

# A Phylogenetic Model for the Prediction of Quantitative Characteristics and Applications to Cardiac Modelling

Richard Mann (Life Sciences Interface DTC)

Supervisors:

Professor Jotun Hein (Oxford Centre for Gene Function)

Dr Jonathan Whiteley (Computing Laboratory)

## **Abstract**

A phylogenetic model for the prediction of quantitative characteristics is developed. The model is tested against a non-phylogenetic control model on simulated and real data and conclusions are drawn as to its theoretical and practical predictive advantage over the control.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Comparative Biology . . . . .	3
1.2	A Theory for the Evolution of Quantitative Characteristics . .	4
1.3	A Mechanical Model of the Heart . . . . .	6
<b>2</b>	<b>Data</b>	<b>8</b>
2.1	Binding Affinity of $\text{Ca}^{2+}$ to TnC . . . . .	8
2.1.1	Preparation . . . . .	8
2.2	Phylogenies . . . . .	8
<b>3</b>	<b>Theory</b>	<b>11</b>
3.1	The 1-Parameter Phylogenetic Model . . . . .	11
3.1.1	Mathematical Foundation . . . . .	11
3.2	The Non-Phylogenetic Model . . . . .	12
3.3	Including Measurement Noise . . . . .	13
3.4	Updating Experimental Measurements . . . . .	13
3.5	Extension to Multiple Correlated Characteristics . . . . .	14
<b>4</b>	<b>Results</b>	<b>15</b>
4.1	Simulated Data . . . . .	15
4.2	Model Validation . . . . .	19
<b>5</b>	<b>Discussion</b>	<b>21</b>
5.1	Summary . . . . .	22
<b>6</b>	<b>Appendices</b>	<b>23</b>

# 1 Introduction

This project details the development of a phylogenetic model for the prediction of quantitative characteristics. The work is motivated by the recent publication by (Niederer et al., 2006) of a mechanical model of the heart including a comprehensive review of measured values of physical quantities used as parameters in the model. The aim of this project is to create a model whereby such parameters can be predicted in a given species by inferring their value from the measured values in other species using phylogenetic information.

## 1.1 Comparative Biology

Comparisons between different species is at the heart of modern biology. Even before the publication of *On the Origin of Species* (Darwin, 1859) taxonomists classified animals by similarities in their morphology, comparing skeletal structures, brain sizes etc., and their behaviour, habitat and geography. Since the discovery of evolution these characteristics have been used by zoologists and paleontologists to estimate the phylogenetic relationships between species, although often in a rather heuristic fashion. With the knowledge that organisms arise from a common ancestor it is natural that much can be learnt about properties and behaviours in one species by looking for homologies in others.

Comparative techniques focus on variation between homologous characteristics in multiple individuals or species. Since the advent of efficient DNA sequencing techniques the majority of comparative work has been done studying the variation of gene structures or other homologous DNA sequences. In the field of phylogenetic inference this has allowed the use of the ‘molecular clock’ whereby the time from divergence between two species is estimated from the absolute amount of variation in a region of the genome that is assumed to vary at a known and constant rate. Comparative genomics also allows geneticists to infer the function of genes in humans from the effect of their removal in other species, typically rodents or primates. At the individual level haplotype mapping (The International HapMap Project) aims to find correlations between single nucleotide polymorphisms in humans with information of medical value such as prevalence of disease or response to treatment, promising an era of individually tailored medicine based on genetic typing. On a larger timescale the study of highly conserved DNA

sequences across species allows the inference of functional regions under the assumption that conservation implies selection and thus function.

Compared to the growth of comparative genetics the growth of macroscopic comparative genetics has been much slower. For extant species it is easier and more reliable to base an inferred phylogeny on molecular data than on quantitative characteristics. In the age of information rich biology even a large study of morphological features will not compete with the sequencing of an genome in terms of the amount of information that can be generated. Nonetheless it is important that the biology does not simply become the study of more and more DNA sequences to the exclusion of everything else. This project aims to demonstrate an application of comparative biology that is applicable to quantitative characteristics at many scales.

A statistical model for the evolution of quantitative characteristics would have a huge range of possible applications. Since the peak of phylogenetic inference from morphological features the statistics of morphology has advanced considerably. There is now a whole field of statistics dedicated to the evaluation and variation of shape. The development of active shape models in areas such as face recognition, for example (Cootes et al., 1999), has enabled variation in shapes and forms to be described by vectors of reference points that vary as a superposition of principal components in some parameter space. Once the variation has been reduced to the magnitude of these components we can ask how these magnitudes evolve just as for any other characteristic.

Essentially we hypothesise that any property of an organism that can be parameterised and quantified is suitable for modelling under a statistical evolutionary framework. The aim of this project is to investigate the properties of simple formulation of this framework and to stimulate further work on such modelling.

## **1.2 A Theory for the Evolution of Quantitative Characteristics**

Even up to the late 1970's and early 1980's, before modern DNA sequencing methods were available, the most statistically rigorous means for inferring a phylogeny was through the use of quantitative characteristics. By understanding how such a characteristic evolves over time along each branch of a topology it is possible to calculate the likelihood, thus allowing a search for

the maximum likelihood estimation of the true phylogeny. While this method was not widely employed, phylogenetics relying largely on more heuristic methods such as minimum evolution, the theory behind it is fully developed.

The discovery of evolution and genetics led to a rapid development of evolutionary statistics in the early 20th century. Fisher (Fisher, 1930) developed a theory of genetic drift where gene frequency varied under Brownian motion mediated by natural selection. For quantitative characteristics where the underlying genetic basis is unknown the value of the characteristic is assumed to be due to the additive contributions of many independent loci, leading to Brownian motion of the characteristic in the absence of selection. This assumes that the contributions from each locus are independent and that they are inherited independently. A model of evolution that uses Brownian motion therefore reflects our ignorance about the underlying genetic cause of a characteristic and its fitness profile. This is largely applicable over small enough evolutionary timescales but can lead to serious problems as timescales become large. As an additional point it should be noted that if the contributions of the loci are multiplicative rather than additive a log-normal distribution will result (Limpert et al., 2001). We will see that this has some relevance in dealing with real data in this project. In the case such data the Brownian motion model can obviously be used once the data has been transformed logarithmically.

Felsenstein (Felsenstein, 2004) gives a thorough review of quantitative characteristic evolution. To summarise, if  $N$  individuals are randomly selected (neutral evolution) from a population with original variance  $V_A$  and mean  $\mu$  then the mean  $\mu'$  and variance  $V'_A$  of the selected sample will be distributed as

$$\mu' \sim N(\mu, V_A/N) \quad (1)$$

$$V'_A = (1 - \frac{1}{2N})V_A \quad (2)$$

We then assume that each of the selected sample reproduce, producing infinitely many offspring in the next generation before another sample is selected. Each individual may suffer a mutation in any allele with some probability  $p$ . The allele will thus have a random quantity added to it, 0 with probability  $(1-p)$  or  $N(0, \sigma_m)$  with probability  $p$ , where  $\sigma_m$  is a genetic variance. The variance of the amount added to the allele is  $p\sigma_m^2$ , so for the whole population with  $n$  loci contributing.

$$V_M = 2np\sigma_m^2 \quad (3)$$

The variance of one generation as a function of the previous is thus

$$V'_A = V_A(1 - \frac{1}{2N}) + V_M \quad (4)$$

and solving for  $V'_A = V_A$  at equilibrium

$$V_A = 2NV_M \quad (5)$$

Population mean shifts by equation (1). Replacing  $V_A$  by equation (5) gives

$$\mu' \sim N(\mu, \sigma^2 = 2V_M) \quad (6)$$

i.e. the population mean undergoes a random walk with variance  $2V_M$ . This is fortuitous as we will find we are bound to measure time in phylogenies in terms of mutations, therefore the the variance per generation is proportional to  $V_M$ , whereas time is inversely proportional to  $V_M$ , leading to a roughly constant mutation rate per unit of phylogenetic time.

