm and corresponds to fully extended domains of about 90 residues. The second one is more disperse, displays lengths of 5 to 20 nm and represents disulfide-containing domains, which make the contour length of the extended peptide shorter, as expected because a part of the domain is arrested by the disulfide. The spread of this population is due to the different position of the cysteines in the different domains [169], as it is shown in **Appendix V**. This is quite significant since disulfide bonds in titin have been suggested to participate in titin mechanical regulation [170].

Overall, the tendencies related to the unfolding forces are different for the two populations. In the case of fully extended domains, the average unfolding force ranges from 180 to 218 pN (**Table 5.1**), depending on the variant, being the lowest stability for orca titin and the highest for LTCA titin domains. Specifically, the average unfolding forces can be grouped in in three well delimited groups:

- A first group with unfolding forces in the range of 210-220 pN. LTCA (218.12 pN), LSCA (213.04 pN), LMCA (212.97 pN) and zebra finch (215.33 pN) titins belong to this group.
- A second group with average unfolding forces around 200 pN that encompasses LPMCA (199.17 pN) and chicken (197.04 pN) titins.
- A third group with unfolding forces between 180-190 pN. All the extant mammals studied are in this group: Rat (187.50 pN), human (184.90 pN) and orca (180.38 pN).

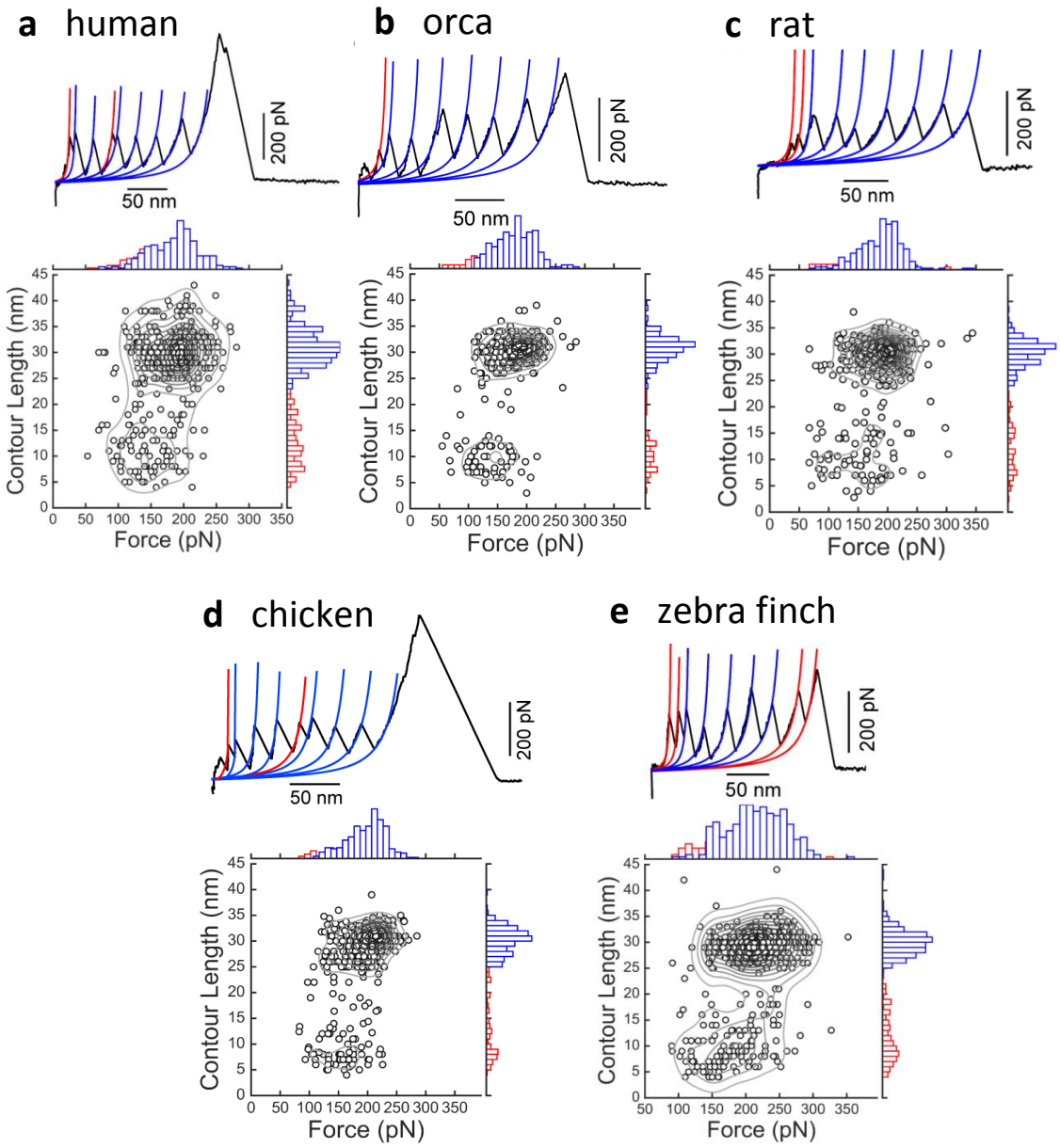


Figure 5.2. *Representative experimental trace and scatter plot of contour lengths and unfolding forces of I65-I72 titins of the five extant species studied: human (n = 347), orca (n = 263), rat (n = 341), chicken (n = 375) and zebra finch (n = 409).*

All the average contour lengths are in the range of 30 nm. Standard deviation (SD) values are in the range of 30-40 pN for unfolding forces and 2-3 nm for contour lengths, which is in accordance with the error values for this type of experiments. Results for this first population of peaks are in agreement with previous mechanical characterization of

I65-I70 purified in reducing conditions where no disulfide bonds could be established [40]. This trend can be visualized more clearly in the cumulative histograms of the unfolding forces (**Fig. 5.3**). The three groups referred above can be seen: the extant mammals first (orange), LPMCA and chicken later and, finally, the zebra finch and the rest of the ancestors.

Species	Force (pN)	SD (pN)	Contour Length (nm)	SD (nm)	n
Non-disulfide bonded domains					
LTCA	218.12	43.67	31.1	2.2	234
LSCA	213.04	40.59	30.0	3.0	376
LMCA	212.97	37.10	30.3	2.3	284
LPMCA	199.17	34.12	30.6	2.2	286
Zebra finch	215.33	42.55	29.5	2.6	281
Rat	187.50	37.75	30.4	2.4	261
Human	184.90	37.86	30.7	3.8	248
Orca	180.38	34.42	30.5	2.4	203
Chicken	197.04	32.69	29.8	2.6	293
Disulfide bonded domains					
LTCA	168.65	45.07	10.9	4.7	140
LSCA	179.22	49.74	11.5	5.7	238
LMCA	168.89	44.26	10.3	4.1	123
LPMCA	167.37	42.04	9.8	5.0	80
Zebra finch	182.72	47.65	10.7	4.6	128
Rat	156.92	52.14	11.5	4.6	80
Human	150.11	38.84	12.6	5.2	99
Orca	134.41	38.34	10.5	4.2	60
Chicken	158.10	37.62	11.7	5.8	82

Table 5.1. *Unfolding force (pN) and contour length (nm) average values of the studied extinct and extant species for non-disulfide bonded and disulfide bonded titin I65-I72 domains. Standard deviation (SD) values for each titin are shown. n values are indicated for each species.*

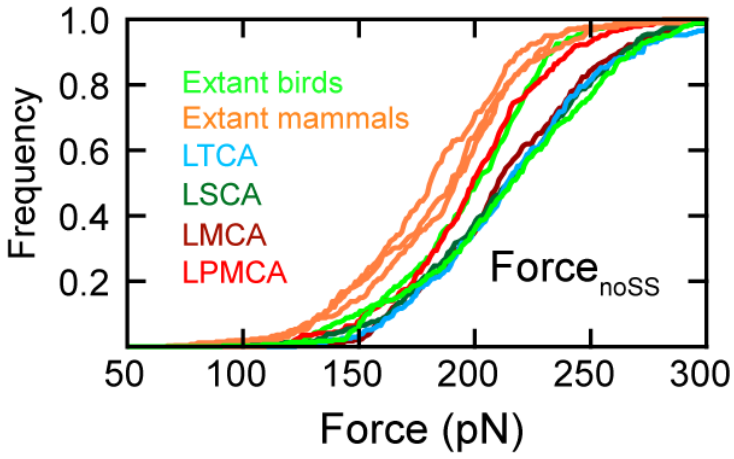


Figure 5.3. *Cumulative histogram of mechanical unfolding force for domains that do not contain disulfide bonds.*

For disulfide bonded domains, we determined an average stability ranging from 134 to 182 pN depending on the variant (**Table 5.1)**, with orca titin having the lowest and zebra finch the highest. Here, the average unfolding forces are not grouped in three clusters as in the case of non-disulfide bonded domains. Instead, the values are continuously distributed. However, the overall trends observed in the reduced form of the protein remain robust: the lowest mechanical stabilities correspond to extant mammals, chicken and LPMCA titin are in a mid-range again and the highest average unfolding forces are related to the rest of the ancestors and to zebra finch, which is perfectly consistent with the values reported for non-disulfide bonded domains. This time, the SD values range between 40-50 pN for unfolding forces and 4-6 nm for contour lengths. These values are sensibly higher that the ones for the non-disulfide bonded population, but it agrees with the spread of this population, that could be related to the isomerizations that most of the domains can suffer according to the positions that cysteines occupy in the domains (see **Appendix V**). Again, this tendency in the unfolding forces is better captured in the cumulative histogram (**Fig. 5.4**).

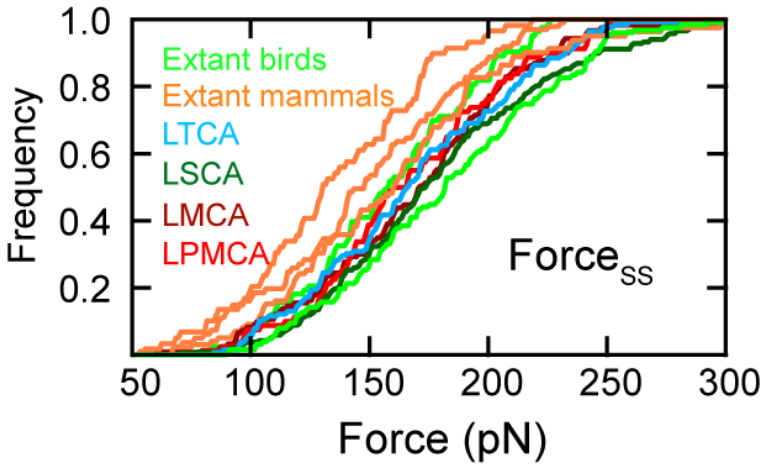


***Figure 5.4**. Cumulative histogram of mechanical unfolding force for disulfide bonded domains.*

5.1.2. Percentage of disulfide bonded domains

The amount of disulfide bonded domains is another piece of information that can be obtained from force extension experiments. By counting the number of domains that contain disulfide bond and comparing it with those that fully extend, it can be deduced that, in general, ancestral proteins LTCA and LSCA have the highest proportion of S-S bonded domains (**Fig. 5.5**).

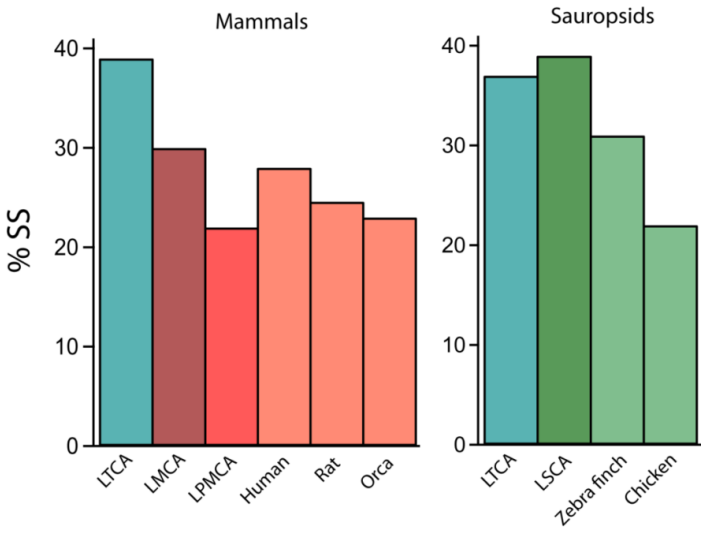


***Figure 5.5.** Percentage of disulfide bonded domains detected in force-extension traces. LTCA and LSCA titin fragments are the ones presenting higher percentage of disulfide bonds in both, mammals and birds. In general, modern animals display fewer disulfide bonds than their ancestors.*

Thus, we conclude that the content of disulfide bonds seem to have decreased for modern species. In fact, the content of disulfide bonds seems to be related to the mechanical stability. More stable titin forms imply more disulfide bonded domains. In the plot of the unfolding forces versus the percentage of disulfide bonded domains (**Fig. 5.6**) we see these forces in both types of domains increase when the percentage of experimental disulfide bonds is higher, following a linear relationship. This is a non-trivial result since it suggests that the amount of (in general) weaker disulfide bonded domains in titin is somehow compensated for a subtle modulation in the strength of both the disulfide bonded and non-disulfide bonded domains depending on the species.

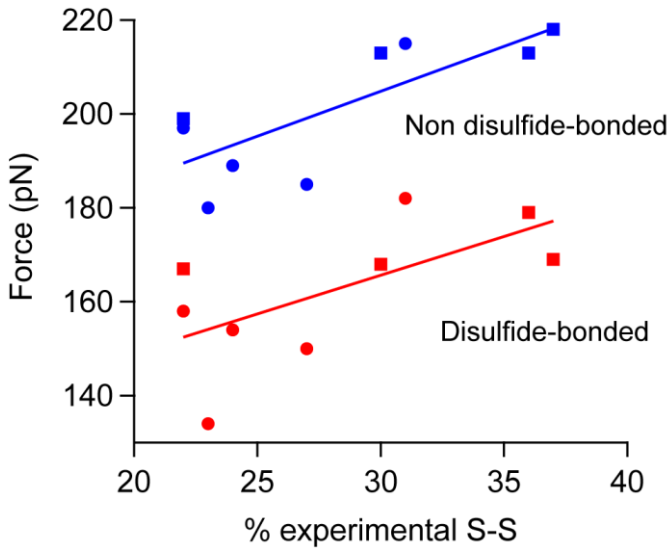


***Figure 5.6**. Unfolding force versus observed disulfide bond percentage. Domains that do not contain disulfide bonds are represented in blue, whereas those showing disulfide bonds are represented in red. Ancestral and extant species are shown with square and circles, respectively.*

5.1.3. Establishing paleomechanical trends

With the aim of searching for possible trends in the evolution of titin unfolding mechanics for various kinds of animals, we plotted the mechanical stabilities for both types of domains versus the age of each species, both for bird (**Fig. 5.7a and 5.7c**) and mammalian lineages (**Fig. 5.7b and 5.7d**). In the case of birds the stability remained stable

for zebra finch and slightly decreased for chicken with respect to LTCA and LSCA in non-disulfide bonded domains. For disulfide bonded domains, we find that zebra finch's mechanical stability is slightly higher than for its ancestors, while for chicken we observe a marginal decrease. In the case of mammals, a clear paleomechanical trend by which stability has decreased over time can be observed in both types of protein domains. This decrease is especially acute for all extant mammals in non-disulfide bonded domains and for orca titin in disulfide bonded domains.

Specifically, while LMCA practically preserved its mechanical stability for non-disulfide bonded domains compared to LTCA during 180 Myr, the extant mammals suffered a drop on their mechanical stability of around 15% in the same period. This singularity also happened for orca titin in disulfide bonded domains. Interestingly, in both birds and mammals, the animals with larger body sizes have lower mechanical stabilities of titin domains with and without disulfide bridges. This phenomenon is discussed in detail in **Chapter 6: Discussion**

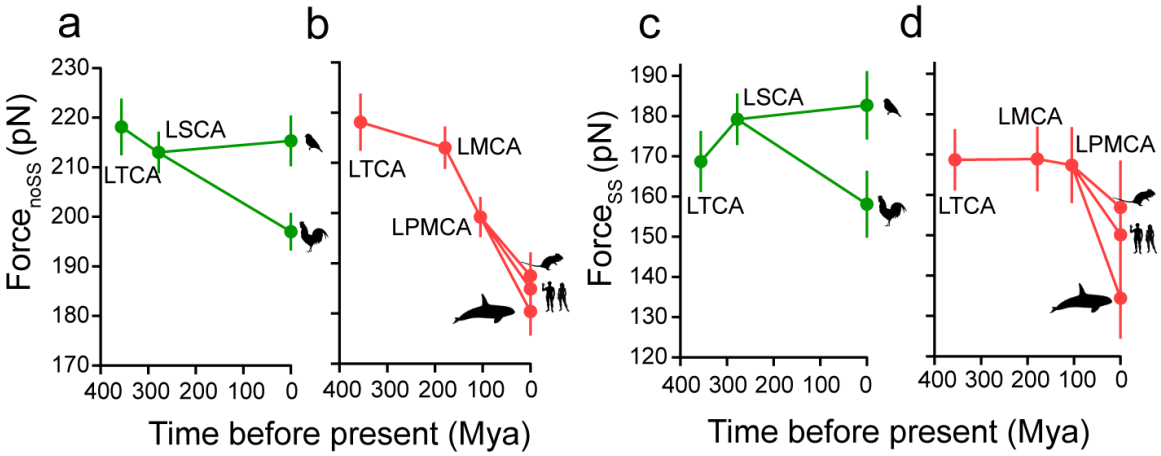


Figure 5.7. *Mechanical stability versus geological time of* ***(a)*** *non-disulfide bonded domains for birds and ancestors and,* ***(b)*** *for mammals and ancestors.* ***(c)*** *Mechanical stability of disulfide-bonded domains for birds and ancestors and,* ***(d)*** *for mammals and ancestors versus geological time. Error bars indicate the 95% confidence intervals for the standard error of the mean.*

.

5.2. smFS force clamp experiments

To confirm the results obtained from the smFS in the force extension mode, we performed single-molecule experiments in the force-clamp modality to capture reductions of disulfide bonds. In this mode, the force applied to the protein can be controlled thanks to PID controller. Force clamp has been widely used to monitor the unfolding of protein in a controlled manner [171]. In this case, the mechanical unfolding is monitored as an increment in length versus time.

5.2.1. Single-molecule disulfide reduction assay

Force clamp techniques have been used in previous works to demonstrate the kinetics of disulfide bond reduction under force by thioredoxin enzymes (Trx) [172, 173]. Thioredoxin is an oxidoreductase enzyme that controls the redox balance in cells by reducing disulfide bonds. The reduction of disulfide bonds by Trx is a force-dependent reaction that can be readily monitored and quantified [173]. By applying force to a disulfide-bonded titin domain one can trigger the unfolding of the domain up to the disulfide bond that becomes exposed to the solution. If Trx is present in the solution the disulfide bond can be reduced and the sequestered residues behind the disulfide bond are released giving rise to an extra extension of the polypeptide chain. With this assay, one can quantify disulfide bonds that, like in the case of many titin Ig domains, are cryptic, i.e., they require mechanical exposure to be reduced. Non-cryptic disulfide bonds will be reduced without the need of mechanical exposure. (**Fig. 5.8**).

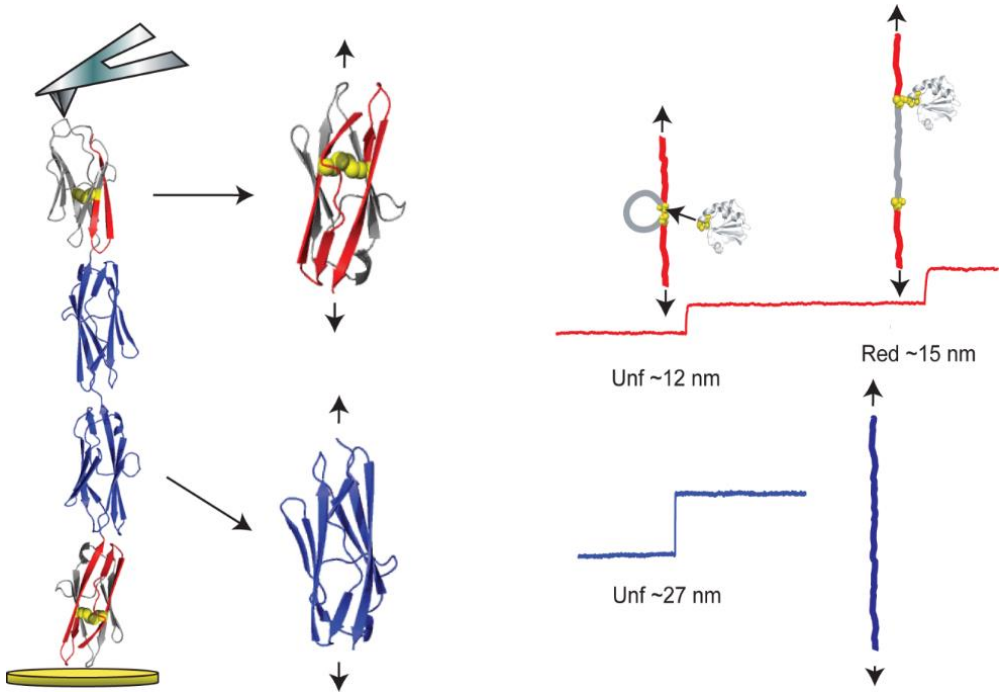


Figure 5.8. *Schematics of force-clamp experiment for detection of single disulfide reduction events. Disulfide bonded domains (red) show a two-step unfolding pattern. The first step (~12 nm) corresponds to the unfolding of the beta sheets that are not trapped in the disulfide bond while the second step (~15nm) shows the unfolding of the rest of the protein after the reduction of the disulfide bond cause by thioredoxin. Not disulfide bonded domains have a single-step (~27nm) unfolding pattern that represents the stretching of the whole domain.*

We have applied this test to LSCA and human titin fragments in the presence of Trx. First, a pulse of force of 135 pN during 2 s that triggers the unfolding of all the domains is applied (**Fig. 5.9 and 5.10 left**). The unfolding of the domains is monitored as a staircase of ~27 nm per step for reduced domains, and shorter steps between 5-20 nm for disulfide-bonded domains (**Fig. 5.9 and 5.10 insets**). The expected length of extended disulfide-containing domains varies due to the different position of the cysteines in the sequences (see **Appendix V**).

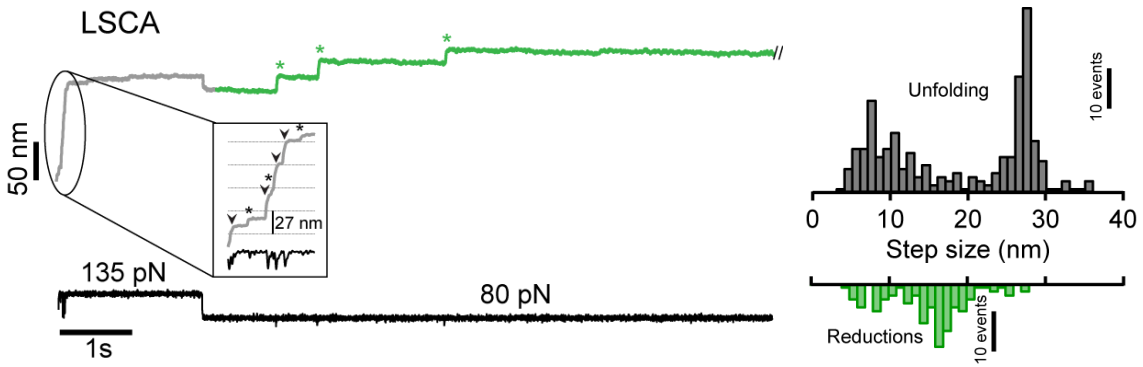


Figure 5.9. ***(Left)*** *Force-clamp experiments for detection of disulfide bond reductions catalyzed by thioredoxin enzymes. Experimental force-clamp trace of LSCA titin. The unfolding of non-disulfide bonded domains is indicated in the inset with arrows whereas the unfolding of disulfide bonded domains up to their S-S bond is shown in the inset with asterisks. The reduction events are monitored at a force of 80 pN, indicated in the green line with green asterisks. (****Right****) Histograms of step size for unfolding events (grey, n=268) and disulfide bond reductions by Trx (green, n=104) in LSCA titin.*

After the unfolding force pulse, we quench the force to 80 pN for 20 s to monitor disulfide bond reductions as steps within the range 5-20 nm for LSCA titin (**Fig. 5.9**) and 15-20 nm for human titin (**Fig. 5.10**). The disulfide bonds can be reduced by Trx enzymes present at 10 µM concentration. Again, the length attributed to reduction events will be different for each domain due to the position of the cysteines. Also, some domains have more than two cysteines which may also imply the possibility of isomerizations [174]. All possible disulfide bond combinations have been estimated and are shown in **Appendix V**. Results show that both disulfide bonded unfolding events and their corresponding reduction events are much higher in LSCA titin that in human titin (**Fig. 5.9 and 5.10 right**). A histogram of the observed reduction events for each construct demonstrate that in LSCA it is common to observe up to 4 reduction events, being 2 reductions the most probable. Contrary to this, for human titin it is common to observe only one reduction event (**Fig. 5.11**), being two reduction a rarely observed phenomenon.

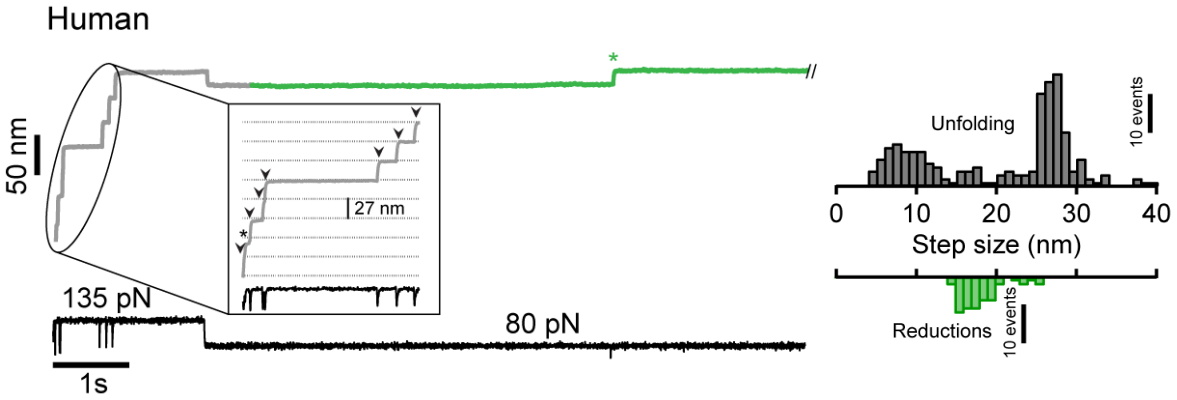


Figure 5.10. ***(Left)*** *Experimental force-clamp trace of human titin. Again, the unfolding of non-disulfide bonded domains is shown in the inset with arrows whereas the unfolding of disulfide bonded domains up to their S-S bond is shown in the inset with asterisks.* ***(Right)*** *Histograms of step size for unfolding events (grey, n=220) and disulfide bond reductions by Trx (green, n=47) in human titin.*

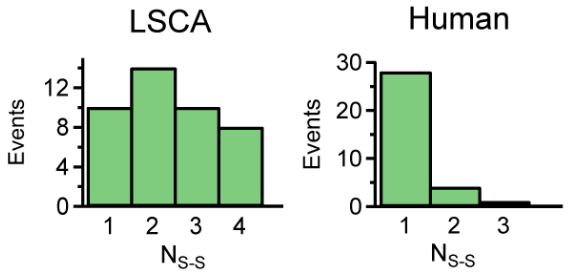


Figure 5.11. *Histogram of the number of reduction events per trace for LSCA and human titin.*

Experiments in absence of Trx were also performed as negative controls following the same force clamp protocol. The first pulse at 135 pN for 2 seconds results in the same staircase pattern (**Fig. 5.12a and 5.12c insets**) for both LSCA and human titin, being the events that correspond to disulfide bonded domains much more common in LSCA titin. Specifically, in LSCA titin up to 4 short events can be seen again, whereas in human titin the most common pattern is 1 short event. As we expected for this control experiments, no reduction events were observed (**Fig. 5.12a and 5.12c**) for both LSCA and human titin. All the events for both species were plotted as step size histograms (**5.12b and 5.12d**). Here, it can be clearly seen that the amount unfolding events for disulfide bonded domains is much higher for LSCA than for human titin. Thus, force-clamp experiments confirm that

LSCA titin contains more disulfide bonds than the human titin, supporting conclusions obtained using force extension for all I65-I72 fragments.

