Downloaded from orbit.dtu.dk on: Apr 18, 2024

DTU Library

# Integrating chemistry, biophysics and physiology in the evolution of mammalian Myoglobins

**Dasmeh, Pouria**

*Publication date:*
2013

*Document Version*
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

*Citation (APA):*
Dasmeh, P. (2013). *Integrating chemistry, biophysics and physiology in the evolution of mammalian Myoglobins*. DTU Chemical Engineering.

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# Integrating chemistry, biophysics and physiology in the evolution of mammalian Myoglobins

**PhD Thesis**

**Pouria Dasmeh**

**Department of Chemistry**

**Technical University of Denmark**

**April 2013**

## Acknowledgement

This thesis is submitted to the Department of Chemistry, Technical University of Denmark, in partial fulfillment of the requirement for the PhD degree in the subject of Biophysical Chemistry. The work of this thesis has been carried out in the period May 2010 to April 2013 as well as 6 months research stay at Harvard University, department of chemistry and chemical biology, from February 2012 to July 2012.

I would first of all like to thank my supervisor Prof. Kasper P. Kepp for providing the opportunity to work with a fascinating, fundamental science and interdisciplinary research project. I have been always keen to work on molecular evolution with insights from chemistry and physics and this challenging dream was financed and supported in this project. Moreover, working with Kasper taught me to be more organized, disciplined, and analytical which are fundamentally important in science.

I had also a great opportunity to experience science atmosphere at Harvard in the group of Prof. Eugene Shakhnovich during my research stay. I will definitely not forget inspiring moments and discussions I had with Eugene, Adrian WR Serohijos, Shimon Bershtein, Murat Cetinbas, Ariel Weinberger and Amy Gilson in this period. I am also grateful to Prof. Randall W. Davis in Texas A&M University, department of marine biology, for the exhilarating collaboration we had in the modeling of dive time in marine mammals.

I would like to thank my officemates and friends Niels Johan Christensen and Alessandro Corozzi for their wonderful company during my PhD program. I had amazing moments discussing almost everything with them during our coffee and lunch hours. I am also very pleased

to acknowledge Pedram Samani and Mohammad Ghahramanpour, two of my old friends that I had, have, and will have discussions with them on physics of life.

I am grateful to many of my Iranian friends at DTU and other Danish institutions that in their company I had time to think of my other hobbies in life, philosophy and literature.

I owe every bit of success in my life to my family; my parents Esmat and Hossein, my siblings Uranus and Ali and my beloved wife, Pardis. She understands me and my interests better than anyone in the world. I am thankful for her unconditional supports and understanding.

You can also be grateful to the prominent people whose life styles were influential in your personal life and career. I owe Richard P. Feynman, the late physicist, a debt of gratitude for his wonderful character in science and his breathtaking lectures on physics. It is usual to be stressful for tons of reasons in your carrier and I could cope with them using his great insight:

"I haven't done anything important, well; I'm never going to do anything important. I used to enjoy physics and mathematical things because I used to play with it. So I decided I'm only going to do things for the fun of it; no importance at all." *Richard P. Feynman*.

Although in preparation of a thesis everything is presented in order, but the truth is a chaotic mind full of new questions and not knowing.

Pouria Dasmeh

April 2013

## Abstract

This work describes an integration between chemistry, molecular biophysics, physiology, sequence evolution and bioinformatics to better understand the evolution of mammalian myoglobins (Mb) in terms of their primary biochemical function (i.e., $O_2$ binding) and their thermodynamic stability (i.e., folding free energy).

First, we merge a large set of previously reported thermochemical data for Mb mutants with a physiological model of $O_2$ delivery in the skeletal muscle cells to quantify the functional proficiency of Mb mutants under various physiological conditions. We find that $O_2$-storage and –transport are distinct functions which depend on $O_2$ partial pressure and conclude that conserved residues in wild type (WT) Mb were fixated under a selection pressure of low $P_{O_2}$.

Second, we present an integrated model of convective $O_2$-transport and $O_2$-affinity of mutant Mb to quantify the impacts of mutations in Mb on the aerobic dive limits (ADL) of Weddell seals (*Leptonychotes weddellii*). We show that wild-type Mb traits are only superior under specific physiological conditions that critically prolong the ADL, action radius, and fitness of the seals.

Third, we deal with the observation of higher folding stabilities (i.e., $\Delta G_{\text{folding}}$) of cetacean Mbs compared to their terrestrial counterparts. Using ancestral sequence reconstruction, maximum likelihood and Bayesian tests to describe the evolution of cetacean Mbs, and experimentally calibrated computation of stability effects of mutations (i.e., $\Delta\Delta G_{\text{folding}}$), we observe accelerated evolution in cetaceans and identify seven positively selected sites in Mb. We show that these sites contribute to Mb stabilization by favoring hydrophobic folding, structural integrity, and intra-helical hydrogen bonds.

Finally, we ask a fundamental question that how a general protein phenotype such as folding stability, that was shown as an example to be positively selected in cetacean Mbs in the

third part of this thesis, affects the rate of protein evolution. Using a model that combines explicit evolution of Mb sequences, folding stability, and application of maximum likelihood (ML) estimation of evolution rate (ER), we find that ER predicted by ML methods is highly correlated with ER from simulations using the explicit sequence information by counting the number of synonymous and nonsynonymous mutations fixed in the population. We show that this agreement is strongest in the regime of high stability where proteins are mostly evolving neutrally. In the unstable regime where protein evolution is dominated by selection for stabilizing mutations we detect a weak yet significant positive selection for specific residues in the sequences from simulations of the order of $dN/dS \sim 1.5$.

Overall, this thesis provides one of the first examples in the study of evolution of function and thermodynamic stability in a mammalian protein with implications for fitness. We quantify and highlight the biological relevance for the selection of a higher concentration of Mb in the skeletal muscle of marine mammals and provide an explanation for the increase in folding stability of Mb. Moreover, this thesis provides number of theoretical findings directly testable with future experimental studies regarding the function, physiology and thermodynamic stability of mammalian Mbs.

## Dansk resumé

Denne afhandling beskriver en integration mellem kemi, biofysik, fysiologi, sekvens-evolution og bioinformatik med det formål at forstå udviklingen af pattedyrs myoglobiner, især med hensyn til deres primære biokemiske funktion, $O_2$-binding, og deres termodynamiske stabilitet.

Først udvikler vi en ny fysiologisk model af $O_2$ transport i muskelceller og benytter et stort termokemisk datasæt for myoglobin-mutanters $O_2$-bindingsstyrke til at rangere mutanternes fysiologiske formåen under forskellige fysiologiske betingelser. Vi viser, at $O_2$-opbevaring og-transport er to separate funktioner, der afhænger af partialtrykket (mængden) af $O_2$, og konkluderer, at bevarede aminosyrer i vildtype-myoglobin gør proteinet til et optimalt transportprotein ved forholdsvis lavt partialtryk af ilt.

For det andet præsenterer vi en integreret model for $O_2$-bindingsstyrke og $O_2$-opbevaring i musklerne til at forstå mutationers indvirkning på den aerobe dykkegrænse hos sæler (*Leptonychotes weddellii*). Vi viser, at vildtypens egenskaber kun er overlegne under særlige fysiologiske betingelser svarende til det rutine-mæssige dyk under udførsel af fysiologiske dykkerespons, der kritisk forlænger dykkelængde, aktionsradius, og dermed indirekte sælernes biologiske fitness.

For det tredje undersøger vi, hvordan stabiliteten (dvs. $\Delta G$ for foldning) af hvalers myoglobin har udviklet sig over tid i forhold til andre pattedyr, der lever på jordoverfladen. Ved at afdække de evolutionære forfædres sekvenser, og brug af statistiske metoder kan vi beskrive udviklingen af hvalers myoglobin og koble den til eksperimentelt kalibrerede stabilitetseffekter af de mutationer, der er fundet sted (dvs. $\Delta\Delta G_{\text{folding}}$). Derved observerer vi en accelereret udvikling af hvalers myoglobin og identificerer syv positivt selekterede positioner i sekvensen. Vi viser, at disse

positioner bidrager til stabilisering af proteinet ved at begunstige hydrofob foldning, strukturel integritet, og intra-helix hydrogenbindinger.

Til sidst stiller vi et grundlæggende spørgsmål: Hvilken betydning har en universel protein fænotype, stabiliteten, der som vist er positivt selekteret i hvalers myoglobin, har betydning for proteinets evolution. Ved hjælp af en model, der kombinerer eksplicit evolution af myoglobin-sekvenser, stabilitetsdata og anvendelse af maksimum-likelihood (ML) estimering af evolutionshastighederne, opdager vi, at hastigheden beskrevet vha ML metoder er stærkt korreleret med den nøjagtige hastighed, som opnås fra simuleringernes eksplicitte sekvensinformation ved at tælle antallet af synonyme og ikke-synonyme mutationer, der er blevet fikseret i populationen. Vi viser, at denne overensstemmelse er stærkest for høj stabilitet, hvor proteinerne for det meste udvikler sig neutralt. I det ustabile regime, hvor proteinevolution er domineret af selektion af stabiliserende mutationer, registrerer vi en svag, men signifikant positiv selektion for specifikke aminosyrer i sekvenserne af størrelsesordenen $dN / dS \sim 1.5$.

Samlet giver denne afhandling et af de første eksempler på udviklingen af funktion og termodynamisk stabilitet i et pattedyrprotein med direkte implikationer for fitness. Vi kvantificerer den biologiske relevans for udvælgelsen af en højere koncentration af myoglobin i havpattedyrs muskler og giver en forklaring på stigningen i proteinernes stabilitet. Desuden giver denne afhandling en række andre, teoretiske resultater der direkte er testbare i fremtidige eksperimenter vedrørende funktion, fysiologi og termodynamisk stabilitet af pattedyrs myoglobiner.

# List of Publications

**Papers included in the thesis**[1]

I. Dasmeh P, Kepp KP (2012) Bridging the gap between chemistry, physiology, and evolution: quantifying the functionality of sperm whale myoglobin mutants. *Compar Biochem Physiol Part A* 161: 9–17.

II. Dasmeh P, Davis RW, Kepp KP (2013) Aerobic Dive Limits of Seals with Mutant Myoglobin Using Combined Thermochemical and Physiological Data. *Compar Biochem Physiol Part A* 164: 119–128.

III. Dasmeh P, Serohijos AWR, Kepp KP, Shakhnovich EI (2013) Positively selected sites in cetacean myoglobins contribute to protein stability. *PLoS Comput Biol* 9(3): e1002929.

IV. Dasmeh P, Serohijos AWR, Kepp KP and Shakhnovich EI. Influence of protein biophysics on inferring molecular clock rates in phylogenetic trees, Submitted to *Molecular Biology and Evolution*.

**Papers not included in the thesis**

V. Kepp KP, Dasmeh P (2013) The effect of distal interactions on $O_2$-binding to Heme. *J Phys Chem B* 117: 3755–3770.

VI. Kepp KP, Dasmeh P. Convergent evolution of whale and seal myoglobins toward higher folding stabilities. *In preparation*.

VII. Dasmeh P, Serohijos AWR, Kepp KP, Shakhnovich EI. Selection for thermodynamic stability produces among-site-rate-variation. *In preparation*.

[1] Materials presented in the first two papers are reused and reprinted in this thesis by permission from Elsevier with license IDs 3137511372302 and 3137511374293. Reuse of the materials in the third paper (PLoS computational biology) is permitted for the authors.

CONTENTS

**Chapter 1: Introduction** ..... **1**
1.1. Introduction ..... 1
1.2. Myoglobin: structure, function, and evolution ..... 3
1.3. Project aims and outlines ..... 7
**Chapter 2: Theoretical methods** ..... **10**
2.1. Overview ..... 10
2.2. Saturation expression ..... 10
2.3. Modeling oxygen delivery in muscle tissues ..... 11
2.4. Quantifying functional proficiency of Mb: oxygen storage and transport ..... 13
2.5. Average Mb saturation in muscle tissue ..... 15
2.6. Physiological model to quantify aerobic dive limit (ADL) ..... 15
2.7. Phylogenetic analysis and ancestral state reconstruction ..... 19
2.8. Estimating evolution rate and detecting adaptive evolution ..... 21
2.9. Estimating effects of point mutations on folding stability (FoldX) ..... 22
2.11. Estimating the effect of point mutations on protein folding stability (ERIS) ..... 24
2.12. Protein evolution model and simulated phylogenies ..... 25
**Chapter 3: Parameters used in the models** ..... **26**
3.1. Overview ..... 26
3.2. Parameters for modeling of muscle tissue ..... 26
3.3. Parameters for modeling of ADL ..... 29
**Chapter 4: Summary of research articles** ..... **31**
4.1. The effect of Mb mutations on cellular and organismal fitness ..... 31
4.1.1. Mutational effects on $O_2$-delivery process ..... 31
4.1.2. Mutational effects on aerobic dive limit (ADL) ..... 39
4.2. Selection for thermodynamic stability ..... 43
4.2.1. Positive selection of folding stability in cetacean Mbs ..... 43
4.3. Influences of selection for thermodynamic stability on evolution rate ..... 53
**Chapter 5: Concluding remarks** ..... **59**
**Chapter 6: Falsifiable predictions of this thesis** ..... **60**
**Appendix A: Dependence of Mb functional proficiencies on relevant parameters** ..... **61**
**Appendix B: Supplementary data for ADL calculations** ..... **67**
**Appendix C: Supplementary data for the study of positive selection in cetacean Mbs** ..... **89**

**Appendix D: Supplementary data for simulated evolution of Mb sequences ..... 105**
**Bibliography ..... 114**
**Publications ..... 136**

"If one asks in what the great achievement of Christopher Columbus when discovering America really consisted, one will have to answer that it was not the idea to make use of the world's spherical shape to travel to India on the Western route; this idea had already been considered by others. Neither was it the expert preparation of his journey or the professional equipment of the ships – this could also have been done by others. The most difficult aspect of this journey certainly was the decision to leave his well-known land behind and to sail so far to the West that a return would not be possible with the provisions available. In a similar way, really new grounds can be found in science only if at a decisive point and time one is prepared to leave the ground on which science has so far been based."

*Werner Heisenberg*

"The secrets eternal neither you know nor I
And answers to the riddle neither you know nor I
Behind the veil there is much talk about us, why
When the veil falls, neither you remain nor I."

*Omar Khayyam*

# CHAPTER 1: INTRODUCTION

## 1.1. Introduction

In natural science macroscopic phenomena have underlying microscopic causes or mechanisms which are hidden from the bare eyes. The most difficult challenge is to unravel this "mysterious" world to understand and predict macroscopic events. This organization of information (i.e., the lower levels influence the higher ones or vice-versa) is one of the most universal characteristics of all observable, falsifiable and repeatable experiences we have had so far. Biological evolution is by no means an exception. Every single change in living systems is principally tractable to lower-level properties from atomic and molecular interactions to cellular organizations. We are thus aftermath effects of the changes at molecular levels that have been selected in evolution. Molecular evolution takes this axiom and aims to relate natural variations in organisms and living systems all the way to the properties of their molecules (i.e., DNA and proteins).

Molecular evolution deals with the evolution of DNA and proteins and evolutionary history of genes. The main goal is to understand patterns of sequence evolution, underlying mechanisms and finally relate them to higher level properties such as protein, cellular, and organismal fitness. This is achieved by analysis of DNA/protein sequences of extant species as the only observables in molecular evolution and to find their similarities using mathematical and statistical methods (i.e., molecular phylogenetics). The main idea behind this approach is very simple and elegant. As proposed in 1962 by Zuckerkandl and Pauling (Zuckerkandl and Pauling, 1962) and subsequently by Margoliash in 1963 (Margoliash, 1963), the number of amino acid differences between two orthologous proteins is apparently proportional to the elapsed time since the organisms carrying them diverged from a common ancestor. In this sense, protein evolution

exhibits a clock-like behavior, ticking when a single amino acid is fixed in the evolving population (Zuckerkandl and Pauling, 1965). This observation, indeed, initiated one of the most provocative ideas in the history of molecular evolution and brought this scientific field into the center of attraction in evolutionary biology.

Once the evolutionary histories of genes are inferred, we can look into the properties of ancestral proteins and investigate how they evolved to their present functions. In this way, molecular records for all proteins from modern species are combined with other information such as geological, paleontological and cosmological records to create a coherent picture of life evolution on earth (Benner et al., 2002). One of the exciting examples of such efforts is the study of Gaucher et al., in correlating the temperature histories of early life from molecular and geological records (Gaucher et al., 2003). By resurrecting candidate sequences for elongation factors of the Tu family (EF-Tu) found at ancient nodes in the bacterial evolutionary tree and measuring their activities, they showed that ancient EF-Tu proteins have temperature optima of 55–65°C. This groundbreaking observation suggests that the ancient bacteria that hosted these particular genes were thermophiles, and neither hyperthermophiles nor mesophiles (Gaucher et al., 2003). A quick survey of literature shows many examples of these studies which I like to call “the full elephant” works[2]. This thesis aims to provide the same picture for the evolution of mammalian Mbs.

[2] It is an old eastern tale that a group of blind men (or men in the dark) touch an elephant to learn what it is like. Each one feels a different part, but only one part, such as the side or the tusk. Finally they start listening and collaborate to "see" the full elephant.

## 1.2. Myoglobin: structure, function, and evolution

Myoglobin (Mb) has a unique role in the history of molecular biology as it is the first protein whose three-dimensional structure was resolved at atomic resolution (Kendrew *et al*., 1958). This discovery initiated the challenging quest for understanding structure and function of proteins as was foreseen by Kendrew in 1962:

"The detailed structures of a few other proteins should soon become known, but it will be clear from many of the topics I have touched upon that we have pressing need to know the structures of very many others, for proteins are unique in combining great diversity of function and complexity of structure with a relative simplicity and uniformity of chemical composition. In determining the structures of only two proteins we have reached, not an end, but a beginning; we have merely sighted the shore of a vast continent, waiting to be explored."

John Kendrew, Nobel Lecture 1962 (Kendrew, 1964)

Mb is indeed one of the most studied proteins often referred to as the 'hydrogen atom of biology' (Frauenfelder et al., 2003). It is a small monomeric protein of ~16 kDa with 145 to 153 residues. Mb is mainly α-helical with eight α-helices named from A to H (Figure 1.1) and binds to small diatomic gases such as $O_2$, CO and NO with a 1:1 stoichiometry. Ligands bind to the Fe atom of the heme group which is placed between E and F helices. The distal and proximal histidines are essential for $O_2$ binding to Mb. The four coordination sites of Fe are filled with electrons of nitrogen atoms from the porphyrin ring, the fifth one with proximal histidine and the sixth site is free for incoming ligands.

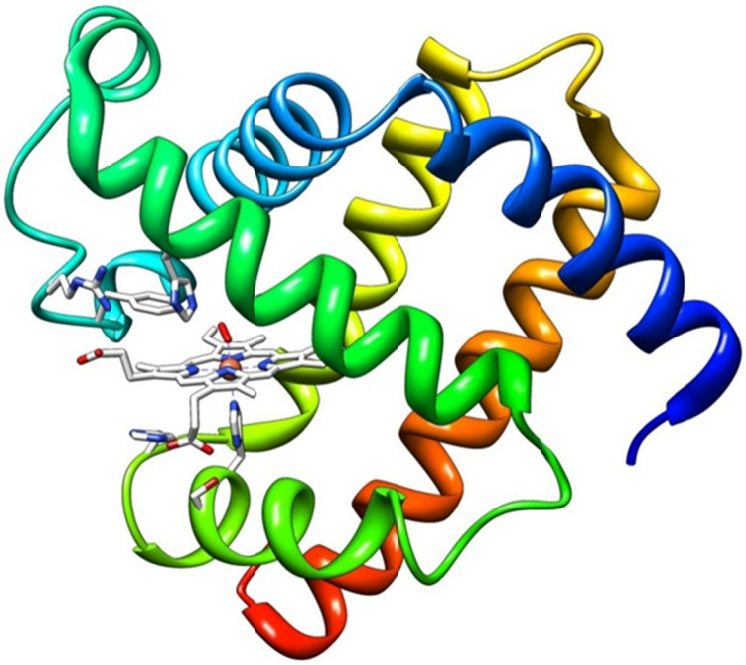


**Figure 1.1**. Structure of oxy-Mb showing the heme and proximal and distal histidines. The image was produced using *Chimera*, University of California, San Francisco (Pettersen *et al*., 2004) (Dasmeh et al., 2013a).

The side-chain of the distal histidine at the E7 helical position was originally thought to have an important role in $O_2$ binding to Mb (Perutz and Mattews, 1966). However, ligands can also escape through the interior of the protein (Elber, 2010). $O_2$ binding to Mb not only involves the primary binding site of this protein but also secondary and tertiary states in the interior of the protein (Tilton *et al*., 1984). As shown in Figure 1.2a, ligand binding to Mb follows a multi-state scheme (Ostermann *et al*., 2000). In the figure, B states are a number of ligand positions in the distal 'pocket' with differing rates of binding to the heme group. Ligands can also relocate between multiple states such as Xe1 and Xe2 cavities within the protein matrix (Scott *et al*., 2001). Overall, the majority of ligands (~70-80%) enters and exits from the distal histidine gate (Scott *et al*., 2001; Elber, 2010) and thus, the kinetics is adequately interpreted in terms of an effective two-step scheme (Figure 1.2b).

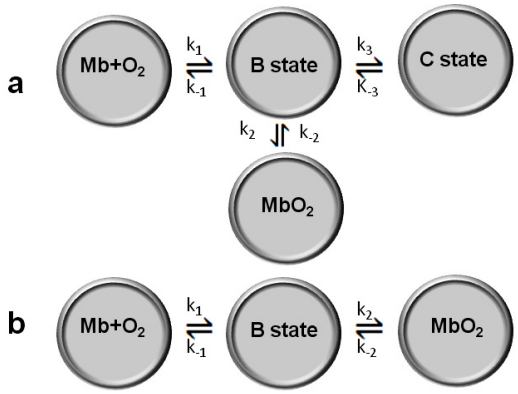



**Figure 1.2**. Kinetics of $O_2$ binding to Mb. a) The C state accounts for internal cavities while in b) mainly the B state is involved under physiological conditions (Dasmeh et al., 2012).

Mb is present in both cardiac and skeletal muscles of vertebrates and in the body walls of invertebrates (Suzuki and Imai, 1998). The primary function of Mb is to increase the availability of $O_2$ in sarcoplasm and thus provides the oxygen flow for oxidative phosphorylation in mitochondria (Wittenberg, 1970). Traditionally, oxygen storage has been considered as the main function of Mb. The main support for this consideration comes from the elevated concentrations of Mb in diving mammals (Guyton *et al*., 1995; Ponganis *et al*., 2002) and also a higher expression level of Mb at higher altitudes (Gimenez *et al*., 1977; Terrados *et al*., 1990). However, it is also proposed that Mb can enhance the $O_2$ diffusive flux by active transport. Although the diffusion coefficient of $O_2$ in muscle cells (i.e., $1.16\times10^{-9}$ $m^2s^{-1}$) is ~140 times higher than that of Mb (i.e., $7.85\times10^{-12}$ $m^2s^{-1}$) (Groebe, 1995; Lin *et al*., 2007b), the high concentration of Mb in heart and skeletal muscle cells provides an advantage in transporting $O_2$ in the cell (Wittenberg and Wittenberg, 2003). *In vitro* studies have indicated that $O_2$ diffuses faster in Mb solution than in Mb-free solution (Wittenberg and Wittenberg, 1989), and Mb shows sufficient mobility to compete with free oxygen. However, *in vivo*, the role of Mb in $O_2$ diffusion was controversial for decades (Gros *et al*., 2010).

Over the last years, protein mutagenesis experiments provided a great insight into the pathways and kinetics of ligand binding to Mb (Varadarajan *et al*., 1985; Springer and Sligar, 1987). To map the pathways for $O_2$ entry and exit in Mb, Olson and co-workers have studies $O_2$ binding parameters of 90 sperm whale Mb mutants at 27 different positions (Scott *et al*., 2003). These investigations revealed that His E7 stabilizes bound $O_2$ about 1000-fold, most likely because of the formation of a strong hydrogen bond between O and $N_\varepsilon H$ of the imidazole group (Olson, 2008).

The evolution of Mb can be understood by an early heme evolution followed by the consequent evolution of apo-Mb. The molecular evolution of the heme group is justified through the selection of the porphyrin structure to satisfy the reversible spin-crossover upon $O_2$ binding to porphyrin, which is a necessary first condition for reversible binding of molecular $O_2$ to Mb (Jensen and Ryde, 2003). For apoMb, evolution can be traced back ~4000 Myr to the common ancestor of globin superfamily (Suzoki and Imai, 1998; Wajcman *et al*., 2009) as one of the basic protein required for life. In mammalian Mbs, residues important for $O_2$ binding (i.e., residues 29, 43, 64 and 68 and the proximal His93) are highly conserved. In a sequence alignment of Sperm whale, Pig, Bovine, Dog, Sheep, Horse and Human Mb, 107 out of 153 residues are identical (see Appendix C, part 2). As a result, the oxygen affinity of mammalian Mbs is nearly unchanged ($K_{O_2} \approx$ 0.8-1.2 $\mu M^{-1}$ at pH 7.0, 20 °C) (Scott et al., 2000).

Mb has a critical role in the evolution of marine mammals. Increased abundance of Mb in skeletal muscle cells is one of the major adaptations of marine mammals to aquatic environment since divergence ~50 Myr ago from their terrestrial ancestors (Kooyman and Ponganis, 1998). In the skeletal muscle cells of deep-diver cetacean and pinniped species, Mb is ~10-20 more abundant than terrestrial mammals (Kooyman and Ponganis, 1998).

### 1.3. Project aims and outlines

The works presented in this thesis can be divided into four parts. In the first part which focuses on $O_2$-binding in Mb, we merge a large set of previously reported thermochemical data for Mb mutants with a physiological model of $O_2$-transport and –storage in the skeletal muscle cells. We quantify the functional proficiency of Mb mutants under various physiological conditions (i.e., $O_2$-consumption rate resembling workload, $O_2$ partial pressure corresponding to hypoxic stress, muscle cell size, and Mb concentration, resembling different organism-specific and compensatory variables) and show that $O_2$-storage and –transport are distinct functions that rank mutants and wild type (WT) differently depending on $O_2$ partial pressure (Dasmeh and Kepp, 2012). We conclude that conserved residues in WT Mb were most likely fixated under a selection pressure of low $P_{O_2}$ (Dasmeh and Kepp, 2012).

In the second part, we present an integrated model of convective $O_2$-transport and $O_2$-affinity of mutant Mb to quantify the impacts of mutations in Mb on the aerobic dive limits (ADL) of Weddell seals (*Leptonychotes weddellii*) (Dasmeh *et al*., 2013a). This integration allows us to show the superiority of WT Mb traits under specific physiological conditions that prolong the dive time, action radius, and thus fitness of the seals. As an extreme example, mutations that destroy the hydrogen bond between His64 and bound $O_2$ reduce ADL up to 14 ± 2 min for routine aerobic dives, whereas many other mutations are nearly neutral in terms of ADL and the inferred fitness. We also find that the cardiac system, the muscle $O_2$-store, animal behavior (i.e., pre-dive ventilation), and the oxygen binding affinity of Mb, $K_{O_2}$, have co-evolved to optimize dive duration at routine aerobic diving conditions, suggesting that such conditions are mostly selected upon in seals.

In the third part, this thesis deals with the observation of higher folding stabilities (i.e., $\Delta G_{\text{folding}}$) of cetacean Mbs compared to their terrestrial counterparts. We employed ancestral

sequence reconstruction, maximum likelihood and Bayesian tests and observed accelerated evolution in cetaceans and identified seven positively selected sites in Mb. Using the experimentally calibrated computation of stability effects of mutations (i.e., $\Delta\Delta G_{\text{folding}}$), we show that these sites contribute to Mb stabilization favoring hydrophobic folding, structural integrity, and intra-helical hydrogen bonds. Moreover, we observe a correlation between Mb folding stability and protein abundance, which suggests a correspondence between the selection pressure for stability and higher expression levels in muscle cells. We explained this observation by the universal selection pressure against misfolding or misfolding prevention hypothesis (Drummond and Wilke 2008; Dasmeh *et al*., 2013b).

In the fourth part, we deal with the fundamental question of how a general protein phenotype such as folding stability, affects the rate of protein evolution. Using a model that combines explicit evolution of Mb sequences, folding stability, and application of maximum likelihood (ML) estimation of evolution rate (ER), we find that ER predicted by ML methods is highly correlated with ER from simulations using the explicit sequence information. We show that this agreement is strongest in the regime of high stability where proteins are mostly evolving neutrally. In the unstable regime where protein evolution is dominated by selection for stabilizing mutations we detect a weak yet significant positive selection for specific residues in the sequences from simulations of the order of $dN/dS \sim 1.5$.

This thesis is divided into six chapters and four appendices. Following this Introduction, in chapter two, we present the theoretical methods employed in this thesis. Chapter three includes details of the parameters used in the modeling. I explain in more depth a summary of research papers published in this project in chapter four. Chapter five presents a brief conclusion of the works presented here and finally in chapter six, I put an emphasis on the two main predictions of my thesis that are easy to falsify by future experiments. Given the different nature of methods

presented in this thesis (from physiological modeling to protein biophysics and phylogenetics) I propose the following structure for connecting different parts with respect to their biological relevance:

**Table 1.1.** Connection between different parts of this thesis with respect to the original biological question.

| **Biological question** | **Theory and parameters** | **Results and discussion** |
|---|---|---|
| How do mutations in Mb influence $O_2$-delivery process in muscle cells? | Sections 2.2, 2.3, 2.4 and 3.1. | Section 4.1.1 and Paper I. |
| How do mutations in Mb influence aerobic dive limit (ADL) in marine mammals? | Sections 2.6 and 3.2. | Section 4.1.2 and Paper II. |
| What is a plausible explanation for the increased stability of cetacean Mbs compared to their terrestrial counterparts? | Sections 2.7, 2.8 and 2.9. | Section 4.2.1 and Paper III |
| How does selection for thermodynamic stability influence evolution rate (by means of *dN/dS*)? | Sections 2.8, 2.10 and 2.11. | Section 4.2.2 and Paper IV. |

# CHAPTER 2: THEORETICAL METHODS

## 2.1. Overview

In this chapter, I explain the theoretical basis of my thesis as follows: In sections 2.2 to 2.5, mathematical details for integrating thermochemistry of $O_2$-binding to Mb with physiological model of muscle tissues are presented. Subsequently, in section 2.6, details for modeling of aerobic dive limits (ADL) in marine mammals are presented. By the end of this section, the reader can get a general picture of Mb function in muscle tissues and its relation to diving in marine mammals.

In sections 2.7 and 2.8, I present the details of phylogenetic and ancestral sequence reconstruction given the sequences of mammalian Mbs and the way we can detect "positive selection" in the evolution of Mb. Once specific mutations of interest are specified, the reader can employ the method presented in section 2.9 to estimate the stability effects of these mutations. Finally, in sections 2.10 to 2.12, I explain how a large scale simulation of Mb sequences is feasible by imposing selection pressure for thermodynamic stability.

## 2.2. Saturation expression

The saturation of Mb by $O_2$ ($S$), is expressed in terms of the Hill equation (Hill, 1936):

$$S = \frac{(P_{O_2})^n}{(P_{50})^n + (P_{O_2})^n} \quad (1)$$

where $P_{O_2}$ is the oxygen partial pressure, $n$ is the oxygen binding cooperativity index, and $P_{50}$ is the value of $P_{O_2}$ at which $S = 0.5$. For Mb ($n$ = 1), saturation can be described using the bimolecular oxygenation equilibrium:

$$Mb + O_2 \rightleftharpoons MbO_2 \quad (2)$$

$$S = \frac{[MbO_2]}{C_{Mb}} = \frac{K_{O_2}[O_2]}{K_{O_2}[O_2]+1} \quad (3)$$

Here, saturation is defined as $S = \frac{[MbO_2]}{C_{Mb}}$, $K_{O_2} = \frac{[MbO_2]}{[O_2][Mb]}$ is the bimolecular oxygenation equilibrium constant, and $C_{Mb}$ is the total concentration of Mb in the cell. Equation (3) can be converted to (1) using $P_{50} = \frac{1}{K_{O_2}\alpha_{O_2}}$ and $[O_2] = \propto_{O_2} P_{O_2}$, where $\propto_{O_2}$ is the oxygen solubility constant in the sarcoplasm. In the general case of two-step binding (Figure 1.2b), using the equilibrium constants $K_1$ and $K_2$, $S$ takes the form:

$$S = \frac{K_1K_2[O_2]}{K_1K_2[O_2]+K_1[O_2]+1} = \frac{K_{O_2}\propto_{O_2}P_{O_2}}{\propto_{O_2}P_{O_2}(K_{O_2}+K_1)+1} \quad (4)$$

**2.3. Modeling oxygen delivery in muscle tissues**

Oxygen delivery process within the muscle cell can be modeled by using the Krogh cylinder model in a revised, state-of-the-art form (Groebe, 1995). As shown in Figure 2.1, the three concentric cylinders represent *capillaries* (inner region), space between red blood cells and sarcolemma, or *carrier free regions* (middle), and the *muscle tissue* as the outer region. We consider only radial oxygen diffusion and assume chemical equilibrium at all times, since both the radial and longitudinal diffusion coefficients of Mb are similar within our scope (Groebe, 1995; Lin *et al*., 2007b). The oxygen gradient from red blood cells (RBC) to mitochondria is calculated from two well-known partial differential equations:

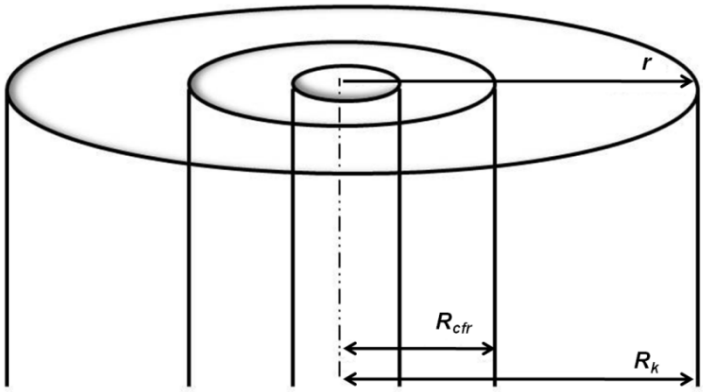



**Figure 2.1**. A section of the Krogh model is shown, composed of three concentric cylinders representing capillary (inner), carrier free region (middle with radius $R_{cfr}$) and the tissue (outer with radius $R_k$). The radial coordinate ($r$) is the generalized coordinate (Dasmeh et al., 2012).

$$D_M \propto_{O_2} \nabla^2 P_{O_2} + D_{Mb} C_{Mb} \nabla^2 S = \dot{V}_{O2} \quad \text{In the muscle cells} \quad (5)$$

$$\nabla^2 P_{O_2} = 0 \quad \text{In the carrier free region} \quad (6)$$

where $S$ is the saturation defined in equation (4), $D_M$ ($\mathrm{cm^2 s^{-1}}$) and $D_{Mb}$ ($\mathrm{cm^2 s^{-1}}$) are the free oxygen and Mb diffusion coefficients in the cell, $C_{Mb}$ ($\mathrm{mol\ L^{-1}}$) is the total concentration of Mb in muscle tissue and $\dot{V}_{O2}$($\mathrm{mol\ L^{-1}\ s^{-1}}$) is the oxygen consumption rate. Transforming equations 5 and 6 into polar coordinates and using appropriate boundary conditions on the continuity of $O_2$ flows at muscle tissue the equations are solved to give:

$$P^*(r) = P^*_{CFR} + \frac{\dot{V}_{O_2}}{2D_M \propto_{O_2}} \left( \frac{1}{2}(r^2 - R^2_{CFR}) - R^2_K ln \frac{r}{R_{CFR}} \right) \quad (7)$$

When r is the radial distance to the capillary here treated as the generalized coordinate which all of integrations are based , $P^*_{CFR}$ is the effective pressure at carrier free region, $P^*(r)$ is the effective pressure inside the tissue and $R_{CFR}$ and $R_K$ are the positions of the carrier free region (i.e., capillary walls) and Krogh cylinder (i.e., the end point). Since the $P_{O_2}$ gradient is much steeper along the

radial coordinate (Groebe, 1995), we neglected the angular terms in solving the equation. An effective pressure (a mathematical transformation of $P_{O_2}$) is defined as:

$$P^*(r) = P_{O_2}(r) + \frac{D_{Mb}C_{Mb}}{D_{O_2}\propto_{O_2}} S\left(P_{O_2}\right) = P_{O_2} + \frac{D_{Mb}C_{Mb}}{D_M\propto_{O_2}} \frac{P_{O_2}}{P_{O_2}+P_{50}} \quad (8)$$

To evaluate $P_{O_2}(r)$ it is necessary to know $\dot{V}_{O2}$, $R_{CFR}$,$R_K$, and $P^*_{CFR}$. The first three are experimentally determined while the last one ($P^*_{CFR}$) should be determined theoretically using the blood flow rate and the hemoglobin (Hb) oxygen binding properties. Since we are interested in the role of Mb in oxygen delivery under various stress conditions such as low $P_{O_2}$, we treat $P^*_{CFR}$ as an external variable which determine the upper oxygen pressure at the sarcolemma ($P_u$). $P_u$ then declines to a lower pressure ($P_l$) through the radial coordinates ($r$). It is also important to note that upper pressure can be affected due to many factors such as changes in blood flow rate, capillary density, Hb oxygen affinity and thus we have simplified all of the mentioned effects into one variable, $P_u$.

### 2.4. Quantifying functional proficiency of Mb: oxygen storage and transport

Using $P_{O_2}$ profile in muscle tissues, we can define Mb functional proficiencies in terms of $O_2$ storage and transport and to investigate how proficient is Mb compared to free diffusing $O_2$ molecules in muscle cells.

The proficiency of Mb as an oxygen storage protein can be quantified by integrating the concentration of $MbO_2$ over the radial distance d$r$ (i.e., the distance between sarcolemma and mitochondria in Figure 2.1) and comparing to the same integral over [$O_2$]. We call this the *oxygen storage ratio* (*OSR*):

$$OSR \equiv \frac{\int [MbO_2](r)dr}{\int [O_2](r)dr} \quad (9)$$

Using equation (4), the definition of saturation as $S = \frac{[MbO_2]}{C_{Mb}}$ and $[O_2] = \alpha_{O_2} P_{O_2}$, OSR takes the following form:

$$OSR \equiv \frac{C_{Mb} \int S(P_{O_2}(r))dr}{\propto_{O_2} \int P_{O_2}(r)dr} \quad (10)$$

Functional proficiency of Mb as a transport protein can also be quantified similar to OSR. We partitioned the overall $O_2$ flux from sarcolemma to mitochondria into contributions from $MbO_2$ and free $O_2$:

$$j_{O_2} = j_{O_2}^{Mb} + j_{O_2}^{O_2} = D_{Mb} C_{Mb} \frac{dS(P_{O_2}(r))}{dr} + D_{O_2} \propto_{O_2} \frac{dP_{O_2}(r)}{dr} \quad (11)$$

where $j_{O_2}$ is the flux density of total $O_2$, i.e., the amount of $O_2$ passing through a unit area of the muscle cell per time unit, and $j_{O_2}^{Mb}$ and $j_{O_2}^{O_2}$ are flux densities of Mb and free $O_2$, respectively. Integrating (11) over the radial distance, the ratio of the total Mb-facilitated oxygen flux $J_{O_2}^{Mb}$, to the free $O_2$ flux $J_{O_2}^{O_2}$ is defined as the oxygen transport ratio (OTR) :

$$OTR = \frac{J_{O_2}^{Mb}}{J_{O_2}^{O_2}} = \frac{D_{Mb} \int \frac{\partial [MbO_2]}{\partial r} dr}{D_{O_2} \int \frac{\partial [O_2]}{\partial r} dr} = \frac{D_{Mb} C_{Mb} \int \frac{\partial S}{\partial r} dr}{D_{O_2} \propto_{O_2} \int \frac{\partial P_{O_2}(r)}{\partial r} dr} = \frac{D_{Mb} C_{Mb} (S_u - S_l)}{D_{O_2} \propto_{O_2} (P_u - P_l)} \quad (12)$$

Here, $S_u$ and $S_l$ are the values of saturation at the limiting values of oxygen pressure within the cell. Knowing $P_{O_2}$ over the radial distance, we calculate the exact value of $P_l$ and thus OTR from equation (12) to consider the effect of $P_{O_2}$ gradient on the transport efficiency of the protein.

### 2.5. Average Mb saturation in muscle tissue

The average saturation of Mb within the muscle cell can be simplified as:

$$\langle S \rangle = \frac{1}{P_c - P_{mit}} \int_{P_{mit}}^{P_c} S\,dP = \frac{1}{P_c - P_{mit}} \int_{P_{mit}}^{P_c} \frac{K_{O_2}\alpha_{O_2}P_{O_2}}{K_{O_2}\alpha_{O_2}P_{O_2}+1}\,dP \tag{13}$$

where $P_c$ and $P_{mit}$ are the partial pressures at the capillary and mitochondria, respectively. Assuming $P_{mit} \sim 0$, equation (13) gives:

$$\langle S \rangle = \frac{1}{P_c} \int_{0}^{P_c} \frac{K_{O_2}\alpha_{O_2}P_{O_2}}{K_{O_2}\alpha_{O_2}P_{O_2}+1}dP = 1 - \frac{\ln(K_{O_2}\alpha_{O_2}P_c+1)}{K_{O_2}\alpha_{O_2}P_c} \tag{14}$$

### 2.6. Physiological model to quantify aerobic dive limit (ADL)

Aerobic dive limits (ADL) as the amount of available $O_2$ for diving in marine mammals can be calculated using an iterative model. In each iteration, a specific volume of blood is passed through different organs and $O_2$ is extracted by knowing the specific oxygen consumption of each organ in the body of Weddell seal (see chapter three for parameters used in this model). Iterations continue until specific termination protocols are met (as is explained below).

During numerical computation, each iteration corresponded to one heartbeat. During this period, convective oxygen transport ($t \times \dot{V}_b \times C_{VO2}$) from the venous blood pool to the arterial pool was calculated, where $t$ (min) is the time of one heartbeat, $\dot{V}_b$ (l min$^{-1}$) is the cardiac output, and $C_{VO2}$ (ml $O_2$ l blood$^{-1}$) is the venous blood oxygen content. Convective transport of oxygen through the main organs and tissues ($t \times \dot{Q}_i \times C_{aO2}$ ml $O_2$) and the amount of oxygen extracted ($t \times \dot{V}_i$ ml $O_2$) were calculated. Here, $\dot{Q}_i$ (l min$^{-1}$), $\dot{V}_i$ (l min$^{-1}$) and $C_{aO2}$ (ml $O_2$ l blood$^{-1}$) are the blood flow rate and

oxygen consumption rate of each organ and tissue and the arterial blood oxygen content (Table 3.1). The extraction coefficient ($E_b$) of oxygen from the blood could not exceed 0.8 (i.e., maximum $E_b$ at critical oxygen delivery) during a single pass of the blood through an organ (Samsel and Schumacker, 1994; Torrance and Wittnich, 1994; Nelson *et al*., 1988). Blood flow decreased proportionately to $\dot{V}_b$ during a dive except for the brain where circulation is maintained at pre-dive levels (Elsner *et al*., 1964; Blix *et al*., 1976).

Only oxygen stored in the blood and skeletal muscle were used to calculate whole body $O_2$ stores. $O_2$ stored as oxy-myoglobin in the heart was neglected as it constitutes less than 2% of the total muscle mass. Due to the complete functional pulmonary shunt in Weddell seal above 3-5 atmospheres (~300-500 kPa), the lung oxygen is not available during the dive (Falke et al., 1985; Reed et al., 1994). A value of 96 liters was used for the blood volume of a standard 450-kg Weddell seal (Kooyman et al., 1980) with 33% and 67% contribution of arterial and venous blood (Rowell 1986; Hurford et al., 1996). The blood Hb concentration was 260 g/l and the oxygen-binding capacity of Hb was 1.34 ml $O_2$ per gram Hb (Kooyman et al., 1980; Qvist et al., 1986). Each liter of blood thus contained 348 ml $O_2$ ml $O_2$ (260 g / l blood × 1.34 ml $O_2$ pr. gram ml $O_2$ / g Hb).

The arterial blood was assumed to be 100% saturated with oxygen at the onset of diving because of pre-dive hyperventilation (Kooyman et al., 1980; Hurford et al., 1996). Venous blood was considered to be 86% saturated at the beginning of a dive with an initial $C_{aO2}$ of 348 ml $O_2$ per liter blood (Kooyman et al., 1980; Ponganis et al., 1993). Arterial and venous muscle oxygen stores were calculated according to Davis et al., (Ponganis et al., 1993), assuming that 35% of the seal's body mass was skeletal muscle with $C_{Mb}$ = 54 g per kg muscle (Kooyman et al., 1980), and with an oxygen-binding capacity of 1.34 ml $O_2$ per gram Mb. The amount of oxygen stored in the muscle of an average adult seal was calculated as $450 \times 0.35 \times 54 \times 1.34 \times \langle S \rangle = 11{,}397 \langle S \rangle$, where

$\langle S \rangle$ is the average saturation of Mb, which is calculated by equation (14). This factor depends not only on $P_{O_2}$, but also on the thermodynamic constant of oxygenation ($K_{O_2}$) which differs in various seal mutants.

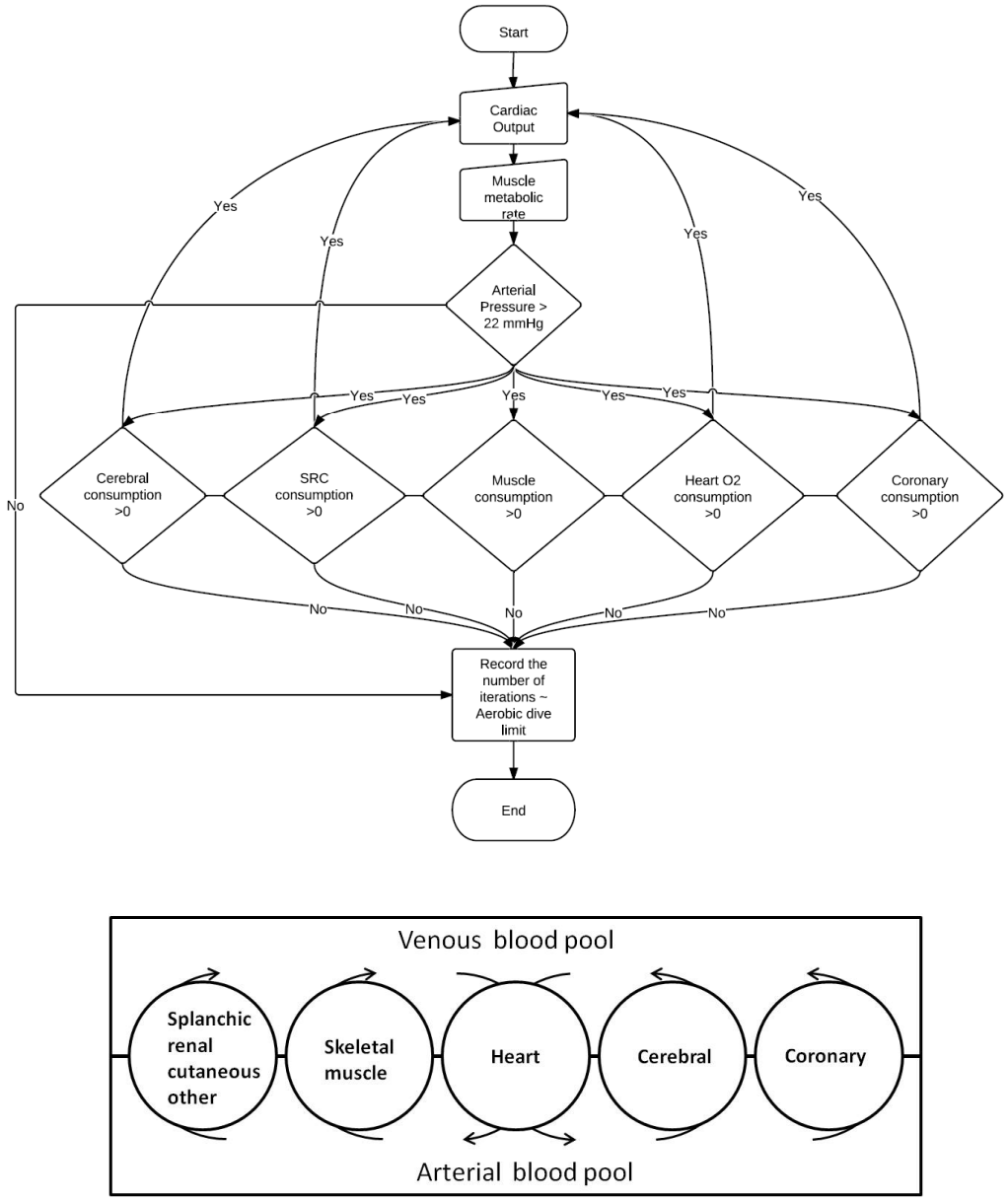



**Figure 2.2.** The algorithm for computing ADL (top) based on the blood circulatory model used in this work simplified into two common venous and arterial (bottom).

A modified model for the ADL of Weddell seals (Davis and Kanatous, 1999) was applied using the circulatory system in Figure 2.2. Based on our previous integrated Krogh model (Dasmeh and Kepp, 2012), which integrates $O_2$ over the actual $P_{O_2}$ profile of the muscle cell to obtain the cell-integrated diffusion and storage as two distinct functions, we noted that diffusion will not affect the long time-average properties such as the total $O_2$ available for diving. Thus, whereas facilitated diffusion is important for short-term response to changes in $P_{O_2}$, for the differential ADLs of Mb mutant animals over longer time-averages (i.e., diving periods), we only need to consider the storage component of our previous model. This also implies that the choice of mitochondrial $P_{O_2}$, while obviously > 0 and dependent on mitochondrial $O_2$ consumption, will not affect the muscle integral of $O_2$-storage for normal dives where the muscles are saturated, and this is where diffusion is most important (Gros et al., 2010; Dasmeh and Kepp, 2012). To quantify this, the difference in average saturation of WT Mb and a His-64 impaired mutant is quite robust to changes in $P_{O_2}$ at the mitochondrial surface, $P_{mit}$, as shown in Appendix B, Figure B24.

Numerical integration procedures for evaluating $\langle S \rangle$ and ADL were performed with MATLAB (Mathworks, vR2010a). As the percentage of cardiac output and muscle metabolic rate were varied in the simulation, the whole body $O_2$ decreased in each iteration. New values for $C_{aO2}$ and $C_{vO2}$ were calculated as $C_{aO2} = 348 \times S_a$ and $C_{VO2} = 300 \times S_v$, where $S_a$ and $S_v$ are the arterial and venous blood saturation, calculated as the ratio of available $O_2$ to the initial total amounts in the arterial and venous blood. We use the reported $P_{50}$ = 26.9 mmHg and Hill coefficient $n$ = 2.39 for an adult Weddell seal to calculate $P_a$ and $P_v$ (Qvist et al., 1986). Within the muscle, the value of $P_{O_2}$ at the mitochondria ($P_{mit}$) was assumed to be zero, and the value of $P_{O_2}$ at the capillaries ($P_c$) was calculated as the average of $P_a$ and $P_v$.

The simulated dive was terminated (i.e., the ADL was reached) when any organ or tissue did not receive sufficient oxygen through convective oxygen transport or $MbO_2$ to maintain aerobic metabolism in any organ or tissue or when the $P_a$ decreased below 22 mmHg (Hurford *et al*., 1996).

To evaluate the effect of oxygen loading at the surface (i.e., forced dives occurring before full $O_2$-reloading due to e.g. sudden threats) on ADL, we evaluated the ADL not only at the spectrum of cardiac and muscle outputs corresponding to variations in circulatory and muscle work, but also at different starting oxygen partial pressures as reflected by changing $P_c$ and re-computing impaired Mb saturations and new ADLs for both mutants and WT.

**2.7. Phylogenetic analysis and ancestral state reconstruction**

To investigate the evolution of mammalian Mbs, nucleotide sequences of mammalian Mbs were used to construct a phylogenetic tree used for evolutionary analyses with codon models (Yang *et al.*, 2000) (Figure 2.3A). To have the highest accuracy in ancestral sequence reconstruction, a larger tree was also constructed from the substantially larger number (82) of available amino acid sequences of mammalian Mbs (Figure 2.3B). For both phylogenies, Zebra finch was the out-group. We divided cetaceans into two major suborders, Mysticeti (minke whale and sei whale) and Odontoceti (sperm whales, beaked whales, dolphins, and porpoises). For the rest of mammals, all the branching patterns followed the known mammalian organism tree with order-specific patterns in primates, rodents, carnivore, cetardiodactylans, and cetaceans (Perelman et al., 2011; Blanga-Kanfi et al., 2009; Bininda-Emonds et al., 1999; Price et al., 2005; Domburg et al., 2011; Hassanin et al., 2012). The accession numbers of all sequences used in this work, as well as full sequences of relevant ancestors are shown in Appendix C. The sequence of ancestral cetacean Mb was inferred from the available mammalian Mb sequences within all orders using the consensus mammalian

species tree. Mb sequences from rodents and primates have minor effects on the most probable inferred ancestral sequence of cetacean Mb (see Appendix C for details).

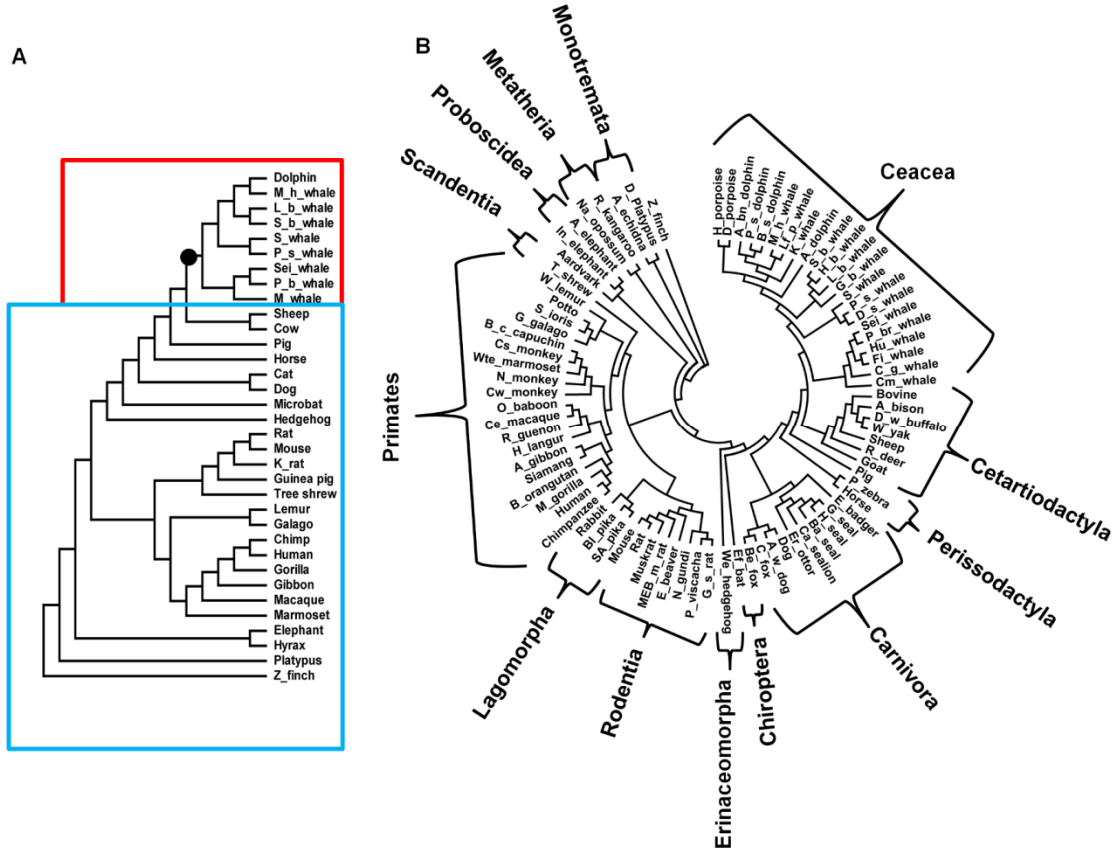



**Figure 2.3.** The mammalian phylogenetic tree constructed from A) nucleotide sequences and B) amino acid sequences. The smaller tree A was used in maximum likelihood tests for adaptive evolution while the tree B was explicitly used for ancestral state reconstruction. The best evolutionary model with the lowest BIC score was Tamura-Nei92 with transition/transversion bias, R = 1.66 in A and Dayhoff in B. Both models allow among-site-rate-variation sampled from a discrete gamma distribution with four categories and shape parameters 0.33 and 0.46 for nucleotide and amino acid sequences respectively. The phylogeny A is divided into two groups of cetaceans (shown in red) and terrestrial mammals (shown in blue) to test the non-uniformity of molecular clock across different lineages and sites. The branch leading to cetaceans is shown with a black circle in Figure 2.3A (Dasmeh et al., 2013b).

The mammalian species tree was analyzed with the MEGA5 package (Tamura et al., 2011) to select the best nucleotide/protein model with the lowest BIC scores, which was the Tamura-Nei92 and Dayhoff model allowing among-site-rate-variation (ASRV) sampled from a discrete gamma distribution with four categories (see Appendix C, part 3 for details) (Tamura et al., 2011; Yang, 1996; Tamura and Nei, 1993). To infer the ancestral sequences of the cetacean clade,

branch lengths were first estimated using the Dayhoff model with ASRV, and the Bayesian posterior probabilities were calculated for each possible ancestral state for each node (Dayhoff et al., 1978).

**2.8. Estimating evolution rate and detecting adaptive evolution**

The pair-wise estimation of *dN/dS* for Mb sequences was done by the Maximum likelihood approach with codon models in CODEML program (Yang, 2007). The equilibrium codon frequencies were estimated from the products of the average observed nucleotide frequencies in the three codon positions (F3X4 model). To detect adaptive evolution, three codon-based models of nucleotide substitutions for the data (Yang, 1997) with the maximum likelihood inference were employed. We first used "branch models" that allow the ω ratio (i.e., *dN*/*dS*) to vary among branches in the phylogeny (Yang, 1997); M0 (one ω ratio for all lineages) and FR (one ω ratio for each branch). Second, we used "site models" that allow the ω ratio to vary among codon sites within the sequence (Yang 1998). Here, we employed five different site models referred to as M1 (nearly neutral), M2 (positive selection), M7 (beta), M8 (beta and ω), and M8fix (M8 with ω fixed at 1) (Yang et al., 2005). Synonymous estimates in both marine and terrestrial mammals were less than 1.5 with the exception of one branch having ω = 1.56, and are thus reliable. We ran the CODEML program several times with different initial values to prevent local optima in the Bayesian identification.

To compare the fit of nested models, classified as null and alternative models, the Likelihood Ratio Tests (LRT) was used (Yang, 1998; Whelan and Goldman, 1999). LRT was applied for models M0 versus FR, and site models M1 versus M2, M7 versus M8, and M8fix versus M8. In cases where the LRT was significant, the Bayes empirical Bayes (BEB) method

implemented for models M2 and M8 was employed to calculate the posterior probabilities for codon classes (Yang et al., 2005).

### 2.9. Estimating effects of point mutations on folding stability (FoldX)

The initial 3D-structures used for calculating the stability of single point mutations were taken from the PDB structures of sperm whale Mb at 1.6 Å (PDB ID: 1MBO) (Phillips 1980) and 1.4 Å resolution (PDB ID: 1U7S) (Kondrashov et al., 2008). These structures were subject to the standard protocol of FoldX (Schymkowitz et al., 2005). FoldX predictions were first validated by predicting $\Delta\Delta G$ values for both PDB structures against a set of experimentally reported Mb mutants (see Appendix C for details). We then finally used the repaired PDB structure at 1.4 Å (Kondrashov et al., 2008) which gave the strongest correlation between calculated and experimental $\Delta\Delta G$s, for computing stabilities within the phylogeny. Individual mutations in the cetacean clade (Figure 4.9) were built using ‘‘Build Model’’ command.