### 1.3 A Mechanical Model of the Heart

The regular rhythmical beating of the heart can be described by an array of coupled differential equations, some describing the electrophysical behaviour of cells, others the chemical reactions within cells and still others describing how chemical changes are transformed into mechanical work. In (Niederer et al., 2006) a model is developed for the mechanical function of the heart, beginning with the binding of free  $\text{Ca}^{2+}$  ions to Troponin C (Tnc) and concluding with the production of tension and then relaxation in the heart muscle. This model contains a large number of parameters. Some of these correspond to measurable physical quantities, others are empirical fitting factors to allow the model to be fitting to known data from the heart function. (Niederer et al., 2006) contains a comprehensive review of all the mammalian data corresponding to measurements of the physical parameters. It is our hypothesis that within mammalian species evolutionary timescales are small enough to model the variation of these parameters between species by the Brownian motion model of evolution discussed above.

Mechanical force in the heart is generated by the binding of crossbridges interweaved myosin and actin filaments. When these crossbridges protruding from the myosin filaments bind to the actin filaments they undergo a conformational change that pulls the filaments in opposite directions, causing

tension. The actin binding sites can also be bound by TnC which in turn can be bound by  $\text{Ca}^{2+}$ . Thus a local change in calcium concentration can lead to force generation as more TnC is bound by  $\text{Ca}^{2+}$ , releasing more actin binding sites for binding with the myosin crossbridges. Tension can be modelled as proportional to the number of free actin binding sites. Therefore understanding the variations in calcium concentration and the the binding of  $\text{Ca}^{2+}$  to TnC is a prerequisite for understanding mechanical force generation in the heart.

The steady state binding of calcium to TnC can be described by a Hill equation

$$\frac{[\text{Ca}^{2+}]_{\text{bound}}}{[\text{Ca}^{2+}]_{\text{maxbound}}} = \frac{[\text{Ca}_{\text{free}}^{2+}]}{[\text{Ca}_{\text{free}}^{2+}] + K(T)} \quad (7)$$

where  $K(T)$  is the tension dependent binding affinity of calcium to TnC. This quantity  $K$  is measurable by experimental observations of the proportion of bound calcium. In (Niederer et al., 2006) this is the only physical quantity measured in a sufficient variety of species to permit phylogenetic analysis. It is noted that for these measurements the calcium concentration was very low, so we can assume that the measurements reflect  $K(0)$  with negligible tension. These measurements are contained in Table 1 of (Niederer et al., 2006) and are also reproduced in the appendices of this work. A summary of the data is presented in Table 1.

Ultimately it is to be hoped that sufficient data will exist on further parameters to enable a phylogenetic analysis of the full parameter set of the model. Further work on the comparative biology of the heart may also shed light of selective pressures or other factors that would modify the naive Brownian motion model of evolution.

## 2 Data

### 2.1 Binding Affinity of $\text{Ca}^{2+}$ to TnC

A complete table of data used from (Niederer et al., 2006) is included in the appendices.

#### 2.1.1 Preparation

Figure 2 of (Niederer et al., 2006) (Appendix A) indicates that the use of different forms of the Troponin complex to measure binding affinity has a systematic effect on the results. To compensate for this effect all measurements carried out using Tn, Tn-TnI or Tn-Tm were multiplied by a factor of 10. Plotting a histogram of the resulting data showed a distinctly non-Gaussian profile. The logarithm of the data did, however, appear to fit a Gaussian curve. Physically this can be explained because the binding affinity is the ratio of two rate constants. each of these is exponentially related to some free energy difference. It is to be expected that molecular changes in the DNA will affect the chemical free energy difference more fundamentally than the rate constants themselves, so we would expect these to follow Brownian motion. Therefore from this point on the parameter of interest is  $\log(k)$ . The data from (Niederer et al., 2006) is summarised in table 1

Table 1: Summary of Binding Affinity Data by Species

Species	$\text{Log}(k)/\mu\text{M}$	$\sigma(\log(k))/\mu\text{M}$
Human	2.2625	1.0456
Bovine	0.8735	1.1730
Canine	0.6244	1.1465
Rat	1.4532	0.4551
Chicken	1.7409	0.7221

### 2.2 Phylogenies

Phylogenies were estimated from the work by (Reyes et al., 2004). A full reproduction of their reconstructed mammalian phylogeny is included in the appendices. The phylogeny was inferred from variation in the 1st and 2nd codon positions of mitochondrial H-stranded protein-coding genes This work

uses 3 phylogenies of varying extent. Figure 1 is the basic phylogeny including only the species for which (Niederer et al., 2006) gives data, i.e. the species in Table 1. For simulation purposes we will want to investigate the

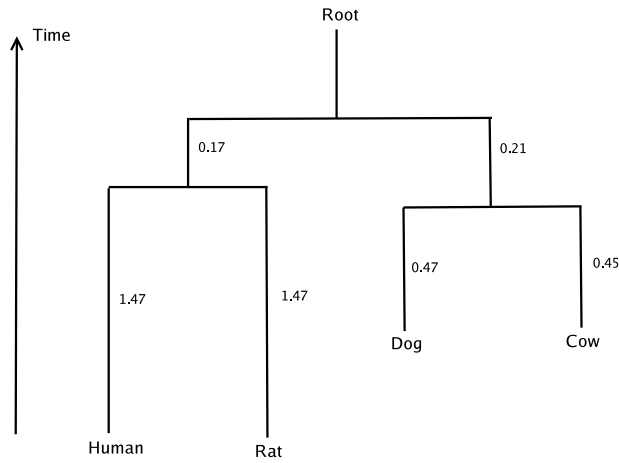


Figure 1: Phylogeny for the species in Table 1

effect of using phylogenies including primate species as these are likely to be necessary for making genuinely accurate predictions of human parameters. Figure 2 extends the basic phylogeny, including the pygmy ape. This is the closest species to humans on the phylogeny in (Reyes et al., 2004). Figure 3 extends the phylogeny further by including more primate species, gorilla and orangutan. All branch lengths in these figures are proportional to the number of nucleotide substitutions per site as per the scale in (Reyes et al., 2004).