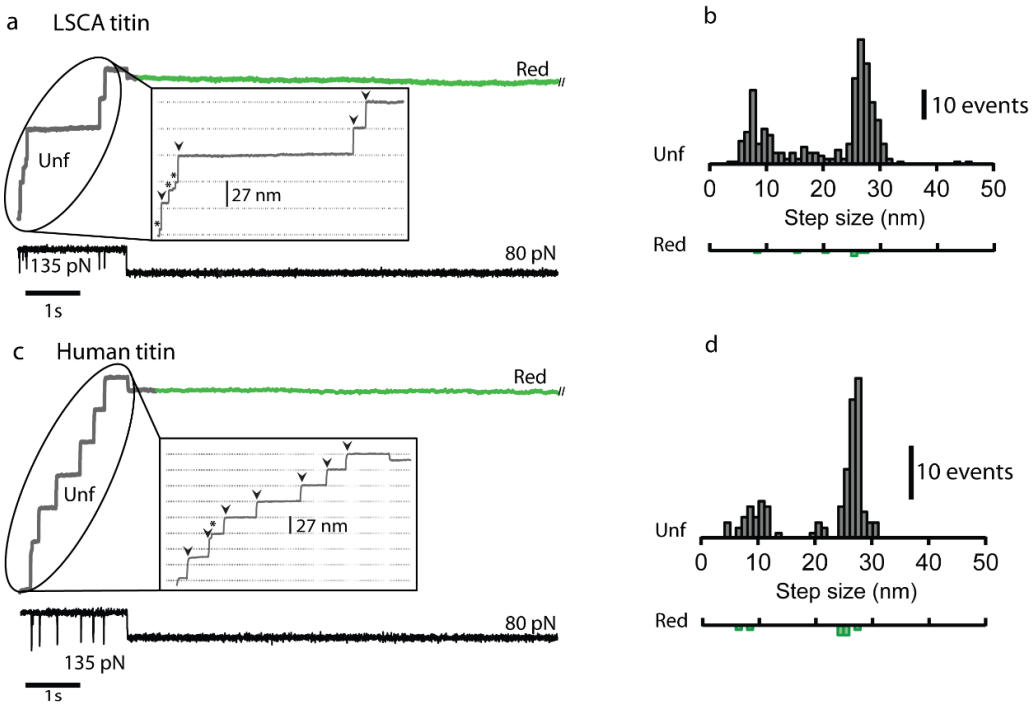


Figure 5.12. *Force-clamp experiment for detection of disulfide reduction events in the absence of thioredoxin enzymes. (**a**) Experimental trace of LSCA titin with seven steps representing each immunoglobulin domain visualized within the first pulse at 135 pN (grey). The fully unfolded domains are marked with an arrow (~27 nm) whereas disulfide bonded domains are marked with an asterisk (~5-20 nm). No steps are detected in the second stage at 80 pN (green) that was kept for 20 s (n=267). (**b**) Two populations of steps can be observed in the histogram. The histogram below shows steps captured in the 80 pN pulse where reductions are generally observed in the presence of Trx. (**c**) and (**d**) Experimental trace and steps histograms for human titin, respectively. In this case eight unfolding events are shown from which only one shows step size corresponding to a disulfide bonded domain. In the traces from human titin is common to observe 1 or 2 disulfide bonded domains. In both cases the histograms of the unfolding pulse resemble the histograms of the force-extension experiments, as expected (n=132).*

5.3. Cysteine quantification biochemical assays

To provide independent measurements of cysteine oxidation, we used a biochemical assay that detects oxidized cysteines [154] by labeling them with the fluorophore monobromobimane (mBBr). In this assay, reduced thiols in the protein are alkylated with an excess of N-ethylmaleimide in denaturing conditions. After running polyacrylamide electrophoresis, the gel is used as a reaction chamber to reduce oxidized thiols by incubation with dithiothreitol (DTT) and subsequent reaction of the newly reduced thiols with mBBr. The resulting fluorescence signal is proportional to the number of oxidized thiols in the sample.

Results show that LTCA and LSCA proteins contain more oxidized cysteines than rat and human proteins (**Fig. 5.13**). In our biochemical assays, we also included a control $(I91\text{-}32/75)_8$ protein (formerly I27) that can be produced in reduced (no disulfides) or oxidized (1 disulfide per domain for a total of 8 disulfides) forms and whose oxidation status can be determined unambiguously by smFS [155]. Since the size of the control protein and the I65-I72 fragments is comparable, we used the normalized fluorescence

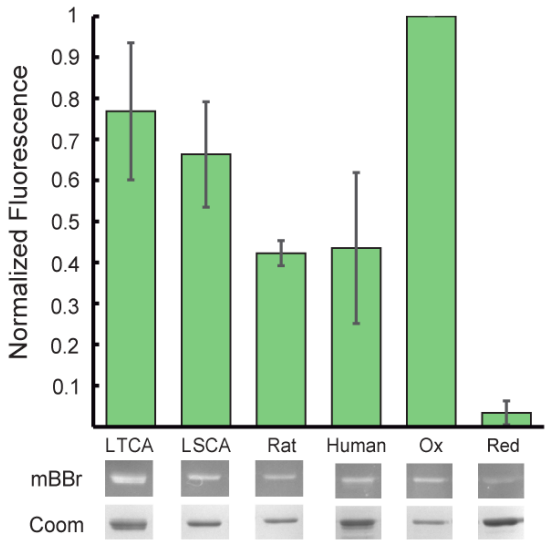


Figure 5.13. *In-gel determination of oxidized thiols for LTCA, LSCA, rat (R) and human (H) I65-I72.Fluorescent bands resulting from the labeling of oxidized thiols with mBBr were normalized by the total quantity of protein as assessed by Coomasie staining and densitometry. The oxidized (Ox) and reduced (Red) versions of* $(I91\text{-}32/75)_8$ *were used as controls. Mean values ± S.D. of at least three independent experiments are represented.*

signals to estimate the number of disulfides in LTCA (6), LSCA (5), human and rat (3-4). This independent method of determining the population of disulfide bonds confirms the trend observed in smFS experiments for more disulfides in the ancestral proteins (**Fig. 5.13**). There are, however, small discrepancies between the exact number of disulfides from the different methods that probably reflect contributions of terminal cysteines needed for attachment in smFS experiments or other forms of oxidation different from disulfides, such as sulfenylation induced by treatment with H_20_2.

5.4. Summary

smFS results in the force extension mode reveal that for all the titin constructs studied there are two different populations: a well-defined and compact one around 30 nm that corresponds to the length of a fully extended Ig domains and another more disperse population in the range of 5-20 nm that is consistent with length of disulfide bonded domains considering the different positions of the cysteines and the part of the Ig domains that they sequester. The mechanical stability of the fully extended domains varies between 180 and 220 pN, whereas the disulfide bonded domains have lower unfolding forces in all the cases (130-180 pN). The percentage of disulfides is in the range of 20-40 % for all the species. Considering these results, different paleomechanical trends have been stablished for bird and mammal lineages. While birds tend to maintain or slightly reduce their mechanical stability, we detected a clear drop in the mechanical properties of mammal Ig domains. The percentage of disulfide bonded domains has decreased over time for both lineages. To corroborate the results related to the disulfide percentages two additional experiments were carried out. The first one is a single-disulfide reduction assay using the smFS on its force clamp mode and Trx as a reducing agent. The second one is a biochemical assay to detect oxidized cysteines and, therefore, disulfide bonds too. Both experiments were consistent with previous data from smFS force extension, detecting more disulfide bonds in ancestral species than in extant ones.

Part IV

Chapter 6: Discussion

The aim of this thesis was to study the the mutational and mechanical history of titin since the Cambrian radiation util nowadays. For that purspose we have used a combination of phylogenetic methods combined with state-of-the-art biophysical and biochemical techniques. Results show that the phylogenies computed for titin and myosin are consistent with previous paleontological data. Reconstruction of titin is specially remarkable, since it is the largest protein whose phylogeny was inferred to the best of our knowledge. Due to high computational costs and difficulties managing long sequences, protein phylogenies rarely exceed 1000 amino acids. Our computational study exceeds this limit in more that an order of magnitude. In this respect, the striking differences found in the mutation rates of the three main muscle filament proteins suggest that titin has been the main evolutionary information carrier not only in terms of genotype, but also phenotype. Moreover, the eight-domain fragments resurrected and studied are significant because they belong to the critical elastic region of titin. Studying additional segments of titin as well as comparing features such as folding kinetics of ancient and modern domains will be interesting to gain a complete understanding of the molecular elements that have driven the molecular evolution of titin and its connection to muscle physiology.

One of the central questions addressed in this thesis is understanding the relationship between disulfide bonds and mechanical stability, and how these elements are linked to muscle physiology. The existence of disulfides in titin was first proposed

following identification of a disulfide bond in the crystal structure of domain I1, and the observation that many domains in titin contain proximal cysteines that can engage in disulfide bonds [175]. The crystal structure of rabbit I65-I70 (PDB code: 3b43) shows that all 6 domains can establish disulfide bonds, although only one is detected experimentally. Indeed, studies of disulfide bonds in titin have been limited by deficient disulfide formation in recombinant expression systems, sometimes leading to contradictory results [176]. Nevertheless, experimental evidence of disulfide formation in native titin is still lacking. Results in titin I65-I72 domains reveal a systematic decrease in their mechanical stability when they are disulfide bonded (**Table 6.1**). This drop is significant for all the species studied and varies between 15 and 25%, being the orca titin the one with the highest decrease (25.49%) and the zebra finch with the lowest one (15.14%).

	F_{noSS} (pN)	F_{SS} (pN)	ΔF (%)
LTCA	218.12	168.65	-22.68
LSCA	213.04	179.22	-15.87
LMCA	212.97	168.89	-20.70
LPMCA	199.17	167.37	-15.97
Zebra finch	215.33	182.72	-15.14
Rat	187.50	156.92	-16.31
Human	184.90	150.11	-18.81
Orca	180.38	134.41	-25.49
Chicken	197.04	158.10	-19.76

Table 6.1. *Mechanical stability of non-disulfide bonded (F_{noSS}) and disulfide bonded (F_{SS}) domains. The change between them (ΔF) is shown in percentage.*

In a recent work of the Nanobiomechanics lab, the effect of disulfides in the overall mechanical stability of the $(I91_{G32C\text{-}A75C})_8$ human cardiac titin mutant polyprotein [169] was measured using smFS in the constant speed mode. The reason to study this protein in terms of mechanical stability was to understand the reduction of the unfolding forces that we have found in all the wild-type titin constructs studied in this thesis and see if this phenomenon can be reproduced in one of the most studied Ig mutant domains. The main advantage of

the $(I91_{G32C-A75C})_8$ polyprotein with respect to the WT sequences studied in this thesis is that the experiment is more controllable, since all of the domains in the polyprotein share the same location of the disulfide bonding cysteines. This drastically reduces the spread in forces and extensions found for the WT, naturally occurring proteins. The histograms of the unfolding force and the scatter plots of the force and contour lengths are shown in **Fig. 6.1**. The reduced state has a mechanical stability of 196±30 pN, while the oxidized form of the protein displays an unfolding force of 171±29 pN. Hence, on average, there is a subtle decrease (25 pN or 13%) in the mechanical stability of the protein upon disulfide bonding, similar to I65-I72 titin domains for all the species characterized in this thesis. Although the 25 pN difference between the oxidized and reduced forms is comparable to the spread in the measured forces, the difference is statistically significant as unambiguously established by a Kolmogorov-Smirnov test [177].

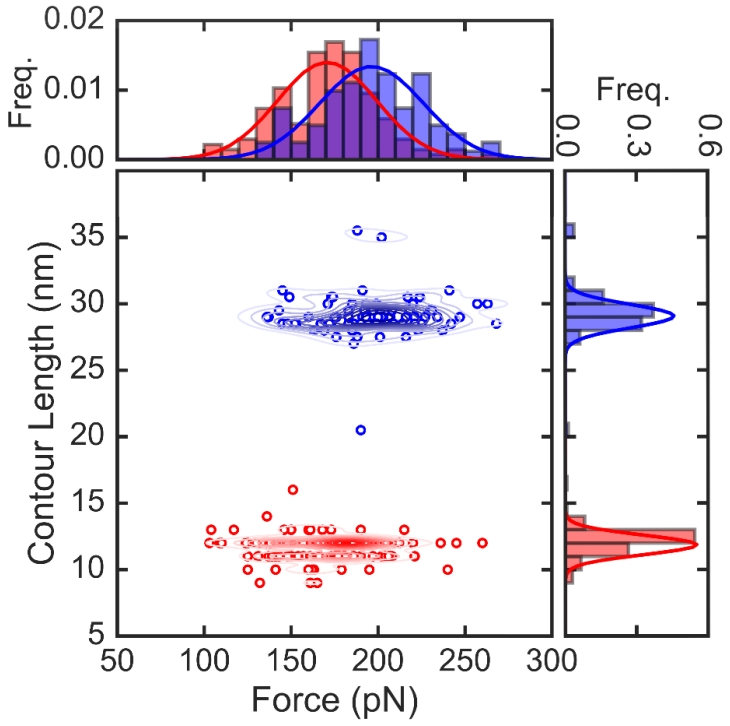


Figure 6.1. *Scatter plot of force versus contour length with corresponding histograms (top and right, respectively) of the* $(I91_{G32C-A75C})_8$ *polyprotein. Lines in the histograms are Gaussian fits to the data. The red and blue colors denote the oxidized and reduced states, respectively. The total number of data points are n=136 and n=81 for oxidized domains and reduced domains, respectively. Contour lines in the scatter plot were generated using kernel density estimates.*

The decrease in mechanical stability upon disulfide bonding is a somewhat surprising result and has not been reported before for other proteins. On the contrary, the results in the same study for FimG, a bacterial fimbriae protein, exhibit the opposite behavior compared to those of I65-I72 and I91 titins. While the mechanical stability of disulfide bonded domains in titin is lower than the reduced ones, for the FimG domain the unfolding force raises for the oxidized state. This observation is reminiscent of previous findings on the de novo designed protein Top7 [7, 8]. Many different contributions to protein energetics may be required to explain the effects of disulfides on mechanical stability. In the case of titin Ig domains disulfides are mainly cryptic and located in the middle region of the domain, and hence, a change in the unfolding mechanism can be ruled out. One possibility is the constrain introduced in the native state vibrational entropy by a disulfide bond [178], which has been found to affect the subpicosecond to 100 ps dynamics of proteins [179]. Another possibility is the presence of mechanical prestress in the disulfide bonded titin [180]. Using a model of titin Ig 65-72 obtained from the Robetta server [181] as input for an on-line tool for the analysis of the stereochemistry of disulfide bonds that implements the formula by Katz and Kossiakoff [182] (https://services.mbi.ucla.edu/disulfide/) it was found that all three disulfide bonds present in the model had χ'_2 angles that are statistically highly unlikely. It has been previously reported that the spatial configuration of the disulfide bonds of intrachain Ig domains could exert a tensile stress on the order of 100 pN [183].

Future work involving a systematic analysis of alternative locations of disulfide bonds across the β-strands in different proteins will enable to establish a consistent theory about the decrease in the mechanical stability of titin induced by disulfide bonds. Given the fundamental role of disulfide bonds in the mechanics of proteins, one may hypothesize that disulfides may be related to the mechanochemical evolution of titin. This supposition is supported by the phylogenetic results we obtained: on one side titin is the most mutated filament by far suggesting that it is the evolutionary information driver of muscle physiology and, on the other side, we also proved that cysteines are only lost in the elastic region of titin. Altogether, these observations suggest that Cys residues played a crucial role in the molecular evolution of the elastic I-band of titin. Concretely, cysteines are involved in the formation of disulfide bonds, which are essential in the conformation of structural proteins, modulating their mechanical stability and constraining their conformational dynamics and biological function, [155, 170, 184].

Our results also show that the evolution of muscle physiology seems to be linked to the molecular evolution of titin in tetrapods. The reconstruction of ancient forms of titin demonstrates that titin domains from the small zebra finch are similar to those of its ancestors, in terms of mechanical properties. However, titin domains from modern mammals have experienced more drastic changes leading to proteins that have lower mechanical stability and fewer disulfide bonds compared to those in their oldest ancestor. This is quite remarkable given that mutation rates for titin in zebra finch and living mammals are quite similar. Most likely, these changes are related to morphological and physiological consequences that derived in the vast diversity of physical and locomotor features found in mammals. This raises the question of whether the relation of titin mechanics with muscle contraction differs across species with different physiological characters. A possible interpretation is that small amniotes with fast muscle contractions rely upon the mechanical response of titin domains. Under physiological forces, titin domains have been shown to unfold and refold during muscle contraction [23]. The presence of disulfide bonds prevents overstretching of titin and increases the recoiling speed of the domains [169], probably increasing also the speed of muscle contraction. Thus, we hypothesize that the balance between mechanical stability and disulfide bonds may be a key factor in titin mechanical regulation and its evolution.

This mechanical regulation seems to hold relation to animal size. In fact, this is consistent with the fact that in small mammal hearts the isoform N2A, which is stiffer than N2BA, is more abundant [163]. Thus, the next logical question is if mechanical stability is related to animal size. This correlation may be related to titin mechanics as small animals have faster muscle contraction and shivering frequency [185]. To probe this idea, we plotted the mechanical stability of domains without and with disulfide bonds, versus body mass. Plots show that there is a correlation between unfolding force and body mass of modern animals (represented in circles in **Fig. 6.2**). In a semi-log scale, the body mass scaled is following an evolutionary power law allometric correlation, (**Fig. 6.2a** and **6.2b**). Allometry is the study of the relationship of body size with many anatomy, physiology and biochemical features and most of the times follows a characteristic power law equation. Concretely, the correlation between body mass and unfolding forces of non-disulfide bonded and disulfide bonded domains can be captured by the next equation:

$$F - F_0 = a * M^b, \qquad (6.1)$$

where F_0 is minimum force value of the curve in pN (taken from the lowest value on the dataset) and M is the body mass (g). Parameters of the equations for both domains are shown in **Table 6.2**. Since there are two allometric scaling plots, a range of body mass for each extinct species was obtained, which is determined by the minimum and maximum value of mass calculated. This is 8-69 g for LTCA, 14-16 g for LSCA, 14-69 g for LMCA and 95-116 g for LPMCA. These weights are typical of small animals and compare surprisingly well with sizes from fossils that could be related to these extinct animals [186-191].

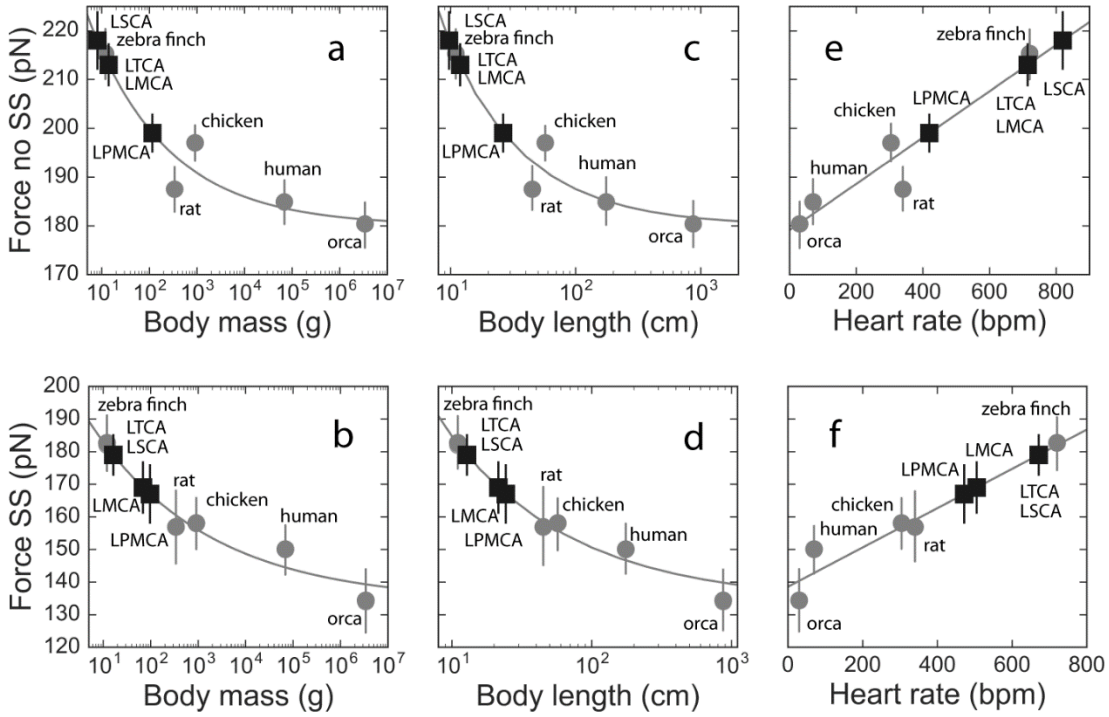


***Figure 6.2**. Correlation of mechanical stability of non-disulfide bonded and bonded domains with animal, body mass (**a, b**), body length (**c, d**) and heart rate (**e, f**). Stabilities versus body mass and length follow a power law correlation in all the cases, whereas stability versus heart rate correlation is linear in both cases. Modern species are represented in grey circles. The values for ancient species LTCA, LSCA, LMCA and LPMCA can be interpolated from the different fittings and are represented in black squares.*

Thus, the observed correlation between mechanical stability and body mass allows to predict the size of extinct species. Body mass has been shown to follow allometric scaling with other physiological traits such as metabolic rate, speed, arterial pressure or heat production [192]. In fact, it is common to observe allometric scaling in biological systems, and even enzyme activity has been suggested to show allometry [193]. However, this is the first observation showing allometric scaling between a physiological feature and molecular-level parameters. Moreover, we also plotted the unfolding force of disulfide bonded and non-bonded domains versus body length. In a semi-log scale, it is observed again that body length scaled the best following an evolutionary power law allometric correlation (**Fig. 6.2c** and **6.2d**):

$$F - F_0 = a * L^b, \qquad (6.2)$$

where L is the length in cm. Parameters for each domain type are shown in **Table 6.2**. Hence, by interpolating the unfolding forces of the ancient species it is possible to obtain the range their body length, which is 10-21cm for LTCA, 12-13 cm for LSCA, 12-21 cm for LMCA, and 24-26 cm for LPMCA. The low body mass and small size that we obtain for ancient species from the allometric correlations may be explained under the light of fossil remains. It has been demonstrated that after the mass extinction occurred in the late Devonian, the so-called Hangenberg event (359 Ma), the majority of taxa found in fossil records were under 40 cm [191]. This is consistent with a global shrinkage process during the early Carboniferous that lasted around 40 Myr. Our predicted data for LTCA, which supposedly lived after the Hangenberg event, are within the range of the sizes reported. Fossils of early amniotes including sauropsids and mammalian-like reptiles from the Carboniferous and Permian, are also within that range [189, 194]. In the case of mammals, numerous findings have shown that early mammals, including placentals, were small rodent-like animals [186-188, 190, 195] like we show in **Fig. 6.3**, which is in line with the sizes that we predict from nanomechanical information of titin. Nevertheless, establishing direct comparison between our estimations and any known fossils is difficult, because it is unlikely that a single fossil could be unambiguously labeled as LTCA, LSCA, LMCA or LPMCA, representing the true common ancestor of different taxonomic groups.

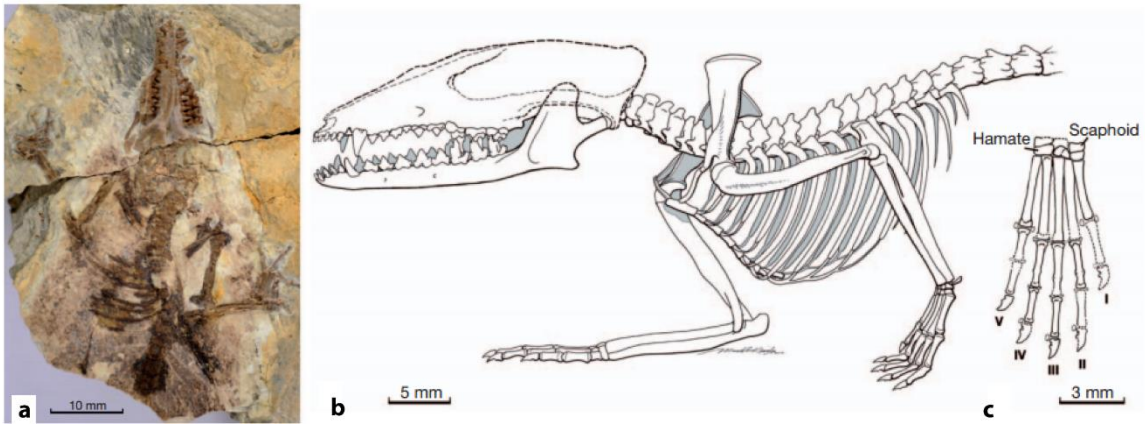


***Figure 6.3**. Holotype specimen of Juramaia sinensis (160 Myr, Jurassic period). (**a**) Specimen photograph and morphological identification. (**b, c**) Restoration of the partly preserved skeleton, skull and hand [187].*

Taking this into account, one can reason that some other physiological and morphological characters in animals may have a correlation with the mechanics and elasticity of titin. The more obvious trait would be heart rate since is directly related to muscle contraction and relaxation and therefore titin elasticity [196]. To test this idea, we plotted values of heart rate [197] against the mechanical stability of non-disulfide bonded (**Fig. 6.2e**) and disulfide bonde domains (**Fig. 6.2f**). Data from modern species fits to a linear expression that relates average unfolding forces of disulfide bonded and non-bonded domains and heart rates:

$$F = a + b * HR \qquad (6.3)$$

Force is introduced in pN and heart rate (HR) in beats per minute (bpm). Parameters for each type of domain are shown in **Table 6.2**. This relationship allows to determine heart rate solely based on unfolding force by simple interpolation. For LTCA the predicted heart rate range is 505-820 bpm, for LSCA 671-714 bpm, 505-714 for LMCA, and 419-472 bpm for LPMCA. Of course, heart rate depends on different factors such as temperature and metabolism, so these values should be simply interpreted as estimates of an approximate range. These heart rates are typical of small fast moving animals.

	a	b	r^2
F_{noSS} vs M	65	-0.26	0.86
F_{SS} vs M	73	-0.17	0.92
F_{noSS} vs L	185	-0.70	0.89
F_{SS} vs L	155	-0.49	0.94
F_{noSS} vs HR	179	0.047	0.87
F_{SS} vs HR	138	0.060	0.89

Table 6.2. *Allometric equations fitting parameters for F_{SS} and F_{noSS} vs mass, length and heart rate, respectively.*

We also used a collection of average heart rate and body mass of extant species [197] and constructed a heart rate vs body mass plot that present the typical -1/4th power allometric scaling (**Fig. 6.4**). Using the average values of HRs estimated from **Fig. 6.2e** and **6.2f** (663 bpm for LTCA, 693 bpm for LSCA, 610 bpm for LMCA, and 446 bpm for LPMCA) one can interpolate the values of body mass in this plot, resulting 19 g for LTCA, 16 g for LSCA, 27 gr for LMCA, and 93 gr for LPMCA. These values are very similar to the ones interpolated using force (**Fig. 6.2a and 6.2b**), which comes to show that the strong correlation between heart rate and mass is also achieved by the analysis of the unfolding force measurements. This suggests that titin mechanics had an important role on the acquisition of muscle diversity in the course of the evolution of tetrapods.