### 2.10. Biophysical model for protein evolution

To investigate the *dN/dS* of a protein evolving under a selection pressure to maintain folding stability, fitness is made proportional to the fraction of folded proteins in the cell defined as $F \propto P_{\mathrm{nat}}$ where $P_{\mathrm{nat}}$ is the probability that a sequence is in the native state at equilibrium given the two-state model for protein unfolding (Privalov and Khechinashvili, 1974; Shakhnovich and Finkelstein, 1989; Goldstein 2011):

$$P_{\mathrm{nat}} = \frac{\exp(-\beta\Delta G)}{1+\exp(-\beta\Delta G)} \qquad (15)$$

Here, $\Delta G$ is the free energy of folding and $\beta=1/RT$. The effect of mutations on folding stability is modeled as:

$$\Delta G_{\mathrm{after}} = \Delta G_{\mathrm{before}} + \Delta\Delta G_{\mathrm{mutation}} \qquad (16)$$

The arising mutation would have a selection coefficient *s* defined as (Goldstein, 2011; Wylie and Shakhnovich, 2011):

$$s = \frac{F_{\text{after}} - F_{\text{before}}}{F_{\text{before}}} \sim e^{\beta \Delta G_{\text{before}}}(1 - e^{\beta \Delta \Delta G_{\text{mutation}}}) \quad (17)$$

which is positive, negative or zero depending on the nature of mutations to be beneficial, deleterious or neutral. For each mutation, the probability of fixation is defined as:

$$P_{\text{fix}} = \frac{1 - \exp(-2s)}{1 - \exp(-2s \times N_{\text{eff}})} \quad (18)$$

where $N_{\text{eff}}$ is the effective population size which is ~$10^4$-$10^5$ for mammals (Lynch and Conery 2003). The effect of all single point mutations on folding stability is assumed to be additive (Fersht et al., 1992).

A population of $10^4$ cells (i.e., $N_{\text{eff}}$) each having Mb gene was evolved in non-overlapping phases (i.e., each generation has Mb sequences with identical $\Delta Gs$) and the folding stability of Mb was updated once a mutation was fixed in the population. A simulated phylogenetic tree was then made by bifurcating the population while keeping $N_{\text{eff}}$ constant after the $\lambda$ arising mutations. The daughter populations were subject to further bifurcations with the same conditions until the limit of bifurcations, here 10 rounds, was reached. We then obtained an exact phylogenetic tree with 1024 external nodes having different branch lengths due to the stochastic nature of fixation of mutations in the population. This tree was analyzed using the ML and Bayesian tests (i.e., M1-M2, M7-M8 and M8-M8fix analysis). Figure 2.4 shows a scheme of the way a bifurcating simulated phylogeny was built from an initial Mb sequence with a specific $\Delta G$.

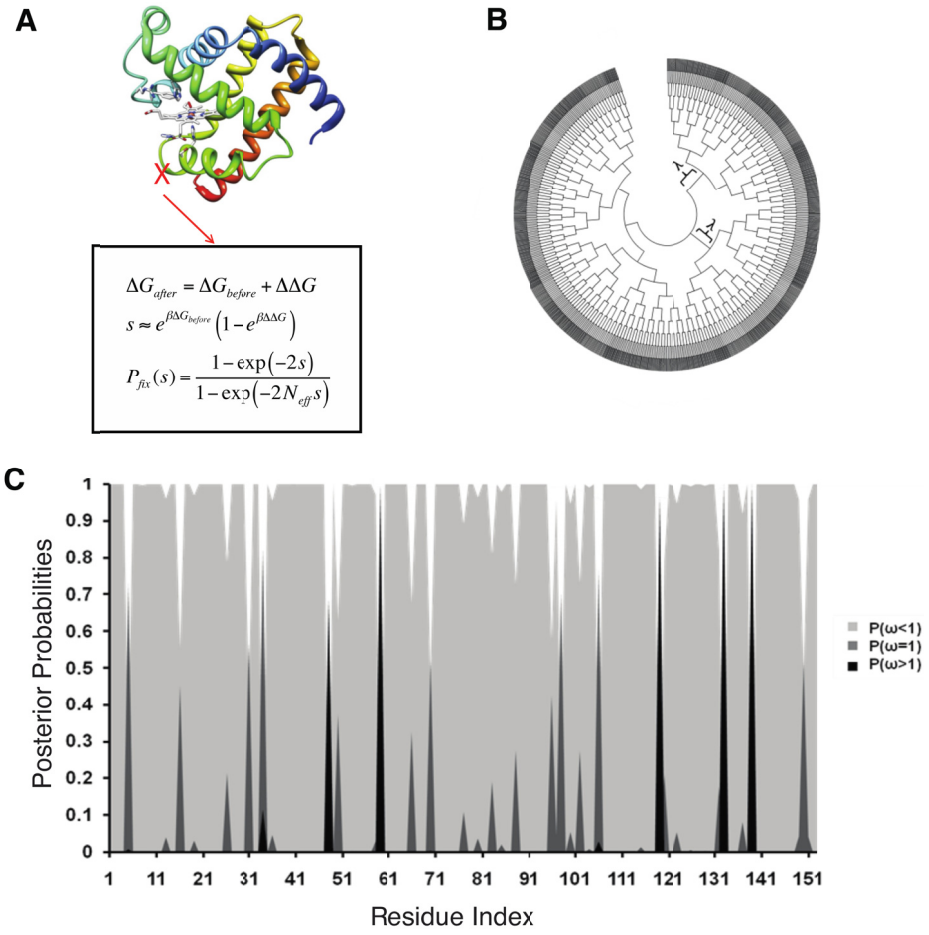



**Figure 2.4.** A) The Mb sequences were evolved in a population of $N_{eff} = 10^4$ cells under monoclonal conditions with selection for folding (equation 17 and 18). B) A bifurcating simulated phylogeny with 1024 external nodes was constructed from an initial Mb sequence with $\Delta G$ = -6.84 kcal/mol. Each bifurcation happens after $\lambda$ arising mutations in the ancestral sequence. C) The resulted phylogenetic tree is then used for further evolutionary rate analysis using the ML and Bayesian methods (Dasmeh et al., 2013c).

## 2.11. Estimating the effect of point mutations on protein folding stability (ERIS)

An initial structure of sperm whale Mb (PDB code=1MBO) (Phillips 1980) was pre-relaxed and subject to mutations by the flexible-back bone method of the ERIS algorithm (Yin et al., 2007a; Yin et al., 2007b). For each mutation, the amino acid rotamers were randomized and minimized using a Monte Carlo simulated annealing approach with Metropolis acceptance criteria based on the change in total free energy. Backbone dihedrals were also allowed to relax to minimize the backbone strains. This procedure was repeated 20 times to take the average of the resulting $\Delta G$ values. In the case of pre-relaxation, the same procedure was repeated 20 times with the wild-type sperm whale

Mb structure and sequence. The $\Delta G$ was calculated for both wild-type and the mutant and $\Delta\Delta G$ was reported as $\Delta G$ (mutants) - $\Delta G$ (WT). To track the effect of mutations on folding stability within simulations, a 154×20 matrix of $\Delta\Delta G$ values was constructed using the described method where each row corresponds to a specific residue in sperm whale Mb and each column to the mutated amino acid (See Appendix D, Table D1). The residues important for heme and $O_2$ binding (i.e., residues 29, 43, 63, 64, 65, 68, 91, 92, and 93) were kept invariable by assigning $\Delta\Delta G$ = 100 kcal/mol for all possible mutations (except a reversion) which gives $P_{\text{fix}} = 0$ in the evolutionary dynamics.

### 2.12. Protein evolution model and simulated phylogenies

The simulated Mb sequence were evolved in a population of $N_{eff} = 10^4$ cells, a relevant to mammalian effective population size (Lynch and Conery, 2003; Charlesworth, 2009; Mesnick et al., 1999) using the nucleotide sequences. As the mutation rate of mammalian globins, $\mu$, is $\sim 10^{-9}$ amino acid per year (Harris and Hey, 1999) and thus $N_{\text{eff}} \times \mu << 1$, the evolution of Mb was assumed to be under monoclonal conditions, i.e., at each generation, population was represented by a single Mb gene. This gene was then let to evolve under the selection for folding stability (i.e., equations 3 and 4) until the fitness reached to a steady value. This fitness-equilibrated sequence which has ~32% identity to sperm whale Mb (see Appendix D, part 2 for details) was then used for making the simulated phylogeny as is shown in Figure 2.4. In this way, we ensured that the protein sequences are sampled near the optimum fitness. The initial population was bifurcated after the specified arising mutations (i.e., the $\lambda$-parameter) defined here as the multiples of population size (e.g. $\lambda = 10N_{\text{eff}} = 10^5$ arising mutations). Bifurcations were then continued to make a simulated phylogeny with 1024 external nodes. For the sake of tractability, more phylogenetic trees were made from the same initial population up to this limit.

# CHAPTER 3: PARAMETERS USED IN THE MODELS

## 3.1. Overview

To model muscle tissue and aerobic dive limit (ADL) we extensively used experimental parameters which are presented in this chapter. Furthermore, the basis of many comparisons and models in this thesis is the available thermodynamic constants of $O_2$-binding to Mb mutants which are tabulated and presented here.

## 3.2. Parameters for modeling of muscle tissue

Table 3.1 shows the parameters used in the calculations to represent cellular conditions of sperm whale Mb. Since no experimental value is reported for $\dot{V}_{O2}$ of sperm-whale muscle cell, we use the metabolic mass adjusted $\dot{V}_{O2}$ due to the available data on the Weddell seal (Davis and Kanatous, 1999):

$$\dot{V}_{O2}(sperm\ whale - organ) = \dot{V}_{O2}(seal - organ) \times \left(\frac{Mass_{sperm\ whale}}{Mass_{whale}}\right)^{-0.25} \qquad (11)$$

Using 35,000 kg as the body weight of sperm whales and 450 kg for a typical Weddell seal (Whitehead, 2002), equation (12) gives $\sim 0.3\ \mu mol\ L^{-1}\ s^{-1}$ for resting $\dot{V}_{O2}$ corrected for extracellular volume (Groebe, 1995). We then apply a 5-fold larger value to mimic diving conditions (Davis and Kanatous, 1999).

To evaluate the functional proficiencies defined in this thesis (OSR and OTR) for Mb mutants, we used the kinetic and thermodynamic oxygenation data for 90 mutants were reported by Olson and coworkers (Scott *et al*., 2003). These mutations occurred in positions both near and far from the heme. OSR and OTR have been calculated at $P_u = 40$ mmHg and $P_u = 5$ mmHg, corresponding to normoxic and hypoxic conditions, respectively. For each mutation, a rank number

is used to describe the approximate distance of the mutated site from the heme group (Table 3.2). This rank is assigned from 1 (wild type), 2 (His64→Gly or H64G) to 90, to include all mutations in the first shell, second shell, proximal site, and far from the heme group, respectively. Table 3.2 also shows the thermodynamic constants of the first equilibrium reaction in $O_2$ binding, $K_1$, the overall oxygenation equilibrium constant $K_{O_2}$ and $P_{50}$ for all mutants.

**Table 3.1**. The list of parameters and variables used to model the physiological conditions of sperm whale Mb.

| Parameter | Symbol | Value |
|---|---|---|
| Average oxygen pressure in cell | $P_{O_2}$ | $P_l < P_{O_2} < P_u$ |
| Oxygen saturation of myoglobin | $S$ | $0 - 1$ |
| $P_{O_2}$ at which S = 0.5 | $P_{50}$ | Typically 1-3 mmHg |
| Upper cellular oxygen pressure (at sarcolemma) | $P_u$ | 5-40 mmHg |
| Lower cellular oxygen pressure (at mitochondria) | $P_l$ | $0 < P_l < P_u$ |
| Myoglobin diffusion coefficient[c] | $D_{Mb}$ | $7.85 \times 10^{-12}$ m$^2$s$^{-1}$ |
| Free oxygen diffusion coefficient[b] | $D_{O_2}$ | $1.16 \times 10^{-9}$ m$^2$s$^{-1}$ |
| Oxygen solubility in the muscle tissue[b] | $\propto_{O_2}$ | $9.4 \times 10^{-7}$mol L$^{-1}$ mmHg$^{-1}$ |
| Mb concentration in sperm whale muscle tissue[a] | $C_{Mb}$ | $3.1 \times 10^{-3}$ mol L$^{-1}$ |
| Concentration of $O_2$-bound myoglobin | $[MbO_2]$ | $\sim S\, C_{Mb}$ |
| Muscle oxygen consumption rate | $\dot{V}_{O2}$ | $0.3 - 1.5 \times 10^{-6}$ mol L$^{-1}$s$^{-1}$ |
| Radius of Krogh cylinder | $R_k$ | 19-100 $\times 10^{-6}$m |
| Radius of cell free region | $R_{cfr}$ | $3.5 \times 10^{-6}$m |

a: Average of two values reported by Scholander, 1940 ; Tawara, 1950. b: Taken from Groebe 1995. c: Lin *et al*., 2007a; Lin *et al*., 2007b.

**Table 3.2**. Thermodynamic constants and half-saturation pressure $P_{50}$ for wild type sperm whale myoglobin and mutants. Each mutated site has a rank, illustrating the approximate distance to heme, classified as either first-shell, second-shell, proximal site, or far-from-heme residue. Data are from (Scott et al., 2003).

| Rank | Position | Mutant | $K_1$ ($\mu M^{-1}$) | $K_{O2}$ ($\mu M^{-1}$) | $P_{50}$ (mmHg)[a] | Rank | Position | Mutant | $K_1$ ($\mu M^{-1}$) | $K_{O2}$ ($\mu M^{-1}$) | $P_{50}$ (mmHg)[a] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | First shell | WT | 5.40E-06 | 1.10 | 0.96 | 46 | Second shell | V66R | 3.71E-06 | 1.08 | 0.98 |
| 2 | | H64G | 2.57E-05 | 0.09 | 12.0 | 47 | | T67F | 3.46E-06 | 1.47 | 0.72 |
| 3 | | H64A | 1.37E-05 | 0.02 | 53.0 | 48 | | T67K | 5.10E-06 | 1.04 | 1.02 |
| 4 | | H64V | 1.79E-05 | 0.01 | 106 | 49 | | T67Q | 3.62E-06 | 0.90 | 1.18 |
| 5 | | H64L | 9.55E-06 | 0.02 | 53.0 | 50 | | T67P | 4.82E-06 | 2.00 | 0.53 |
| 6 | | H64F | 1.73E-05 | 0.01 | 106 | 51 | | T67A | 1.18E-05 | 1.10 | 0.96 |
| 7 | | H64W | 8.60E-07 | 0.07 | 15.1 | 52 | | I107V | 1.14E-05 | 1.20 | 0.88 |
| 8 | | V68A | 2.93E-06 | 1.20 | 0.88 | 53 | | I107T | 9.67E-06 | 1.40 | 0.76 |
| 9 | | V68T | 5.38E-07 | 0.07 | 15.1 | 54 | | I107L | 7.92E-06 | 1.20 | 0.88 |
| 10 | | V68I | 1.29E-06 | 0.23 | 4.61 | 55 | | I107F | 2.88E-06 | 1.60 | 0.66 |
| 11 | | V68L | 3.87E-06 | 3.40 | 0.31 | 56 | | I107W | 1.62E-06 | 3.30 | 0.32 |
| 12 | | V68F | 1.00E-07 | 0.48 | 2.21 | 57 | | I107V | 5.83E-06 | 1.20 | 0.88 |
| 13 | | V68W | 2.00E-08 | 0.59 | 1.80 | 58 | | I111L | 6.83E-06 | 0.90 | 1.18 |
| 14 | | L29A | 9.39E-06 | 0.78 | 1.36 | 59 | | I111F | 1.02E-05 | 0.94 | 1.13 |
| 15 | | L29V | 6.17E-06 | 1.10 | 0.96 | 60 | | I111M | 4.50E-05 | 0.63 | 1.68 |
| 16 | | F L29 | 1.68E-05 | 15.0 | 0.07 | 61 | | I111W | 6.81E-06 | 1.50 | 0.71 |
| 17 | | L29W | 6.73E-06 | 0.03 | 35.3 | 62 | Prox-imal site | L89G | 3.10E-06 | 1.30 | 0.82 |
| 18 | | F43V | 5.17E-05 | 0.16 | 6.63 | 63 | | L89W | 5.18E-06 | 0.28 | 3.79 |
| 19 | | F43L | 1.09E-05 | 0.21 | 5.05 | 64 | | H97A | 6.93E-06 | 2.00 | 0.53 |
| 20 | | F43I | 5.18E-06 | 0.10 | 10.6 | 65 | | H97V | 6.14E-06 | 1.10 | 0.96 |
| 21 | | F43W | 6.73E-06 | 0.17 | 6.24 | 66 | | H97D | 7.00E-06 | 1.90 | 0.56 |
| 22 | Second shell | I28A | 8.45E-06 | 0.96 | 1.10 | 67 | | H97F | 4.42E-06 | 1.50 | 0.71 |
| 23 | | I28W | 4.67E-06 | 2.30 | 0.46 | 68 | | H97Q | 5.83E-06 | 0.48 | 2.21 |
| 24 | | L32A | 8.70E-06 | 0.89 | 1.19 | 69 | | I99A | 1.80E-06 | 2.10 | 0.50 |
| 25 | | L32V | 7.62E-06 | 0.93 | 1.14 | 70 | | I99V | 4.52E-06 | 2.10 | 0.50 |
| 26 | | L32I | 5.25E-06 | 0.94 | 1.13 | 71 | | I99L | 6.94E-06 | 0.48 | 2.21 |
| 27 | | L32F | 3.91E-06 | 1.00 | 1.06 | 72 | | I99N | 4.91E-06 | 3.00 | 0.35 |
| 28 | | L32M | 3.14E-06 | 1.10 | 0.96 | 73 | | L104A | 1.04E-05 | 3.10 | 0.34 |
| 29 | | L32W | 6.13E-07 | 2.70 | 0.39 | 74 | | L104V | 4.24E-06 | 2.00 | 0.53 |
| 30 | | R45A | 6.41E-06 | 0.26 | 4.08 | 75 | | L104W | 3.01E-06 | 5.8 | 0.18 |
| 31 | | R45L | 4.66E-06 | 0.26 | 4.08 | 76 | | F138A | 1.33E-05 | 1.20 | 0.88 |
| 32 | | R45T | 6.89E-06 | 0.17 | 6.24 | 77 | | F138W | 3.14E-06 | 0.79 | 1.34 |
| 33 | | R45K | 4.66E-06 | 0.31 | 3.42 | 78 | Far from heme | W7F | 5.08E-06 | 0.80 | 1.33 |
| 34 | | R45S | 8.10E-06 | 0.42 | 2.52 | 79 | | Q8V | 6.67E-06 | 0.49 | 2.16 |
| 35 | | R45A | 1.23E-05 | 0.06 | 17.7 | 80 | | W14F | 6.99E-06 | 0.66 | 1.61 |
| 36 | | F46V | 7.40E-05 | 0.07 | 15.1 | 81 | | M55A | 3.46E-06 | 1.20 | 0.88 |
| 37 | | F46L | 2.42E-05 | 0.18 | 5.89 | 82 | | M55L | 2.21E-06 | 0.90 | 1.18 |
| 38 | | F46W | 9.72E-06 | 0.28 | 3.79 | 83 | | M55W | 2.57E-06 | 0.83 | 1.28 |
| 39 | | F46A | 2.75E-06 | 0.43 | 2.47 | 84 | | A71F | 6.59E-06 | 0.78 | 1.36 |
| 40 | | L61F | 2.50E-06 | 1.10 | 0.96 | 85 | | L72V | 5.87E-06 | 1.60 | 0.66 |
| 41 | | L61A | 4.56E-06 | 0.88 | 1.20 | 86 | | L72W | 4.71E-06 | 0.73 | 1.45 |
| 42 | | G65I | 4.93E-06 | 0.52 | 2.04 | 87 | | K79A | 7.21E-06 | 1.00 | 1.06 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 43 | | G65T | 1.38E-05 | 1.54 | 0.69 | 88 | | K79L | 6.54E-06 | 1.00 | 1.06 |
| 44 | | G65G | 4.39E-06 | 1.79 | 0.59 | 89 | | M131L | 8.00E-06 | 0.32 | 3.31 |
| 45 | | V66K | 4.24E-06 | 1.25 | 0.85 | 90 | | A144V | 6.41E-06 | 0.83 | 1.28 |

*a*: calculated due to $P_{50} = \frac{1}{\propto_{O_2}(K_{O_2} - K_1)}$

### 3.3. Parameters for modeling of ADL

The values of $K_{O_2}$ for single point mutants of sperm whale Mb (Scott *et al*., 2001) were used to calculate the seal Mb mutant ADLs (see section 2.6 and Appendix B, Table B2). From a Blast ClustalW alignment from the UniProt interface (UniProt Consortium, 2011) of harbor seal (*Phoca vitulina*), gray seal (*Halichoerus grypus*) and sperm whale Mb, all have identical sequence lengths (154), including 128 identical positions and 22 similar positions (Bradshaw and Gurd, 1969). Notably, all the sites studied in this work are identical in WT whales and seals.

The sperm whale $K_{O_2}$ values were first corrected from 20°C to 37°C by a factor of 0.2457 based on the temperature dependence of $P_{50}$ (Schenkman *et al*., 1997) and $\alpha_{O2}$ (Mahler *et al*., 1985). Then, using the $P_{50}$ values of 3.75 mmHg and 3.45 mmHg for sperm whale and Weddell seal WT Mb, all mutant seal Mb $K_{O_2}$-values were calculated by multiplying the sperm whale mutant $K_{O_2}$ at 37°C by 1.085 (the ratio of WT seal $K_{O_2}$ to WT whale $K_{O_2}$) (see Appendix B for uncorrected mutant ADLs). Given the small difference between $P_{50}$ for whale and seal, which is within the experimental uncertainty of thermochemically measured $K_{O_2}$, the use of whale mutant data is a good approximation and does not affect the significance of the conclusions, although the quantitative uncertainties in animal-specific mutant ADL is up to ~2 minutes (*vide infra*).

**Table 3.3**. List of parameters and variables used to model ADL.

| Parameter | Symbol | Values used in the present work |
|---|---|---|
| Cardiac output | $\dot{V}_b$ | 42.00 l $min^{-1}$ at rest |
| Brain blood flow rate | $\dot{Q}_B$ | 0.36 l $min^{-1}$ at rest |
| Heart blood flow rate | $\dot{Q}_H$ | 1.84 l $min^{-1}$ at rest |
| Skeletal muscle blood flow rate | $\dot{Q}_M$ | 7.90 l $min^{-1}$ at rest |
| Blood flow rate for splanchnic, renal, cutaneous, and other peripheral tissues | $\dot{Q}_{SRC}$ | 32.63 l $min^{-1}$ at rest |
| Brain oxygen consumption rate | $\dot{V}_{B_{O_2}}$ | 13.3 ml $O_2$ $min^{-1}$ at rest |
| Heart oxygen consumption rate | $\dot{V}_{H_{O_2}}$ | 112.5 ml $O_2$ $min^{-1}$ at rest |
| Skeletal muscle oxygen consumption rate | $\dot{V}_{M_{O_2}}$ | 216.6 ml $O_2$ $min^{-1}$ at rest |
| $O_2$-consumption rate for splanchnic, renal, cutaneous, and other peripheral tissues | $\dot{V}_{SRC_{O_2}}$ | 555 ml $O_2$ $min^{-1}$ at rest |
| Heart beat rate | $f_h$ | 51.5 beats $min^{-1}$ at rest |
| Arterial blood oxygen saturation | $S_a$ | 100-38 % |
| Venous blood oxygen saturation | $S_v$ | 86-36 % |
| Arterial blood $P_{O_2}$ | $P_a$ | 119-22 mmHg |
| Venous blood $P_{O_2}$ | $P_v$ | 55-21 mmHg |
| $P_{O_2}$ at mitochondria | $P_{mit}$ | ~0 mmHg |
| $P_{O_2}$ at capillary | $P_c$ | 87-21 mmHg |
| Average Mb saturation for mutants | $<S>$ | 99-27 % |
| Average oxygen pressure in cell | $P_{O_2}$ | $P_{mit} < P_{O_2} < P_c$ |
| $P_{O_2}$ at which $S = 0.5$ | $P_{50}$ | ~1-3 mmHg for wild type Mb |
| Bimolecular Mb oxygenation constant | $K_{O_2}$ | ~1 $\mu M^{-1}$ |
| Oxygen solubility in the muscle tissue | $\alpha_{O2}$ | 9.4 x $10^{-7}$ mol $L^{-1}$ $mmHg^{-1}$ |
| Mb concentration in Weddell seal muscle tissue | $C_{Mb}$ | 54 g $kg^{-1}$ muscle |

# CHAPTER 4: SUMMARY OF RESEARCH ARTICLES

In the previous chapters the theoretical basis for this thesis as well as details of the parameters used in the models were presented. In this chapter we aim to describe in more depth the scientific articles published in this project and thus make a connection to the publication part of this thesis. The articles have been divided into two groups; articles related to the $O_2$-binding properties of Mb and articles related to folding stability in mammalian Mbs including the simulated evolution of Mb sequences.

## 4.1. The effect of Mb mutations on cellular and organismal fitness

### 4.1.1. Mutational effects on $O_2$-delivery process

One of the most genuine approaches in physical and mathematical sciences for understanding more about the system of interest is to "perturb" its current state. We then gain valuable information about the functioning, stability and possible transitions of physical systems to different states. In molecular evolution we tend to apply the same approach to proteins by studying their biochemical functions and biophysical properties upon mutations. Once mutations (either randomly or rationally generated) occur in the protein of interest, one can ask the following questions: i) how is function or stability of the protein affected by these mutations, ii) how is a cellular process influenced by these mutations and finally iii) whether these mutations eventually alter fitness in living organisms or not.

In the case of Mb, the first question can be answered by either experimental or theoretical studies of mutational effects on thermodynamic or kinetic constants of $O_2$-binding to the protein or folding/unfolding events. This approach has been extensively attempted since the time Mb structure was resolved (Kendrew et al., 1958). As explained in the Introduction, Mb has been the perfect model system to understand mutational effects by theoretical and experimental studies

(Frauenfelder et al., 2003). For this reason, we skipped this level of investigation and used the available literature for Mb mutants to formulate our hypothesis and scientific work[3].

In the next level of analysis, we estimated the effect of mutations on $O_2$-delivery process in muscle tissues by first quantifying the proficiencies of Mb mutants as an oxygen storage and transport protein (OSR and OTR defined in section 2.4). The main purpose of this work was to quantify and understand the functional proficiency of various Mb mutants (Table 3.2) under changing environmental conditions, and subsequently, to deduce the selection-pressure that conserves parts of the WT protein.

We calculated values of OSR for sperm whale Mb at $P_u$ = 40 mmHg and $P_u$ = 5 mmHg at $\dot{V}_{O2} = 0.3\ \mu\text{mol}\ \text{L}^{-1}\text{s}^{-1}$, which corresponds to normoxic rest conditions, versus the rank of each mutant as is shown in Figure 4.1, top. Because of the well-known saturation behavior of Mb, lower $P_{O_2}$ causes a higher contribution of Mb to the overall oxygen storage, compared to free $O_2$, so that Mb strongly compensates the depletion of free $O_2$. The average OSR for the whole data set increases from ~80 to ~553, suggesting a 6-fold increase in the ratio of stored $O_2$ relative to free $O_2$ upon transition from normoxic to hypoxic conditions. In physiological terms, this implies that the relative importance as an $O_2$-storage of an average Mb mutant increases ~6-fold upon transition to hypoxic conditions.

The major changes in OSR and OTR are due to the mutations of residues His-64, Val-68, Leu-29, and Phe-43 in the first shell, and Arg-45, Phe-46, Leu-61, and Gly-65 in the second shell of the distal pocket. Mutations in Ile-111, Leu-89, His-97, Ile-99, and Phe-138 in the proximal site, and in Met-131 far from heme show small effects on OSR and OTR. Almost all these

[3] As presented in the publication list of this thesis, we also worked on the role of distal interactions on $O_2$-binding to heme by quantum mechanical (DFT) computations. However, for the sake of integrity and the fact that this thesis is concerned with protein evolution, I did not include the paper in the present thesis.

mutations impair the ability of protein to store oxygen, and the impairment is highly dependent on $P_{O_2}$. The severe effect of mutations affecting His-64 can be understood from the existence of a hydrogen bond from site 64 to $O_2$ (Olson 2008). For example, the H64V mutation completely lacks the hydrogen bond, and the thermochemical data reveal a two-orders-of-magnitude decrease in the binding constant. However, we can now see that the physiological effect on $O_2$-storage is somewhat smaller in the whale: OSR decreases from ~80 to ~23 at normoxic conditions, and from ~553 to ~30 under hypoxic conditions, given the very small whale muscle oxygen consumption rate of $\dot{V}_{O2} = 0.3\ \mu\text{mol}\ \text{L}^{-1}\text{s}^{-1}$.

Using equation (12) in chapter two, OTR can be calculated for the whole mutant data set for any given $P_u$ of the cell. The bottom of Figure 4.1 shows the OTR values for the WT and mutants at $P_u$ = 40 mmHg and $P_u$ = 5 mmHg at $\dot{V}_{O2} = 0.3\ \mu\text{mol}\ \text{L}^{-1}\text{s}^{-1}$. At $P_u$ = 40 mmHg, facilitated $O_2$-transport is negligible. However, at low $P_{O_2}$, Mb has a significant role in $O_2$-transport. This can be explained by the fact that *S* changes over lower $P_{O_2}$ regions of the cell, giving rise to a larger [$MbO_2$] gradient and hence, a larger flux of $MbO_2$. Thus, we identify a crucial difference between the behavior of the transport and storage functions under changing environmental conditions, showing that the two functions are intrinsically distinct. In Figure 4.2, bottom, OTR is compared for the WT and mutants in the range from $P_u$ = 40 mmHg to $P_u$ = 1 mmHg. It can be seen that Mb contributes substantially to $O_2$-transport only below $P_u$ ~ 10 mmHg. Similar to OSR, OTR is most affected by replacements within the first and second shells, while mutations within the third shell generally retain WT function.

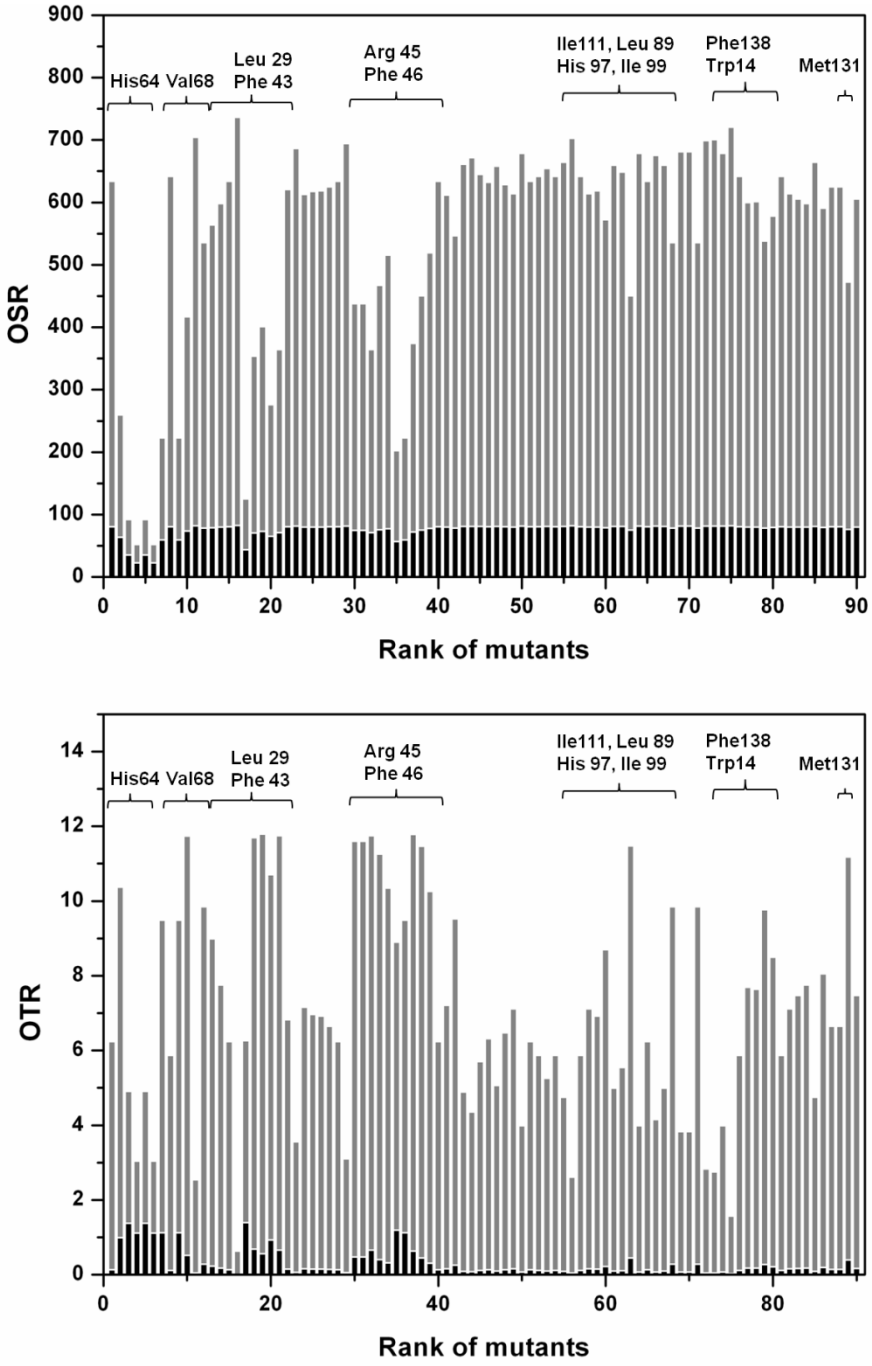



**Figure 4.1.** The OSR (top) and OTR (bottom) values at upper pressure $P_u$ = 40 mmHg (black bars) and $P_u$ = 5 mmHg (gray bars), corresponding to normoxic and hypoxic conditions. Both OSR and OTR are calculated for sperm whale having $\dot{V}_{O2} = 0.3$ μ mol $L^{-1}s^{-1}$, $C_{Mb}$ = 3.1 m mol $L^{-1}$ and $R_k$ = 19 μm (Dasmeh et al., 2012).

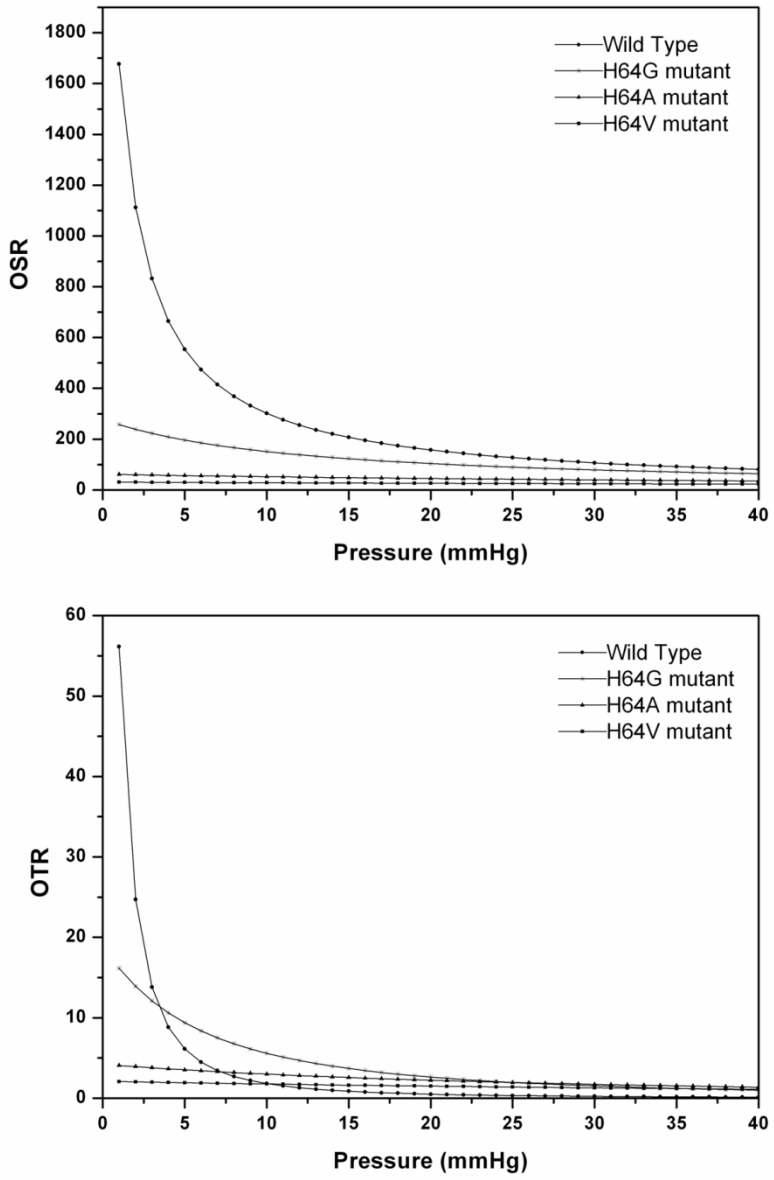



**Figure 4.2.** The OSR (top) and OTR (bottom) vs. $P_{O_2}$ for wild type Mb and H64G, V68A, and H64V mutants. The calculations are based on $\dot{V}_{O2} = 0.4$ μ mol $L^{-1}s^{-1}$, $C_{Mb}$ = 3.1 m mol $L^{-1}$, and $R_k$ = 19 μm (Dasmeh et al., 2012).

An important observation from our model is that the proficiency of the WT as compared to other mutants depends very much on $P_{O_2}$. As seen in Figure 4.1 bottom, OTR for low-affinity mutants (primarily those with mutations in sites 64, 68, 29, and 43) is in fact *greater* than WT at higher $P_{O_2}$. Furthermore, this behavior is inversed at lower $P_{O_2}$ where the WT is much more

proficient in terms of transport, compared to these mutants. In fact, over the $P_{O_2}$ range, there are "critical pressures" at which two mutants are equally proficient despite their different $K_{O_2}$, as is shown in Figure 4.2. The reason for this surprising observation is that OTR, as calculated from equation (10), depends mainly on the derivative of saturation (i.e., $\frac{dS}{dP}$), which has a $P_{O_2}$-dependent behavior. This behavior is shown in Figure 4.3, where the slopes of the saturation curve of the WT and the H64G mutant are compared at $P_u$ = 20 mmHg and $P_u$ = 2 mmHg, close to the $P_{50}$ of ~1 mmHg of the wild type protein. Saturation curves for mutants with higher $P_{50}$ are less steep at lower $P_{O_2}$ and steeper at normoxic conditions, compared to the WT. This situation causes the WT to be a better transport protein at lower $P_{O_2}$, whereas low-affinity mutants are in fact better transporters at normoxic conditions.

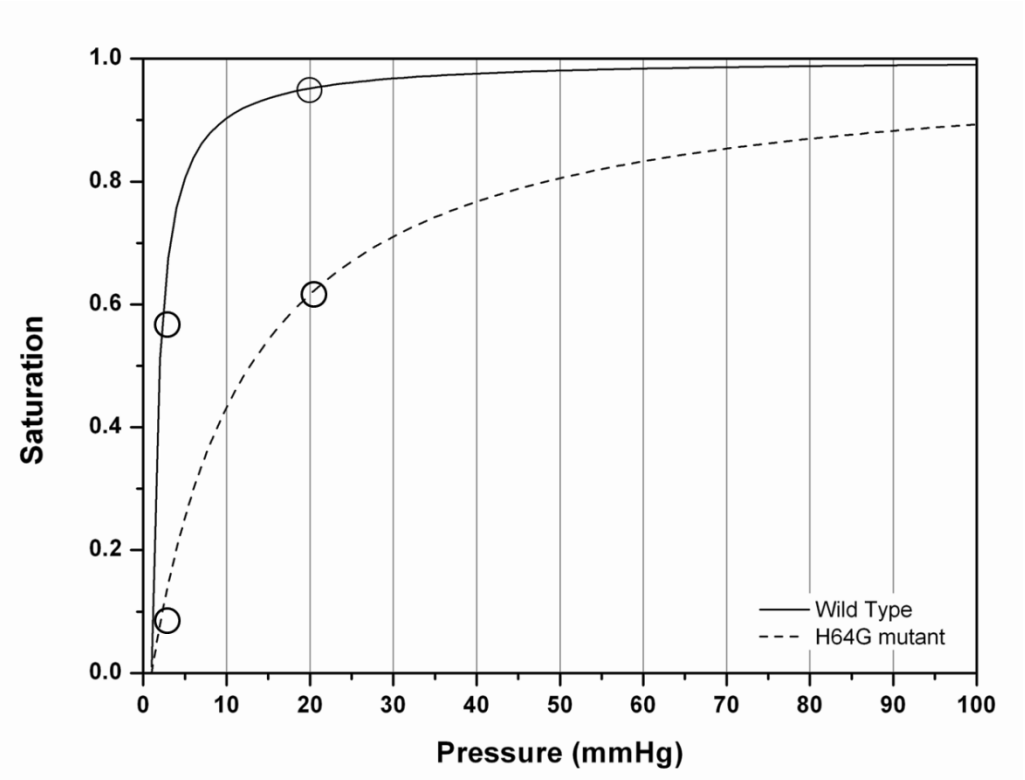



**Figure 4.3.** The saturation curve for WT Mb (solid line) and the H64G mutant (dashed line). The circles compare saturation of two proteins at $P_{O_2}$ = 20 mmHg and 2 mmHg. Calculations are based on sperm whale having $\dot{V}_{O2}$ = 0.3 μ mol $L^{-1}s^{-1}$, $C_{Mb}$ = 3.1 m mol $L^{-1}$, and $R_k$ = 19 μm (Dasmeh et al., 2012).

We can also look at the distribution of functionality as we change the physiological parameters. The main idea behind this approach is to find the physiological conditions (i.e., here $P_{O_2}$) that correspond to the maximum variation (and thus differences) among Mb mutants. Figure 4.4 shows the distribution of OSR and OTR at different $P_u$ levels in the cell. The distributions of functional proficiency tend to be centered near the WT proficiency, showed in dashed lines. This observation is due to fact that many mutations are less likely to affect protein function although, these mutations occurred in the positions that are important for $O_2$-binding. We define these mutations as "neutral in term of function". Interestingly, higher $\dot{V}_{O2}$ or lower $P_{O_2}$ not only increases the role of Mb in active oxygen delivery, but also increases the *variation* in functionality among all mutants. The very narrow distribution of functionality at $P_u$ = 40 mmHg changes to a much wider distribution at $P_u$ = 5 mmHg, accompanied by an increased variance from ~148 to ~2526 for OSR and from ~0.13 to ~7.2 for OTR. It is tempting to deduce from this that the selection pressure is much stronger where the variation is large, i.e., where more mutants are impaired, relative to the WT: From an evolutionary point of view, larger functional variation favors the selection of beneficial mutations, including the WT protein.

Besides investigating OSR and OTR at different $P_{O_2}$ levels, we can also investigate the effect of increased $\dot{V}_{O2}$ that resembles work conditions, e.g. during diving, or otherwise elevated metabolism on the functional proficiencies of Mb mutants. Moreover, one can also change cell size parameters and Mb concentration in comparative-like studies within this framework (see Appendix A for details).

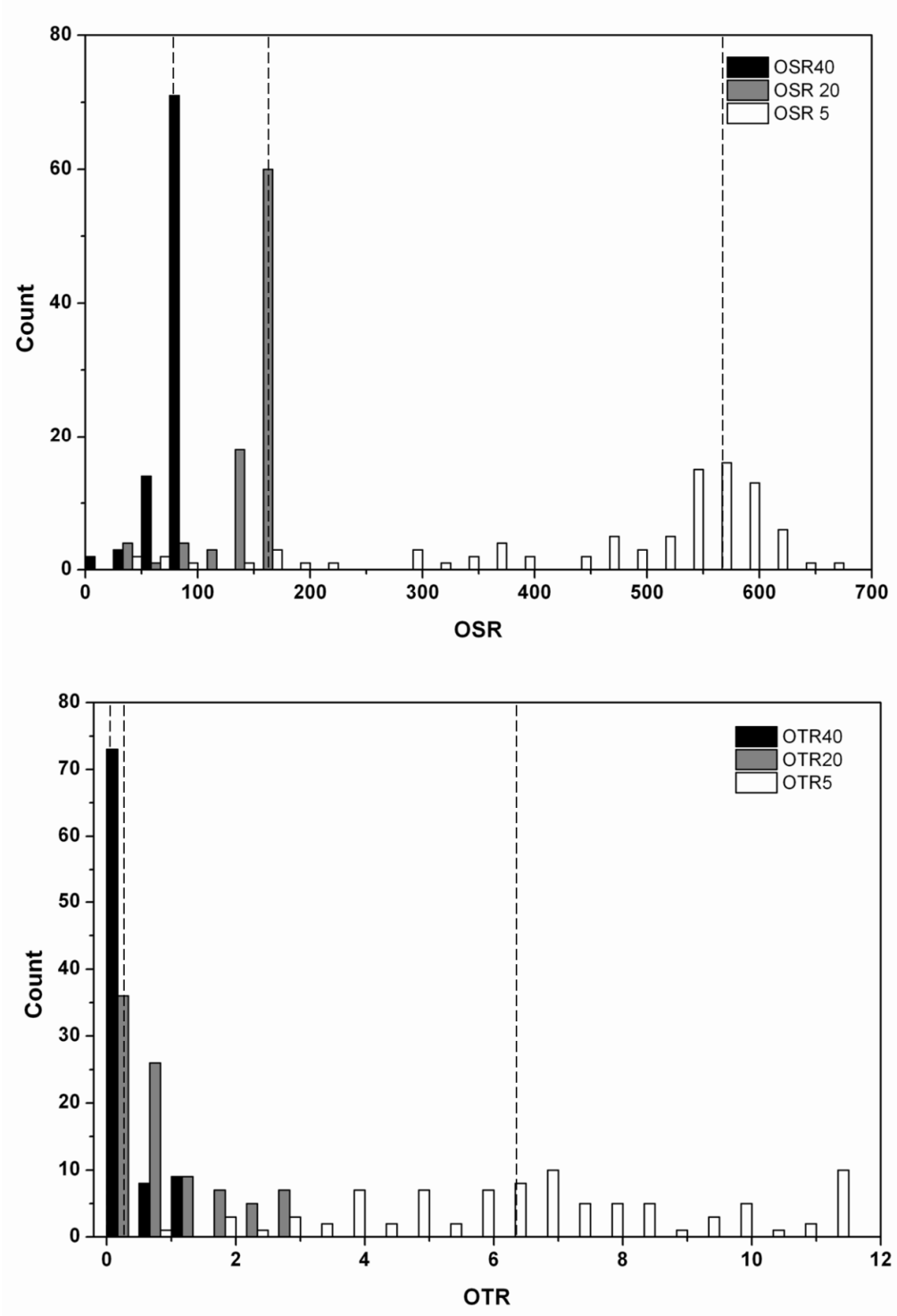



**Figure 4.4.** The OSR (up) and OTR (down) distributions among mutants for three upper pressures of $P_u$ = 40 mmHg (black), 20 mmHg (gray), and 5 mmHg (white). The calculations are based on sperm whale having $\dot{V}_{O2} = 0.3$ μ mol $L^{-1}s^{-1}$, $C_{Mb}$ = 3.1 m mol $L^{-1}$, and $R_k$ = 19 μm. All the dashed lines show the position of WT Mb in the distributions (Dasmeh et al., 2012).

### 4.1.2. Mutational effects on aerobic dive limit (ADL)

Here we investigate the effect of Mb mutations on the dive time of Weddell seal as one of the model systems in animal physiology. Basically, diving times in marine mammals are the limits that animals have enough $O_2$ to consume while submerged. Any impairment of muscle oxygen storage (by mutations in Mb) can reduce the amount of available $O_2$ for diving and thus influences the dive time *ceteris paribus*. We take the muscle model developed in this work (section 2.3) and integrate it with the convective $O_2$ transport in Weddell seal (section 2.6).

As presented in chapter 2 of this thesis, we can model aerobic dive limits of marine mammals by using an iterative model of convective oxygen transport through different organs (section 2.6). In this model, any mutation in Mb directly affects the average oxygen saturation of muscle tissues (section 2.5) and thus influences ADL. However, this is not a linear effect in all possible behavioral and physiological conditions. To better compare the effectiveness of low vs. high $O_2$-affinity Mbs (either from different organisms or e.g. for two mutants) at various physiological conditions, we computed the change in ADL when $K_{O_2}$ changes from 0.01 to 1 $\mu M^{-1}$ at different combinations of $\dot{V}_{M_{O_2}}$ and $\dot{V}_b$, as shown in Figure 4.5. The maximum effectiveness of a high-affinity Mb with $K_{O_2}$ ~1 $\mu M^{-1}$ occurs at $\dot{V}_{M_{O_2}} \sim 3\ \dot{V}_{M_{O_2}}$ *(rest)* and $\dot{V}_b \sim 0.15\ \dot{V}_b$ *(rest)*. Under these conditions, and with fully oxygenated blood at the beginning of a dive, the ADL can be increased by the most effective Mbs up to 14 min compared to impaired mutants with $K_{O_2} \sim 0.01\ \mu M^{-1}$. The single hydrogen bond between His-64 and $O_2$ now reveals its physiological importance. This single residue appears to be crucial for the survival of the animal since a 14 minute reduction in ADL would have a substantial effect on foraging, mating, and predator evading abilities.

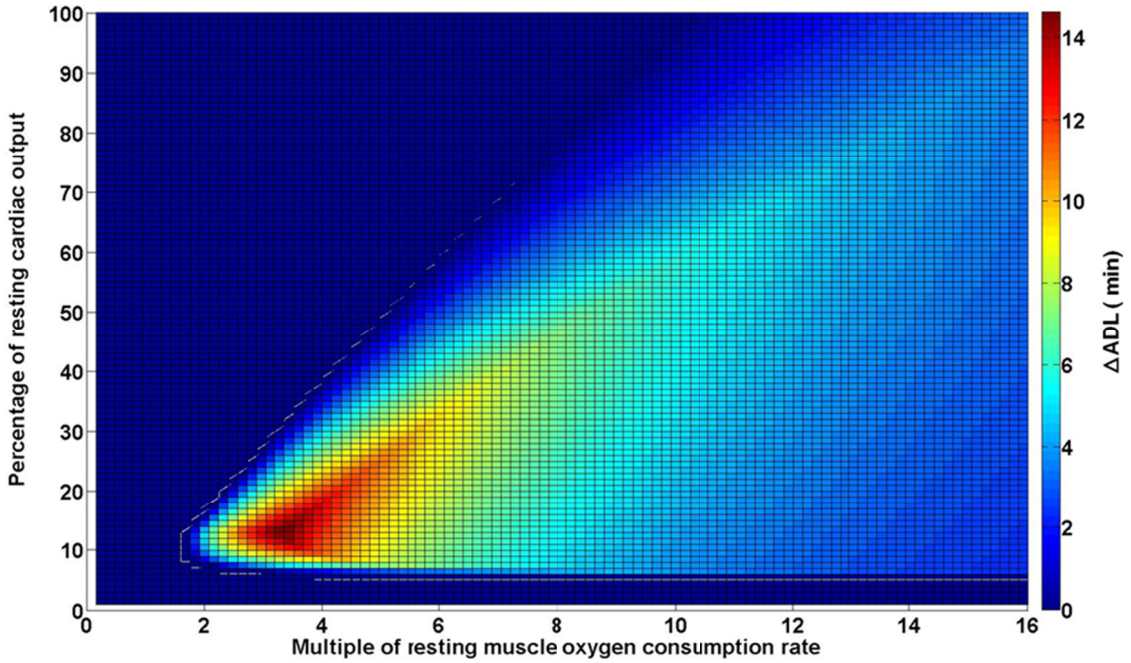



**Figure 4.5.** Plot of change in ADL (in minutes, color bar) going from $K_{O_2}$ = 0.01 to 1 μM$^{-1}$, over possible combinations of $\dot{V}_{M_{O_2}}$ and $\dot{V}_b$, using temperature-corrected seal data. The maximum occurs at $\dot{V}_{M_{O_2}}$ ~3 $\dot{V}_{M_{O_2}}$ *(rest)* and $\dot{V}_b$ ~0.15 $\dot{V}_b$ *(rest)*. $P_a$ = 119 mmHg and $P_v$ = 55 mmHg, corresponding to fully oxygenated, hyperventilated starting dive conditions (Dasmeh et al., 2013a).

Figure 4.5 can be thought of as a differential fitness landscape for the evolution of Mb in closely related marine mammals (e.g. seals and whales). The vast majority of conditions in Figure 4.5 can be seen to not effectively distinguish low-affinity Mb mutants from high-affinity mutants and WT. However, under very specific conditions, most pronounced at $\dot{V}_{M_{O_2}}$ ~3-4 $\dot{V}_{M_{O_2}}$ *(rest)* and $\dot{V}_b$ ~ 0.15 $\dot{V}_b$ *(rest)*, high affinity Mbs are maximally proficient, compared to alternatives. Importantly, these conditions of optimal differential ADL are very similar to those prevailing under routine dives ($\dot{V}_{M_{O_2}}$ ~5$\dot{V}_{M_{O_2}}$ *(rest)*). We might identify these conditions as what molecular

evolutionists would call the *selection pressure*: the physiological and environmental conditions where the WT protein is most proficient compared to impaired mutants. Although the exact values of the parameters will change in various diving mammals, this qualitative observation is general and confirms previous explanations of why $C_{Mb}$ is so much higher in diving marine mammals (Kooyman and Ponganis, 1998; Helbo and Fago, 2012) and how cardiac and muscle output has co-evolved not only with routine diving behavior (Davis et al., 2004) but also directly with the oxygen-storing capacity of WT Mb, quantified by experimental $K_{O_2}$.

We can use the model described above to quantify the effect of single point mutations in Mb on the ADL for the Weddell seal. However, the conclusions are valid for other animals with proper adjustment of physiological and thermochemical parameters. To calculate the average Mb saturation and ADL, we used a routine aerobic diving $\dot{V}_{M_{O_2}}$ of five times resting and $\dot{V}_b$ ~ 0.27 times that at rest (Davis et al., 2004).

ADLs for specific mutants are shown in Figure 4.6. The error bars represent the uncertainty in ADL caused by 20% experimental errors in $K_{O_2}$-values. It can be seen from Figure 4.6 that the most deleterious single-point mutations in residues 64 or 29 decrease ADL from 17 minutes of the WT to e.g. ~3-4 minutes under routine diving conditions. Such a reduction in ADL would greatly reduce the foraging ability of the seal and probably results in an inability to forage which would be lethal. This conclusion is based on the routine swimming speed (1.2 m $s^{-1}$) (Davis *et al*., 2003). For a 4 min dive, the seal would be able to swim to a depth of about 144 m, which is shallower than the routine depth of its primary prey (*Pleuragramma antarcticum*) at depths greater than 160 m (Fuiman *et al*., 2007). In addition, it would have no time to pursue prey even if they were encountered. Hence the seal would be unable to feed effectively with such a severe reduction in the ADL.

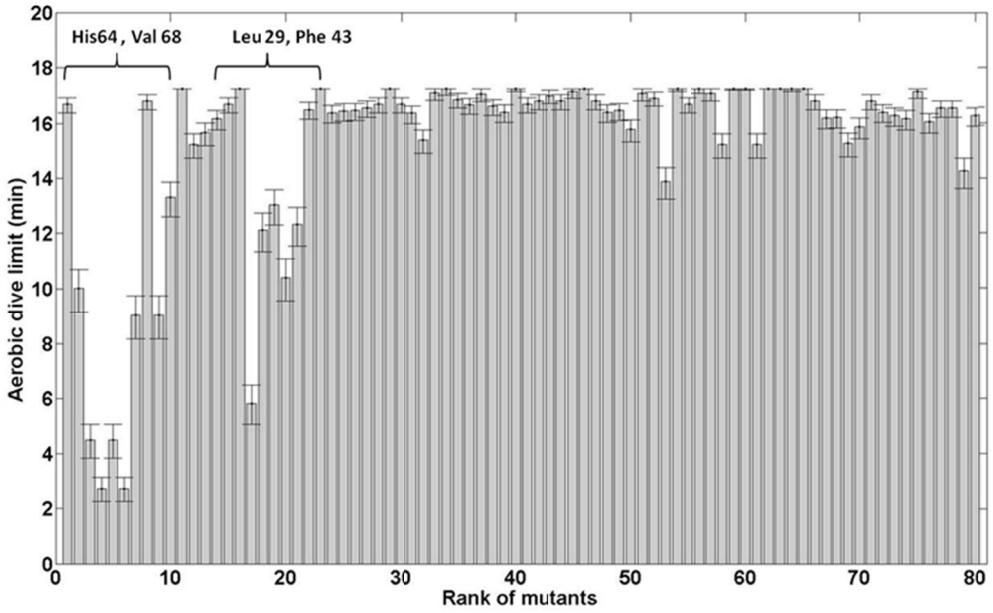



**Figure 4.6.** The calculated ADL (min) of the simulated dive of a Weddell seal with $\dot{V}_{M_{O_2}} = 5\ \dot{V}_{M_{O_2}}(rest)$ and $\dot{V}_b = 0.27\ \dot{V}_b(rest)$ at $37^0C$ for different Mb mutants, using converted seal mutant $K_{O_2}$ data. Error bars are calculated from 20% experimental error in $K_{O_2}$ values (Dasmeh et al., 2013a).

From the calculations, the most impairing mutations in sites 64, 68, 43, and 29 of Mb cause up to five-fold reductions in available muscle $O_2$ and ADL. From the figure, for high-affinity mutants with larger $K_{O_2}$ than WT (e.g. V68L and L29F), muscle $O_2$ is not a limiting factor for ADL under routine dive conditions, whereas cardiac output becomes a limiting factor, and thus ADL is not significantly higher than for the WT. This important result shows that the cardiac, muscular, and behavioral adaptations of seals are correlated with the WT proficiency, suggesting that all these aspects most likely have co-evolved.

The uncertainties of the experimental oxygenation constants correspond to uncertainties of computed ADLs of ca. $\pm$1 minute. In addition, uncertainties in the initial parameters and the choice of $P_{O_2}$ in the mitochondria and capillary during various scenarios affected the saturation curve and ADL. A sensitivity analysis using ±10% change in the initial parameters showed that Mb concentration and muscle $O_2$ consumption rate were the main sources of potential

errors, with ~15% and ~10% effect on ADL, respectively. Uncertainties in other parameters had minor effects on the ADL (see Appendix B, Figure B23). The model provided ADLs with uncertainties of approximately ±10% or ca. 2 minutes. Thus, the qualitative conclusions are not affected by any reasonable change in input parameters arising from individual variation or inherent uncertainties in experimental data.

### 4.2. Selection for thermodynamic stability

#### 4.2.1. Positive selection of folding stability in cetacean Mbs

As was described early in the introduction and later quantified in the second chapter of this thesis, the increased concentration of Mb in the skeletal muscle of deep-diving cetacean species is one of the main evolutionary adaptations during the transition of cetaceans to aquatic environment. Moreover, despite nearly unchanged $O_2$-binding affinity of Mb among mammals (Scott et al., 2000, Helbo and Fago, 2012) cetacean Mbs are more stable than their terrestrial counterparts by ~2-4 kcal/mol (Scott et al., 2000). The current hypothesis for this increased stability is due to the sustained anaerobic and acidic conditions in the skeletal muscle of marine mammals as they experience prolonged dives (Scott et al., 2000). Mbs in skeletal muscle cells of these species are then suggested to be under selective pressure for increased resistance to acid induced unfolding (Scott et al., 2000).

This hypothesis is in contrast with several observations. First, marine mammals generally stay under aerobic metabolism due to the high cost of recovery after switch to anaerobic conditions (Ponganis, 2011). The longest dives recorded for large whales such as blue and fin whales are much shorter than predicted the dive limits under aerobic conditions (ADL) (Croll et al., 2001). In similar studies of sperm whales and seals, almost all the dives were found to not greatly exceed ADL (Kooyman et al., 1980; Watwood et al., 2006). Second, even in the case of switch to

anaerobic metabolism, the pH-fall in muscle and blood of marine mammals after the long dives (less than one unit from its physiological value) is too small to initiate unfolding (Hughson and Baldwin, 1989). These observations show that a switch to anaerobic metabolism and sustained acidosis is far from the *modus operandi* of marine mammals as observed in the wild (Kooyman et al., 1980).