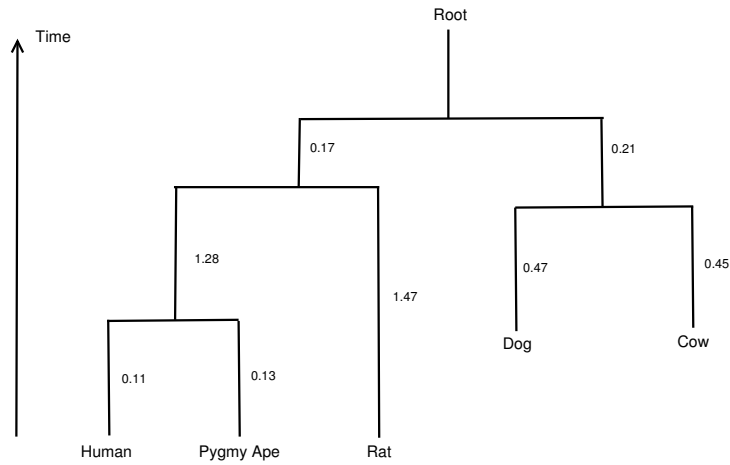


Figure 2: Extension of Figure 1 to include the pygmy ape

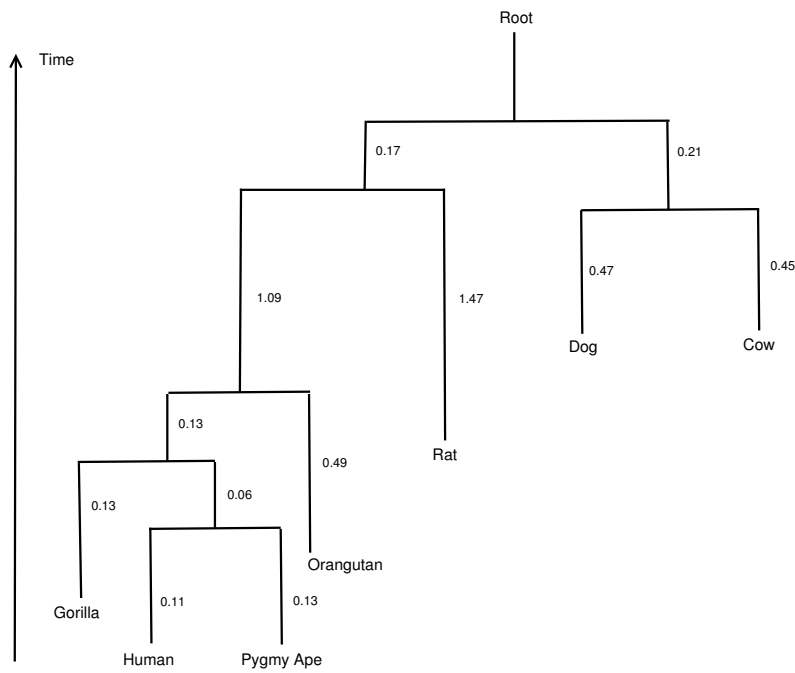


Figure 3: Further extension of Figure 1 to include more primate species

### 3 Theory

The aim of this work is to determine whether the use of prior knowledge about the phylogeny significantly affects our prediction of unknown parameters. Therefore we must set up 2 competing models, with and without phylogenetic components. This section discusses the formulation of the models. The only mathematical difference between them is the covariance of parameters between species. As such the phylogenetic model is formulated and the non-phylogenetic model discussed as a special case of this model.

#### 3.1 The 1-Parameter Phylogenetic Model

##### 3.1.1 Mathematical Foundation

The theory presented in the introduction describes quantitative characteristics evolving under Brownian motion as the result of random genetic drift. In the absence of knowledge of selective pressures neutral evolution is assumed. This leads to a multivariate normal distribution of the parameters for each species. The covariance of the distribution is defined by the phylogeny, each entry being proportional to the shared evolutionary time of the 2 species.

$$\mathbf{k} \sim \text{MVN}(\mathbf{k}_{\text{root}}, \Sigma) \quad (8)$$

The pdf of the multivariate normal distribution for  $N$  variates is a generalisation of the standard Gaussian pdf.

$$P(\mathbf{k} = \mathbf{x}) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{k}_{\text{root}})^T \Sigma^{-1} (\mathbf{x} - \mathbf{k}_{\text{root}})\right) \quad (9)$$

There is a standard result for the conditional probability of one variable in a MVN distribution. Let us assume that the parameters of all species are collected into a vector  $\mathbf{k}$  where the 1st element is the human parameter,  $k_H$  and the parameters for the other species form a vector  $\mathbf{k}_{\bar{H}}$ . Let us also decompose the covariance matrix into 4 components. A  $(1 \times 1)$  scalar  $\Sigma_{11}$  which determines the self-variance of the human parameter, an  $(N - 1 \times 1)$  vector  $\Sigma_{21}$  and its transpose  $\Sigma_{12}$  which give the covariance between the human and non-human parameters and an  $(N - 1 \times N - 1)$  matrix  $\Sigma_{22}$ , the covariance matrix of non-human parameters.

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \quad (10)$$

With these definitions we can write the distribution of all the parameters and the conditional distribution of the human parameter as a function of the non-human parameters

$$\mathbf{k} \sim \text{MVN}(k_{\text{root}}, \Sigma) \quad (11)$$

$$k_H | \mathbf{k}_{\bar{\mathbf{H}}} \sim N(\bar{k}, \bar{\Sigma}) \quad (12)$$

Where

$$\bar{k} = k_{\text{root}} + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{k}_{\bar{\mathbf{H}}} - k_{\text{root}}) \quad (13)$$

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \quad (14)$$

Up to this point we have assumed complete knowledge of  $k_{\text{root}}$  and  $\Sigma$ . In truth we can only estimate these from current information. A simple way to do this would be to use  $\mathbf{k}_{\bar{\mathbf{H}}}$  to obtain a Maximum Likelihood Estimate (MLE) of  $k_{\text{root}}$  under the covariance  $\Sigma_{22}$ . An MLE approach does not, however, fully consistently propagate our uncertainty. A better approach is a Bayesian model with an uninformative prior on the unknown parameters. This essentially amounts to an integration over the unknowns weighted by the likelihood. This integration gives the central equation for our model. Defining  $\Sigma = \sigma^2 \Sigma'$  where  $\sigma^2$  is a varying normalising variance and  $\Sigma'$  contains the topological information that does not vary.

$$P(k_H | \mathbf{k}_{\bar{\mathbf{H}}}) = \frac{\int P_N(k_H | k_{\text{root}}, \sigma, \bar{\Sigma}') P_{\text{MVN}}(\mathbf{k}_{\bar{\mathbf{H}}} | k_{\text{root}}, \sigma, \Sigma'_{22}) dk_{\text{root}} d\sigma}{\int P_{\text{MVN}}(\mathbf{k}_{\bar{\mathbf{H}}} | k_{\text{root}}, \sigma, \Sigma'_{22}) dk_{\text{root}} d\sigma} \quad (15)$$

For numerical integration there needs to be a specified range over which the sum is taken. In this integration this means specifying the range of  $\sigma$  and  $k_{\text{root}}$ . Simulations of the likelihood under the PM produce the likelihood profile in Figure 4 (a). It can be seen that the likelihood is very sharply peaked on the mean and with fairly small width on the standard deviation. Figure 4 (b) shows that this is true of the likelihood under the NPM too. Therefore we can perform the integration to include all of this peak and obtain a very good estimate to the true integral.