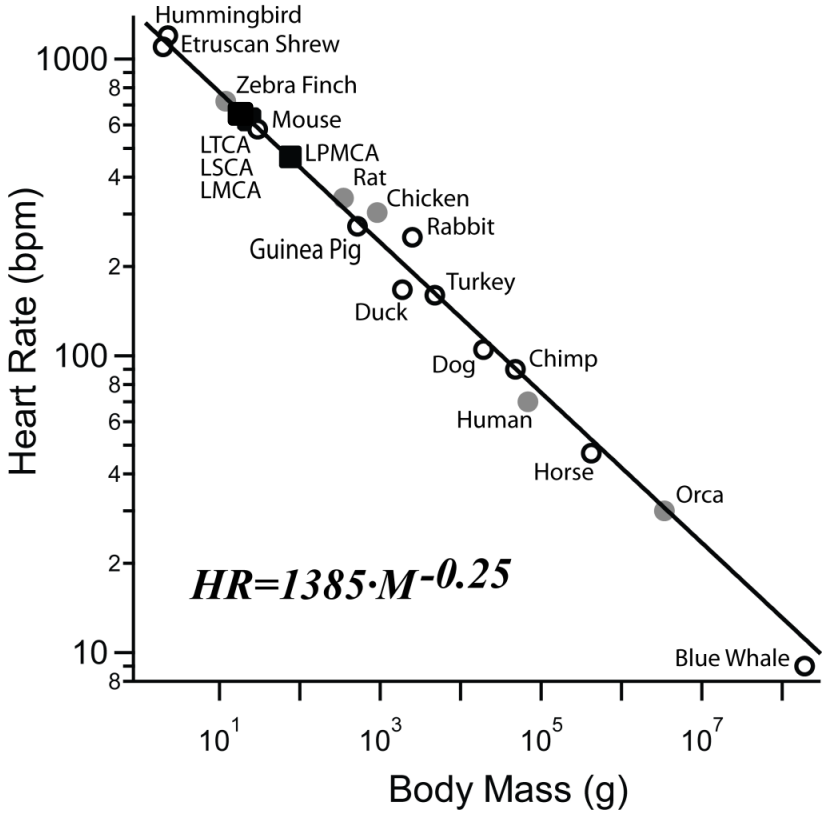


Figure 6.4. *Heart rate versus body mass for a collection of species including the predicted values for LTCA, LSCA LMCA and LPMCA. The correlation can be fitted using a power law that yields the typical -1/4 power for allometric scaling. Extant species used in this study are* shown in grey circles, ancestral species in black squares and other species in empty circles.

Moreover, under physiological forces, titin domains have been shown to unfold and refold in muscle contraction. The presence of disulfide bonds may increase the recoiling speed of titin domains. However, having a large number of domains unfolding may generate an overstretching of titin. Therefore, the balance between mechanical stability and disulfide bonds may be the key driving factor in titin mechanical regulation. A large number of disulfides decreases the stretching length and speeds up domain recoiling but also decreases the average unfolding force. This opens up the question of whether the relation of titin mechanics with muscle contraction differs across different species with different physiological characters.

Chapter 7: Conclusions

- Our results show that the evolution of muscle physiology is strongly linked to the molecular evolution of titin. Titin domains from different lineages seem to have evolved differently, indicating a major role in diversification. On the other hand, actin is extremely conserved during evolution. Myosin's mutation rate is lower than titin's, suggesting a less important role is muscle physiology evolution.

- The differences in mechanical stability and disulfide bond occurrence between titins from living species and their ancestors appear as key elements in the mechanical evolution of titin. These differences illustrate a paleomechanical trend along the different animal lineages that allow us to establish physiological correlations purely based on the molecular properties of titin.

- We observe a systematic decrease in the mechanical stability of Ig domains for all the species when they are disulfide bonded. We hypothesize that disulfides could play a global role in the mechanical stability of proteins, rather than being simple mechanical lockers.

- The reconstruction of ancient forms of titin demonstrates that titin domains from zebra finch are similar to their common ancestors, in terms of their mechanical

properties. Titin domains from modern mammals, as diverse as orca whale, rat, and human, have experienced more drastic changes, leading to proteins that have lower mechanical stability and fewer disulfide bonds. This is remarkable, given that the mutation rate for titin molecules from zebra finch and mammals are quite similar. Most likely, these changes had morphological and physiological consequences that derived in the vast diversity of physical and locomotor features found in mammals.

- The observed correlations between mechanical force and heart rate, body mass and body length may have a physiological significance. Ancestral and zebra finch titin fragments are richer in disulfide bonds and have higher mechanical stability which make them stiffer than mammal's titin, especially than those of large mammals. This is consistent with the fact that in the heart of small mammals the isoform N2A, which is stiffer than N2BA, is more abundant. Results demostrate that even in the N2BA isoform, titin domains are also stiffer. Animals with high heart pulse and fast muscle contraction rely upon the mechanical response of titin.

- The predicted data por LTCA, which lived after the Hangenberg event, are within the range of the sizes reported from fossil records. Fossils of early reptiles are also within this range. In the case of mammals, numerous findings have shown that early mammals were shrew- or mouse-like animals, with sizes between 10 and 30 cm. Again, this is in line with the sizes that we predict from nanomechanical information of titin.

- These calculations are approximations based on nanomechanical and biochemical properties. The most striking conclusion is that using nanomechanical measurements and ancestral sequence reconstruction we are able to determine that the ancestors of tetrapods, sauropsids and mammals were rather small animals, comparable to those found in fossil records.

Part V

Appendix I

List of titin proteins from the species used in the construction of phylogenetic trees

The sequences were retrieved from the Uniprot and GenomeNET databases using BLAST. The asterisks indicate sequences that were updated or added during the study and considered for the tree in Figure 4.4:

B0S758	*Danio rerio* (Zebrafish)
gi 348541917	*Oreochromis niloticus* (Tilapia)
F7EAV6	*Xenopus tropicalis* (Western clawed frog)
gi 103063864	*Phyton bivittatus* (Burmese python *)
gi 100560476	*Anolis carolinensis* (American chameleon *)
H9GP88	*Anolis carolinensis* (American chameleon)
K7G060	*Pelodiscus sinensis* (Chinese softshell turtle)
gi 102559759	*Alligator mississippiensis* (American alligator)
gi 465955284	*Chelonia mydas* (Green sea turtle)
gi 449507164	*Taeniopygia guttata* (Zebrafinch)
gi 483520158	*Anas platyrhynchos* (Wild duck)
G1NAX9	*Meleagris gallopavo* (Wild turkey)
gi 100546807	*Meleagris gallopavo* (Wild turkey *)
gi 101921795	*Falco peregrinus* (Peregrine falcon *)
gi 101819370	*Ficedulaalbicollis* (Collared flycatcher *)
gi 363735918	*Gallus* (Chicken)
F6VRV4	*Ornithorhynchus anatinus* (Platypus)
gi 103171044	*Ornithorhynchus anatinus* (Platypus*)
gi 395519871	*Sarcophilus harrisii* (Tasmanian devil)
G3UK67	*Loxodonta africana* (African elephant)
gi 471370017	*Trichechus manatus* (West Indian manatee)
L5K2L4	*Pteropus alecto* (Black flying fox)
G1P5X9	*Myotis lucifugus* (Little brown bat)
M3WG03	*Felis catus* (Cat)
gi 359323893	*Canis lupus* (Dog)
G1L1P3	*Ailuropoda melanoleuca* (Giant panda)

F6VG02	*Equus caballus* (Horse)
gi 465995183	*Orcinus orca* (Orca)
F1N757	*Bos Taurus* (Cattle)
gi 426220782	*Ovis aries* (Sheep)
G1U9S3	*Oryctolagus cuniculus (*Rabbit)
G3HAC6	*Cricetulus griseus* (Chinese hamster)
A2ASS6	*Mus musculus* (Mouse)
gi 392339498	*Rattus norvegicus* (Brown Rat)
L9KLA3	*Tupaiabelangeri chinensis* (Chinese tree shrew)
F7IGY9	*Callithrix jacchus* (Marmoset)
H2P803	*Pongo abelii* (Orangutan)
G3QYH8	*Gorilla gorilla* (Gorilla)
H2QJ24	*Pan troglodytes* (Chimpanzee)
D3DPG0	*Homo sapiens* (Human)

List of myosin II proteins from the species used in the construction of the phylogenetic tree.

The sequences were retrieved from the Uniprot and GenomeNET databases using BLAST.

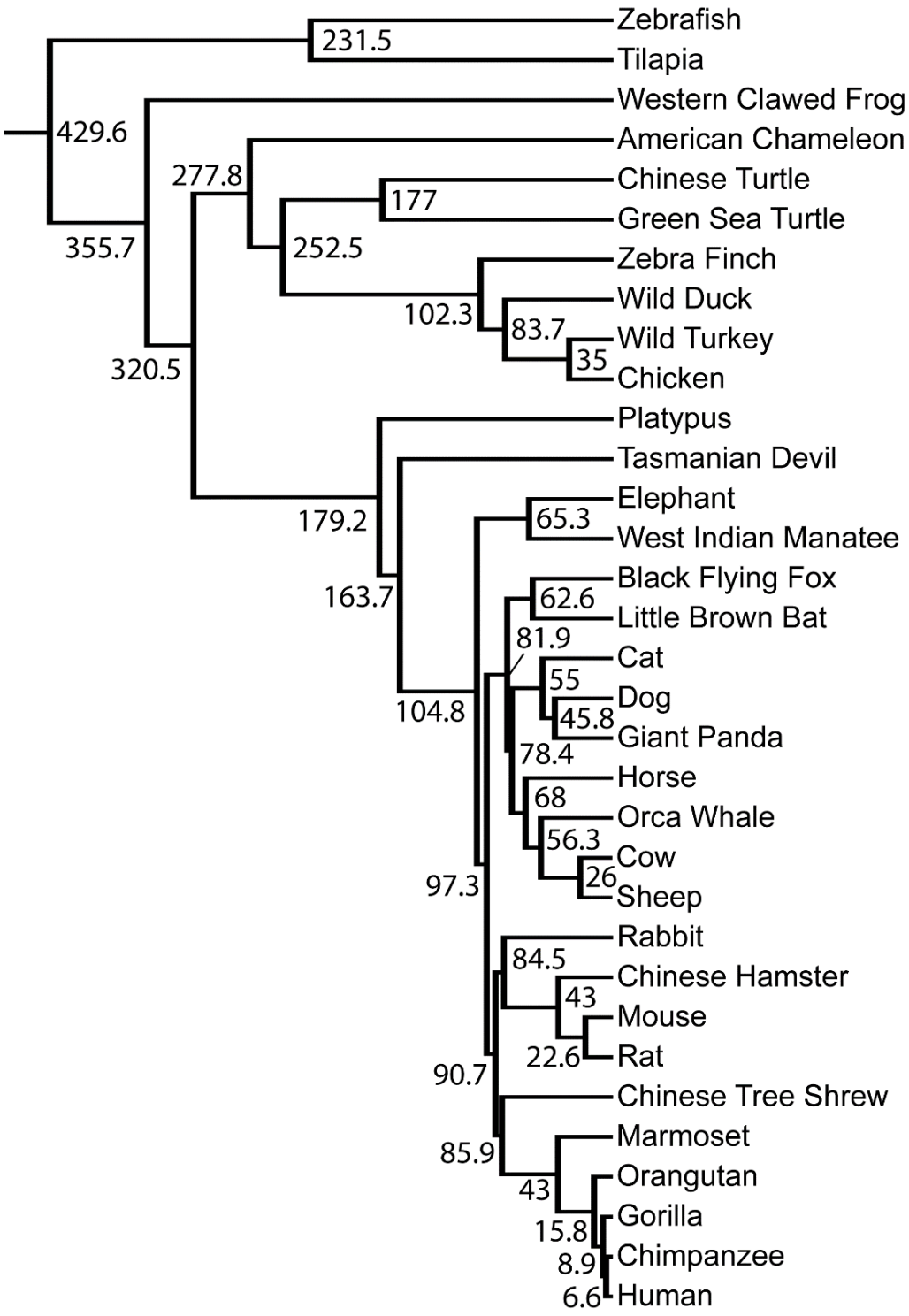
W5N278	*Lepisosteus oculatus* (Spotted gar)
I3KU01	*Tilapia nilotica* (Nile tilapia)
X1WF87	*Danio rerio* (Zebrafish)
H3A7R4	*Latimeria chalumnae* (Coelacanth)
B4F6Y1	*Xenopus tropicalis* (Western clawed frog)
L7N035	*Anolis carolinensis* (Arboreal lizard)
K7FJP0	*Pelodiscus sinensis* (Chinese softshell turtle)
F1P3W8	*Gallus gallus* (Chicken)
R7VQ57	*Columba livia* (Pigeon)
U3JD47	*Ficedula albicollis* (Flycatcher)
H0Z9U7	*Taeniopygia guttata* (Zebrafinch)
G3W6X6	*Sarcophilus harrisii* (Tasmanian devil)
G1PUN7	*Myotis lucifugus* (Little brown bat)
Q8MJV1	*Equus caballus* (Horse)

W5PT09	*Ovis aries* (Sheep)
Q9BE41	*Bos Taurus* (Cattle)
M3WDH9	*Felis catus* (Cat)
G1L2C8	*Ailuropoda melanoleuca* (Giant panda)
Q076A7	*Canis lupus* (Dog)
G1SJQ4	*Oryctolagus cuniculus (*Rabbit)
F1LRV9	*Rattus norvegicus* (Brown Rat)
G3UW82	*Mus musculus* (Mouse)
F7DRK7	*Callithrix jacchus* (Marmoset)
H2NSR2	*Pongo abelii* (Orangutan)
G3RN81	*Gorilla gorilla* (Gorilla)
Q9UKX2	*Homo sapiens* (Human)

Appendix II

Datation of titin chronogram according to fossil records and TTOL.

Titin chronogram (Node ages are reported in Myr).



Appendix III

Sequences of modern and ancestral titin fragments studied

The nomenclature of the fragment corresponds to the canonical human titin with UniProt code Q8W242 from residues 8232 to 8976, which correspond to domains I65 to I72. The equivalent domains are used for the rest of the species and ancestral proteins.

Human

EPPRFIKKLEPSRIVKQDEFTRYECKIGGSPEIKVLWYKDETEIQESSKFRMSFVDS
VAVLEMHNLSVEDSGDYTCEAHNAAGSASSSTSLKVKEPPIFRKKPHPIETLKGA
DVHLECELQGTPPFHVSWYKDKRELRSGKKYKIMSENFLTSIHILNVDAADIGEY
QCKATNDVGSDTCVGSIALKAPPRFVKKLSDISTVVGKEVQLQTTIEGAEPISVV
WFKDKGEIVRESDNIWISYSENIATLQFSRVEPANAGKYTCQIKNDAGMQECFAT
LSVLEPATIVEKPESIKVTTGDTCTLECTVAGTPELSTKWFKDGKELTSDNKYKIS
FFNKVSGLKIINVAPSDSGVYSFEVQNPVGKDSCTASLQVSDRTVPPSFTRKLKET
NGLSGSSVVMECKVYGSPPISVSWFHEGNEISSGRKYQTTLTDNTCALTVNMLEE
SDSGDYTCIATNMAGSDECSAPLTVREPPSFVQKPDPMDVLTGTNVTFTSIVKGT
PPFSVSWFKGSSELVPGDRCNVSLEDSVAELELFDVDTSQSGEYTCIVSNEAGKA
SCTTHLYIKAPAKFVKRLNDYSIEKGKPLILEGTFTGTPPISVTWKKNGINVTPSQR
CNITTTEKSAILEIPSSTVEDAGQYNCYIENASGKDSCSAQILILEPPYFVKQLEPVK
VSVGDSASLQCQLAGTPEIGVSWYKGDTKLRPTTTYKMHFRNNVATLVFNQVDI
NDSGEYICKAENSVGEVSASTFLT

Rat

EPPRFIKKLDQSRIVKQDEYTRYECKIGGSPEIKVLWYKDEVEIQESSKFRMSFED
SVAILEMHNLSVEDSGDYTCEARNAAGSASSSTSLKVKEPPVFRKKPFPVETLKG
ADVHLECELQGTPPFQVSWYKDKRELRSGKKYKIMSENLLTSIHILNVDTADIGE
YQCKATNDVGSDTCVGSVTLKAPPQFVKKLSDVSTIIGKEVQLQTTIEGAEPISVA
WFKDKGEIVRESDNIWISHSENVATLHFSRAEPANAGKYTCQIKNDAGVQECYA
TLSVLEPATIVEKPESIKVTTGDTCTLECMVSGTPELSTKWFKDGKELTGDSKYKI
SFFNKVSGLKIISVAPGDSGVYSFEVQNPVGKDSCTVSIQVSDRIIPPSFTRKLKET
NGLSGSSVVMECKVFGSPPISVLWLHDGNAISSGRKYQTTLTDNTCALTVNMLE

DADAGDYTCIATNVAGSDECSAPLTVREPPSFVQKPDPMDVLTGSNVTFTSIVKG
TPPFTVSWFKGSSELVPGARCNVSLQDSVAELELFDVDTSQSGDYTCIVSNEAGR
ASCTTQLFVKAPAIFVKRLNDYSIEKGKPLILEGTFSGTPPISVTWKKNGVNVTAS
QRCNITTTEKSAILEILSSTVEDSGQYNCYIENASGKDSCSAQILILEPPYFVKQLEP
LKVTVGDSASLQCQLAGTPEIGVSWYKGDTKLRPTTTCKMHFKNNVATLVFTQ
VDSNDSGEYICRAENSVGEVSSSTFLT

Orca

EPPRFIKKLEPSRIMKQGESTRYECKVGGSPEIKVLWYKDETEIQESSKFRMSFHD
SVAVLEMHALSVEDSGDYTCEARNAAGRASSSTTLKVKEPPVFRKKPRPVETLE
GADVHLECELQGTPPFQVSWHKDKRELRSGKKYKIMSENFLTSIHILSVSAADVG
EYQCKATNDVGGDTCVGSITLKAPPRFVKKPSDISAIVGEEVRLQAAIEGTEPISM
VWFKDKGEMVRESDNIWISYSENIATLQFSRVETANAGKYTCQIKNDAGMQECF
ATVSILEPAAIVEKPESIKVTTGDTCTLECMVTGTPELTTKWFKDAKELTSDSKYK
ISFFNKISGLKIINVAPSDSGVYSFEVQNPVGKDSCTASVHVSDRIVPPSFTRKLKE
TNGLSGSSVVMECKVYGSPPISVSWFHEGNEISSGRKYQTTLTDNTCALTVNMLE
DSDTGDYTCIATNVAGSDECSAPLTVREPPSFVQKPDPMDVLTGANVTFTSLVKG
TPPFSVSWFKGSSELVPGDRCNVSLEDSVAELELFDVDTSQSGEYTCIVSNEAGK
ASCTTHLYVKAPAKFVKKLNDYSLEKGKPLILEATYTGTLPISVTWKKNGTNITP
SQRCHITTTEKSAILEIPSSTVEDAGQYNCYIENASGKDSCSAQILILEPPYFVKQLE
PVKVTIGDSASLQCQLAGTPEIGVSWYKGDTKLRPTTTYKMHFKNNVATLVFNQ
VDSNDSGEYICRAENSVGEVSSSTFLT

Zebra Finch

EPPRFIKKLDSSKLVKQHDSTTYECKIGGSPEIKVTWYKGETEIHPSEKYRMSFVD
SVAVIEMHNLSVEDSGDYTCEAQNPAGSASTSTTLRVKAPPIFTKKPHPVETLKG
SDIHLECKLQGTPPFQISWYKDKREIRSSKKYKVMSENYIASLHILSVDTADVGEY
HCKAVNDVGSDSCIGSVTLRAPPTFVKKLSDLTVVVGESIELQAAVQGSQPISVL
WLKDKGEIIRESDNLWISYSENIATMQIGNAEPTNAGKYICQIKNDAGIQECFAML
KVLEPAVIVEKPGPVKVTAGDSCTLECTVDGTPELTARWFKDGNELSTDHKYKIS
FFNKVSGLKILNATLEDSGEYTFEVKNSVGKSSCTASVHVSDRIIPPSFTRKLKET
YGQLGSSAVLECKVYGSPPILVSWFHDGQEITSGEKYQATLTDNTCSLKVNGLQE
SDMGTYLCTATNAAGSDECSAFLSVIESPSFVKKPEPLDVLSGANITFTSIVKGSPP

LEVKWFRGSVELVPGPRCNITLQDSVAELELFDVHPLESGDYTCQVSNEAGKISC
TTHLFVKEPAKFVKKVNDLSVEKGKNLILECTYMGTPPISVTWKKNGVKIMHSE
KCSITTTDTSAILEIPNSKLEDQGQYSCHIENDSGKDTCHGTITILEPPYFIRSLEPV
QVTVGDSASLQCQVAGTPEMIVSWYKGDTKLRGTATMKMHFRNQIATLVFSQV
DGSDSGEYICKVENSVGEASSSSLLT

Chicken

EPPRFIKKLDSSRLVKQHDSTRYECKVGGSPEIKVTWYKGETEIHPSEKYSMSFV
DSVAVLEMHNLSVEDSGDYSCEAQNPAGSASTSTSLKVKAPPAFTKKPHPVQTL
KGSDVHLECELQGTPPFQISWYKDKREIRSSKKYKVMSENYLASIHILNVDTADV
GEYHCKAVNDVGSDSCIGSVTLRAPPTFVKKLSDVTVVVGETIELQAAVEGAQPI
SVLWLKDKGEIIRESENLWISYSENVASLKIGNAEPTNAGKYICQIKNDAGFQECF
AKLTVLEPAVIVEKPGPVKVTAGDSCTLECTVDGTPELTARWFKDGNELSTDHK
YKISFFNKVSGLKILNAGLEDSGEYTFEVKNSVGKSSCTASLQVSDRIMPPSFTRK
LKETYGQLGSSAVLECKVYGSPPILVSWFHDGQEITSGDKYQATLTDNTCSLKVN
GLQESDMGTYSCTATNVAGSDECSAFLSVREPPSFVKKPEPFNVLSGENITFTSIV
KGSPPLEVKWFRGSIELAPGHKCNITLQDSVAELELFDVQPLQSGDYTCQVSNEA
GKISCTTHLFVKEPAKFVMKVNDLSVEKGKNLILECTYTGTPPISVTWKKNGVIL
KHSEKCSITTTETSAILEIPNSKLEDQGQYSCHIENDSGQDNCHGAITILEPPYFVTP
LEPVQVTVGDSASLQCQVAGTPEMIVSWYKGDTKLRGTATVKMHFKNQVATLV
FSQVDSDDSGEYICKVENTVGEATSSSLLT

LTCA

EPPRFVKKLESSKVVKQGDSTRFECKISGSPEIRVLWYKNDAEIQHGGKYRMSFV
DSVAVLEISNASVEDSGDYTCEAHNDAGSASCSTSLKVKEPPVFIKKPHPVETLK
GSDVSLQCELKGTPPFQVSWYKDKREIKSSKKYKIMSENYLASIHILKVDAADIG
EYQCKAVNDVGSDTCLGSIKLKEPPRFVKKLSDASAVVGEPVELQATVEGAQPIS
VTWLKDKEEIVRESENIWISFSDNVATLQFLNAEPANAGKYTCQIKNDAGVQECF
ATLSVLEPAVIVEKPESMKVTSGDTCTLECTVSGTPELSAKWFKDGKELSSDHKY
KISFHNKVSGLKILNAAPNDSGEYTFEVKNNVGKDSCAMSVQVSDRIIPPSFTRKL
KETHGLLGSSVVLECKVSGSPPISVSWFHNGNEITSGGKYQATLTDNTCTLTVSA
LETSDAGKYSCTATNVAGSDECSAALTVKEPPSFVEKPEPLEVLPGATVTFTAIIK
GTPPFKVKWFRGSTELVPGRRCNISLEDSVAVLELYNVDTSQSGDYTCQITNDAG

KDSCTTHLFVKEPAKFVKKLNDYSIEKGKPLILECTYTGTPPISVTWKKNGVEITQ
SEKCSITTTEKSCILEIPNSTMEDAGQYTCHVENASGHDTCQATISILEPPYFIKPLE
PVEVTAGDAASLQCQIAGTPEIKVSWYKGDTKLRATATSKMHFKNNVATLVFA
QVDSNDSGEYICKAENSVGEASSSTSLT

LSCA

EPPRFIKKLDSSRVVKQHDSTRYECKIGGSPEIKVIWYKNETEIHPSSKYRMSFVD
SVAVIEMHNLSVEDSGDYTCEAQNAAGSASSSTSLKVKEPPVFSKKPHPVETLKG
SDVHLECELKGTPPFQVSWYKDKREIRSSKKYKIMSENYLASIHILNVDAADIGE
YHCKAVNDVGSDTCIGSITLKAPPRFVKKLSDLTAVVGEPVELQATVEGAQPISV
LWLKDKGEIVRESDNLWISYSENVATLQIGNAEPANAGKYTCQIKNDAGVQECF
ATLSVLEPAVIVEKPGPMKVTSGDSCTLECTVAGTPELTARWFKDGNELSTDHK
YKISFFNKVSGLKILNAGPEDSGEYTFEVKNSVGKSSCTASVHVSDRIIPPSFTRKL
KETHGLLGSSVVLECKVYGSPPISVSWFHDGQEITSGDKYQATLTDNTCSLTVNA
LEESDAGNYSCTATNVAGSDECSAYLTVREPPSFVKKPEPLQVLSGANITFTSIIK
GTPPFDVKWFRGSVELVPGHRCNISLEDSVAELELFDVHPLQSGDYTCLVTNEAG
KISCTTHLFVKEPAKFVKKLNDFSVEKGKPLILECTYTGTPPISVTWKKNGVKITQ
SEKCSITTTETSAILEIPNSKMEDAGQYTCHIENDSGQDNCHATISILEPPYFVRPLE
PVQVTVGDSASLQCQVAGTPEIIVSWYKGDTKLRATATSKMHFKNNVATLVFN
QVDSNDSGEYICKAENSVGEASSSALLS

LMCA

EPPRFIKKLESSRVVKQHDSTRYECKIGGSPEIKVLWYKNETEIQSSSKFRMSFVD
SVAVIEMHNLSVEDSGDYTCEAHNAAGSASSSTSLKVKEPPVFSKKPHPVETLKG
SDVHLECELRGTPPFQVSWHKDKREIRSGKKYKIMSENFLTSIHILNVDASDIGEY
QCKAVNDVGSDTCVGSITLKAPPRFVKKLSDLSAVVGDQVQLQATIEGAEPISVV
WFKDKGEIVRESDNIWISYSENVATLQFANAEPANAGKYTCQIKNDAGMQECFA
TLSVLEPAVIVEKPESMKVTSGDTCTLECTVAGTPELSTKWFKDGKELTSDSKYK
ISFFNKVSGLKIINVAPNDSGVYTFEVQNSVGKDSCTASVQVSDRIVPPSFTRKLK
ETNGLFGSSVVLECKVYGSPPISVSWFHEGNEITSGRKYQATLTDNTCSLTVNAL
EESDAGDYTCVATNVAGSDECSAALTVREPPSFVQKPDPLDVLTGTNVTFTSIIK
GTPPFSVSWFKGSSELVPGDRCNISLDDSVAELELFDVDTSQSGDYTCVVTNEAG
KASCTTHLYVKAPAKFVKKLNDYSIEKGKPLILEGTYTGTPPISVTWKKNGLNITP

SQKCSITTTEKSAILEIPSSTVEDAGQYTCYIENASGKDSCHAQILILEPPYFVKQLE
PVKVTVGDSASLQCQVAGTPEIAVSWYKGDTKLRATATSKMHFRNNVATLVFN
QVDSNDSGEYICKAENSVGEVSSSTFLT

LPMCA

EPPRFIKKLEPSRIVKQDEYTRYECKIGGSPEIKVLWYKDETEIQESSKFRMSFVDS
VAVLEMHNLSVEDSGDYTCEAHNAAGSASSSTSLKVKEPPIFRKKPHPVETLKG
ADVHLECELQGTPPFQVSWHKDKRELRSGKKYKIMSENFLTSIHILNVDAADIGE
YQCKATNDVGSDTCVGSITLKAPPRFVKKLSDISTIVGEEVQLQTTIEGAEPISVV
WFKDKGEIVRESDNIWISYSENIATLQFSRAEPANAGKYTCQIKNDAGMQECFAT
LSVLEPAAIVEKPESIKVTSGDTCTLECTVTGTPELSTKWFKDGKELTSDSKYKISF
FNKVSGLKIINVAPNDSGVYSFEVQNPVGKDSCTASVQVSDRIVPPSFTRKLKETN
GLSGSSVVMECKVYGSPPISVSWFHEGNEISSGRKYQTTLTDNTCALTVNMLEES
DAGDYTCVATNVAGSDECSAPLTVREPPSFVQKPDPMDVLTGTNVTFTSIIKGTP
PFSVSWFKGSSELVPGDRCNVSLEDSVAELELFDVDTSQSGEYTCIVSNEAGKAS
CTTHLYVKAPAKFVKRLNDYSIEKGKPLILEGTYTGTPPISVTWKKNGINITPSQR
CNITTTEKSAILEIPSSTVEDAGQYNCYIENASGKDSCSAQILILEPPYFVKQLEPVK
VTVGDSASLQCQLAGTPEIAVSWYKGDTKLRPTATYKMHFRNNVATLVFNQVD
SNDSGEYICRAENSVGEVSSSTFLT

Appendix IV

Sequences of modern and ancestral myosin II sequences studied.