Interestingly, we observed a correlation between Mb concentrations and stability (i.e., $\Delta G$) across mammalian species (Figure 4.7). This observation can be explained by the recently proposed hypothesis for the increased tendency of highly expressed protein to be more stable (Drummond and Wilke, 2008; Yang et al., 2010; Serohijos et al., 2013). However, the crux is to show that the extra stability of cetacean Mbs is "selected" during the evolution and is not a result of neutral evolution.

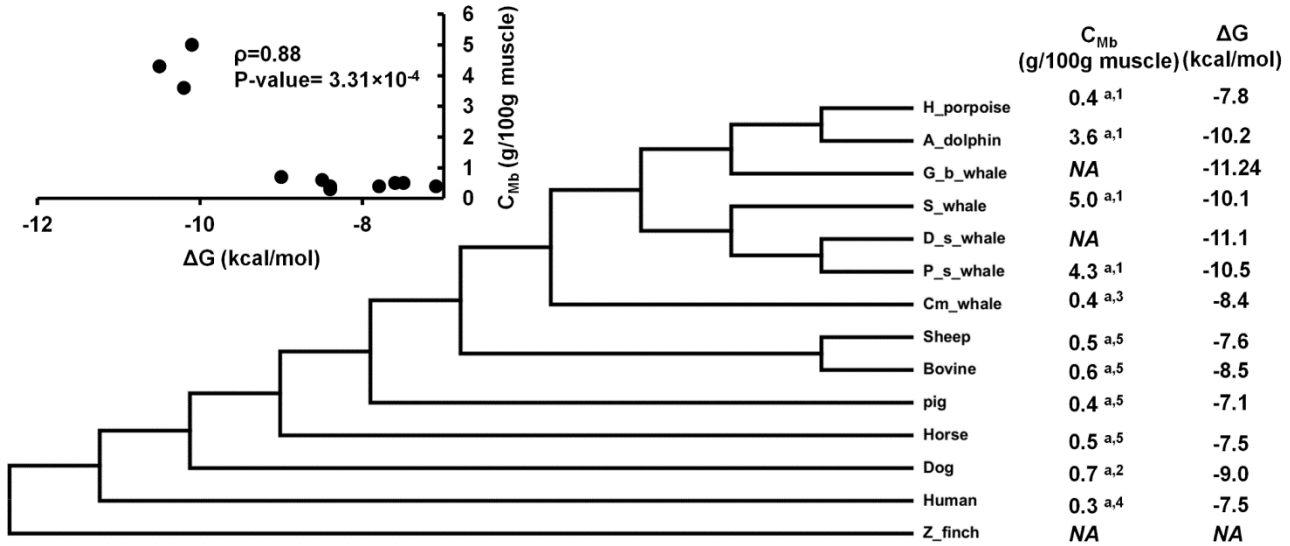



**Figure 4.7.** Divergence of cetaceans and the increase in Mb concentration by ~10–20 fold. The experimental folding stability of apoMb is added to the difference in stability of holo and apoMb reported for horse heart Mb (2.7 kcal/mol). Stability is highly correlated with Mb concentration with correlation coefficient ρ = 0.88 and p-value = 0.000331. The Mb concentration has been measured in [a]dorsi and in [b]psaos muscle types. Data are taken from 1: (Noren and Williams 2000), 2: (Reynafarje 1963), 3: (Dolar et al., 1999), 4: (Gros et al., 2010) and 5: (Lawrie 1953). All the folding stabilities are taken from (Scott et al., 2000) (Dasmeh et al., 2013b).

The first step in dealing with this concern (i.e., non-neutral evolution of folding stability in cetacean Mbs) is to simply check the evolution rate quantified as *dN/dS,* i.e., rate ratio of nonsynonymous to synonymous substitutions, between the cetacean clade and another terrestrial clade of mammalian phylogeny where nucleotide sequences are available. As presented in Table 4.1, we observed a higher rate of evolution in the whole-gene *dN/dS* pairwise comparison of cetaceans compared to primates. The null hypothesis of two sets of *dN/dS* in primate and cetacean Mbs being similar is strongly rejected with the P-value of ~ $1.33\times10^{-16}$ using the two-sample t-test.

**Table 4.1**. The pair-wise evolution rate (i.e., dN/dS) among the cetacean and primate Mbs using the maximum likelihood approach described in chapter 2 (Dasmeh et al., 2013b).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| L_b_whale | | | | | | | | | |
| S_whale | 0.2761 | | | | | | | | |
| P_s_whale | 0.2259 | 0.2122 | | | | | | | |
| M_whale | 0.2647 | 0.2741 | 0.2735 | | | | | | |
| M_h_whale | 0.2433 | 0.1950 | 0.1890 | 0.1754 | | | | | |
| P_b_whale | 0.3057 | 0.2324 | 0.2636 | 0.1566 | 0.2386 | | | | |
| Sei_whale | 0.3469 | 0.2538 | 0.2832 | 0.1262 | 0.2173 | 0.001 | | | |
| S_b_whale | 0.1079 | 0.2641 | 0.2166 | 0.2796 | 0.2723 | 0.3261 | 0.3705 | | |
| Dolphin | 0.2805 | 0.2536 | 0.2374 | 0.2096 | 0.3328 | 0.2865 | 0.2592 | 0.3176 | |
| | L_b_whale | S_whale | P_s_whale | M_whale | M_h_whale | P_b_whale | Sei_whale | S_b_whale | Dolphin |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Human | | | | | | | | |
| Chimpanzee | 0.0312 | | | | | | | |
| Macaque | 0.0635 | 0.0860 | | | | | | |
| Gibbon | 0.0532 | 0.0774 | 0.0738 | | | | | |
| Marmoset | 0.1272 | 0.1480 | 0.0647 | 0.1101 | | | | |
| Gorilla | 0.0435 | 0.0941 | 0.0949 | 0.1053 | 0.1666 | | | |
| Lemur | 0.0487 | 0.0514 | 0.0566 | 0.0511 | 0.0537 | 0.0513 | | |
| Galago | 0.0964 | 0.0900 | 0.0742 | 0.1138 | 0.0753 | 0.0911 | 0.0905 | |
| | Human | Chimpanzee | Macaque | Gibbon | Marmoset | Gorilla | Lemur | Galago |

We also constrained ω to be the same in the whole cetacean clade ($\omega_1$) and different for the rest of the mammals ($\omega_0$). LRT for this comparison was significant when it is compared with the one-ratio test with P-value < $10^{-16}$. For ~26% of sites in Mb, $\omega_1$=0.43 and $\omega_0$=0.19, testifying to a significantly higher evolution rate in cetaceans (Table 4.2).

The observation of the higher rate of evolution in the cetacean clade could suggest accelerated evolution of cetacean Mbs by positive selection of specific residues. To test this, we compared three site pair-models as M1-M2, M7-M8 and M8fix-M8 to identify sites under positive selection, as presented in Table 4.2 (see section 2.8 for details). From Table 4.2, M8 vs. M8fix comparison as the most stringent test indicated that seven sites (5, 22, 35, 51, 66, 121, and 129) are under positive selection with high posterior probabilities using the Bayes empirical Bayes (BEB) test (Yang et al., 2005). Residue 21 was also detected to have a substantially high *dN/dS*, but its rate was not significantly greater than 1 and thus this residue was not detected by the BEB test. Figure 4.8 shows these eight sites with their posterior BEB probabilities using the M8 model, mapped onto the structure of sperm-whale Mb (Dasmeh et al., 2013b).

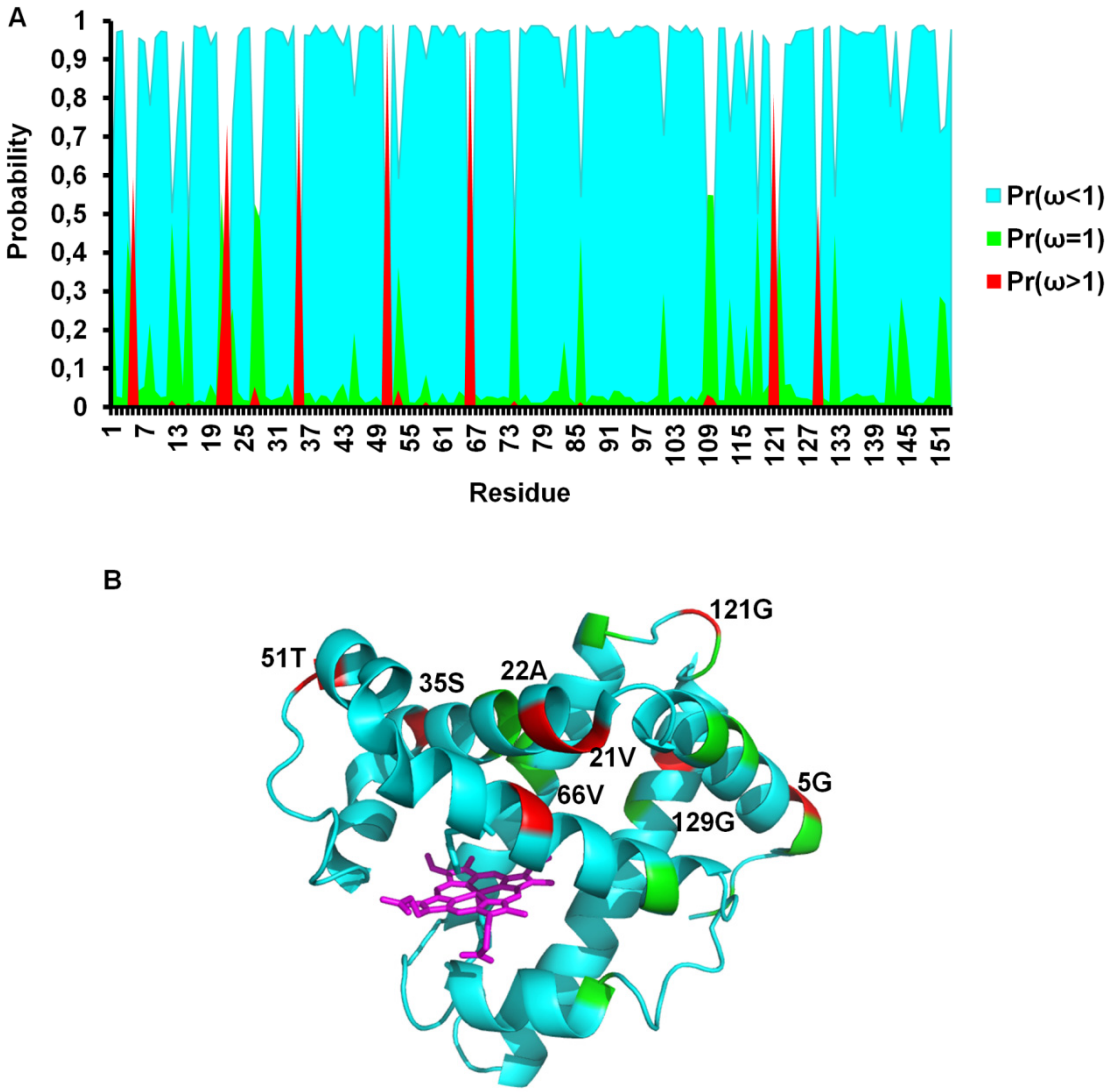



**Figure 4.8.** The Bayes empirical Bayes predictions for ω values for each site in cetacean Mb.A) For each residue p(ω<1), p(ω = 1) and p(ω>1) are shown in cyan, green and red respectively. Residues 5, 21, 22, 35, 51, 66, 121, and 129 have probabilities (ω>1)>0.5 with <ω> = 5.86 from the M8 model using the ML-estimated branch lengths under the M0 model. B) Crystal structure of sperm whale Mb taken from the protein data bank (ID = 1U7S) (Kondrashov et al., 2008) with residues color coded by p(ω). The figure was created using PyMOL (http://www.pymol.org) (Dasmeh et al., 2013b).

**Table 4.2.** Log likelihood values and parameter estimates of the site models, and branch-site models. Branch lengths are first estimated by the M0 model and then used in the more advanced model (Dasmeh et al., 2013b).

| Clades | Model | ln L | Estimates of parameters | 2Δl | P-value | Positively selected sites (BEB: P(ω>1)>0.50)[a] |
|---|---|---|---|---|---|---|
| **Cetacea** | M0 (one ratio) | -1241.82 | $\omega_0$=0.1980 | | | |
| | Free ratio | -1236.39 | See Appendix C | (M0 vs. Free ratio) 10.86 | 0.69 | - |
| | **Site models** | | | | | |
| | M1a | -1251.18 | $p_0$=0.83845, $p_1$=0.16155, $\omega_0$=0.02688, $\omega_1$=1 | | | - |
| | M2a | -1248.47 | $p_0$=0.84199, $p_1$=0.14878, $p_2$=0.00922, $\omega_0$=0.03212, $\omega_1$=1.00000, $\omega_2$=4.91963 | (M1a vs M2a) 5.42 | 0.06 | **5**, 22, 35, **51**, **66**, 121, **129** |
| | M7 | -1251.39 | p=0.06085 q=0.29213 | | | - |
| | M8 | -1247.47 | $p_0$=0.98777, p=0.11682, q=0.66881, $p_1$=0.01223, ω=4.33010 | (M7 vs. M8) 7.84 | 0.019 | **5**, 22, 35, **51**, **66**, **121**, **129** |
| | M8fix | -1251.06 | $p_0$=0.86441, p=0.11615, q=2.08136, $p_1$=0.13559, ω=1.00000 | (M8fix vs M8) 7.18 | $7.37\times10^{-3}$ | - |
| **Terrestrial mammals** | M0 (one ratio) | -4499.91 | $\omega_0$=0.1062 | | | - |
| | Free ratio | -4469.29 | See Appendix C | (M0 vs. Free ratio) 61.24 | 0.065 | - |
| **Mammals** | M0 (one ratio) | -4926.63 | $\omega_0$=0.08 | | | - |
| | Free ratio | -4872.64 | See Appendix C | (M0 vs. Free ratio) 107.98 | $4.8\times10^{-4}$ | - |
| | **Site models** | | | | | |
| | M1a | -4646.77 | $p_0$=0.88207, $p_1$=0.11793, $\omega_0$=0.05590, $\omega_1$=1 | | | - |
| | Clade model ( cetaceans) | -4594.72 | $p_0$=0.68694, $p_1$=0.04973,$p_2$=0.26333, branch type 0: $\omega_0$=0.02043,$\omega_1$=1.00000, $\omega_2$=0.19272, branch type 1: $\omega_0$=0.02043, $\omega_1$=1.00000, $\omega_2$=0.43113 | (M1a vs. Clade Model) 104.1 | $<10^{-16}$ | - |
| | **Branch-site models** | | | | | |
| | Model A | -4643.53 | $p_0$=0.74119, $p_1$=0.09943, $p_2$=0.14053,$p_3$=0.01885, $\omega_0$=0.05388, $\omega_1$=1, $\omega_2$=1 | (M1a vs Model A) 486 | $<10^{-16}$ | - |
| | Null model A (ω=1) | -4643.53 | $p_0$=0.62272, $p_1$=0.08364, $p_2$=0.25887, $p_3$=0.03477, $\omega_0$=0.05392, $\omega_1$=1, $\omega_2$=1 | (model A vs Null model A) 2 | 1 | 15, 27, 28, 101, 118, 140 |

a: P(ω>1)>0.95 is shown in bold.

We were then interested to track the mutational pathways across different lineages of cetaceans by constructing ancestral sequences (Figure 4.9). To infer ancestral states we used the large species tree in Figure 2.3B constructed from 82 Mb amino acid sequences, applying the Dayhoff substitution matrix allowing for among-site-rate-variation as explained in section 2.7. The probability of inference was 1 for all sites except in the sites 1, 13 and 28 where it was 0.5-0.9. In all of these sites, the alternative preferred amino acid was the initial mutated amino acid and thus our results did not encounter the problem of combinatorial ancestral characters (Gaucher, 2007).

Using the FoldX algorithm (section 2.9), we computed the $\Delta\Delta G$ of the mutations in each branch of phylogeny as is shown in Figure 4.9. The overall stabilization or destabilization of each branch is depicted in red or blue, and the branch height is proportional to the absolute computed $\Delta\Delta G$ value of that specific branch. We can see that the overall stability increases in seven branches distributed from -0.3 to -5.1 kcal/mol.

From Figure 4.9, a substantial increase of ~5.1 kcal/mol was gained by mutations G15A, E27D, V28I, V101I, K118R, and G129A upon divergence of cetaceans from the rest of mammals. From Table 4.2, the total ω is not significantly greater than 1, but this may be an unrealistically strict criterion for a small, highly constrained protein such as Mb, as evolutionary rate is strongly correlated to protein size due to the fraction of near-neutral sites increasing with size. Instead, LRT was significant when the branch-site test for positive selection (model A) was compared with the nearly neutral model (M1a), which indicates a higher ω in this first branch leading to cetaceans. The higher ω along this ancestral branch is consistent with positive selection under a new arising selection pressure. As presented in Table 4.2, selection is further supported by the identified amino acid sites in the BEB test having high probabilities along this specific branch, and by the massive increase in the stability phenotype of ~5 kcal/mol occuring in this branching.

After this early divergence that presumably established the majority of the new Mb stability, throughout the cetacean lineages, we found that folding stability is maintained by fixation of several stabilizing mutations. From Figure 4.9A, the key mutations preserving this tendency are G5A, V13I, V21I, V21L, E27D, G35S, S35H, N66V, N66H, N66I, G74A, D83E, K118R, G121S, and G129A mutations. Eight of these mutations occur in the five sites 5, 35, 66, 121, and 129 which were detected by to be under positive selection. Thus, the pure sequence-based maximum likelihood methods, amino acid substitution probabilities, and changes in biophysical stability as detected by structure-based approaches all point to the same interpretation of positive selection to obtain and maintain a higher Mb stability for the whales. As a further support for the link, G5A, G35S, and G129A mutations have been observed in more stable Mbs in comparative studies (Scott et al., 2000).

Figure 4.9B shows *dN/dS* values for the variable sites in the cetacean clade versus the inferred average $\Delta\Delta G$ of the mutations within the cetacean clade. Four of the residues detected to be under positive selection (i.e., residues 5, 35, 66, and 121) show an effect on folding stability > 0.5 kcal/mol, with 5 and 66 being most significant, both towards stabilization (~0.7 and ~1.0 kcal/mol). The G129A substitution in the first branch leading to cetaceans (see Figure 4.9A), is stabilizing (i.e., $\Delta\Delta G$ = -0.69 kcal/mol), but undergoes three inversions from Ala to Gly in the branches leading to sperm whales, beaked whales and the suborder of *Delphinidae*, which makes it net destabilizing when summing over occurrences, although this is less significant and could reflect a partial relaxation of stability selection.

Figure 4.9B and 4.9C show an interesting feature of the evolutionary dynamics of protein stability. As was recently shown by relating protein stability (i.e., $\Delta G$) and evolution rate (i.e., *dN/dS*), proteins may evolve to a stability regime having a detailed balance between stabilizing and destabilizing mutations (Serohijos et al., 2012). Without the stability effects of sites detected to

be under positive selection, the $\Delta\Delta G$ distribution of mutations is nearly symmetric with an average mutation having $\Delta\Delta G$ = 0.1 kcal/mol. The average $\Delta\Delta G$ of an arising mutation in Mb is estimated to be ~1.2 kcal/mol (Tokuriki et al., 2007). These values suggest a balance between stabilizing and destabilizing mutations in the late branches of the cetacean clade (i.e., a pendulum-like behavior).

This balance can be shifted by positive selection of fixating stabilizing mutations such as G5A, G35S, S35H, N66V, N66H, N66I, G121S and G129A in the cetacean Mbs which provides a further stabilization of -1.7 kcal/mol for the whole clade and -4.4 kcal/mol when the branches leading to harbor porpoise and common minke whale are removed. Mbs in these animals have $\Delta G$ similar to that of terrestrials confirmed both from experimental mutagenesis and stability measurements and from the FoldX computations. The role of positive selection is also reflected in the probability of stabilization (i.e., $\Delta\Delta G < 0$ kcal/mol) conditional of positive selection, $pr(\Delta\Delta G < 0 \mid \omega > 1)$, using the Bayes rule (Bayes 1763), being ~0.80 (see Appendix C for details).

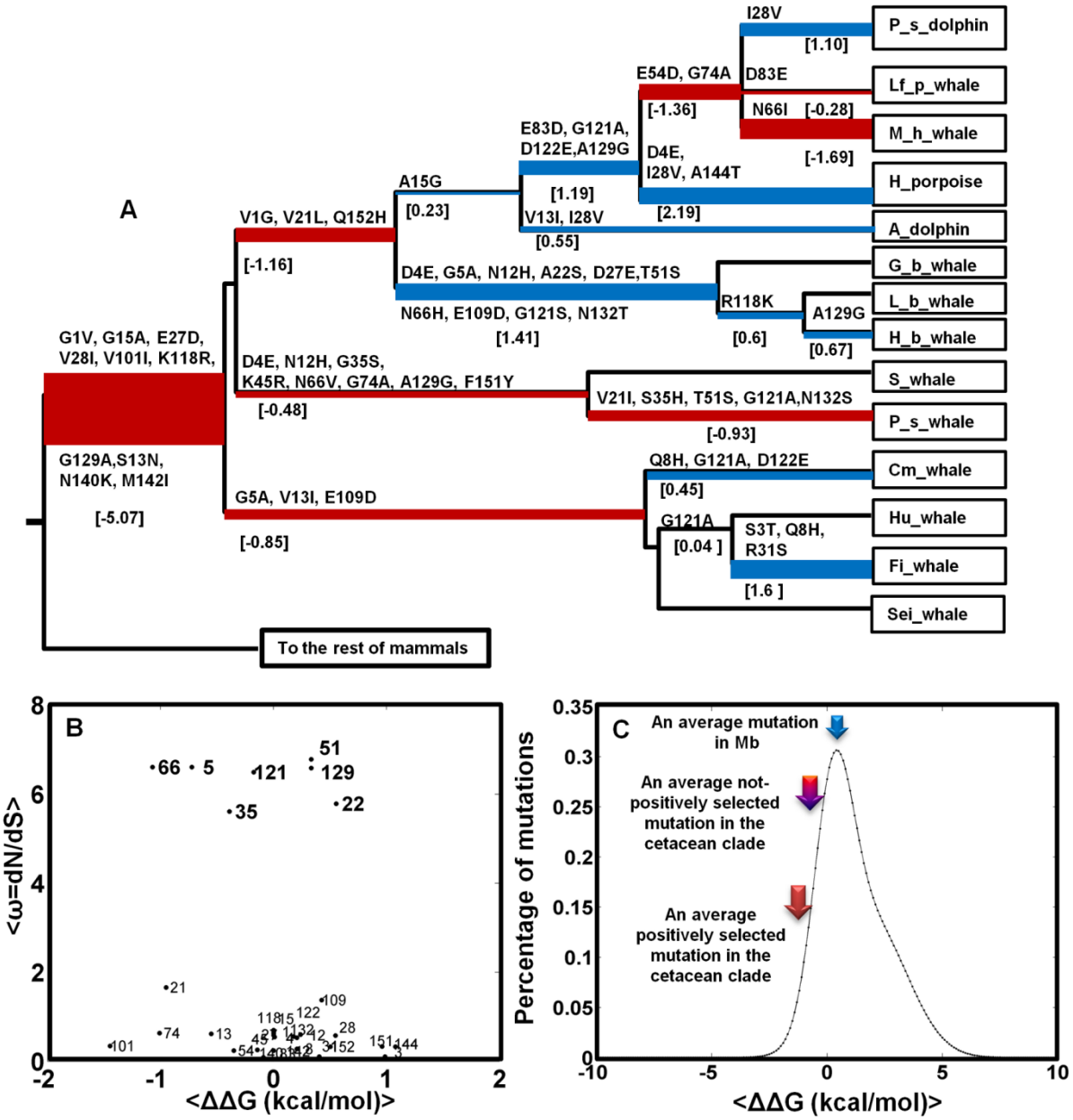



**Figure 4.9. A) The Phylogenetic tree of cetacean Mb upon the divergence from terrestrial counterparts.** Ancestral states were inferred using the maximum likelihood (ML) approach described in section 2.7 (Tamura et al., 2011). Amino acid changes in each branch are shown with the respective changes in free energy of folding, $\Delta\Delta G$ in kcal/mol calculated from the FoldX force field (Schymkowitz et al., 2005). Stabilization and destabilization is presented by red and blue colors respectively across the phylogeny, with branch height proportional to $|\Delta\Delta G|$ of that specific branch. B) The average $\omega = dN/dS$ for the variable sites in A from the M8 model is plotted versus the average $\Delta\Delta G$ of mutations in these sites. C) The distribution of mutational effects in Mb from (Tokuriki et al., 2007) is shown with the solid black line where arrows show the average $\Delta\Delta G$ for an average mutation in Mb (~1.22 kcal/mol), in the cetacean clade among not-positively selected mutations (~0.06 kcal/mol) and, among the positively selected residues (~−0.26 kcal/mol). The probability of stabilization caused by positive selection is ~0.8 (Dasmeh et al., 2013b).

### 4.3. Influences of selection for thermodynamic stability on evolution rate

Observing positive selection in cetacean Mbs prompted us to look into the role of selection for thermodynamic stability and its influences on the rate of protein evolution (i.e., *dN/dS*). As was shown by Tokuriki et al., (Tokuriki et al., 2007), and in an agreement with experimental mutagenesis experiments (Kumar et al., 2006), an average mutation in an average protein is ~ +1 kcal/mol destabilizing. Given the marginal stability of proteins which is ~5-10 kcal/mol, most proteins are prone to misfolding and thus fitness decline by ~5-10 mutations (Tokuriki and Tawfik, 2009). To prevent this in evolution, protein stability should be maintained by selection of compensatory stabilizing mutations. Although the existence of these compensatory mutations has been addressed in previous works, their effect on the rate of protein evolution has not been systematically studied.

From a molecular biophysics perspective, the folding stability (folding free energy, i.e., $\Delta G$) is one of the major determinants of sequence evolution and maintenance of stability imposes a selection pressure for many proteins (Dokholyan and Shakhnovich, 2001; Taverna and Goldstein, a2002; Taverna and Goldstein, b2002; Pal et al., 2006; Zeldovich et al., 2007; Goldstein, 2008; Chen and Dokholyan, 2008; Goldstein, 2011). Here we study the large scale evolution of simulated Mb sequences by imposing the selection pressure for folding stability and thus investigate this problem in more details.

We are now interested in estimating *dN/dS* among different branches of the simulated Mb phylogeny. For each branch, the change in nucleotides results in synonymous or nonsynonymous mutations. Nonsynonymous mutations are either purged or fixed in the population depending on the magnitude of $P_{fix}$ defined in equation 18 in section 2.10. We then use the codon models and ML estimation implemented in CODEML (Yang, 2007) to compute *dN/dS* (i.e., $\omega_{ML}$)

for the sequences resulting from simulation. We also define the *dN/dS* from simulations as $\omega_{pop}$ by counting the number of synonymous and nonsynonymous substitutions and normalizing by the number of synonymous and nonsynonymous sites using the sequence information in the simulation trajectories.

The top and right panels in Figure 4.10 show the histograms of $\omega_{ML}$ and $\Delta G$ for 20887 branches of 12 independent phylogenetic trees all having the same initial ancestral Mb sequence with $\Delta G$ = -6.84 kcal/mol (i.e., the expected value of $\Delta G$ after fitness equilibration, see chapter two for details). Bifurcations happen after $10^5$ mutational attempts (i.e., the $\lambda$ parameter) in the Mb sequence, which corresponds to ~5 fixed amino acid mutations. From the figure, most branches are under purifying selection having $\omega_{ML} < 1$ with an average of 0.55 and standard deviation of 0.51. However, there is a finite probability of observing $\omega_{ML} > 1$. Overall, 3035 out of 20887 branches have an elevated rate of nonsynonymous vs. synonymous substitutions. Folding stabilities span from ~ -4 kcal/mol to ~ -10 kcal/mol with an average of -6.34 kcal/mol and a standard deviation of 0.83 kcal/mol. The obtained skewed distribution of $\Delta G$ is in agreement with the empirical distribution of folding stabilities of proteins from the Protherm database (Kumar et al., 2006).

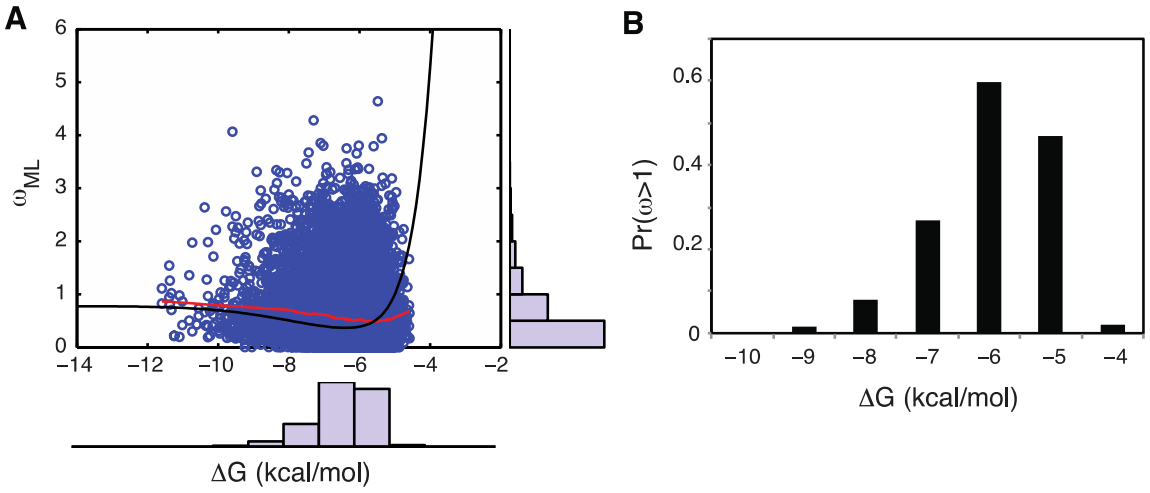



**Figure 4.10.** A) Top panel: distribution of ω inferred by maximum likelihood estimation for 20887 branches of 12 independent phylogenetic trees. Right panel: Folding stabilities of internal branches of the phylogenies. Main figure: The analytical molecular clock curve (in black) (Serohijos et al., 2012) with the scattered $\Delta G$ and $\omega_{ML}$ from the internal nodes of simulations (in red). The locally weighted scatterplot smoothing (LOWESS) line is shown in blue. B) The probability of observing $\omega_{ML} > 1$, $pr(\omega_{ML} > 1)$, at different folding stabilities (Dasmeh et al., 2013c).

The central panel in Figure 4.10A shows the scatter plot of folding stabilities, $\Delta G$ vs. $\omega_{ML}$. Each point corresponds to an ancestral $\Delta G$ with the $\omega_{ML}$ value calculated to its closest extant sequence in 12 simulated phylogenetic trees all having the same initial ancestral Mb sequence with $\Delta G$ = -6.84 kcal/mol, and all the bifurcations happens with $\lambda = 10^5$ mutations in the Mb sequence. Here, the molecular clock curve[4], which is an average of all arising mutations, is shown in black and the blue line shows the locally weighted scatter plot smoothing (LOWESS). As shown in the figure, there is a higher probability of observing $\omega_{ML} > 1$ at lower stabilities. Within the phylogeny, Mb spends much of its time under purifying selection (i.e., $\omega_{ML} < 1$) while traversing to very high and low stabilities with lower probabilities as reflected in $\omega_{ML} \sim 1$ and $\omega_{ML} > 1$ respectively. From Figure 4.10B, the probability of observing $\omega_{ML} > 1$ is increased at lower stabilities up to its maximum at $\Delta G \sim$ -6 kcal/mol where $\Delta G$ has its most probable value. Overall, $pr(\omega_{ML} > 1)$ is ~ 0.15 for Mb evolving under selection for stability with the described evolutionary dynamics. While

[4] For the derivation of "molecular clock curve" please refer to Serohijos *et al.*, 2012.

the molecular clock is expected to tick fastest at the least stable regime (Serohijos et al., 2012), the probability of observing $\omega_{ML} > 1$ decreases because the probability density (i.e., distribution function of $\Delta G$) approaches to 0 at $\Delta G = 0$ kcal/mol (Goldstein 2011; Bloom et al., 2007; Bloom et al., 2005; Zeldovich et al., 2007).

The statistical significance of the observation of $dN/dS > 1$ is usually inferred by using the likelihood ratio test (LRT) (Yang 1998). In a LRT test, twice the log-likelihood difference between two nested models has a $\chi^2$-distribution with a number of degrees of freedom equal to the difference in the number of free parameters in both models (Whelan and Goldman, 1999). We have evaluated the statistical significance of *dN/dS* by comparing two models when $\omega_{ML}$ is set to 1 and is left to vary (Yang 1998). For simulated sequences, the cases where $\omega_{ML} > 1$ are not statistically significant as judged by the LRT (see Appendix D, Figure D3). However, it has been shown that proteins with the whole gene-*dN/dS* values in the range of ~0.25 can have signatures of positive selection in specific parts of the sequence (Swanson et al., 2004; Sawyer and Malik, 2006).

To test for positive selection among the different residues in the simulated Mb sequence, we used the codon-based models of nucleotide substitutions to estimate the rate of nonsynonymous to synonymous mutations, *dN/dS*, across different sites (i.e., site models). For an evolving Mb sequence with $\lambda = 10^5$ mutational attempts, three pair-models as M1-M2, M7-M8 and M8fix-M8 are employed to identify sites under positive selection as presented in Table 4.3 (see chapter 2 for details). As shown in Table 4.3, the LRT gives a significant result, with six sites detected to be under positive selection having high posterior probabilities in the Bayes Emprical Bayes test (BEB) (Yang et al., 2005).

**Table 4.3.** Log likelihood values of the site models with detected sites under positive selection (Dasmeh et al., 2013c).

| Phylogeny | Models (number of parameters) | ln L | 2Δl | P value | Positively selected sites (BEB: $Pr(\omega > 1) > 0.5$)[a] [$\omega_{ML}$] |
|---|---|---|---|---|---|
| Simulated ($\lambda=10^5$) | M1a (2) | -65183.82 | - | - | - |
| | M2a (4) | -65141.86 | (M1a vs. M2a) 83.92 | $<10^{-16}$ | 34 [1.47], **48** [1.49], **59** [1.50], **119** [1.49], **133** [1.50], **139** [1.50] |
| | M7 (2) | -64591.18 | - | - | - |
| | M8 (4) | -64563.17 | (M7 vs. M8) 56.02 | $6.84\times10^{-13}$ | 48[1.32], **59** [1.50], **119** [1.48], **133** [1.50], **139** [1.50] |
| | M8fix (3) | -64586.49 | (M8 vs. M8fix) 46.64 | $8.53\times10^{-12}$ | - |

a: $Pr(\omega_{ML} > 1) > 0.95$ is shown in bold.

To investigate the reproducibility of the results, we have analyzed ten different phylogenetic trees with evolving Mb sequence with $\lambda = 10^5$. LRT was significant in all cases and different sites were detected to be under positive selection (see Appendix C for details). As is presented in Table 4.3, the maximum $\omega_{ML}$ for the sites under positive selection is 1.5 pointing to a weak yet significant elevated rate of nonsynonymous to synonymous mutations in these positions. This condition imposes a lower bound for the observation of positive selection of $dN/dS \sim 1.5$. Any signal of positive selection, judged by the significant LRT at specific residues with $dN/dS \sim 1.5$ can thus be due to maintaining stability rather than gaining new functions, in particular if mutations occur far from the active site.

To explore the sensitivity of our method to larger population sizes, we simulated also a phylogenetic tree with 1024 extant sequences and $N_{eff} = 10^5$. The average and the variance of $dN/dS$ was 0.51 and 0.22, respectively, with the larger population size (i.e., $N_{eff} = 10^5$), significantly smaller than 0.55 and 0.26 at $N_{eff} = 10^4$ (two sample t-test at the significance level of 0.05). Furthermore, $P(\omega_{ML} > 1)$ was slightly higher at the smaller population size with 0.14 and 0.13 for $N_{eff} = 10^4$ and $N_{eff} = 10^5$, respectively. With the larger population size, the average $\Delta G$ decreased to ~ -7.66 kcal/mol, consistent with previous studies on the relation between population size and the strength of selection for folding stability (Goldstein, 2011; Wylie and Shakhnovich, 2011). This

effect is mainly caused by the fact that in smaller populations, deleterious mutations have a higher chance of fixation. Therefore, on average, proteins have lower $\Delta G$ at $N_{eff} = 10^5$. Since proteins are more stable at this condition, we observed a lower probability of $\omega_{ML} > 1$.

# CHAPTER 5: CONCLUDING REMARKS

In previous chapters of this thesis, I described an integration of chemistry, biophysics, physiology and molecular phylogenetics to address some aspects of Mb evolution in mammalian species. We showed via an elaborated model of muscle tissue that the fixation of the distal pocket in mammalian Mb in general and specially His-64 occurs under hypoxic conditions, where WT is significantly more proficient. This result is further supported by observing specific physiological and behavioral conditions (i.e., cardiac output and muscle metabolic rates at routine dive conditions) where WT Mb contributes significantly to dive time compared to highly impaired mutants. The conservation of specific residues in the heme pocket of mammalian Mbs can be explained and justified by this approach.

Furthermore, we proposed an alternative hypothesis for the increased stability of deep-diving cetacean Mbs compared to their terrestrial counterparts. We provided evidence in an important real case that folding stability could be selected for in response to the selection for higher abundances due to $O_2$-consumption demands. To our knowledge, this is the first time that positive selection is reported for protein folding stability rather than biochemical functions.

Another important result of this thesis is that selection for folding stability is strong enough to influence the rate of protein evolution (i.e., *dN/dS*). For proteins at the precipices of unfolding, stabilizing mutations are fixed at a higher rate which leaves an imprint on the coding sequences through a significantly elevated rate of nonsynonymous mutations compared to synonymous ones. For Mb sequences evolved under a realistic selection pressure for maintaining folding stability, *dN/dS* was ~1.5 for different residues across Mb sequences.

# CHAPTER 6: FALSIFIABLE PREDICTIONS OF THIS THESIS

Besides the main results presented in previous chapters, we also propose two main theoretical predictions that are easy to falsify by future experimental studies.

First, Mb mutants with severely impaired oxygen storage would have a higher transport proficiency compared to WT protein at normoxic conditions. This finding is justified by the non-linear form of saturation expression (i.e., the Hill equation) and the fact that low-affinity mutants are in fact better $O_2$-transporters because they still have empty sites for $O_2$, giving rise to a larger [$MbO_2$] gradient (more varying saturation curve). We anticipate that it will be possible to verify this theoretical finding in a steady-state $O_2$-flow set up where the active $O_2$-diffusion of His-64 mutants vs. WT is measured. The best proof-of-concept for this prediction comes from comparing $O_2$ affinity of Hb and Mb. Hb has a lower affinity for $O_2$ and thus physiologically is a better transport protein.

Second, similar to whale species, seal Mbs should also be more stable compared to terrestrial mammalian Mbs as they have experienced the same convergent evolution to aquatic environment. In fact it is shown experimentally that harbor seal Mb is as stable as sperm whale Mb (Puett et al., 1973). However, more stability measurements are needed to verify the significance of this prediction. If misfolding prevention is the underlying cause for the selection of higher stabilities in whale Mbs (as we claimed in this thesis) the same applies to seal Mbs.

If any of the above predictions fail to come true, the theories presented in this thesis to support these arguments need minor or major revisions as *nature cannot be fooled*.[5]

[5] From the commentary of Richard P. Feynman in the final report of the commission for investigation of the space shuttle Challenger disaster (Feynman, 1986).

## APPENDIX A: DEPENDENCE OF MB FUNCTIONAL PROFICIENCIES ON RELEVANT PARAMETERS

In the following diagrams, the dependence of OSR and OTR on the relevant parameters of the theoretical model is investigated. It is more illustrative to compare the functional proficiency of WT and the mutants with respect to the variations in these parameters.

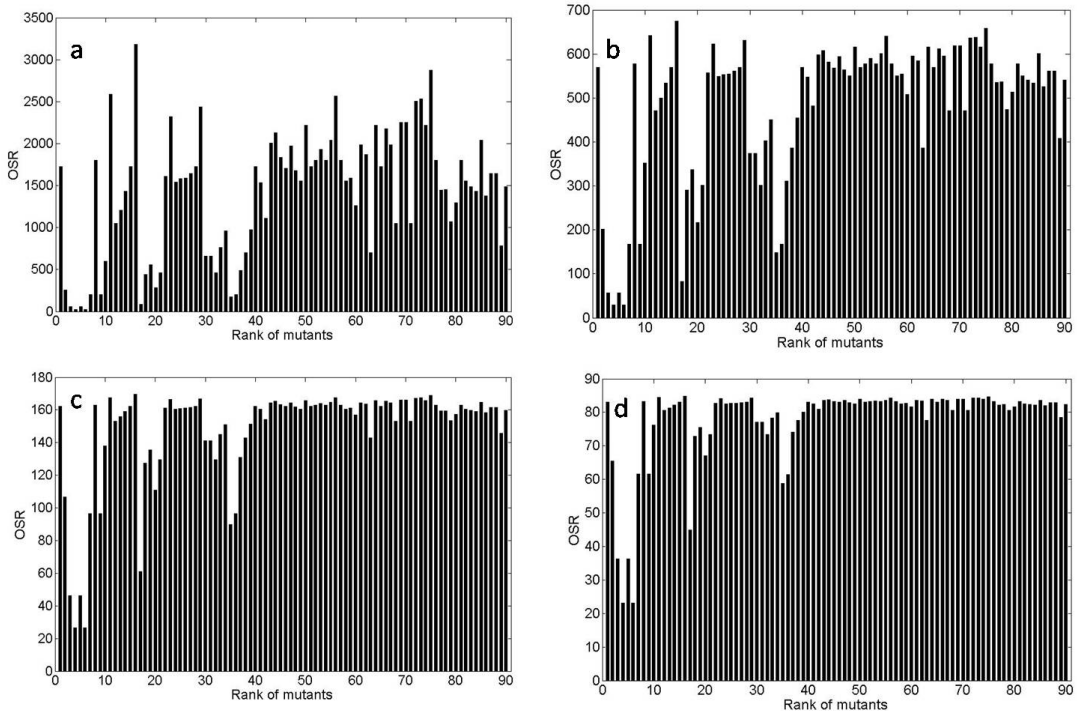



**Figure A1**. The OSR values at upper pressure a) $P_u$ = 1 mmHg; b) $P_u$ = 5 mmHg; c) $P_u$ = 20 mmHg; and d) $P_u$ = 40 mmHg. For all charts, $C_{Mb}$ = 3.1 $m\ mol\ L^{-1}$, $R_k$ = 19 µm, and $\dot{V}_{O_2} = 0.3\ \mu\ mol\ L^{-1}s^{-1}$.

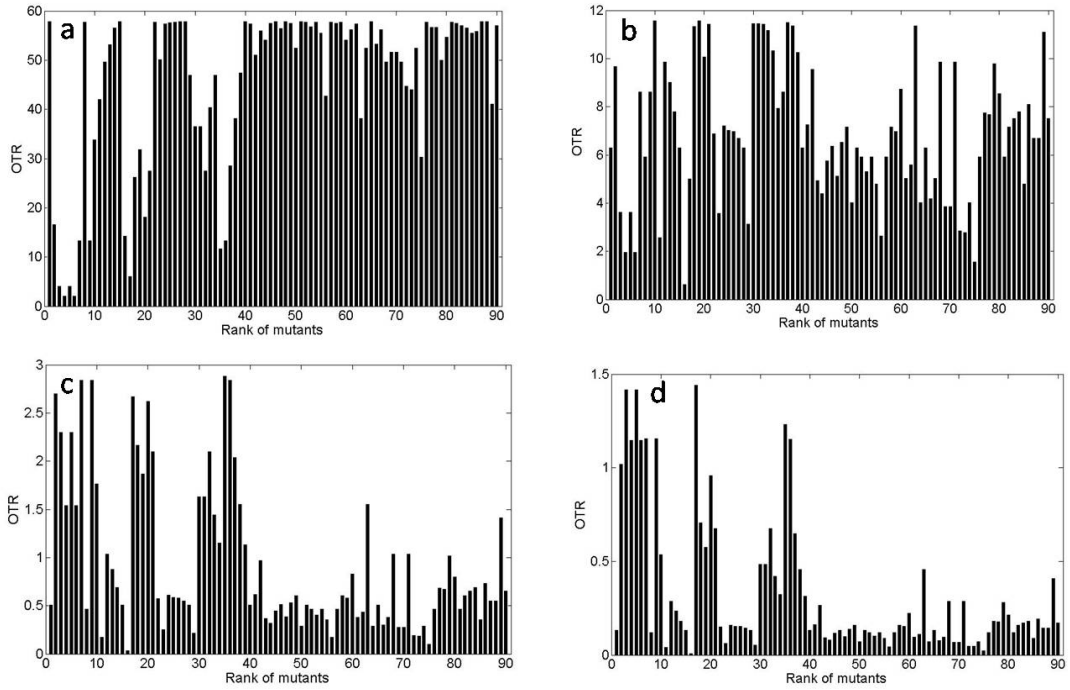



**Figure A2**. The OTR values at different upper pressure a) $P_u$ = 1 mmHg; b) $P_u$ = 5 mmHg; c) $P_u$ = 20 mmHg; and d) $P_u$ = 40 mmHg. For all charts, $C_{Mb}$ = 3.1 $m\ mol\ L^{-1}$, $R_k$ = 19 µm, and $\dot{V}_{O_2}$ = 0.3 $\mu\ mol\ L^{-1}s^{-1}$ .

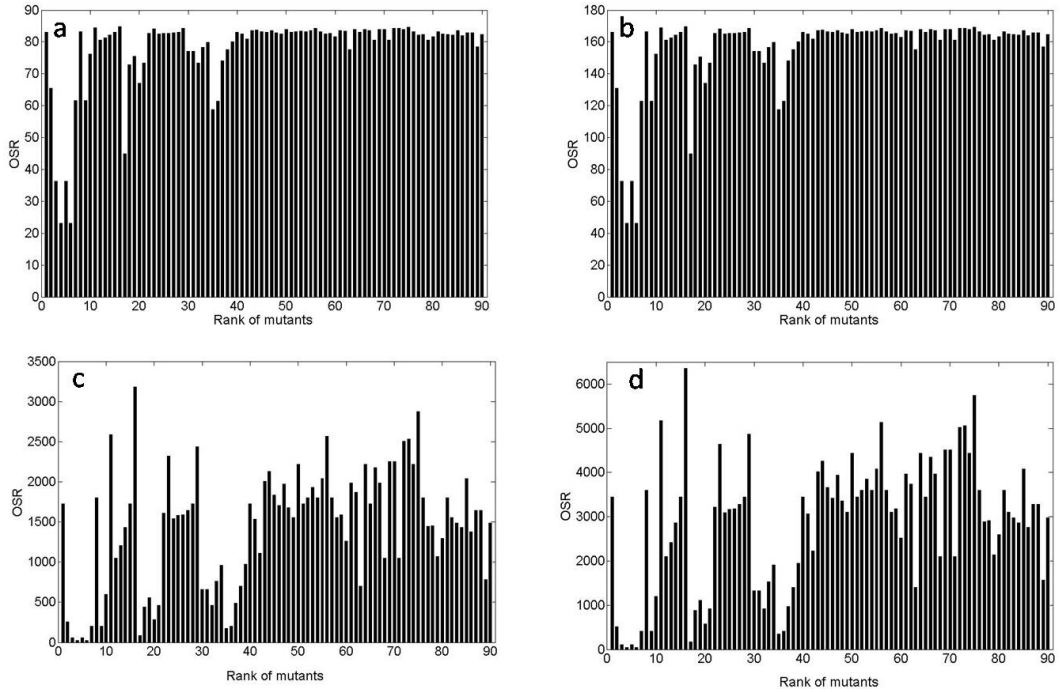



**Figure A3**. The OSR values at a) $P_u$ = 40 mmHg and $C_{Mb}$ = 3.1 $mol\ L^{-1}$; b) $P_u$ = 40 mmHg and $C_{Mb}$ = 6.2 $m\ mol\ L^{-1}$; c) $P_u$ = 5 mmHg and $C_{Mb}$ = 3.1 mM; and d) $P_u$ = 5 mmHg and $C_{Mb}$ = 6.2 $m\ mol\ L^{-1}$. For all charts $R_k$ = 19 μm and $\dot{V}_{O_2} = 0.3\ \mu\ mol\ L^{-1}s^{-1}$ .

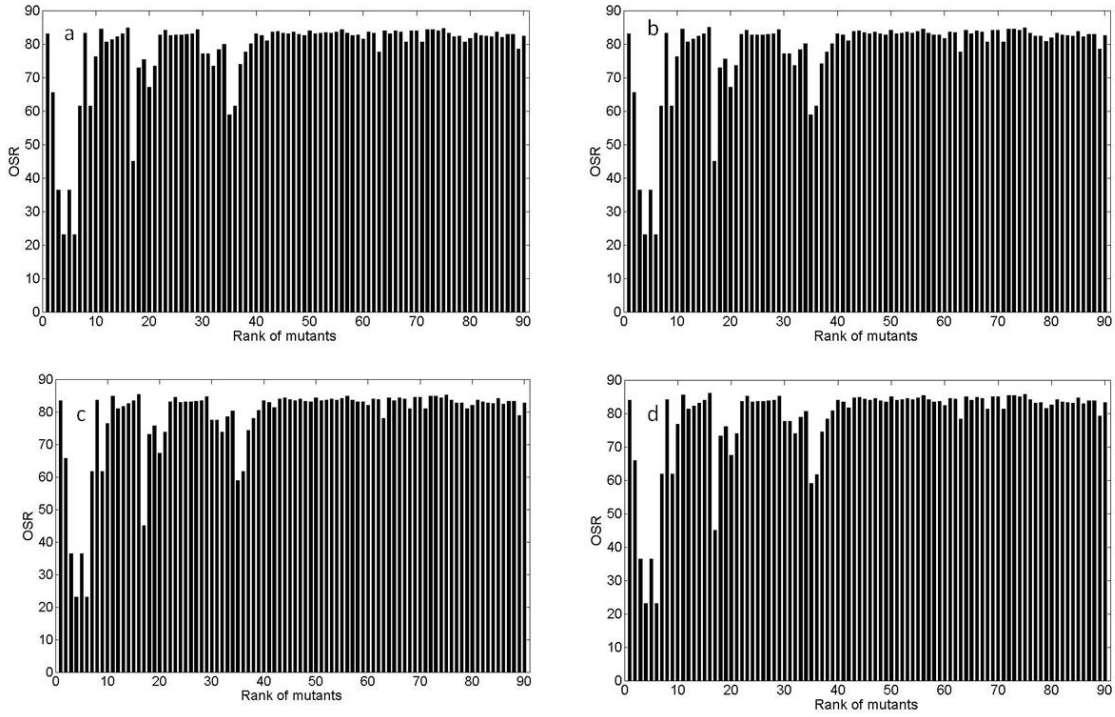



**Figure A4**. The OSR values at a) $\dot{V}_{O2} = 0.3\,\mu\,mol\,L^{-1}s^{-1}$; b) $\dot{V}_{O2} = 0.8\,\mu\,mol\,L^{-1}s^{-1}$; c) $\dot{V}_{O2} = 2\,\mu\,mol\,L^{-1}s^{-1}$; and d) $\dot{V}_{O2} = 4\,\mu\,mol\,L^{-1}s^{-1}$. For all charts, $P_u$= 40 mmHg, $C_{Mb}$= 3.1 $mol\,L^{-1}$, and $R_k$= 19μm.

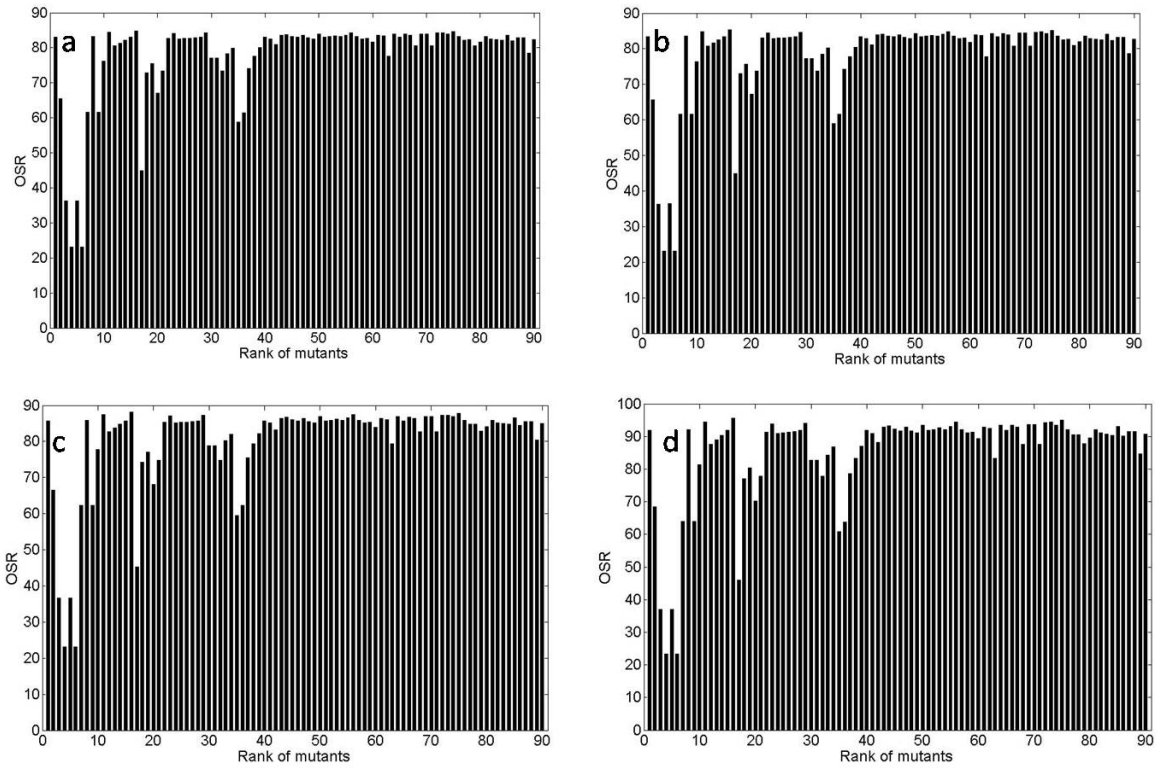



**Figure A5**. The OSR values having a) $R_k = 19\,\mu m$; b) $R_k = 34\,\mu m$; c) $R_k = 64\,\mu m$; and d) $R_k = 100\,\mu m$. For all charts, $C_{Mb}$ = 3.1 $m\,mol\,L^{-1}$, $P_u$ = 40 mmHg, and $\dot{V}_{O_2} = 0.3\,\mu\,mol\,L^{-1}s^{-1}$.

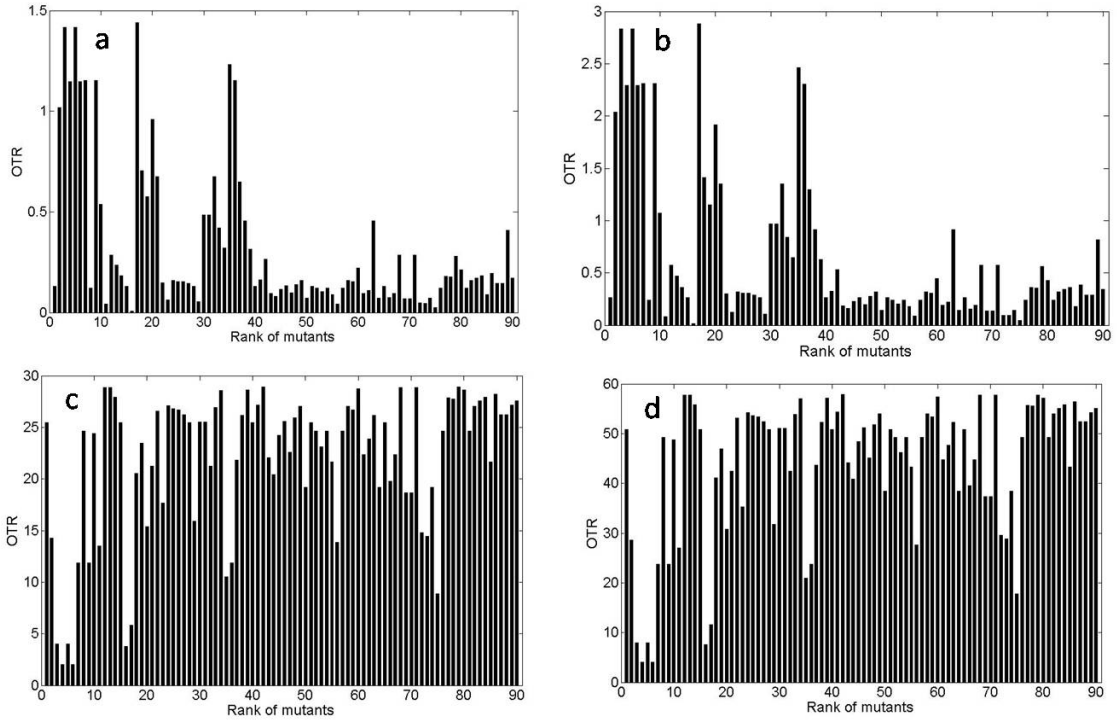



**Figure A6**. The OTR values at a) $P_u$ = 40 mmHg and $C_{Mb}$ = 3.1 $mol\ L^{-1}$; b) $P_u$ = 40 mmHg and $C_{Mb}$ = 6.2 $m\ mol\ L^{-1}$; c) $P_u$ = 5 mmHg and $C_{Mb}$ = 3.1 $m\ mol\ L^{-1}$; and d) $P_u$ = 5 mmHg and $C_{Mb}$ = 6.2 $m\ mol\ L^{-1}$. For all charts, $R_k$ = 19 μm and $\dot{V}_{O_2} = 0.3\ \mu\ mol\ L^{-1}s^{-1}$ .

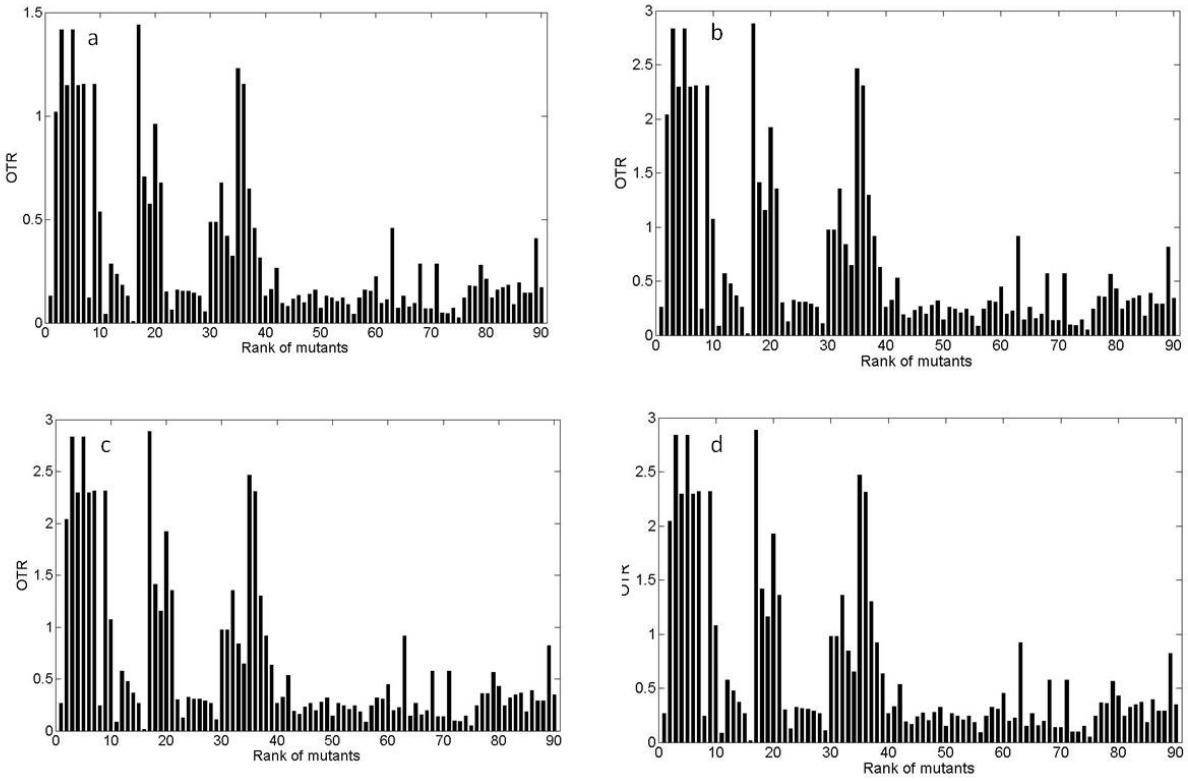



**Figure A7**. The OTR values at a) $\dot{V}_{O2} = 0.3\ \mu\ mol\ L^{-1}s^{-1}$; b) $\dot{V}_{O2} = 0.8\ \mu\ mol\ L^{-1}s^{-1}$; c) $\dot{V}_{O2} = 2\ \mu\ mol\ L^{-1}s^{-1}$; and d) $\dot{V}_{O2} = 4\ \mu\ mol\ L^{-1}s^{-1}$. For all charts, $P_u$ = 40 mmHg, $C_{Mb}$ = 3.1 $m\ mol\ L^{-1}$, and $R_k$ = 19 μm.