## 3.2 The Non-Phylogenetic Model

Ignoring knowledge about the phylogeny is to assume that each parameter has evolved independently. The parameters still form a draw from a MVN distribution, this time with a diagonal covariance. Using our previous definition of  $\Sigma'$  as the covariance normalised by the variance per unit time the

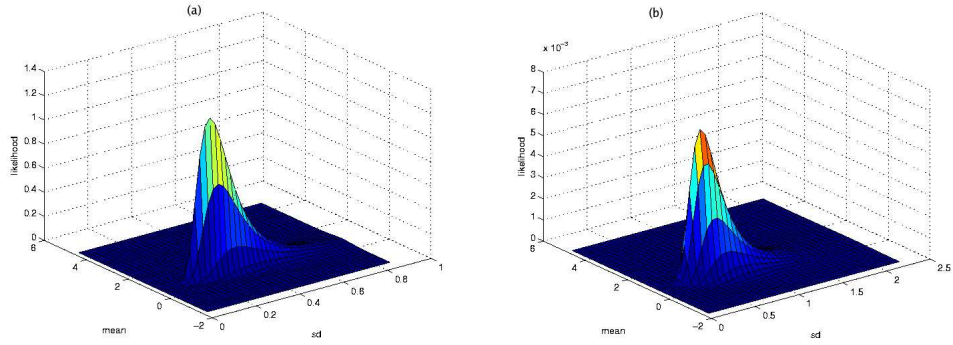


Figure 4: The likelihood as a function of the root mean and the standard deviation under the phylogenetic model (a) and the non-phylogenetic model (b)

non-phylogenetic model (NPM) would have  $\Sigma' = \mathbf{I}$ . With this covariance we can apply equation (15) to obtain our NPM estimate for the human parameter.

### 3.3 Including Measurement Noise

A crucial part of our uncertainty comes not only from transferring measurements to humans but also through the errors in the measurement process itself. These errors can be included in the framework of the multivariate model. The covariance matrix due to measurement error forms a diagonal matrix, where we assume that measurement errors are independent between species. The values for this matrix can be estimated via the sample variance for the measurements on each species, which is the maximum likelihood estimate for the measurement noise. This matrix is then added to the phylogenetic covariance to form the full covariance matrix.

### 3.4 Updating Experimental Measurements

In the case where experimental measurements exist for the human parameter we can use data from the rest of the phylogeny to improve our estimate, reducing the variation due to measurement error. Equivalently we can use the measurement data to update our phylogenetic estimate.

We treat the measured data as prior information on  $k_H$ . We can then update the measurement via Bayes' rule. If  $D$  is our measurement data,

$$P(k_H|\mathbf{k}_{\bar{\mathbf{H}}}, D) = \frac{P(\mathbf{k}_{\bar{\mathbf{H}}}|k_H)P(k_H|D)}{\sum_{k_H} P(\mathbf{k}_{\bar{\mathbf{H}}}|k_H)} \quad (16)$$

### 3.5 Extension to Multiple Correlated Characteristics

If we were to obtain measurements of the whole parameter set we may find that some vary in correlation with each other. This should provide extra information to improve the estimate of the human parameter set rather than treating each character independently. It is relatively simple to extend the one parameter framework to many parameters. If we assume a known covariance between characteristics within one species  $\Sigma_{\text{internal}}$ , and label the phylogenetic covariance  $\Sigma_{\text{external}}$ , then the total covariance is the Kronecker product

$$\Sigma = \Sigma_{\text{internal}} \otimes \Sigma_{\text{external}} \quad (17)$$

Now we have a multivariate distribution over the human parameter set

$$\mathbf{k}_{\mathbf{H}} \sim \text{MVN}(\bar{\mathbf{k}}, \bar{\Sigma}) \quad (18)$$

With parameters

$$\bar{\mathbf{k}} = \mathbf{k}_{\text{root}} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{k}_{\bar{\mathbf{H}}} - \mathbf{k}_{\text{root}}) \quad (19)$$

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \quad (20)$$

$\mathbf{k}_{\bar{\mathbf{H}}}$  is now a stacked vector containing the vectors of parameters in each non-human species. To reflect the uncertainty in ancestral parameter values and variance per unit time equation (15) can be updated to give an integration over unknowns

$$P(\mathbf{k}_{\mathbf{H}}|\mathbf{k}_{\bar{\mathbf{H}}}) = \frac{\int P_{\text{MVN}}(\mathbf{k}_{\mathbf{H}}|\mathbf{k}_{\text{root}}, \sigma, \bar{\Sigma}')P_{\text{MVN}}(\mathbf{k}_{\bar{\mathbf{H}}}| \mathbf{k}_{\text{root}}, \sigma, \Sigma'_{22})d\mathbf{k}_{\text{root}}d\sigma}{\int P_{\text{MVN}}(\mathbf{k}_{\bar{\mathbf{H}}}| \mathbf{k}_{\text{root}}, \sigma, \Sigma'_{22})d\mathbf{k}_{\text{root}}d\sigma} \quad (21)$$

## 4 Results

### 4.1 Simulated Data

The ideal case was simulated where we assume no measurement noise and perfect Brownian motion describes the evolution. To this end data was simulated by drawing from a MVN using a standard technique where the covariance of the distribution was provided by the phylogenetic topology with a standard deviation of 0.5 per unit length and a root mean of 1. Both the phylogenetic model and the non-phylogenetic model were used to infer a distribution of the human parameter conditional upon the values for the other species from this simulated data. This gives a measure of the maximum predictive advantage the phylogenetic model could provide if it described the evolution perfectly. Figure 5 shows a typical pair of posteriors generated by both models for simulation using the basic phylogeny in Figure 1. The measure of success will be the ratio between the likelihood of the simulated human data under each model. Repeating the simulation and prediction cycle 1000 times generates a distribution of the likelihood ratio that approximates a Gaussian profile (Figure 6). The results can thus be characterised by a mean value and the standard deviation. For the basic phylogeny the average likelihood ratio is

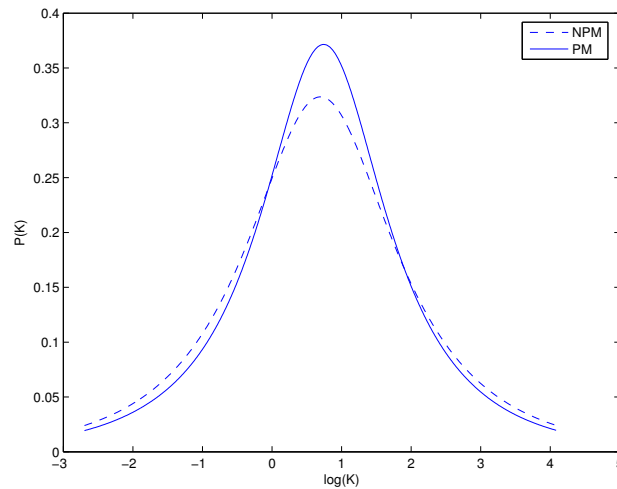


Figure 5: A typical pair of posteriors from the non-phylogenetic (NPM) and phylogenetic (PM) models based on simulated data

marginally greater than 1.