Human

MSSDSELAVFGEAAPFLRKSERERIEAQNRPFDAKTSVFVAEPKESFVKGTIQSRE
GGKVTVKTEGGATLTVKDDQVFPMNPPKYDKIEDMAMMTHLHEPAVLYNLKE
RYAAWMIYTYSGLFCVTVNPYKWLPVYKPEVVTAYRGKKRQEAPPHIFSISDNA
YQFMLTDRENQSILITGESGAGKTVNTKRVIQYFATIAVTGEKKKEEITSGKIQGT
LEDQIISANPLLEAFGNAKTVRNDNSSRFGKFIRIHFGTTGKLASADIETYLLEKSR
VVFQLKAERSYHIFYQITSNKKPELIEMLLITTNPYDYPFVSQGEISVASIDDQEEL
MATDSAIDILGFTNEEKVSIYKLTGAVMHYGNLKFKQKQREEQAEPDGTEVADK
AAYLQSLNSADLLKALCYPRVKVGNEYVTKGQTVEQVSNAVGALAKAVYEKM
FLWMVARINQQLDTKQPRQYFIGVLDIAGFEIFDFNSLEQLCINFTNEKLQQFFNH
HMFVLEQEEYKKEGIEWTFIDFGMDLAACIELIEKPMGIFSILEEECMFPKATDTS
FKNKLYDQHLGKSANFQKPKVVKGKAEAHFALIHYAGVVDYNITGWLEKNKDP
LNETVVGLYQKSENLNKLMTNLRSTHPHFVRCIIPNETKTPGAMEHELVLHQLRC
NGVLEGIRICRKGFPSRILYADFKQRYKVLNASAIPEGQFIDSKKASEKLLASIDID
HTQYKFGHTKVFFKAGLLGLLEEMRDDKLAQLITRTQARCRGFLARVEYQRMV
ERREAIFCIQYNIRSFMNVKHWPWMKLFFKIKPLLKSAETEKEMATMKEEFQKIK
DELAKSEAKRKELEEKMVTLLKEKNDLQLQVQAEAEGLADAEERCDQLIKTKIQ
LEAKIKEVTERAEDEEEINAELTAKKRKLEDECSELKKDIDDLELTLAKVEKEKH
ATENKVKNLTEEMAGLDETIAKLTKEKKALQEAHQQTLDDLQAEEDKVNTLTK
AKIKLEQQVDDLEGSLEQEKKLRMDLERAKRKLEGDLKLAQESIMDIENEKQQL
DEKLKKKEFEISNLQSKIEDEQALGIQLQKKIKELQARIEELEEEIEAERASRAKAE
KQRSDLSRELEEISERLEEAGGATSAQIEMNKKREAEFQKMRRDLEEATLQHEAT
AATLRKKHADSVAELGEQIDNLQRVKQKLEKEKSEMKMEIDDLASNVETVSKA
KGNLEKMCRTLEDQLSELKSKEEEQQRLINDLTAQRGRLQTESGEFSRQLDEKEA
LVSQLSRGKQAFTQQIEELKRQLEEEIKAKNALAHALQSSRHDCDLLREQYEEEQ
ESKAELQRALSKANTEVAQWRTKYETDAIQRTEELEEAKKKLAQRLQAAEEHV
EAVNAKCASLEKTKQRLQNEVEDLMLDVERTNAACAALDKKQRNFDKILAEW
KQKCEETHAELEASQKEARSLGTELFKIKNAYEESLDQLETLKRENKNLQQEISD
LTEQIAEGGKRIHELEKIKKQVEQEKCELQAALEEAEASLEHEEGKILRIQLELNQ

VKSEVDRKIAEKDEEIDQLKRNHIRIVESMQSTLDAEIRSRNDAIRLKKKMEGDL
NEMEIQLNHANRMAAEALRNYRNTQGILKDTQIHLDDALRSQEDLKEQLAMVE
RRANLLQAEIEELRATLEQTERSRKIAEQELLDASERVQLLHTQNTSLINTKKKLE
TDISQMQGEMEDILQEARNAEEKAKKAITDAAMMAEELKKEQDTSAHLERMKK
NMEQTVKDLQLRLDEAEQLALKGGKKQIQKLEARVRELEGEVESEQKRNAEAV
KGLRKHERRVKELTYQTEEDRKNILRLQDLVDKLQAKVKSYKRQAEEAEEQSN
TNLAKFRKLQHELEEAEERADIAESQVNKLRVKSREVHTKVISEE

Rat

MSSDAEMAVFGEAAPYLRKSEKERIEAQNKPFDAKSSVFVVDAKESFVKATVQS
REGGKVTAKTEGGATVTVKDDQVFPMNPPKYDKIEDMAMMTHLHEPAVLYNL
KERYAAWMIYTYSGLFCVTVNPYKWLPVYNAEVVAAYRGKKRQEAPPHIFSISD
NAYQFMLTDRENQSILITGESGAGKTVNTKRVIQYFATIAVTGEKKKEEAPSGKM
QGTLEDQIISANPLLEAFGNAKTVRNDNSSRFGKFIRIHFGTTGKLASADIETYLLE
KSRVTFQLKAERSYHIFYQIMSNKKPDLIEMLLITTNPYDYAFVSQGEITVPSIDDQ
EELMATDSAIDILGFTSDERVSIYKLTGAVMHYGNMKFKQKQREEQAEPDGTEV
ADKAAYLQNLNSADLLKALCYPRVKVGNEYVTKGQTVQQVYNAVGALAKAV
YEKMFLWMVTRINQQLDTKQPRQYFIGVLDIAGFEIFDFNSLEQLCINFTNEKLQ
QFFNHHMFVLEQEEYKKEGIEWEFIDFGMDLAACIELIEKPMGIFSILEEECMFPK
ATDTSFKNKLYEQHLGKSNNFQKPKPAKGKVEAHFSLVHYAGTVDYNIAGWLD
KNKDPLNETVVGLYQKSENLNKLMTNLRSTHPHFVRCIIPNETKTPGAMEHELV
LHQLRCNGVLEGIRICRKGFPSRILYADFKQRYKVLNASAIPEGQFIDSKKASEKL
LGSIDIDHTQYKFGHTKVFFKAGLLGLLEEMRDDKLAQLITRTQAMCRGYLARV
EYQKMVERRESIFCIQYNVRAFMNVKHWPWMKLYFKIKPLLKSAETEKEMANM
KEEFEKTKENLAKAEAKRKELEEKMVALMQEKNDLQLQVQSEADSLADAEERC
DQLIKTKIQLEAKIKEVTERAEDEEEINAELTAKKRKLEDECSELKKDIDDLELTL
AKVEKEKHATENKVKNLTEEMAGLDETIAKLTKEKKALQEAHQQTLDDLQAEE
DKVNTLTKAKIKLEQQVDDLEGSLEQEKKIRMDLERAKRKLEGDLKLAQESTM
DVENDKQQLDEKLKKKEFEMSNLQSKIEDEQALGMQLQKKIKELQARIEELEEEI
EAERASRAKAEKQRSDLSRELEEISERLEEAGGATSAQIEMNKKREAEFQKMRR
DLEEATLQHEATAATLRKKHADSVAELGEQIDNLQRVKQKLEKEKSEMKMEID
DLASNMEVISKSKGNLEKMCRTLEDQVSELKTKEEEQQRLINELTAQRGRLQTES
GEYSRQLDEKDSLVSQLSRGKQAFTQQIEELKRQLEEEVKAKSALAHALQSSRH

DCDLLREQYEEEQEAKAELQRAMSKANSEVAQWRTKYETDAIQRTEELEEAKK
KLAQRLQDAEEHVEAVNAKCASLEKTKQRLQNEVEDLMIDVERTNAACAALDK
KQRNFDKILAEWKQKYEETHAELEASQKESRSLSTELFKIKNAYEESLDQLETLK
RENKNLQQEISDLTEQIAEGGKRIHELEKIKKQIEQEKSELQAALEEAEASLEHEE
GKILRIQLELNQVKSEIDRKIAEKDEEIDQLKRNHIRVVESMQSTLDAEIRSRNDAI
RIKKKMEGDLNEMEIQLNHSNRMAAEALRNYRNTQGILKDTQLHLDDALRGQE
DLKEQLAMVERRANLLQAEIEELRATLEQTERSRKIAEQELLDASERVQLLHTQN
TSLINTKKKLETDISQIQGEMEDIVQEARNAEEKAKKAITDAAMMAEELKKEQDT
SAHLERMKKNLEQTVKDLQHRLDEAEQLALKGGKKQIQKLEARVRELEGEVEN
EQKRNVEAIKGLRKHERRVKELTYQTEEDRKNVLRLQDLVDKLQSKVKAYKRQ
AEEAEEQSNVNLAKFRKIQHELEEAEERADIAESQVNKLRVKSREVHTKIISEE

Zebra finch

MSSDAEMAIFGEAAPYLRKSEKERIEAQNKPFDAKSSVFVVHAKESFVKGTITSR
ESGKVTVKTEGGETLTVKDDQIFSMNPPKYDKIEDMAMMTHLHEPAVLYNLKE
RYAAWMIYTYSGLFCVTVNPYKWLPVYNPEVVLAYRGKKRQEAPPHIFSISDNA
YQFMLTDRENQSILITGESGAGKTVNTKRVIQYFATIAASGEKKKEEQTSGKMQG
TLEDQIISANPLLEAFGNAKTVRNDNSSRFGKFIRIHFGATGKLASADIETYLLEKS
RVTFQLKAERSYHIFYQIMSNKKPELIDMLLITTNPYDYQFVSQGEITVASINDQE
ELMATDSAIDILGFSADEKTAIYKLTGAVMHYGNLKFKQKQREEQAEPDGTEVA
DKAAYLMGLNSADLLKALCYPRVKVGNEYVTKGQTVQQVYNSVGALAKAVYE
KMFLWMVVRINEQLDTKQPRQYFIGVLDIAGFEIFDFNSLEQLCINFTNEKLQQFF
NHHMFVLEQEEYKKEGIEWTFIDFGMDLAACIELIEKPMGIFSILEEECMFPKATD
TSFKNKLYDQHLGKSNNFQKPKPAKGKAEAHFSLVHYAGTVDYNITGWLEKNK
DPLNETVIGLYQKSENLNKLMTNLRSTHPHFVRCIIPNETKTPGAMEHELVLHQL
RCNGVLEGIRICRKGFPSRVLYADFKQRYKVLNASAIPEGQFIDSKKASEKLLGSI
DVDHTQYKFGHTKVFFKAGLIGILEEMRDEKLAQLITRTQARCRGFLMRVEYQR
MVERRESIFCIQYNIRAFMNVKHWPWMKLFFKIKPLLKSAESEKEMANMKEEFE
KTKEELAKSEAKRKELEEKMASLMKEKNDLQLQVQAEADALADAEERCDQLIK
TKIQLEAKVKEVTERAEDEEEINAELTAKKRKLEDECSELKKDIDDLELTLAKVE
KEKHATENKVKNLTEEMAALDETIAKLTKEKKALQEAHQQTLDDLQAEEDKVN
TLTKAKTKLEQQVDDLEGSLEQEKKLRMDLERAKRKLEGDLKLAQDSIMDLEN
DKQQLDEKLKKKDFEISQIQSKIEDEQALGMQLQKKIKELQARIEELEEEIEAERT

SRAKAEKHRADLSRELEEISERLEEAGGATAAQVEMNKKREAEFQKMRRDLEEA
TLQHEATASALRKKHADSTAELGEQIDNLQRVKQKLEKEKSEMKMEIDDLASN
MESVSKAKANLEKMCRTLEDQLSEIKTKEEEHQRMINDLNAQRARLQTEAGEFS
RQVDEKDALISQLSRGKQAFTQQIEELKRHLEEEIKAKNALAHALQSARHDCDLL
REQYEEEQEAKGELQRALSKANSEVAQWRTKYETDAIQRTEELEEAKKKLAQRL
QDAEEHVEAVNAKCASLEKTKQRLQNEVEDLMIDVERSNAACAALDKKQKNFD
KILAEWKQKYEETQAELEASQKESRSLSTELFKMKNAYEESLDHLETMKRENKN
LQQEISDLTEQIAEGGKAIHELEKVKKQIEQEKSEIQAALEEAEASLEHEEGKILRL
QLELNQVKSEIDRKIAEKDEEIDQMKRNHLRIVDSMQSTLDAEIRSRNEALRLKK
KMEGDLNEMEIQLSHANRVAAEAQKNLRNTQAVLKDTQIHLDDALRTQEDLKE
QVAMVERRANLLQAEIEELRAALEQTERSRKVAEQELLDASERVQLLHTQNTSLI
NTKKKLETDIAQIQGEMEDTIQEARNAEEKAKKAITDAAMMAEELKKEQDTSAH
LERMKKNLDQTVKDLQHRLDEAEQLALKGGKKQIQKLEARVRELEGEVDAEQK
RSAEAVKGVRKYERRVKELTYQSEEDRKNILRLQDLEDKLQRKVKSYKRQAEE
AEELSNVNLSKFRKIQHELEEAEERADIAESQVNKLRAKSREISKKIEEEE

Chicken

ASPDAEMAAFGEAAPYLRKSEKERIEAQNKPFDAKSSVFVVHPKESFVKGTIQSK
ETGKVTVKTEGGETLTVKEDQIFSMNPPKYDKIEDMAMMTHLHEPAVLYNLKE
RYAAWMIYTYSGLFCVTVNPYKWLPVYNPEVVLAYRGKKRQEAPPHIFSISDNA
YQFMLTDRENQSILITGESGAGKTVNTKRVIQYFATIAASGEKKKEEQP-
GKMQGTLEDQIISANPLLEAFGNAKTVRNDNSSRFGKFIRIHFGATGKLASADIET
YLLEKSRVTFQLKAERSYHIFYQIMSNKKPELIDMLLITTNPYDYHFVSQGEITVP
SINDQEELMATDSAIDILGFTADEKVAIYKLTGAVMHYGNLKFKQKQREEQAEP
DGTEVADKAAYLMGLNSADLLKALCYPRVKVGNEYVTKGQTVQQVNNSVGAL
AKAVYEKMFLWMVVRINQQLDTKQPRQYFIGVLDIAGFEIFDFNSLEQLCINFTN
EKLQQFFNHHMFVLEQEEYKKEGIEWEFIDFGMDLAACIELIEKPMGIFSILEEEC
MFPKATDTSFKNKLYDQHLGKSSNFQKPKPAKGKAEAHFSLVHYAGTVDYNIT
GWLEKNKDPLNETVIGLYQKSENLNKLMTNLRSTHPHFVRCIIPNETKTPGAMEH
ELVLHQLRCNGVLEGIRICRKGFPSRVLYADFKQRYKVLNASAIPEGQFIDSKKAS
EKLLGSIDVDHTQYKFGHTKVFFKAGLLGLLEEMRDEKLAQLITRTQARCRGFL
MRVEYQRMVERRESIFCIQYNVRAFMNVKHWPWMKLFFKIKPLLKSAESEKEM
ANMKEEFEKTKEELAKSEAKRKELEEKMVKLVQEKNDLQLQVQAEADALADA

EERCDQLIKTKIQLEAKIKEVTERAEDEEEINAELTAKKRKLEDECSELKKDIDDL
ELTLAKVEKEKHATENKVKNLTEEMAALDETIAKLTKEKKALQEAHQQTLDDL
QAEEDKVNTLTKAKTKLEQQVDDLEGSLEQEKKLRMDLERAKRKLEGDLKMSQ
DTIMDLENDKQQLDEKLKKKDFEISQIQSKIEDEQALGMQLQKKIKELQARIEELE
EEIEAERTSRAKAEKHRADLSRELEEISERLEEAGGATAAQIDMNKKREAEFQKM
RRDLEEATLQHEATAAALRKKHADSTAELGEQIDNLQRVKQKLEKEKSELKMEI
DDLASNMESVSKAKANLEKMCRTLEDQLSEIKSKEEEHQRMINDLSTQRARLQT
ESGEYSRQVEEKDALISQLSRGKQAFTQQIEELKRHLEEEIKAKNALAHALQSAR
HDCDLLREQYEEEQEAKGELQRALSKANSEVAQWRTKYETDAIQRTEELEEAKK
KLAQRLQDAEEHVEAVNAKCASLEKTKQRLQNEVEDLMIDVERANAACAALD
KKQKNFDKILAEWKQKYEETQAELEASQKESRSLSTELFKMKNAYEESLDHLET
LKRENKNLQQEISDLTEQIAEGGKAIHELEKVKKQIEQEKSEIQAALEEAEASLEH
EEGKILRLQLELNQVKSEIDRKIAEKDEEIDQLKRNHLRIVESMQSTLDAEIRSRNE
ALRLKKKMEGDLNEMEIQLNHANRVAAEAQKNLRNTQGVLKDTQIHLDDALRT
QEDLKEQVAMVERRANLLQAEIEELRAALEQTERSRKVAEQELMDASERVQLL
HTQNTSLINTKKKLETDIAQIQSEMEDTIQEARNAEEKAKKAITDAAMMAEELKK
EQDTSAHLERMKKNLDQTVKDLQLRLDEAEQLALKGGKKQIQKLEARVRELEG
EVDAEQKRSAEAVKGVRKYERRVKELTYQSEEDRKNILRLQDLVDKLQMKVKS
YKRQAEEAEELSNVNLTKFRKIQHELEEAEERADIAESQVNKLRAKSREFHKKIE
EEE

LTCA

MSSDAEMAVFGVAAPYLRKSEKERIEAQNRPFDAKTSVFVVDPKEMYVKGTIQS
KEGGKVTVKTEDGKTLTVKDDEVFPMNPPKYDKIEDMAMMTHLNEPSVLYNLK
ERYAAWMIYTYSGLFCVTVNPYKWLPVYNPEVVAAYRGKKRQEAPPHIFSISDN
AYQFMLTDRENQSILITGESGAGKTVNTKRVIQYFATIAATGEKKKEEAPPGKMQ
GTLEDQIIQANPLLEAFGNAKTVRNDNSSRFGKFIRIHFGTTGKLASADIETYLLE
KSRVTFQLSAERSYHIFYQIMSNKKPELIEMLLITTNPYDYPFVSQGEITVASIDDQ
EELMATDSAIDILGFTADEKVGIYKLTGAVMHYGNMKFKQKQREEQAEPDGTE
VADKAAYLMGLNSADLLKALCYPRVKVGNEYVTKGQTVQQVYNSVGALAKSV
YEKMFLWMVVRINQQLDTKQPRQYFIGVLDIAGFEIFDFNSLEQLCINFTNEKLQ
QFFNHHMFVLEQEEYKKEGIEWEFIDFGMDLAACIELIEKPMGIFSILEEECMFPK
ATDTSFKNKLYDQHLGKSNNFQKPKPAKGKAEAHFSLVHYAGTVDYNISGWLD

KNKDPLNETVVGLYQKSENLNKLMTNLRSTHPHFVRCIIPNETKTPGAMDHQLV
LHQLRCNGVLEGIRICRKGFPSRILYGDFKQRYKVLNASAIPEGQFIDSKKASEKL
LGSIDVDHTQYKFGHTKVFFKAGLLGTLEEMRDDKLAQLITRTQAMCRGYLMR
VEFQKMMERRESIFCIQYNIRSFMNVKHWPWMKLYFKIKPLLKSAESEKEMANM
KEEFEKTKEELAKSEAKRKELEEKMVTLLQEKNDLQLQVQSETENLADAEERCE
GLIKTKIQLEAKIKELNERLEDEEEMNAELTAKKRKLEDECSELKKDIDDLELTLA
KVEKEKHATENKVKNLTEEMAALDETIAKLTKEKKALQEAHQQTLDDLQAEED
KVNTLTKAKTKLEQQVDDLEGSLEQEKKLRMDLERAKRKLEGDLKLAQESIMD
LENDKQQLDEKLKKKDFEISQLQSKIEDEQALGAQLQKKIKELQARIEELEEEIEA
ERAARAKAEKQRSDLSRELEEISERLEEAGGATSAQIEMNKKREAEFQKMRRDL
EEATLQHEATAAALRKKQADSVAELGEQIDNLQRVKQKLEKEKSELKMEIDDLA
SNMESVSKAKANLEKMCRTLEDQLSEVKTKEDEHQRLINDLSAQRARLQTENGE
FSRQLEEKESLISQLSRGKQAFTQQIEELKRQLEEEIKAKNALAHALQSARHDCDL
LREQYEEEQEAKAELQRSMSKANSEVAQWRTKYETDAIQRTEELEEAKKKLAQ
RLQDAEEQIEAVNSKCASLEKTKQRLQGEVEDLMIDVERSNAACAALDKKQRNF
DKILAEWKQKYEESQAELEASQKESRSLSTELFKMKNAYEESLDHLETLKRENK
NLQQEISDLTEQIAESGKAIHELEKAKKQIEQEKSELQAALEEAEASLEHEEGKIL
RIQLELNQVKSEIDRKIAEKDEEIDQLKRNHQRVVESMQSTLDAEIRSRNDALRIK
KKMEGDLNEMEIQLSHANRQAAEAQKHLRNVQGQLKDTQLHLDDALRGQEDL
KEQLAMVERRNNLMQAEIEELRAALEQTERGRKVAEQELIDASERVQLLHSQNT
SLINTKKKLEADISQLQGEVEDAIQEARNAEEKAKKAITDAAMMAEELKKEQDT
SAHLERMKKNLDQTVKDLQHRLDEAEQLALKGGKKQLQKLEARVRELESELEA
EQKRGADAIKGVRKYERRVKELTYQSEEDRKNVLRLQDLVDKLQLKVKAYKRQ
AEEAEEQANVNLSRFRKVQHELEEAEERADIAESQVNKLRAKSRDIGTKVVSEE

LSCA

MSSDAEMAVFGVAAPYLRKSEKERIEAQNKPFDAKTSVFVVDPKESFVKGTIQS
REGGKVTVKTEGGATLTVKDDQVFPMNPPKYDKIEDMAMMTHLHEPAVLYNL
KERYAAWMIYTYSGLFCVTVNPYKWLPVYNPEVVAAYRGKKRQEAPPHIFSISD
NAYQFMLTDRENQSILITGESGAGKTVNTKRVIQYFATIAATGEKKKEEPPSGKM
QGTLEDQIISANPLLEAFGNAKTVRNDNSSRFGKFIRIHFGTTGKLASADIETYLLE
KSRVTFQLKAERSYHIFYQIMSNKKPELIEMLLITTNPYDYPFVSQGEITVASIDDQ
EELMATDSAIDILGFTADEKVAIYKLTGAVMHYGNMKFKQKQREEQAEPDGTE

VADKAAYLMNLNSADLLKALCYPRVKVGNEYVTKGQTVQQVYNSVGALAKA
VYEKMFLWMVVRINQQLDTKQPRQYFIGVLDIAGFEIFDFNSLEQLCINFTNEKL
QQFFNHHMFVLEQEEYKKEGIEWEFIDFGMDLAACIELIEKPMGIFSILEEECMFP
KATDTSFKNKLYDQHLGKSNNFQKPKPAKGKAEAHFSLVHYAGTVDYNITGWL
DKNKDPLNETVVGLYQKSENLNKLMTNLRSTHPHFVRCIIPNETKTPGAMEHEL
VLHQLRCNGVLEGIRICRKGFPSRILYADFKQRYKVLNASAIPEGQFIDSKKASEK
LLGSIDVDHTQYKFGHTKVFFKAGLLGLLEEMRDDKLAQLITRTQAMCRGYLM
RVEYQKMMERRESIFCIQYNIRSFMNVKHWPWMKLYFKIKPLLKSAESEKEMAN
MKEEFEKTKEELAKSEAKRKELEEKMVSLMQEKNDLQLQVQSEAEGLADAEER
CDQLIKTKIQLEAKIKELTERAEDEEEMNAELTAKKRKLEDECSELKKDIDDLELT
LAKVEKEKHATENKVKNLTEEMAALDETIAKLTKEKKALQEAHQQTLDDLQAE
EDKVNTLTKAKTKLEQQVDDLEGSLEQEKKLRMDLERAKRKLEGDLKLAQESI
MDLENDKQQLDEKLKKKDFEISQLQSKIEDEQALGMQLQKKIKELQARIEELEEE
IEAERASRAKAEKQRSDLSRELEEISERLEEAGGATSAQIEMNKKREAEFQKMRR
DLEEATLQHEATAAALRKKHADSVAELGEQIDNLQRVKQKLEKEKSELKMEIDD
LASNMETVSKAKANLEKMCRTLEDQLSEVKTKEEEHQRLINDLSAQRARLQTES
GEFSRQLEEKDSLISQLSRGKQAFTQQIEELKRQLEEEIKAKNALAHALQSARHD
CDLLREQYEEEQEAKAELQRAMSKANSEVAQWRTKYETDAIQRTEELEEAKKK
LAQRLQDAEEHVEAVNSKCASLEKTKQRLQNEVEDLMIDVERSNAACAALDKK
QRNFDKILAEWKQKYEETQAELEASQKESRSLSTELFKMKNAYEESLDHLETLK
RENKNLQQEISDLTEQIAEGGKAIHELEKVKKQIEQEKSELQAALEEAEASLEHEE
GKILRIQLELNQVKSEIDRKIAEKDEEIDQLKRNHLRVVESMQSTLDAEIRSRNDA
LRIKKKMEGDLNEMEIQLSHANRQAAEAQKNLRNTQGILKDTQLHLDDALRGQ
EDLKEQLAMVERRANLMQAEIEELRAALEQTERSRKVAEQELLDASERVQLLHT
QNTSLINTKKKLETDISQIQGEMEDTIQEARNAEEKAKKAITDAAMMAEELKKEQ
DTSAHLERMKKNLDQTVKDLQHRLDEAEQLALKGGKKQIQKLEARVRELEGEV
ENEQKRSAEAIKGVRKYERRVKELTYQSEEDRKNVLRLQDLVDKLQMKVKAYK
RQAEEAEEQSNVNLSKFRKIQHELEEAEERADIAESQVNKLRVKSREIHTKIVSEE