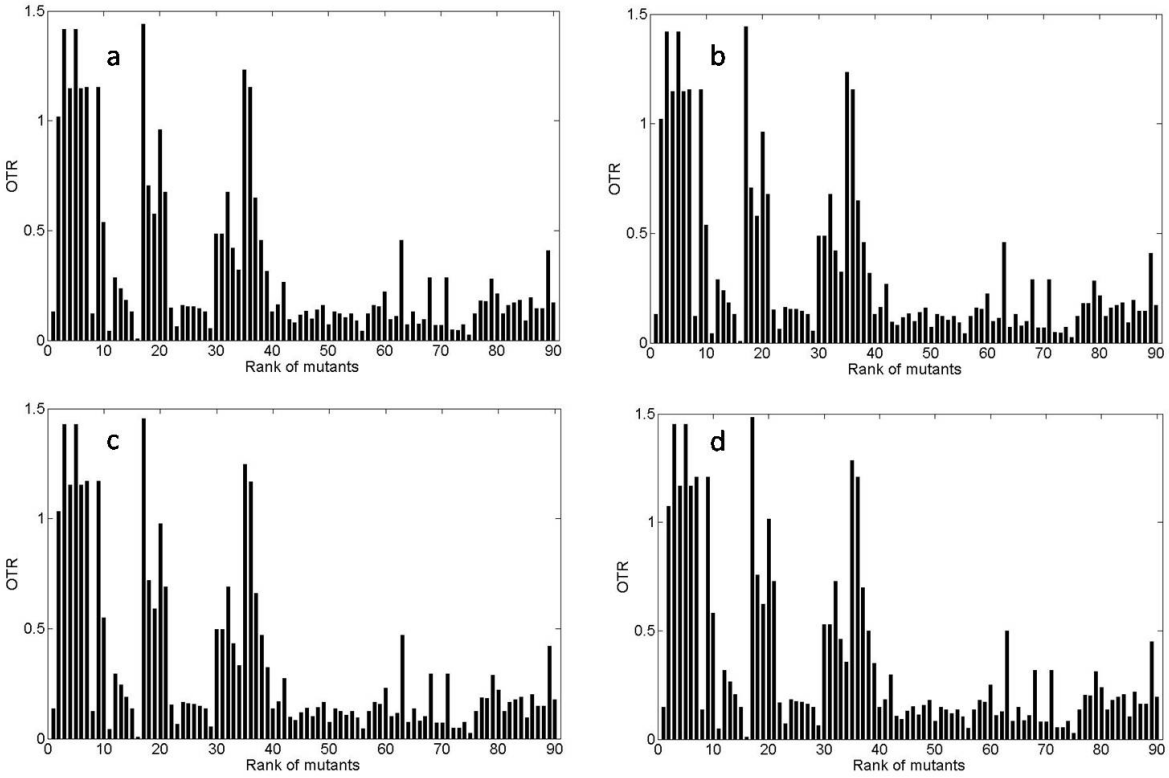



**Figure A8**. The OTR values having a) $R_k = 19\ \mu m$; b) $R_k = 34\ \mu m$; c) $R_k = 64\ \mu m$; and d) $R_k = 100\ \mu m$. For all charts, $C_{Mb}$ = 3.1 $m\ mol\ L^{-1}$, $P_u$ = 40 mmHg, and $\dot{V}_{O_2} = 0.3\ \mu\ mol\ L^{-1}s^{-1}$.

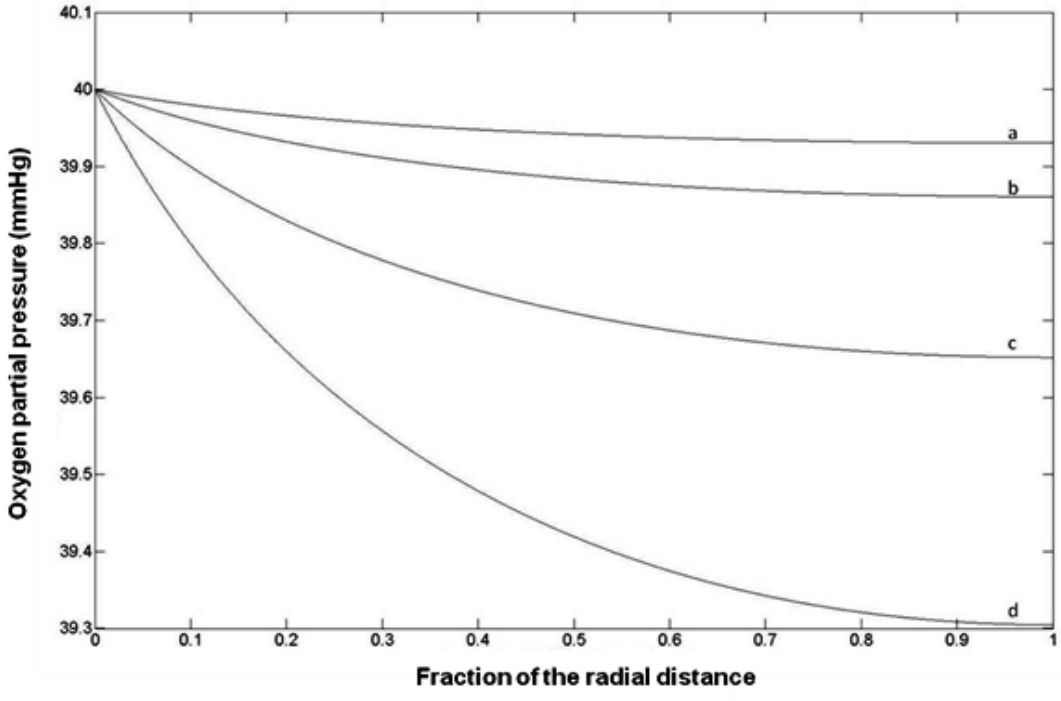



**Figure A9**. The $P_{O_2}(r)$ for a) $\dot{V}_{O2} = 0.3\ \mu\ mol\ L^{-1}s^{-1}$; b) $\dot{V}_{O2} = 0.8\ \mu\ mol\ L^{-1}s^{-1}$; c) $\dot{V}_{O2} = 2\ \mu\ mol\ L^{-1}s^{-1}$; and d) $\dot{V}_{O2} = 4\ \mu\ mol\ L^{-1}s^{-1}$. For all diagrams, $P_u$ = 40 mmHg, $C_{Mb}$ = 3.1 $m\ mol\ L^{-1}$, and $R_k$ = 19 μm.

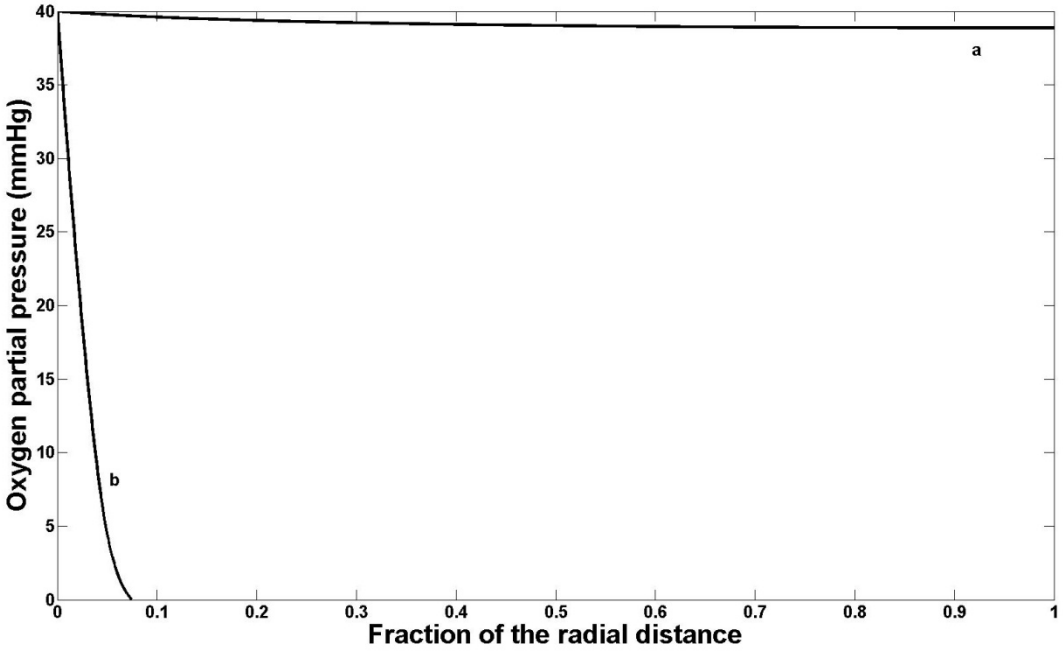



**Figure A10**. The $P_{O_2}(r)$ for a) Sperm whale at diving conditions: $\dot{V}_{O2} = 2\ \mu\ mol\ L^{-1}s^{-1}$; $P_u$ = 40 mmHg,; $C_{Mb}$ = 3.1 $m\ mol\ L^{-1}$; and $R_k$ = 19 μm; and b) Human skeletal muscle at maximum workload: $\dot{V}_{O2} = 201\ \mu\ mol\ L^{-1}s^{-1}$; $P_u$ = 40 mmHg; $C_{Mb}$ = 0.25 $m\ mol\ L^{-1}$; and $R_k$ = 19 μm.

## APPENDIX B: SUPPLEMENTARY DATA FOR ADL CALCULATIONS

**Table B1**. Thermodynamic constants and half-saturation pressure $P_{50}$ for wild type and mutant sperm whale Mbs. Each mutated site is ranked according to its distance to heme, classified as either first-shell, second-shell, proximal site, or far-from-heme residue. $K_{O2}$ values at $20^0C$ from Scott *et al*., 2003 were converted to $37^0C$ by the factor 0.2457 based on the temperature effect on P50 (Schenkman et.al, 1997).

| Rank | Mutant | $K_{O2}$ ($\mu M^{-1}$), $20^0C$ | $K_{O2}$ ($\mu M^{-1}$), $37^0C$ | $P_{50}$ (mmHg), $37^0C$ | Rank | Mutant | $K_{O2}$ ($\mu M^{-1}$), $20^0C$ | $K_{O2}$ ($\mu M^{-1}$), $37^0C$ | $P_{50}$ (mmHg), $37^0C$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | WT | 1.1 | 0.27 | 3.9 | 46 | V66R | 1.08 | 0.27 | 4.0 |
| 2 | H64G | 0.09 | 0.02 | 48.1 | 47 | T67F | 1.47 | 0.36 | 2.9 |
| 3 | H64A | 0.02 | 0.00 | 216.5 | 48 | T67K | 1.04 | 0.26 | 4.2 |
| 4 | H64V | 0.01 | 0.00 | 433.0 | 49 | T67Q | 0.9 | 0.22 | 4.8 |
| 5 | H64L | 0.02 | 0.00 | 216.5 | 50 | T67P | 2 | 0.49 | 2.2 |
| 6 | H64F | 0.01 | 0.00 | 433.0 | 51 | T67A | 1.1 | 0.27 | 3.9 |
| 7 | H64W | 0.07 | 0.02 | 61.9 | 52 | I107V | 1.2 | 0.29 | 3.6 |
| 8 | V68A | 1.2 | 0.29 | 3.6 | 53 | I107T | 1.4 | 0.34 | 3.1 |
| 9 | V68T | 0.07 | 0.02 | 61.9 | 54 | I107L | 1.2 | 0.29 | 3.6 |
| 10 | V68I | 0.23 | 0.06 | 18.8 | 55 | I107F | 1.6 | 0.39 | 2.7 |
| 11 | V68L | 3.4 | 0.84 | 1.3 | 56 | I107W | 3.3 | 0.81 | 1.3 |
| 12 | V68F | 0.48 | 0.12 | 9.0 | 57 | I107V | 1.2 | 0.29 | 3.6 |
| 13 | V68W | 0.59 | 0.14 | 7.3 | 58 | I111L | 0.9 | 0.22 | 4.8 |
| 14 | L29A | 0.78 | 0.19 | 5.6 | 59 | I111F | 0.94 | 0.23 | 4.6 |
| 15 | L29V | 1.1 | 0.27 | 3.9 | 60 | I111M | 0.63 | 0.15 | 6.9 |
| 16 | L29F | 15 | 3.69 | 0.3 | 61 | I111W | 1.5 | 0.37 | 2.9 |
| 17 | L29W | 0.03 | 0.01 | 144.3 | 62 | L89G | 1.3 | 0.32 | 3.3 |
| 18 | F43V | 0.16 | 0.04 | 27.1 | 63 | L89W | 0.28 | 0.07 | 15.5 |
| 19 | F43L | 0.21 | 0.05 | 20.6 | 64 | H97A | 2 | 0.49 | 2.2 |
| 20 | F43I | 0.1 | 0.02 | 43.3 | 65 | H97V | 1.1 | 0.27 | 3.9 |
| 21 | F43W | 0.17 | 0.04 | 25.5 | 66 | H97D | 1.9 | 0.47 | 2.3 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 22 | I28A | 0.96 | 0.24 | 4.5 | 67 | H97F | 1.5 | 0.37 | 2.9 |
| 23 | I28W | 2.3 | 0.57 | 1.9 | 68 | H97Q | 0.48 | 0.12 | 9.0 |
| 24 | L32A | 0.89 | 0.22 | 4.9 | 69 | I99A | 2.1 | 0.52 | 2.1 |
| 25 | L32V | 0.93 | 0.23 | 4.7 | 70 | I99V | 2.1 | 0.52 | 2.1 |
| 26 | L32I | 0.94 | 0.23 | 4.6 | 71 | I99L | 0.48 | 0.12 | 9.0 |
| 27 | L32F | 1 | 0.25 | 4.3 | 72 | I99N | 3 | 0.74 | 1.4 |
| 28 | L32M | 1.1 | 0.27 | 3.9 | 73 | L104A | 3.1 | 0.76 | 1.4 |
| 29 | L32W | 2.7 | 0.66 | 1.6 | 74 | L104V | 2 | 0.49 | 2.2 |
| 30 | R45A | 0.26 | 0.06 | 16.7 | 75 | L104W | 5.8 | 1.43 | 0.7 |
| 31 | R45L | 0.26 | 0.06 | 16.7 | 76 | F138A | 1.2 | 0.29 | 3.6 |
| 32 | R45T | 0.17 | 0.04 | 25.5 | 77 | F138W | 0.79 | 0.19 | 5.5 |
| 33 | R45K | 0.31 | 0.08 | 14.0 | 78 | W7F | 0.8 | 0.20 | 5.4 |
| 34 | R45S | 0.42 | 0.10 | 10.3 | 79 | Q8V | 0.49 | 0.12 | 8.8 |
| 35 | R45A | 0.06 | 0.01 | 72.2 | 80 | W14F | 0.66 | 0.16 | 6.6 |
| 36 | F46V | 0.07 | 0.02 | 61.9 | 81 | M55A | 1.2 | 0.29 | 3.6 |
| 37 | F46L | 0.18 | 0.04 | 24.1 | 82 | M55L | 0.9 | 0.22 | 4.8 |
| 38 | F46W | 0.28 | 0.07 | 15.5 | 83 | M55W | 0.83 | 0.20 | 5.2 |
| 39 | F46A | 0.43 | 0.11 | 10.1 | 84 | A71F | 0.78 | 0.19 | 5.6 |
| 40 | L61F | 1.1 | 0.27 | 3.9 | 85 | L72V | 1.6 | 0.39 | 2.7 |
| 41 | L61A | 0.88 | 0.22 | 4.9 | 86 | L72W | 0.73 | 0.18 | 5.9 |
| 42 | G65I | 0.52 | 0.13 | 8.3 | 87 | K79A | 1 | 0.25 | 4.3 |
| 43 | G65T | 1.54 | 0.38 | 2.8 | 88 | K79L | 1 | 0.25 | 4.3 |
| 44 | G65G | 1.79 | 0.44 | 2.4 | 89 | M131L | 0.32 | 0.08 | 13.5 |
| 45 | V66K | 1.25 | 0.31 | 3.5 | 90 | A144V | 0.83 | 0.20 | 5.2 |

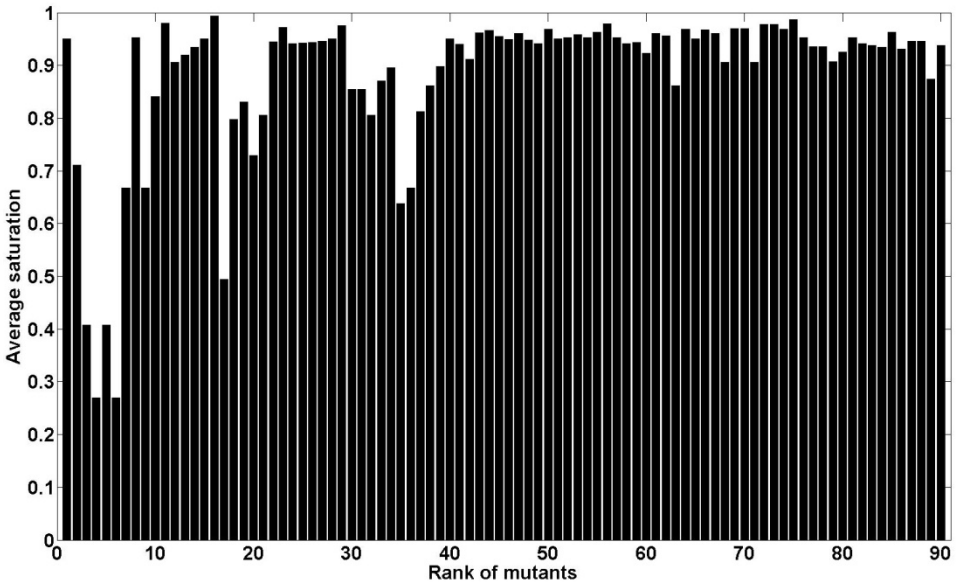



**Figure B1**. The average saturation of Mbs within the seal muscle cell in the pressure range, $P_{mit}$ = 0 mmHg to $P_c$=87 mmHg, at $20^0C$, using sperm whale mutant data directly.

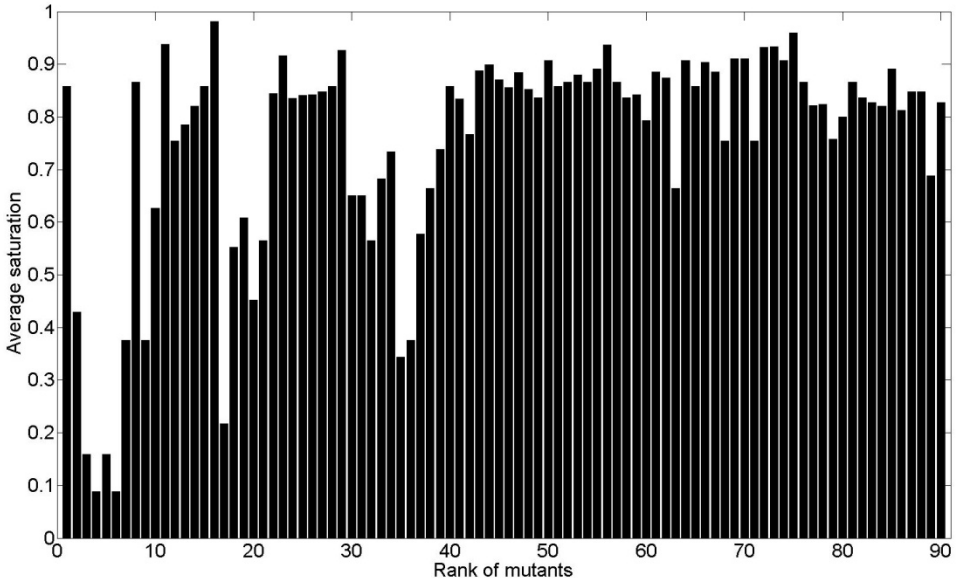



**Figure B2**. The average saturation of Mbs within the muscle cell in the pressure range, $P_{mit}$ = 0 mmHg to $P_c$=87 mmHg, at $37^0C$, using temperature-corrected sperm whale data.

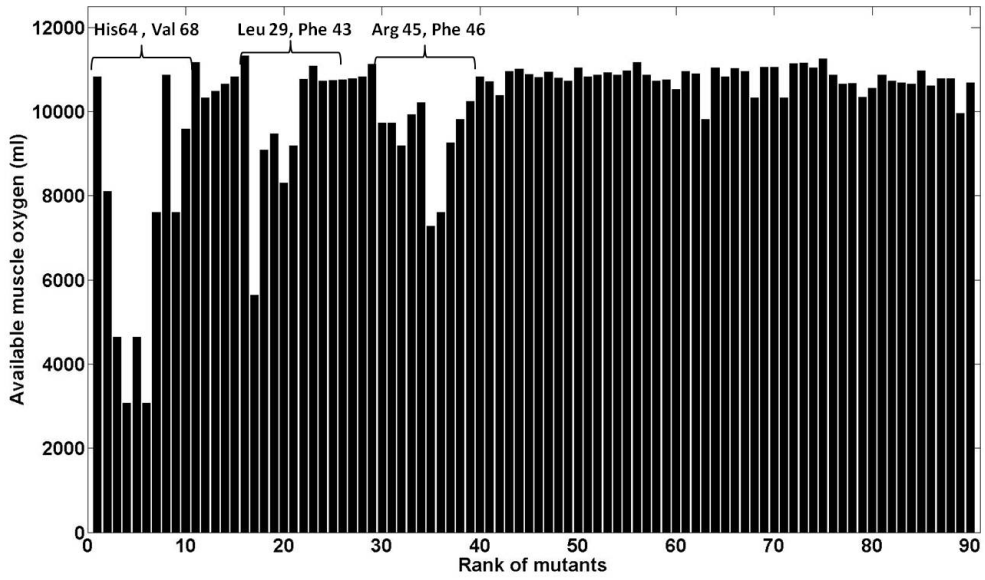



**Figure B3**. The amount of available $O_2$ (ml) at $20^0C$ in the muscle of Weddell seal mutants at the beginning of a dive, using whale mutant data directly.

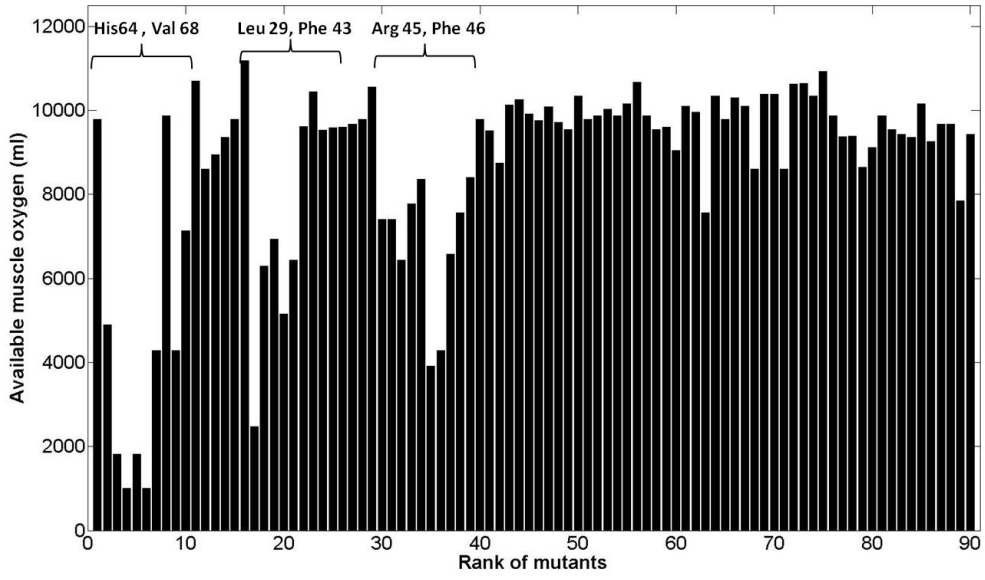



**Figure B4**. The amount of available $O_2$ (ml) at $37^0C$ in the muscle of Weddell seal mutants at the beginning of a dive, using whale mutant data directly.

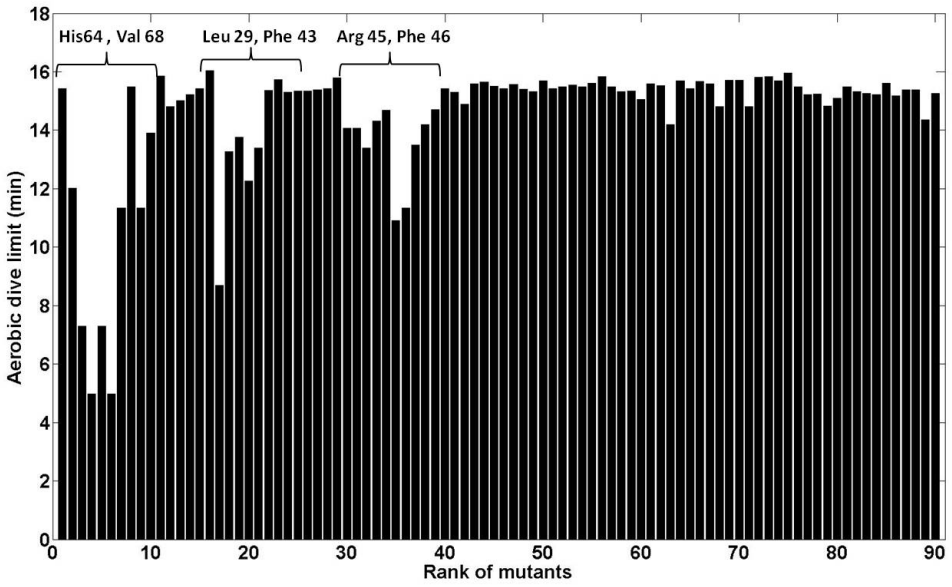



**Figure B5**. The calculated ADL (min) of the simulated dive of a Weddell seal with $\dot{V}_{MO2} = 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} = 0.19\,\dot{V}_{bO2}(rest)$ at $20^0C$ for different Mb mutants, using whale mutant $K_{O2}$ data directly.

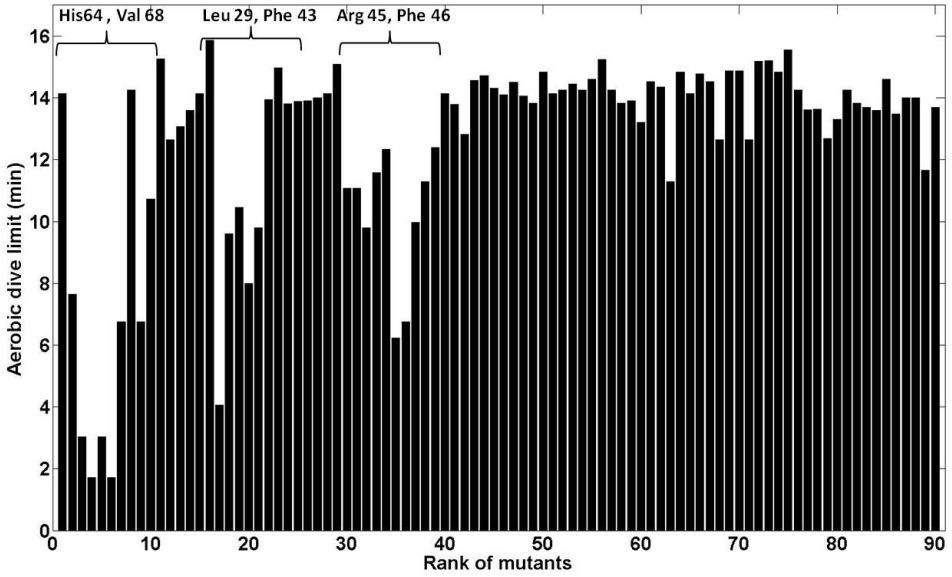



**Figure B6**. The calculated ADL (min) of the simulated dive of a Weddell seal with $\dot{V}_{MO2} = 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} = 0.19\dot{V}_{bO2}(rest)$ at $37^0C$ for different Mb mutants, using whale mutant $K_{O2}$ data directly.

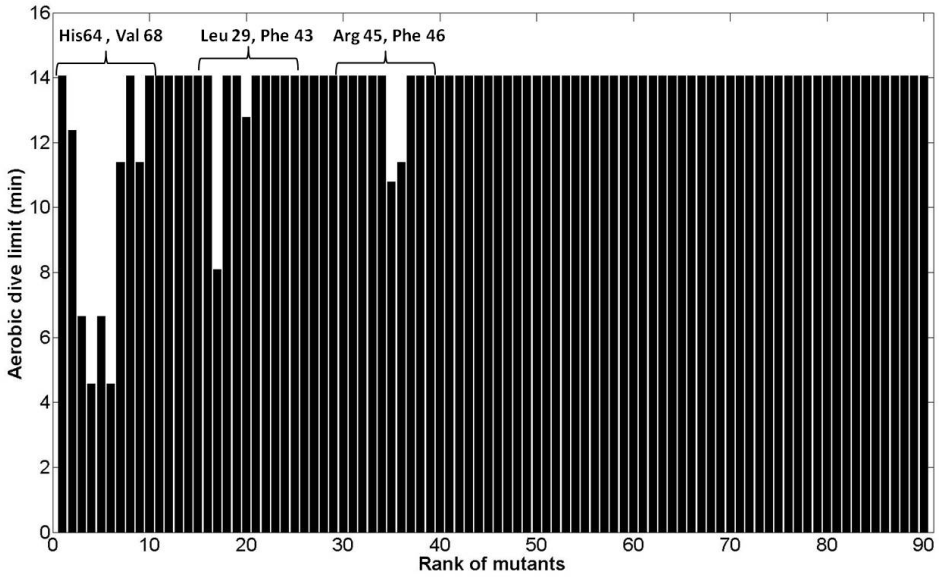



**Figure B7**. The calculated ADL (min) of the simulated dive of a Weddell seal with $\dot{V}_{MO2} = 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} = 0.37\dot{V}_{bO2}(rest)$ at $37^0C$ for different Mb mutants, using whale mutant $K_{O2}$ data directly.

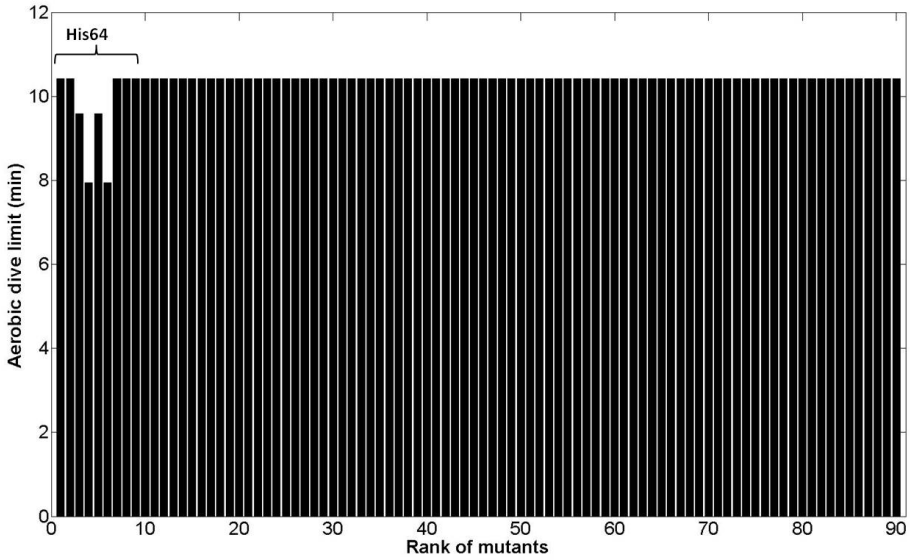



**Figure B8**. The calculated ADL (min) of the simulated dive of a Weddell seal with $\dot{V}_{MO2} = 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} = 0.55\dot{V}_{bO2}(rest)$ at $37^0C$ for different Mb mutants, using whale mutant $K_{O2}$ data directly.

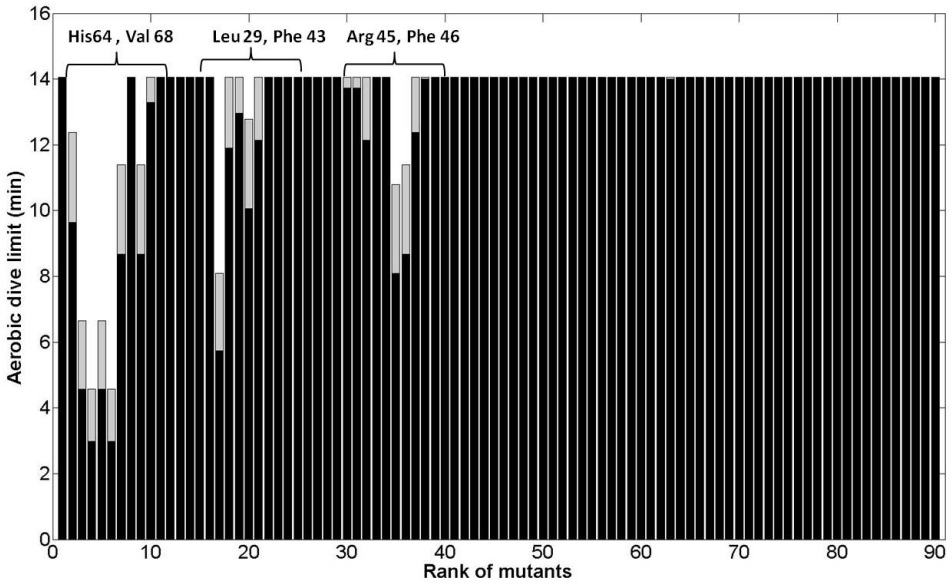



**Figure B9**. The calculated ADL (min) of the simulated dive of a Weddell seal with $\dot{V}_{MO2} = 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} = 0.37\dot{V}_{bO2}(rest)$ at $37^0C$ for different Mb mutants, using whale mutant $K_{O2}$ data directly at $P_a$ = 119 mmHg and $P_v$ = 55 mmHg (gray bars) and $P_a$ = 59.5 mmHg and $P_v$ =27.5 mmHg (black bars).

**Table B2**. Thermodynamic constants and half-saturation pressure $P_{50}$ for wild type and mutant seal Mbs. Each mutated site is ranked according to its distance to heme, classified as either first-shell, second-shell, proximal site, or far-from-heme residue. $K_{O2}$ values for seal mutants are converted from sperm whale data by the factor of $(\frac{K_{O2}(Seal\,WT\,Mb)}{K_{O2}(Whale\,WT\,Mb)} = 1.085)$.

| Rank | Mutant | $K_{O2}$ ($\mu M^{-1}$), $37^0C$ | $P_{50}$ (mmHg), $37^0C$ | Rank | Mutant | $K_{O2}$ ($\mu M^{-1}$), $37^0C$ | $P_{50}$ (mmHg), $37^0C$ |
|---|---|---|---|---|---|---|---|
| 1 | WT | 0.29 | 3.6 | 41 | T67A | 0.29 | 3.6 |
| 2 | H64G | 0.02 | 44.3 | 42 | I107V | 0.32 | 3.3 |
| 3 | H64A | 0.01 | 199.5 | 43 | I107T | 0.37 | 2.9 |
| 4 | H64V | 0.00 | 399.1 | 44 | I107L | 0.32 | 3.3 |
| 5 | H64L | 0.01 | 199.5 | 45 | I107F | 0.43 | 2.5 |
| 6 | H64F | 0.00 | 399.1 | 46 | I107W | 0.88 | 1.2 |
| 7 | H64W | 0.02 | 57.0 | 47 | I107V | 0.32 | 3.3 |
| 8 | V68A | 0.32 | 3.3 | 48 | I111L | 0.24 | 4.4 |
| 9 | V68T | 0.02 | 57.0 | 49 | I111F | 0.25 | 4.2 |
| 10 | V68I | 0.06 | 17.4 | 50 | I111M | 0.17 | 6.3 |
| 11 | V68L | 0.91 | 1.2 | 51 | I111W | 0.40 | 2.7 |
| 12 | V68F | 0.13 | 8.3 | 52 | L89G | 0.35 | 3.1 |
| 13 | V68W | 0.16 | 6.8 | 53 | L89W | 0.07 | 14.3 |
| 14 | L29A | 0.21 | 5.1 | 54 | H97A | 0.53 | 2.0 |
| 15 | L29V | 0.29 | 3.6 | 55 | H97V | 0.29 | 3.6 |
| 16 | L29F | 4.00 | 0.3 | 56 | H97D | 0.51 | 2.1 |
| 17 | L29W | 0.01 | 133.0 | 57 | H97F | 0.40 | 2.7 |
| 18 | F43V | 0.04 | 24.9 | 58 | H97Q | 0.13 | 8.3 |
| 19 | F43L | 0.06 | 19.0 | 59 | I99A | 0.56 | 1.9 |
| 20 | F43I | 0.03 | 39.9 | 60 | I99V | 0.56 | 1.9 |
| 21 | F43W | 0.05 | 23.5 | 61 | I99L | 0.13 | 8.3 |
| 22 | I28A | 0.26 | 4.2 | 62 | I99N | 0.80 | 1.3 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 23 | I28W | 0.61 | 1.7 | 63 | L104A | 0.83 | 1.3 |
| 24 | L32A | 0.24 | 4.5 | 64 | L104V | 0.53 | 2.0 |
| 25 | L32V | 0.25 | 4.3 | 65 | L104W | 1.55 | 0.7 |
| 26 | L32I | 0.25 | 4.2 | 66 | F138A | 0.32 | 3.3 |
| 27 | L32F | 0.27 | 4.0 | 67 | F138W | 0.21 | 5.1 |
| 28 | L32M | 0.29 | 3.6 | 68 | W7F | 0.21 | 5.0 |
| 29 | L32W | 0.72 | 1.5 | 69 | Q8V | 0.13 | 8.1 |
| 30 | L61F | 0.29 | 3.6 | 70 | W14F | 0.18 | 6.0 |
| 31 | L61A | 0.23 | 4.5 | 71 | M55A | 0.32 | 3.3 |
| 32 | G65I | 0.14 | 7.7 | 72 | M55L | 0.24 | 4.4 |
| 33 | G65T | 0.41 | 2.6 | 73 | M55W | 0.22 | 4.8 |
| 34 | G65G | 0.48 | 2.2 | 74 | A71F | 0.21 | 5.1 |
| 35 | V66K | 0.33 | 3.2 | 75 | L72V | 0.43 | 2.5 |
| 36 | V66R | 0.29 | 3.7 | 76 | L72W | 0.19 | 5.5 |
| 37 | T67F | 0.39 | 2.7 | 77 | K79A | 0.27 | 4.0 |
| 38 | T67K | 0.28 | 3.8 | 78 | K79L | 0.27 | 4.0 |
| 39 | T67Q | 0.24 | 4.4 | 79 | M131L | 0.09 | 12.5 |
| 40 | T67P | 0.53 | 2.0 | 80 | A144V | 0.22 | 4.8 |

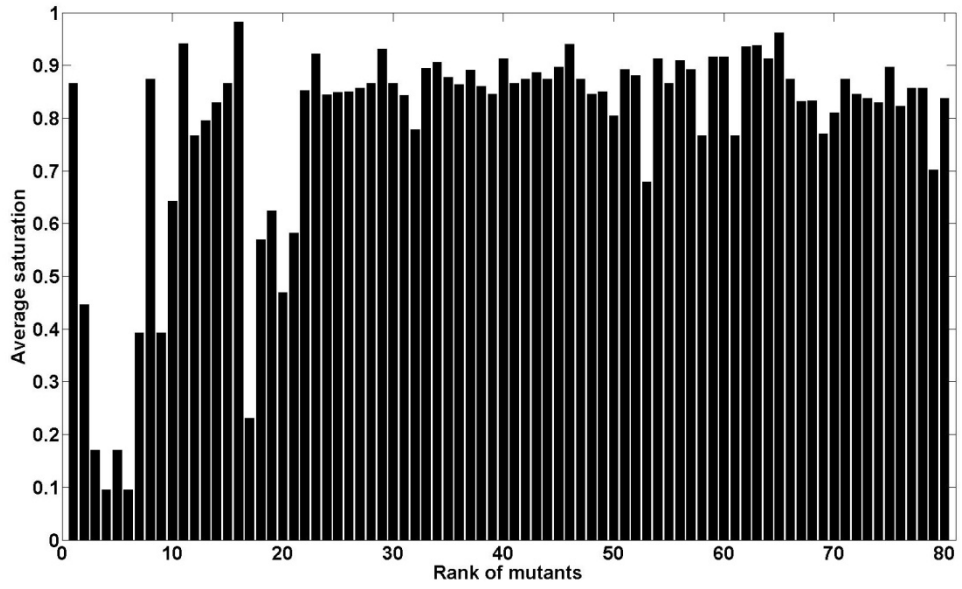



**Figure B10**. The average saturation of converted seal Mbs within the muscle cell in the pressure range, $P_{mit}$ = 0 mmHg to $P_c$=87 mmHg at $37^0C$.

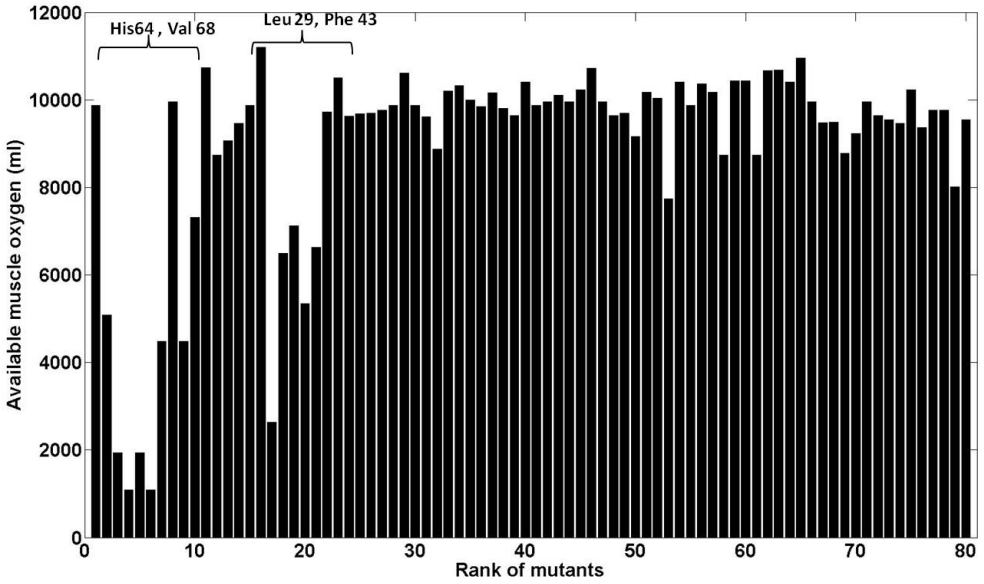



**Figure B11**. The amount of available $O_2$ (ml) at $37^0C$ in the muscle of Weddell seal mutants at the beginning of a dive, using converted seal mutant data.

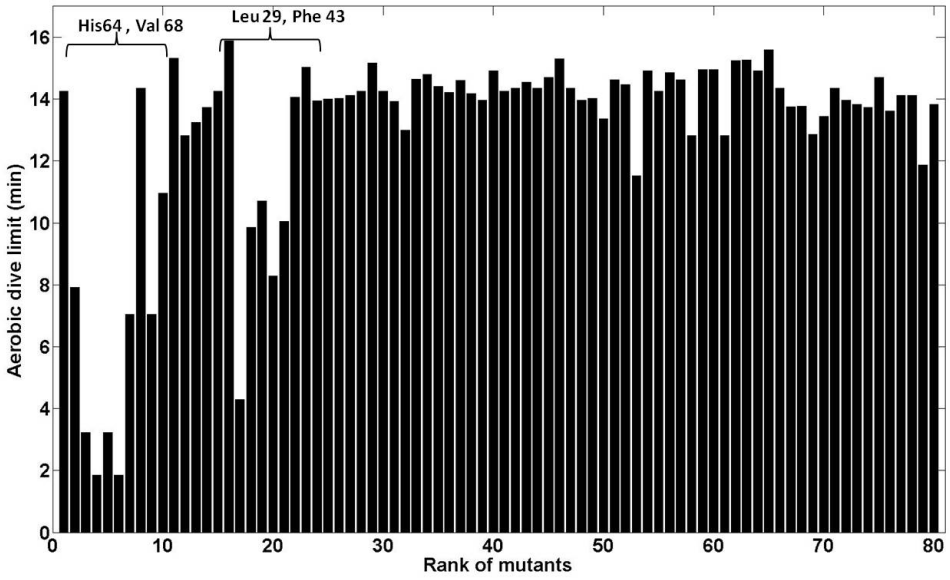



**Figure B12**. The calculated ADL (min) of the simulated dive of a Weddell seal with $\dot{V}_{MO2} = 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} = 0.19\dot{V}_{bO2}(rest)$ at $37^0C$ for different Mb mutants, using converted seal mutant $K_{O2}$ data.

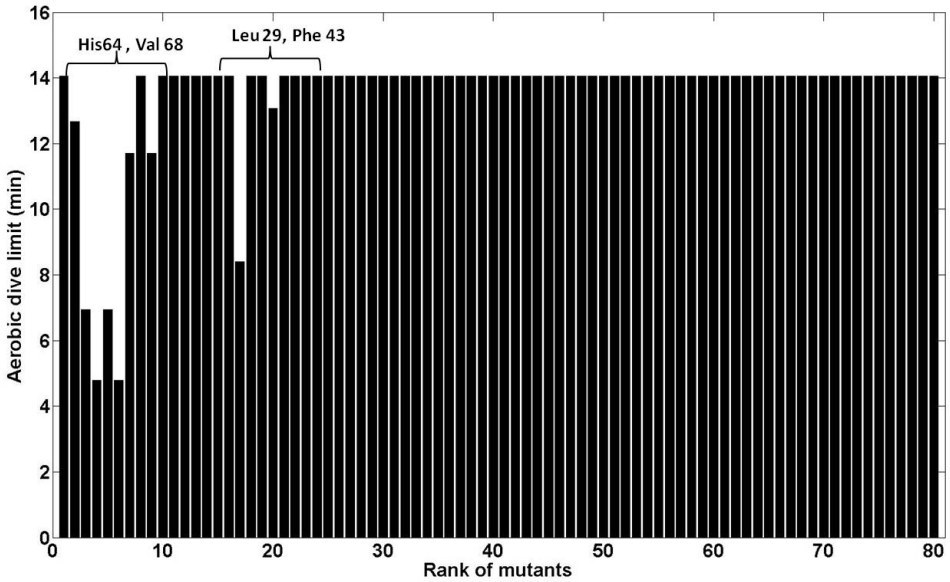



**Figure B13**. The calculated ADL (min) of the simulated dive of a Weddell seal with $\dot{V}_{MO2} = 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} = 0.37\dot{V}_{bO2}(rest)$ at $37^0C$ for different Mb mutants, using converted seal mutant $K_{O2}$ data directly.

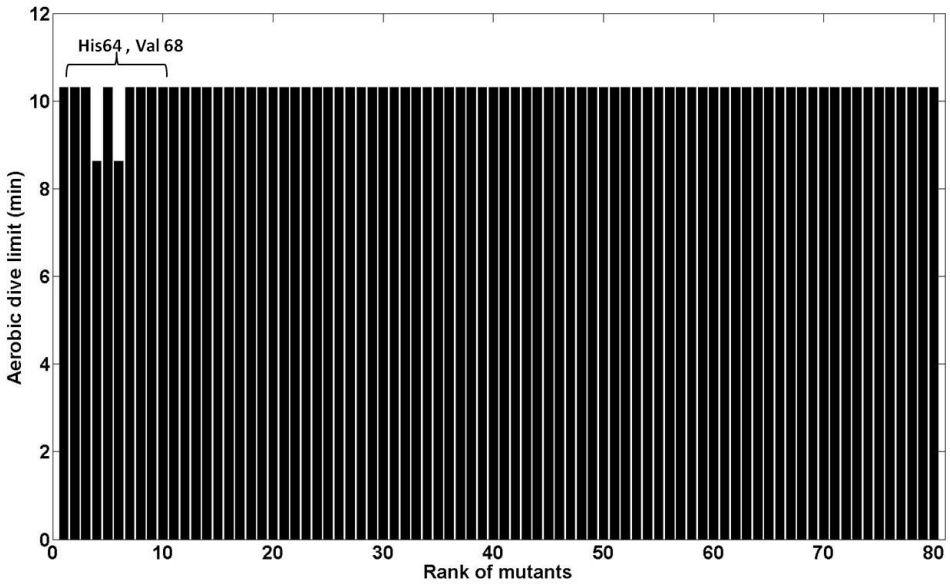



**Figure B14**. The calculated ADL (min) of the simulated dive of a Weddell seal with $\dot{V}_{MO2} = 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} = 0.55\dot{V}_{bO2}(rest)$ at $37^0C$ for different Mb mutants, using converted seal mutant $K_{O2}$ data.

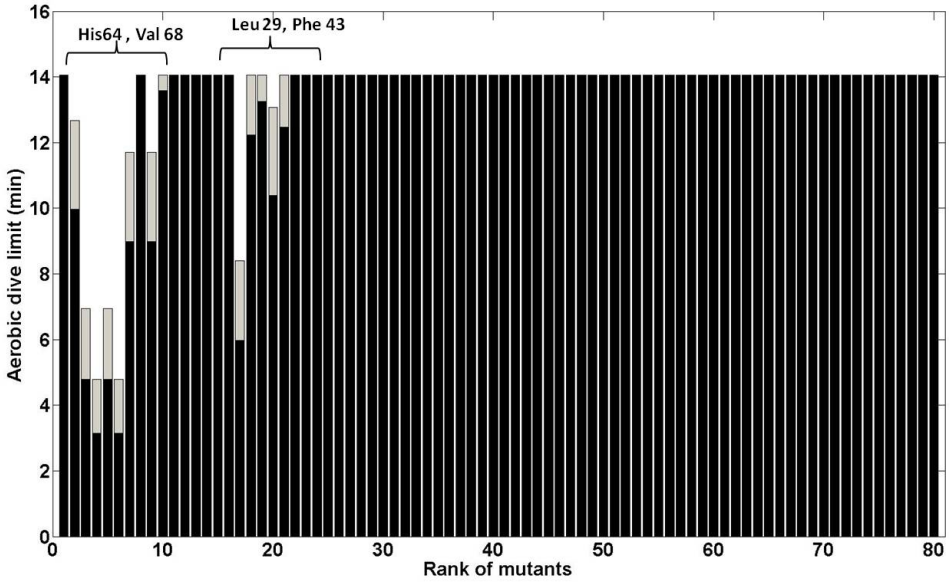



**Figure B15**. The calculated ADL (min) of the simulated dive of a Weddell seal with $\dot{V}_{MO2} = 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} = 0.37\dot{V}_{bO2}(rest)$ at $37^0C$ for different Mb mutants, using converted seal mutant $K_{O2}$ data, at $P_a$ =119 mmHg and $P_v$ = 55 mmHg (gray bars) and $P_a$ = 59.5 mmHg and $P_v$ = 27.5 mmHg (black bars).

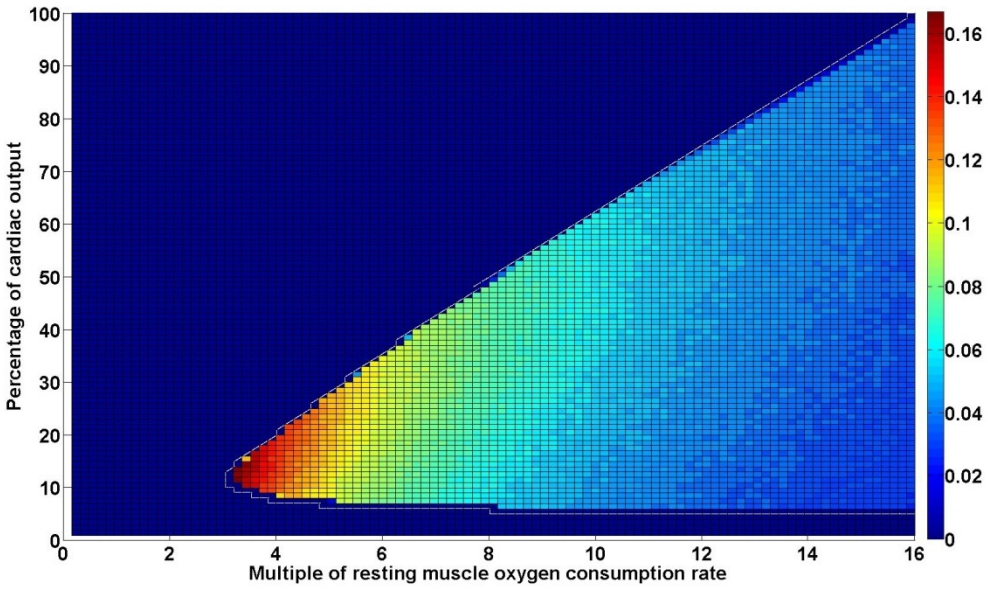



**Figure B16**. The difference in ADL from using whale instead of seal data. Plot of $\Delta ADL=ADL(seal\ Mb)\ ADL(whale\ Mb)$ over possible combinations of $\dot{V}_{MO2}$ and $\dot{V}_{bO2}$. The maximum $\Delta ADL$ occurs at $\dot{V}_{MO2} \sim 3.5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} \sim 0.1\dot{V}_{bO2}(rest)$.

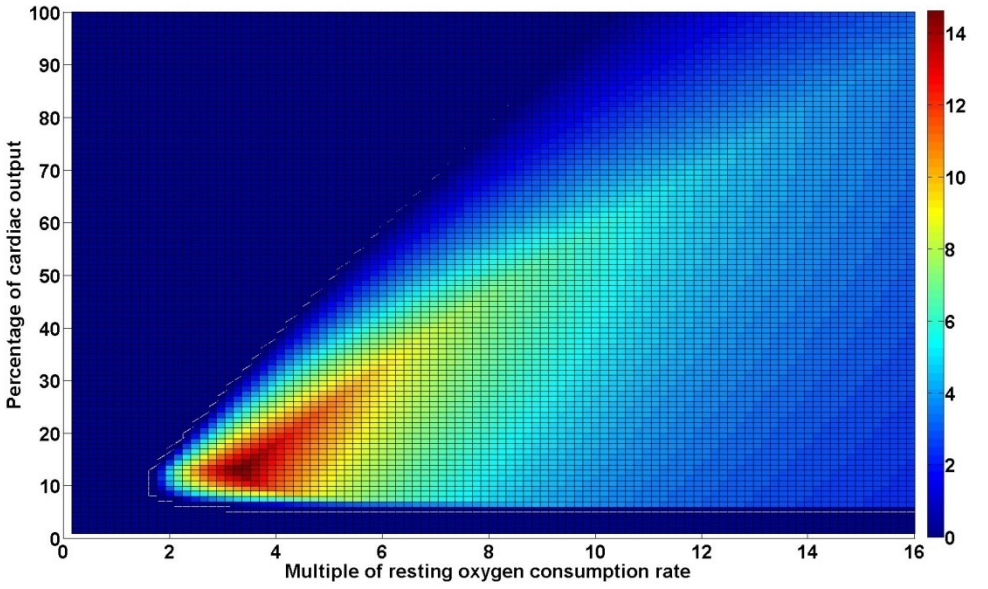



**Figure B17**. Plot of $\Delta ADL_{0.01 \to 1}$ over possible combinations of $\dot{V}_{MO2}$ and $\dot{V}_{bO2}$ at $P_a$=119 mmHg and $P_v$=55 mmHg. The maximum $\Delta ADL_{0.01 \to 1}$ occurs at $\dot{V}_{MO2} \sim 3\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} \sim 0.15\dot{V}_{bO2}(rest)$.

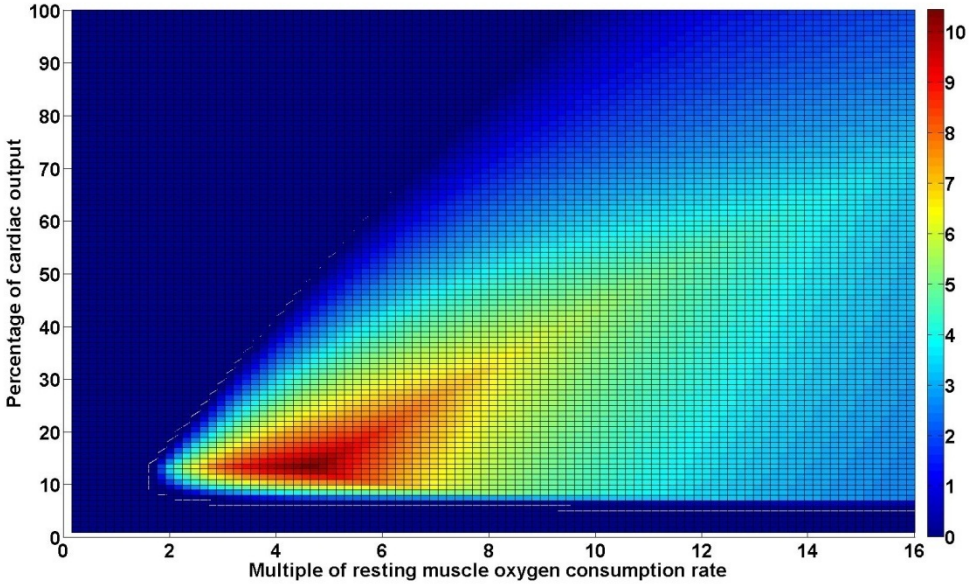



**Figure B18**. Plot of $\Delta ADL_{0.01 \to 1}$ over possible combinations of $\dot{V}_{MO2}$ and $\dot{V}_{bO2}$ at $P_a$=59.5 mmHg and $P_v$=27 mmHg. The maximum $\Delta ADL_{0.01 \to 1}$ occurs at $\dot{V}_{MO2} \sim 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} \sim 0.15\dot{V}_{bO2}(rest)$.

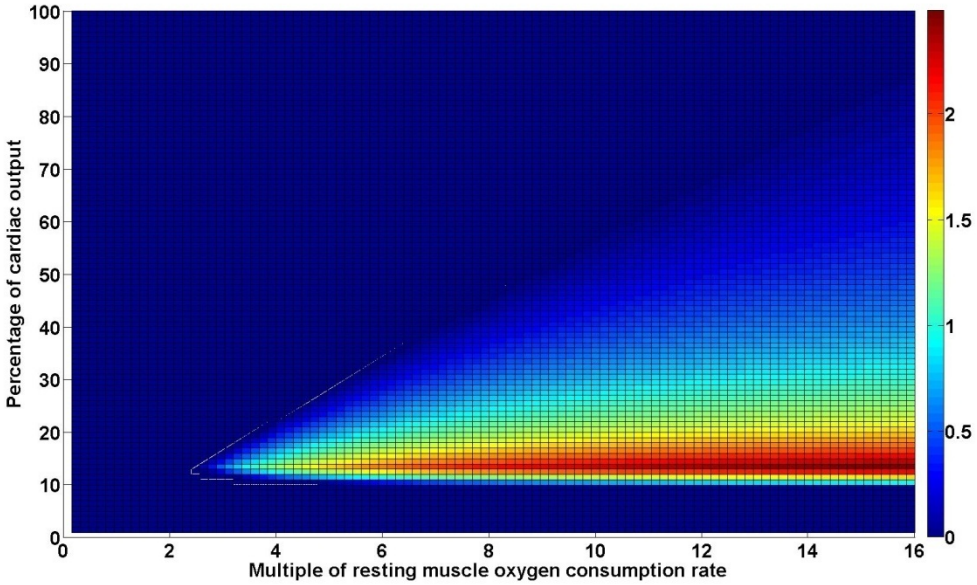



**Figure B19**. Plot of $\Delta ADL_{0.01 \to 1}$ over possible combinations of $\dot{V}_{MO2}$ and $\dot{V}_{bO2}$ at $P_a$=29.75 mmHg and $P_v$=13.75 mmHg. The maximum $\Delta ADL_{0.01 \to 1}$ occurs at $\dot{V}_{MO2} \sim 12-16\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} \sim 0.15\dot{V}_{bO2}(rest)$.