$$\left\langle \frac{L(\text{PM})}{L(\text{NPM})} \right\rangle = 1.0163 \pm 0.1361 \quad (22)$$

This confirms, as one would expect, that if the characteristics in question genuinely evolve under Brownian motion then a phylogenetic model will, on average, predict the true value of a parameter with greater likelihood. However, the predictive advantage is small and variable enough to encompass a large region where the NPM is more successful. Under this phylogeny it will be very difficult to distinguish between the 2 models. Further sim-

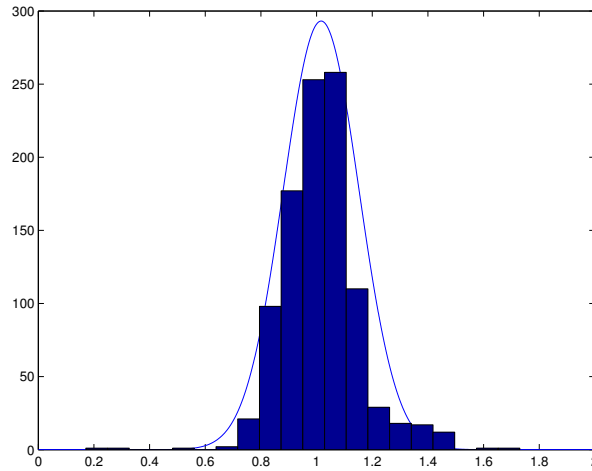


Figure 6: A histogram showing the frequency of likelihood ratios measured on generated data

ulations are performed on the extended phylogenies. Inclusion of primate species with recent common ancestor with humans cause the 2 models to diverge in their predictions. Figures 7 and 8 show that this is the case. The posteriors under the 2 models are more divergent and the average likelihood ratio has increased. The greatest difference between the models is found when just one species close to the human lineage is added as in Figure 2. The distribution of likelihood ratios in Figure 7 (a) is characterised by

$$\left\langle \frac{L(\text{PM})}{L(\text{NPM})} \right\rangle = 2.0252 \pm 0.8980 \quad (23)$$

One might expect that the further addition of more primate species would

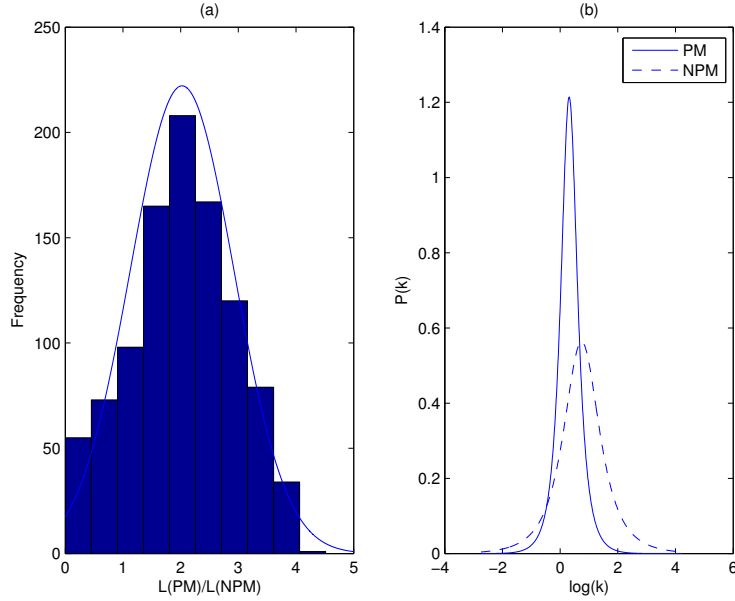


Figure 7: Distribution of predictive likelihoods (a) and typical posterior predictions (b) for a phylogeny containing pygmy ape

increase the relative advantage of the phylogenetic model. In fact it does increase the absolute predictive power, shown by the smaller variance in the posterior in Figure 8 (b), but the distribution of likelihood ratios is characterised by

$$\left\langle \frac{L(\text{PM})}{L(\text{NPM})} \right\rangle = 1.9172 \pm 0.8813 \quad (24)$$

Under the assumption of Brownian motion information will decay rapidly with the evolutionary time between 2 species, hence the large improvement of a phylogenetic model when there is a large discrepancy between the divergence times of each species with humans. This might also lead us to expect that a model containing only those species within some limited distance of humans would be almost as good as the full model. This is indeed what we find in the likelihood ratio distribution between the full model and the a

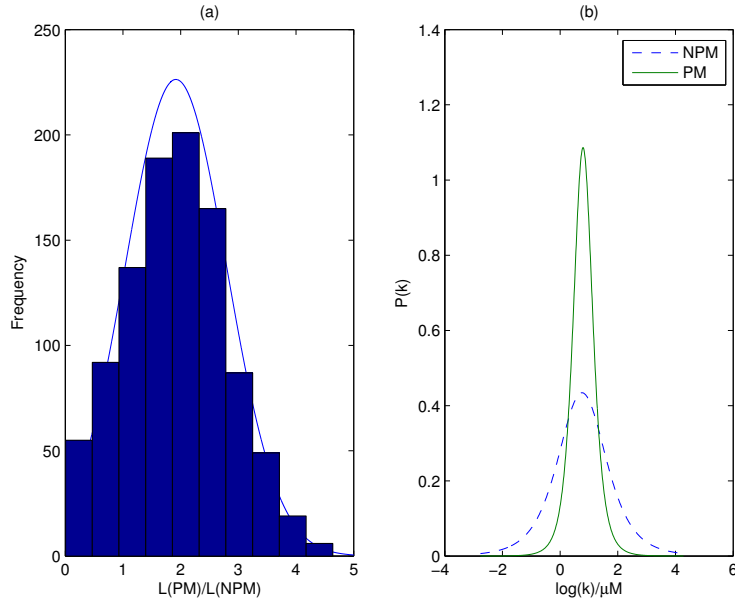


Figure 8: Distribution of predictive likelihoods (a) and typical posterior predictions (b) for a phylogeny containing pygmy ape, gorilla and orangutan

model limited only to primates (Figure 9). This ratio is characterised by

$$\left\langle \frac{L(\text{Full})}{L(\text{Limited})} \right\rangle = 1.1186 \pm 0.5439 \quad (25)$$

The likelihood ratio between these 2 models is small enough to make them practically indistinguishable, especially considering that the Brownian motion assumption becomes more questionable with greater evolutionary separation. Where there are multiple species close to humans from which to infer it is likely that the best model will only include those species. Only in the case where, like in Figure 2, only 1 species is close to the human line is it advisable to include the wider phylogeny as it is not possible to estimate the variance per unit time from a single sample species.