LMCA

MSSDAEMAVFGEAAPYLRKSEKERIEAQNKPFDAKTSVFVVDPKESFVKGTIQSR
EGGKVTVKTEGGATLTVKDDQVFPMNPPKYDKIEDMAMMTHLHEPAVLYNLK
ERYAAWMIYTYSGLFCVTVNPYKWLPVYNPEVVAAYRGKKRQEAPPHIFSISDN

AYQFMLTDRENQSILITGESGAGKTVNTKRVIQYFATIAVTGEKKKEEPPSGKMQ
GTLEDQIISANPLLEAFGNAKTVRNDNSSRFGKFIRIHFGTTGKLASADIETYLLEK
SRVTFQLKAERSYHIFYQIMSNKKPELIEMLLITTNPYDYAFVSQGEITVPSIDDQE
ELMATDSAIDILGFTSDEKVAIYKLTGAVMHYGNMKFKQKQREEQAEPDGTEVA
DKAAYLQNLNSADLLKALCYPRVKVGNEYVTKGQTVQQVYNAVGALAKAVYE
KMFLWMVTRINQQLDTKQPRQYFIGVLDIAGFEIFDFNSLEQLCINFTNEKLQQFF
NHHMFVLEQEEYKKEGIEWEFIDFGMDLAACIELIEKPMGIFSILEEECMFPKATD
TSFKNKLYEQHLGKSNNFQKPKPAKGKAEAHFSLVHYAGTVDYNITGWLDKNK
DPLNETVVGLYQKSENLNKLMTNLRSTHPHFVRCIIPNETKTPGAMEHELVLHQL
RCNGVLEGIRICRKGFPSRILYADFKQRYKVLNASAIPEGQFIDSKKASEKLLGSID
IDHTQYKFGHTKVFFKAGLLGLLEEMRDDKLAQLITRTQAMCRGYLMRVEYQK
MMERRESIFCIQYNIRAFMNVKHWPWMKLYFKIKPLLKSAETEKEMANMKEEFE
KTKEELAKSEAKRKELEEKMVSLMQEKNDLQLQVQSEAEGLADAEERCDQLIK
TKIQLEAKIKEVTERAEDEEEINAELTAKKRKLEDECSELKKDIDDLELTLAKVEK
EKHATENKVKNLTEEMAGLDETIAKLTKEKKALQEAHQQTLDDLQAEEDKVNT
LTKAKTKLEQQVDDLEGSLEQEKKLRMDLERAKRKLEGDLKLAQESIMDIENDK
QQLDEKLKKKEFEMSNLQSKIEDEQALGMQLQKKIKELQARIEELEEEIEAERAS
RAKAEKQRSDLSRELEEISERLEEAGGATSAQIEMNKKREAEFQKMRRDLEEATL
QHEATAAALRKKHADSVAELGEQIDNLQRVKQKLEKEKSELKMEIDDLASNME
TVSKAKANLEKMCRTLEDQLSEVKTKEEEQQRLINDLSAQRARLQTESGEFSRQ
LDEKDSLVSQLSRGKQAFTQQIEELKRQLEEEIKAKNALAHALQSARHDCDLLRE
QYEEEQEAKAELQRAMSKANSEVAQWRTKYETDAIQRTEELEEAKKKLAQRLQ
DAEEHVEAVNSKCASLEKTKQRLQNEVEDLMIDVERSNAACAALDKKQRNFDK
ILAEWKQKYEETQAELEASQKESRSLSTELFKMKNAYEESLDQLETLKRENKNL
QQEISDLTEQIAEGGKRIHELEKIKKQIEQEKSELQAALEEAEASLEHEEGKILRIQ
LELNQVKSEIDRKIAEKDEEIDQLKRNHLRVVESMQSTLDAEIRSRNDALRIKKK
MEGDLNEMEIQLNHANRQAAEALRNLRNTQGILKDTQLHLDDALRGQEDLKEQ
LAMVERRANLMQAEIEELRATLEQTERSRKVAEQELLDASERVQLLHTQNTSLIN
TKKKLETDISQIQGEMEDIVQEARNAEEKAKKAITDAAMMAEELKKEQDTSAHL
ERMKKNLEQTVKDLQHRLDEAEQLALKGGKKQIQKLEARVRELEGEVENEQKR
NVEAIKGLRKHERRVKELTYQTEEDRKNVLRLQDLVDKLQTKVKAYKRQAEEA
EEQSNVNLSKFRKIQHELEEAEERADIAESQVNKLRVKSREVHTKIISEE

LPMCA

MSSDQEMAIFGEAAPYLRKSEKERIEAQNRPFDAKTSVFVVEPKESFVKGTIQSRE
GGKVTVKTEAGATLTVKDDQVFPMNPPKYDKIEDMAMMTHLHEPAVLYNLKE
RYAAWMIYTYSGLFCVTVNPYKWLPVYNPEVVAAYRGKKRQEAPPHIFSISDNA
YQFMLTDRENQSILITGESGAGKTVNTKRVIQYFATIAVTGEKKKEEPTSGKMQG
TLEDQIISANPLLEAFGNAKTVRNDNSSRFGKFIRIHFGTTGKLASADIETYLLEKS
RVTFQLKAERSYHIFYQIMSNKKPELIEMLLITTNPYDYPFVSQGEITVPSIDDQEE
LMATDSAIDILGFTNEEKVSIYKLTGAVMHYGNMKFKQKQREEQAEPDGTEVAD
KAAYLQGLNSADLLKALCYPRVKVGNEYVTKGQTVEQVTNAVGALAKAVYEK
MFLWMVTRINQQLDTKQPRQYFIGVLDIAGFEIFDFNSLEQLCINFTNEKLQQFFN
HHMFVLEQEEYKKEGIEWEFIDFGMDLAACIELIEKPMGIFSILEEECMFPKATDT
SFKNKLYEQHLGKSNNFQKPKVVKGKAEAHFSLVHYAGTVDYNITGWLDKNK
DPLNETVVGLYQKSENLNKLMTNLRSTHPHFVRCIIPNETKTPGAMEHELVLHQL
RCNGVLEGIRICRKGFPSRILYADFKQRYKVLNASAIPEGQFIDSKKASEKLLASID
IDHTQYKFGHTKVFFKAGLLGLLEEMRDDKLAQLITRTQARCRGFLARVEYQKM
VERRESIFCIQYNIRSFMNVKHWPWMKLFFKIKPLLKSAETEKEMATMKEEFEKT
KEELAKSEAKRKELEEKMVSLLQEKNDLQLQVQSEAEGLADAEERCDQLIKTKI
QLEAKIKEVTERAEDEEEINAELTAKKRKLEDECSELKKDIDDLELTLAKVEKEK
HATENKVKNLTEEMAGLDETIAKLTKEKKALQEAHQQTLDDLQAEEDKVNTLT
KAKIKLEQQVDDLEGSLEQEKKLRMDLERAKRKLEGDLKLAQESIMDIENEKQQ
LDEKLKKKEFEMSNLQSKIEDEQALGMQLQKKIKELQARIEELEEEIEAERASRA
KAEKQRSDLSRELEEISERLEEAGGATSAQIEMNKKREAEFQKMRRDLEEATLQH
EATAAALRKKHADSVAELGEQIDNLQRVKQKLEKEKSEMKMEIDDLASNVETV
SKAKGNLEKMCRTLEDQVSELKSKEEEHQRLINDLTAQRGRLQTESGEFSRQLD
EKEALVSQLSRGKQAFTQQIEELKRQLEEEIKAKNALAHALQSARHDCDLLREQ
YEEEQESKAELQRALSKANSEVAQWRTKYETDAIQRTEELEEAKKKLAQRLQDA
EEHVEAVNAKCASLEKTKQRLQNEVEDLMLDVERTNAACAALDKKQRNFDKIL
AEWKQKYEETHAELEASQKEARSLGTELFKMKNAYEESLDQLETLKRENKNLQ
QEISDLTEQIAEGGKRIHELEKIKKQIEQEKSELQAALEEAEASLEHEEGKILRIQLE
LNQVKSEIDRKIAEKDEEIDQLKRNHIRVVESMQSTLDAEIRSRNDAIRIKKKMEG
DLNEMEIQLNHANRMAAEALRNYRNTQGILKDTQLHLDDALRGQEDLKEQLAM
VERRANLLQAEIEELRATLEQTERSRKVAEQELLDASERVQLLHTQNTSLINTKK
KLETDISQIQGEMEDIVQEARNAEEKAKKAITDAAMMAEELKKEQDTSAHLERM

KKNMEQTVKDLQHRLDEAEQLALKGGKKQIQKLEARVRELEGEVESEQKRNAE
AVKGLRKHERRVKELTYQTEEDRKNILRLQDLVDKLQAKVKSYKRQAEEAEEQ
SNTNLSKFRKLQHELEEAEERADIAESQVNKLRVKSREVHTKIISEE

Appendix V

Estimation of force-clamp lengths for unfolding and reductions of the titin domains I65 to I72.

The columns indicate species, number of residues of the domain, expected length of full domain and possible pairs of cysteines forming a disulfide bond. The length of the unfolding up to the disulfide bond is indicated as $L_{S\text{-}S}$, and the reduction associated to the particular disulfide bond as L_{red} in square brackets. Lengths have been estimated as reported elsewhere [174].

I65

Species	N aa	L full domain (nm)	$L_{S\text{-}S}$ [L_{red}] (nm)		
			23-74	23-85	74-85
LTCA	89	26	9 [17]	5 [21]	23 [3]
LSCA	89	26	9 [17]	-	-
LMCA	89	26	9 [17]	-	-
LPMCA	89	26	9 [17]	-	-
Zebra finch	89	26	9 [17]	-	-
Human	89	26	9 [17]	-	-
Whale	89	26	9 [17]	-	-
Rat	89	26	9 [17]	-	-
Chicken	89	26	9 [17]	-	-

I66

Species	N aa	L full domain (nm)	$L_{S\text{-}S}$ [L_{red}] (nm)		
			22-73	22-84	73-84
LTCA	88	26	9 [17]	5 [21]	23 [3]
LSCA	88	26	9 [17]	5 [21]	23 [3]
LMCA	88	26	9 [17]	5 [21]	23 [3]
LPMCA	88	26	9 [17]	5 [21]	23 [3]
Zebra finch	88	26	9 [17]	5 [21]	23 [3]
Human	88	26	9 [17]	5 [21]	23 [3]
Whale	88	26	9 [17]	5 [21]	23 [3]
Rat	88	26	9 [17]	5 [21]	23 [3]
Chicken	88	26	9 [17]	5 [21]	23 [3]

I67

Species	N aa	L full domain (nm)	$L_{S\text{-}S}$ [L_{red}] (nm)
			74-85
LTCA	89	26	23 [3]
LSCA	89	26	23 [3]
LMCA	89	26	23 [3]
LPMCA	89	26	23 [3]
Zebra finch	89	26	23 [3]
Human	89	26	23 [3]
Whale	89	26	23 [3]
Rat	89	26	23 [3]
Chicken	89	26	23 [3]

I68

Species	N aa	L full domain (nm)	$L_{S\text{-}S}$ [L_{red}] (nm)		
			19-85	23-85	19-23
LTCA	91	27	5 [22]	6 [21]	26 [1]
LSCA	91	27	5 [22]	6 [21]	26 [1]
LMCA	91	27	5 [22]	6 [21]	26 [1]
LPMCA	91	27	5 [22]	6 [21]	26 [1]
Zebra finch	91	27	5 [22]	6 [21]	26 [1]
Human	91	27	5 [22]	6 [21]	26 [1]
Whale	91	27	5 [22]	6 [21]	26 [1]
Rat	91	27	5 [22]	6 [21]	26 [1]
Chicken	91	27	5 [22]	6 [21]	26 [1]

I69

Species	N aa	L full domain (nm)	$L_{S\text{-}S}$ [L_{red}] (nm)					
			22-73	22-84	73-84	22-56	56-73	56-84
LTCA	88	26	9 [17]	5 [21]	23 [3]	14 [12]	20 [6]	17 [9]
LSCA	88	26	9 [17]	5 [21]	23 [3]	14 [12]	20 [6]	17 [9]
LMCA	88	26	9 [17]	5 [21]	23 [3]	14 [12]	20 [6]	17 [9]
LPMCA	88	26	9 [17]	5 [21]	23 [3]	14 [12]	20 [6]	17 [9]
Zebra finch	88	26	9 [17]	5 [21]	23 [3]	14 [12]	20 [6]	17 [9]
Human	88	26	9 [17]	5 [21]	23 [3]	14 [12]	20 [6]	17 [9]
Whale	88	26	9 [17]	5 [21]	23 [3]	14 [12]	20 [6]	17 [9]
Rat	88	26	9 [17]	5 [21]	23 [3]	14 [12]	20 [6]	17 [9]
Chicken	88	26	9 [17]	5 [21]	23 [3]	14 [12]	20 [6]	17 [9]

I70

Species	N aa	L full domain (nm)	$L_{S\text{-}S}$ [L_{red}] (nm)		
			47-73	47-84	73-84
LTCA	88	26	17 [9]	13 [13]	23 [3]
LSCA	88	26	17 [9]	13 [13]	23 [3]
LMCA	88	26	17 [9]	13 [13]	23 [3]
LPMCA	88	26	17 [9]	13 [13]	23 [3]
Zebra finch	88	26	17 [9]	13 [13]	23 [3]
Human	88	26	17 [9]	13 [13]	23 [3]
Whale	88	26	17 [9]	13 [13]	23 [3]
Rat	88	26	17 [9]	13 [13]	23 [3]
Chicken	88	26	17 [9]	13 [13]	23 [3]

I71

Species	N aa	L full domain (nm)	$L_{S\text{-}S}$ [L_{red}] (nm)				
			48-74	48-85	74-85	23-48	23-74
LTCA	89	26	17 [9]	14 [12]	23 [3]	17 [9]	9 [17]
LSCA	89	26	17 [9]	14 [12]	23 [3]	17 [9]	9 [17]
LMCA	89	26	17 [9]	14 [12]	23 [3]	-	-
LPMCA	89	26	17 [9]	14 [12]	23 [3]	-	-
Zebra finch	89	26	17 [9]	14 [12]	23 [3]	18 [9]	9 [17]
Human	89	26	17 [9]	14 [12]	23 [3]	-	-
Whale	89	26	17 [9]	14 [12]	23 [3]	-	-
Rat	89	26	17 [9]	14 [12]	23 [3]	-	-
Chicken	89	26	17 [9]	14 [12]	23 [3]	17 [9]	9 [17]

Species	N aa	L full domain (nm)	$L_{S\text{-}S}$ [L_{red}] (nm)				
			23-85	23-57	48-57	57-74	57-85
LTCA	89	26	5 [21]	15 [11]	23 [3]	20 [6]	17 [19]
LSCA	89	26	5 [21]	-	-	-	-
LMCA	89	26	-	-	-	-	-
LPMCA	89	26	-	-	-	-	-
Zebra finch	89	26	5 [21]	-	-	-	-
Human	89	26	-	-	-	-	-
Whale	89	26	-	-	-	-	-
Rat	89	26	-	-	-	-	-
Chicken	89	26	5 [21]	-	-	-	-

I72

Species	N aa	L full domain (nm)	L_{S-S} [L_{red}] (nm)		
			22-73	22-47	47-73
LTCA	88	26	9 [17]	-	-
LSCA	88	26	9 [17]	-	-
LMCA	88	26	9 [17]	-	-
LPMCA	88	26	9 [17]	-	-
Zebra finch	88	26	9 [17]	-	-
Human	88	26	9 [17]	-	-
Whale	88	26	9 [17]	-	-
Rat	88	26	9 [17]	17 [9]	17 [9]
Chicken	88	26	9 [17]	-	-

Bibliography

1. Basmajian, J.V. and C.J. De Luca, *Muscles alive: their functions revealed by electromyography*. 1985: Williams & Wilkins.
2. Hanson, J. and H.E. Huxley, *Structural basis of the cross-striations in muscle*. Nature, 1953. **172** (4377): 530-532.
3. Huxley, H., *Electron microscope studies of the organisation of the filaments in striated muscle.* Biochim Biophys Acta, 1953. **12** (1-2): 387-394.
4. Huxley, A.F. and R. Niedergerke, *Structural changes in muscle during contraction; interference microscopy of living muscle fibres.* Nature, 1954. **173** (4412): 971.
5. Huxley, H. and J. Hanson, *Changes in the cross-striations of muscle during contraction and stretch and their structural interpretation.* Nature, 1954. **173** (4412): 973.
6. Wang, K., J. McClure, and A. Tu, *Titin: major myofibrillar components of striated muscle.* Proc Natl Acad Sci U S A, 1979. **76** (8): 3698-3702.
7. Furst, D.O., M. Osborn, R. Nave, and K. Weber, *The organization of titin filaments in the half-sarcomere revealed by monoclonal antibodies in immunoelectron microscopy: a map of ten nonrepetitive epitopes starting at the Z line extends close to the M line.* J Cell Biol, 1988. **106** (5): 1563-1572.
8. *Huxleys' Missing Filament: Form and Function of Titin in Vertebrate Striated Muscle* Annu Rev Physiol, 2017 **79**:145-166
9. Bang, M.-L., T. Centner, F. Fornoff, A.J. Geach, M. Gotthardt, M. McNabb, et al., *The complete gene sequence of titin, expression of an unusual≈ 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system.* Circ Res, 2001. **89** (11): 1065-1072.
10. Wang, K., R. Ramirez-Mitchell, and D. Palter, *Titin is an extraordinarily long, flexible, and slender myofibrillar protein.* Proc Natl Acad Sci U S A, 1984. **81** (12): 3685-3689.
11. Labeit, S. and B. Kolmerer, *Titins: giant proteins in charge of muscle ultrastructure and elasticity.* Science, 1995. **270** (5234): 293-296.
12. Houmeida, A., J. Holt, L. Tskhovrebova, and J. Trinick, *Studies of the interaction between titin and myosin.* J Cell Biol, 1995. **131** (6): 1471-1481.
13. Eilertsen, K.J., S.T. Kazmierski, and T. Keller, *Cellular titin localization in stress fibers and interaction with myosin II filaments in vitro.* J Cell Biol, 1994. **126** (5): 1201-1210.
14. Improta, S., A.S. Politou, and A. Pastore, *Immunoglobulin-like modules from titin I-band: extensible components of muscle elasticity.* Structure, 1996. **4** (3): 323-337.
15. Labeit, S., M. Gautel, A. Lakey, and J. Trinick, *Towards a molecular understanding of titin.* EMBO J, 1992. **11** (5): 1711.
16. Sanger, J.W. and J.M. Sanger, *Fishing out proteins that bind to titin.* J Cell Biol, 2001. **154** (1): 21-24.
17. Ryle, A.P., F. Sanger, L.F. Smith, and R. Kitai, *The disulphide bonds of insulin.* Biochem J, 1955. **60** (4): 541-556.
18. Kellermayer, M.S., S.B. Smith, H.L. Granzier, and C. Bustamante, *Folding-unfolding transitions in single titin molecules characterized with laser tweezers.* Science, 1997. **276** (5315): 1112-1116.
19. Li, H., W.A. Linke, A.F. Oberhauser, M. Carrion-Vazquez, J.G. Kerkvliet, H. Lu, et al., *Reverse engineering of the giant muscle protein titin.* Nature, 2002. **418** (6901): 998-1002.

20. Hsin, J., J. Strumpfer, E.H. Lee, and K. Schulten, *Molecular origin of the hierarchical elasticity of titin: simulation, experiment, and theory.* Annu Rev Biophys, 2011. **40**: 187-203.
21. Neagoe, C., M. Kulke, F. del Monte, J.K. Gwathmey, P.P. de Tombe, R.J. Hajjar, et al., *Titin isoform switch in ischemic human heart disease.* Circulation, 2002. **106** (11): 1333-1341.
22. Makarenko, I., C. Opitz, M. Leake, C. Neagoe, M. Kulke, J. Gwathmey, et al., *Passive stiffness changes caused by upregulation of compliant titin isoforms in human dilated cardiomyopathy hearts.* Circ Res, 2004. **95** (7): 708-716.
23. Rivas-Pardo, J.A., E.C. Eckels, I. Popa, P. Kosuri, W.A. Linke, and J.M. Fernandez, *Work Done by Titin Protein Folding Assists Muscle Contraction.* Cell Rep, 2016. **14** (6): 1339-1347.
24. Rief, M., M. Gautel, F. Oesterhelt, J.M. Fernandez, and H.E. Gaub, *Reversible unfolding of individual titin immunoglobulin domains by AFM.* Science, 1997. **276** (5315): 1109-1112.
25. Alegre-Cebollada, J., P. Kosuri, D. Giganti, E. Eckels, J.A. Rivas-Pardo, N. Hamdani, et al., *S-glutathionylation of cryptic cysteines enhances titin elasticity by blocking protein folding.* Cell, 2014. **156** (6): 1235-1246.
26. Lee, G.U., L.A. Chrisey, and R.J. Colton, *Direct measurement of the forces between complementary strands of DNA.* Science, 1994. **266** (5186): 771.
27. Li, H. and J.M. Fernandez, *Mechanical design of the first proximal Ig domain of human cardiac titin revealed by single molecule force spectroscopy.* J Mol Biol, 2003. **334** (1): 75-86.
28. Tskhovrebova, L., J. Trinick, J.A. Sleep, and R.M. Simmons, *Elasticity and unfolding of single molecules of the giant muscle protein titin.* Nature, 1997. **387** (6630): 308-312.
29. Carrion-Vazquez, M., A.F. Oberhauser, S.B. Fowler, P.E. Marszalek, S.E. Broedel, J. Clarke, et al., *Mechanical and chemical unfolding of a single protein: a comparison.* Proc Natl Acad Sci U S A, 1999. **96** (7): 3694-3699.
30. Fowler, S.B., R.B. Best, J.L.T. Herrera, T.J. Rutherford, A. Steward, E. Paci, et al., *Mechanical unfolding of a titin Ig domain: structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering.* J Mol Biol, 2002. **322** (4): 841-849.
31. Brockwell, D.J., G.S. Beddard, J. Clarkson, R.C. Zinober, A.W. Blake, J. Trinick, et al., *The effect of core destabilization on the mechanical resistance of I27.* Biophys J, 2002. **83** (1): 458-472.
32. Carrion-Vazquez, M., P.E. Marszalek, A.F. Oberhauser, and J.M. Fernandez, *Atomic force microscopy captures length phenotypes in single proteins.* Proc Natl Acad Sci U S A, 1999. **96** (20): 11288-11292
33. Cecconi, C., E.A. Shank, C. Bustamante, and S. Marqusee, *Direct observation of the three-state folding of a single protein molecule.* Science, 2005. **309** (5743): 2057-2060.
34. Erickson, H.P., *Reversible unfolding of fibronectin type III and immunoglobulin domains provides the structural basis for stretch and elasticity of titin and fibronectin.* Proc Natl Acad Sci U S A, 1994. **91** (21): 10114-10118.
35. Li, L., H.H.-L. Huang, C.L. Badilla, and J.M. Fernandez, *Mechanical unfolding intermediates observed by single-molecule force spectroscopy in a fibronectin type III module.* J Mol Biol, 2005. **345** (4): 817-826.

36. Williams, P.M., S.B. Fowler, R.B. Best, J.L. Toca-Herrera, K.A. Scott, A. Steward, et al., *Hidden complexity in the mechanical properties of titin.* Nature, 2003. **422** (6930): 446-449.
37. Li, H., A.F. Oberhauser, S.B. Fowler, J. Clarke, and J.M. Fernandez, *Atomic force microscopy reveals the mechanical design of a modular protein.* Proc Natl Acad Sci U S A, 2000. **97** (12): 6527-6531
38. Rief, M., M. Gautel, A. Schemmel, and H.E. Gaub, *The mechanical stability of immunoglobulin and fibronectin III domains in the muscle protein titin measured by atomic force microscopy.* Biophys J, 1998. **75** (6): 3008-3014.
39. Watanabe, K., C. Muhle-Goll, M.S. Kellermayer, S. Labeit, and H. Granzier, *Different molecular mechanics displayed by titin's constitutively and differentially expressed tandem Ig segments.* J Struct Biol, 2002. **137** (1-2): 248-258.
40. Watanabe, K., C. Muhle-Goll, M.S. Kellermayer, S. Labeit, and H. Granzier, *Different molecular mechanics displayed by titin's constitutively and differentially expressed tandem Ig segments.* J Struct Biol, 2002. **137** (1-2): 248-258.
41. Oberdörfer, Y., H. Fuchs, and A. Janshoff, *Conformational analysis of native fibronectin by means of force spectroscopy.* Langmuir, 2000. **16** (26): 9955-9958.
42. Oberhauser, A.F., C. Badilla-Fernandez, M. Carrion-Vazquez, and J.M. Fernandez, *The mechanical hierarchies of fibronectin observed with single-molecule AFM.* J Mol Biol, 2002. **319** (2): 433-447.
43. Watanabe, K., P. Nair, D. Labeit, M.S. Kellermayer, M. Greaser, S. Labeit, et al., *Molecular mechanics of cardiac titin's PEVK and N2B spring elements.* J Biol Chem, 2002. **277** (13): 11549-11558.
44. Li, H., A.F. Oberhauser, S.D. Redick, M. Carrion-Vazquez, H.P. Erickson, and J.M. Fernandez, *Multiple conformations of PEVK proteins detected by single-molecule techniques.* Proc Natl Acad Sci U S A, 2001. **98** (19): 10682-10686.
45. Sarkar, A., S. Caamano, and J.M. Fernandez, *The elasticity of individual titin PEVK exons measured by single molecule atomic force microscopy.* J Biol Chem, 2005. **280** (8): 6261-6264.
46. Ott, W., M.A. Jobst, C. Schoeler, H.E. Gaub, and M.A. Nash, *Single-molecule force spectroscopy on polyproteins and receptor–ligand complexes: The current toolbox.* J Struct Biol, 2017. **197** (1): 3-12.
47. Best, R.B., D.J. Brockwell, J.L. Toca-Herrera, A.W. Blake, D.A. Smith, S.E. Radford, et al., *Force mode atomic force microscopy as a tool for protein folding studies.* Anal Chim Acta, 2003. **479** (1): 87-105.
48. Lu, H. and K. Schulten, *The key event in force-induced unfolding of Titin's immunoglobulin domains.* Biophys J, 2000. **79** (1): 51-65.
49. Best, R.B., B. Li, A. Steward, V. Daggett, and J. Clarke, *Can non-mechanical proteins withstand force? Stretching barnase by atomic force microscopy and molecular dynamics simulation.* Biophys J, 2001. **81** (4): 2344-2356.
50. Perez-Jimenez, R., A. Alonso-Caballero, R. Berkovich, D. Franco, M.W. Chen, P. Richard, et al., *Probing the effect of force on HIV-1 receptor CD4.* ACS Nano, 2014. **8** (10): 10313-10320.
51. Dietz, H. and M. Rief, *Exploring the energy landscape of GFP by single-molecule mechanical experiments.* Proc Natl Acad Sci U S A, 2004. **101** (46): 16192-16197.
52. Hall, B.G., *Simple and accurate estimation of ancestral protein sequences.* Proc Natl Acad Sci U S A, 2006. **103** (14): 5431-5436.