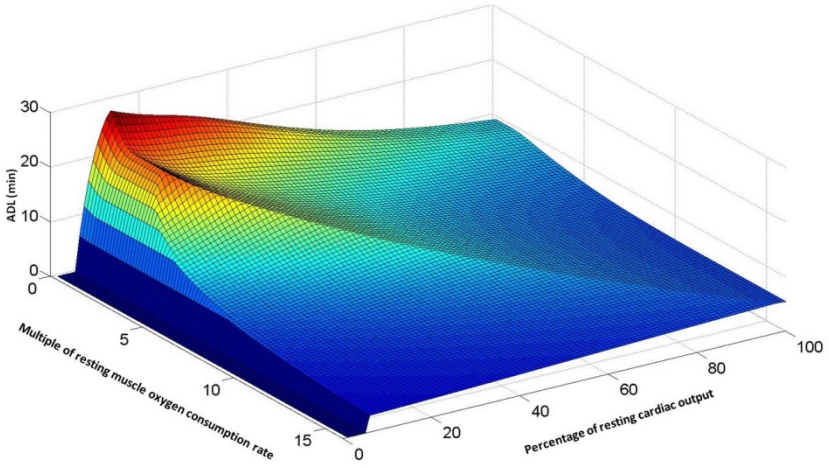



**Figure B20**. Plot of ADL for WT seal Mb, over possible combinations of $\dot{V}_{MO2}$ and $\dot{V}_{bO2}$ at $P_a$=119 mmHg and $P_v$=55 mmHg. The maximum *ADL* occurs at $\dot{V}_{MO2} \sim 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} \sim 0.2\dot{V}_{bO2}(rest)$.

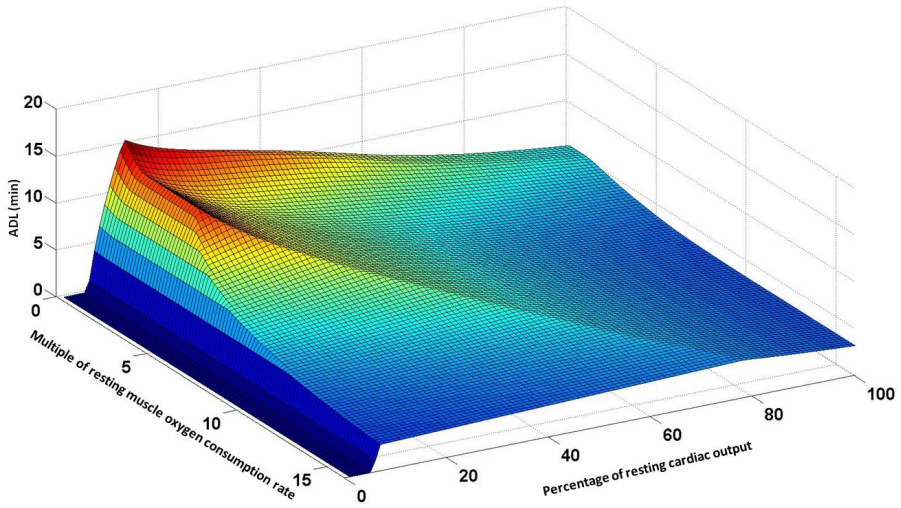



**Figure B21**. Plot of *ADL* for WT seal Mb, over possible combinations of $\dot{V}_{MO2}$ and $\dot{V}_{bO2}$ at $P_a$=59.5 mmHg and $P_v$=27.5 mmHg. The maximum *ADL* occurs at $\dot{V}_{MO2} \sim 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} \sim 0.15\dot{V}_{bO2}(rest)$.

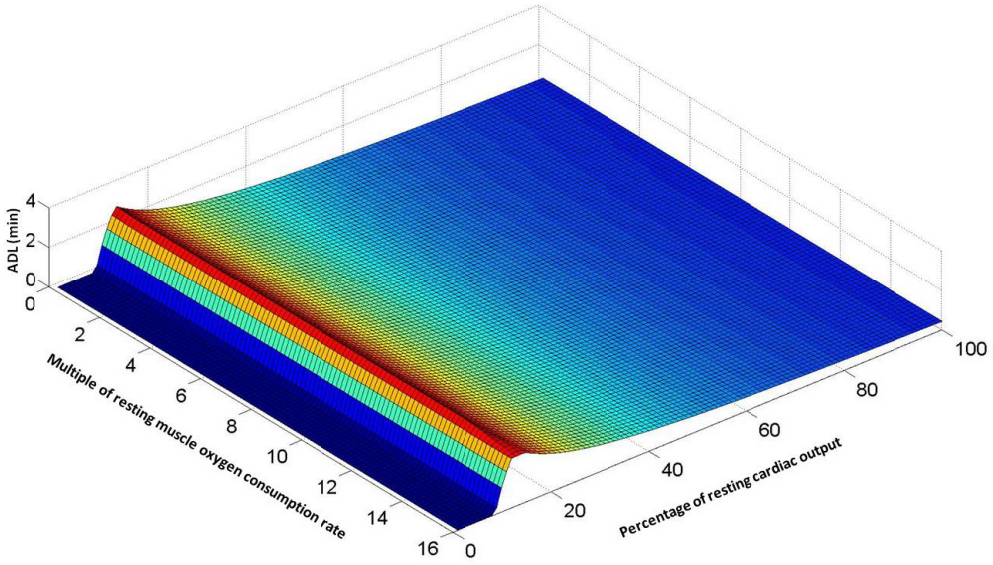



**Figure B22**. Plot of *ADL* for WT seal Mb, over possible combinations of $P_a$=29.75 mmHg and $P_v$=13.75 mmHg. The maximum *ADL* occurs at $\dot{V}_{bO2} \sim 0.15\dot{V}_{bO2}(rest)$ .

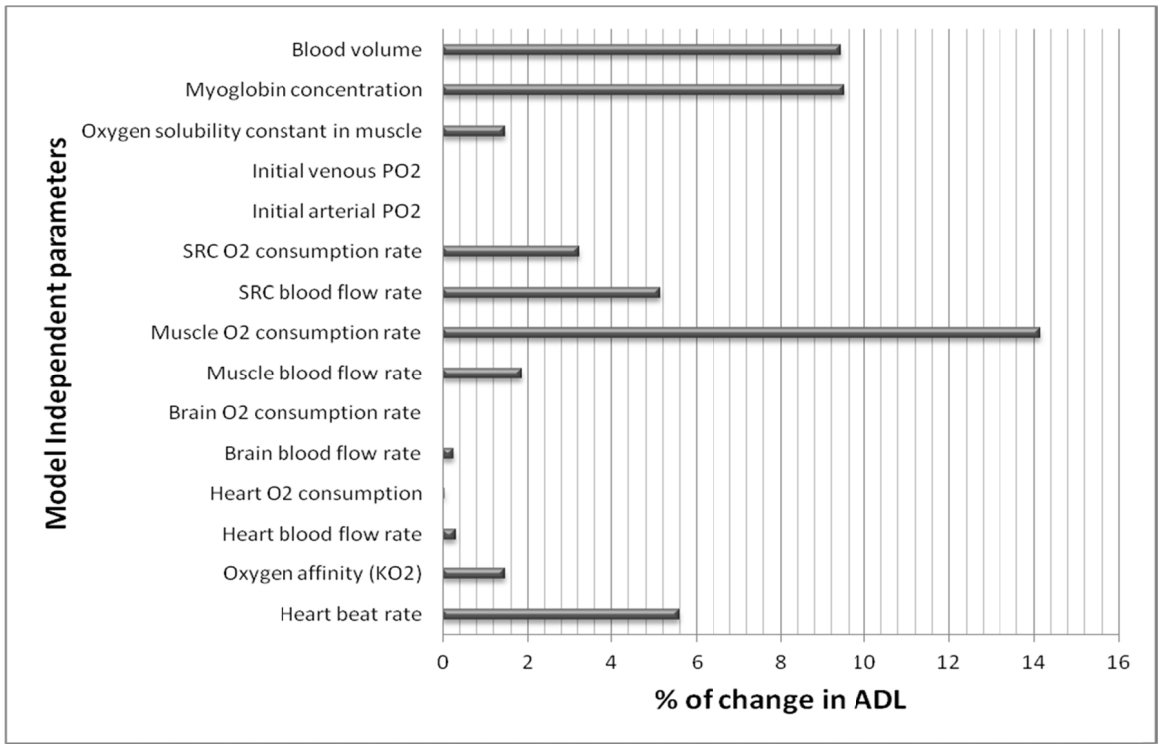



**Figure B23**. Percentage of change in ADL by ±10% change in the values of initial parameters presented in table S1. The results are under the routine dive conditions, $\dot{V}_{MO2} \sim 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} \sim 0.15\dot{V}_{bO2}(rest)$ .

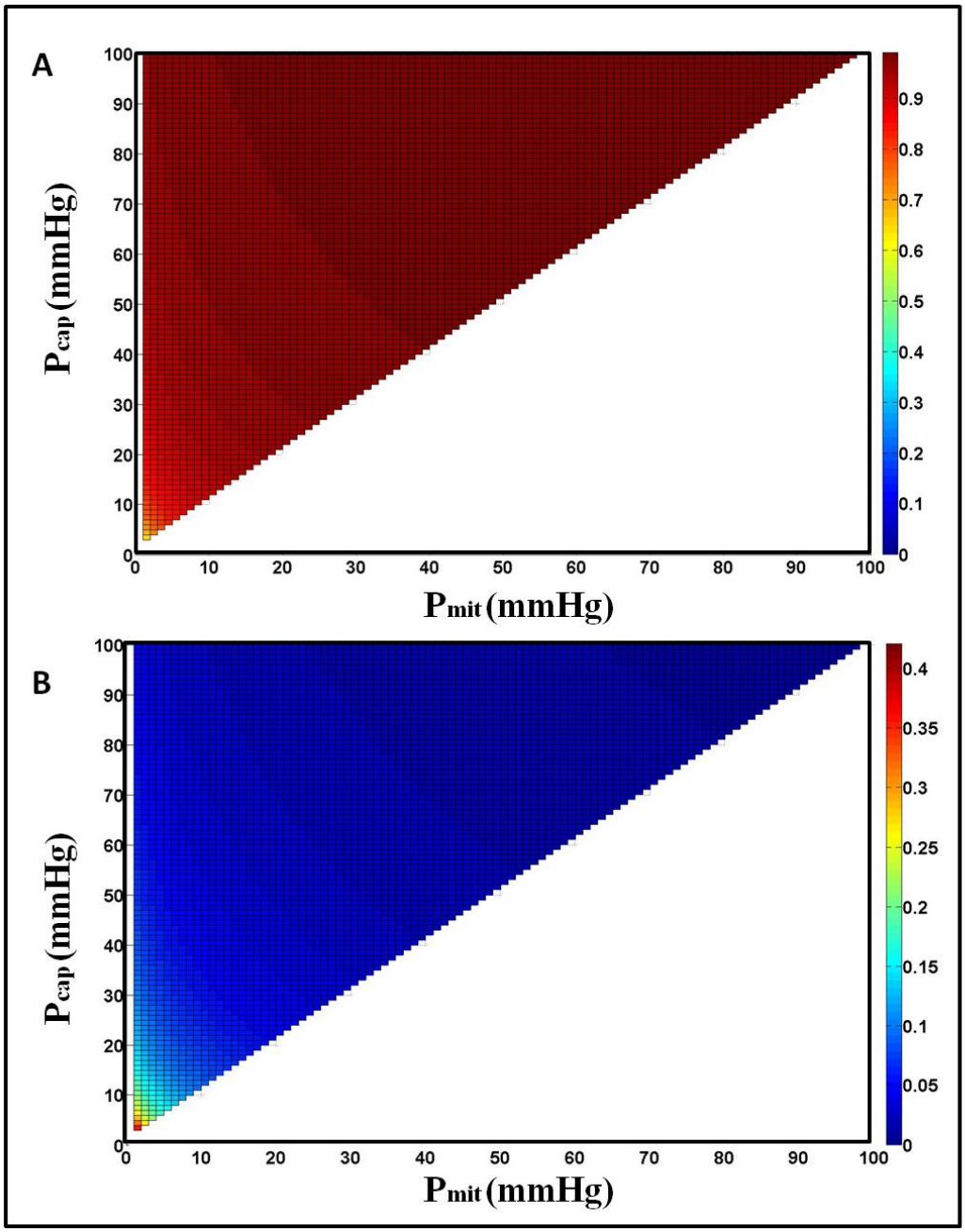



**Figure B24**. Sensitivity analysis for changes in mitochondrial $P_{O2}$, $P_{mit}$: Changes in average saturation at A) $K_{O2}$ = 1 $\mu M^{-1}$ and B) going from $K_{O2}$ = 0.01 to 1 $\mu M^{-1}$, over possible combinations of $P_{O2}$ at mitochondria, $P_{mit}$, and $P_{O2}$ at capillary, $P_{cap}$. The average saturation is calculated from Equation 4 and the oxygen solubility in the muscle tissue, $\alpha_{O2}$ = 9.4 x $10^{-7}$ mol $L^{-1}$ $mmHg^{-1}$.

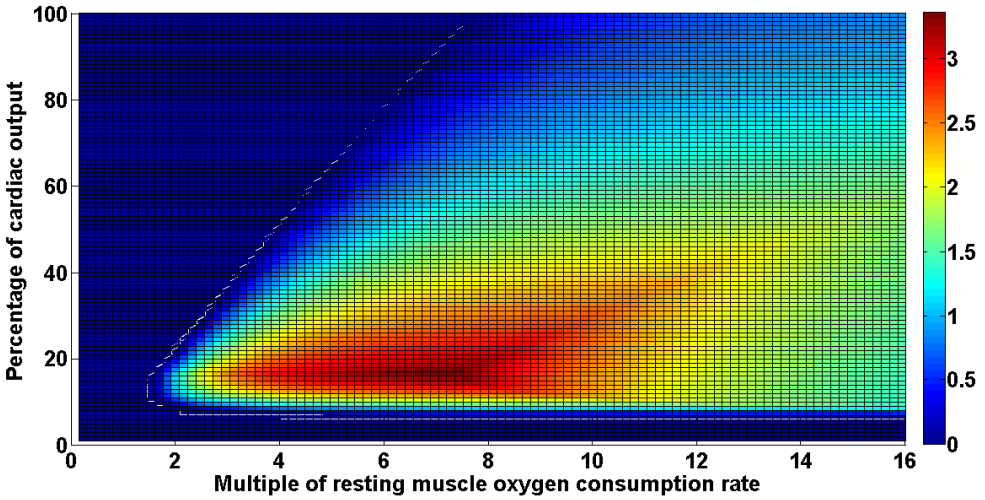



**Figure B25**. Plot of $\Delta \mathrm{ADL}_{0.002 \to 0.29}$ for a male sea lion over possible combinations of $\dot{V}_{MO2}$ and $\dot{V}_{bO2}$ at $P_a$=119 mmHg and $P_v$=55 mmHg. The maximum $\Delta \mathrm{ADL}_{0.002 \to 0.29}$ occurs at $\dot{V}_{MO2} \sim 4-8\, \dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} \sim 0.2\, \dot{V}_{bO2}(rest)$.

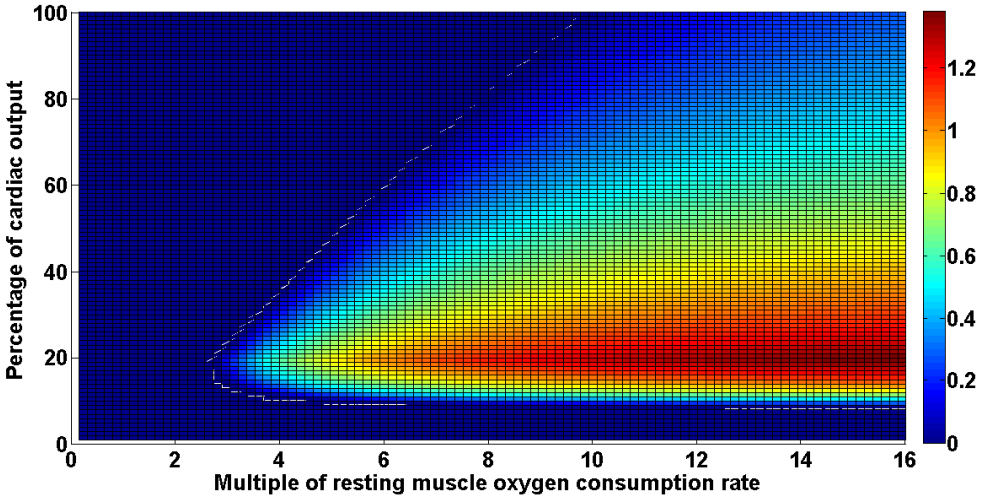



**Figure B26**. Plot of $\Delta \mathrm{ADL}_{0.002 \to 0.29}$ for a female California sea lion over possible combinations of $\dot{V}_{MO2}$ and $\dot{V}_{bO2}$ at $P_a$=119 mmHg and $P_v$=55 mmHg. The maximum $\Delta \mathrm{ADL}_{0.002 \to 0.29}$ occurs at $\dot{V}_{MO2} \sim 8-16\, \dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} \sim 0.2\, \dot{V}_{bO2}(rest)$.

**Table B3**. Computed aerobic dive limit (ADL), parameters and variables for a 450 kg Weddell seal, a 200 kg Weddell seal and male and female California sea lions. ADL is estimated under routine dive conditions of $\dot{V}_{MO2} \sim 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} \sim 0.27\dot{V}_{bO2}(rest)$.

| | | 450 kg Weddell seal | | Hypothetical 200 kg Weddell seal [Corrected for blood volume, Mb concentration and heart bit rate] | | Male California sea lion (175 kg) | | Female California sea lion (86.7 kg) | |
|---|---|---|---|---|---|---|---|---|---|
| | | WT | H64V | WT | H64V | WT | H64V | WT | H64V |
| | | $K_{O2}$=0.29 | $K_{O2}$=0.0027 | $K_{O2}$=0.29 | $K_{O2}$=0.0027 | $K_{O2}$=0.29 | $K_{O2}$=0.0027 | $K_{O2}$=0.29 | $K_{O2}$=0.0027 |
| Aerobic dive limit (min) | ADL | 16.7 | 2.72 | 12.4 | 1.74 | 3.17 | 0.98 | 2.87 | 0.95 |
| Mb concentration (g kg$^{-1}$ muscle) | $C_{Mb}$ | 54.0 | 54.0 | 54.0 | 54.0 | 35.0 | 35.0 | 50.0 | 50.0 |
| Heart beat rate (beats min$^{-1}$) | $f_h$ | 13.9 | 13.9 | 31.3 | 31.3 | 35.8 | 35.8 | 72.2 | 72.2 |
| Blood volume (l) | $BV$ | 96.0 | 96.0 | 42.7 | 42.7 | 20.0 | 20.0 | 10.0 | 10.0 |
| Cardiac output (l min$^{-1}$) | $\dot{V}_b$ | 11.5 | 11.5 | 5.12 | 5.12 | 4.48 | 4.48 | 2.22 | 2.22 |
| Brain blood flow rate (l min$^{-1}$) | $\dot{Q}_B$ | 0.360 | 0.360 | 0.160 | 0.160 | 0.140 | 0.140 | 0.070 | 0.070 |
| Heart blood flow rate (l min$^{-1}$) | $\dot{Q}_H$ | 0.497 | 0.497 | 0.221 | 0.221 | 0.193 | 0.193 | 0.097 | 0.097 |
| Skeletal muscle blood flow rate (l min$^{-1}$) | $\dot{Q}_M$ | 2.13 | 2.13 | 0.94 | 0.94 | 0.83 | 0.83 | 0.411 | 0.411 |
| Blood flow rate for splanchnic, renal, cutaneous, and other peripheral tissues (l min$^{-1}$) | $\dot{Q}_{SRC}$ | 8.81 | 8.81 | 3.91 | 3.91 | 3.43 | 3.43 | 1.69 | 1.69 |
| Brain oxygen consumption rate (ml $O_2$ min$^{-1}$) | $\dot{V}_{Bo_2}$ | 13.3 | 13.3 | 7.23 | 7.23 | 6.55 | 6.55 | 3.86 | 3.86 |
| Heart oxygen consumption rate (ml $O_2$ min$^{-1}$) | $\dot{V}_{Ho_2}$ | 30.4 | 30.4 | 16.5 | 16.5 | 14.9 | 14.9 | 8.83 | 8.83 |
| Skeletal muscle oxygen consumption rate (ml $O_2$ min$^{-1}$) | $\dot{V}_{Mo_2}$ | $1.08\times10^3$ | $1.08\times10^3$ | 588 | 588 | 532 | 532 | 314 | 314 |
| $O_2$-consumption rate for splanchnic, renal, cutaneous, and other peripheral tissues (ml $O_2$ min$^{-1}$) | $\dot{V}_{SRCo_2}$ | 555 | 555 | 302 | 302 | 273 | 273 | 161 | 161 |
| Average Mb saturation | $<S>$ | 0.865 | 0.0964 | 0.864 | 0.0964 | 0.864 | 0.0964 | 0.864 | 0.0964 |

## Multiple sequence alignment of seals and whales myoglobins

```
CLUSTAL O(1.1.0) multiple sequence alignment

Harbor_porpoise               MGLSEGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKFKHLKTEAEMKASE 60
Dall's_porpoise               MGLSEGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKFKHLKTEAEMKASE 60
Sperm_whale                   MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 60
Goose-beaked_whale            MGLSEAEWQLVLHVWAKVEADLSGHGQEILIRLFKGHPETLEKFDKFKHLKSEAEMKASE 60
Killer_whale                  MGLSDGEWQLVLNVWGKVEADLAGHGQDILIRLFKGHPETLEKFDKFKHLKTEADMKASE 60
Common_minke_whale            MVLSDAEWHLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASE 60
Humpback_whale                MVLSDAEWQLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASE 60
Finback_whale                 MVLTDAEWHLVLNIWAKVEADVAGHGQDILISLFKGHPETLEKFDKFKHLKTEAEMKASE 60
California_gray_whale         MVLSDAEWQLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASE 60
Long_finned_pilot_whale       MGLSDGEWQLVLNVWGKVEADLAGHGQDILIRLFKGHPETLEKFDKFKHLKTEADMKASE 60
Dwarf_sperm_whale             MVLSEGEWQLVLHVWAKVEADIAGHGQDILIRLFKHHPETLEKFDRFKHLKSEAEMKASE 60
Hubb's_beaked_whale           MGLSEAEWQLVLHVWAKVEADLSGHGQEILIRLFKGHPETLEKFDKFKHLKSEAEMKASE 60
Harbor_seal                   MGLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKFKHLKSEDDMRRSE 60
Gray_seal                     MGLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKFKHLKSEDDMRRSE 60
Baikal_seal                   MGLSDGEWHLVLNVWGKWETDLAGHGQEVLIRLFKSHPETLEKFDKFKHLKSEDDMRRSF 60
California_sea_lion           MGLSDGEWQLVLNIWGKVEADLVGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKRSE 60
Saddleback_dolphin            MGLSDGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKFKHLKTEADMKASE 60
Sei_whale                     MVLSDAEWQLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASE 60
Pigmy_bryde's_whale           MVLSDAEWQLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASE 60
Longman's                     MGLSEAEWQLVLHVWAKVEADLSGHGQEILIRLFKGHPETLEKFDKFKHLKSEAEMKASE 60
Pygmy_sperm_whale             MVLSEGEWQLVLHVWAKVEADIAGHGQDILIRLFKHHPETLEKFDRFKHLKSEAEMKASE 60
Stejneger's_beaked_whale      MGLSEAEWQLVLHVWAKVEADLSGHGQEILIRLFKGHPETLEKFDKFKHLKSEAEMKASE 60
Melon_headed_whale            MGLSDGEWQLVLNVWGKVEADLAGHGQDILIRLFKGHPETLEKFDKFKHLKTEADMKASE 60
Amazon_dolphin                MGLSDGEWQLVLNIWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKFKHLKTEAEMKASE 60
Aatlantic_bottle_nose_dolphin MGLSDGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKFKHLKTEADMKASE 60
Pantropical_spotted_dolphin   MGLSDGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKFKHLKTEADMKASE 60
                              * *::.**:***.:*.* *:*: ****::** *** *********:*****:* :*: *

Harbor_porpoise               DLKKHGNTVLTALGGILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
Dall's_porpoise               DLKKHGNTVLTALGGILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
Sperm_whale                   DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
Goose-beaked_whale            DLKKHGHTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRH 120
Killer_whale                  DLKKHGNTVLTALGAILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
Common_minke_whale            DLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRH 120
Humpback_whale                DLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRH 120
Finback_whale                 DLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRH 120
California_gray_whale         DLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRH 120
Long_finned_pilot_whale       DLKKHGNTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
Dwarf_sperm_whale             DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
Hubb's_beaked_whale           DLKKHGHTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 120
Harbor_seal                   DLRKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSKH 120
Gray_seal                     DLRKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSKH 120
Baikal_seal                   DLRKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSKH 120
California_sea_lion           DLKKHGKTVLTALGGILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLQSKH 120
Saddleback_dolphin            DLKKHGNTVLTALGAILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
Sei_whale                     DLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRH 120
Pigmy_bryde's_whale           DLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRH 120
Longman's                     DLKKHGHTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 120
Pygmy_sperm_whale             DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
Stejneger's_beaked_whale      DLKKHGHTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 120
Melon_headed_whale            DLKKHGITVLTALGAILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
Amazon_dolphin                DLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
Aatlantic_bottle_nose_dolphin DLKKHGNTVLTALGAILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
Pantropical_spotted_dolphin   DLKKHGNTVLTALGAILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
                              **:*** *******.********:*************************:******:*:*

Harbor_porpoise               PAEFGADAQGAMNKALELFRKDIATKYKELGFHG 154
Dall's_porpoise               PAEFGADAQGAMNKALELFRKDIATKYKELGFHG 154
Sperm_whale                   PGDFGADAQGAMNKALELFRKDIAAKYKELGYQG 154
Goose-beaked_whale            PSDFGADAQAAMTKALELFRKDIAAKYKELGFHG 154
Killer_whale                  PAEFGADAQGAMNKALELFRKDIAAKYKELGFHG 154
Common_minke_whale            PAEFGADAQAAMNKALELFRKDIAAKYKELGFQG 154
Humpback_whale                PADFGADAQAAMNKALELFRKDIAAKYKELGFQG 154
Finback_whale                 PADFGADAQAAMNKALELFRKDIAAKYKELGFQG 154
California_gray_whale         PGDFGADAQAAMNKALELFRKDIAAKYKELGFQG 154
Long_finned_pilot_whale       PAEFGADAQGAMNKALELFRKDIAAKYKELGFHG 154
Dwarf_sperm_whale             PADFGADAQGAMSKALELFRKDIAAKYKELGYQG 154
Hubb's_beaked_whale           PSDFGADAQGAMTKALELFRKDIAAKYKELGFHG 154
```

```
Harbor_seal                    PAEFGADAQAAMKKALELFRNDIAAKYKELGFHG 154
Gray_seal                      PAEFGADAQAAMKKALELFRNDIAAKYKELGFHG 154
Baikal_seal                    PAEFGADAQAAMKKALELFRNDIAAKIKELGFHG 154
California_sea_lion            PGDFGADTHAAMKKALELFRNDIAAKYRELGFQG 154
Saddleback_dolphin             PAEFGADAQGAMNKALELFRKDIAAKYKELGFHG 154
Sei_whale                      PGDFGADAQAAMNKALELFRKDIAAKYKELGFQG 154
Pigmy_bryde's_whale            PGDFGADAQAAMNKALELFRKDIAAKYKELGFQG 154
Longman's                      PSDFGADAQAAMTKALELFRKDIAAKYKELGFHG 154
Pygmy_sperm_whale              PADFGADAQGAMSKALELFRKDIAAKYKELGYQG 154
Stejneger's_beaked_whale       PSDFGADAQGAMTKALELFRKDIAAKYKELGFHG 154
Melon_headed_whale             PAEFGADAQGAMNKALELFRKDIAAKYKELGFHG 154
Amazon_dolphin                 PGDFGADAQAAMNKALELFRKDIAAKYKELGFHG 154
Aatlantic_bottle_nose_dolphin  PAEFGADAQGAMNKALELFRKDIAAKYKELGFHG 154
Pantropical_spotted_dolphin    PAEFGADAQGAMNKALELFRKDIAAKYKELGFHG 154
                               *.:****::.**.*******:***:* :***::*
```

**Table B4**. Natural variation in the residues 8, 28, 66 and 144 in whale, seal and sea lion Mbs.

| Organism | Variable sites in the mutant data set | | | | Accession code |
|---|---|---|---|---|---|
| | 8 | 28 | 66 | 144 | |
| Harbor porpoise | Q | V | N | T | P68278 |
| Dall's porpoise | Q | V | N | T | P68277 |
| Sperm whale | Q | I | V | A | P02185 |
| Goose beaked whale | Q | I | H | A | P02182 |
| Killer whale | Q | I | N | A | P02173 |
| Common minke whale | H | I | N | A | P02179 |
| Humpback whale | Q | I | N | A | P02178 |
| Finback whale | H | I | N | A | P02180 |
| California gray whale | Q | I | N | A | P02177 |
| Long finned pilot whale | Q | I | N | A | P02174 |
| Dwarf sperm whale | Q | I | V | A | P02184 |
| Hubbs beaked whale | Q | I | H | A | P02183 |
| Harbor seal | H | V | N | A | P68080 |
| Gray seal | H | V | N | A | P68081 |
| Baikal seal | H | V | N | A | P30562 |
| California sealion | Q | V | K | A | P02161 |
| Saddleback dolphin | Q | V | N | A | P68276 |
| Sei whale | Q | I | N | A | Q0KIY1 |
| Pigmy bryde's whale | Q | I | N | A | Q0KIY2 |
| Longmans beaked whale | Q | I | H | A | Q0KIY9 |
| Pygmy sperm whale | Q | I | V | A | Q0KIY5 |
| Stejnegers beaked whale | Q | I | H | A | Q0KIY0 |
| Melon headed whale | Q | I | I | A | Q0KIY3 |
| Amazon dolphin | Q | V | N | A | P02181 |
| Atlantic bottle nose dolphin | Q | V | N | A | P68279 |
| Pantropical spotted dolphin 1 | Q | V | N | A | Q0KIY7 |
| Pantropical spotted dolphin 2 | Q | V | N | T | Q0KIY6 |

**Table B5**. Aerobic dive limit (ADL) calculated for mutants having mutations in the residues 8, 28, 66 and 144. ADL for the WT seal is 16.68 min under routine dive conditions of $\dot{V}_{MO2} \sim 5\dot{V}_{MO2}(rest)$ and $\dot{V}_{bO2} \sim 0.27\dot{V}_{bO2}(rest)$.

| Site | Mutants | |
|---|---|---|
| 28 | A28 | W28 |
| | ADL (min)= 16.5 | ADL (min)=17.2 |
| 66 | K66 | R66 |
| | ADL (min)=16.9 | ADL (min)=16.7 |
| 8 | V8 | |
| | ADL (min)=15.3 | |
| 144 | V144 | |
| | ADL (min)=16.3 | |

# APPENDIX C: SUPPLEMENTARY DATA FOR THE STUDY OF POSITIVE SELECTION IN CETACEAN MBS

## 1. FoldX calculated versus the experimental ΔΔ*G*s on a validation set of 16 mutations

**Table C1**. FoldX calculations for mutations with experimental ΔΔ*G* values with PDB structures (1MBO (Phillips 1980)) and (1U7S (Kondrashov et al., 2008)).

| Nr | **Reference** | Mutation | $\Delta\Delta G_{experimental}$ (unfolding) | FoldX+ Repaired+1MBO (kcal/mol) | FoldX+ Repaired+1u7s (kcal/mol) |
|---|---|---|---|---|---|
| 1 | (Nishimura et al., 2003) | **WT**[a] | 0 | 0 | 0 |
| 2 | (Nishimura et al., 2003) | I28A | 2.06 | 0.74 | 1.08 |
| 3 | (Nishimura et al., 2003) | L29A | 0.39 | 2.69 | 2.81 |
| 4 | (Nishimura et al., 2003) | I30A | 1.9 | 1.6 | 2.32 |
| 5 | (Nishimura et al., 2003) | L32A | 2.04 | 2.91 | 2.81 |
| 6 | (Hughson et al., 1991) | A130L | 2.3 | 4.29 | 2.11 |
| 7 | (Hughson et al., 1991) | A130K | 3.7 | 2.55 | 4.43 |
| 8 | (Hughson et al., 1991) | F123T | 3.5 | 4.11 | 4.75 |
| 9 | (Barrick et al., 1994) | H24V | 0.52 | 0.14 | -1.23 |
| 10 | (Barrick et al., 1994) | H36Q | 1.3 | 0.54 | 0.91 |
| 11 | (Barrick et al., 1994) | H48Q | 0.62 | -0.29 | -0.42 |
| 12 | (Barrick et al., 1994) | H64Q | 0.45 | -0.29 | -0.52 |
| 13 | (Barrick et al., 1994) | H82Q | 0.05 | 2.24 | 1.34 |
| 14 | (Barrick et al., 1994) | H93G | -0.04 | 2.07 | 2.12 |
| 15 | (Barrick et al., 1994) | H97Q | 0.11 | -0.23 | -0.26 |
| 16 | (Barrick et al., 1994) | H113Q | 0.26 | 0.14 | -0.48 |
| 17 | (Barrick et al., 1994) | H119F | 0.68 | -0.25 | -0.5 |

a: Experimental ΔΔ*G* values for the mutants are calculated from Δ*G*(mutant)- Δ*G*(WT) where Δ*G*(WT) is the respective WT free energy of unfolding for each group.

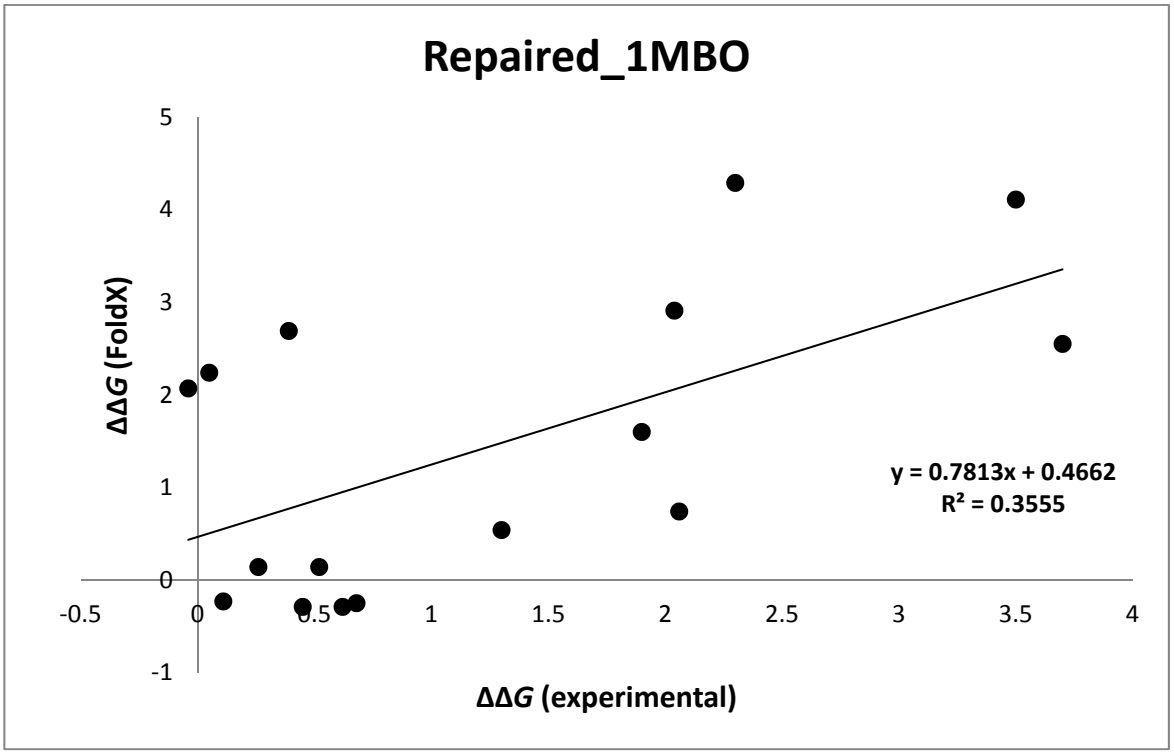



**Figure C1**. ΔΔ*G* values predicted by FoldX versus experimental ΔΔ*G*s (kcal/mol) for the validation set (pdb=1MBO).

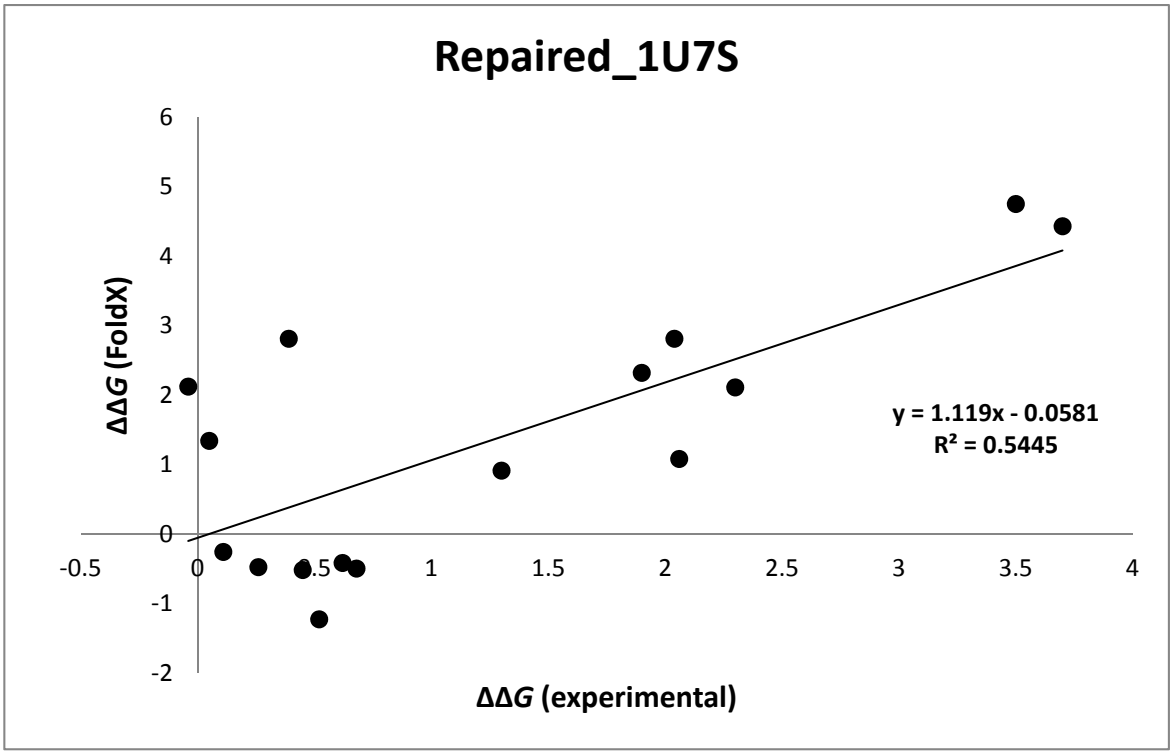



**Figure C2**. ΔΔ*G* values predicted by FoldX versus experimental ΔΔ*G*s (kcal/mol) for the validation set (pdb=1U7S).

**Table C2.** FoldX calculations for all mutations in the Cetacean clade with PDB structures (1U7S (Kondrashov et al., 2008)) (Mutations in the sites detected to be under positive selection are shown in grey).

| Mutation | $\Delta\Delta G$ | Mutation | $\Delta\Delta G$ |
|---|---|---|---|
| V1G | -0.444 | N66V | -1.074 |
| G1V | 0.444 | N66H | -0.444 |
| S3T | 0.988 | N66I | -1.696 |
| D4E | 0.010 | G74A | -1.008 |
| G5A | -0.722 | E83D | 0.282 |
| Q8H | 0.208 | D83E | -0.282 |
| N12H | 0.240 | V101I | -1.446 |
| V13I | -0.550 | E109D | 0.426 |
| G15A | -0.230 | K118R | -0.670 |
| A15G | 0.230 | R118K | 0.670 |
| V21I | -0.678 | G121A | 0.038 |
| V21L | -1.224 | G121S | -1.032 |
| A22S | 0.556 | D122E | 0.208 |
| E27D | -1.518 | G129A | -0.670 |
| D27E | 1.518 | A129G | 0.670 |
| V28I | -1.098 | S132N | -0.007 |
| I28V | 1.098 | N132S | 0.007 |
| R31S | 0.404 | N132T | 0.530 |
| G35S | -0.142 | N140K | -0.090 |
| S35H | -0.636 | M142I | 0.212 |
| K45R | -0.142 | A144T | 1.080 |
| T51S | 0.334 | F151Y | 0.962 |
| E54D | -0.352 | Q152H | 0.504 |

Probability of stabilization being conditional on the positive selection can be calculated as:

$$\Pr(\Delta\Delta G < 0|\omega > 1) = \frac{\Pr(\omega>1|\Delta\Delta G<0)\Pr(\Delta\Delta G<0)}{\Pr(\omega>1)} \qquad \text{(S1)}$$

Overall, there are 63 different mutations in the whale, 26 mutations with $\Delta\Delta G < 0$ and 17 mutations in the sites detected to be under positive selection with nine of them having $\Delta\Delta G < 0$ (see Figure 4.9 in the main text). Equation S1 thus gives:

$$\Pr(\Delta\Delta G < 0|\omega > 1) = \frac{\binom{9}{17}\binom{26}{63}}{\binom{17}{63}} = 0.8090$$

## 2. Alignment for sperm whale, pig, bovine, dog, sheep, horse and human myoglobin (Mb) sequences

```
CLUSTAL O(1.1.0) multiple sequence alignment

SP|sp|P02185|MYG_PHYMC|MYG_PHYMC MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE 60
SP|sp|P02189|MYG_PIG|MYG_PIG     MGLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
SP|sp|P02192|MYG_BOVIN|MYG_BOVIN MGLSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 60
SP|sp|P02159|MYG_LYCPI|MYG_LYCPI MGLSDGEWQIVLNIWGKVETDLAGHGQEVLIRLFKNHPETLDKFDKFKHLKTEDEMKGSE 60
SP|sp|P02190|MYG_SHEEP|MYG_SHEEP MGLSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 60
SP|sp|P68082|MYG_HORSE|MYG_HORSE MGLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 60
SP|sp|P02144|MYG_HUMAN|MYG_HUMAN MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
                                 * **:**** **. *.***:*: ****::*****. *****:***:*****:* ***.**

SP|sp|P02185|MYG_PHYMC|MYG_PHYMC DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
SP|sp|P02189|MYG_PIG|MYG_PIG     DLKKHGNTVLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH 120
SP|sp|P02192|MYG_BOVIN|MYG_BOVIN DLKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKH 120
SP|sp|P02159|MYG_LYCPI|MYG_LYCPI DLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPVKYLEFISDAIIQVLQNKH 120
SP|sp|P02190|MYG_SHEEP|MYG_SHEEP DLKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKH 120
SP|sp|P68082|MYG_HORSE|MYG_HORSE DLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH 120
SP|sp|P02144|MYG_HUMAN|MYG_HUMAN DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH 120
                                 ****** .******.***********:. **:***.*****:*******:.**:**: :*

SP|sp|P02185|MYG_PHYMC|MYG_PHYMC PGDFGADAQGAMNKALELFRKDIAAKYKELGYQG 154
SP|sp|P02189|MYG_PIG|MYG_PIG     PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG 154
SP|sp|P02192|MYG_BOVIN|MYG_BOVIN PSDFGADAQAAMSKALELFRNDMAAQYKVLGFHG 154
SP|sp|P02159|MYG_LYCPI|MYG_LYCPI SGDFHADTEAAMKKALELFRNDIAAKYKELGFQG 154
SP|sp|P02190|MYG_SHEEP|MYG_SHEEP PSDFGADAQGAMSKALELFRNDMAAQYKVLGFQG 154
SP|sp|P68082|MYG_HORSE|MYG_HORSE PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG 154
SP|sp|P02144|MYG_HUMAN|MYG_HUMAN PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
                                  .** **::.**.*******:*:*::** **::
```

## 3. The best nucleotide and amino acid models fitted to the data

### 3.1 Amino acid substitution models for all mammals

| Model | #Param | BIC | AICc | lnL | Invariant | Gamma |
|---|---|---|---|---|---|---|
| Dayhoff+G | 164 | 7024.6 | 5806.3 | -2737 | n/a | 0.49985 |
| Dayhoff+G+I | 165 | 7032.9 | 5807.1 | -2736.4 | 0.136066 | 0.67094 |
| JTT+G | 164 | 7040.8 | 5822.4 | -2745.1 | n/a | 0.47808 |
| JTT+G+I | 165 | 7049.4 | 5823.6 | -2744.6 | 0.145221 | 0.6543 |
| WAG+G | 164 | 7074.9 | 5856.5 | -2762.1 | n/a | 0.45491 |
| WAG+G+I | 165 | 7083.2 | 5857.4 | -2761.5 | 0.127089 | 0.58426 |
| rtREV+G | 164 | 7106.2 | 5887.8 | -2777.8 | n/a | 0.44271 |
| rtREV+G+I | 165 | 7113.5 | 5887.8 | -2776.7 | 0.155407 | 0.60074 |
| Dayhoff+I | 164 | 7163.5 | 5945.1 | -2806.4 | 0.363738 | n/a |
| JTT+I | 164 | 7182.6 | 5964.2 | -2816 | 0.374813 | n/a |
| cpREV+G | 164 | 7190.7 | 5972.4 | -2820 | n/a | 0.38501 |
| cpREV+G+I | 165 | 7196.4 | 5970.6 | -2818.1 | 0.209722 | 0.55634 |
| JTT+G+F | 183 | 7200.3 | 5841.3 | -2735 | n/a | 0.48313 |
| JTT+G+I+F | 184 | 7208.6 | 5842.2 | -2734.4 | 0.124288 | 0.61995 |
| Dayhoff+G+F | 183 | 7219.2 | 5860.3 | -2744.5 | n/a | 0.49825 |
| Dayhoff+G+I+F | 184 | 7227.6 | 5861.2 | -2743.9 | 0.117732 | 0.63561 |
| WAG+I | 164 | 7232.1 | 6013.7 | -2840.7 | 0.374551 | n/a |
| rtREV+I | 164 | 7269.4 | 6051 | -2859.4 | 0.374902 | n/a |
| rtREV+G+F | 183 | 7272 | 5913 | -2770.8 | n/a | 0.44466 |
| rtREV+G+I+F | 184 | 7279.2 | 5912.9 | -2769.7 | 0.149784 | 0.59206 |
| WAG+G+F | 183 | 7287.9 | 5928.9 | -2778.8 | n/a | 0.45483 |
| mtREV24+G | 164 | 7289.7 | 6071.3 | -2869.5 | n/a | 0.41432 |
| WAG+G+I+F | 184 | 7296.1 | 5929.7 | -2778.2 | 0.120563 | 0.56934 |

| | | | | | | |
|---|---|---|---|---|---|---|
| mtREV24+G+I | 165 | 7296.8 | 6071 | -2868.3 | 0.146911 | 0.54439 |
| mtREV24+G+F | 183 | 7314.9 | 5956 | -2792.3 | n/a | 0.43429 |
| mtREV24+G+I+F | 184 | 7323.5 | 5957.2 | -2791.9 | 0.093653 | 0.5111 |
| Dayhoff | 163 | 7328.7 | 6117.7 | -2893.7 | n/a | n/a |
| JTT+I+F | 183 | 7347.1 | 5988.1 | -2808.4 | 0.368742 | n/a |
| JTT | 163 | 7361.5 | 6150.5 | -2910.1 | n/a | n/a |
| Dayhoff+I+F | 183 | 7366.9 | 6007.9 | -2818.3 | 0.358554 | n/a |
| cpREV+I | 164 | 7373.1 | 6154.7 | -2911.2 | 0.395134 | n/a |
| cpREV+G+F | 183 | 7403.5 | 6044.5 | -2836.6 | n/a | 0.39089 |
| cpREV+G+I+F | 184 | 7406 | 6039.6 | -2833.1 | 0.266494 | 0.60489 |
| WAG | 163 | 7408.5 | 6197.5 | -2933.6 | n/a | n/a |
| rtREV+I+F | 183 | 7434 | 6075 | -2851.8 | 0.373158 | n/a |
| WAG+I+F | 183 | 7450.6 | 6091.7 | -2860.2 | 0.369681 | n/a |
| rtREV | 163 | 7465.3 | 6254.3 | -2962 | n/a | n/a |
| mtREV24+I | 164 | 7496.3 | 6277.9 | -2972.8 | 0.362499 | n/a |
| JTT+F | 182 | 7508.4 | 6156.8 | -2893.8 | n/a | n/a |
| mtREV24+I+F | 183 | 7509 | 6150.1 | -2889.4 | 0.357606 | n/a |
| Dayhoff+F | 182 | 7522 | 6170.5 | -2900.6 | n/a | n/a |
| cpREV | 163 | 7546.3 | 6335.3 | -3002.5 | n/a | n/a |
| cpREV+I+F | 183 | 7576.7 | 6217.7 | -2923.2 | 0.396104 | n/a |
| WAG+F | 182 | 7614.2 | 6262.6 | -2946.7 | n/a | n/a |
| rtREV+F | 182 | 7618.3 | 6266.7 | -2948.7 | n/a | n/a |
| mtREV24+F | 182 | 7675.5 | 6324 | -2977.3 | n/a | n/a |
| mtREV24 | 163 | 7723.3 | 6512.3 | -3091 | n/a | n/a |
| cpREV+F | 182 | 7731.3 | 6379.7 | -3005.2 | n/a | n/a |

## 3.2. Amino acid substitution models for the whale clade

| Model | #Param | BIC | AICc | lnL |
|---|---|---|---|---|
| Dayhoff+G | 18 | 1772.3 | 1676.6 | -820.08 |
| Dayhoff+G+I | 19 | 1779.5 | 1678.6 | -820.04 |
| JTT+G | 18 | 1784.5 | 1688.9 | -826.22 |
| JTT+G+I | 19 | 1791.9 | 1690.9 | -826.22 |
| WAG+G | 18 | 1796.7 | 1701 | -832.27 |
| rtREV+G | 18 | 1799.9 | 1704.2 | -833.87 |
| WAG+G+I | 19 | 1803.7 | 1702.7 | -832.12 |
| rtREV+G+I | 19 | 1806.4 | 1705.5 | -833.49 |
| Dayhoff | 17 | 1821.8 | 1731.4 | -848.5 |
| Dayhoff+I | 18 | 1829.1 | 1733.4 | -848.5 |
| cpREV+G | 18 | 1832.9 | 1737.3 | -850.41 |
| cpREV+I | 18 | 1837 | 1741.4 | -852.47 |
| cpREV+G+I | 19 | 1838.4 | 1737.4 | -849.47 |
| JTT | 17 | 1847.7 | 1757.4 | -861.48 |
| JTT+I | 18 | 1855.1 | 1759.4 | -861.48 |
| WAG | 17 | 1864.5 | 1774.1 | -869.87 |
| Dayhoff+G+F | 37 | 1868 | 1672.3 | -798.2 |
| WAG+I | 18 | 1871.8 | 1776.2 | -869.87 |
| JTT+G+F | 37 | 1873.7 | 1678 | -801.06 |
| Dayhoff+G+I+F | 38 | 1875.3 | 1674.4 | -798.2 |

| | | | | |
|---|---|---|---|---|
| rtREV | 17 | 1877.4 | 1787 | -876.3 |
| JTT+G+I+F | 38 | 1881 | 1680.1 | -801.06 |
| rtREV+I | 18 | 1884.7 | 1789.1 | -876.3 |
| mtREV24+G | 18 | 1886.1 | 1790.5 | -877.01 |
| rtREV+G+F | 37 | 1886.6 | 1690.9 | -807.53 |
| mtREV24+G+I | 19 | 1890.8 | 1789.9 | -875.7 |
| rtREV+G+I+F | 38 | 1893.7 | 1692.8 | -807.42 |
| mtREV24+G+F | 37 | 1900.7 | 1705 | -814.58 |
| WAG+G+F | 37 | 1901.7 | 1706 | -815.05 |
| mtREV24+G+I+F | 38 | 1902.7 | 1701.7 | -811.89 |
| WAG+G+I+F | 38 | 1908.9 | 1708 | -815.02 |
| cpREV | 17 | 1915.8 | 1825.4 | -895.5 |
| Dayhoff+F | 36 | 1925.3 | 1734.8 | -830.54 |
| cpREV+G+F | 37 | 1925.6 | 1729.9 | -827.02 |
| Dayhoff+I+F | 37 | 1932.6 | 1736.9 | -830.54 |
| cpREV+G+I+F | 38 | 1932.9 | 1732 | -827.02 |
| JTT+F | 36 | 1937.1 | 1746.7 | -836.46 |
| JTT+I+F | 37 | 1944.5 | 1748.8 | -836.45 |
| WAG+F | 36 | 1968.8 | 1778.4 | -852.3 |
| WAG+I+F | 37 | 1976.2 | 1780.5 | -852.3 |
| rtREV+F | 36 | 1978.4 | 1787.9 | -857.08 |
| mtREV24 | 17 | 1979.5 | 1889.1 | -927.35 |
| mtREV24+F | 36 | 1980.5 | 1790.1 | -858.16 |
| rtREV+I+F | 37 | 1985.7 | 1790 | -857.08 |
| mtREV24+I | 18 | 1986.8 | 1891.1 | -927.34 |
| mtREV24+I+F | 37 | 1987.9 | 1792.2 | -858.16 |
| cpREV+F | 36 | 1996.9 | 1806.5 | -866.34 |
| cpREV+I+F | 37 | 2004.3 | 1808.6 | -866.34 |

### 3.3. Nucleotide substitution models for the whale clade

| Model | #Param | BIC | AICc | lnL | Invariant | Gamma | R |
|---|---|---|---|---|---|---|---|
| T92+G+I | 21 | 2821.1 | 2686 | -1321.9 | 0.581175 | 0.22705 | 1.4379 |
| T92+G | 20 | 2833.5 | 2704.8 | -1332.3 | n/a | 0.09193 | 1.0517 |
| K2+G+I | 20 | 2834 | 2705.3 | -1332.5 | 0.603384 | 0.2703 | 1.3252 |
| HKY+G+I | 23 | 2834.9 | 2686.9 | -1320.3 | 0.578942 | 0.22564 | 1.4735 |
| TN93+G+I | 24 | 2839.4 | 2685 | -1318.4 | 0.58886 | 0.22944 | 1.4997 |
| K2+G | 19 | 2839.6 | 2717.3 | -1339.6 | n/a | 0.09298 | 0.9933 |
| JC+G | 18 | 2842.6 | 2726.7 | -1345.3 | n/a | 0.09076 | 0.5 |
| JC+I | 18 | 2843.8 | 2727.9 | -1345.9 | 0.788337 | n/a | 0.5 |
| JC+G+I | 19 | 2844.2 | 2721.9 | -1341.9 | 0.596368 | 0.27056 | 0.5 |
| HKY+G | 22 | 2846.3 | 2704.7 | -1330.3 | n/a | 0.09227 | 1.0712 |
| HKY+I | 22 | 2848.2 | 2706.6 | -1331.2 | 0.791017 | n/a | 1.0625 |
| TN93+G | 23 | 2850.5 | 2702.5 | -1328.1 | n/a | 0.10033 | 1.0612 |
| GTR+G+I | 27 | 2875.6 | 2701.9 | -1323.8 | 0.613329 | 0.3932 | 1.1952 |
| GTR+G | 26 | 2875.8 | 2708.5 | -1328.1 | n/a | 0.10037 | 1.0594 |
| T92 | 19 | 2973.3 | 2851 | -1406.4 | n/a | n/a | 0.916 |
| K2 | 18 | 2974.2 | 2858.4 | -1411.1 | n/a | n/a | 0.9147 |
| JC | 17 | 2977.7 | 2868.3 | -1417.1 | n/a | n/a | 0.5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| T92+I | 20 | 2982.1 | 2853.4 | -1406.6 | 0.00001 | n/a | 0.916 |
| K2+I | 19 | 2982.7 | 2860.4 | -1411.1 | 0.00001 | n/a | 0.9147 |
| HKY | 21 | 2987.1 | 2851.9 | -1404.9 | n/a | n/a | 0.9134 |
| TN93 | 22 | 2992.5 | 2850.9 | -1403.4 | n/a | n/a | 0.9175 |
| TN93+I | 23 | 2999.4 | 2851.5 | -1402.6 | 0.00001 | n/a | 0.9173 |
| GTR | 25 | 3016 | 2855.2 | -1402.5 | n/a | n/a | 0.9275 |
| GTR+I | 26 | 3022.9 | 2855.7 | -1401.7 | 0.00001 | n/a | 0.9275 |

## 4. CODEML output for ML estimation of dN/dS for the whole mammalian tree

TREE # 1: ((((((((((((6, 8), 1), (5, 9)), (2, 3)), (4, 7)), (20, (18, 31))), 19), (23, 24)), 33), 26), (((((21, 22), 28), 34), 29), ((16, 17), (14, (12, (13, (15, (10, 11)))))))), (25, 27)), 32, 30); MP score: 810

lnL(ntime: 65 np:130): -4872.649004 +0.000000

35..36 36..37 37..38 38..39 39..40 40..41 41..42 42..43 43..44 44..45 45..46 46..47 47..6 47..8 46..1 45..48 48..5 48..9 44..49 49..2 49..3 43..50 50..4 50..7 42..51 51..20 51..52 52..18 52..31 41..19 40..53 53..23 53..24 39..33 38..26 37..54 54..55 55..56 56..57 57..58 58..21 58..22 57..28 56..34 55..29 54..59 59..60 60..16 60..17 59..61 61..14 61..62 62..12 62..63 63..13 63..64 64..15 64..65 65..10 65..11 36..66 66..25 66..27 35..32 35..30

0.584809 0.074657 0.042898 0.000004 0.010941 0.034051 0.060302 0.110256 0.000004 0.013758 0.101162 0.000004 0.144838 0.007257 0.016031 0.090745 0.013375 0.019617 0.089965 0.013714 0.068115 0.053076 0.047203 0.000004 0.016442 0.136696 0.205366 0.072308 0.031778 0.294705 0.100081 0.271188 0.129317 0.429174 0.468055 0.000004 0.070835 0.048026 0.023908 0.409510 0.122394 0.107222 0.310555 0.403000 0.167701 0.000004 0.133405 0.153733 0.244236 0.133067 0.177681 0.062640 0.104159 0.040638 0.023365 0.006463 0.019474 0.008152 0.036874 0.014403 0.045353 0.274313 0.174368 0.316042 1.076215 0.030549 0.104180 0.082713 0.000100 999.000000 0.163035 0.048265 0.151570 0.000100 999.000000 0.623015 0.000100 0.397076 0.000100 0.133631 0.152928 0.183448 0.399032 0.358904 0.000100 0.265186 0.291258 0.137223 0.000100 0.137245 0.056099 0.157465 0.041776 0.147009 0.037426 0.107210 0.033873 0.328011 0.152701 0.034302 7.069095 0.086784 0.073461 999.000000 0.058117 0.078421 0.124204 0.053060 0.041973 0.021899 0.000100 0.032308 0.101873 0.104325 0.081870 0.122353 0.039095 0.014248 0.718144 0.000100 999.000000 0.096304 0.000100 0.000100 0.156876 0.255623 0.103774 0.084310 0.048824 0.080762

Note: Branch length is defined as number of nucleotide substitutions per codon (not per neucleotide site).

tree length = 8.45964

((((((((((((6: 0.144838, 8: 0.007257): 0.000004, 1: 0.016031): 0.101162, (5: 0.013375, 9: 0.019617): 0.090745): 0.013758, (2: 0.013714, 3: 0.068115): 0.089965): 0.000004, (4: 0.047203, 7: 0.000004): 0.053076): 0.110256, (20: 0.136696, (18: 0.072308, 31: 0.031778): 0.205366): 0.016442): 0.060302, 19: 0.294705): 0.034051, (23: 0.271188, 24: 0.129317): 0.100081): 0.010941, 33: 0.429174): 0.000004, 26: 0.468055): 0.042898, (((((21: 0.122394, 22: 0.107222): 0.409510, 28:

0.310555): 0.023908, 34: 0.403000): 0.048026, 29: 0.167701): 0.070835, ((16: 0.153733, 17: 0.244236): 0.133405, (14: 0.177681, (12: 0.104159, (13: 0.023365, (15: 0.019474, (10: 0.036874, 11: 0.014403): 0.008152): 0.006463): 0.040638): 0.062640): 0.133067): 0.000004): 0.000004): 0.074657, (25: 0.274313, 27: 0.174368): 0.045353): 0.584809, 32: 0.316042, 30: 1.076215);

((((((((((((P_b_whale: 0.144838, S_b_whale: 0.007257): 0.000004, L_b_whale: 0.016031): 0.101162, (M_h_whale: 0.013375, Dolphin: 0.019617): 0.090745): 0.013758, (S_whale: 0.013714, P_s_whale: 0.068115): 0.089965): 0.000004, (M_whale: 0.047203, Sei_whale: 0.000004): 0.053076): 0.110256, (Pig: 0.136696, (Sheep: 0.072308, Cow: 0.031778): 0.205366): 0.016442): 0.060302, Horse: 0.294705): 0.034051, (Cat: 0.271188, Dog: 0.129317): 0.100081): 0.010941, Microbat: 0.429174): 0.000004, Hedgehog: 0.468055): 0.042898, (((((Rat: 0.122394, Mouse: 0.107222): 0.409510, K_rat: 0.310555): 0.023908, Guinea_pig: 0.403000): 0.048026, Tree_shrew: 0.167701): 0.070835, ((Lemur: 0.153733, Galago: 0.244236): 0.133405, (Marmoset: 0.177681, (Macaque: 0.104159, (Gibbon: 0.023365, (Gorilla: 0.019474, (Human: 0.036874, Chimp: 0.014403): 0.008152): 0.006463): 0.040638): 0.062640): 0.133067): 0.000004): 0.000004): 0.074657, (Elephant: 0.274313, Hyrax: 0.174368): 0.045353): 0.584809, Platypus: 0.316042, Z_finch: 1.076215);

Detailed output identifying parameters

w (dN/dS) for branches: 0.03055 0.10418 0.08271 0.00010 999.00000 0.16303 0.04826 0.15157 0.00010 999.00000 0.62301 0.00010 0.39708 0.00010 0.13363 0.15293 0.18345 0.39903 0.35890 0.00010 0.26519 0.29126 0.13722 0.00010 0.13725 0.05610 0.15746 0.04178 0.14701 0.03743 0.10721 0.03387 0.32801 0.15270 0.03430 7.06910 0.08678 0.07346 999.00000 0.05812 0.07842 0.12420 0.05306 0.04197 0.02190 0.00010 0.03231 0.10187 0.10433 0.08187 0.12235 0.03909 0.01425 0.71814 0.00010 999.00000 0.09630 0.00010 0.00010 0.15688 0.25562 0.10377 0.08431 0.04882 0.08076

## 5. CODEML output for ML estimation of dN/dS for the whale clade of mammalian tree

TREE # 1: ((((1, 8), (5, 9)), (2, 3)), (6, (4, 7)), 10); MP score: 110
lnL(ntime: 17 np: 34): -1236.397819 +0.000000
11..12 12..13 13..14 14..1 14..8 13..15 15..5 15..9 12..16 16..2 16..3 11..17 17..6 17..18 18..4 18..7 11..10
0.000004 0.015419 0.108174 0.020475 0.000004 0.080726 0.013206 0.019818 0.099070 0.013827 0.068445 0.061192 0.000004 0.000004 0.040176 0.006770 0.338735 0.000100 999.000000 0.468419 0.092380 0.000100 0.215621 0.181910 0.403131 0.253993 0.000100 0.264662 0.239922 0.000100 0.000100 0.186734 0.000100 0.123082

Note: Branch length is defined as number of nucleotide substitutions per codon (not per neucleotide site).

tree length = 0.88605

((((1: 0.020475, 8: 0.000004): 0.108174, (5: 0.013206, 9: 0.019818): 0.080726): 0.015419, (2: 0.013827, 3: 0.068445): 0.099070): 0.000004, (6: 0.000004, (4: 0.040176, 7: 0.006770): 0.000004): 0.061192, 10: 0.338735);

((((L_b_Whale: 0.020475, S_b_whale: 0.000004): 0.108174, (M_h_whale: 0.013206, Dolphin: 0.019818): 0.080726): 0.015419, (S_whale: 0.013827, P_s_whale: 0.068445): 0.099070): 0.000004, (P_b_whale: 0.000004, (M_whale: 0.040176, Sei_whale: 0.006770): 0.000004): 0.061192, Human: 0.338735);

Detailed output identifying parameters

w (dN/dS) for branches: 0.00010 999.00000 0.46842 0.09238 0.00010 0.21562 0.18191 0.40313 0.25399 0.00010 0.26466 0.23992 0.00010 0.00010 0.18673 0.00010 0.12308

## 6. CODEML output for ML estimation of dN/dS for the terrestrial clade of mammalian tree

TREE # 1: (((((((((22, 9), 11), 10), (14, 15)), 24), 17), (((((13, 12), 19), 25), 20), ((8, 7), (5, (3, (4, (6, (2, 1)))))))), (16, 18)), 23, 21); MP score: 698

lnL(ntime: 0 np: 48): -4469.293861 +0.000000

2.132874 0.205803 0.177989 0.027162 0.000100 0.244110 0.023127 0.069289 0.239596 0.117675 0.052670 0.033534 0.098425 0.101625 0.071061 0.267009 0.295372 0.054535 0.176578 0.157013 0.107313 0.246148 0.090720 0.094597 0.141986 0.106614 0.065221 0.036812 0.000100 0.064889 0.117602 0.199427 0.134152 0.154916 0.035965 0.021182 0.414814 0.000100 999.000000 0.103703 0.000100 0.177970 0.000100 0.117072 0.130268 0.097307 0.055631 0.125422

tree length = 2.89227

(((((((((22: 0.014664, 9: 0.024835): 0.107857, 11: 0.033307): 0.024740, 10: 0.096967): 0.022258, (14: 0.061968, 15: 0.071931): 0.125947): 0.013695, 24: 0.159251): 0.020547, 17: 0.104168): 0.032203, (((((13: 0.059425, 12: 0.029436): 0.162509, 19: 0.077621): 0.017955, 25: 0.128501): 0.017460, 20: 0.044689): 0.011044, ((8: 0.125528, 7: 0.027376): 0.024531, (5: 0.079872, (3: 0.031680, (4: 0.006093, (6: 0.007188, (2: 0.006173, 1: 0.011604): 0.002617): 0.004348): 0.019334): 0.016000): 0.024475): 0.014916): 0.018029): 0.038019, (16: 0.112693, 18: 0.069882): 0.047069): 0.024887, 23: 0.161489, 21: 0.555492);

(((((((((Cow: 0.014664, Sheep: 0.024835): 0.107857, Pig: 0.033307): 0.024740, Horse: 0.096967): 0.022258, (Cat: 0.061968, Dog: 0.071931): 0.125947): 0.013695, Microbat: 0.159251): 0.020547, Hedgehog: 0.104168): 0.032203, (((((Mouse: 0.059425, Rat: 0.029436): 0.162509, K_rat: 0.077621): 0.017955, Guinea_pig: 0.128501): 0.017460, Tree_shrew: 0.044689): 0.011044, ((Galago: 0.125528, Lemur: 0.027376): 0.024531, (Marmoset: 0.079872, (Macaque: 0.031680, (Gibbon: 0.006093, (Gorilla: 0.007188, (Chimp: 0.006173, Human: 0.011604): 0.002617):

0.004348): 0.019334): 0.016000): 0.024475): 0.014916): 0.018029): 0.038019, (Elephant: 0.112693, Hyrax: 0.069882): 0.047069): 0.024887, Platypus: 0.161489, Z_finch: 0.555492);

Detailed output identifying parameters

kappa (ts/tv) = 2.13287

w (dN/dS) for branches: 0.20580 0.17799 0.02716 0.00010 0.24411 0.02313 0.06929 0.23960 0.11767 0.05267 0.03353 0.09843 0.10162 0.07106 0.26701 0.29537 0.05454 0.17658 0.15701 0.10731 0.24615 0.09072 0.09460 0.14199 0.10661 0.06522 0.03681 0.00010 0.06489 0.11760 0.19943 0.13415 0.15492 0.03597 0.02118 0.41481 0.00010 999.00000 0.10370 0.00010 0.17797 0.00010 0.11707 0.13027 0.09731 0.05563 0.12542

## 7. Likelihood ratio test for site models when branch lengths are estimated for each model rather taking the ML-estimated branch lengths from the M0 model

**Table C3**. LRT values for M7 vs. M8 and M8 vs. M8fix. Branch lengths are estimated by ML method for each model rather taking the ML-estimated under the M0 model.

| Clades | Model | ln L | 2Δl | P value | Positively selected sites (BEB: P(ω>1)>0.50) |
|---|---|---|---|---|---|
| **Cetaceans** | **Site models (number of parameters)** | | | | |
| | M7 | -1215.04 | | | - |
| | M8 | -1211.16 | (M7 vs. M8) 7.76 | 0.0206 | 5, 22, 35, 51, 66, 121, 129 |
| | M8fix | -1214.71 | (M8fix vs M8) 7.1 | 0.007 | |

## 8. The most probable cetacean ancestor with the complete phylogentic tree (Figure 2.3-B), primate-rodent truncated tree and just the cetacean clade

```
Truncated tree          MVLSDGEWQLVLNVWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASE 60
Cetacean clade          MVLSDAEWQLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASE 60
Complete tree           MVLSDGEWQLVLNVWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASE 60
                        *****.*******:**********************************************

Truncated tree          DLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
Cetacen clade           DLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRH 120
Complete tree           DLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH 120
                        ************************************************:**********

Truncated tree          PGDFGADAQAAMNKALELFRKDIAAKYKELGFQG 154
Cetacean clade          PGDFGADAQAAMNKALELFRKDIAAKYKELGFQG 154
Complete tree           PGDFGADAQAAMNKALELFRKDIAAKYKELGFQG 154
                         ***********************************
```

## 9. Evaluating the robustness of positive selection with the gene-tree rather organism-tree for cetacean Mbs

To evaluate the robustness of positive selection with the gene-tree rather the species-tree, we have used the Maximum Likelihood method based on the Dayhoff matrix based model to make the phylogeny as shown in Figure C3. The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. A discrete Gamma distribution was used to model evolutionary rate differences among sites (4 categories (+*G*, parameter = 0.6640)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 10 amino acid sequences. There were a total of 154 positions in the final dataset. Evolutionary analyses were conducted in MEGA5.

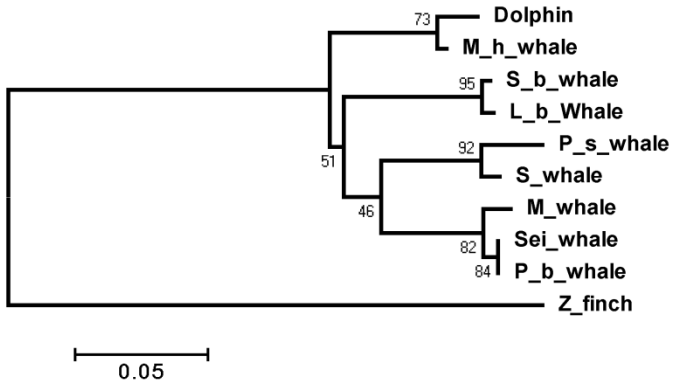



**Figure C3**. The gene-tree for the cetacean Mbs using the maximum likelihood estimation based on Dayhoff substitution model. Rate heterogeneity is allowed by using a discrete gamma distribution with four categories. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches.

Positive selection inferred in amino acid sites and judged by Likelihood ratio tests is also significant by using the gene-tree (see Table C4 below).

**Table C4**. LRT values for M7 vs. M8 and M8 vs. M8fix for the gene tree of cetaceans rather using the species tree.

| Model | ln L | 2Δl | P-value | Positively selected sites (BEB: P(ω>1)>0.50) |
|---|---|---|---|---|
| M7 | -1399.00 | | | --- |
| M8 | -1391.65 | (M7 vs. M8) 14.7 | 0.00064 | 5, 22, 35, 51, 66, 121, 129 |
| M8fix | -1396.86 | (M8fix vs M8) 10.42 | 0.0012 | --- |

**Table C5.** Species name and accession number of Mb sequences used in this study.

| Rank | Name in the phylogeny (full common name) | Species name | Accession number-Protein sequence[a] | Accession number-Nucleotide sequence |
|---|---|---|---|---|
| 1 | B_s_dolphin (Black sea dolphin) | Delphinus delphis | P68276 | *NA* |
| 2 | P_s_dolphin (Pantropical spotted dolphin) | Stenella attenuata | Q0KIY7 | BAF03580 [b] |
| 3 | A_bn_dolphin (Atlantic bottle-nosed dolphin) | Tursiops truncatus | P68279 | *NA* |
| 4 | H_porpoise (Harbor porpoise) | Phocoenoides | P68278 | *NA* |
| 5 | D_porpoise (Dall's porpoise) | Phocoenoides dalli dalli | P68277 | *NA* |
| 6 | M_h_whale (Melon-headed whale) | Peponocephala electra | Q0KIY3 | BAF03584 [b] |
| 7 | Lf_p_whale (Long-finned pilot whale) | Globicephala melas | P02174 | *NA* |
| 8 | K_whale (Killer whale) | Orcinus orca | P02173 | *NA* |
| 9 | A_dolphin (Amazon dolphin) | Inia geoffrensis | P02181 | *NA* |
| 10 | S_b_whale (Stejneger's beaked whale) | Mesoplodon stejnegeri | Q0KIY0 | BAF03587 [b] |

| | | | | |
|---|---|---|---|---|
| 11 | H_b_whale (Hubb's beaked whale) | Mesoplodon carlhubbsi | P02183 | *NA* |
| 12 | L_b_whale (Longman's beaked whale) | Indopacetus pacificus | Q0KIY9 | BAF03578 [b] |
| 13 | G_b_whale (Goose-beaked whale) | Ziphius cavirostris | P02182 | *NA* |
| 14 | P_b_whale (Pigmy Bryde's whale) | Balaenoptera edeni | Q0KIY2 | BAF03585 [b] |
| 15 | S_whale (Sperm whale) | Physeter macrocephalus | P02185 | BAF03579 [b] |
| 16 | P_s_whale (Pygmy sperm whale) | Kogia breviceps | Q0KIY5 | BAF03582 [b] |
| 17 | D_s_whale (Dwarf sperm whale) | Kogia sima | P02184 | *NA* |
| 18 | Hu_whale (Humpback whale) | Megaptera novaeangliae | P02178 | *NA* |
| 19 | Fi_whale (Finback whale) | Balaenoptera physalus | P02180 | *NA* |
| 20 | C_g_whale (California gray whale) | Eschrichtius gibbosus | P02177 | *NA* |
| 21 | Sei_whale (Sei whale) | Balaenoptera borealis | Q0KIY1 | BAF03586 [b] |
| 22 | C_m_whale (Common minke whale) | Balaenoptera acutorostrata | P02179 | BAF03583 [b] |
| 23 | Bovine (Bovine) | Bos taurus | P02192 | BAA00311 [b] |
| 24 | A_bison (American bison) | Bison bison | P86873 | *NA* |
| 25 | D_w_buffalo (Domestic water buffalo) | Bubalus bubalis | P84997 | *NA* |
| 26 | W_yak (Wild yak) | Bos mutus grunniens | Q2MJN4 | *NA* |
| 27 | Sheep (Sheep) | Ovis aries | P02190 | ABJ97274 [b] |
| 28 | R_deer (Red deer) | Cervus elaphus | P02191 | *NA* |
| 29 | Goat (Goat) | Capra hircus | B7U9B5 | *NA* |
| 30 | Pig (Pig) | Sus scrofa | P02189 | AAA31073 [b] |
| 31 | P_zebera (Plains zebra) | Equus burchelli | P68083 | *NA* |
| 32 | Horse (Horse) | Equus caballus | P68082 | NM_001164016.1 [c] |
| 33 | E_badger | Meles meles | P02157 | *NA* |
| 34 | G_seal (Gray seal) | Halichoerus grypus | P68081 | *NA* |
| 35 | H_seal (Harbor seal) | Phoca vitulina | P68080 | *NA* |
| 36 | B_seal (Baikal seal) | Phoca sibirica | P30562 | *NA* |
| 37 | Ca_sealion (California | Zalophus | P02161 | *NA* |

| | | | | |
|---|---|---|---|---|
| | sealion) | californianus | | |
| 38 | E_r_otter (Eueopean river otter) | Lutra lutra | P11343 | *NA* |
| 39 | Dog (Dog) | Canis familiaris | P63113 | *NA* |
| 40 | Cat (Cat) | Felis catus | *NU* | ENSFCAT00000010057 [d] |
| 41 | A_w_dog (African wild dog) | Lycaon pictus | P02159 | *NA* |
| 42 | C_fox (Cape fox) | Vulpes chama | P02160 | *NA* |
| 43 | B_e_folx (Bat-eared fox) | Otocyon megalotis | P63114 | *NA* |
| 44 | Microbat (Microbat) | Corynorhinus townsendii | *NU* | ENSMLUG00000013313 [c] |
| 45 | E_f_bat (Egyptian fruit bat) | Rousettus aegyptiacus | P02163 | *NA* |
| 46 | G_s_rat (Guaira spiny rat) | Proechimys guairae | P04249 | *NA* |
| 47 | P_viscacha (Plains viscacha) | Lagostomus maximus | P04250 | *NA* |
| 48 | N-gundi (Northern gundi) | Ctenodactylus gundi | P20856 | *NA* |
| 49 | E_beaver (Eurasian beaver) | Castor fiber | P14396 | *NA* |
| 50 | MBE_rat (Middle East blind mole rat) | Spalax ehrenbergi | P04248 | *NA* |
| 51 | Muskrat (Muskrat) | Ondatra zibethicus | P32428 | *NA* |
| 52 | Rat (Rat) | Rattus norvegicus | Q9QZ76 | ENSDORG00000014500 [c] |
| 53 | K_rat (Kangaroo rat) | Dipodomys | *NU* | ENSDORG00000014500 [c] |
| 54 | Ginea_pig (Ginea pig) | Cavia porcellus | *NU* | ENSCPOG00000006864 [c] |
| 55 | Mouse (Mouse) | Mus musculus | P04247 | ENSMUSG00000018893 [c] |
| 56 | S_a_pika (Southern American pika) | Ochotona princeps | P02171 | *NA* |
| 57 | B_l_pika (Black-lipped pika) | Ochotona curzoniae | Q6PL31 | *NA* |
| 58 | Rabbit (Rabbit) | Oryctolagus cuniculus | P02170 | *NA* |
| 59 | Chimpanzee (Chimpanzee) | Pan troglodytes | P02145 | ENSPTRG00000023553 [c] |
| 60 | Human (Human) | Homo sapiens | P02144 | ENSG00000198125 [c] |
| 61 | M_gorilla (Mountain gorilla) | Gorilla gorilla beringei | P02147 | ENSGGOG00000011478 [c] |
| 62 | B_orangutan (Bornean orangutan) | Pongo pygmaeus | P02148 | *NA* |
| 63 | Siamang (Siamang) | Hylobates syndactylus | P62735 | *NA* |

| | | | | |
|---|---|---|---|---|
| 64 | A_gibbon (Agile gibbon) | Hylobates agilis | P62734 | ENSNLEG00000014375 [c] |
| 65 | H_langur (Hanuman langur) | Semnopithecus entellus | P68085 | *NA* |
| 66 | R_guenon (Red guenon) | Erythrocebus patas | P68086 | *NA* |
| 67 | C_e_macaque (Crab-eating macaque) | Macaca fascicularis | P02150 | ENSMMUG00000005034 [c] |
| 68 | O_baboon (Olive baboon) | Papio anubis | P68084 | *NA* |
| 69 | B_w_monkey (Brown woolly monkey) | Lagothrix lagotricha | P02154 | *NA* |
| 70 | N_monkey (Night monkey) | Aotus trivirgatus | P02151 | *NA* |
| 71 | Wte_marmoset (White-tufted-ear marmoset) | Callithrix jacchus | P02152 | ENSCJAG00000000506 [c] |
| 72 | Cs_monkey (Common squirrel monkey) | Saimiri sciureus | P02155 | *NA* |
| 73 | B_c_capuchin (Brown-capped capuchin) | Cebus apella | P02153 | *NA* |
| 74 | G_galago (Greater galago) | Otolemur crassicaudatus | P02168 | ENSOGAG00000005651 [c] |
| 75 | S_ioris (Slow loris) | Nycticebus coucang | P02167 | *NA* |
| 76 | Potto (Potto) | Perodicticus potto edwarsi | P02166 | *NA* |
| 77 | W_lemur (Weasel sportive lemur) | Lepilemur mustelinus | P02169 | ENSMICG00000014107 [c] |
| 78 | T_shrew (Tree shrew) | Tupaia glis | P02165 | ENSTBEG00000002813 [c] |
| 79 | Aardvark (Aardvark) | Orycteropus afer | P02164 | |
| 80 | In_elephant (Indian elephant) | Elephas maximus | P02186 | ENSLAFG00000023176 [c] |
| 81 | A_elephant (African elephant) | Loxodonta africana | P02187 | *NA* |
| 82 | Hyrax (Hyrax) | Procavia capensis | *NU* | ENSPCAG00000003717 [c] |
| 83 | Na_opossum (North American opossum) | Didelphis marsupialis virginiana | P02193 | *NA* |
| 84 | R_Kangaroo (Red kangaroo) | Macropus rufus | P02194 | *NA* |
| 85 | A_echidna (Australian echidna) | Tachyglossus aculeatus aculeatus | P02195 | *NA* |

***Continued Table S5.***

| Rank | Name in the phylogeny (full common name) | Species name | Accession number-Protein sequence | Accession number-Nucleotide sequence |
|---|---|---|---|---|
| 86 | D_platypus (Duckbill platypus) | Ornithorhynchus anatinus | P02196 | ENSOANG00000010874 [c] |
| 87 | We_hedgehog (Western european hedgehog) | Erinaceus europaeus | P02156 | ENSEEUG00000005138 [c] |
| 88 | Z_finch (Zebra finch) | Taeniopygia guttata | H0ZKN4 | ENSTGUG00000010818 [c] |

*NU*=Not used, *NA*=Not Available, a: Taken from Uniprot database (Uniprot consortium, 2011), b: Take from EMBL database (Kanz et al., 2005), c:Taken from Ensembl genome browser (Hubbard et al., 2009).

# APPENDIX D: SUPPLEMENTARY DATA FOR SIMULATED EVOLUTION OF MB SEQUENCES

**Table D1.** $\Delta\Delta G$ matrix for single point mutations on sperm whale Mb calculated by ERIS (Yin et al., 2007). The rows shown in gray are the residues around the heme group in the apo-myoglobin which are highly conserved in all mammals. These residues are kept invariable by assigning $\Delta\Delta G$ = 100 kcal/mol for all possible mutations (except a mutation to itself) which gives *P*fix = 0 in the evolutionary dynamics.

| | Mutant amino acids | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Residue** | **A** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **K** | **L** | **M** | **N** | **P** | **Q** | **R** | **S** | **T** | **V** | **W** | **Y** |
| **1** | -0.81 | 7.87 | -0.83 | -0.47 | 2.08 | -0.56 | 0.76 | 0.93 | 0.05 | 0.34 | 0.67 | 0.00 | 1.04 | -0.31 | 0.59 | -0.82 | -0.36 | 0.00 | 2.46 | 2.01 |
| **2** | 3.30 | 11.78 | 3.40 | 5.20 | 20.06 | 4.59 | 10.76 | 2.68 | 3.64 | 0.00 | 1.97 | 2.14 | 4.14 | 4.50 | 13.37 | 3.22 | 4.28 | 2.94 | 43.74 | 35.21 |
| **3** | 0.93 | 9.46 | -0.98 | 0.31 | 2.93 | 0.13 | 1.85 | 1.31 | 0.48 | 0.73 | 1.58 | 0.71 | 5.64 | 1.73 | 0.55 | 0.00 | -0.06 | 1.57 | -0.81 | 2.93 |
| **4** | 2.67 | 12.02 | 3.11 | 0.00 | 6.11 | 3.57 | 4.69 | 4.07 | 2.83 | 3.71 | 3.92 | 3.36 | 4.74 | 3.11 | 5.13 | 3.13 | 3.32 | 3.48 | 6.20 | 5.59 |
| **5** | -1.34 | 7.84 | -0.83 | -1.24 | 0.42 | 0.00 | 0.46 | -0.21 | -0.84 | -0.18 | 0.81 | -0.71 | 2.18 | 0.12 | 0.65 | -0.67 | -0.63 | -0.17 | 1.56 | 0.62 |
| **6** | 3.74 | 13.58 | 2.81 | 0.00 | 5.70 | 5.79 | 4.95 | 3.83 | 4.97 | 2.53 | 5.48 | 4.25 | 0.95 | 3.03 | 4.75 | 4.93 | 3.38 | 3.47 | 6.04 | 5.50 |
| **7** | 4.01 | 13.55 | 4.73 | 4.03 | 1.20 | 5.79 | 2.95 | 5.22 | 5.13 | 2.71 | 3.75 | 3.94 | 12.44 | 4.19 | 5.07 | 4.46 | 5.52 | 5.26 | 0.00 | 2.24 |
| **8** | 0.38 | 10.18 | 0.45 | 0.33 | 3.01 | 1.93 | 2.81 | 1.53 | 1.60 | 1.33 | 1.98 | 0.54 | 12.07 | 0.00 | 0.05 | 0.17 | 0.37 | 1.16 | 2.97 | 2.65 |
| **9** | 0.22 | 9.90 | 0.54 | 0.53 | 2.26 | 1.78 | 0.76 | 0.06 | 0.48 | 0.00 | 0.07 | 0.15 | 11.90 | 0.89 | 1.15 | 0.12 | -0.08 | 0.58 | 1.75 | 2.33 |
| **10** | 2.34 | 12.28 | 4.64 | 8.41 | 21.73 | 4.95 | 14.81 | 1.89 | 13.04 | 3.65 | 7.13 | 2.52 | 14.51 | 8.14 | 21.10 | 1.93 | 0.58 | 0.00 | 39.39 | 27.13 |
| **11** | 2.51 | 11.79 | 2.55 | 2.61 | 5.90 | 3.54 | 4.51 | 2.15 | 2.92 | 0.00 | 2.75 | 0.93 | 13.42 | 1.70 | 2.69 | 3.01 | 1.58 | 2.72 | 5.51 | 5.66 |
| **12** | -2.39 | 7.56 | -1.56 | -1.52 | -0.04 | -0.56 | 0.00 | -0.49 | -1.74 | -1.43 | -1.21 | -1.60 | 12.54 | -1.25 | -1.13 | -2.33 | -2.31 | -0.86 | 0.90 | -1.97 |
| **13** | -1.21 | 8.93 | 1.73 | 4.15 | 11.74 | -0.04 | 10.10 | 7.64 | 5.62 | 4.08 | 3.12 | 1.24 | 6.02 | 3.89 | 16.13 | 0.46 | -0.49 | 0.00 | 28.64 | 14.81 |
| **14** | 2.51 | 13.11 | 4.31 | 2.65 | 2.41 | 4.36 | 2.98 | 2.38 | 3.34 | 0.62 | 2.29 | 3.02 | 17.78 | 2.54 | 3.85 | 2.73 | 1.95 | 2.42 | 0.00 | 2.54 |
| **15** | 0.00 | 8.82 | -0.62 | -0.40 | 1.20 | -0.09 | 0.12 | 0.88 | -0.30 | -0.26 | 0.76 | -0.33 | 12.65 | 0.08 | 0.33 | -1.52 | -1.15 | 0.38 | 1.70 | 1.18 |
| **16** | -0.51 | 8.88 | 0.01 | -0.12 | 0.67 | 1.08 | 0.21 | 0.05 | 0.00 | 0.41 | 0.07 | 0.30 | 8.35 | -0.08 | 1.35 | 1.13 | 0.34 | -0.63 | 1.91 | 0.26 |
| **17** | 1.11 | -0.37 | 2.23 | 1.26 | 6.76 | 3.12 | 3.20 | 2.39 | 8.19 | 0.03 | 4.76 | 0.68 | 10.94 | 1.46 | 12.26 | 2.73 | 0.70 | 0.00 | 15.75 | 19.30 |
| **18** | 1.81 | 11.62 | 1.23 | 0.00 | 4.51 | 2.21 | 4.10 | 3.03 | 2.15 | 1.00 | 2.37 | 1.29 | 13.09 | 0.91 | 2.59 | 1.83 | 2.44 | 3.71 | 4.00 | 4.93 |
| **19** | 0.00 | 9.04 | -0.44 | -0.51 | 1.96 | 0.26 | 0.91 | 1.50 | -0.50 | 0.34 | 1.18 | 0.15 | 14.26 | -0.33 | 0.54 | -0.44 | 0.60 | 1.85 | 3.38 | 2.05 |
| **20** | 3.69 | 12.47 | 0.00 | 3.50 | 6.32 | 3.77 | 4.98 | 5.51 | 4.77 | 5.99 | 5.28 | 2.06 | -0.47 | 4.85 | 5.05 | 3.59 | 4.98 | 5.35 | 8.40 | 5.69 |
| **21** | 0.19 | 9.75 | 1.31 | -0.11 | 6.98 | 2.25 | 3.47 | 0.08 | 0.67 | -1.11 | 0.62 | 1.07 | 2.71 | 0.39 | 0.24 | 1.66 | 0.36 | 0.00 | 11.44 | 6.98 |
| **22** | 0.00 | 9.50 | 1.90 | 0.43 | 4.37 | 0.57 | 3.08 | 5.45 | 0.62 | 2.72 | 0.60 | 3.46 | 2.33 | 1.12 | 1.65 | 0.85 | 2.23 | 2.90 | 5.68 | 4.53 |
| **23** | 5.53 | 13.86 | 5.93 | 5.83 | 10.31 | 0.00 | 6.45 | 11.56 | 6.78 | 9.85 | 7.04 | 5.81 | 21.99 | 5.40 | 6.70 | 5.32 | 6.15 | 11.67 | 9.81 | 8.92 |
| **24** | 3.27 | 12.62 | 4.08 | 4.88 | 0.60 | 4.67 | 0.00 | 3.78 | 6.49 | 4.42 | 3.79 | 2.53 | 12.75 | 4.05 | 11.09 | 3.76 | 2.75 | 2.79 | 9.40 | 6.45 |
| **25** | 3.93 | 17.89 | 25.30 | 39.10 | 49.65 | 0.00 | 43.42 | 33.13 | 47.32 | 38.20 | 25.86 | 25.75 | 17.28 | 38.46 | 52.17 | 5.09 | 12.80 | 17.31 | 81.99 | 57.13 |
| **26** | 0.80 | 10.39 | 2.30 | 0.13 | 13.94 | 2.43 | 3.96 | 10.05 | 0.57 | -1.31 | 1.11 | 1.16 | 12.64 | 0.00 | 1.13 | 0.87 | 2.48 | 2.90 | 19.21 | 30.29 |
| **27** | -0.41 | 9.34 | 0.00 | -0.21 | 1.12 | 1.53 | -0.12 | 2.54 | 0.07 | -0.60 | 0.30 | -0.83 | 10.78 | 0.49 | 0.93 | -0.99 | 0.16 | 2.10 | 2.13 | 1.10 |
| **28** | 3.15 | 12.23 | 5.55 | 5.90 | 8.17 | 4.85 | 6.02 | 0.00 | 6.39 | 4.88 | 2.08 | 4.81 | -0.45 | 4.91 | 11.35 | 2.34 | 1.62 | 0.83 | 15.01 | 11.23 |
| **29** | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| **30** | 1.11 | 11.03 | 3.27 | 1.98 | 12.24 | 2.92 | 5.64 | 0.00 | 4.40 | 1.25 | 2.43 | 2.90 | 10.87 | 1.92 | 4.14 | 1.39 | 0.51 | -0.04 | 18.40 | 13.14 |
| **31** | -0.83 | 8.49 | 0.38 | 0.16 | -0.66 | 0.47 | -0.32 | 0.71 | -0.72 | -1.53 | -0.03 | -1.04 | 10.99 | -0.68 | 0.00 | -0.97 | -1.46 | -0.62 | 0.79 | -0.68 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **32** | 3.46 | 12.31 | 4.31 | 2.42 | 4.48 | 4.91 | 3.43 | 5.53 | 2.49 | 0.00 | 1.94 | 3.45 | 15.63 | 1.69 | 2.68 | 3.49 | 2.20 | 2.43 | 2.88 | 10.57 |
| **33** | 4.45 | 13.35 | 4.86 | 5.30 | 0.00 | 6.13 | 2.45 | 2.74 | 5.86 | 3.09 | 3.23 | 4.13 | 15.23 | 4.50 | 10.87 | 4.48 | 4.93 | 3.31 | 5.20 | -0.21 |
| **34** | 0.91 | 10.99 | 1.58 | 1.01 | 3.63 | 2.03 | 2.92 | 1.35 | 0.00 | -0.06 | 1.87 | 2.29 | 12.29 | 0.55 | 1.77 | 0.99 | 1.18 | 1.41 | 4.43 | 3.27 |
| **35** | 1.81 | 11.78 | 0.18 | 0.43 | 3.54 | 3.05 | 3.26 | 2.69 | 2.05 | 2.44 | 1.81 | 3.13 | 14.20 | 0.03 | 1.02 | 0.00 | 0.05 | 1.98 | 3.69 | 3.43 |
| **36** | -0.20 | 9.02 | 0.35 | -1.29 | -0.53 | 1.69 | 0.00 | -0.14 | -0.60 | 2.46 | -0.39 | 0.03 | 0.07 | -0.31 | -0.22 | 0.54 | -0.69 | -1.15 | 1.09 | 0.59 |
| **37** | -1.21 | 8.46 | -0.60 | -0.88 | 0.47 | -0.01 | 0.21 | -0.04 | -0.58 | 0.08 | 0.18 | -0.04 | 0.00 | -0.28 | 0.14 | -0.62 | 0.05 | 0.34 | -0.41 | 0.08 |
| **38** | 0.65 | 9.61 | -0.05 | 0.00 | 2.92 | 2.01 | 2.27 | 2.55 | 0.66 | 0.85 | 1.40 | 0.88 | 3.13 | 1.01 | 1.42 | 0.31 | 2.13 | 2.39 | 3.94 | 2.53 |
| **39** | -0.60 | 9.65 | 2.95 | 6.11 | 2.13 | 1.63 | 1.51 | 0.39 | 3.39 | 4.41 | 1.04 | 2.42 | 12.87 | 3.38 | 6.02 | -1.03 | 0.00 | -0.71 | 12.57 | 2.01 |
| **40** | 1.21 | 11.57 | 2.97 | 0.88 | 8.57 | 2.78 | 7.14 | 2.23 | 0.67 | 0.00 | 1.63 | 1.50 | 8.35 | -1.26 | 1.35 | 0.99 | 2.80 | 2.82 | 12.71 | 8.12 |
| **41** | -0.23 | 9.59 | 0.24 | 0.00 | 2.20 | 1.64 | 1.81 | 2.08 | -0.15 | 1.07 | 1.05 | 0.33 | 7.52 | 1.07 | 1.81 | 0.14 | 0.83 | 1.57 | 2.59 | 1.10 |
| **42** | 1.27 | 10.65 | 2.75 | 0.71 | 6.04 | 2.45 | 3.34 | 3.13 | 0.00 | 2.73 | 1.07 | 2.53 | 10.69 | 0.51 | 1.17 | 0.25 | 1.56 | 2.84 | 8.30 | 6.88 |
| **43** | 100.00 | 100.00 | 100.00 | 100.00 | 0.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| **44** | 1.24 | 10.57 | 0.00 | 1.61 | 4.39 | 1.76 | 3.77 | 2.78 | 1.83 | 2.32 | 3.27 | 0.42 | 2.66 | 2.85 | 3.16 | 2.64 | 2.38 | 2.31 | 5.09 | 3.99 |
| **45** | -0.64 | 9.48 | -0.28 | -0.94 | 2.64 | 1.23 | 0.73 | 0.93 | -0.52 | -0.21 | 0.74 | 0.07 | 2.94 | -0.11 | 0.00 | 0.52 | -0.27 | -0.06 | 0.50 | 2.26 |
| **46** | 4.39 | 13.15 | 4.19 | 2.85 | 0.00 | 5.99 | 1.34 | 9.78 | 4.27 | 1.61 | 4.13 | 2.94 | 0.30 | 3.49 | 5.18 | 5.27 | 4.49 | 4.77 | 6.56 | 0.29 |
| **47** | 1.05 | 10.62 | 1.21 | 1.32 | 2.95 | 1.93 | 0.88 | 2.84 | 0.00 | 2.35 | 4.08 | 1.52 | 11.30 | 2.72 | 1.91 | 0.53 | 1.68 | 2.81 | 4.14 | 3.53 |
| **48** | -1.14 | 7.42 | -1.56 | -1.29 | 1.01 | -0.90 | 0.00 | -0.53 | -1.27 | -0.12 | -0.22 | -0.92 | 2.46 | -0.72 | 0.13 | -1.37 | -0.71 | -0.82 | 1.96 | 0.37 |
| **49** | 3.27 | 12.05 | 3.20 | 2.70 | 3.87 | 5.22 | 2.72 | 1.97 | 2.12 | 0.00 | 1.98 | 1.84 | 15.39 | 2.41 | 4.15 | 3.07 | 2.43 | 3.18 | 5.30 | 4.19 |
| **50** | 0.14 | 8.55 | -0.16 | 0.00 | 1.84 | 0.62 | 0.56 | 1.00 | 0.00 | 0.77 | 1.35 | -0.02 | -0.24 | 0.33 | 0.19 | -1.71 | -1.49 | 1.47 | 3.03 | 2.04 |
| **51** | 0.33 | 8.91 | -1.00 | -0.14 | 1.17 | 0.81 | 0.73 | 1.19 | -0.75 | 0.68 | 0.92 | -0.45 | 0.14 | -0.54 | -1.17 | 0.38 | 0.00 | 0.44 | 4.09 | 1.00 |
| **52** | 1.93 | 11.27 | 2.20 | 0.00 | 5.74 | 3.49 | 5.20 | 2.65 | 2.50 | 2.45 | 3.26 | 2.95 | 3.15 | 1.10 | 3.17 | 2.14 | 1.13 | 1.52 | 5.94 | 5.29 |
| **53** | 0.00 | 9.92 | 0.58 | 0.28 | 2.60 | 0.98 | 2.17 | 1.76 | 0.80 | 2.03 | 1.97 | 0.97 | 2.79 | 1.62 | 2.19 | 0.36 | 1.03 | 0.94 | 3.68 | 2.99 |
| **54** | 1.76 | 11.21 | 1.24 | 0.00 | 3.96 | 2.95 | 3.34 | 3.42 | 1.24 | 1.96 | 2.05 | 1.48 | 14.59 | 0.75 | 2.46 | 1.13 | 2.12 | 3.47 | 4.49 | 4.01 |
| **55** | 0.35 | 10.15 | 2.48 | -0.06 | -0.07 | 2.48 | -2.56 | 3.73 | 0.30 | -1.51 | 0.00 | -1.14 | 10.99 | -0.59 | 0.38 | 0.32 | 2.75 | 3.83 | 1.43 | 4.71 |
| **56** | 1.78 | 11.18 | 2.58 | 0.98 | 3.25 | 2.82 | 1.27 | 1.64 | 0.00 | 1.45 | 2.35 | 1.17 | 12.20 | 1.30 | 0.73 | 1.28 | 1.64 | 2.09 | 5.02 | 3.36 |
| **57** | 0.00 | 8.91 | -0.47 | 0.05 | 2.54 | 0.85 | 2.13 | 2.84 | -0.16 | 0.44 | 1.03 | 0.17 | 16.29 | 0.98 | 0.78 | -0.78 | 0.83 | 1.93 | 2.80 | 2.32 |
| **58** | 0.69 | 9.42 | -0.49 | -1.02 | 9.43 | 2.36 | 3.93 | 4.03 | 0.00 | -0.22 | 0.69 | -0.70 | 16.61 | 0.02 | 0.66 | 0.00 | -0.01 | 2.42 | 12.36 | 15.25 |
| **59** | -0.38 | 8.64 | -0.41 | 0.00 | 2.94 | 0.31 | 1.53 | 0.81 | 0.41 | 1.08 | 0.54 | 0.14 | 0.91 | 0.72 | 0.82 | -0.01 | 0.08 | 0.18 | 3.78 | 2.77 |
| **60** | -0.30 | 9.02 | 0.00 | -2.40 | 2.25 | 1.12 | 2.13 | 1.60 | 0.07 | 0.24 | 0.86 | 0.05 | 3.18 | 0.13 | 0.95 | 0.31 | 0.78 | 1.73 | 3.15 | 2.34 |
| **61** | 3.30 | 12.83 | 4.55 | 3.76 | 9.95 | 5.19 | 4.79 | 5.10 | 7.49 | 0.00 | 3.59 | 3.16 | 16.27 | 3.57 | 9.67 | 4.28 | 2.62 | 4.26 | 21.16 | 20.72 |
| **62** | 0.84 | 10.64 | 2.20 | 0.17 | 15.40 | 1.38 | 9.93 | 0.19 | 0.00 | 0.66 | 0.51 | 0.23 | 10.08 | -1.08 | -0.30 | 0.69 | -0.08 | 0.86 | 15.88 | 18.73 |
| **63** | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| **64** | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| **65** | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| **66** | 0.48 | 10.11 | 2.24 | 0.08 | 2.63 | 1.45 | 1.94 | 0.16 | -0.31 | 1.53 | 0.77 | 2.24 | 8.78 | 0.86 | 1.16 | 1.40 | 0.77 | 0.00 | 3.08 | 1.52 |
| **67** | -1.00 | 8.88 | -0.96 | -3.08 | 0.38 | 0.55 | 0.56 | -1.26 | -0.94 | -0.51 | 0.43 | -0.79 | 11.07 | -2.01 | 0.18 | -0.12 | 0.00 | -0.91 | 0.40 | -0.13 |
| **68** | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.00 | 100.00 | 100.00 |
| **69** | 3.51 | 12.43 | 3.14 | 3.60 | 9.27 | 4.85 | 5.91 | 2.41 | 5.73 | 0.00 | 2.86 | 1.91 | 12.31 | 2.30 | 11.32 | 3.37 | 2.39 | 3.87 | 17.57 | 19.43 |
| **70** | 0.02 | 9.28 | 0.75 | -0.02 | 2.67 | 1.17 | 2.04 | 1.68 | 0.50 | 0.39 | 1.38 | 0.84 | 13.24 | 1.49 | 0.79 | 0.20 | 0.00 | 0.98 | 3.92 | 2.60 |
| **71** | 0.00 | 10.13 | 1.11 | 0.26 | 1.89 | 1.92 | 0.19 | 5.03 | 1.11 | 1.10 | 0.82 | 0.12 | 10.82 | 0.82 | 0.98 | 0.20 | 1.90 | 3.76 | 2.86 | 1.50 |
| **72** | 3.32 | 13.01 | 3.76 | 3.48 | 2.91 | 5.26 | 3.76 | 0.68 | 3.29 | 0.00 | 3.08 | 2.12 | 13.46 | 2.70 | 5.19 | 3.20 | 1.64 | 1.78 | 5.88 | 1.76 |
| **73** | 1.28 | 12.52 | 10.26 | 8.40 | 15.97 | 0.00 | 17.92 | 20.94 | 13.81 | 15.15 | 6.79 | 9.54 | 19.26 | 14.80 | 13.33 | 2.87 | 15.43 | 10.28 | 20.04 | 15.01 |
| **74** | 0.00 | 9.12 | -0.25 | -0.64 | 2.01 | 1.26 | 0.73 | 0.77 | 0.07 | 0.58 | 1.04 | -0.74 | 10.91 | 0.52 | -1.16 | -0.35 | -1.03 | 0.57 | 2.74 | 2.13 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **75** | 2.60 | 11.52 | 3.93 | 5.18 | 11.22 | 4.78 | 7.85 | 0.00 | 6.91 | 2.53 | 3.01 | 1.50 | -0.33 | 4.86 | 7.51 | 3.54 | 2.43 | 0.71 | 22.08 | 22.36 |
| **76** | 3.58 | 12.74 | 3.87 | 2.16 | -0.25 | 5.38 | 1.25 | 2.98 | 4.40 | 0.00 | 2.32 | 2.82 | 12.27 | 1.35 | 9.15 | 4.51 | 4.21 | 4.68 | 3.21 | 4.24 |
| **77** | 1.24 | 11.09 | 3.05 | 1.69 | 1.60 | 2.48 | 0.70 | 3.44 | 0.00 | 1.65 | 1.75 | 0.36 | 11.55 | 2.05 | 1.59 | 0.67 | 2.02 | 3.03 | 2.66 | 1.28 |
| **78** | 0.18 | 10.06 | -0.17 | -0.36 | 1.50 | 1.43 | 1.11 | 10.54 | 0.00 | 0.64 | 1.06 | -0.19 | 16.41 | -0.36 | 0.57 | 0.16 | 3.66 | 5.64 | 3.89 | 1.36 |
| **79** | 3.25 | 11.78 | 2.62 | 1.82 | 5.48 | 2.91 | 3.27 | 0.00 | 0.00 | 1.93 | 3.23 | 1.61 | 0.13 | 1.99 | 2.81 | 3.35 | -0.09 | 0.07 | 0.15 | 7.52 |
| **80** | 0.41 | 0.21 | 0.31 | -0.03 | 0.14 | 0.00 | 0.19 | 0.18 | 0.18 | 0.24 | 0.57 | 0.25 | 4.87 | 0.10 | 0.23 | 0.86 | 0.21 | 0.15 | 0.36 | 0.07 |
| **81** | -0.94 | 8.16 | -2.09 | -1.45 | 0.77 | -1.42 | 0.00 | 0.29 | -1.38 | -0.51 | -0.16 | -1.08 | 14.47 | -0.94 | -0.53 | -1.70 | -0.79 | -0.35 | 2.75 | 1.08 |
| **82** | -0.71 | 9.53 | 0.80 | 0.81 | 2.74 | 1.34 | 0.00 | 3.94 | 2.91 | 19.89 | 4.88 | -0.11 | -0.22 | 0.16 | 0.22 | 0.46 | 2.30 | 2.34 | 14.37 | 6.40 |
| **83** | 0.54 | 0.11 | 1.41 | 0.00 | 4.81 | 1.89 | 3.21 | 1.94 | 1.84 | 1.56 | 2.39 | 2.16 | 5.10 | 1.92 | 2.84 | 1.97 | 1.66 | 1.76 | 5.04 | 4.29 |
| **84** | 0.00 | 8.63 | -0.68 | -0.57 | 2.44 | 0.33 | 1.40 | 1.05 | 0.08 | 1.24 | 0.69 | -0.20 | 2.66 | 0.40 | 0.28 | -0.47 | 0.23 | 0.70 | 3.25 | 2.47 |
| **85** | 0.32 | 8.88 | -0.12 | 0.00 | 1.81 | 2.14 | 0.85 | 0.11 | 0.12 | 0.43 | 0.10 | -0.02 | 0.16 | -0.11 | 0.92 | 0.99 | 0.32 | -0.19 | 1.88 | 1.47 |
| **86** | 1.24 | 11.31 | 3.01 | 2.46 | 17.61 | 2.91 | 10.61 | -0.02 | 3.72 | 0.00 | 1.75 | 1.60 | -0.29 | 2.15 | 1.86 | 0.43 | -0.89 | 0.87 | 12.48 | 33.36 |
| **87** | 1.15 | 10.68 | 1.18 | -0.05 | 4.06 | 1.50 | 3.61 | 1.74 | 0.00 | 0.73 | 1.59 | 1.39 | 14.33 | 1.43 | 0.92 | 0.27 | -0.55 | 1.73 | 5.64 | 6.04 |
| **88** | -2.84 | 6.61 | -2.73 | -2.51 | -0.59 | -2.47 | -0.44 | -1.44 | -1.97 | -1.54 | -1.46 | -1.86 | 0.00 | -1.98 | -1.75 | -1.46 | -2.03 | -2.45 | 0.00 | -1.18 |
| **89** | 0.40 | 9.91 | 1.22 | 0.24 | 0.58 | 1.62 | 0.68 | -1.12 | 1.39 | 0.00 | 0.26 | 0.59 | 0.03 | -0.25 | 1.00 | 1.32 | 0.68 | -0.42 | 2.74 | 0.14 |
| **90** | 0.00 | 11.90 | 13.91 | 10.42 | 38.35 | 2.15 | 33.99 | 4.48 | 6.70 | 14.32 | 5.15 | 12.48 | 10.37 | 11.00 | 5.59 | 1.86 | 3.25 | 4.18 | 52.15 | 46.20 |
| **91** | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| **92** | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| **93** | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| **94** | 0.00 | 12.20 | 13.72 | 22.43 | 67.14 | 1.32 | 22.27 | 18.60 | 22.31 | 14.14 | 14.27 | 20.07 | 13.93 | 22.90 | 19.93 | 0.13 | 7.80 | 12.86 | 36.41 | 87.87 |
| **95** | 2.05 | 10.65 | 1.71 | 0.65 | 5.17 | 3.45 | 2.77 | 2.25 | 1.45 | 3.15 | 2.50 | 1.97 | -0.08 | 1.79 | 2.01 | 0.99 | 0.00 | 2.30 | 15.77 | 5.38 |
| **96** | 0.68 | 9.22 | 0.33 | -0.29 | 2.25 | 1.56 | 1.42 | 0.71 | 0.00 | 0.74 | 1.84 | 1.28 | 10.27 | -0.40 | 0.53 | 0.71 | 0.52 | 0.31 | 4.00 | 1.93 |
| **97** | 0.05 | 9.53 | -0.04 | -0.35 | 0.57 | 1.09 | 0.00 | 1.44 | 0.06 | 0.00 | -0.36 | 0.05 | 9.27 | -0.34 | -0.56 | 1.00 | 0.88 | 1.33 | 0.33 | -0.52 |
| **98** | 0.57 | 9.74 | 0.66 | -0.10 | 1.68 | 0.77 | 0.36 | -0.34 | 0.00 | 0.85 | 1.67 | -0.28 | -0.16 | 0.06 | 0.09 | 0.87 | 1.36 | -0.22 | -0.30 | 1.91 |
| **99** | 2.52 | 11.06 | 2.28 | 1.95 | 1.12 | 3.93 | 1.85 | 0.00 | 1.75 | 0.72 | 1.71 | 2.37 | -0.49 | 1.18 | 1.35 | 3.39 | 1.61 | 0.05 | 4.81 | 1.93 |
| **100** | 0.66 | 9.88 | 0.01 | 0.42 | 2.86 | 2.23 | 2.17 | 0.66 | 0.76 | 1.01 | 1.46 | 0.21 | 0.00 | 0.52 | 1.34 | 0.83 | 0.85 | 0.86 | 2.42 | 2.10 |
| **101** | 0.45 | 10.43 | 1.11 | 1.00 | 2.80 | 2.08 | 2.34 | 0.00 | 0.20 | 0.32 | 2.15 | 1.39 | -0.35 | 0.82 | 1.76 | 1.26 | 1.21 | 0.57 | 6.24 | 2.83 |
| **102** | 0.86 | 9.62 | 0.32 | 0.01 | 2.89 | 2.09 | 1.98 | 0.87 | 0.00 | 0.64 | 1.11 | 0.40 | 4.53 | 0.65 | 1.13 | 0.76 | 0.87 | 0.85 | 4.15 | 2.40 |
| **103** | 2.57 | 12.03 | 3.22 | 2.12 | 0.17 | 4.37 | 1.78 | 1.17 | 3.35 | 1.69 | 1.66 | 1.97 | 10.98 | 1.64 | 4.70 | 3.70 | 3.20 | 2.07 | 6.28 | 0.00 |
| **104** | 2.13 | 11.55 | 3.27 | 1.83 | 0.43 | 3.42 | 0.62 | 0.51 | 2.25 | 0.00 | 1.18 | 1.04 | 10.65 | 1.24 | 2.66 | 3.05 | 2.47 | 0.71 | -0.13 | 4.50 |
| **105** | 1.68 | 11.29 | 1.26 | 0.00 | 3.34 | 2.15 | 2.47 | 10.71 | 1.05 | 1.28 | 2.80 | 1.98 | 14.81 | 0.75 | 2.11 | 2.72 | 3.88 | 5.37 | 2.91 | 3.13 |
| **106** | 0.54 | 9.17 | 0.18 | 0.28 | 0.00 | 1.45 | -1.40 | 0.08 | 0.07 | -0.59 | 0.37 | -0.21 | 9.65 | -0.03 | 0.19 | 1.21 | 1.01 | 0.70 | 0.27 | -1.01 |
| **107** | 1.07 | 10.62 | 2.43 | 1.24 | 2.57 | 3.22 | 2.62 | 0.00 | 2.10 | -0.71 | 1.29 | 1.72 | 10.67 | 1.26 | 1.69 | 2.59 | 1.12 | 1.29 | 10.15 | 2.83 |
| **108** | 0.12 | 9.94 | 0.97 | 1.10 | 1.32 | 1.74 | 0.33 | 0.26 | 3.19 | 0.06 | 1.90 | 0.79 | 15.47 | 1.37 | 2.89 | 0.00 | 2.38 | 1.83 | 17.02 | 3.12 |
| **109** | -0.14 | 9.44 | 0.02 | 0.00 | 1.03 | 0.94 | -0.02 | 0.27 | 0.10 | 0.27 | 0.58 | 0.12 | 10.06 | 0.48 | -0.64 | 0.14 | 0.95 | 0.51 | 1.95 | 0.66 |
| **110** | 0.00 | 10.35 | 5.74 | 9.99 | 10.90 | 2.52 | 9.49 | 5.23 | 13.97 | 11.40 | 2.47 | 5.63 | 11.44 | 8.44 | 22.17 | 0.11 | 3.01 | 3.37 | 32.18 | 13.33 |
| **111** | 3.74 | 12.77 | 4.61 | 3.71 | 4.03 | 4.97 | 4.63 | 0.00 | 2.71 | 0.36 | 2.28 | 3.98 | 15.49 | 3.10 | 3.23 | 3.48 | 1.30 | 1.19 | 8.08 | 7.14 |
| **112** | 1.48 | 11.14 | 2.25 | 1.30 | 5.74 | 3.55 | 4.08 | 0.00 | 0.03 | 0.37 | 1.45 | 1.04 | 10.13 | 0.38 | 0.49 | 1.59 | 1.49 | 0.69 | 8.67 | 14.93 |
| **113** | -0.72 | 7.93 | -1.56 | -1.71 | 0.53 | 0.28 | 0.00 | -0.05 | -0.15 | -0.17 | -0.65 | -1.65 | 10.13 | -0.70 | 0.20 | -1.46 | -1.26 | -1.36 | 1.40 | 0.11 |
| **114** | 1.14 | 10.97 | 2.27 | 1.27 | 14.29 | 2.84 | 7.33 | 0.20 | 2.45 | 4.94 | 2.05 | 1.90 | 10.94 | 1.63 | 1.60 | 1.11 | 0.48 | 0.00 | 15.86 | 20.52 |
| **115** | 1.38 | 10.83 | 1.54 | 0.49 | -0.67 | 3.07 | 1.08 | 9.86 | 3.73 | 0.00 | 0.85 | 0.78 | 11.07 | 0.77 | 10.65 | 2.10 | 2.21 | 4.25 | 4.53 | 1.04 |
| **116** | -0.79 | 7.93 | -0.50 | -1.31 | 0.08 | 0.03 | 0.00 | -1.03 | -1.33 | -1.57 | -0.78 | -0.63 | 9.35 | -0.12 | -1.25 | -2.12 | -2.05 | -1.98 | 0.96 | -0.29 |
| **117** | 0.60 | 9.64 | 0.27 | 0.61 | 2.09 | 1.84 | 1.47 | 1.10 | 0.60 | 1.07 | 1.58 | 0.73 | 14.37 | 1.63 | 1.33 | 0.00 | 0.28 | 1.60 | 2.54 | 1.18 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **118** | 1.57 | 10.67 | 1.38 | 1.54 | 4.24 | 3.21 | 1.09 | 1.65 | 1.66 | 0.27 | 1.84 | 0.51 | -0.19 | 0.98 | 0.00 | -0.65 | -1.41 | 1.83 | 4.26 | 5.45 |
| **119** | 2.65 | 9.97 | 1.95 | 1.55 | 0.85 | 0.83 | 0.00 | 3.01 | 1.07 | 1.08 | 0.97 | 0.77 | 0.60 | 0.71 | 0.50 | 2.45 | 3.44 | 1.40 | 0.18 | 0.37 |
| **120** | -0.25 | -0.03 | 0.71 | -0.81 | 2.43 | 0.10 | 1.23 | 1.05 | 0.17 | 0.77 | 1.18 | 0.80 | 0.00 | 0.88 | 0.76 | 0.07 | 0.40 | 0.33 | 4.04 | 2.11 |
| **121** | -0.69 | 8.15 | -1.87 | -1.09 | 1.97 | 0.00 | 1.17 | 0.31 | -0.92 | -0.10 | 0.63 | -0.55 | 3.43 | 0.52 | 0.57 | -0.31 | -0.26 | 0.13 | 2.14 | 0.71 |
| **122** | 0.14 | 9.17 | 0.00 | -0.57 | 0.37 | 1.56 | 0.36 | 3.33 | -0.47 | 0.41 | 0.79 | -0.14 | -0.35 | 0.33 | 0.37 | 1.16 | 0.79 | 0.45 | 2.04 | 0.54 |
| **123** | 2.15 | 11.05 | 3.27 | 2.75 | 0.00 | 2.78 | 0.90 | 1.56 | 2.39 | 6.34 | 1.01 | 1.52 | -0.22 | 2.19 | 2.85 | 2.33 | 3.01 | 1.52 | 0.41 | 2.66 |
| **124** | 0.29 | 9.16 | -1.09 | -0.08 | 1.82 | 0.00 | 1.27 | 1.01 | 0.13 | 1.06 | 1.45 | -0.66 | 1.59 | 0.63 | 0.15 | 0.18 | 0.36 | 0.74 | 0.37 | -0.48 |
| **125** | 0.00 | 9.42 | -0.22 | -0.47 | 2.39 | 0.31 | 1.64 | 0.69 | 0.49 | 0.86 | 1.13 | 0.88 | 1.41 | 0.97 | 1.74 | 0.56 | 0.84 | 0.53 | 3.71 | 2.91 |
| **126** | -0.10 | 9.45 | 0.00 | -0.17 | 3.15 | 0.97 | 2.37 | 0.80 | 0.16 | 0.84 | 1.32 | 0.40 | 3.00 | 1.84 | 1.46 | 0.51 | 0.36 | 0.56 | 4.89 | 2.91 |
| **127** | 0.00 | 8.81 | 1.09 | -0.24 | -0.07 | 1.66 | 0.78 | 2.00 | 4.87 | 4.71 | 2.05 | 0.13 | 2.49 | 0.63 | 4.64 | 1.13 | 1.03 | 1.03 | 3.20 | -0.40 |
| **128** | 0.99 | 10.39 | 0.63 | 0.44 | 4.09 | 2.31 | 1.79 | 0.33 | 0.61 | -0.33 | 0.47 | 0.59 | 13.15 | 0.00 | 0.90 | 1.01 | 1.50 | 0.60 | 3.43 | 4.67 |
| **129** | -1.11 | 7.93 | -1.37 | -1.43 | 1.92 | 0.00 | -0.14 | -0.16 | -0.77 | -0.70 | 0.34 | -1.42 | 10.88 | -0.18 | -1.60 | -1.77 | -1.66 | -0.11 | 1.81 | 1.43 |
| **130** | 0.00 | 11.34 | 10.50 | 11.83 | 16.03 | 2.97 | 15.77 | 9.35 | 11.71 | 13.02 | 5.51 | 9.33 | 11.40 | 11.35 | 7.47 | 0.84 | 3.67 | 6.16 | 21.01 | 15.92 |
| **131** | 1.39 | 10.53 | 2.85 | 1.17 | 0.75 | 2.91 | 1.67 | -0.44 | 2.41 | -1.85 | 0.00 | 1.06 | 11.27 | 0.08 | 3.54 | 1.02 | -0.23 | -0.25 | 3.32 | 3.17 |
| **132** | 0.17 | 9.85 | 0.65 | 0.15 | 3.50 | 1.53 | 2.17 | 0.90 | -0.60 | 0.31 | 0.85 | 0.00 | 12.91 | -0.57 | -0.19 | -0.18 | -0.40 | 1.15 | 2.05 | 2.83 |
| **133** | 0.79 | 10.68 | 1.13 | 1.58 | 3.10 | 2.24 | 1.96 | 1.93 | 0.00 | 1.20 | 1.85 | 1.18 | 12.13 | 0.57 | 1.64 | 1.23 | 0.62 | 2.49 | 3.69 | 2.74 |
| **134** | 0.00 | 13.01 | 15.45 | 14.53 | 28.03 | 3.00 | 25.96 | 11.61 | 16.64 | 14.89 | 9.74 | 12.71 | 13.92 | 14.13 | 18.27 | 1.17 | 6.77 | 6.43 | 46.51 | 32.35 |
| **135** | 3.46 | 12.27 | 3.94 | 3.41 | 1.41 | 4.56 | 2.21 | 5.30 | 5.40 | 0.00 | 2.08 | 2.81 | 14.03 | 3.52 | 6.43 | 3.09 | 3.51 | 5.32 | 11.12 | 2.90 |
| **136** | 1.01 | 10.07 | 1.25 | 0.00 | 2.84 | 2.14 | 2.53 | 1.87 | 1.17 | 1.33 | 2.01 | 0.91 | 14.51 | 1.44 | 2.14 | 0.23 | -0.06 | 2.38 | 3.68 | 2.67 |
| **137** | 0.97 | 10.53 | 1.34 | 1.01 | 1.91 | 1.90 | 1.05 | 2.36 | 0.68 | 0.00 | 1.07 | 1.50 | 14.87 | -0.47 | 0.33 | 1.09 | 0.25 | 1.75 | 1.41 | 1.73 |
| **138** | 2.47 | 11.79 | 3.64 | 3.09 | 0.00 | 4.13 | 2.75 | 2.10 | 2.20 | -0.42 | 1.56 | 2.29 | 15.47 | 2.37 | 2.06 | 1.28 | 0.40 | 0.90 | 1.16 | 2.34 |
| **139** | 0.16 | 9.45 | 0.63 | -0.26 | 10.06 | 1.32 | 5.41 | -1.81 | -0.09 | -1.02 | 0.58 | -1.11 | 11.57 | -0.41 | 0.00 | -0.14 | -0.61 | -0.41 | 8.93 | 15.12 |
| **140** | -0.83 | 8.53 | 0.13 | 0.04 | 1.64 | 0.38 | 0.30 | 0.52 | 0.00 | -0.27 | 0.82 | -0.54 | 13.15 | 0.87 | 0.86 | -0.76 | -0.93 | 0.36 | 2.24 | 1.68 |
| **141** | -0.84 | 8.99 | 0.00 | -1.32 | 0.57 | 0.64 | -0.65 | -0.36 | 0.05 | 1.10 | -0.10 | -1.21 | 9.53 | -0.40 | 0.78 | -0.53 | -1.46 | -1.20 | 2.30 | 0.77 |
| **142** | 1.84 | 11.56 | 3.90 | 2.41 | 4.84 | 4.66 | 0.98 | 0.00 | 2.78 | 0.18 | 1.87 | 1.64 | 11.46 | 1.96 | 5.07 | 2.16 | 1.18 | 0.04 | 17.26 | 8.02 |
| **143** | 0.00 | 12.27 | 3.36 | 2.70 | 19.04 | 1.67 | 20.66 | 6.07 | 3.11 | 3.65 | 2.94 | 4.74 | 15.47 | 3.20 | 3.83 | 1.70 | 6.39 | 5.63 | 26.50 | 19.69 |
| **144** | 0.00 | 9.22 | -0.26 | -0.78 | 1.32 | 0.82 | 0.39 | 0.01 | -0.24 | 0.06 | 0.77 | 0.05 | 11.08 | 0.54 | 0.14 | -0.07 | -0.37 | 0.01 | 3.45 | 1.60 |
| **145** | 1.30 | 10.92 | 2.22 | 1.41 | 0.91 | 2.87 | 0.56 | -0.25 | 0.00 | 0.75 | 1.42 | 1.36 | 11.29 | -0.49 | 2.23 | 2.60 | 1.65 | 0.39 | 3.21 | 1.21 |
| **146** | 3.07 | -0.70 | 3.57 | 1.91 | 0.41 | 5.11 | 1.36 | 7.46 | 2.58 | 0.99 | 1.73 | 2.13 | 13.30 | 0.72 | 2.77 | 4.14 | 6.29 | 7.97 | 7.49 | 0.00 |
| **147** | -0.56 | 9.52 | -0.02 | -0.66 | 1.55 | 0.35 | 1.01 | 1.98 | 0.00 | -0.36 | 0.75 | -0.71 | 12.76 | 0.24 | -0.63 | 0.32 | 0.18 | 2.16 | 2.20 | 1.18 |
| **148** | 1.27 | 10.20 | 0.48 | 0.00 | 2.55 | 1.90 | 2.42 | 1.19 | 1.12 | 2.02 | 1.75 | 1.05 | 14.61 | 0.76 | 2.39 | 0.93 | 0.84 | 1.29 | 3.89 | 2.64 |
| **149** | 1.79 | 11.28 | 2.06 | 1.62 | 2.37 | 3.10 | 1.64 | 3.00 | 1.96 | 0.00 | 1.48 | 1.42 | 12.53 | 1.18 | 3.10 | 2.48 | 1.85 | 2.90 | 4.39 | 1.65 |
| **150** | 1.51 | 0.23 | 1.29 | 0.14 | -0.07 | 0.00 | 0.36 | 0.08 | 1.72 | 0.19 | 0.05 | 1.50 | 0.57 | 0.14 | 0.41 | 0.06 | -0.13 | 0.01 | 0.16 | 0.01 |
| 151 | 1.51 | 9.47 | 0.78 | 0.85 | 0.12 | 2.30 | -0.69 | 7.15 | 0.63 | 2.20 | 1.61 | 0.94 | 0.21 | 1.11 | 0.99 | 1.73 | 1.03 | 3.40 | 1.00 | 0.00 |
| 152 | -0.44 | 9.56 | 0.02 | -0.62 | 0.29 | 0.43 | 0.59 | -0.47 | -0.19 | -0.94 | 0.47 | 0.33 | 2.39 | 0.00 | 0.89 | 0.11 | 1.51 | 0.19 | 1.86 | 0.69 |
| 153 | 1.23 | -0.40 | 0.97 | 1.34 | -0.07 | 0.00 | -0.77 | -0.45 | 1.48 | 1.80 | -0.67 | 0.87 | -0.42 | -0.30 | 2.53 | 1.18 | 1.58 | 1.83 | -0.76 | -0.18 |

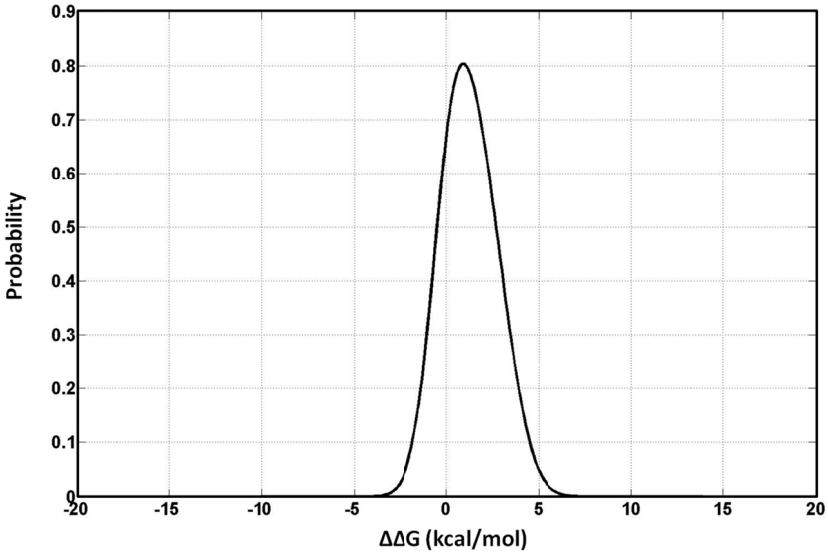



**Figure D2.** Probability distribution function of mutational effects on WT horse Mb, taken from (Tokuriki 2007).

The relation between the instantaneous *dN/dS* for each substitution (Nielsen and Yang 2003) and $\Delta G$ and $\Delta\Delta G$ can be defined as:

$$\frac{dN}{dS} = \omega(s) = N_{eff}\frac{1-\exp(-2s)}{1-\exp(-2s\times N_{eff})} \qquad (1)$$

while:

$$s = \frac{F_{\text{after}}-F_{\text{before}}}{F_{\text{before}}} \sim e^{\beta\Delta G_{\text{before}}}\left(1 - e^{\beta\Delta\Delta G_{\text{mutation}}}\right) \qquad (2)$$

One can then evaluate *dN/dS* at different stabilities (i.e., $\Delta G$) and with different mutational effects ($\Delta\Delta G$).