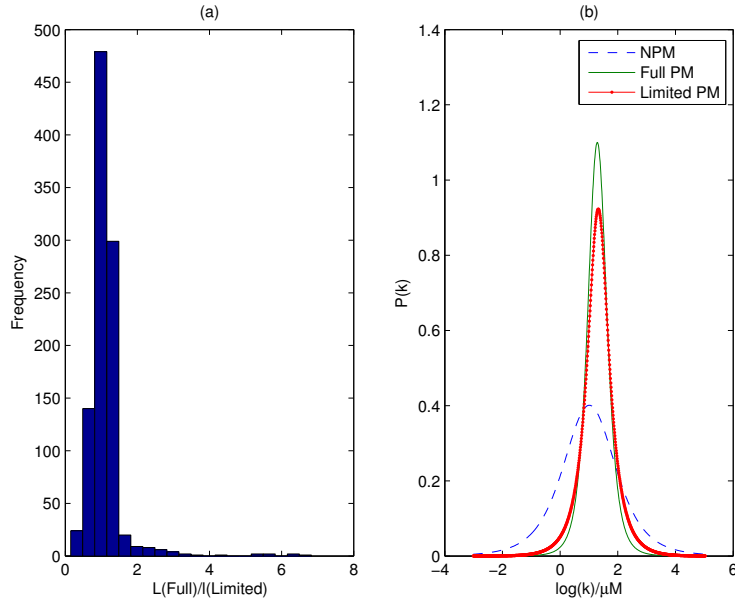


Figure 9: Distribution of predictive likelihoods (a) and typical posterior predictions (b) for the full phylogenetic model and the model limited only to primate species for the phylogeny in Figure 3

## 4.2 Model Validation

A goodness-of-fit test was performed on the data for a MVN and a simple one-dimensional Gaussian. The likelihoods under each were

$$L(\mathbf{k}|\Sigma) = 0.0300 \quad (26)$$

$$L(\mathbf{k}|I) = 0.0296 \quad (27)$$

$$\frac{L(\mathbf{k}|\Sigma)}{L(\mathbf{k}|I)} = 1.0153 > 1 \quad (28)$$

As we would expect for such a small data set the results are inconclusive. The same is true of when the predictive power of the 2 models is tested. Figure 10 shows the posterior distribution on the parameter for each species inferred from the measured values in the other 3 species under each model. The combined likelihood ratio gives a measure of the predictive success

$$\prod \frac{L(\text{PM})}{L(\text{NPM})} = 0.9842 \quad (29)$$

Again the result is inconclusive as we would expect from the simulated

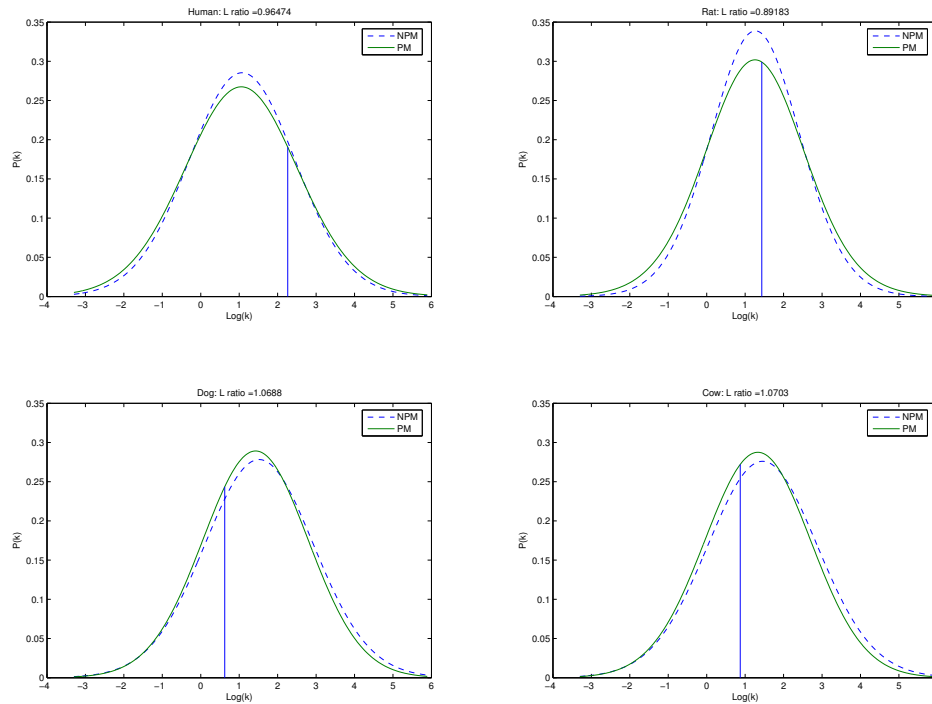


Figure 10: Inferred posterior distributions for each species under both models conditional on the measured values from the other 3 species. Actual measured value marked by the vertical lines.

example on this phylogeny. The best that can be said is that there is no significant evidence (at the 95% level) that these likelihood ratios are not drawn from the distribution inferred from Figure 6 with mean and standard deviation given by equation 22. It is certainly impossible to make any positive statement about the relative effectiveness of the models from this data alone. As can be seen from the simulated data it is highly unlikely that any statistically significant evidence for the superiority of a phylogenetic model can be found without the use of species much close to the human lineage.

## 5 Discussion

It was not possible with the available data to make a judgement on the effectiveness of a phylogenetic model for parameter prediction. Simulated examples showed that even if the parameters of interest evolved exactly as the model predicts it would be nearly impossible to distinguish between the predictive accuracy of the phylogenetic model and the control. An attempt was made to find more data over a wider phylogeny by reviewing cardiac electrophysiology models (Winslow et al., 1999; Puglisi and Bers, 2001; Pandit et al., 2001; ten Tusscher et al., 2004). However, although homologous parameters could be found in these models, attempts to validate the model under this data proved unsuccessful. It is believed that this is due to a lack of independence between the electrophysiology models, many of which use available parameter measurements from other species or transplant parameter sets from previous models in other species. Hence parameter set variation could not correspond to phylogenetic relation.

It is to be hoped that in time larger data sets may become available to test the model on. The supposed wide applicability suggests that such a data set must exist for some measured property of related organisms. The work on simulated data shows that if evolutionary timescales across the phylogeny are sufficiently varied it would be possible to assess whether the quantities in question had evolved as the model suggests. Suggested fields of investigation would included morphology via the statistics of shapes, especially active shape models, which in the context of comparative heart research would describe the evolution of cardiac morphology, representing changes in size, shape and structure over evolutionary time. Paleontology and anthropology may also provide viable data sets such as measurements of the variation in human skulls or other bone and fossil measurements. Of course it would be necessary to have an alternative method of constructing a phylogeny for these cases that was independent of the quantity being modelled.

It remains belief of the authors that this type of model is the natural formulation of variation in such parameter sets and may in the future prove of value in predicting values, especially in humans, where direct experimentation is not possible. However, the conclusions of this project suggest that at this time the model will be of little use in the absence of further data. To assess the possible usefulness to heart modelling a full stability analysis would be required on the models in question. At this time there is little idea exactly how sensitive cardiac models are to variations in each of their

parameters. There is little reason to be excited by a 5% improvement in the prediction of a parameter that can vary by twice its own value with negligible effect. Likewise even a 1% improvement could be useful for a parameter whose influence is sufficiently strong and non-linear.

## 5.1 Summary

A phylogenetic model was formulated for the prediction of quantitative characteristics based on the approximation of characteristics evolving under Brownian motion. The conditional probability of one variable in a multivariate normal distribution provided the mechanism for generating posterior distributions of the parameter of interest conditioned on the values of the parameter in other species. The covariance of the distribution was defined by the shared evolutionary time of pairs of species and an estimation of the measurement error in each species. Likelihood weighted summation over possible ancestral parameter values and variance per unit time completed the posterior calculation.