53. Merkl, R. and R. Sterner, *Ancestral protein reconstruction: techniques and applications.* Biol Chem, 2016. **397** (1): 1-21.
54. Hedges, S.B., J. Marin, M. Suleski, M. Paymer, and S. Kumar, *Tree of life reveals clock-like speciation and diversification.* Mol Biol Evol, 2015. **32** (4): 835-845.
55. Kratzer, J.T., M.A. Lanaspa, M.N. Murphy, C. Cicerchi, C.L. Graves, P.A. Tipton, et al., *Evolutionary history and metabolic insights of ancient mammalian uricases.* Proc Natl Acad Sci U S A, 2014. **111** (10): 3763-3768.
56. Zakas, P.M., H.C. Brown, K. Knight, S.L. Meeks, H.T. Spencer, E.A. Gaucher, et al., *Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction.* Nat Biotechnol, 2016.
57. Gaucher, E.A., S. Govindarajan, and O.K. Ganesh, *Palaeotemperature trend for Precambrian life inferred from resurrected proteins.* Nature, 2008. **451** (7179): 704-707.
58. Perez-Jimenez, R., A. Ingles-Prieto, Z.M. Zhao, I. Sanchez-Romero, J. Alegre-Cebollada, P. Kosuri, et al., *Single-molecule paleoenzymology probes the chemistry of resurrected enzymes.* Nat Struct Mol Biol, 2011. **18** (5): 592-596.
59. Shindyalov, I., N. Kolchanov, and C. Sander, *Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?* Protein Eng, 1994. **7** (3): 349-358.
60. Ingles-Prieto, A., B. Ibarra-Molero, A. Delgado-Delgado, R. Perez-Jimenez, J.M. Fernandez, E.A. Gaucher, et al., *Conservation of protein structure over four billion years.* Structure, 2013. **21** (9): 1690-1697.
61. Risso, V.A., J.A. Gavira, D.F. Mejia-Carmona, E.A. Gaucher, and J.M. Sanchez-Ruiz, *Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian beta-lactamases.* J Am Chem Soc, 2013. **135** (8): 2899-2902.
62. Reisinger, B., J. Sperl, A. Holinski, V. Schmid, C. Rajendran, L. Carstensen, et al., *Evidence for the existence of elaborate enzyme complexes in the Paleoarchean era.* J Am Chem Soc, 2014. **136** (1): 122-129.
63. Akanuma, S., Y. Nakajima, S. Yokobori, M. Kimura, N. Nemoto, T. Mase, et al., *Experimental evidence for the thermophilicity of ancestral life.* Proc Natl Acad Sci U S A, 2013. **110** (27): 11067-11072.
64. Robert, F. and M. Chaussidon, *A palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts.* Nature, 2006. **443** (7114): 969-972.
65. Nguyen, V., C. Wilson, M. Hoemberger, J.B. Stiller, R.V. Agafonov, S. Kutter, et al., *Evolutionary drivers of thermoadaptation in enzyme catalysis.* Science, 2017. **355** (6322): 289-294.
66. Higgins, D.G., S. Labeit, M. Gautel, and T.J. Gibson, *The evolution of titin and related giant muscle proteins.* J Mol Evol, 1994. **38** (4): 395-404.
67. Ohtsuka, S., A. Hanashima, K. Kubokawa, Y. Bao, Y. Tando, J. Kohmaru, et al., *Amphioxus connectin exhibits merged structure as invertebrate connectin in I-band region and vertebrate connectin in A-band region.* J Mol Biol, 2011. **409** (3): 415-426.
68. Hanashima, A., M. Ogasawara, Y. Nomiya, T. Sasaki, Y. Bao, and S. Kimura, *Genomic- and protein-based approaches for connectin (titin) identification in the ascidian Ciona intestinalis.* Methods, 2012. **56** (1): 18-24.
69. Kenny, P.A., E.M. Liston, and D.G. Higgins, *Molecular evolution of immunoglobulin and fibronectin domains in titin and related muscle proteins.* Gene, 1999. **232** (1): 11-23.
70. Ziegler, C., *Titin-related proteins in invertebrate muscles.* Comp Biochem Physiol A Mol Integr Physiol, 1994. **109** (4): 823-833.

71. Hooper, S.L. and J.B. Thuma, *Invertebrate muscles: muscle specific genes and proteins.* Physiol rev, 2005. **85** (3): 1001-1060.
72. Erwin, D.H., M. Laflamme, S.M. Tweedt, E.A. Sperling, D. Pisani, and K.J. Peterson, *The Cambrian conundrum: early divergence and later ecological success in the early history of animals.* Science, 2011. **334** (6059): 1091-1097.
73. Platnick, N.I. and H.D. Cameron, *Cladistic Methods in Textual, Linguistic, and Phylogenetic Analysis.* Syst Zool, 1977. **26** (4): 380-385.
74. Tehrani, J.J., *The phylogeny of Little Red Riding Hood.* PLoS One, 2013. **8** (11): e78871.
75. Walker, R.S., K.R. Hill, M.V. Flinn, and R.M. Ellsworth, *Evolutionary history of hunter-gatherer marriage practices.* PLoS One, 2011. **6** (4): e19066.
76. Schuh, R.T., *Biological systematics: principles and applications*. 2000: Cornell University Press.
77. Folinsbee, K.E., D.C. Evans, J. Fröbisch, L.A. Tsuji, and D.R. Brooks, *5 Quantitative Approaches to Phylogenetics*, in *Handbook of Paleoanthropology*. 2007, Springer.
78. Craw, R., *Margins of Cladistics: Identity, Difference and Place in the Emergence of Phylogenetic Systematics 1864–1975*, in *Trees of Life*. 1992, Springer.
79. Sanger, F., E.O. Thompson, and R. Kitai, *The amide groups of insulin.* Biochem J, 1955. **59** (3): 509-518.
80. Morgan, G.J., *Emile Zuckerkandl, Linus Pauling, and the molecular evolutionary clock, 1959–1965.* J Hist Biol, 1998. **31** (2): 155-178.
81. Pauling, L. and E. Zuckerkandl, *Chemical paleogenetics.* Acta Chem Scand, 1963. **17**: 9-16.
82. Fitch, W.M., *Toward defining the course of evolution: minimum change for a specific tree topology.* Syst Biol, 1971. **20** (4): 406-416.
83. Sankoff, D., *Minimal mutation trees of sequences.* SIAM J Appl Math, 1975. **28** (1): 35-42.
84. Swofford, D.L. and B. Documentation, *Phylogenetic analysis using parsimony.* Illinois Natural History Survey, Champaign, 1989.
85. Eyre-Walker, A., *Problems with parsimony in sequences of biased base composition.* J Mol Evol, 1998. **47** (6): 686-690.
86. Yang, Z., S. Kumar, and M. Nei, *A new method of inference of ancestral nucleotide and amino acid sequences.* Genetics, 1995. **141** (4): 1641-1650.
87. Koshi, J.M. and R.A. Goldstein, *Probabilistic reconstruction of ancestral protein sequences.* J Mol Evol, 1996. **42** (2): 313-320.
88. Pupko, T., I. Pe, R. Shamir, and D. Graur, *A fast algorithm for joint reconstruction of ancestral amino acid sequences.* Mol Biol Evol, 2000. **17** (6): 890-896.
89. Schultz, T.R., R.B. Cocroft, and G.A. Churchill, *The reconstruction of ancestral character states.* Evolution, 1996: 504-511.
90. Huelsenbeck, J.P., B. Larget, and D. Swofford, *A compound Poisson process for relaxing the molecular clock.* Genetics, 2000. **154** (4): 1879-1892.
91. Huelsenbeck, J.P., B. Rannala, and B. Larget, *A Bayesian framework for the analysis of cospeciation.* Evolution, 2000. **54** (2): 352-364.
92. Huelsenbeck, J.P., B. Rannala, and J.P. Masly, *Accommodating phylogenetic uncertainty in evolutionary studies.* Science, 2000. **288** (5475): 2349-2350.
93. Huelsenbeck, J.P., F. Ronquist, R. Nielsen, and J.P. Bollback, *Bayesian inference of phylogeny and its impact on evolutionary biology.* Science, 2001. **294** (5550): 2310-2314.

94. Yokoyama, S. and F.B. Radlwimmer, *The molecular genetics and evolution of red and green color vision in vertebrates.* Genetics, 2001. **158** (4): 1697-1710.
95. Field, S.F. and M.V. Matz, *Retracing evolution of red fluorescence in GFP-like proteins from Faviina corals.* Mol Biol Evol, 2010. **27** (2): 225-233.
96. Wilson, C., R. Agafonov, M. Hoemberger, S. Kutter, A. Zorba, J. Halpin, et al., *Using ancient protein kinases to unravel a modern cancer drug's mechanism.* Science, 2015. **347** (6224): 882-886.
97. Kratzer, J.T., M.A. Lanaspa, M.N. Murphy, C. Cicerchi, C.L. Graves, P.A. Tipton, et al., *Evolutionary history and metabolic insights of ancient mammalian uricases.* Proc Nat Acad Sci U S A, 2014. **111** (10): 3763-3768.
98. Risso, V.A., F. Manssour-Triedo, A. Delgado-Delgado, R. Arco, A. Barroso-delJesus, A. Ingles-Prieto, et al., *Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history.* Mol Biol Evol, 2015. **32** (2): 440-455.
99. Mirceta, S., A.V. Signore, J.M. Burns, A.R. Cossins, K.L. Campbell, and M. Berenbrink, *Evolution of mammalian diving capacity traced by myoglobin net surface charge.* Science, 2013. **340** (6138): 1234192.
100. Lemey, P., A. Rambaut, A.J. Drummond, and M.A. Suchard, *Bayesian phylogeography finds its roots.* PLoS Comput Biol, 2009. **5** (9): e1000520.
101. Clark, J.R., R.H. Ree, M.E. Alfaro, M.G. King, W.L. Wagner, and E.H. Roalson, *A comparative study in ancestral range reconstruction methods: retracing the uncertain histories of insular lineages.* Syst Biol, 2008. **57** (5): 693-707.
102. Bourque, G. and P.A. Pevzner, *Genome-scale evolution: reconstructing gene orders in the ancestral species.* Genome Res, 2002. **12** (1): 26-36.
103. Joy, J.B., R.H. Liang, R.M. McCloskey, T. Nguyen, and A.F. Poon, *Ancestral reconstruction.* PLoS Comput Biol, 2016. **12** (7): e1004763.
104. Swofford, D.L., G.J. Olsen, P.J. Waddell, and D.M. Hillis, *Phylogenetic inference.* 1996.
105. Stamatakis, A., *RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.* Bioinformatics, 2006. **22** (21): 2688-2690.
106. Frumhoff, P.C. and H.K. Reeve, *Using phylogenies to test hypotheses of adaptation: a critique of some current proposals.* Evolution, 1994. **48** (1): 172-180.
107. Cunningham, C.W., *Some limitations of ancestral character-state reconstruction when testing evolutionary hypotheses.* Syst Biol, 1999. **48** (3): 665-674.
108. Felsenstein, J., *Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters.* Syst Biol, 1973. **22** (3): 240-249.
109. Schluter, D., T. Price, A.Ø. Mooers, and D. Ludwig, *Likelihood of ancestor states in adaptive radiation.* Evolution, 1997: 1699-1711.
110. GoJOBORI, T. and S. Yokoyama, *Rates of evolution of the retroviral oncogene of Moloney murine sarcoma virus and of its cellular homologues.* Proc Natl Acad Sci U S A, 1985. **82** (12): 4198-4201.
111. Cunningham, C.W., K.E. Omland, and T.H. Oakley, *Reconstructing ancestral character states: a critical reappraisal.* Trends Ecol Evolut, 1998. **13** (9): 361-366.
112. Li, G., M. Steel, and L. Zhang, *More taxa are not necessarily better for the reconstruction of ancestral character states.* Syst Biol, 2008. **57** (4): 647-653.
113. Dayhoff M.O., S.R.M., Orcutt B.C., *Amino acid scale: Relative mutability of amino acids (Ala=100)*. Atlas of Protein Sequence and Structure. Vol. 5. 1978. National Biomedical Research Foundation.

114. Jones, D.T., W.R. Taylor, and J.M. Thornton, *The rapid generation of mutation data matrices from protein sequences.* CABIOS, 1992. **8** (3): 275-282.
115. Whelan, S. and N. Goldman, *A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.* Mol Biol Evol, 2001. **18** (5): 691-699.
116. Liò, P. and M. Bishop, *Modeling sequence evolution.* Bioinformatics: Data, Sequence Analysis and Evolution, 2008: 255-285. Springer
117. Felsenstein, J., *Evolutionary trees from DNA sequences: a maximum likelihood approach.* J Mol Evol, 1981. **17** (6): 368-376.
118. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24** (8): 1586-1591.
119. Huelsenbeck, J.P. and J.P. Bollback, *Empirical and hierarchical Bayesian estimation of ancestral states.* Syst Biol, 2001. **50** (3): 351-366.
120. Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, *Basic local alignment search tool.* J Mol Biol, 1990. **215** (3): 403-410.
121. Boutet, E., D. Lieberherr, M. Tognolli, M. Schneider, and A. Bairoch, *UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase.* Plant bioinformatics: methods and protocols, 2007: 89-112. Springer
122. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89** (22): 10915-10919.
123. Kanehisa, M., S. Goto, S. Kawashima, and A. Nakaya, *The KEGG databases at GenomeNet.* Nucleic Acids Res, 2002. **30** (1): 42-46.
124. Larkin, M.A., G. Blackshields, N. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23** (21): 2947-2948.
125. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22** (22): 4673-4680.
126. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32** (5): 1792-1797.
127. Kumar, S., M. Nei, J. Dudley, and K. Tamura, *MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences.* Brief Bioinform, 2008. **9** (4): 299-306.
128. Tamura, K., G. Stecher, D. Peterson, A. Filipski, and S. Kumar, *MEGA6: molecular evolutionary genetics analysis version 6.0.* Mol Biol Evol, 2013. **30** (12): 2725-2729.
129. Castresana, J., *Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.* Mol Biol Evol, 2000. **17** (4): 540-552.
130. Wilgenbusch, J.C. and D. Swofford, *Inferring evolutionary trees with PAUP*.* Current Protocols Bioinformatics, 2003: Unit 6.4. Wiley
131. Swofford, D.L., *PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.* 2003.
132. Drummond, A.J., M.A. Suchard, D. Xie, and A. Rambaut, *Bayesian phylogenetics with BEAUti and the BEAST 1.7.* Mol Biol Evol, 2012. **29** (8): 1969-1973.
133. Drummond, A.J. and A. Rambaut, *BEAST: Bayesian evolutionary analysis by sampling trees.* BMC Evol Biol, 2007. **7** (1): 214.
134. Ayres, D.L., A. Darling, D.J. Zwickl, P. Beerli, M.T. Holder, P.O. Lewis, et al., *BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics.* Syst Biol, 2011: syr100.

135. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24** (8): 1586-1591.
136. Yang, Z., *PAML: a program package for phylogenetic analysis by maximum likelihood.* Computer applications in the biosciences: CABIOS, 1997. **13** (5): 555-556.
137. Paradis, E., J. Claude, and K. Strimmer, *APE: analyses of phylogenetics and evolution in R language.* Bioinformatics, 2004. **20** (2): 289-290.
138. Ashkenazy, H., O. Penn, A. Doron-Faigenboim, O. Cohen, G. Cannarozzi, O. Zomer, et al., *FastML: a web server for probabilistic reconstruction of ancestral sequences.* Nucleic Acids Res, 2012. **40**: 580-584.
139. Pond, S.L.K. and S.V. Muse, *HyPhy: hypothesis testing using phylogenies,* in *Statistical methods in molecular evolution*. 2005, 125-181. Springer.
140. Maddison, W. and D. Maddison, *Mesquite: a modular system for evolutionary analysis. Version 2.75. 2011. 2015*. 2016.
141. Ronquist, F. and J.P. Huelsenbeck, *MrBayes 3: Bayesian phylogenetic inference under mixed models.* Bioinformatics, 2003. **19** (12): 1572-1574.
142. Huelsenbeck, J.P. and F. Ronquist, *MRBAYES: Bayesian inference of phylogenetic trees.* Bioinformatics, 2001. **17** (8): 754-755.
143. Hubisz, M.J., K.S. Pollard, and A. Siepel, *PHAST and RPHAST: phylogenetic analysis with space/time models.* Brief Bioinform, 2010: bbq072.
144. http://www.chem-agilent.com/pdf/strata/200249.pdf, *Agilent Technologies. XL1-Blue Competent Cells.*
145. https://tools.thermofisher.com/content/sfs/manuals/MAN0012655_GeneJET_Plasmid_Miniprep_UG.pdf, *Thermo Scientific. GeneJET Plasmid Miniprep.*
146. https://tools.thermofisher.com/content/sfs/manuals/MAN0012661_GeneJET_Gel_Extraction_UG.pdf, *Thermo Scientific. GeneJET Gel Extraction Kit.*
147. https://tools.thermofisher.com/content/sfs/manuals/MAN0011906_DNAsert_Ligation_Vector_DNA_UG.pdf, *Thermo Scientific. T4 ligase DNA Insert Ligation*
148. https://www.merckmillipore.com/GB/en/product/Origami%E2%84%A2-2-Competent-Cells---Novagen, E.B.-. *Origami™ 2 Competent Cells.*
149. Hutter, J.L. and J. Bechhoefer, *Calibration of atomic-force microscope tips.* Rev Sci Instrum, 1993. **64** (7): 1868-1873.
150. Willett, R., K. Baldwin, K. West, and L. Pfeiffer, *Differential adhesion of amino acids to inorganic surfaces.* Proc Natl Acad Sci U S A, 2005. **102** (22): 7817-7822.
151. Marszalek, P.E., H. Li, A.F. Oberhauser, and J.M. Fernandez, *Chair-boat transitions in single polysaccharide molecules observed with force-ramp AFM.* Proc Natl Acad Sci U S A, 2002. **99** (7): 4278-4283.
152. Kratky, O. and G. Porod, *Röntgenuntersuchung gelöster fadenmoleküle.* RRecl. Trav. Chim. Pays-Bas, 1949. **68** (12): 1106-1122.
153. Bouchiat, C., M. Wang, J.-F. Allemand, T. Strick, S. Block, and V. Croquette, *Estimating the persistence length of a worm-like chain molecule from force-extension measurements.* Biophys J, 1999. **76** (1): 409-413.
154. Rogers, L.K., B.L. Leinweber, and C.V. Smith, *Detection of reversible protein thiol modifications in tissues.* Anal Biochem, 2006. **358** (2): 171-184.
155. Kosuri, P., J. Alegre-Cebollada, J. Feng, A. Kaplan, A. Ingles-Prieto, C.L. Badilla, et al., *Protein folding drives disulfide formation.* Cell, 2012. **151** (4): 794-806.

156. Benton, M.J., P.C. Donoghue, R.J. Asher, M. Friedman, T.J. Near, and J. Vinther, *Constraints on the timescale of animal evolutionary history.* Palaeontol Electron, 2015. **18** (1): 1-106.
157. Welker, F., M.J. Collins, J.A. Thomas, M. Wadsley, S. Brace, E. Cappellini, et al., *Ancient proteins resolve the evolutionary history of Darwin's South American ungulates.* Nature, 2015. **522** (7554): 81-84.
158. Abascal, F., R. Zardoya, and D. Posada, *ProtTest: selection of best-fit models of protein evolution.* Bioinformatics, 2005. **21** (9): 2104-2105.
159. Drummond, A.J., M.A. Suchard, D. Xie, and A. Rambaut, *Bayesian phylogenetics with BEAUti and the BEAST 1.7.* Mol Biol Evol, 2012. **29** (8): 1969-1973.
160. Tabuce, R., R.J. Asher, and T. Lehmann, *Afrotherian mammals: a review of current data.* Mammalia, 2008. **72** (1): 2-14.
161. Sigé, B., J.-J. Jaeger, J. Sudre, and M. Vianey-Liaud, *Altiatlasius koulchii n. gen. et sp., primate omomyidé du Paléocène supérieur du Maroc, et les origines des euprimates.* Palaeontographica Abteilung A, 1990: 31-56.
162. Seiffert, E.R., E.L. Simons, W.C. Clyde, J.B. Rossie, Y. Attia, T.M. Bown, et al., *Basal anthropoids from Egypt and the antiquity of Africa's higher primate radiation.* Science, 2005. **310** (5746): 300-304.
163. Neagoe, C., C.A. Opitz, I. Makarenko, and W.A. Linke, *Gigantic variety: expression patterns of titin isoforms in striated muscles and consequences for myofibrillar passive stiffness.* J Muscle Res Cell Motil, 2003. **24** (2-3): 175-189.
164. von Castelmur, E., M. Marino, D.I. Svergun, L. Kreplak, Z. Ucurum-Fotiadis, P.V. Konarev, et al., *A regular pattern of Ig super-motifs defines segmental flexibility as the elastic mechanism of the titin chain.* Proc Natl Acad Sci U S A, 2008. **105** (4): 1186-1191.
165. Wong, J.W., S.Y. Ho, and P.J. Hogg, *Disulfide bond acquisition through eukaryotic protein evolution.* Mol Biol Evol, 2011. **28** (1): 327-334.
166. Fiser, A. and I. Simon, *Predicting redox state of cysteines in proteins.* Methods Enzymol. 2002: **353**, 10-21
167. Fiser, A. and I. Simon, *Predicting the oxidation state of cysteines by multiple sequence alignment.* Bioinformatics, 2000. **16** (3): 251-256.
168. Perez-Jimenez, R., S. Garcia-Manyes, S.R. Ainavarapu, and J.M. Fernandez, *Mechanical unfolding pathways of the enhanced yellow fluorescent protein revealed by single molecule force spectroscopy.* J Biol Chem, 2006. **281** (52): 40010-40014.
169. Ainavarapu, S.R., J. Brujic, H.H. Huang, A.P. Wiita, H. Lu, L. Li, et al., *Contour length and refolding rate of a small protein controlled by engineered disulfide bonds.* Biophys J, 2007. **92** (1): 225-233.
170. Grutzner, A., S. Garcia-Manyes, S. Kotter, C.L. Badilla, J.M. Fernandez, and W.A. Linke, *Modulation of titin-based stiffness by disulfide bonding in the cardiac titin N2-B unique sequence.* Biophys J, 2009. **97** (3): 825-834.
171. Fernandez, J.M. and H. Li, *Force-clamp spectroscopy monitors the folding trajectory of a single protein.* Science, 2004. **303** (5664): 1674-1678.
172. Perez-Jimenez, R., J. Li, P. Kosuri, I. Sanchez-Romero, A.P. Wiita, D. Rodriguez-Larrea, et al., *Diversity of chemical mechanisms in thioredoxin catalysis revealed by single-molecule force spectroscopy.* Nat Struct Mol Biol, 2009. **16** (8): 890-896.
173. Wiita, A.P., R. Perez-Jimenez, K.A. Walther, F. Grater, B.J. Berne, A. Holmgren, et al., *Probing the chemistry of thioredoxin catalysis with force.* Nature, 2007. **450** (7166): 124-127.

174. Alegre-Cebollada, J., P. Kosuri, J.A. Rivas-Pardo, and J.M. Fernandez, *Direct observation of disulfide isomerization in a single protein.* Nat Chem, 2011. **3** (11): 882-887.
175. Mayans, O., J. Wuerges, S. Canela, M. Gautel, and M. Wilmanns, *Structural evidence for a possible role of reversible disulphide bridge formation in the elasticity of the muscle protein titin.* Structure, 2001. **9** (4): 331-340.
176. Li, H. and J.M. Fernandez, *Mechanical design of the first proximal Ig domain of human cardiac titin revealed by single molecule force spectroscopy.* J Mol Biol, 2003. **334** (1): 75-86.
177. Justel, A., D. Peña, and R. Zamar, *A multivariate Kolmogorov-Smirnov test of goodness of fit.* Stat Probab Lett 1997. **35** (3): 251-259.
178. Tidor, B. and M. Karplus, *The contribution of cross-links to protein stability: A normal mode analysis of the configurational entropy of the native state.* Proteins, 1993. **15** (1): 71-79.
179. Ishikawa, H., S. Kim, K. Kwak, K. Wakasugi, and M.D. Fayer, *Disulfide bond influence on protein structural dynamics probed with 2D-IR vibrational echo spectroscopy.* Proc Natl Acad Sci U S A, 2007. **104** (49): 19309-19314.
180. Zhou, B., I.B. Baldus, W. Li, S.A. Edwards, and F. Grater, *Identification of allosteric disulfides from prestress analysis.* Biophys J, 2014. **107** (3): 672-681.
181. Kim, D.E., D. Chivian, and D. Baker, *Protein structure prediction and analysis using the Robetta server.* Nucleic Acids Res, 2004. **32** (Web Server issue): W526-531.
182. Katz, B.A. and A. Kossiakoff, *The crystallographically determined structures of atypical strained disulfides engineered into subtilisin.* J Biol Chem, 1986. **261** (33): 15480-15485.
183. Anjukandi, P., P. Dopieralski, J. Ribas-Arino, and D. Marx, *The effect of tensile stress on the conformational free energy landscape of disulfide bonds.* PLoS One, 2014. **9** (10): e108812.
184. Betz, S.F., *Disulfide bonds and the stability of globular proteins.* Protein Sci, 1993. **2** (10): 1551-1558.
185. Legendre, L.J., G. Guenard, J. Botha-Brink, and J. Cubo, *Palaeohistological Evidence for Ancestral High Metabolic Rate in Archosaurs.* Syst Biol, 2016.
186. Luo, Z.X., Q.J. Meng, Q. Ji, D. Liu, Y.G. Zhang, and A.I. Neander, *Mammalian evolution. Evolutionary development in basal mammaliaforms as revealed by a docodontan.* Science, 2015. **347** (6223): 760-764.
187. Luo, Z.X., C.X. Yuan, Q.J. Meng, and Q. Ji, *A Jurassic eutherian mammal and divergence of marsupials and placentals.* Nature, 2011. **476** (7361): 442-445.
188. Martin, T., J. Marugan-Lobon, R. Vullo, H. Martin-Abad, Z.X. Luo, and A.D. Buscalioni, *A Cretaceous eutriconodont and integument evolution in early mammals.* Nature, 2015. **526** (7573): 380-384.
189. Carroll, R.L., *A Middle Pennsylvanian captorhinomorph, and the interrelationships of primitive reptiles.* J Paleo, 1969: 151-170.
190. Bi, S., Y. Wang, J. Guan, X. Sheng, and J. Meng, *Three new Jurassic euharamiyidan species reinforce early divergence of mammals.* Nature, 2014. **514** (7524): 579-584.
191. Sallan, L. and A.K. Galimberti, *Body-size reduction in vertebrates following the end-Devonian mass extinction.* Science, 2015. **350** (6262): 812-815.
192. Lindstedt, S.L. and P.J. Schaeffer, *Use of allometry in predicting anatomical and physiological parameters of mammals.* Lab Anim, 2002. **36** (1): 1-19.