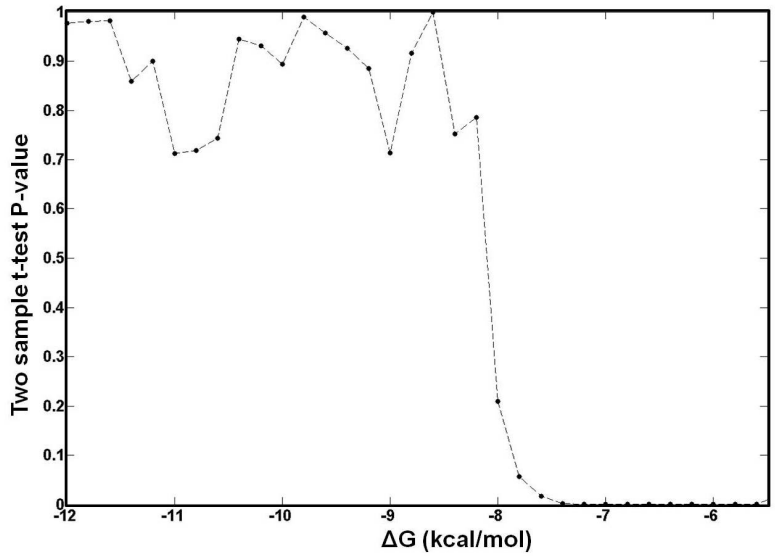



**Figure D2**. P-value of two sample t-test for the null hypothesis of $\omega_{pop}$ and $\omega_{ML}$ being independent random samples from normal distributions with equal means and equal but unknown variances at different $\Delta G$ values. All the analyses are done among branches of 12 simulated bifurcating phylogenetic trees each having 1024 external nodes. In maximum likelihood inference, the transition-transversion rate ratio is set to 1 as there is no preference assigned to transition rates in the simulation protocol and the equilibrium codon frequencies were estimated from the products of the average observed nucleotide frequencies in the three codon positions (F3X4 model).

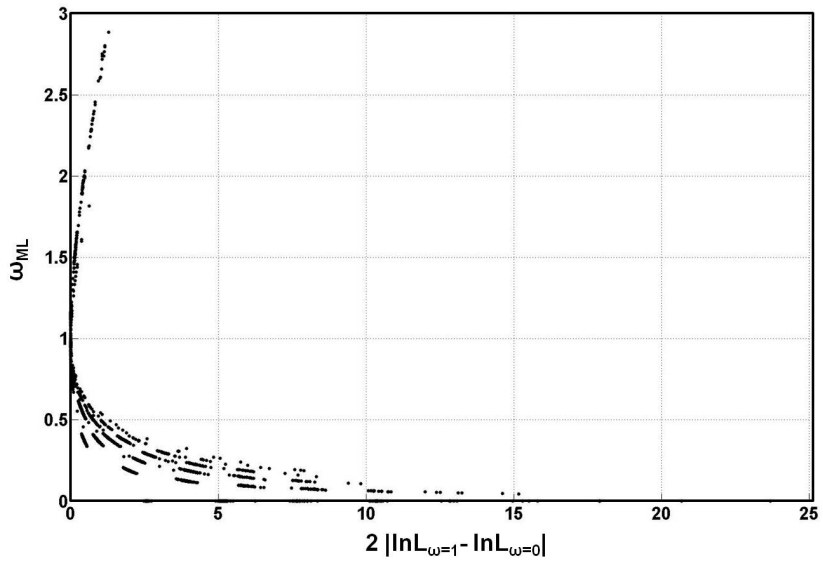



**Figure D3.** ER estimated by ML method, $\omega_{ML}$, versus twice the difference in log likelihoods of the models when *dN/dS* is set to 1 versus the case it is left to vary. The value of $2|\ln L_{\omega=1} - \ln L_{\omega=0}| > 6.64$ corresponds to the P-values smaller than 0.001 for rejecting the null hypothesis of ER being 1.

**1. CODEML output for positively selected residues in Mb sequences from 10 different phylogenetic trees all starting from the same ancestral sequence with $\lambda=10^5$ mutational attempts.**

**Tree 1:**

Positively selected sites (*: P>95%; **: P>99%)

(amino acids refer to 1st sequence: mb1)

| | Pr(w>1) | post mean +- SE for w |
|---|---|---|
| 48 H | 0.687 | 1.326 +- 0.257 |
| 59 K | 1.000** | 1.500 +- 0.004 |
| 119 R | 0.967* | 1.481 +- 0.100 |
| 133 R | 1.000** | 1.500 +- 0.003 |
| 139 S | 1.000** | 1.500 +- 0.000 |

**Tree 2:**

Positively selected sites (*: P>95%; **: P>99%)

(amino acids refer to 1st sequence: mb1)

| | Pr(w>1) | post mean +- SE for w |
|---|---|---|
| 106 T | 0.894 | 1.441 +- 0.171 |
| 133 R | 1.000** | 1.500 +- 0.002 |
| 139 S | 0.928 | 1.460 +- 0.144 |

**Tree 3:**

Positively selected sites (*: P>95%; **: P>99%)

(amino acids refer to 1st sequence: mb1)

| | Pr(w>1) | post mean +- SE for w |
|---|---|---|
| 106 T | 0.798 | 1.388 +- 0.224 |
| 133 K | 1.000** | 1.500 +- 0.003 |
| 139 S | 0.986* | 1.492 +- 0.065 |

**Tree 4:**

Positively selected sites (*: P>95%; **: P>99%)

(amino acids refer to 1st sequence: mb1)

| | Pr(w>1) | post mean +- SE for w |
|---|---|---|
| 106 T | 0.999** | 1.500 +- 0.017 |

133 R 1.000** 1.501 +- 0.023

139 S 0.996** 1.498 +- 0.033

**Tree 5:**

Positively selected sites (*: P>95%; **: P>99%)

(amino acids refer to 1st sequence: mb1)

Pr(w>1) post mean +- SE for w

133 R 1.000** 1.500 +- 0.002

139 S 0.966* 1.481 +- 0.100

**Tree 6:**

Positively selected sites (*: P>95%; **: P>99%)

(amino acids refer to 1st sequence: mb1)

Pr(w>1) post mean +- SE for w

106 T 0.995** 1.497 +- 0.038

133 G 1.000** 1.500 +- 0.001

139 S 0.910 1.450 +- 0.160

**Tree 7:**

Positively selected sites (*: P>95%; **: P>99%)

(amino acids refer to 1st sequence: mb1)

Pr(w>1) post mean +- SE for w

133 R 1.000** 1.500 +- 0.002

139 I 1.000** 1.500 +- 0.007

**Tree 8:**

Positively selected sites (*: P>95%; **: P>99%)

(amino acids refer to 1st sequence: mb1)

Pr(w>1) post mean +- SE for w

59 K 0.693 1.329 +- 0.257

106 T 0.569 1.258 +- 0.279

133 K 0.997** 1.499 +- 0.028

139 I 0.999** 1.500 +- 0.015

**Tree 9:**

Positively selected sites (*: P>95%; **: P>99%)

(amino acids refer to 1st sequence: mb1)

| Pr(w>1) | post mean +- SE for w |
|---|---|
| 106 T | 0.665 1.313 +- 0.263 |
| 133 K | 0.998** 1.499 +- 0.021 |
| 139 S | 0.993** 1.496 +- 0.047 |

**Tree 10:**

Positively selected sites (*: P>95%; **: P>99%)

(amino acids refer to 1st sequence: mb1)

| Pr(w>1) | post mean +- SE for w |
|---|---|
| 106 T | 0.665 1.313 +- 0.263 |
| 133 K | 0.998** 1.499 +- 0.021 |
| 139 S | 0.993** 1.496 +- 0.047 |

**2. Alignment for sperm whale Mb (SW) and ancestral Mb in explicit-sequence simulations (Ancestor) with 32% identity with SW Mb.**

```
CLUSTAL O(1.1.0) multiple sequence alignment

Ancestor DLWEIEFDAVLSVWSTLEKDLAGHGRLVLVFLFMSAASAQAQFDSYSHLARTAAAKSSKA 60
SW       VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED 60
          * * *:: ** **:.:* *:****: :*: ** *  .:  :** :.**   *  *:*:

Ancestor LQKHGSIVLRSLGRVLRAVTSADDQPKGIAQSHSPVQLPQIIYVETLSRSLKQTSVTSRP 120
SW       LKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP 120
         *:***  ** :** :*:     : : * :****:  :   * *:* :*.:: :.  : :*

Ancestor DKISAEFLEAVNRALTRLSSTVAEALETQIRLP- 153
SW       GDFGADAQGAMNKALELFRKDIAAKYKE-LGYQG 153
          .:.*:   *:*:**  : . :*   :  :
```

# BIBLIOGRAPHY

Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* 18: 1585–1592.

Anisimova M, Yang Z (2007) Multiple hypotheses testing to detect adaptive protein evolution affecting individual branches and sites. *Mol Biol Evol* 24:1219–1228.

Barrick D, Hughson FM, Baldwin RL (1994) Molecular mechanisms of acid denaturation. The role of histidine residues in the partial unfolding of apomyoglobin. *J Mol Biol* 237: 588–601.

Bangsbo J (2000) Muscle oxygen uptake in humans at onset of and during intense exercise. *Acta Physiol Scand* 168: 457–464.

Bayes T (1763) An essay toward solving a problem in the doctrine of chances. *Philos Trans R Soc Lond* 53: 370–418.

Benner SA, Caraco MD, Thomson JM, and Gaucher EA (2002) Planetary biology: paleontological, geological, and molecular histories of life. *Science* 296: 864–868.

Beard DA (2006) Modeling of oxygen transport and cellular energetics explains observations on in vivo cardiac energy metabolism. *PLoS Comput Biol* 2: e107.

Bedford T, Hartl DL (2008) Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in *DROSOPHILA*. *Mol Biol Evol* 25: 1631–1638.

Berenbrink M, Mirceta S (2009) How to make a whale: Molecular signature of myoglobin in diving birds and mammals. *Com Biochem Physiol Part A* 153: S44.

Bininda–Emonds ORP, Gittleman JL, Purvis A (1999) Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biol Rev* 74: 143–175.

Biswas S, Akey JM (2006) Genomic insights into positive selection. *Trends Genet* 22: 437–446.

Blanga–Kanfi S, Miranda H, Penn O, Pupko T, DeBry RW, Huchon D (2009) Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. *BMC Evol Biol* 9: 71.

Blix AS, Kjekshus JK, Enge I, Bergan A (1976) Myocardial blood flow in the diving seal. *Acta Physiol Scand* 96: 277–280.

Bollback JP (2006) SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* 7: 88.

Bradshaw RA, Gurd FRN (1969) Comparison of myoglobins from harbor seal, porpoise, and sperm whale. *V. J Biol Chem* 244: 2167–2181.

Bogardt RA, Jones BN, Dwulet FE, Garner WH, Lehman LD, et al., (1980) Evolution of the amino acid substitution in the mammalian myoglobin gene. *J Mol Evol* 15: 197–218.

Bucci E (2009) Thermodynamic approach to oxygen delivery in vivo by natural and artificial oxygen carriers. *Biophys Chem* 142:1–6.

Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10: 195–205.

Chen Y, Dokholyan NV (2008) Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. *Mol Biol Evol* 25: 1530–1533.

Chen P, Shakhnovich EI (2009) Lethal mutagenesis in viruses and bacteria. *Genetics* 183: 639–650.

Cherry JL (2010) Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol Biol Evol* 27:735–741.

Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem 75*: 333–366.

Christensen NJ, Kepp KP (2012) Accurate Stabilities of Laccase Mutants Predicted with a Modified FoldX Protocol. *J Chem Inf Model* 52: 3028–3042.

Croll DA, Acevedo–Gutiérrez A, Tershy BR, Urbán–Ramírez J (2001) The diving behavior of blue and fin whales: is dive duration shorter than expected based on oxygen stores? *Comp Biochem Physiol A* 129: 797–809.

Clark CA, Burns JM, Schreer JF, Hammill MO (2007) A longitudinal and cross–sectional analysis of total body oxygen store development in nursing harbor seals (Phoca vitulina). *J Comp Physiol B* 177: 217–227.

Cutler DJ (2000) Understanding the over-dispersed molecular clock. *Genetics* 154: 1403–1417.

Dasmeh P, Kepp KP (2012) Bridging the gap between chemistry, physiology, and evolution: Quantifying the functionality of sperm whale myoglobin mutants. *Compar Biochem Physiol Part A* 161: 9–17.

Dasmeh P, Kepp KP, Davis RW (2013a) Aerobic dive limits of seals with mutant myoglobin using combined thermochemical and physiologivcal data. *Compar Biochem Physiol Part A* 164: 119–128.

Dasmeh P, Serohijos AWR, Kepp KP, Shakhnovich EI (2013b) Positively selected sites in cetacean myoglobins contribute to protein stability. *PLoS Comput Biol* 9(3): e1002929.

Dasmeh P, Serohijos AWR, Kepp KP and Shakhnovich EI (2013c) Influence of protein biophysics on inferring molecular clock rates in phylogenetic trees, Submitted to *Molecular Biology and Evolution*.

Davis RW, Kanatous SB (1999) Convective oxygen transport and tissue oxygen consumption in Weddell seals during aerobic dive. *J Exp Biol* 202: 1091–1113.

Davis RW, Fuiman LA, Williams TM, Horning M, Hagey W (2003) Classification of Weddell seal dives based on three-dimensional movements and video recorded observations. *Marine Ecology Progress Series* 264: 109–122.

Davis RW, Polasek L,Watson R, Fuson A, Williams TM, Kanatous SB (2004) The diving paradox: New insights into the role of the dive response in air-breathing vertebrates. *Compar Biochem Physiol Part A* 138: 263–268.

Davis RW, Kanatous SB (1999) Convective oxygen transport and tissue oxygen consumption in Weddell seals during aerobic dives. *J Exp Biol* 202: 1091–1113.

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Natl Biomedical Research pp. 345–352.

Dokholyan NV, Shakhnovich EI (2001) Understanding hierarchical protein evolution from first principles. *J Mol Biol* 312: 289–307.

Dolar ML, Suarez P, Ponganis PJ, Kooyman GL (1999) Myoglobin in pelagic small cetaceans. *J Exp Biol* 202: 227–236.

Drummond DA, Wilke CO (2008) Mistranslation–induced protein misfolding as a dominant constraint on coding–sequence evolution. *Cell* 134: 341–352.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A* 102: 14338–14343.

Duteil S, Bourrilhon C, Raynaud JS, Wary C, Richardson RS, Leroy-Willig A, Jouanin JC, Guezennec CY, Carlier PG (2004) Metabolic and vascular support for the role of myoglobin in humans: a multiparametric NMR study. *Am J Physiol Regul Integr Comp Physiol* 287: R1441–1449.

Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6: 197–208.

Elber R (2010) Ligand diffusion in globins: simulations versus experiment. *Curr Opin Struct Biol* 20: 162–167.

Elsner RW, Franklin DL, VanCritters RL (1964) Cardiac output during diving in an unrestrained sea lion. *Nature* 202: 809–810.

Endeward V, Gros G, Jürgens KD (2010) Significance of myoglobin as an oxygen store and oxygen transporter in the intermittently perfused human heart: a model study. *Cardiovasc Res* 87,

Falke KJ, Hill RD, Qvist J, Schneider RC, Guppy M, Liggins GC, Hochachka PW, Elliot RE, Zapol WM (1985) Seal lungs collapse during free diving: Evidence from arterial nitrogen tensions. *Science* 229: 556–558.

Fersht AR, Matouschek A, Serrano L (1992) The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol* 224: 771–782.

Feynman, R. P. (1986). Personal Observation on Reliability of Shuttle. In Rogers, W. P., et. al. Report of the Presidential Commission on the Space Shuttle Challenger Accident. Volume II, Appendix F. Washington, D.C.: Government Printing Agency, 1986.

Felsenstein J, Churchill GA (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol Biol Evol* 13: 93–104.

Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford.

Frauenfelder H, McMahon BH, Fenimore PW (2003) Myoglobin: The hydrogen atom of biology and a paradigm of complexity. *Proc Natl Acad Sci USA* 100: 8615–8617.

Fuiman LA, Kiersten MM, Williams TM, Davis RW (2007) Structure of foraging dives in the Antarctic fast-ice environment. *Deep-Sea Res II* 54: 270–289.

Garry DJ, Ordway GA, Lorenz JN, Radford NB, Chin ER, Grange RW, Bassel-Duby R, Williams RS (1998) Mice without myoglobin. *Nature* 395: 905–908.

Gaucher EA (2007) in Ancestral Sequence Reconstruction (ed. Liberles DA). Oxford: Oxford University Press.

Gaucher EA, Thomson JM, Burgan MF, and Benner SA (2003) Inferring the paleoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 425: 285–288.

Gillespie JH (1984) The molecular clock may be an episodic clock. *Proc Natl Acad Sci USA* 81: 8009–8013.

Gillespie JH (1986) Natural selection and the molecular clock. *Mol Biol Evol* 3: 138–155.

Gimenez M, Sanderson RJ, Reiss OK, Banchero N (1977) Effects of altitude on myoglobin and mitochondrial protein in canine skeletal muscle. *Respiration* 34: 171–176.

Goldstein RA (2008) The structure of protein evolution and the evolution of protein structure. *Curr Opin Struct Biol* 18: 170–177.

Goldstein RA (2011) The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins* 79: 1396–1407.

Groebe K (1995) An easy-to-use model for O2 supply to red muscle. Validity of assumptions, sensitivity to errors in data. *Biophys J* 68: 1246–1269.

Gros G, Wittenberg BA, Jue T (2010) Myoglobin's old and new clothes: from molecular structure to function in living cell. *J Exp Biol* 213: 2713–2715.

Guyton GP, Stanek KS, Schneider RC, Hochachka PW, Hurford WE, Zapol DG, Liggins GC, Zapol WM (1995) Myoglobin saturation in free-diving Weddell seals. *J Appl Physiol* 79: 1148–1155.

Harris EE, Hey J (1999) X chromosome evidence for ancient human histories. *Proc Natl Acad Sci USA* 96: 3320–24.

Hassanin A, Delsuc F, Ropiquet A, Hammer C, Jansen van Vuuren B, Matthee C, Ruiz–Garcia M, Catzeflis F, Areskoug V, Nguyen TT (2012) Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. *C R Biol* 335: 32–50.

Heo M, Maslov S, Shakhnovich EI (2011) Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proc Natl Acad Sci USA* 108: 4258–4263.

Helbo S, Fago A (2012) Functional properties of myoglobins from five whale species with different diving capacities. *J Exp Biol 215*: 3403–3410.

Hendgen-Cotta UB, Merx MW, Shiva S, Schmitz J, Becher S, Klare JP, Steinhoff HJ, Goedecke A, Schrader J, Gladwin MT, Kelm M, Rassaf T (2008) Nitrite reductase activity of myoglobin regulates respiration and cellular viability in myocardial ischemia-reperfusion injury. *Proc Natl Acad Sci USA* 105: 10256–10261.

Ho BK, Dill KA (2006) Folding very short peptides using molecular dynamics. *PLoS Comput Biol* 2: e27.

Holder M, Lewis PO (2003) Phylogeny estimation: Traditional and Bayesian approaches. *Nat Rev Genet* 4: 275–284.

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K et al. (2009) Ensembl 2009, *Nucleic Acids Res* 37:D690–D697.

Hughson FM, Barrick D, Baldwin RL (1991) Probing the stability of a partly folded apomyoglobin intermediate by site-directed mutagenesis. *Biochemistry* 30: 4113–4118.

Hughson FM, Baldwin RL (1989) Using of site–directed mutagenesis to destabilize native apomyoglobin relative to folding intermediates. *Biochemistry* 28: 4415–4422.

Hurford WE, Hochachka PW, Schneider RC, Guyton GP, Stanek KS, Zapol DG, Liggins GC, Zapol WM (1996) Splenic contraction, catecholamine release and blood volume redistribution during diving in the Weddell seal. *J Appl Physiol* 80: 298–306.

Jensen KP, Ryde U (2004) How heme binds $O_2$: Reasons for reversible binding and spin inversion. *J Biol Chem* 279: 14561–14569.

Jensen KP, Ryde U (2003) Comparison of the Chemical Properties of Iron and Cobalt Porphyrins and Corrins. *ChemBioChem* 4: 413–424.

Kanatous SB, Hawke TJ, Trumble SJ, Pearson LE, Watson RR, Garry DJ, Williams TM, Davis R. W (2008) The ontogeny of aerobic and diving capacity in the skeletal muscles of Weddell seals. *J Exp Biol* 211: 2559–2565.

Kanatous SB, Mammen PPA (2010) Regulation of myoglobin expression. *J Exp Biol* 213: 2741–2747.

Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A et al. (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* 33:D29–D33.

Kendrew JC (1964) 'Myoglobin and the Structure of Proteins. Nobel Lecture, December 11, 1962', in Nobel Lectures, Chemistry 1942–1962 (Amsterdam: Elsevier Publishing Company): 676–98.

Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* 181: 662–666.

Kiel C, Aydin D, Serrano L (2008) Association rate constants of ras–effector interactions are evolutionarily conserved. *PLoS Computational Biology* 4: e1000245.

Kimura M (1983) The Neutral Theory of Molecular Evolution (Cambridge Univ. Press, Cambridge, 1983).

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 267: 275–276.

Kondrashov DA, Zhang W, Aranda IV R, Stec B, Phillips GN (2008) Sampling of the native conformational ensemble of myoglobin via structures in different crystalline environments. *Proteins Struct Funct Bioinf* 70: 353–362.

Kooyman GL, Ponganis PJ (1998) The physiological basis of diving to depth: Birds and mammals. *Annu Rev Physiol* 60: 19–32.

Kooyman GL, Wahrenbrock EA, Castellini MA, Davis RW, Sinnett EE (1980) Aerobic and anaerobic metabolism during voluntary diving in Weddell seals: Evidence of preferred pathways from blood chemistry and behavior. *J Comp Physiol B* 138: 335–346.

Krogh A (1919) The number and the distribution of capillaries in muscle with the calculation of the oxygen pressure necessary for supplying the tissue. *J Physiol* (*Lond)* 52: 409–515.

Kumar S (2005) Molecular clocks: four decades of evolution. *Nat Rev Genet* 6: 654–662.

Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res* 34: D204–206.

Lawrie RA (1953) The activity of the cytochrome system in muscle and its relation to myoglobin. *Biochem J* 55: 298–305.

Lin PC, Kreutzer U, Jue T (2007a) Anisotropy and temperature dependence of myoglobin translational diffusion in myocardium: implication for oxygen transport and cellular architecture. *Biophys J* 92: 2608–2620.

Lin PC, Kreutzer U, Jue T (2007b) Myoglobin translational diffusion in rat myocardium and its implication on intracellular oxygen transport. *J Physiol* 578: 595–603.

Lio` P, Goldman N (1998) Models of molecular evolution and phylogeny. *Genome Res* 8:1223–1244.

Lobkovsky AE, Wolf YI, Koonin EV (2010) Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc Natl Acad Sci USA* 107: 2983–2988.

Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302: 1401–1404.

Mahler M, Louy C, Homsher E, Peskoff A (1985). Reappraisal of Diffusion, Solubility, and Consumption of Oxygen in Frog Skeletal Muscle, with Applications to Muscle Energy Balance. *J Gen Physiol* 86: 105–134.

Mailund T, Dutheil JY, Hobolth A, Lunter G, Schierup MH (2011) Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS Genet* 7: e1001319.

Margoliash E (1963) Primary structure and evolution of cytochrome c. *Proc Natl Acad Sci USA* 50: 672–679.

Masuda K, Truscott K, Lin PC, Kreutzer U, Chung Y, Sriram R, Jue T (2008) Determination of myoglobin concentration in blood-perfused tissue. *Eur J Appl Physiol* 104: 41–48.

McGowen MR, Spaulding M, Gatesy J (2009) Divergence date estimation and a comprehensive molecular tree of extant cetaceans. *Mol Phylogenet Evol* 53: 891–906.

McGuire BJ, Secomb TW (2001) A theoretical model for oxygen transport in skeletal muscle under conditions of high oxygen demand. *J Appl Physiol* 91: 2255–2265.

Mendez J, Keys A (1960) Density and composition of mammalian muscle. *Metabolism* 9: 184–188.

Mesnick SL, Taylor BL, Duc RGL, Trevino SE, O'Corry-Crowe GM, Dizon AE (1999) Culture and genetic evolution in whales. *Science* 284:2055a.

Mirceta S, Campbell KL, Berenbrink M (2009) Molecular evolution of myoglobin in small diving mammals. *Com Biochem Physiol Part A* 153: S98–S99.

Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol* 291: 177–196.

Mustonen V, Lassig M (2009) From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet* 25: 111–119.

Naylor GJP, Gerstein M (2000) Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins. *J Mol Evol* 51: 223–233.

Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics. New York: Oxford University Press.

Nelson DP, Samsel RW, Wood LD, Schumacker PT (1988) Pathological supply dependence of systemic and intestinal $O_2$ uptake during endotoxemia. *J Appl Physiol* 64: 2410–2419.

Nemeth PM, Lowry OH (1984) Myoglobin levels in individual human skeleton-muscle fibers of different types. *J Hist Cytochem* 132: 1211–1216.

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148: 929–936.

Nielsen R, Yang Z (2003) Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol* 20: 1231–1239.

Nishimura C, Wright PE, Dyson HJ (2003) Role of the B helix in early folding events in apomyoglobin: evidence from site-directed mutagenesis for native-like long range interactions. *J Mol Biol* 334: 293–307.

Noren SR, Williams EE (2000) Body size and skeletal muscle myoglobin of cetaceans: adaptations for maximum dive duration. *Comp Biochem Physiol A* 126: 181–191.

Ohta T (1992) The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst* 23:263–286.

Ohta T, Kimura M (1971) On the constancy of the evolutionary rate of cistrons. *J Mol Evol* 1: 18–25.

Olson JS (2008) Protein Reviews: Dioxygen binding and sensing proteins. Section 14: "From $O_2$ binding diffusion into red blood cells to ligand pathways in globins", Springer, Milan, Italy.

O'Neil KT, DeGrado WF (1990) A thermodynamic scale for the helix–forming tendencies of the commonly occurring amino acids. *Science* 250: 646–651.

Ostermann A, Waschipky R, Parak FG, Nienhaus GU (2000) Ligand binding and conformational motions in myoglobin. *Nature* 404: 205–208.

Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158: 927–931.

Pal C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. *Nat Rev Genet* 7: 337–348.

Papadopoulos S, Endeward V, Revesz-Walker B, Jürgens KD, Gros G (2001) Radial and longitudinal diffusion of myoglobin in single living heart and skeletal muscle cells *Proc Natl Acad Sci USA* 98: 5904–5909.

Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, et al., (2011) A molecular phylogeny of living primates. *PLoS Genetics* 7: e1001342.

Perutz MF, Mattews FS (1966) An X-ray study of azide methaemoglobin. *J Mol Biol* 21: 199–207.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin, TE (2004) UCSF Chimera-a visualization system for exploratory research and analysis. *J Comput Chem* 25: 1605–1612.

Phillips SEV (1980) Structure and refinement of oxymyoglobin at 1.6 Å resolutions. *J Mol Biol* 142: 531–554.

Polasek LK, Dickson KA, Davis RW (2006) Metabolic indicators in the skeletal muscles of harbor seals (Phoca vitulina). *Am J Physiol Regul Integr Comp Physiol* 290: R1720–R1727.

Ponganis PJ (2011) Diving mammals. In Comprehensive Physiology (ed. R. Terjung), 447–465. Hoboken, NJ: John Wiley & Sons, Inc.

Pollock DD, Thiltgen G, Goldstein RA (2012) Amino acid coevolution induces an evolutionary stokes shift. *Proc Natl Acad Sci USA* 109: E1352–E1359.

Ponganis PJ, Kooyman GL, Castellini MA (1993) Determinants of the aerobic dive limit of Weddell seals: analysis of diving metabolic rates, post dive end tidal $P_{O2}$'s and blood and muscle oxygen stores. *Physiol Zool* 66: 732–749.

Ponganis PJ, Kreutzer U, Sailasuta N, Knower T, Hurd R, Jue T (2002) Detection of myoglobin desaturation in Mirounga angustirostris during apnea. *Am J Physiol Regul Integr Comp Physiol* 282: R267–R272.

Ponganis PJ, Kreutzer U, Stockard TK, Lin PC, Sailasuta N, Tran TK, Hurd R, Jue T (2008) Blood flow and metabolic regulation in seal muscle during apnea. *J Exp Biol* 211: 3323–3332.

Prasad AB, Allard MW, Green ED (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol Biol Evol* 25: 1795–1808.

Price SA, Bininda–Emonds ORP, Gittleman JL (2005) A complete phylogeny of the whales, dolphins and even–toed hoofed mammals (Cetartiodactyla). *Biol Rev Comb Philos Soc* 80: 445–473.

Privalov PL, Khechinashvili NN (1974) A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *J Mol Biol* 86: 665–684.

Puett D, Friebele E, Hammonds RG., Jr (1973) A comparison of the conformational stabilities of homologous hemoproteins. Myoglobin from several species, human hemoglobin and subunits. *Biochim Biophys Acta* 328: 261–277.

Qvist J, Hill RD, Schneider RC, Falke KJ, Liggins GC, Guppy M, Elliot RL, Hochachka PW, Zapol WM (1986) Hemoglobin concentrations and blood gas tensions of free diving Weddell seals. *J Appl Physiol* 61: 1560–1569.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164: 1645–1656.

Reed JZ, Chambers C, Fedak MA, Butler PJ (1994) Gas exchange of freely diving grey seals (Halichoerus grypus). *J Exp Biol* 191: 1–18.

Regis WCB, Fattori J, Santoro MM, Jamin M, Ramos CHI (2005) On the difference in stability between horse and sperm whale myoglobins. *Arch Biochem Biophys* 436: 168–177.

Reynafarje B (1963) Simplified method for the determination of myoglobin. *J Lab Clin Med* 6: 138–145.

Romero-Herrera AE, Lehmann H, Joysey KA, Fridays AE (1973) Molecular Evolution of Myoglobin and the Fossil Record: a Phylogenetic Synthesis. *Nature* 246: 389–395.

Rossi-Fanelli A, Antonini E (1958) Studies on the oxygen and carbon monoxide equilibria of human myoglobin. *Arch Biochem Biophys* 77: 478–492.

Rowell LB (1986) Human circulation: regulation during physical Stress. Oxford, New York: Oxford University Press. pp 415.

Samsel RW, Schumacker PT (1994) Systemic hemorrhage augments local $O_2$ extraction in canine intestine. *J Appl Physiol* 77: 2291–2298.

Sánchez IE, Beltrao P, Stricher F, Schymkowitz J, Ferkinghoff–Borg J, et al., (2008) Genome–wide prediction of SH2 domain targets using structural information and the FoldX algorithm. *PLoS Comput Biol* 4: e1000052.

Sawyer SL, Malik HS (2006) Positive selection of yeast nonhomologous endjoining genes and a retrotransposon conflict hypothesis. *Proc Natl Acad Sci USA* 103: 17614–17619.

Sawyer SA, Parsch J, Zhang Z, Hartl DL (2007) Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc Natl Acad Sci USA* 104: 6504–6510.

Schenkman KA, Marble DR, Burns DH, Feigl EO (1997) Myoglobin oxygen dissociation by multiwavelength spectroscopy. *J Appl Physiol* 82: 86–92.

Scholander PF (1940) Experimental investigation onthe respiratory function in diving mammals and birds. *Hvalrad Skr* 22: 1–131.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. *Nucl Acids Res* 33: W382–W388.

Scott EE (1998) Apoglobin Stability and Ligand Movements in Mammalian Myoglobins. Ph.D. dissertation, Rice University, Houston, TX.

Scott EE, Gibson QH, Olson JS (2001) Mapping the pathways for O2 entry into and exit from myoglobin. *J Biol Chem* 276: 5177–5188.

Scott EE, Paster EV, Olson JS (2000) the stabilities of mammalian apomyoglobin vary over a 600–fold range and can be enhanced by comparative mutagenesis. *J Biol Chem* 275: 27129–27136.

Serohijos AWR, Hegedus T, Aleksandrov AA, He L, Cui L, Dokholyan NV, Riordan JR (2008) Phenylalanine-508 mediates a cytoplasmic-membrane domain contact in the CFTR 3D structure crucial to assembly and channel function. *Proc Natl Acad Sci USA* 105: 3256–3261.

Serohijos AWR, Rimas Z, Shakhnovich EI (2012) Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly. *Cell report* 2: 249–256.

Serohijos AWR, Lee SY and Shakhnovich EI (2013) Highly Abundant Proteins Favor More Stable 3D Structures in Yeast. *Biophys J* 104: L1–L3.

Shakhnovich EI, Finkelstein AV (1989) Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers* 28: 1667–1680.

Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. *Nat Rev Genet* 11: 572–582.

Springer BA, Sligar SG (1987) High-level expression of sperm whale myoglobin in Escherichia coli. *Proc Natl Acad Sci USA* 84: 8961–8965.

Stern JS (1992) Surfacing rates and surfacing patterns of minke whales (Balaenoptera acutorostrata) off central California, and the probability of a whale surfacing within visual range. *Reports of the International Whaling Commission* 42: 379–385.

Stewart JM, Blakely JA, Karpowicz PA, Kalanxhi E, Thatcher BJ, Martine BM (2004) Unusually weak oxygen binding, physical properties, partial sequence, autoxidation rate and potential phosphorylation sites of beluga what (Delphinapterus leucas) myoglobin. *Comp Biochem Physiol* B 137: 401– 412.

Suzuki T, Imai K (1998) Evolution of myoglobin. *CMLS Cell Mol Life Sci*. 54: 979–1004.

Swanson WJ, Wong A, Wolfner MF, Aquadro CF (2004) Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. *Genetics* 168: 1457–1465.

Takahata N (1987) On the overdispersed molecular clock. *Genetics* 116: 169–179.

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees *Mol Biol Evol* 10: 512–526.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods *Mol Biol Evol* 28: 2731–2739.

Tamuri AU, Dos Reis M, Goldstein RA (2012) Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics* 190: 1101–1115.

Taverna DM, Goldstein RA (2002a) Why are proteins marginally stable? *Proteins* 46: 105–109.

Taverna DM, Goldstein RA (2002b) Why are proteins so robust to site mutations? *J Mol Biol* 315: 479–484.

Tawara T (1950) On the respiratory pigments of whale (Studies on whale blood II). *Sci Rep Whales Res Inst* 3: 96–101.

Terrados N, Jansson E, Sylven C, Kaijser L (1990) Is hypoxia a stimulus for synthesis of oxidative enzymes and myoglobin? *J Appl Physiol* 68: 2369–2372.

Tilton RF, Kuntz ID, Petsko GA (1984) Cavities in proteins-structure of a metmyoglobin-xenon complex solved to 1.9 Å. *Biochemistry* 23: 2849–2857.

Torrance SM, Wittnich C (1994) Blood lactate and acid–base balance in graded neonatal hypoxia: Evidence for oxygen-restricted metabolism. *J Appl Physiol* 77: 2318–2324.

Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. *J Mol Biol* 369: 1318–1332.

Tokuriki N, Stricher F, Serrano L, Tawfik DS (2008) How protein stability and new functions trade off. *PLoS Comput Biol* 4: e1000002.

Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. *J Mol Biol* 369: 1318–1332.

Tokuriki N, Tawfik DS (2009) Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol* 19: 596–604.

UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* 39: D214–D219.

Varadarajan R, Szabo A, Boxer SG (1985) Cloning, expression in Escherichia coli, and reconstitution of human myoglobi. *Proc Natl Acad Sci USA* 82: 5681–5684.

Wajcman H, Kiger L, Marden MC (2009) Structure and function evolution in the superfamily of globins. *C R Biologies* 332: 273–282.

Watwood SL, Miller PJO, Johnson M, Madsen PT, Tyack PL (2006) Deep–diving foraging behaviour of sperm whales (Physeter macrocephalus). *J Anim Ecol* 75: 814–825.

Weber RE, Hemmingsen EA, Johansen K (1974) Functional and biochemical studies of penguin myoglobin. *Comp Biochem Physiol* 49 B: 197–214.

Westgate AJ, Read AJ, Berggren P, Koopman HN, Gaskin DE (1995) Diving behaviour of harbour porpoise, Phocoena phocoena. *Can Fish Aquat Sci* 52: 1064–1073.

Whelan S, Goldman N (1999) Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol Biol Evol* 16: 1292–1299.

Whitehead, H (2002) In Perrin, W., Würsig B., and Thewissen, J. Encycl. Mar. Mam. Academic Press, 1165–1172.

Wittenberg JB (1970) Myoglobin-facilitated oxygen diffusion: role of myoglobin in oxygen entry into muscle. *Physiol Rev* 50: 559–636.

Wittenberg BA, Wittenberg JB (1989) Transport of Oxygen in Muscle. *Rev Physiol* 51: 857–878.

Wittenberg JB, Wittenberg BA (2003) Myoglobin function reassessed. *J Exp Biol* 206: 2011–2020.

Williams TM, Davis RW, Fuiman LA, Francis J, Le Boeuf B, Horning M, Calambokidis J, Croll DA (2000) Sink or Swim: strategies for cost efficient diving by marine mammals. *Science* 288: 133–136.

Williams TM (2001) Intermittent swimming by mammals: a strategy for increasing energetic efficiency during diving. *Am Zool* 41: 166–176.

Williams CL, Meir JU, Ponganis PJ (2011) What triggers the aerobic dive limit? Patterns of muscle oxygen depletion during dives of emperor penguins. *J Exp Biol* 214: 1802–1812.

Wright TJ, Davis RW (2006) The effect of myoglobin concentration on aerobic dive limit in a Weddell seal. *J Exp Biol* 209: 2576–2585.

Wylie CS, Shakhnovich EI (2011) A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc Natl Acad Sci USA* 108: 9916–9921.

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 15: 568–573.

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 15: 496–503.

Yang Z (2006) Computational Molecular Evolution. (Oxford University Press, Oxford).

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24: 1586–1591.

Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon–substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155: 431–449.

Yang Z, Wong WS, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22: 1107–1118.

Yang JR, Zhuang SM, Zhang J. (2010) Impact of translational error–induced and error–free misfolding on the rate of protein evolution. *Mol Syst Biol* 6: 421.

Yang Z (1996) Among–site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* 11: 367–372.

Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141: 1641–1650.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13: 555–556.

Yin S, Ding F, Dokholyan NV (2007a) Eris: an automated estimator of protein stability. *Nat Methods* 4: 466–467.

Yin S, Ding F, Dokholyan NV (2007b) Modeling Backbone Flexibility Improves Protein Stability Estimation. *Structure* 15: 1567–1576.

Zaia J, Annan RS, Biemann K (1992) The correct molecular weight of myoglobin, a common calibrant for mass spectrometry. *Rapid commun mass spectrum* 6: 32–36.

Zeldovich KB, Chen P, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci USA* 104: 16152–16157.

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch–site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 22: 2472–2479.

Zuckerkandl E, Pauling L (1962) in Horizons in Biochemistry. eds Kasha M and Pullman B (Academic Press, New York), pp 189–225.

Zuckerkandl E, Pauling L (1965) in Evolving Genes and Proteins. eds Bryson V and Vogel HJ. (Academic Press, New York), pp 97–166.

# PUBLICATIONS

Comparative Biochemistry and Physiology, Part A 161 (2012) 9–17

Contents lists available at ScienceDirect

Comparative Biochemistry and Physiology, Part A

journal homepage: www.elsevier.com/locate/cbpa

# Bridging the gap between chemistry, physiology, and evolution: Quantifying the functionality of sperm whale myoglobin mutants

Pouria Dasmeh, Kasper P. Kepp *

*Technical University of Denmark, DTU Chemistry, Kemitorvet 206, DK-2800 Kongens Lyngby, Denmark*

ARTICLE INFO

*Article history:*
Received 29 April 2011
Received in revised form 26 July 2011
Accepted 29 July 2011
Available online 8 August 2011

*Keywords:*
Myoglobin
Whale
Oxygen storage
Oxygen transport
Muscle cell
Physiology
Diving

ABSTRACT

This work merges a large set of previously reported thermochemical data for myoglobin (Mb) mutants with a physiological model of $O_2$-transport and -storage. The model allows a quantification of the functional proficiency of myoglobin (Mb) mutants under various physiological conditions, *i.e.* $O_2$-consumption rate resembling workload, $O_2$ partial pressure resembling hypoxic stress, muscle cell size, and Mb concentration, resembling different organism-specific and compensatory variables. We find that $O_2$-storage and -transport are distinct functions that rank mutants and wild type differently depending on $O_2$ partial pressure. Specifically, the wild type is near-optimal for storage at all conditions, but for transport only at severely hypoxic conditions. At normoxic conditions, low-affinity mutants are in fact better $O_2$-transporters because they still have empty sites for $O_2$, giving rise to a larger [$MbO_2$] gradient (more varying saturation curve). The distributions of functionality reveal that many mutants are near-neutral with respect to function, whereas only a few are strongly affected, and the variation in functionality increases dramatically at lower $O_2$ pressure. These results together show that conserved residues in wild type (WT) Mb were fixated under a selection pressure of low $P_{O2}$.

© 2011 Elsevier Inc. All rights reserved.

## 1. Introduction

Due to its central position in protein science, myoglobin (Mb) is often called the 'hydrogen atom of biology' (Frauenfelder et al., 2003). As the first protein whose three-dimensional structure was revealed at atomic resolution (Kendrew et al., 1958), it has been the subject of extensive experimental and theoretical research. Mb is a monomeric protein of approximately 16 kDa which is generally abundant in muscles (both cardiac and skeletal) of vertebrates and in the body wall of invertebrates (Suzuki and Imai, 1998). Mb increases the availability of oxygen in muscle cells by taking up the oxygen released by hemoglobin in blood, thereby providing $O_2$ during muscle contraction, when blood flow through capillaries is restricted. A cornerstone of muscle metabolism is thus the ability of Mb to enhance the supply of $O_2$ to the mitochondria (Wittenberg, 1970).

Direct storage of $O_2$ has been considered the main function of Mb, based on the observation of elevated concentrations of Mb in diving mammals (Guyton et al., 1995; Ponganis et al., 2002) and increased expression of Mb at higher altitudes (Gimenez et al., 1977; Terrados et al., 1990). In contrast, it has been a matter of controversy whether Mb can contribute significantly to active $O_2$ transport within the cell: Despite the smaller diffusion coefficient of Mb compared to free $O_2$, the high concentration of Mb in working heart and muscle cells, more than thirty-fold than that of free $O_2$, confers an advantage in transporting $O_2$ from the sarcolemma to the mitochondria (Wittenberg and Wittenberg, 2003). *In vitro* studies have indeed confirmed that $O_2$ diffuses faster in Mb solution than in Mb-free solution (Wittenberg and Wittenberg, 1989), and Mb exhibits sufficient mobility and $O_2$-carrying ability to compete effectively with free $O_2$. *In vivo*, however, the role of Mb in $O_2$ diffusion has remained obscure for decades (Gros et al., 2010).

From a theoretical perspective, models have elaborated on the classical work by Krogh (1919) to clarify the problem (Groebe, 1995). Since these models were based on variable quantities such as concentrations and diffusion coefficients in muscle tissues, acceptance of Mb-facilitated diffusion has languished in absence of definitive experimental confirmation (Lin et al., 2007a).

Recently, pulsed-field gradient NMR and Flourescence Recovery After Photobleaching (FRAP) methods enabled precise measurement of the endogenous diffusion coefficient of Mb in myocardial and skeletal muscle cells (Papadopoulos et al., 2001; Lin et al., 2007a,b). Following these results, it was then argued that Mb probably has no significant contribution to $O_2$ transport under normoxic conditions, while its importance increases as the oxygen pressure declines and the cell experiences hypoxia.

Mb binds $O_2$ with a 1:1 stoichiometry at its heme group. This side of the heme is referred to as distal, the other as proximal. More than 40 years ago, it was proposed that $O_2$-binding involves the side-chain of the distal histidine 64 at the E7 helical position (Perutz and Mathews, 1966). Beside this 'histidine-gate' hypothesis, it was also shown that ligands may escape through the interior of the protein

* Corresponding author. Tel.: +45 45 25 24 09.
*E-mail address:* kpj@kemi.dtu.dk (K.P. Kepp).

1095-6433/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cbpa.2011.07.027

where hydrophobic Xe-binding cavities are observed (Tilton et al., 1984). Ligand binding to Mb then follows a multi-state scheme where internal states and distal cavities play a major role (Fig. 1-a) (Ostermann et al., 2000). Here, B states represent a number of ligand positions in the distal 'pocket' with differing rates of binding to the heme group. In the C state, the ligand is relocating between multiple states such as Xe1 and Xe2 cavities within the protein matrix (Scott et al., 2001). The current view is that the majority of ligands (~70–80%) enter and exit from the distal histidine gate (Scott et al., 2001; Elber, 2010) and thus, the kinetics is adequately interpreted in terms of an effective two-step scheme (Fig. 1-b).

Much insight into the pathways and kinetics of ligand binding to Mb has been obtained from studies of its site-directed mutants (Varadarajan et al., 1985; Springer and Sligar, 1987). By measuring the $O_2$ binding parameters of 90 sperm whale Mb mutants at 27 different positions, a mapping of the pathways for $O_2$ entry and exit was achieved (Scott et al., 2001). These investigations show that His E7 stabilizes bound $O_2$ about 1000-fold, most likely because of the formation of a strong hydrogen bond between O and $N_\varepsilon H$ of the imidazole group (Olson, 2008).

In general, the molecular evolution of the globin superfamily can be traced back ~4000 Myr to a common ancestor, which was among the basic protein components required for life (Suzuki and Imai, 1998; Wajcman et al., 2009). Mammalian Mb is thought to have diverged from another common ancestor about 600 Myr ago at the beginning of Cambrian period, and sequence alignments reveal a highly conserved distal pocket within all of these (Romero-Herrera et al., 1973).

The molecular evolution of the heme group can be partly rationalized through the positive selection of the porphyrin structure to satisfy the reversible spin-crossover upon $O_2$ binding to porphyrin, which is a necessary first condition for reversible binding of molecular $O_2$ to Mb (Jensen and Ryde, 2004). In this work, we investigate how the physiological function is further refined by the properties of the surrounding protein, *i.e.* the sequence-dependent thermochemistry of $O_2$-binding, by putting these data into a physiological framework and changing the physiological variables of $O_2$ partial pressure ($P_{O2}$), cell size, Mb concentration, and metabolic rate.

To this end, we quantify transport and storage proficiency as functions of physiological variables in the same way that "reversible binding" and "near-degeneracy" of spin states were used to quantify the functional proficiency of porphyrin and iron, specifically (Jensen and Ryde, 2003, 2004). We then investigate why and how selection pressure conserves the wild-type Mb. It is found that the relative importance of storage and transport varies from one mutant to another and is greatly affected by physiological and environmental conditions, and that the WT appears to have been selected under hypoxic conditions where other mutants are much less proficient.

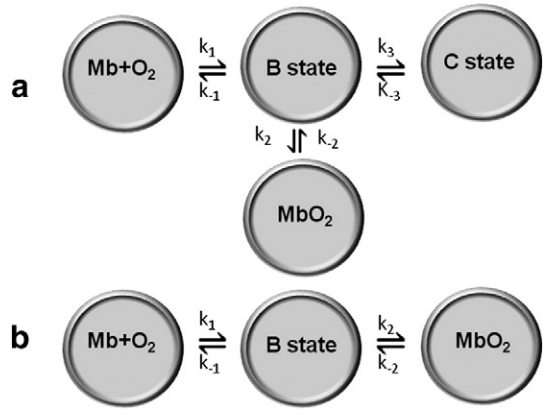



**Fig. 1.** Kinetics of $O_2$ binding to Mb. a) The C state accounts for internal cavities while in b) mainly the B state is involved under physiological conditions.

## 2. Theoretical background

### 2.1. Saturation expression

The saturation of Mb by $O_2$ ($S$), can be expressed in terms of the Hill equation (Hill, 1936):

$$S = \frac{\left(P_{O_2}\right)^n}{(P_{50})^n + \left(P_{O_2}\right)^n} \tag{1}$$

where $P_{O_2}$ is the oxygen partial pressure, $n$ is the oxygen binding cooperativity index, and $P_{50}$ is the value of $P_{O_2}$ at which $S = 0.5$. For Mb with only one subunit, $n = 1$ and saturation can be described using the bimolecular oxygenation equilibrium:

$$Mb + O_2 \rightleftharpoons MbO_2 \tag{2}$$

$$S = \frac{[MbO_2]}{C_{Mb}} = \frac{K_{O_2}[O_2]}{K_{O_2}[O_2] + 1}. \tag{3}$$

Here, saturation is defined as $S = \frac{[MbO_2]}{C_{Mb}}$, $K_{O_2} = \frac{[MbO_2]}{[O_2][Mb]}$ is the bimolecular oxygenation equilibrium constant, and $C_{Mb}$ is the total concentration of Mb in the cell, which is equal to the sum of [Mb] and [$MbO_2$]. Eq. (3) can be converted to (1) using $P_{50} = \frac{1}{K_{O_2}\alpha_{O_2}}$ and $[O_2] = \alpha_{O_2}P_{O_2}$, where $\alpha_{O_2}$ is the oxygen solubility constant in the sarcoplasm. In the general case of two-step binding (Fig. 1-b), using the equilibrium constants $K_1$ and $K_2$, $S$ takes the form:

$$S = \frac{K_1K_2[O_2]}{K_1K_2[O_2] + K_1[O_2] + 1} = \frac{K_{O_2}\alpha_{O_2}P_{O_2}}{\alpha_{O_2}P_{O_2}\left(K_{O_2} + K_1\right) + 1}. \tag{4}$$

The second equality in Eq. (4) holds by using $K_{O_2} = K_1K_2$ and $[O_2] = \alpha_{O_2}P_{O_2}$, and the half saturation pressure can be calculated as $P_{50} = \frac{1}{\alpha_{O_2}(K_{O_2} - K_1)}$. For all mutants studied here $K_{O_2} \gg K_1$ and Eq. (4) reduces to Eq. (3), but this may not always be the case.

### 2.2. Muscle tissue and oxygen delivery

To model oxygen delivery within the muscle cell, we applied the Krogh cylinder model in a revised, state-of-the-art form (Groebe, 1995). As shown in Fig. 2, the model consists of three concentric cylinders representing *capillaries* (inner region), space between red blood cells and sarcolemma, or *cell free regions* (middle), and the *muscle tissue* as the outer region. We consider only radial oxygen diffusion and assume chemical equilibrium at all times, since both the radial and longitudinal diffusion coefficients of Mb are similar within our scope (Groebe, 1995; Lin et al., 2007b). The radial distance ($r$) is used as a generalized coordinate. With this theoretical framework, $P_{O_2}$ is derived as:

$$P_{O_2}(r) = \frac{1}{2}\left(P^*(r) - \frac{D_{Mb}C_{Mb}}{D_{O_2}\alpha_{O_2}} - P_{50}\right) + \frac{1}{2}\sqrt{\left(P^*(r) - \frac{D_{Mb}C_{Mb}}{D_{O_2}\alpha_{O_2}} - P_{50}\right)^2 + 4P^*(r)P_{50}} \tag{5}$$

where $D_{O_2}\,(m^2\,s^{-1})$ and $D_{Mb}\,(m^2\,s^{-1})$ are the free oxygen and Mb diffusion coefficients in the cell, $C_{Mb}\,(mol\ L^{-1})$ and $\alpha_{O_2}\,(mol\ L^{-1}\,mmHg^{-1})$ are the total concentration of Mb and the oxygen solubility in muscle tissue.