Tests on simulated data suggested that the available data would be insufficient to distinguish the predictive accuracy of the phylogenetic model from the non-phylogenetic control. This proved to be true. Further tests on simulated data under the hypothesis of further data revealed that with sufficient variation in the phylogenetic relationships across the tree, specifically by including species much closer to humans, it would be possible to assess whether the model was more accurate in its predictions than the control, and thus whether the parameters under consideration had indeed evolved as the model predicts.

## Acknowledgements

Alongside help from the project supervisors the author would also like to acknowledge Steven Niederer and Nicolas Smith of the Bioengineering Institute, University of Auckland for their help with the cardiac modelling aspect of the project.

## 6 Appendices

### A: (Niederer et al., 2006) Table 1 and Figure 2

**TABLE 1** Binding affinities of  $\text{Ca}^{2+}$  to site II of cardiac troponin C

Species	Temp ( $^{\circ}\text{C}$ )	Troponin complex	Bound $\text{Ca}^{2+}$ measure	Mg (mM)	$K_d$ ( $\text{M}^{-1}$ )	$K$ ( $\mu\text{M}$ )	Ref.
Human	30	NTnC	NMRs	—	$4 \times 10^5$	2.5	(18)
Human	15	TnC	F27W	None	$4.2 \times 10^4$	24	(25)
Human	15	TnC	F27W	3	$1.4 \times 10^5$	7.1	(25)
Human	30	TnC	NMRs	—	$5 \times 10^4$	20	(24)
Bovine	4	TnC	SC	None	$2.5 \times 10^5$	4.0	(13)
Bovine	4	TnC	SC	4	$2.5 \times 10^5$	4.0	(13)
Bovine	—	TnC	IAANS	3	$7 \times 10^5$	1.4	(35)
Bovine	7	TnC	F27W	—	$9.3 \times 10^4$	11	(17)
Bovine	21	TnC	F27W	—	$1.9 \times 10^5$	5.3	(17)
Bovine	37	TnC	F27W	—	$2.6 \times 10^5$	3.9	(17)
Mammal	RT	TnC	IAANS	3	$2.5 \times 10^5$	4.0	(23)
Mammal	RT	TnC	IAANS	None	$4.5 \times 10^5$	2.2	(23)
Mammal	21	TnC	F27W	None	$2 \times 10^5$	5.0	(19)
Rat	4	TnC	IAANS (84)	3	$3.2 \times 10^5$	3.1	(20)
R/B	23	TnC	IAANS	3	$7.2 \times 10^5$	1.4	(21)
Chicken	23	TnC	IAANS (84)	None	$2.9 \times 10^5$	3.5	(22)
Chicken	23	TnC	IAANS	None	$3.6 \times 10^5$	2.8	(22)
Chicken	23	TnC-TnI	IAANS (84)	5	$8 \times 10^5$	1.3	(22)
Bovine	—	TnC-TnI	IAANS	3	$1.5 \times 10^7$	0.07	(35)
Rat	—	TnC-TnI	IAANS	3	$1.7 \times 10^6$	0.59	(26)
R/B	23	TnC-TnI	IAANS	3	$1.5 \times 10^6$	0.67	(21)
Mammal	RT	TnC-TnI	IAANS	None	$3 \times 10^6$	0.33	(23)
Bovine	4	TnC-TnI	SC	4	$1 \times 10^6$	1.0	(13)
Bovine	4	TnC-TnI	SC	None	$1 \times 10^6$	1.0	(13)
Chicken	23	Tn	IAANS	5	$1.2 \times 10^6$	0.83	(22)
Bovine	4	Tn	SC	None	$2.5 \times 10^6$	0.40	(13)
Bovine	4	Tn	SC	4	$2.5 \times 10^6$	0.40	(13)
Bovine	25	Tn-Tm	IAANS	2.5	$1.2 \times 10^6$	0.83	(27)
P/C*	RT	SP	IAANS (84)	1	$6.3 \times 10^5$	1.6	(29)
R/C	RT	SP	IAANS (84)	1	$6.3 \times 10^5$	1.6	(29)
Bovine	25	SP	SC	5	$4 \times 10^{6\dagger}$	0.25	(33)
Bovine	25	SP	SC	5	$2 \times 10^{6\dagger}$	0.50	(32)
Bovine <sup>‡</sup>	25	Tn-Tm-A	SC	2.5	$4 \times 10^5$	2.5	(27)
Bovine	25	Tn-Tm-A	SC	2.5	$9.6 \times 10^5$	1.0	(27)
Bovine	25	Tn-Tm-A	IAANS	2.5	$1.1 \times 10^6$	0.91	(27)
Bovine	25	SP	SC	5	$2 \times 10^{6\dagger}$	0.5	(34)
R/C	23	SP	IAANS	1	$2 \times 10^5$	0.5	(22)
R/C	23	SP	IAANS (84)	1	$4.7 \times 10^5$	2.1	(22)
Canine	25	SP	SC	2, 10	$1.2 \times 10^6$	0.83	(28)
Canine	25	SP	SC	2	$2.36 \times 10^5$	4.2	(31)

IAANS is IAANS-labeled TnC; IAANS (84) is IAANS-labeled TnC, with Cys amino acids at residue 84; NMRs is NMR spectroscopy; None =  $<1 \times 10^{-3}$  mM; P/C is porcine fiber with chicken TnC; R/B is rat fiber/bovine Tn-Tm; R/C is rat fiber with chicken TnC; SC is scintillation counting; and SP is skinned preparation. RT is room temperature.

\*BDM added.

<sup>†</sup>Affinity for sites II, III, and IV combined.

<sup>‡</sup>IAANS bound, but not used to measure affinity.

Figure 11: Reproduction of Table 1 from (Niederer et al., 2006) showing full data on measured binding affinities