193. Burness, G.P., S.C. Leary, P.W. Hochachka, and C.D. Moyes, *Allometric scaling of RNA, DNA, and enzyme levels: an intraspecific study.* Am J Physiol, 1999. **277** (4 Pt 2): R1164-1170.
194. Paton, R.L., T.R. Smithson, and J.A. Clack, *An amniote-like skeleton from the Early Carboniferous of Scotland.* Nature, 1999. **398** (6727): 508-513.
195. O'Leary, M.A., J.I. Bloch, J.J. Flynn, T.J. Gaudin, A. Giallombardo, N.P. Giannini, et al., *The Placental Mammal Ancestor and the Post–K-Pg Radiation of Placentals.* Science, 2013. **339** (6120): 662-667.
196. Granzier, H. and S. Labeit, *Cardiac titin: an adjustable multi-functional spring.* J Physiol, 2002. **541** (Pt 2): 335-342.
197. Seymour, R.S. and A.J. Blaylock, *The principle of laplace and scaling of ventricular wall stress and blood pressure in mammals and birds.* Physiol Biochem Zool, 2000. **73** (4): 389-405.

Acknowledgements

I would like to start this section thanking Raul Perez-Jimenez for giving me the opportunity to carry out this thesis under his supervision. During the last years I have learnt things that I could not even imagine four years ago from and with him. I also want to acknowledge my codirector, David de Sancho, for all the support, help and the constructive and beneficial criticism since he came to our group. Many thanks to Txema Pitarke for trusting me and letting me to develop my career in CIC nanoGUNE.

I also feel very lucky with the colleagues I had during my Thesis. Mila esker Nerea, nire eskuin eskua izan zara lau urte hauetan. Muchas gracias Álvaro, has sido un gran apoyo estos años. Merci Marie, pour la patience que vous aviez avec moi. Danke Jörg, ich werde deine Unterstützung nie vergessen. Gracias Borja por ayudarme siempre que has podido. Eskerrik asko ere Leireei, beti irribarre bat ateratzeagatik. Gracias también al resto de mi grupo por todo: Patricia, Bárbara, Anne, Susana, Simon… Thanks also to the rest of nanoGUNE for making me feel home these four years. I would like to think that I made some friends that will last forever. A special mention to Paulo, Iban, Jon, Unai and César, for putting up with me.

My acknowledges to Jorge Alegre-Cebollada and Elías Herrero-Galán for the biochemical experiments and the nice talks. I am very grateful with the people from RUB University in Bochum for giving me the opportunity to do a stage there. Many thanks to Professor Wolfgang Linke for accepting me on his group. I learnt a lot about titin and muscle physiology. I am also very happy that I met Nazha Hamdani, who helped me in everything and cared about me. I also realized during my stage that I have very good friends in Germany. Danke, Fritz, спасибо, Igor. Nire tesiaren azal dotorea Juleneri esker da. Eskerrik asko Erik, euskaraz daogen zatia zuzentzeagatik eta orrialde bat baino gehiago irakurtzeagatik. Eta, nola ez, mila esker nire familiari, eta batez ere nire gurasoei, beti hor egon direlako.

Miren eta Robertori, izango garelako.

List of publications

1. ***Mechanochemical Evolution of the Giant Muscle Protein Titin as Inferred from Resurrected Proteins***

   Aitor Manteca, Jörg Schönfelder, Álvaro Alonso-Caballero, Marie J Fertin, Nerea Barruetabeña, Bruna F Faria, Elias Herrero-Galán, Jorge Alegre-Cebollada, David De Sancho and Raul Perez-Jimenez

   Nature Structural and Molecular Biology. Article in Press. DOI: 10.1038/3426

2. ***The Influence of Disulfide Bonds on the Mechanical Stability of Proteins is Context Dependent***

   Aitor Manteca, Álvaro Alonso-Caballero, Marie J Fertin, Simon Poly, David de Sancho and Raul Perez-Jimenez

   Under revision in Journal of Biological Chemistry

Mechanochemical evolution of the giant muscle protein titin as inferred from resurrected proteins

Q1 Aitor Manteca[1], J Schönfelder[1], Alvaro Alonso-Caballero[1], Marie J Fertin[1], Nerea Barruetabeña[1], Bruna F Faria[2], Q2 Elias Herrero-Galán[3], Jorge Alegre-Cebollada[3], David De Sancho[1,4] & Raul Perez-Jimenez[1,4,5]

The sarcomere-based structure of muscles is conserved among vertebrates; however, vertebrate muscle physiology is extremely diverse. A molecular explanation for this diversity and its evolution has not been proposed. We use phylogenetic analyses and single-molecule force spectroscopy (smFS) to investigate the mechanochemical evolution of titin, a giant protein responsible for the elasticity of muscle filaments. We resurrect eight-domain fragments of titin corresponding to the common ancestors to mammals, sauropsids, and tetrapods, which lived 105–356 Myr, and compare them with titin fragments from some of their modern descendants. We demonstrate that the resurrected titin molecules are rich in disulfide bonds and display high mechanical stability. These mechanochemical elements have changed over time, creating a paleomechanical trend that seems to correlate with animal body size, allowing us to estimate the sizes of extinct species. We hypothesize that mechanical adjustments in titin contributed to physiological changes that allowed the muscular development and diversity of modern tetrapods.

Titin is a micrometer-long muscle protein composed of hundreds of individually folded domains[1]. Titin is present in all vertebrates, being one of the main components of the sarcomere together with actin and myosin. The main constituents of titin are immunoglobulin (Ig) domains, fibronectin type 3 domains and the unstructured PEVK region[2]. In the sarcomere, titin connects the Z disk to the M line (**Fig. 1a**). The main function of titin is providing passive elasticity to the muscle through acting like a spring. In addition, recent studies support an important role of the refolding of titin during contraction[3]. Although titin has been studied for decades, there is yet much to be explored regarding the correlation between muscle physiological diversity in animals and the biochemistry and nanomechanics of titin. Given the morphological and locomotor diversity in vertebrates, titin may hold the key for some phenotypes displayed by animals in terms of muscle physiology. In this respect, it can be hypothesized that the evolution of titin has been central to the acquisition of muscle diversity in animals. However, the role of titin in evolution, for instance, during the huge physiological outbreak after the Cambrian explosion 542 Myr[4], remains unexplored.

During the past two decades, our knowledge of the mechanical properties of titin has increased dramatically due to the use of smFS techniques, which make it possible to apply calibrated mechanical forces to titin domains[5,6]. In addition, increasing amounts of genetic data offer new avenues for comparative biology to better understand biological systems. In this regard, phylogenetic methods applied to genomic information have made it possible to establish evolutionary relationships among different living organisms, including the possibility to infer the putative sequences of the genes of their extinct ancestors[7,8]. Ancestral sequence reconstruction (ASR) allows us to track the evolutionary history of genes and proteins to obtain information about extinct species. This information relates to physiological and metabolic features[9,10] as well as to the environmental conditions that hosted ancestral organisms[11,12].

Here, we have combined smFS and ASR to investigate the evolution of the mechanical and biochemical properties of titin. We have used phylogeny to travel back in time and reconstruct a fragment of titin from various extinct species, including the last common ancestors of tetrapods, sauropsids and mammals. We expressed the ancient proteins in the laboratory and measured their mechanical properties using smFS. We found that differences in mechanical stability and disulfide bond occurrence between titin from living species and their ancestors appear as key elements that have driven the mechanical evolution of titin. These differences illustrate a paleomechanical trend, that is, a mechanical trend from the ancestral to modern animals, which allows us to establish physiological correlations purely based on the nanomechanical properties of titin. These correlations allow us to predict body sizes of ancestral species that are comparable to those found in fossil records. Our experiments shed light on some of the molecular features that have driven the diversification and speciation of titin and possibly muscle physiology in vertebrates.

RESULTS

Reconstructing ancestral titin molecules

We used 33 protein sequences of titin from modern vertebrate species, with most of them corresponding to the complete protein sequence composed of more than 30,000 residues. The amino acid

[1]CIC nanoGUNE, San Sebastian, Spain. [2]Laboratory of Molecular Modeling, Federal University of São João del-Rei, São João del-Rei, Brazil. [3]Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Madrid, Spain. [4]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain. [5]Evolution and Genomics Technologies, S. L. (Evolgene), San Sebastian, Spain. Correspondence should be addressed to R.P.-J. (r.perezjimenez@nanogune.eu).

Received 21 February; accepted 2 June; published online XX XX 2017; doi:10.1038/nsmb.3426

sequences were retrieved from Uniprot and GenomeNet databases containing titin from mammals, sauropsids, amphibians and ray-finned fishes (Online Methods). Using these sequences, we generated a sequence alignment and constructed a phylogenetic chronogram using Bayesian inference as well as maximum parsimony (**Fig. 1b** and **Supplementary Fig. 1a**,**b**).

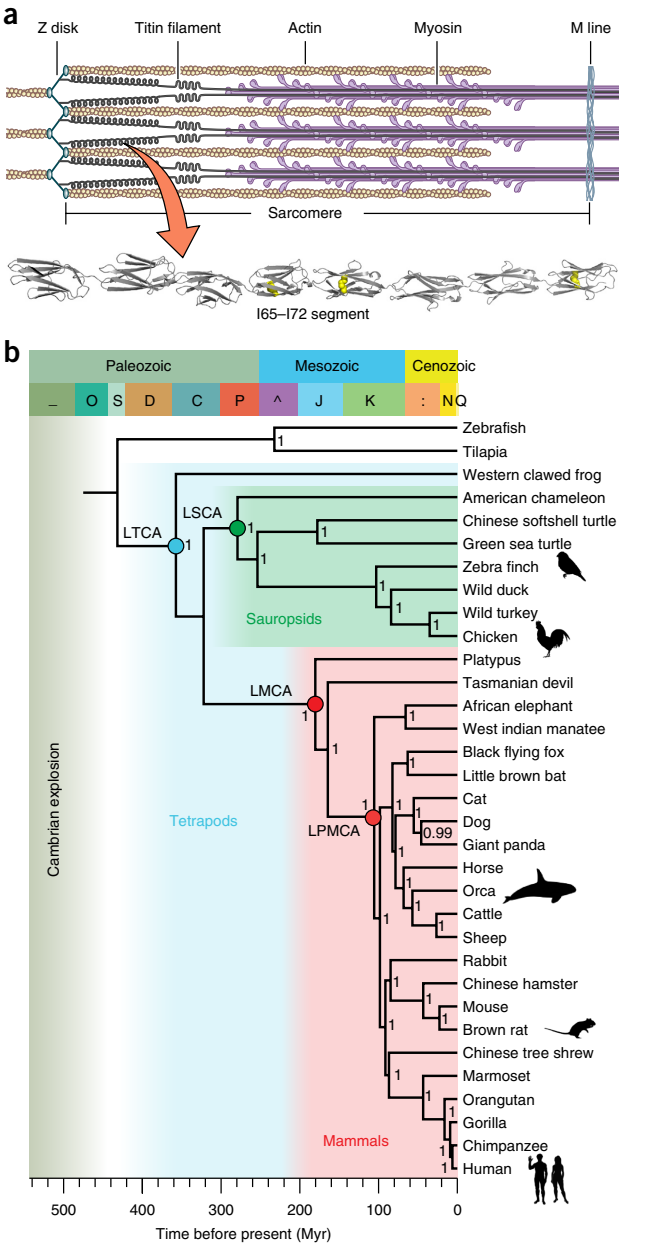


Figure 1 Reconstruction of ancestral titin fragments. (**a**) Scheme of one half of the sarcomere from Z disk to M line. The three main sarcomeric proteins, actin, myosin and titin, are shown. The segment encompassing domains I65–I72 from the elastic part of titin has been selected for ancestral reconstruction and testing. (**b**) Uncorrelated log-normal relaxed-clock chronogram of titin with geological time inferred with Bayesian inference. A total of 33 titin genes were used. The modern species studied are indicated by the animal outlines: zebra finch, chicken, orca, rat and human. The internal nodes (circles) were selected for resurrection and laboratory testing and represent the last common ancestors of tetrapods (LTCA, 356 ± 11 Myr), sauropsids (LSCA, 278 ± 14 Myr), mammals (LMCA, 179 ± 38 Myr) and placental mammals (LPMCA, 105 ± 17 Myr). Posterior probabilities for branch support are shown in the nodes. Geological periods are shown in the upper bar. Outlines were retrieved from http://www.phylopic.org.

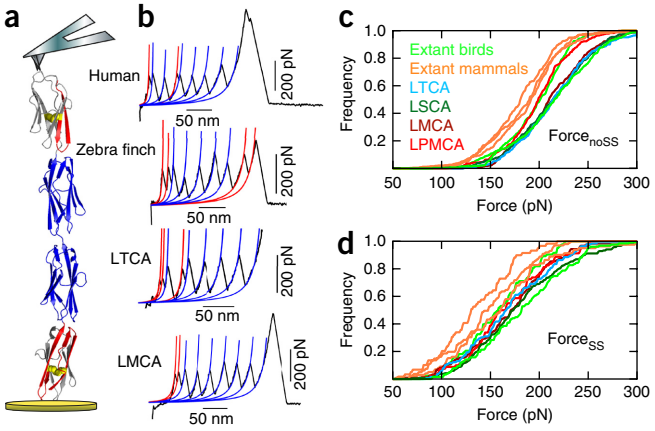


Figure 2 Single-molecule experiments with titin. (**a**) Schematic representation of a single-molecule experiment using the smFS (not to scale). Although the segment I65–I72 contains eight domains, for clarity only four are represented. Disulfide-bonded domains, red and gray; cysteine, yellow highlight; non-disulfide-bonded domains, blue. The protein is mechanically stretched between a cantilever tip and a gold-coated surface. (**b**) Representative experimental traces of the polyprotein I65–I72. The unfolding of domains is monitored as a sawtooth pattern of force versus extension peaks. The worm-like chain model was used to fit the data. Fully extended domains (blue lines) show extension of about 30 nm, whereas disulfide-bonded domains (red lines) show shorter contour lengths of 5–20 nm. (**c**,**d**) Cumulative histogram of mechanical unfolding force for domains that do not contain disulfide bonds, noSS (**c**) and domains that are disulfide bonded, SS (**d**).

From this tree, we sampled several internal nodes for reconstruction of the most probable ancestral sequence. In particular, we chose the tree nodes corresponding to the last tetrapod common ancestor (LTCA), which lived in the early Carboniferous period around 356 Myr ago; the last sauropsid common ancestor (LSCA), which is thought to have lived in the Permian period ~278 Myr ago; the last mammal common ancestor (LMCA), which lived during the Jurassic period ~179 Myr ago; and finally, the last placental mammal common ancestor (LPMCA), from the mid-Cretaceous period ~105 Myr ago (**Fig. 1b**). We used maximum likelihood to infer the most probable sequences of the ancestral nodes, following well-described procedures[9,11,12] (Online Methods). Posterior probability distributions across all sites for ancestral sequences are reported in **Supplementary Figure 2**. The overall posterior probability of the sites lies between 0.90 and 0.99. Of note, the determination of ancestral sequences is not free from uncertainty. In experimental ASR, the goal is to determine the phenotypes and characteristics of proteins rather than the exact and accurate sequences. There is not a single exactly reconstructed sequence that represents the true ancestor of a particular set of organisms, but the phenotype and characteristics associated with those sequences must be unique. Interestingly, ASR is able to capture such phenotypes. Over the years, numerous studies have been carried out in order to assess the robustness of ASR[11,13–15].

We dated the nodes using multiple sources from the Time Tree of Life[16] as well as paleontological data[17]. We selected for resurrection and laboratory testing of an eight-domain fragment of titin encompassing domains I65–I72 in the canonical human titin sequence (Uniprot Q8WZ42) and the homologous fragment in other species. This fragment is a good proxy of the elastic I-band region, located in the proximal

tandem-Ig region of N2A skeletal titin and N2BA cardiac titin isoforms[18]. Part of this fragment has been characterized in terms of structure and mechanics[19,20]. Additionally, the alignment of this fragment is well resolved, suggesting that it is structurally conserved.

Sequence conservation and the role of cysteines in sarcomeric proteins

The comparison of sequences of the inferred ancestral titin I65–I72 with those of the modern counterparts yielded amino acid identities ranging from ~76% to 90% (**Supplementary Table 1**). We also compared titin sequence diversity levels with those of myosin II and actin, the other two most abundant proteins in the sarcomere (**Supplementary Table 1**). Actin is an extremely conserved protein that shows almost 99% identity across most living vertebrates, suggesting that it had very little influence on the molecular diversification of the sarcomere. In the case of myosin, identities of modern forms reach over 90%. We reconstructed a phylogenetic tree of myosin II from a similar pool of vertebrates as in the case of titin (**Supplementary Fig. 3**) and inferred ancient myosin sequences. Using ancestral and modern sequences, we observed that the titin I65–I72 fragment has approximately twice the mutation rate of myosin II (**Supplementary Fig. 4a**,**b**). This suggests that titin has contributed to the molecular diversification of the sarcomere more extensively than myosin.

We determined the pattern of amino acid replacement in I65–I72 in the transition from LTCA to human. Amino acid replacements weighted by their relative mutability[21] show that cysteine residues display mutability different from what is expected during the evolution of titin I65–I72 from LTCA to human. Human titin contains considerably fewer cysteine residues than LTCA titin (**Supplementary Fig. 4c**–**f**). Assignment of cysteine residues in the ancestral sequences are supported by values of posterior probability close to 1. This is noteworthy, given that cysteine is one of the least mutable residues in proteins, second only to tryptophan[21]; cysteine residues are rarely lost once acquired[22]. We carried out a similar analysis for a titin fragment from the distal region of the elastic I band (I88–I95) and observed similar behavior for cysteine (**Supplementary Fig. 4c**–**f**). In contrast, the same analysis in a fragment from the rigid A-band titin (I126–Fn90–94–I128) shows the opposite behavior in regard to the mutability of cysteine, similar to that found in myosin. Altogether, these observations suggest that cysteine residues played a crucial role in the molecular evolution of the elastic I band of titin. Interestingly, cysteines are involved in the formation of disulfide bonds, efficiently modulating the mechanical properties of the parent protein[23,24].

The existence of disulfides in titin was first proposed following the identification of a disulfide bond in the crystal structure of domain I1 and the observation that many domains in titin contain proximal cysteines that can engage in disulfide bonds[25]. The crystal structures of different titin domains (PDB 3B34, 2RIK, 2RJM, 1G1C) have shown the existence of such disulfides. Indeed, studies of disulfide bonds in titin have been limited, owing to deficient disulfide formation in recombinant expression systems, sometimes leading to contradictory results[26]. Nevertheless, experimental evidence of disulfide formation in native titin is still lacking.

Given the fundamental role of disulfide bonds in the mechanics of proteins, we hypothesize that disulfides may be related to the mechanochemical evolution of titin. The genes encoding the four ancestral I65–I72 fragments were synthesized, and the gene products were expressed in *Escherichia coli* and purified under equal oxidative stress conditions to overcome limited disulfide bond formation in the host. We also expressed and purified I65–I72 from representative modern

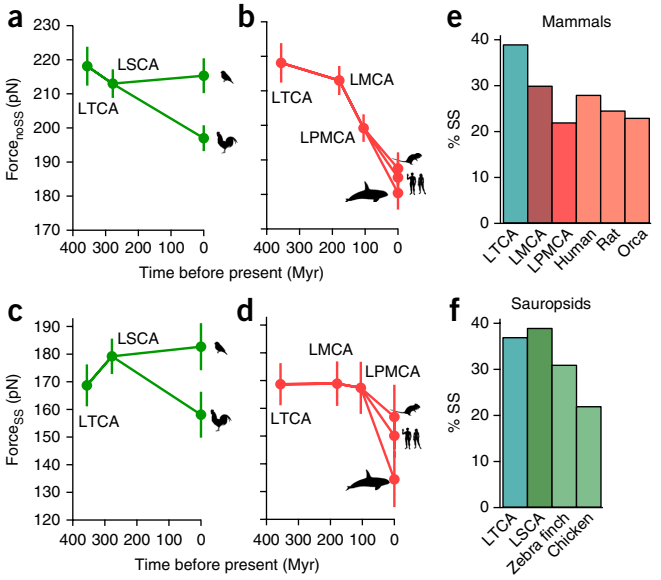


Figure 3 Mechanical stability versus geological time. (**a**,**b**) Unfolding forces of non-disulfide-bonded (noSS) domains for birds and ancestors (**a**) and for mammals and ancestors (**b**). (**c**,**d**) Unfolding forces of disulfide-bonded (SS) domains for birds and ancestors (**c**) and for mammals and ancestors (**d**). Error bars in **a**–**d** indicate 95% confidence intervals for the sample mean. (**e**,**f**) Percentage of disulfide-bonded domains detected in force–extension traces for mammals (**e**) and sauropsids (**f**). LTCA and LSCA titin fragments have the higher percentage of disulfide bonds in both groups. In general, modern animals display fewer disulfide bonds than their ancestors. Data collected from $n = 374$ for LTCA, $n = 614$ for LSCA, $n = 407$ for LMCA, $n = 366$ for LPMCA, $n = 409$ for zebra finch, $n = 375$ for chicken, $n = 263$ for orca, $n = 341$ for rat and $n = 347$ for human titin fragment.

amniote species covering different clades in our tree: human, brown rat, orca, chicken and zebra finch.

smFS reveals different paleomechanical trends for birds and mammals

To investigate the mechanical properties of the titin variants, we performed smFS by mechanically stretching the proteins at a constant speed of 400 nm/s (**Fig. 2a**, Online Methods). The stretching of titin domains leads to saw-tooth patterns in force–extension recordings, in which each peak represents the mechanical unfolding of an individual domain (**Fig. 2b**). Analysis of the traces using well-established procedures[27] allowed us to determine the mechanical stabilities and contour lengths of the constituent domains.

For all of the variants tested, we observed two distinct populations of peaks. The first one had contour lengths of ~30 nm, and the second showed lengths of 5–20 nm (**Supplementary Fig. 5**). We believe that the first population corresponds to fully extended domains of about 90 residues, whereas the second population represents disulfide-containing domains that make the contour length of the extended peptide shorter, as expected, considering the different positions of the cysteines in the different domains[28]. This is notable because disulfide bonds in titin have been suggested to participate in the mechanical regulation of the protein[24]. We constructed cumulative histograms of mechanical stability for both populations. In the case of fully extended domains, we estimated an average unfolding force ranging from 180 to 218 pN (**Fig. 2c**), depending on the variant, with the lowest stability for orca titin and the highest for LTCA titin domains. Results for this first population of peaks are in agreement with previous mechanical

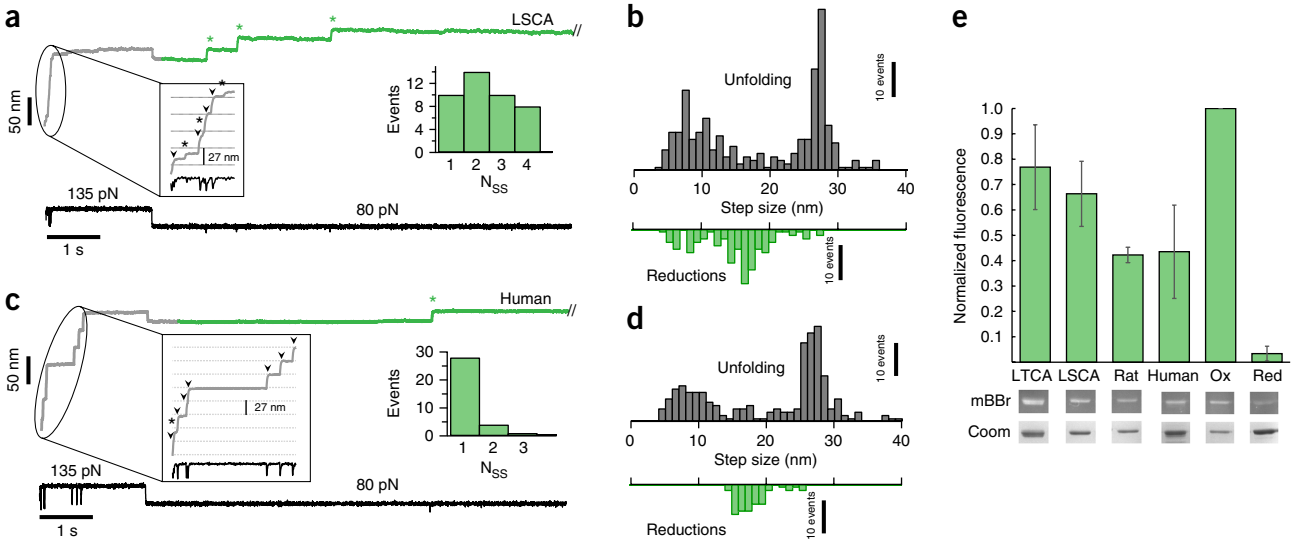


Figure 4 Force-clamp experiments for detection of disulfide-bond reductions catalyzed by Trx. (**a**) Experimental force-clamp trace of LSCA titin. We first apply a pulse of force of 135 pN during 2 s that triggers unfolding of non-disulfide-bonded domains (inset arrows) and disulfide-bonded domains up to their disulfide bond (asterisks in inset). The disulfide bonds can be reduced by Trx enzymes present at 10 μM concentration. The reduction events are monitored at a force of 80 pN (green line). A histogram of the number of reduction events per trace is shown. (**b**) Histograms of step size for unfolding events (gray) and disulfide bond reductions by Trx (green) in LSCA titin ($n = 372$). (**c**) Experimental force-clamp trace for human titin. We rarely observe more than one reduction. (**d**) Step size histograms for unfolding events (gray) and reduction events (green) ($n = 267$). (**e**) In-gel determination of oxidized thiols for LTCA, LSCA, rat and human I65–I72. Fluorescent bands resulting from the labeling of oxidized thiols with mBBr were normalized by the total quantity of protein as assessed by Coomassie staining and densitometry. The oxidized (Ox) and reduced (Red) versions of $(I91–32/75)_8$ were used as controls. Mean values ± s.d. of $n = 3$ independent experiments are represented (for human titin, $n = 4$).

characterization of I65–I70 purified in reducing conditions, in which no disulfide bonds can be established[20]. For the collection of domains that contain disulfide bonds, we determined average stability ranging from 134 to 182 pN, depending on the variant (**Fig. 2d**), with orca titin having the lowest stability and zebra finch the highest.

We plotted stabilities for both types of domains versus the age of each titin fragment for both bird (**Fig. 3a**,**b**) and mammalian lineages (**Fig. 3c**,**d**). In the case of birds, the stability changed slightly for zebra finch and chicken with respect to LTCA and LSCA but in opposite directions. In the case of mammals, we see a clear paleomechanical trend by which stability has decreased over time. Interestingly, in both birds and mammals, we observed that animals with larger body sizes have lower mechanical stabilities of titin domains with and without disulfide bridges. By counting the number of domains that contain disulfide bonds and comparing it with the number of those that fully extend, we observed that, in general, ancestral proteins of LTCA and LSCA have the highest proportions of disulfide-bonded domains (**Fig. 3e**,**f**).