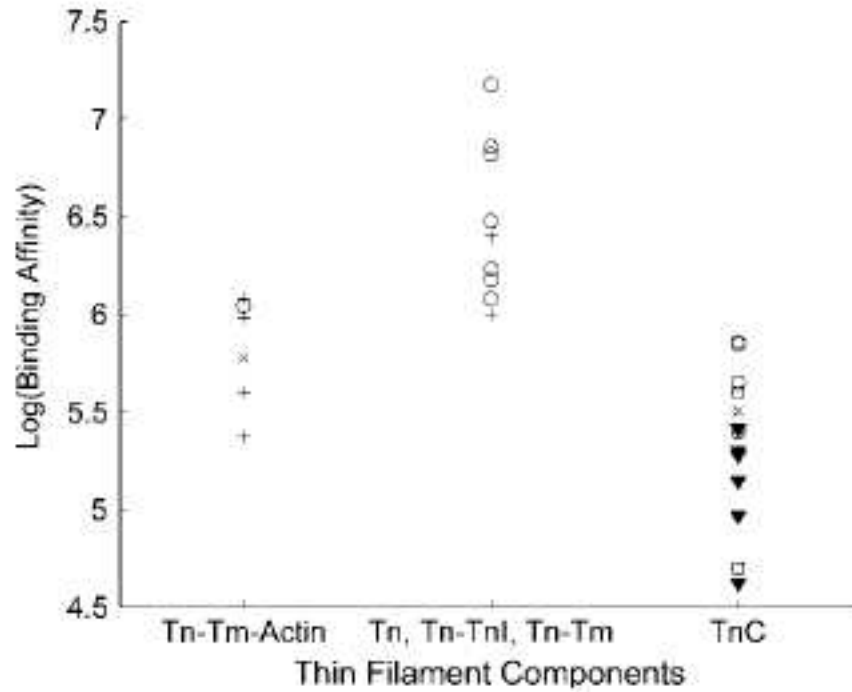


FIGURE 2 Affinity of  $\text{Ca}^{2+}$  to TnC contained in various components of mammalian cardiac thin filaments, from studies listed in Table 1. The plus-symbol (+) is scintillation counting;  $\circ$  is IAANS (Cys-35, Cys-84);  $\times$  is IAANS (Cys-35);  $\blacktriangledown$  is F27W; and  $\square$  is MRI spectroscopy.

Figure 12: Reproduction of Figure 2 from (Niederer et al., 2006) showing systematic differences in measured binding affinity using different Troponin complexes NB: According to the terminology in this work the vertical axis represents  $\log(k^{-1})$

## B: Full Mammalian Phylogeny from (Reyes et al., 2004)

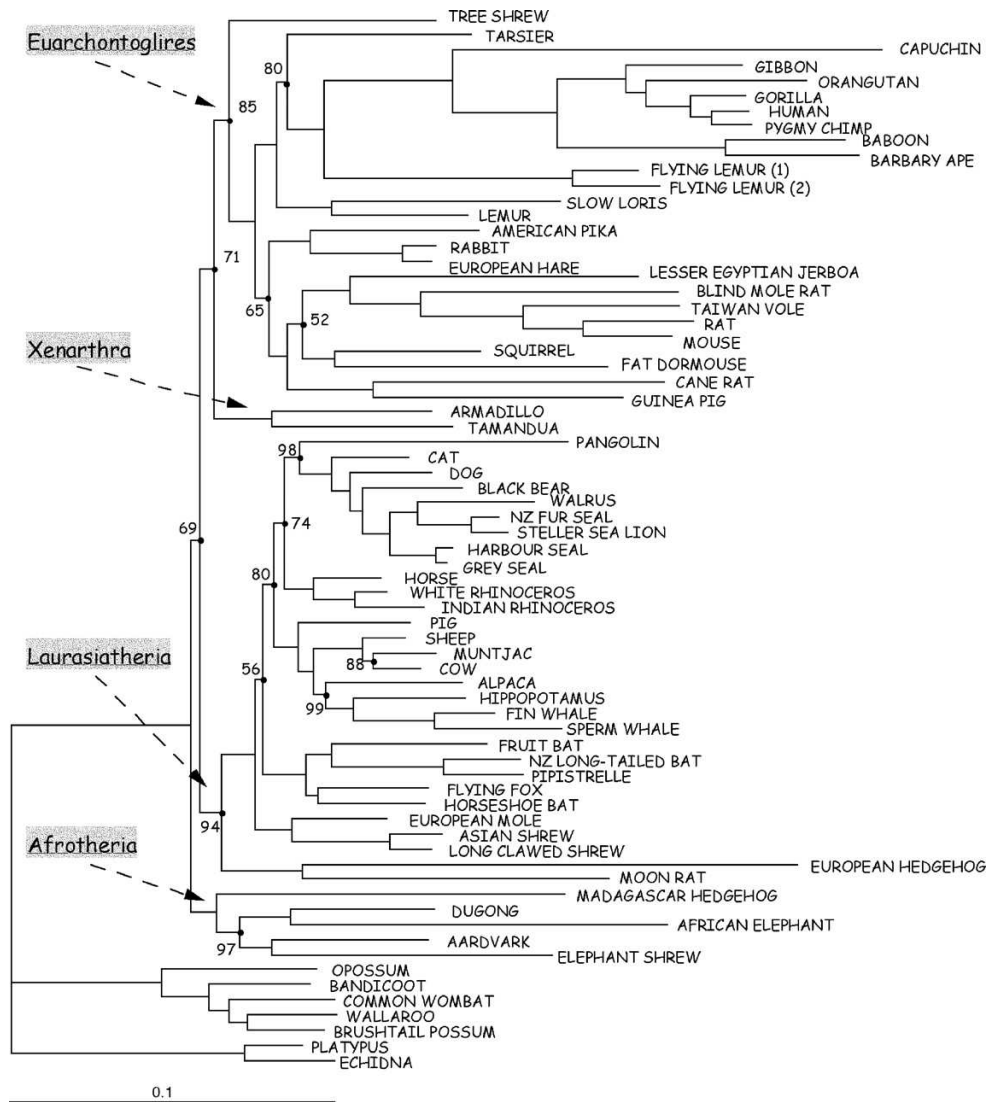


Figure 13: Phylogenetic tree reproduced from (Reyes et al., 2004). Lengths of branches are proportional to the number of nucleotide substitutions in the first and second codon positions of the mitochondrial H-standed protein-coding genes with the exclusion of Leu synonymous sites

## References

- T. F. Cootes, G. J. Edwards, and C. J. Taylor. Comparing active shape models with active appearance models. *Proc. British Machine Vision Conference*, 1:173–182, 1999.
- C. Darwin. *On the Origin of Species*. London: John Murray, 1859.
- J. Felsenstein. *Inferring Phylogenies*. Sinauer, 2004.
- R. A. Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, 1930.
- E. Limpert, W. A. Stahel, and M. Abbt. Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51:341–352, 2001.
- S. A. Niederer, P. J. Hunter, and N. P. Smith. A quantitative analysis of cardiac myocyte relaxation: A simulation study. *Biophysical Journal*, 90:1697–1722, 2006.
- S. V. Pandit, R. B. Clark, W. R. Giles, and S. S. Demir. A mathematical model of action potential heterogeneity in adult rat left ventricular myocytes. *Biophysical Journal*, 81:3029–3051, 2001.
- J. L. Puglisi and D. M. Bers. LabHEART: an interactive computer model of rabbit ventricular myocyte ion channels and Ca transport. *Am. J. Physiol. Cell Physiol.*, 281:C2049–C2060, 2001.
- A. Reyes, C. Gissi, F. Catzeflis, E. Nevo, G. Pesole, and C. Saccone. Congruent mammalian trees from mitochondrial and nuclear genes using bayesian methods. *Mol. Biol. Evol.*, 21(2):397–403, 2004.
- K. H. W. J. ten Tusscher, D. Noble, P. J. Noble, and A. V. Panfilov. A model for human ventricular tissue. *Am. J. Physiol. Heart Circ. Physiol.*, 286c:H1573–H1589, 2004.
- The International HapMap Project. [www.hapmap.org](http://www.hapmap.org).
- R. L. Winslow, J. Rice, S. Jafri, E. Marbán, and B. O’Rourke. Mechanisms of altered excitation-contraction coupling in canine tachycardia-induced heart failure, ii: Model studies. *Circulation Research*, 84:571–586, 1999.