Force-clamp analysis confirms the evolutionary decay of disulfide bonds

The occurrence of disulfide bonds seems to have decreased in modern species. In fact, the content of disulfide bonds seems to be related to the mechanical stability. More stable titin forms imply more disulfide-bonded domains (**Supplementary Fig. 5**). To better identify disulfide bonds, we performed single-molecule experiments in the force-clamp modality in order to capture reduction of disulfide bonds. In this mode, the force applied to the protein can be controlled. The mechanical unfolding is monitored as an increment in length versus time. We have used force-clamp techniques in previous work to demonstrate the kinetics of disulfide bond reduction under force by thioredoxin enzymes (Trx)[29,30]. Trx is an enzyme that controls the redox balance in cells by reducing disulfide bonds. The reduction of disulfide bonds by Trx is a force-dependent reaction that can be readily monitored and quantified[30]. By applying force to a disulfide-bonded titin domain, we can trigger the unfolding of the domain up to the disulfide bond, which becomes exposed to the solution. If Trx is present in the solution, the disulfide bond can be reduced, and the sequestered residues behind the disulfide bond are released, giving rise to an extra extension of the polypeptide chain (**Supplementary Fig. 6a**). With this assay, we can quantify disulfide bonds that are cryptic, requiring exposure in order to be reduced. Non-cryptic disulfide bonds are reduced without the need of mechanical exposure.

We applied this test to LSCA and human titin fragments in the presence of Trx. We first applied a pulse of force of 135 pN during 0.5 s that triggers unfolding of all the domains (**Fig. 4a**,**c**). The unfolding of the domains is monitored as a staircase of ~27 nm per step for reduced domains and shorter steps of between 5 and 20 nm for disulfide-bonded domains (**Fig. 4b**,**d**). The expected length of extended disulfide-containing domains varies due to the different positions of cysteines in the sequences. After the unfolding force pulse, we quenched the force to 80 pN for 20 s to monitor disulfide bond reductions as steps within the range of 5–20 nm for LSCA titin and 15–20 nm for human titin (**Fig. 4a**,**c**). Again, the length attributed to reduction events is different for each domain. Additionally, some domains have more than two cysteines, which may imply the possibility of isomerizations[31]. All possible disulfide bond combinations have been estimated and are shown in **Supplementary Table 2**. A histogram of the observed reduction events demonstrates that in LSCA, it is common to observe up to four reduction events (**Fig. 4a**), whereas in human titin, it is common to observe only one reduction event (**Fig. 4c**). In the absence of Trx, no reduction events were observed (**Supplementary Fig. 6b–e**). Thus, force-clamp experiments confirm that LSCA titin contains more disulfide bonds than does

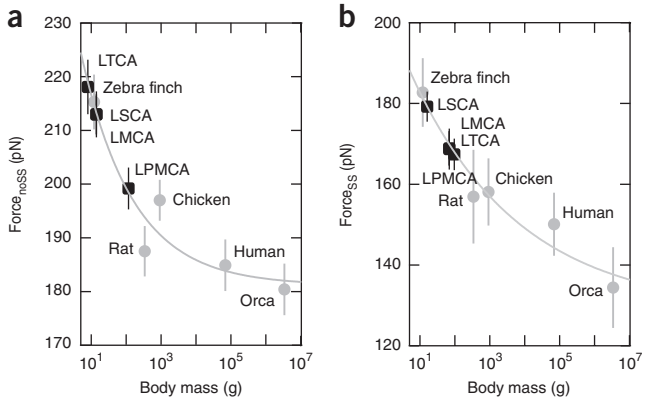


Figure 5 Correlation of mechanical stability with animal body mass for (**a**) non-disulfide-bonded domains (noSS) and (**b**) disulfide bonded domains (SS). Stability versus body mass follows a power-law correlation: for disulfide-bonded domains, $F - F_0 = 73 \times M^{-0.17}$ and for domains without disulfides, $F - F_0 = 65 \times M^{-0.26}$. Modern species are represented as gray circles. The values of body mass for ancient species LTCA, LSCA, LMCA and LPMCA can be interpolated from the different fittings and are represented in black squares. Error bars indicate 95% confidence intervals for the sample mean.

human titin, supporting conclusions obtained using force extension for all I65–I72 fragments.

To provide independent measurements of cysteine oxidation, we used a biochemical assay that detects oxidized cysteines[32] by labeling them with the fluorophore monobromobimane (mBBr). In this assay, reduced thiols in the protein are alkylated with an excess of *N*-ethylmaleimide in denaturing conditions. After polyacrylamide electrophoresis, the gel is used as a reaction chamber to reduce oxidized thiols through incubation with DTT and for subsequent reaction of the newly reduced thiols with mBBr. The resulting fluorescence signal is proportional to the amount of oxidized thiols in the sample. Results show that LTCA and LSCA proteins contain more oxidized cysteines than rat and human proteins (**Fig. 4e**). In our biochemical assays, we included a control $(I91–32/75)_8$ protein (formerly I27) that can be produced in reduced (no disulfides) or oxidized (one disulfide per domain for a total of eight disulfides) forms and whose oxidation status can be determined unambiguously by smFS[23]. Because the sizes of the control protein and the I65–I72 fragments are comparable, we used the normalized fluorescence signals to estimate the number of disulfides in LTCA (six), LSCA (five), human and rat (three or four). The results from this independent method of determining the population of disulfide bonds confirms the trend observed in smFS experiments of more disulfides in the ancestral proteins (**Fig. 4e**). There are, however, small discrepancies between the exact number of disulfides from the different methods that probably reflect contributions of terminal cysteines needed for attachment in smFS experiments or other forms of oxidation different from disulfides, such as sulfenylation induced by treatment with H_2O_2.

DISCUSSION

Our data suggest that small animals have titin domains with greater mechanical resistance than those in large animals. Indeed, this is consistent with the fact that in small mammal hearts, the titin isoform N2A, which is stiffer than N2BA, is more abundant[18]. We wondered whether mechanical stability is related to animal size. This correlation may be related to titin mechanics, as small animals have faster muscle contractions and shivering frequencies[33].

To probe this idea, we plotted the mechanical stability of domains, with and without disulfide bonds, versus body mass. We observed that the dependence of the average unfolding force on body mass can be fitted to a power law, suggesting an evolutionary allometric relationship (**Fig. 5a**,**b**). Body mass has been shown to follow allometric scaling with other physiological traits such as metabolic rate, speed, arterial pressure or heat production[34]. In fact, it is common to observe allometric scaling in biological systems, and even enzyme activity has been suggested to show allometry[35]. However, to the best of our knowledge, this is the first observation showing allometric scaling between a physiological feature and molecular-level parameters.

From these correlations, body mass of extinct species could be predicted. Inasmuch as we have two allometric scaling plots (**Fig. 5a**,**b**), we determined a range of body mass for each extinct species that is determined by the minimum and maximum values of mass obtained. This is 8–70 g for LTCA, 14–16 g for LSCA, 14–70 g for LMCA and 95–116 g for LPMCA. These weights are typical of small animals, perhaps between 10 and 40 cm in length. These estimates compare surprisingly well with sizes from fossils that could be related to these extinct animals[36–41]. Thus, the observed correlation between mechanical stability and body mass allows us to predict the sizes of extinct species. However, it must be noted that these correlations seem to be more sensitive for the high mechanical stability of titin molecules.

Our results show that the evolution of muscle physiology seems to be linked to the molecular evolution of titin in tetrapods. The reconstruction of ancient forms of titin demonstrates that titin domains from the small zebra finch are similar to those of its ancestors in terms of mechanical properties. However, titin domains from modern mammals have experienced more drastic changes, leading to proteins that have lower mechanical stability and fewer disulfide bonds compared to those in their oldest ancestor. This is remarkable given that mutation rates for titin in zebra finch and living mammals are quite similar. Most likely, these changes are related to morphological and physiological consequences that derived in the vast diversity of physical and locomotor features found in mammals.

This observation raises the question of whether the relation of titin mechanics with muscle contraction differs across species with different physiological features. A possible interpretation is that small amniotes with fast muscle contractions rely upon the mechanical response of titin domains. Under physiological forces, titin domains have been shown to unfold and refold during muscle contraction[3]. The presence of disulfide bonds prevents overstretching of titin and increases the recoiling speed of the domains[28], probably increasing the speed of muscle contraction as well. Thus, we hypothesize that the balance between mechanical stability and disulfide bonds may be a key factor in titin mechanical regulation and its evolution. Another possible interpretation relates to the fact that chemical modifications such as S-glutathionylation and oxidative stress have been suggested to occur in titin, and both involve cysteine residues[6], but even these phenomena are related to titin mechanics. Although other mechanisms might be related to the mechanochemical differences observed, it is hard to speculate with an interpretation other than the purely mechanical, given the clear elastic character of titin.

The size that we obtained for ancient species from the allometric correlations may be explained under the light of fossil remains. It has been shown that after mass extinction occurred in the late Devonian period, the so-called Hangenberg event (359 Ma), the majority of taxa found in fossil records were under 40 cm[41]. This is consistent with a global shrinkage process that occurred during the early Carboniferous period that lasted around 40 Myr. Our predicted data for LTCA, which supposedly lived after the Hangenberg event, are within the range of

the sizes reported. Fossils of early amniotes, including sauropsids and mammalian-like reptiles from the Carboniferous and Permian periods, are also within that range[39,42]. In the case of mammals, numerous findings have shown that early mammals, including placentals, were small rodent-like animals[36–38,40,43], which is in line with the sizes that we predict from nanomechanical information of titin. Nevertheless, establishing direct comparison between our estimations and any known fossils is difficult, because it is unlikely that a single fossil could be unambiguously labeled as LTCA, LSCA, LMCA or LPMCA, representing the true common ancestor of different taxonomic groups.

The eight-domain fragments of titin that we have studied are important because they belong to the critical elastic region of titin. Studying additional segments of titin as well as comparing features such as folding kinetics of ancient and modern domains will be interesting in order to gain a complete understanding of the molecular elements that have driven the molecular evolution of titin and its connection to muscle physiology.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Q3 ACKNOWLEDGMENTS

Research has been supported by the Ministry of Economy and Competitiveness (MINECO) grant BIO2016-77390-R, BFU2015-71964 to R. P.-J., BIO2014-54768-P and RYC-2014-16604 to J.A-C., and CTQ2015-65320-R to D.D.S., and the European Commission grant CIG Marie Curie Reintegration program FP7-PEOPLE-2014 to R.P.-J. A.A.-C. is funded by the predoctoral program of the Basque Government. R.P.-J. and D.D.S., thank CIC nanoGUNE and the Ikerbasque Foundation for Science for financial support. CNIC is supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and the Pro-CNIC Foundation and is a Severo Ochoa Center of Excellence (MINECO award SEV-2015-0505). Plasmid pQE80-(I91-32/75)$_8$ was a kind gift from J. Fernández (Columbia University). We thank R. Zardoya (National Museum of Natural Sciences, Madrid) for helpful discussions and comments. The authors acknowledge technical support provided by IZO-SGI SGIker of UPV/EHU and European funding (ERDF and ESF) for the use of the Arina HPC cluster and the assistance provided by T. Mercero and E. Ogando.

AUTHOR CONTRIBUTIONS

R.P.-J. designed the research. A.M., B.F.F., N.B., D.D.S. and R.P.-J. conducted phylogenetic analysis. A.M. and M.J.F. cloned and expressed proteins. A.M., A.A.-C., J.S., D.D.S. and R.P.-J. performed AFM experiments and data analysis. E.H.-G. and J. A.-C. performed biochemical determination of disulfides in titin fragments. R.P-J. drafted the paper and all authors contributed in revising and editing the manuscript.

Q4 COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interest.

1. Fürst, D.O., Osborn, M., Nave, R. & Weber, K. The organization of titin filaments in the half-sarcomere revealed by monoclonal antibodies in immunoelectron microscopy: a map of ten nonrepetitive epitopes starting at the Z line extends close to the M line. *J. Cell Biol.* **106**, 1563–1572 (1988).
2. Labeit, S. & Kolmerer, B. Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science* **270**, 293–296 (1995).
3. Rivas-Pardo, J.A. *et al.*Work done by titin protein folding assists muscle contraction. *Cell Rep.* **14**, 1339–1347 (2016).
4. Erwin, D.H. *et al.* The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science* **334**, 1091–1097 (2011).
5. Rief, M., Gautel, M., Oesterhelt, F., Fernandez, J.M. & Gaub, H.E. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science* **276**, 1109–1112 (1997).
6. Alegre-Cebollada, J. *et al.* S-glutathionylation of cryptic cysteines enhances titin elasticity by blocking protein folding. *Cell* **156**, 1235–1246 (2014).
7. Hall, B.G. Simple and accurate estimation of ancestral protein sequences. *Proc. Natl. Acad. Sci. USA* **103**, 5431–5436 (2006).
8. Merkl, R. & Sterner, R. Ancestral protein reconstruction: techniques and applications. *Biol. Chem.* **397**, 1–21 (2016).
9. Kratzer, J.T. *et al.* Evolutionary history and metabolic insights of ancient mammalian uricases. *Proc. Natl. Acad. Sci. USA* **111**, 3763–3768 (2014).
10. Zakas, P.M. *et al.* Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nat. Biotechnol.* **35**, 35–37(2017).
11. Gaucher, E.A., Govindarajan, S. & Ganesh, O.K.Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451**, 704–707 (2008).
12. Perez-Jimenez, R. *et al.* Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat. Struct. Mol. Biol.* **18**, 592–596 (2011).
13. Eick, G.N., Bridgham, J.T., Anderson, D.P., Harms, M.J. & Thornton, J.W. Robustness of reconstructed ancestral protein functions to statistical uncertainty. *Mol. Biol. Evol.* **34**, 247–261 (2017).
14. Hanson-Smith, V., Kolaczkowski, B. & Thornton, J.W. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol. Biol. Evol.* **27**, 1988–1999 (2010).
15. Randall, R.N., Radford, C.E., Roof, K.A., Natarajan, D.K. & Gaucher, E.A. An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat. Commun.* **7**, 12847 (2016).
16. Hedges, S.B., Marin, J., Suleski, M., Paymer, M. & Kumar, S. Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.* **32**, 835–845 (2015).
17. Benton, M.J. *et al.* Constraints on the timescale of animal evolutionary history. *Palaeontologica Electronica* **18.1**, 18.1.1FC (2015).
18. Neagoe, C., Opitz, C.A., Makarenko, I. & Linke, W.A. Gigantic variety: expression patterns of titin isoforms in striated muscles and consequences for myofibrillar passive stiffness. *J. Muscle Res. Cell Motil.* **24**, 175–189 (2003).
19. von Castelmur, E. *et al.* A regular pattern of Ig super-motifs defines segmental flexibility as the elastic mechanism of the titin chain. *Proc. Natl. Acad. Sci. USA* **105**, 1186–1191 (2008).
20. Watanabe, K., Muhle-Goll, C., Kellermayer, M.S., Labeit, S. & Granzier, H. Different molecular mechanics displayed by titin's constitutively and differentially expressed tandem Ig segments. *J. Struct. Biol.* **137**, 248–258 (2002).
21. Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. in *Atlas of Protein Sequence and Structure* Vol. 5, Suppl. 3 (National Biomedical Research Foundation, 1978).
22. Wong, J.W., Ho, S.Y. & Hogg, P.J. Disulfide bond acquisition through eukaryotic protein evolution. *Mol. Biol. Evol.***28**, 327–334 (2011).
23. Kosuri, P. *et al.* Protein folding drives disulfide formation. *Cell* **151**, 794–806 (2012).
24. Grützner, A. *et al.* Modulation of titin-based stiffness by disulfide bonding in the cardiac titin N2-B unique sequence. *Biophys. J.* **97**, 825–834 (2009).
25. Mayans, O., Wuerges, J., Canela, S., Gautel, M. & Wilmanns, M. Structural evidence for a possible role of reversible disulphide bridge formation in the elasticity of the muscle protein titin. *Structure* **9**, 331–340 (2001).
26. Li, H. & Fernandez, J.M. Mechanical design of the first proximal Ig domain of human cardiac titin revealed by single molecule force spectroscopy. *J. Mol. Biol.* **334**, 75–86 (2003).
27. Perez-Jimenez, R., Garcia-Manyes, S., Ainavarapu, S.R. & Fernandez, J.M. Mechanical unfolding pathways of the enhanced yellow fluorescent protein revealed by single molecule force spectroscopy. *J. Biol. Chem.* **281**, 40010–40014 (2006).
28. Ainavarapu, S.R. *et al.* Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. *Biophys. J.* **92**, 225–233 (2007).
29. Perez-Jimenez, R. *et al.* Diversity of chemical mechanisms in thioredoxin catalysis revealed by single-molecule force spectroscopy. *Nat. Struct. Mol. Biol.* **16**, 890–896 (2009).
30. Wiita, A.P. *et al.* Probing the chemistry of thioredoxin catalysis with force. *Nature* **450**, 124–127 (2007).
31. Alegre-Cebollada, J., Kosuri, P., Rivas-Pardo, J.A. & Fernández, J.M.Direct observation of disulfide isomerization in a single protein. *Nat. Chem.* **3**, 882–887 (2011).
32. Rogers, L.K., Leinweber, B.L. & Smith, C.V. Detection of reversible protein thiol modifications in tissues. *Anal. Biochem.* **358**, 171–184 (2006).
33. Legendre, L.J., Guénard, G., Botha-Brink, J. & Cubo, J. Palaeohistological evidence for ancestral high metabolic rate in Archosaurs. *Syst. Biol.* **65**, 989–996 (2016).
34. Lindstedt, S.L. & Schaeffer, P.J. Use of allometry in predicting anatomical and physiological parameters of mammals. *Lab. Anim.* **36**, 1–19 (2002).
35. Burness, G.P., Leary, S.C., Hochachka, P.W. & Moyes, C.D. Allometric scaling of RNA, DNA, and enzyme levels: an intraspecific study. *Am. J. Physiol.* **277**, R1164–R1170 (1999).
36. Luo, Z.X. *et al.*Mammalian evolution. Evolutionary development in basal mammaliaforms as revealed by a docodontan. *Science* **347**, 760–764 (2015).
37. Luo, Z.X., Yuan, C.X., Meng, Q.J. & Ji, Q.A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature* **476**, 442–445 (2011).
38. Martin, T. *et al.* A Cretaceous eutriconodont and integument evolution in early mammals. *Nature* **526**, 380–384 (2015).
39. Carroll, R.L. A Middle Pennsylvanian captorhinomorph, and the interrelationships of primitive reptiles. *J. Paleo.* **43**, 151–170 (1969).
40. Bi, S., Wang, Y., Guan, J., Sheng, X. & Meng, J. Three new Jurassic euharamiyidan species reinforce early divergence of mammals. *Nature* **514**, 579–584 (2014).
41. Sallan, L. & Galimberti, A.K. Body-size reduction in vertebrates following the end-Devonian mass extinction. *Science* **350**, 812–815 (2015).
42. Paton, R.L., Smithson, T.R. & Clack, J.A. An amniote-like skeleton from the Early Carboniferous of Scotland. *Nature* **398**, 508–513 (1999).
43. O'Leary, M.A. *et al.* The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* **339**, 662–667 (2013).

ONLINE METHODS

Phylogenetic analysis and ancestral sequence reconstruction. A set of 33 titin sequences were used from which 28 correspond to the full sequence of titin with over 30,000 residues. The sequences represent five different classes of vertebrate animals, mammals, amphibians, reptiles, birds and bony fishes, and were retrieved from the UniProt and GenomeNet databases. All sequence ID numbers are listed in **Supplementary Note**. The sequences were aligned using the MUSCLE software and further edited manually. We tested the alignment for best model of protein evolution using ProTest[44], resulting with the Jones-Taylor-Thornton (JTT) with gamma distribution model as the best evolution model. A second set of sequences was used containing new sequences that were added to databases at a later stage during the course of this study. We decided to include these new sequences to test the robustness of our tree, as most additions were in the Sauropsida and Amphibia clades that were less represented in our initial set. Two phylogenies were performed, the first using Bayesian inference using Markov Chain Monte Carlo, and the second, by the maximum parsimony criterion, with identical results in tree topology in all cases. For Bayesian inference, we used BEAST v1.8.2 package software[45] incorporating the BEAGLE library for parallel processing. We set monophyletic groups for primates, rodentia, carnivora, chiroptera, cetartiodactyla, archosaurs, testudines, squamata and fishes, being the latest the selected outgroup. We set the JTT model with 8 categories gamma distribution, Yule model for speciation and length chain of 25 million generations, sampling every 1000 generations. Calculations were run for 12 days in a single node of an HPC cluster of Intel Xeon 2680v2 processors, using 16 cores at 2.6 GHz and 64 GB of memory. A 12-core iMac computer was used for smaller trees such as myosin. We discarded the initial 30% of trees as burn-in. All nodes were supported by posterior probabilities above 0.63 with most of them nearly 1. The myosin tree was performed using only BEAST. Tree Annotator was used to estimate a maximum clade credibility tree removing 30% of initial trees as burn-in. FigTree v1.4.2 was used for tree representation and editing. For parsimony we used PAUP* 4.0 software[46] using the heuristic search option and performing 2,000 bootstrap replicates. All bifurcations showed high bootstrap support, with most of them around 100% and a minimum of 67%. Divergence times were collected from different sources using both molecular clocks as well as paleontological records[17]. Finally, ancestral sequence reconstruction is performed by maximum likelihood using PAML 4.8[47], incorporating a gamma distribution for variable replacement rates across sites and the JTT model. The ancestral sequences are listed in the **Supplementary Note**. Posterior probabilities were calculated for all 20 amino acids. In each site, the residue with the highest posterior probability was selected. We have resurrected titin I65–I72 fragments that belong to the last tetrapod common ancestor (LTCA) which lived around 356 ± 11 Ma; the last sauropsida common ancestor LSCA, 278 ± 14 Ma; the last mammal common ancestor LMCA, 179 ± 38 Ma; the last placental mammal common ancestor LPMCA, 105 ± 17 Ma.

Protein expression and purification. Genes encoding the ancestral and extant titin proteins were synthesized and codon optimized for expression in *E. coli* cells (Life Technologies). The genes were cloned into pQE80L vector (Qiagen) and transformed onto *E. coli* Origami2 cells with enhanced disulfide bond formation machinery (Merck Millipore). Bacteria were incubated overnight in LB medium at 37 °C, and 1 mM IPTG was add after reaching OD of 0.5 to induce protein expression. After centrifugation, cell pellets were lysed with French pressure cell press and the His_6-tagged proteins were loaded onto His GraviTrap affinity column (GE Healthcare). Oxidation was triggered by addition of 0.5% H_2O_2 overnight at room temperature. The proteins were then further purified by size exclusion chromatography using a Superdex 200HR column (GE Healthcare). The buffer used was 10 mM HEPES, pH 7.2, 150 mM NaCl, 1 mM EDTA at pH 7.0. The purified proteins were finally verified by SDS-PAGE. Trx and oxidized and reduced $(I91–32/75)_8$ were purified as described before[29,48].

Biochemical determination of oxidized cysteines. A protocol for in-gel determination of oxidized thiols was adapted and optimized from previous reports[32]. 1 μg of each protein was incubated with 10 mM N-ethylmaleimide in HEPES buffer in the presence of 3% w/v SDS for 30 min at 60 °C, in order to block all initially reduced thiols by irreversible alkylation. Samples were subsequently run on a 12% SDS-PAGE gel, and oxidized thiols were then reduced by incubation of the gel with 10 mM DTT for 1 h at 60 °C. After three washes with 50 mM Tris-HCl, 10 mM EDTA, 3% w/v SDS, pH 6.8, the gel was incubated with a 5 mM mBBr solution in the same buffer for 2 h in the dark. Excess mBBr was removed by destaining the gel with 40% ethanol, 10% acetic acid overnight (3 changes). Fluorescent bands resulting from the reaction of the newly reduced thiols with mBBr were visualized on a Gel-Doc with UV excitation using standard filters for ethidium bromide emission. Quantification of the bands was performed by densitometry using the Quantity One software. The amount of protein in each well was later assessed by Coomassie staining and densitometry and was used to normalize fluorescence signals. The fluorescent background of a replicate gel that was not reduced with DTT was subtracted. Oxidized and reduced $(I91–32/75)_8$ control proteins were also included in the experiment. Using force-clamp measurements[23], we estimated that 99% of I91–32/75 domains of the oxidized sample were oxidized, whereas 95% of the domains in the reduced sample were reduced. These experimental values were used to estimate the number of disulfides in the I65–I72 samples.

Single-molecule force spectroscopy experiments. We performed the experiments on a commercial atomic force spectrometer (AFS) from Luigs & Neumann. Cantilever models MLCT and OBL-10 of silicon nitride were used (Bruker). We calibrated the cantilevers using the equipartition theorem giving rise to a typical spring constant of 0.02 N m^{-1} for MCLT and 0.006 N m^{-1} for OBL-10. The AFS works in the force–extension mode at a pulling speed of 400 nm/s and amplitude of 400 nm. In the force-clamp mode, the length resolution is 0.5 nm and the piezo feedback response can be as low as 1 ms. The buffer used in the experiment was 10 mM HEPES, pH 7.2, 150 mM NaCl, 1 mM EDTA and 2 mM NADPH. We added eukaryotic Trx enzyme to a final concentration of 10 μM. The buffer also contains eukaryotic Trx reductase (50 nM) to keep Trx enzyme in the reduced state. To perform the experiment, we deposited 10–20 μl of substrate ~0.1 mg ml^{-1} on a gold-covered coverslide. A drop of ~100 μl containing the Trx solution was then added. The force-clamp protocol consists of three pulses of force. In the first pulse, the cantilever tip was pressed against the surface at 800 pN for 2 s. In the second pulse, the attached I65–I72 titin was stretched at 135 pN for 2 s. The third pulse was the test force in which the reduction events were captured. This pulse was applied at 80 pN for 20 s to capture all possible reduction events. The traces were collected and analyzed using custom-written software in Igor Pro 6.37 (Wavemetrics). All proteins were measured following a double-blind protocol in which three independent researchers measured all samples labeled as A, B, C, etc., from different expressions, using several cantilevers for each protein and two AFS instruments. All figures were generated using Igor Pro and Adobe Illustrator CS6.

Data analysis. When we represent averages of the experimental observables (unfolding forces for the disulfide bonded and non-disulfide-bonded domains) the error bars indicate the s.e.m. The percentage of disulfide bonds was estimated as the fraction of unfolding events corresponding to the disulfide bonded (i.e. low extension) population. For the correlation between unfolding forces and mass, we use a power law expression including a predefined offset to fit the data, $F–F_0 = a \times m^b$, where F_0 corresponds to the lowest measured value for the force, reflecting a lower limit in the unfolding force of the titin domains. Least-squares fits were performed using the Levenberg-Marquardt algorithm.

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

44. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).
45. Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A.Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
46. Wilgenbusch, J.C. & Swofford, D. Inferring evolutionary trees with PAUP*. *Curr. Protoc. Bioinformatics* Chapter 6: Unit 6.4.1–6.4.28 (2003).
47. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
48. Kahn, T.B., Fernández, J.M. & Perez-Jimenez, R.Monitoring oxidative folding of a single protein catalyzed by the disulfide oxidoreductase DsbA. *J. Biol. Chem.* **290**, 14518–14527 (2015).

doi:10.1038/nsmb.3426