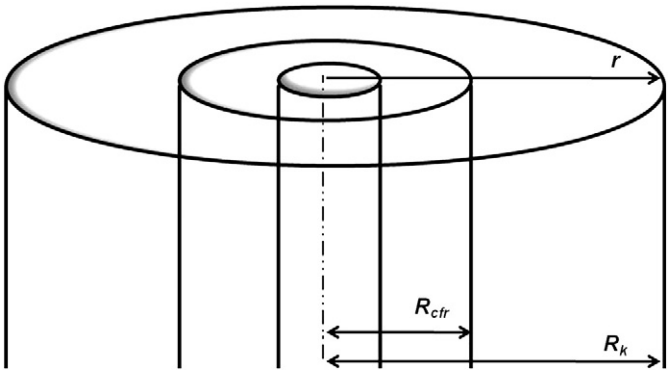



**Fig. 2.** A section of the Krogh model is shown, composed of three concentric cylinders representing capillary (inner), cell free region (middle with radius $R_{cfr}$) and the tissue (outer with radius $R_k$). The radial coordinate ($r$) is the generalized coordinate.

The effective pressure, $P^*(r)$ is then calculated within the radial distance as (Groebe, 1995):

$$P^*(r) = P_u + \frac{D_{Mb}C_{Mb}}{D_{O_2}\alpha_{O_2}}S(P_u) + \frac{\dot{V}_{O_2}}{2D_{Mb}\alpha_{O_2}}\left(\frac{1}{2}\left(r^2 - R_{cfr}^2\right) - R_k^2 ln\frac{r}{R_{cfr}}\right) \quad (6)$$

Here, $P_u = P_{O_2}(R_{cfr})$ is the upper pressure in the cell, $\dot{V}_{O_2}$ $(mol\ L^{-1}s^{-1})$ is the oxygen consumption rate, $S$ is the saturation defined in Eq. (4) and $R_{cfr}$ and $R_k$ are the positions of the carrier free region (*i.e.* capillary walls) and mitochondrial membrane. The upper pressure $P_u$ then declines to a lower pressure, $P_l$, along the radial coordinate ($r$). Importantly, $P_u$ can be affected by many factors such as changes in blood flow rate, capillary density, or hemoglobin oxygen affinity, and thus all "circulatory" effects are assimilated into one variable, $P_u$.

### 2.3. Myoglobin function

#### 2.3.1. Oxygen storage

The proficiency of Mb as an oxygen storage protein can be quantified by integrating the concentration of $MbO_2$ over the radial distance d$r$ (*i.e.* the distance between sarcolemma and mitochondria in Fig. 2) and comparing to the same integral over $[O_2]$. We call this the *oxygen storage ratio* (OSR):

$$OSR \equiv \frac{\int [MbO_2](r)dr}{\int [O_2](r)dr}. \quad (7)$$

Using Eq. (4), the definition of saturation as $S = \frac{[MbO_2]}{C_{Mb}}$ and $[O_2] = \alpha_{O_2}P_{O_2}$, OSR takes the form:

$$OSR \equiv \frac{C_{Mb}\int S\left(P_{O_2}(r)\right)dr}{\alpha_{O_2}\int P_{O_2}(r)dr}. \quad (8)$$

It is then possible to numerically integrate (8) based on $S$ in (4) and $P_{O_2}(r)$ in (5). The efficiency of an $O_2$-storage protein is thus reflected in Eq. (8) which measures the amount of stored $O_2$ compared to free $O_2$ globally within the cell.

#### 2.3.2. Oxygen transport

The overall oxygen transport from sarcolemma to mitochondria can be partitioned into contributions from $MbO_2$ and free $O_2$:

$$j_{O_2} = j_{O_2}^{Mb} + j_{O_2}^{O_2} = D_{Mb}C_{Mb}\frac{dS\left(P_{O_2}(r)\right)}{dr} + D_{O_2}\alpha_{O_2}\frac{dP_{O_2}(r)}{dr} \quad (9)$$

where $j_{O_2}$ is the flux density of total $O_2$, *i.e.* the amount of $O_2$ passing through a unit area of the muscle cell per time unit, and $j_{O_2}^{Mb}$ and $j_{O_2}^{O_2}$ are flux densities of Mb and free $O_2$, respectively. Integrating (9) over the radial distance, the ratio of the total Mb-facilitated oxygen flux $J_{O_2}^{Mb}$, to the free $O_2$ flux $J_{O_2}^{O_2}$ is defined as the oxygen transport ratio (OTR):

$$OTR = \frac{J_{O_2}^{Mb}}{J_{O_2}^{O_2}} = \frac{D_{Mb}\int\frac{\partial[MbO_2]}{\partial r}dr}{D_{O_2}\int\frac{\partial[O_2]}{\partial r}dr} = \frac{D_{Mb}C_{Mb}\int\frac{\partial S}{\partial r}dr}{D_{O_2}\alpha_{O_2}\int\frac{\partial P_{O_2}(r)}{\partial r}dr} = \frac{D_{Mb}C_{Mb}(S_u - S_l)}{D_{O_2}\alpha_{O_2}(P_u - P_l)}. \quad (10)$$

Here, $S_u$ and $S_l$ are the values of saturation at the limiting values of oxygen pressure within the cell. In previous works, Eq. (10) was simplified by assuming $P_l = 0$ mm Hg and $S_l = 0$ to give (Lin et al., 2007a; Gros et al., 2010):

$$OTR = \frac{J_{O_2}^{Mb}}{J_{O_2}^{O_2}} = \frac{D_{Mb}C_{Mb}S_u}{D_M\alpha_{O_2}P_u} = \frac{D_{Mb}C_{Mb}}{D_M\alpha_{O_2}(P_u + P_{50})}. \quad (11)$$

Knowing $P_{O_2}$ over the radial distance, we calculate the exact value of $P_l$ and thus OTR from Eq. (10) to consider the effect of $P_{O_2}$ gradient on the transport efficiency of the protein.

In addition to the storage and transport of $O_2$ within the muscle cell, Mb is thought to have another $O_2$-related function which is best described as a "buffer capacity" (Bucci, 2009). According to this concept, the oxygen gradient within the cell is buffered at values of $P_{O_2}$, determined by the sensitivity of the Mb to changes of $P_{O_2}$. A buffer capacity can then be defined as $\frac{\partial S}{\partial P}$. However, since such buffer capacity and the $O_2$ flux density in Eq. (9) both depend on the $\frac{\partial S}{\partial P}$ term (*i.e.* saturation gradient over the pressure range), the buffer capacity is redundant with the instantaneous $O_2$ transport.

### 2.4. Method and parameters used

Parameters used in the calculations have been chosen to represent cellular conditions of sperm whale Mb and are summarized in Table 1. For comparison, we also later consider other typical $O_2$ consumption rates and Mb concentrations reflecting human fiber type I as an example of a terrestial mammal. For $C_{Mb}$ in sperm whale, a value of 5.03 g per 100 g wet muscle was used. This is the average of two reported values (Scholander, 1940; Tawara, 1950) which gives $C_{Mb} = 3.1\times10^{-3}$ mol $L^{-1}$ using a tissue density of 1.06 kg $L^{-1}$ and 17,200 Da as molecular weight of sperm whale Mb (Mendez and Keys, 1960; Zaia et al., 1992). In comparison, the concentration range is 1.8–5.8 g per 100 g muscle in cetaceans and 2.7–8.1 g per 100 g muscle in

**Table 1**
The list of parameters and variables used to model the physiological conditions of sperm whale Mb.

| Parameter | Symbol | Value |
|---|---|---|
| Average oxygen pressure in cell | $P_{O_2}$ | $P_l < P_{O_2} < P_u$ |
| Oxygen saturation of myoglobin | $S$ | 0–1 |
| $P_{O_2}$ at which $S = 0.5$ | $P_{50}$ | ~1–3 mm Hg for wild type Mb |
| Upper cellular oxygen pressure (at sarcolemma) | $P_u$ | 5–40 mmHg |
| Lower cellular oxygen pressure (at mitochondria) | $P_l$ | $0 < P_l < P_u$ |
| Myoglobin diffusion coefficient | $D_{Mb}$ | $7.85\times10^{-12}$ $m^2s^{-1}$ |
| Free oxygen diffusion coefficient | $D_{O_2}$ | $1.16\times10^{-9}$ $m^2s^{-1}$ |
| Oxygen solubility in the muscle tissue | $\alpha_{O_2}$ | $9.4\times10^{-7}$ $molL^{-1}mmHg^{-1}$ |
| Mb concentration in sperm whale muscle tissue | $C_{Mb}$ | $3.1\times10^{-3}$ $molL^{-1}$ |
| Concentration of $O_2$-bound myoglobin | $[MbO_2]$ | ~$S$ $C_{Mb}$ |
| Muscle oxygen consumption rate | $\dot{V}_{O_2}$ | $0.3-1.5\times10^{-6}$ $molL^{-1}s^{-1}$ |
| Radius of Krogh cylinder | $R_k$ | $19-100\times10^{-6}$ m |
| Radius of cell free region | $R_{cfr}$ | $3.5\times10^{-6}$ m |

seals (Noren and Williams, 2000). For human fiber type I, the value $C_{Mb} = 0.25 \times 10^{-3}$ mol $L^{-1}$ is used (Nemeth and Lowry, 1984, and within the range of 0.2–0.28 range of Duteil et al., 2004).

Since no experimental value has been reported for $\dot{V}_{O_2}$ in the working or resting sperm-whale muscle cell, we use the metabolic mass adjusted $\dot{V}_{O_2}$ due to the available data on the Weddell seal (Davis and Kanatous, 1999):

$$\dot{V}_{O_2}(sperm\ whale-\text{organ}) = \dot{V}_{O_2}(\text{seal}-\text{organ}) \times \left(\frac{\text{Mass}_{sperm\ whale}}{\text{Mass}_{\text{whale}}}\right)^{-0.25}. \quad (12)$$

Using 35,000 kg as the body weight of sperm whales and 450 kg for a typical Weddell seal (Whitehead, 2002), Eq. (12) gives ~0.3 μmol $L^{-1}$ $s^{-1}$ for resting $\dot{V}_{O_2}$ corrected for extracellular volume (Groebe, 1995). We then apply a 5-fold larger value to mimic diving conditions (Davis and Kanatous, 1999). For the muscle tissue of a 70-kg human, Eq. (12) gives $\dot{V}_{O_2} \cong 1.5\ \mu mol\ L^{-1} s^{-1}$. To account for the role of the metabolic rate, the reported value of $\dot{V}_{O_2} \cong 300$ mL $min^{-1}$ $kg^{-1} = 201$ μmol $L^{-1}$ $s^{-1}$ is employed for the maximally working human skeletal muscle (Bangsbo, 2000). A value of $7.85 \times 10^{-12}$ $m^2$ $s^{-1}$ is also selected for the Mb diffusion constant, based on the most up-to-date reported *in vivo* experiments (Lin et al., 2007a,b).

For free oxygen within the cell, a solubility constant of 0.94 μmol $L^{-1}$ $mmHg^{-1}$ and a diffusion coefficient of $1.16 \times 10^{-9}$ $m^2$ $s^{-1}$ were used (Groebe, 1995). The values of 3.5 μm and 19 μm were also used for the radius of cell free region ($R_{cfr}$) and Krogh cylinder ($R_k$) as shown in Fig. 2. $R_K$ is allowed to change up to 100 μm, to evaluate the effects of cell size on the storage and transport proficiency.

The kinetic and thermodynamic oxygenation data for 90 mutants were reported by Olson and coworkers (Scott et al., 2001). The mutations occur in positions both near and far from the heme. Descriptors (OSR and OTR) have been calculated at $P_u = 40$ mm Hg and $P_u = 5$ mm Hg, corresponding to normoxic and hypoxic conditions, respectively.

For each mutation, a rank number is used to describe the approximate distance of the mutated site from the heme group (Table 2). This rank is assigned from 1 (wild type), 2 (His64 → Gly or H64G) to 90, to include all mutations in the first shell, second shell, proximal site, and far from the heme group, respectively. Table 2 also shows the thermodynamic constants of the first equilibrium reaction in $O_2$ binding, $K_1$, the overall oxygenation equilibrium constant $K_{O_2}$ and $P_{50}$ for all mutants.

**Table 2**
Thermodynamic constants and half-saturation pressure $P_{50}$ for wild type sperm whale myoglobin and mutants. Each mutated site has a rank, illustrating the approximate distance to heme, classified as either first-shell, second-shell, proximal site, or far-from-heme residue. Data are from Scott et al., 2001.

| Rank | Position | Mutant | $K_1$ ($\mu M^{-1}$) | $K_{O2}$ ($\mu M^{-1}$) | $P_{50}$ (mm Hg)[a] | Rank | Position | Mutant | $K_1$ ($\mu M^{-1}$) | $K_{O2}$ ($\mu M^{-1}$) | $P_{50}$ (mm Hg)[a] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | First shell | WT | 5.40E-06 | 1.10 | 0.96 | 46 | Second shell | V66R | 3.71E-06 | 1.08 | 0.98 |
| 2 | | H64G | 2.57E-05 | 0.09 | 12.0 | 47 | | T67F | 3.46E-06 | 1.47 | 0.72 |
| 3 | | H64A | 1.37E-05 | 0.02 | 53.0 | 48 | | T67K | 5.10E-06 | 1.04 | 1.02 |
| 4 | | H64V | 1.79E-05 | 0.01 | 106 | 49 | | T67Q | 3.62E-06 | 0.90 | 1.18 |
| 5 | | H64L | 9.55E-06 | 0.02 | 53.0 | 50 | | T67P | 4.82E-06 | 2.00 | 0.53 |
| 6 | | H64F | 1.73E-05 | 0.01 | 106 | 51 | | T67A | 1.18E-05 | 1.10 | 0.96 |
| 7 | | H64W | 8.60E-07 | 0.07 | 15.1 | 52 | | I107V | 1.14E-05 | 1.20 | 0.88 |
| 8 | | V68A | 2.93E-06 | 1.20 | 0.88 | 53 | | I107T | 9.67E-06 | 1.40 | 0.76 |
| 9 | | V68T | 5.38E-07 | 0.07 | 15.1 | 54 | | I107L | 7.92E-06 | 1.20 | 0.88 |
| 10 | | V68I | 1.29E-06 | 0.23 | 4.61 | 55 | | I107F | 2.88E-06 | 1.60 | 0.66 |
| 11 | | V68L | 3.87E-06 | 3.40 | 0.31 | 56 | | I107W | 1.62E-06 | 3.30 | 0.32 |
| 12 | | V68F | 1.00E-07 | 0.48 | 2.21 | 57 | | I107V | 5.83E-06 | 1.20 | 0.88 |
| 13 | | V68W | 2.00E-08 | 0.59 | 1.80 | 58 | | I111L | 6.83E-06 | 0.90 | 1.18 |
| 14 | | L29A | 9.39E-06 | 0.78 | 1.36 | 59 | | I111F | 1.02E-05 | 0.94 | 1.13 |
| 15 | | L29V | 6.17E-06 | 1.10 | 0.96 | 60 | | I111M | 4.50E-05 | 0.63 | 1.68 |
| 16 | | FL29 | 1.68E-05 | 15.0 | 0.07 | 61 | | I111W | 6.81E-06 | 1.50 | 0.71 |
| 17 | | L29W | 6.73E-06 | 0.03 | 35.3 | 62 | Proximal site | L89G | 3.10E-06 | 1.30 | 0.82 |
| 18 | | F43V | 5.17E-05 | 0.16 | 6.63 | 63 | | L89W | 5.18E-06 | 0.28 | 3.79 |
| 19 | | F43L | 1.09E-05 | 0.21 | 5.05 | 64 | | H97A | 6.93E-06 | 2.00 | 0.53 |
| 20 | | F43I | 5.18E-06 | 0.10 | 10.6 | 65 | | H97V | 6.14E-06 | 1.10 | 0.96 |
| 21 | | F43W | 6.73E-06 | 0.17 | 6.24 | 66 | | H97D | 7.00E-06 | 1.90 | 0.56 |
| 22 | Second shell | I28A | 8.45E-06 | 0.96 | 1.10 | 67 | | H97F | 4.42E-06 | 1.50 | 0.71 |
| 23 | | I28W | 4.67E-06 | 2.30 | 0.46 | 68 | | H97Q | 5.83E-06 | 0.48 | 2.21 |
| 24 | | L32A | 8.70E-06 | 0.89 | 1.19 | 69 | | I99A | 1.80E-06 | 2.10 | 0.50 |
| 25 | | L32V | 7.62E-06 | 0.93 | 1.14 | 70 | | I99V | 4.52E-06 | 2.10 | 0.50 |
| 26 | | L32I | 5.25E-06 | 0.94 | 1.13 | 71 | | I99L | 6.94E-06 | 0.48 | 2.21 |
| 27 | | L32F | 3.91E-06 | 1.00 | 1.06 | 72 | | I99N | 4.91E-06 | 3.00 | 0.35 |
| 28 | | L32M | 3.14E-06 | 1.10 | 0.96 | 73 | | L104A | 1.04E-05 | 3.10 | 0.34 |
| 29 | | L32W | 6.13E-07 | 2.70 | 0.39 | 74 | | L104V | 4.24E-06 | 2.00 | 0.53 |
| 30 | | R45A | 6.41E-06 | 0.26 | 4.08 | 75 | | L104W | 3.01E-06 | 5.8 | 0.18 |
| 31 | | R45L | 4.66E-06 | 0.26 | 4.08 | 76 | | F138A | 1.33E-05 | 1.20 | 0.88 |
| 32 | | R45T | 6.89E-06 | 0.17 | 6.24 | 77 | | F138W | 3.14E-06 | 0.79 | 1.34 |
| 33 | | R45K | 4.66E-06 | 0.31 | 3.42 | 78 | Far from heme | W7F | 5.08E-06 | 0.80 | 1.33 |
| 34 | | R45S | 8.10E-06 | 0.42 | 2.52 | 79 | | Q8V | 6.67E-06 | 0.49 | 2.16 |
| 35 | | R45A | 1.23E-05 | 0.06 | 17.7 | 80 | | W14F | 6.99E-06 | 0.66 | 1.61 |
| 36 | | F46V | 7.40E-05 | 0.07 | 15.1 | 81 | | M55A | 3.46E-06 | 1.20 | 0.88 |
| 37 | | F46L | 2.42E-05 | 0.18 | 5.89 | 82 | | M55L | 2.21E-06 | 0.90 | 1.18 |
| 38 | | F46W | 9.72E-06 | 0.28 | 3.79 | 83 | | M55W | 2.57E-06 | 0.83 | 1.28 |
| 39 | | F46A | 2.75E-06 | 0.43 | 2.47 | 84 | | A71F | 6.59E-06 | 0.78 | 1.36 |
| 40 | | L61F | 2.50E-06 | 1.10 | 0.96 | 85 | | L72V | 5.87E-06 | 1.60 | 0.66 |
| 41 | | L61A | 4.56E-06 | 0.88 | 1.20 | 86 | | L72W | 4.71E-06 | 0.73 | 1.45 |
| 42 | | G65I | 4.93E-06 | 0.52 | 2.04 | 87 | | K79A | 7.21E-06 | 1.00 | 1.06 |
| 43 | | G65T | 1.38E-05 | 1.54 | 0.69 | 88 | | K79L | 6.54E-06 | 1.00 | 1.06 |
| 44 | | G65G | 4.39E-06 | 1.79 | 0.59 | 89 | | M131L | 8.00E-06 | 0.32 | 3.31 |
| 45 | | V66K | 4.24E-06 | 1.25 | 0.85 | 90 | | A144V | 6.41E-06 | 0.83 | 1.28 |

[a] Calculated due to $P_{50} = \frac{1}{\alpha_{O_2}(K_{O_2} - K_1)}$.

To calculate OSR one obtains $P_{O_2}(r)$ within the cell from Eq. (9), $S$ from Eq. (4) and then OSR from Eq. (8), using numerical integration. For OTR, $P_l$ and $S_l$ in Eq. (10) are also calculated from $P_{O_2}(r)$ at $r = R_k$ and $S$ ($P(R_k)$). The ratios thus implicitly depend on the muscle oxygen consumption rate ($\dot{V}_{O_2}$) that differs at rest and workload conditions. Numerical integration procedures for evaluating $P_{O_2}(r)$ and the functional descriptors were carried out with the software MATLAB (Mathworks, vR2010a). A termination protocol was implemented for the integration over the radial distance when $P_{O_2}(r)<0$. This corresponds to effectively zero oxygen partial pressure near the mitochondria under working conditions of large $\dot{V}_{O_2}$.

## 3. Results and discussion

The main purpose of the present work is to quantify and understand the functional proficiency of various Mb mutants under changing environmental conditions, which has not been attempted before, and subsequently, to deduce the selection-pressure that conserves parts of the WT protein. Using the two main functional descriptors, OSR and OTR, we evaluate the storage and transport ability of the various proteins at different physiological conditions. The use of different parameters in the model also enables us to compare different species. The dependence of OSR and OTR on $P_{O_2}$ and $\dot{V}_{O_2}$ is described in detail, where the effects of other relevant parameters are presented in the supplementary material.

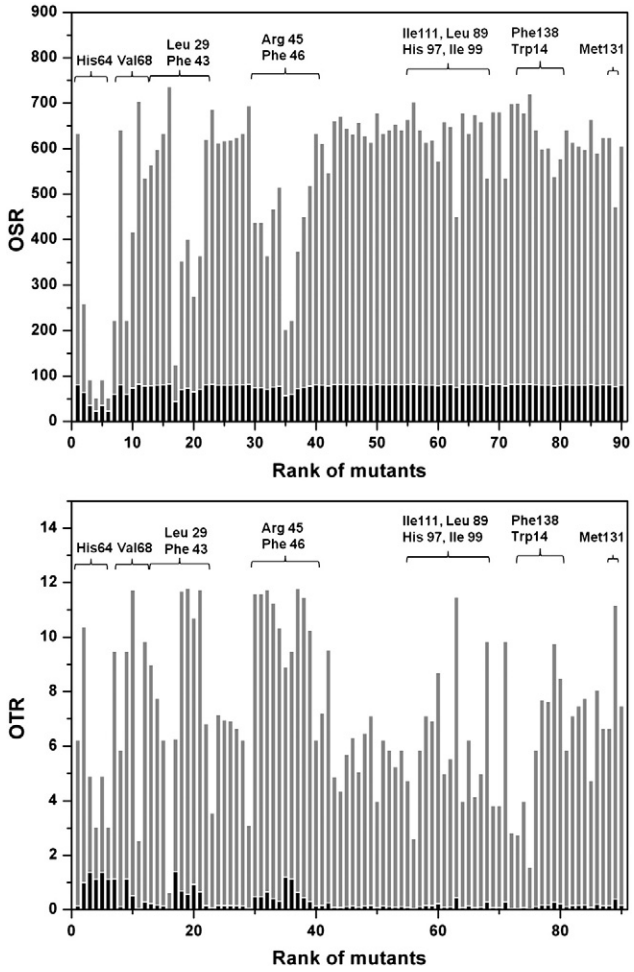



**Fig. 3.** The OSR (top) and OTR (bottom) values at upper pressure $P_u = 40$ mm Hg (black bars) and $P_u = 5$ mm Hg (gray bars), corresponding to normoxic and hypoxic conditions. Both OSR and OTR are calculated for sperm whale having $\dot{V}_{O_2} = 0.3$ μ mol $L^{-1}$ $s^{-1}$, $C_{Mb} = 3.1$ m mol $L^{-1}$ and $R_k = 19$ μm.

### 3.1. Oxygen storage of sperm-whale Mb

Fig. 3, top, shows the values of OSR for sperm whale Mb at $P_u = 40$ mm Hg and $P_u = 5$ mm Hg at $\dot{V}_{O_2} = 0.3$ μmol $L^{-1}$ $s^{-1}$, which corresponds to normoxic rest conditions, *versus* the rank of each mutant as given in Table 2. Because of the well-known saturation behavior of Mb, lower $P_{O_2}$ causes a higher contribution of Mb to the overall oxygen storage, compared to free $O_2$, so that Mb strongly compensates the depletion of free $O_2$. As anticipated directly from the binding constants, OSR is severely affected by mutations close to the distal pocket, but we can now quantify the physiological effect directly: the average OSR for the whole data set increases from ~80 to ~553, suggesting a 6-fold increase in the ratio of stored $O_2$ relative to free $O_2$ upon transition from normoxic to hypoxic conditions. In physiological terms, this implies that the relative importance as an $O_2$-storage of an average Mb mutant (including the WT, since the average mutant functionality is close to that of the WT) increases ~6-fold upon transition to hypoxic conditions.

The major changes are due to the mutations of residues His-64, Val-68, Leu-29, and Phe-43 in the first shell, and Arg-45, Phe-46, Leu-61, and Gly-65 in the second shell of the distal pocket. Small effects are observed for mutations in Ile-111, Leu-89, His-97, Ile-99, and Phe-138 in the proximal site, and in Met-131 far from heme. Almost all these mutations impair the ability of protein to store oxygen, and the impairment is highly dependent on $P_{O_2}$.

The mutations affecting His-64 can be understood from a structure–function relationship involving the hydrogen bond from site 64 to $O_2$ as suggested by Olson (2008). For example, the H64V mutation completely lacks the hydrogen bond, and the thermochemical data reveal a two-orders-of-magnitude decrease in the binding constant. However, we can now see that the physiological effect on $O_2$-storage is somewhat smaller in the whale: OSR decreases from ~80 to ~23 at normoxic conditions, and from ~553 to ~30 under hypoxic conditions, given the very small whale muscle oxygen consumption rate of $\dot{V}_{O_2} = 0.3$ μmol $L^{-1}$ $s^{-1}$. We shall see later that the effect is much larger in smaller terrestial mammals (humans) with substantially higher $O_2$-consumption rates.

Fig. 4 shows the $P_{O_2}$-dependence of the OSR for WT Mb and three low-affinity mutants (H64G, H64A and H64V). Importantly, as is not clear from the thermochemical data directly, the difference between WT and mutants is insignificant above $P_u = 20$ mm Hg. The relative importance of Mb storage is largest at low $P_{O_2}$ and increases dramatically as $P_{O_2}$ decreases, in particular for the WT and less so for low-affinity mutants. Thus, under hypoxic stress, Mb has a significant role in providing oxygen to maintain $O_2$ consumption in the muscle cell and furthermore, under hypoxia the WT "shows its value", *i.e.* the hydrogen bond to the His-64 becomes much more significant. Accordingly, we conclude that hypoxia is the dominant selection pressure for the preservation of His-64 in WT Mb, given that storage is a critical function of Mb. Importantly, the WT is superior at all $P_{O_2}$ when it comes to storage, but as we shall see, the case is very different in terms of active $O_2$-transport.

### 3.2. Oxygen transport of sperm-whale Mb

Using Eq. (10), OTR values can similarly be calculated for the whole mutant data set for any given $P_u$ of the cell. The bottom of Fig. 3 shows the OTR values for the WT and mutants at $P_u = 40$ mm Hg and $P_u = 5$ mm Hg at $\dot{V}_{O_2} = 0.3$ μmol $L^{-1}$ $s^{-1}$. At $P_u = 40$ mm Hg, facilitated $O_2$-transport is negligible, since most mutants contribute ~20% or less to the overall $O_2$ flux. The average OTR is about 0.32, and most surprisingly, the WT has only 11% contribution to the total $O_2$ flux in the cell under normoxic conditions, much less than the low-affinity mutants.

However, at low $P_{O_2}$, Mb has a significant role in $O_2$-transport. The explanation is that $S$ changes over lower $P_{O_2}$ regions of the cell, giving rise to a larger [$MbO_2$] gradient and hence, a larger flux of $MbO_2$. Thus,

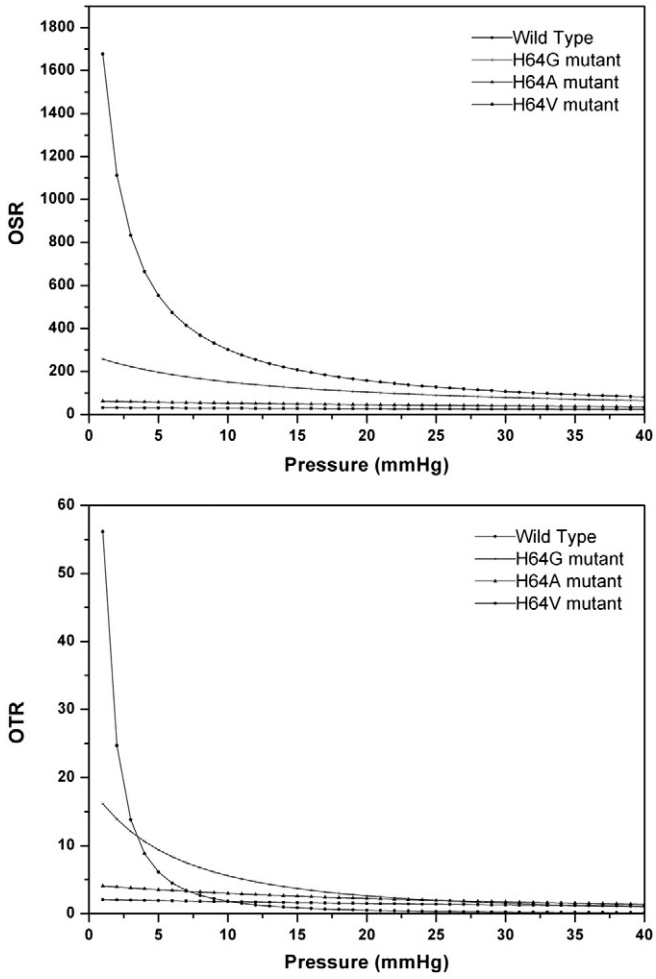



**Fig. 4.** The OSR (top) and OTR (bottom) *vs.* $P_{O_2}$ for wild type Mb and H64G, V68A, and H64V mutants. The calculations are based on $\dot{V}_{O_2} = 0.4$ μ mol $L^{-1}$ $s^{-1}$, $C_{Mb} = 3.1$ m mol $L^{-1}$, and $R_k = 19$ μm.

we identify a crucial difference between the behavior of the transport and storage functions under changing environmental conditions, showing that the two functions are intrinsically distinct.

In Fig. 4, bottom, OTR is compared for the WT and mutants in the range from $P_u = 40$ mm Hg to $P_u = 1$ mm Hg. It can be seen that Mb contributes substantially to $O_2$-transport only below $P_u \sim 10$ mm Hg. Similar to OSR, OTR is most affected by replacements within the first and second shells, while mutations within the third shell generally retain WT function.

An important observation from our model is that the proficiency of the WT as compared to other mutants depends very much on $P_{O_2}$. As seen in Fig. 3 bottom, OTR for low-affinity mutants (primarily those with mutations in sites 64, 68, 29, and 43) is in fact *greater* than WT at higher $P_{O_2}$. Furthermore, this behavior is inversed at lower $P_{O_2}$ where the WT is much more proficient in terms of transport, compared to these mutants. In fact, over the $P_{O_2}$ range, there are "critical pressures" at which two mutants are equally proficient despite their different $K_{O_2}$, as is shown in Fig. 4.

The reason for this surprising observation is that OTR, as calculated from Eq. (10), depends mainly on the derivative of saturation $\left(i.e.\ \frac{dS}{dP}\right)$, which has a $P_{O_2}$-dependent behavior. This behavior is shown in Fig. 5, where the slopes of the saturation curve of the WT and the H64G mutant are compared at $P_u = 20$ mm Hg and $P_u = 2$ mm Hg, close to the $P_{50}$ of ~1 mm Hg of the wild type protein. It can be seen that the saturation curves for mutants with higher $P_{50}$ are less steep at lower $P_{O_2}$ and steeper at normoxic conditions, compared to the WT. This situation causes the WT to be a better transport protein at lower $P_{O_2}$,

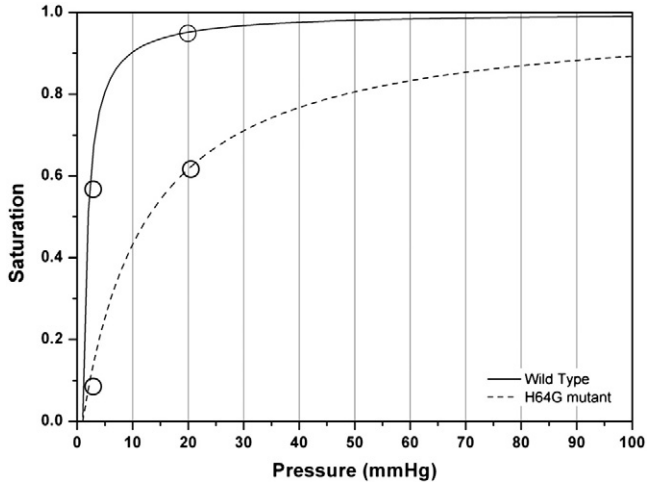



**Fig. 5.** The saturation curve for WT Mb (solid line) and the H64G mutant (dashed line). The circles compare saturation of two proteins at $P_{O2} = 20$ mm Hg and 2 mm Hg. Calculations are based on sperm whale having $\dot{V}_{O_2} = 0.3$ μ mol $L^{-1}$ $s^{-1}$, $C_{Mb} = 3.1$ m mol $L^{-1}$, and $R_k = 19$ μm.

whereas low-affinity mutants are in fact better transporters at normoxic conditions.

Two important consequences of this result are that i) the proficiency of a mutant depends on environmental conditions and ii) it can surpass that of WT. The general fact that the functional proficiency of a mutant can surpass that of the WT is well-known from the directed-evolution field, where mutants of enzymes are often sought that outperform the WT at *e.g.* rough, industrial conditions. The observation also has important implications for the natural evolution of Mb, as we will discuss below.

### 3.3. The proficiency of mutants under varying conditions

So far we have examined only rest conditions of sperm whale Mb. Now we will evaluate the effect of increased $\dot{V}_{O_2}$, resembling hard work, *e.g.* during diving, or otherwise elevated metabolism, on the function of Mb mutants. From Eq. (6), a higher $\dot{V}_{O_2}$ results in lower $P_{O_2}$ within the cell and a steeper gradient. This in turn results in a larger significance of Mb, both in terms of storage and function.

As shown in Table 3, whale Mb globally within the cell contains 80 times more $O_2$ than is freely available, due to the high concentration of Mb, and thus the role of Mb in $O_2$-storage is substantial. This ratio is largely unaffected by changes in consumption rate, and the change is marginal (from 485 to 488 as consumption increases) even at hypoxic conditions. The same lack of effect of consumption rate is seen on the active transport ability. OSR and OTR both increase less than 2% as a result of increased $\dot{V}_{O_2}$ during diving. The reason is that sperm whale has one of the lowest mass-specific metabolic rates known, and the $\dot{V}_{O_2}$ values are thus so small that a 5-fold increase in $\dot{V}_{O_2}$ does not affect the $P_{O_2}$ gradient enough to change OSR and OTR.

From Eqs. (8) and (10), OSR and OTR also depend on $C_{Mb}$ and $R_k$. Both functions scale linearly with $C_{Mb}$ at all physiological conditions. The WT Mb shows an increased OSR from 80 to 162 and OTR from 0.13 to 0.26 when $C_{Mb}$ is doubled at $P_u = 40$ mm Hg and $R_k = 19$ μm. It is thus not surprising that the adaptation strategy of marine mammals has been toward a largely increased $C_{Mb}$ compared to the terrestrial mammals. This is also in agreement with the compensatory, increased expression of Mb as a result of living under hypoxic conditions (*e.g.* high altitude).

Both OSR and OTR depend on changes in $P_l$, the oxygen partial pressure near the mitochondria. In addition to increased muscle consumption rate, a larger cell size $R_k$ will also cause lower $P_l$ values and thus, due to both lower global $P_{O_2}$, cause higher storage and

**Table 3**
OSR and OTR for sperm whale at rest and diving conditions, and at normoxic ($P_u = 40$ mm Hg) and hypoxic ($P_u = 5$ mm Hg) conditions.

| | Resting metabolic rate $\left(\dot{V}_{O_2} = 0.3\ \mu mol\ L^{-1}\ s^{-1}\right)$ | | Diving metabolic rate $\left(\dot{V}_{O_2} = 1.5\ \mu mol\ L^{-1}\ s^{-1}\right)$ | |
|---|---|---|---|---|
| | $P_u = 40$ mm Hg | $P_u = 5$ mm Hg | $P_u = 40$ mm Hg | $P_u = 5$ mm Hg |
| OSR | 80 | 553 | 81 | 556 |
| OTR | 0.13 | 6.10 | 0.13 | 6.14 |

transport proficiency. However, in the sperm whale with its very low metabolic rate, these effects become relatively unimportant: Increasing $R_k$ from 19 μm to 100 μm only changes the average OSR from ~76 to ~107 and the average OTR from ~0.32 to ~0.42 at $P_u = 40$ mm Hg and $\dot{V}_{O_2} = 0.3\ \mu mol\ L^{-1} s^{-1}$. However, in other (terrestrial) animals with much larger consumption rates, these parameters play a key role in the physiological proficiency of Mb (*vide infra*), and it is meaningless to discuss "rest" or "work" without referring to the general state of these parameters, since they all affect the $O_2$-profile and Mb function. The various effects of the parameters on the whole mutant data set are presented in a set of figures in the supplementary material.

### 3.4. Distribution of protein function

In this section, we look at the distribution of functionality as we change the physiological parameters. As is shown in Fig. 6, for all $P_u$ levels in the cell, the distributions of functional proficiency tend to be centered near the WT proficiency, showed in dashed lines. This observation is due to the fact that many mutations are less likely to affect protein function even when these mutations were selected for their possible impairment of $O_2$-binding: In fact, the large majority of all mutants will have $O_2$-binding characteristics similar to the WT (the neutral theory of evolution). We define these mutations as "neutral in term of function" and thus make an analogy to the neutral theory of evolution, which states that most possible mutants will have similar fitness to the WT and thus, high mutation rates in genes. Importantly, we find that the functionality of the WT is close to the average functionality of all mutants. This will be even truer in large data sets, since the mutants in the data set were generally close to heme or in the pathways of $O_2$ migration within the protein, compared to the "average" amino acid.

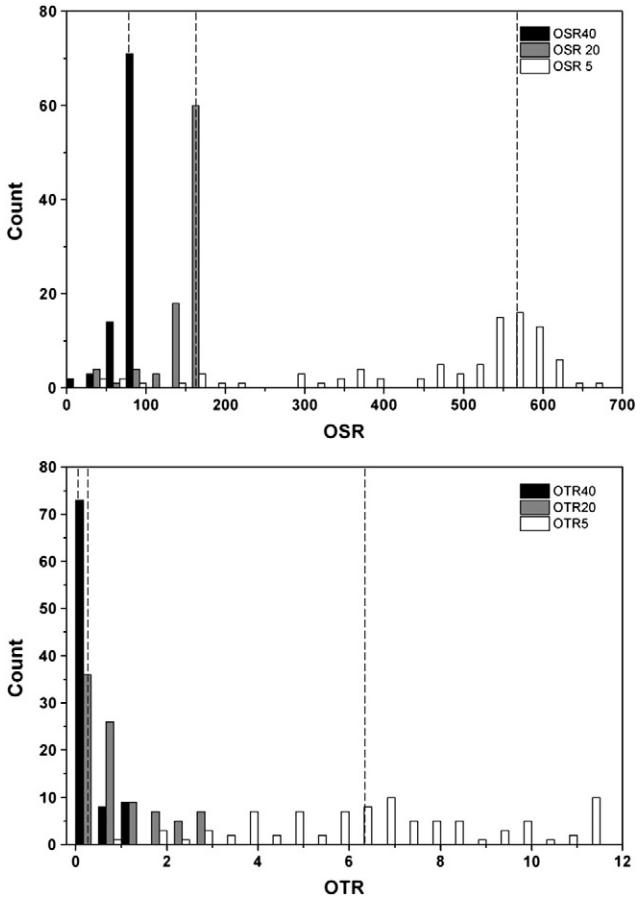



**Fig. 6.** The OSR (up) and OTR (down) distributions among mutants for three upper pressures of $P_u = 40$ mm Hg (black), 20 mm Hg (gray), and 5 mm Hg (white). The calculations.

Some mutants differ substantially from the WT, as seen in Fig. 5. These mutations are either negative (*i.e.* display impaired functionality) or positive (*i.e.* display enhanced functionality) and form the tails of the distributions. Interestingly, higher $\dot{V}_{O_2}$ or lower $P_{O_2}$ not only increases the role of Mb in active oxygen delivery, but also increases the *variation* in functionality among all mutants. In fact, variation scales approximately as ~ $\frac{1}{P_u}$. The very narrow distribution of functionality at $P_u = 40$ mm Hg changes to a much wider distribution at $P_u = 5$ mm Hg, accompanied by an increased variance from ~148 to ~2526 for OSR and from ~0.13 to ~7.2 for OTR. It is tempting to deduce from this that the selection pressure is much stronger where the variation is large, *i.e.* where more mutants are impaired, relative to the WT: From an evolutionary point of view, larger functional variation favors the selection of beneficial mutations, including the WT protein. Thus, the distributions of mutant proficiency supports the hypothesis that hypoxia is the main selection pressure for WT Mb. It also suggests a correlation between the concept of neutral mutations by high mutation rate and our concept of "neutral in terms of function".

### 3.5. Comparison of Mb functions in whales and in humans

It is possible to distinguish between the major roles of Mb in different species via comparison of the OSR and OTR values using organism specific parameters. Assuming an identical $D_{Mb}$ in different species, the relevant parameters are $C_{Mb}$, cell size, and oxygen consumption rate in the muscle, $\dot{V}_{O_2}$, as well as $P_{O_2}$ of the habitat. Any change in these parameters can affect the cellular $P_{O_2}$ gradient and thus OSR and OTR. Since the major driving force of diffusion flux is the concentration gradient along the diffusion path, transport proficiency of WT Mb is more significant in species that have generically higher $\dot{V}_{O_2}$ or at conditions where either consumption rate is increased or $P_{O_2}$ is low. When $P_l$ and consequently $S_l$ go toward zero (see the supplementary materials) OTR only depends on $P_u$ and $S_u$ as represented by Eq. (11) and is independent of the $P_{O_2}$ gradient.

Table 4 compares the $P_{50}$, OSR, and OTR values for sperm whale and human WT myoglobin at rest and work conditions. Here, $P_{50}$ is assumed constant for the aerobic metabolism and the Bohr effect is neglected. $C_{Mb}$ is also considered to be constant during the transition from rest to work (Duteil et al., 2004). Since $C_{Mb}$ for sperm whale is ~12-fold higher than that of human, both OSR and OTR are greater in sperm whale by the same factor under resting conditions. Importantly, due to the low mass-specific metabolic rate of whale, the OSR and OTR change very little from rest to dive conditions. In contrast, for humans, the more than 100-fold larger $\dot{V}_{O_2}$ during work causes significant changes in OSR and OTR from rest to workload conditions. In the whale, the very high $C_{Mb}$ means that Mb is always the dominating storage for $O_2$. At hypoxic conditions (during dives), Mb supplies the majority (about six times more, *i.e.* 6/7) of the $O_2$ to the mitochondria, whereas Mb plays no role in active $O_2$-transport at normoxic conditions (*e.g.* at the surface or early during the dive). In humans, Mb only plays a significant role in active transport at severely hypoxic work conditions.

**Table 4**
Comparison of OSR and OTR for whale and human at normoxic ($P_u = 40$ mm Hg) and hypoxic ($P_u = 5$ mm Hg) $O_2$-partial pressure.

| Quantities | Species | | | |
|---|---|---|---|---|
| | Sperm whale | | Human | |
| | Rest | Dive | Rest | Work |
| $P_{50}$ (mm Hg) | 0.96[a] | 0.96 | 2.75[b] | 2.75 |
| $C_{Mb}$ ($10^{-3}$ mol $L^{-1}$) | 3.1 | 3.1 | 0.25 | 0.25 |
| $\dot{V}_{O_2}$ (μM $s^{-1}$) [c] | 0.3 | 1.5 | 1.5 | 201 |
| OSR 40 | 80.9 | 82.6 | 6.7 | 15.5 |
| OSR 5 | 555 | 567 | 49.6 | 79.3 |
| OTR 40 | 0.13 | 0.13 | 0.01 | 0.43 |
| OTR 5 | 6.14 | 6.32 | 0.56 | 2.99 |

[a] Calculated due to $P_{50} = \frac{1}{\alpha_{O_2}(K_{O_2} - K_1)}$.
[b] Taken from Rossi-Fanelli and Antonini (1958).
[c] Calculated from the Eq. (12).

### 3.6. Perspectives

The question whether a selection pressure operates to fixate the conserved residues in mammalian WT Mb can be answered in a new way by comparing the functional proficiency of relevant mutants. In sperm whale (and possibly other marine mammals) it is mainly the $O_2$-storage proficiency of Mb that justifies the protein. However, at severely hypoxic conditions, active transport becomes very distinct from the storage function and helps to increase the availability of $O_2$ near the mitochondria. A selection pressure thus operates to favor Mb at very low $P_{O_2}$ mainly due to active transport. Since the functional variation is also substantially larger at low $P_{O_2}$, we conclude that the fixation of the distal pocket in mammalian Mb in general and specially His-64 occurs under hypoxic conditions, where the WT is significantly more proficient.

Negative mutations impair the fitness by decreasing the reproduction or survival probabilities, *e.g.* via marginal changes to the aerobic dive limit, or for terrestrial animals, the maximum running distance, which both adversely affect the action radius. Since about 26–47% of stored $O_2$ in diving mammals resides in the muscle tissue (Kooyman and Ponganis, 1998), insufficient $O_2$-storage of Mb due to mutations near or at His-64 reduces the action radius, and consequently the feeding and mating success. The sequence–function relations developed in this work show that modern biology can bridge all the way from chemical biology to physiology in order to understand the evolution of biological matter.

We have shown in this work that the functional proficiency of WT Mb (OSR and OTR) varies in relation to $P_{O_2}$, $C_{Mb}$, $\dot{V}_{O_2}$, and the cell size, $R_k$. Since OTR scales linearly with $D_{Mb}$, it would be interesting to investigate whether $D_{Mb}$ has also been adapted in terrestrial *vs.* marine mammals. $D_{Mb}$ should perceivably be optimized by mutations on the protein surface to reduce the effective viscosity of the sarcoplasm. This would be in accordance with the recent observation of changes in Mb solubility upon increased surface charge in Mbs of birds and marine mammals (Berenbrink and Mirceta, 2009; Mirceta et al., 2009).

Using the computed OTR values, we can also identify the physiological conditions where Mb mutants are more proficient than WT. One of the most striking results is that low-affinity mutants may in fact be better transport proteins at intermediate $P_{O_2}$ (Fig. 3, bottom), since the WT reaches saturation (and hence induces a more shallow $P_{O_2}$-gradient) already at low $P_{O_2}$. WT Mb is thus not proficient in transporting $O_2$ at normal $P_{O_2}$ compared to many low-affinity mutants, especially those involving destruction of the hydrogen bond to His-64. We anticipate that it will be possible to verify this theoretical finding in a steady-state $O_2$-flow set up where the active $O_2$-diffusion of His-64 mutants *vs.* WT is measured.

## 4. Conclusion

Using the theory of oxygen delivery within the muscle cell, we have ranked 90 myoglobin mutants whose sequences and thermochemical properties are known (Scott et al., 2001), according to their physiological proficiency, using a global model of the muscle applicable to a range of organism-specific and environmental variables and parameters.

We have shown that in the whale, with its extremely low mass-specific metabolic rate, the WT is only a superior transport protein at low $P_{O_2}$ (<3 mm Hg). The transport ability of the WT, which we find is distinct from the storage ability, is surpassed by low-affinity mutants at intermediate and normoxic conditions, suggesting that hypoxia is the selection pressure that preserves WT Mb if transport is critical.

Instead, WT Mb is marginally superior as an $O_2$-storage protein at all $P_{O_2}$ but only significantly (*i.e.* two times or more) better than low-affinity mutants below $P_u \sim 5$ mm Hg (Fig. 3, top). Assuming that selection occurs due to these two functions, we deduce based on the above observations, that the $O_2$ affinity must has been tuned during the course of evolution to maintain the most proficient storage protein and to additionally enhance $O_2$-transport primarily under hypoxic conditions.

For other animals, specifically smaller terrestrial mammals with much higher metabolic rates, the WT Mb contributes substantially to both storage and transport within the muscle cell at normoxic conditions, as seen in Table 4, and the role is now much enhanced during hypoxia. Thus, our work quantifies organism-specific differences in the function of Mbs.

The identified distinct influences of metabolic rate and hypoxia on the two Mb functions may have relevance to other fields of science, *e.g.* relating to hypoxic stress and metabolic disorders in medicine.

Within the protein structure, different shells around the center of reaction can be classified due to their effects on the function of the protein. For Mb, about 60% of the replacements in the first shell impair the storage function but improve transport at intermediate $P_{O_2}$. In the second shell and the proximal sites of the heme group, replacements are less likely to affect any of the functions, since His-64 is a key residue preserving the balance between the two functions. Replacements far from heme are almost neutral with respect to both functions, since they perturb the proximal site less. Such mutations potentially constitute the vast majority of mutations possible, thus providing an argument for the neutral theory of molecular evolution from a structure-function perspective.

## Acknowledgments

This work was made possible by a grant from the Danish National Science Research Council, Project Case 272-08-0041.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.cbpa.2011.07.027.

## References

Bangsbo, J., 2000. Muscle oxygen uptake in humans at onset of and during intense exercise. Acta Physiol. Scand. 168, 457–464.
Berenbrink, M., Mirceta, S., 2009. How to make a whale: molecular signature of myoglobin in diving birds and mammals. Comp. Biochem. Physiol. A 153, S44.
Bucci, E., 2009. Thermodynamic approach to oxygen delivery in vivo by natural and artificial oxygen carriers. Biophys. Chem. 142, 1–6.
Davis, R.W., Kanatous, S.B., 1999. Convective oxygen transport and tissue oxygen consumption in Weddell seals during aerobic dive. J. Exp. Biol. 202, 1091–1113.
Duteil, S., Bourrilhon, C., Raynaud, J.S., Wary, C., Richardson, R.S., Leroy-Willig, A., Jouanin, J.C., Guezennec, C.Y., Carlier, P.G., 2004. Metabolic and vascular support for the role of myoglobin in humans: a multiparametric NMR study. Am. J. Physiol. Regul. Integr. Comp. Physiol. 287 (6), R1441–1449.

Elber, R., 2010. Ligand diffusion in globins: simulations versus experiment. Curr. Opin. Struct. Biol. 20, 162–167.
Frauenfelder, H., McMahon, B.H., Fenimore, P.W., 2003. Myoglobin: the hydrogen atom of biology and a paradigm of complexity. Proc. Natl. Acad. Sci. U. S. A. 100, 8615–8617.
Gimenez, M., Sanderson, R.J., Reiss, O.K., Banchero, N., 1977. Effects of altitude on myoglobin and mitochondrial protein in canine skeletal muscle. Respiration 34, 171–176.
Groebe, K., 1995. An easy-to-use model for O2 supply to red muscle. Validity of assumptions, sensitivity to errors in data. Biophys. J. 68, 1246–1269.
Gros, G., Wittenberg, B.A., Jue, T., 2010. Myoglobin's old and new clothes: from molecular structure to function in living cell. J. Exp. Biol. 213, 2713–2715.
Guyton, G.P., Stanek, K.S., Schneider, R.C., Hochachka, P.W., Hurford, W.E., Zapol, D.G., Liggins, G.C., Zapol, W.M., 1995. Myoglobin saturation in free-diving Weddell seals. J. Appl. Physiol. 79, 1148–1155.
Hill, R., 1936. Oxygen dissociation curves of muscle haemoglobin. Proc. R. Soc. Lond. B 120, 472–483.
Jensen, K.P., Ryde, U., 2003. Comparison of the chemical properties of iron and cobalt porphyrins and corrins. ChemBioChem 4, 413–424.
Jensen, K.P., Ryde, U., 2004. How heme binds $O_2$: reasons for reversible binding and spin inversion. J. Biol. Chem. 279, 14561–14569.
Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., Phillips, D.C., 1958. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature 181, 662–666.
Kooyman, G.L., Ponganis, P.J., 1998. The physiological basis of diving to depth: birds and mammals. Annu. Rev. Physiol. 60, 19–32.
Krogh, A., 1919. The number and the distribution of capillaries in muscle with the calculation of the oxygen pressure necessary for supplying the tissue. J. Physiol. (London) 52, 409–515.
Lin, P.C., Kreutzer, U., Jue, T., 2007a. Anisotropy and temperature dependence of myoglobin translational diffusion in myocardium: implication for oxygen transport and cellular architecture. Biophys. J. 92, 2608–2620.
Lin, P.C., Kreutzer, U., Jue, T., 2007b. Myoglobin translational diffusion in rat myocardium and its implication on intracellular oxygen transport. J. Physiol. 578, 595–603.
Mendez, J., Keys, A., 1960. Density and composition of mammalian muscle. Metabolism 9, 184–188.
Mirceta, S., Campbell, K.L., Berenbrink, M., 2009. Molecular evolution of myoglobin in small diving mammals. Comp. Biochem. Physiol. A 153, S98–S99.
Nemeth, P.M., Lowry, O.H., 1984. Myoglobin levels in individual human skeleton-muscle fibers of different types. J. Histochem. Cytochem. 132, 1211–1216.
Noren, S.R., Williams, T.M., 2000. Body size and skeletal muscle myoglobin of cetaceans: adaptations for maximizing dive duration. Comp. Biochem. Physiol. A 126, 181–191.
Olson, J.S., 2008. Protein reviews: dioxygen binding and sensing proteins. Section 14: "From $O_2$ binding diffusion into red blood cells to ligand pathways in globins". Springer, Milan, Italy.
Ostermann, A., Waschipky, R., Parak, F.G., Nienhaus, G.U., 2000. Ligand binding and conformational motions in myoglobin. Nature 404, 205–208.
Papadopoulos, S., Endeward, V., Revesz-Walker, B., Jürgens, K.D., Gros, G., 2001. Radial and longitudinal diffusion of myoglobin in single living heart and skeletal muscle cells. Proc. Natl. Acad. Sci. U. S. A. 98, 5904–5909.
Perutz, M.F., Mathews, F.S., 1966. An X-ray study of azide methaemoglobin. J. Mol. Biol. 21, 199–207.
Ponganis, P.J., Kreutzer, U., Sailasuta, N., Knower, T., Hurd, R., Jue, T., 2002. Detection of myoglobin desaturation in Mirounga angustirostris during apnea. Am. J. Physiol. Regul. Integr. Comp. Physiol. 282, R267–R272.
Romero-Herrera, A.E., Lehmann, H., Joysey, K.A., Fridays, A.E., 1973. Molecular evolution of myoglobin and the fossil record: a phylogenetic synthesis. Nature 246, 389–395.
Rossi-Fanelli, A., Antonini, E., 1958. Studies on the oxygen and carbon monoxide equilibria of human myoglobin. Arch. Biochem. Biophys. 77, 478–492.
Scholander, P.F., 1940. Experimental investigation on the respiratory function in diving mammals and birds. Hvalrad. Skr. 22, 1–131.
Scott, E.E., Gibson, Q.H., Olson, J.S., 2001. Mapping the pathways for O2 entry into and exit from myoglobin. J. Biol. Chem. 276, 5177–5188.
Springer, B.A., Sligar, S.G., 1987. High-level expression of sperm whale myoglobin in *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. 84, 8961–8965.
Suzuki, T., Imai, K., 1998. Evolution of myoglobin CMLS. Cell. Mol. Life Sci. 54, 979–1004.
Tawara, T., 1950. On the respiratory pigments of whale(Studies on whale blood II.). Sci. Rep. Whales Res. Inst. 3, 96–101.
Terrados, N., Jansson, E., Sylven, C., Kaijser, L., 1990. Is hypoxia a stimulus for synthesis of oxidative enzymes and myoglobin? J. Appl. Physiol. 68, 2369–2372.
Tilton, R.F., Kuntz, I.D., Petsko, G.A., 1984. Cavities in proteins-structure of a metmyoglobin-xenon complex solved to 1.9 A-. Biochemistry 23, 2849–2857.
Varadarajan, R., Szabo, A., Boxer, S.G., 1985. Cloning, expression in *Escherichia coli*, and reconstitution of human myoglobin. Proc. Natl. Acad. Sci. U. S. A. 82, 5681–5684.
Wajcman, H., Kiger, L., Marden, M.C., 2009. Structure and function evolution in the superfamily of globins. C. R. Biol. 332, 273–282.
Whitehead, H., 2002. In: Perrin, W., Würsig, B., Thewissen, J. (Eds.), Encycl. Mar. Mam. Academic Press, pp. 1165–1172.
Wittenberg, J.B., 1970. Myoglobin-facilitated oxygen diffusion: role of myoglobin in oxygen entry into muscle. Physiol. Rev. 50, 559–636.
Wittenberg, B.A., Wittenberg, J.B., 1989. Transport of oxygen in muscle. Rev. Physiol. 51, 857–878.
Wittenberg, J.B., Wittenberg, B.A., 2003. Myoglobin function reassessed. J. Exp. Biol. 206, 2011–2020.
Zaia, J., Annan, R.S., Biemann, K., 1992. The correct molecular weight of myoglobin, a common calibrant for mass spectrometry. Rapid Commun. Mass Spectrom. 6, 32–36.

Comparative Biochemistry and Physiology, Part A 164 (2013) 119–128

Contents lists available at SciVerse ScienceDirect

Comparative Biochemistry and Physiology, Part A

journal homepage: www.elsevier.com/locate/cbpa

# Aerobic dive limits of seals with mutant myoglobin using combined thermochemical and physiological data

Pouria Dasmeh [a], Randall W. Davis [b], Kasper P. Kepp [a,*]

[a] Technical University of Denmark, DTU Chemistry, DK 2800 Kongens Lyngby, Denmark
[b] Texas A&M University, Department of Marine Biology, OCSB, 200 Seawolf Parkway, Galveston, TX 77553, USA

ARTICLE INFO

*Article history:*
Received 11 August 2012
Received in revised form 12 October 2012
Accepted 14 October 2012
Available online 18 October 2012

*Keywords:*
Myoglobin
Aerobic dive limit
Protein mutations
Fitness
Seals

ABSTRACT

This paper presents an integrated model of convective $O_2$-transport, aerobic dive limits (ADL), and thermochemical data for oxygen binding to mutant myoglobin (Mb), used to quantify the impact of mutations in Mb on the dive limits of Weddell seals (*Leptonychotes weddellii*). We find that wild-type Mb traits are only superior under specific behavioral and physiological conditions that critically prolong the ADL, action radius, and fitness of the seals. As an extreme example, the mutations in the conserved His-64 reduce ADL up to $14 \pm 2$ min for routine aerobic dives, whereas many other mutations are nearly neutral in terms of ADL and the inferred fitness. We also find that the cardiac system, the muscle $O_2$-store, animal behavior (i.e. pre-dive ventilation), and the oxygen binding affinity of Mb, $K_{O_2}$, have co-evolved to optimize dive duration at routine aerobic diving conditions, suggesting that such conditions are mostly selected upon in seals. The model is capable of roughly quantifying the physiological impact of single-protein mutations and thus bridges an important gap between animal physiology and molecular (protein) evolution.

© 2012 Elsevier Inc. All rights reserved.

## 1. Introduction

A good example of physiological adaptation in vertebrates is the ~10 times higher myoglobin (Mb) concentration in the skeletal muscles of some diving marine mammals compared to terrestrial mammals (Polasek et al., 2006). For aquatic, air-breathing animals that spend much of their lives submerged, oxy-Mb contributes significantly to total body oxygen stores. This adaptation, which increases aerobic breath-hold duration, is critical for the survival of diving mammals, but apparently not for terrestrial mammals where the role of Mb has been extensively studied (Weber et al., 1974; Noren and Williams, 2000; Lin et al., 2007; Endeward et al., 2010; Gros et al., 2010; Helbo and Fago, 2012), despite having similar properties such as $P_{50}$ (i.e. $P_{O_2}$ at half saturation of Mb) values and diffusion constants (Ponganis et al., 2008). Indeed, Mb knock-out mice with essentially unaffected phenotypes (Garry et al., 1998) stirred debate about other possible roles of Mb in terrestrial animals, which have later been confirmed (Cossins and Berenbrink, 2008; Hendgen-Cotta et al., 2008). These differences indicate that Mb plays a role for the fitness of diving mammals that is not shared by terrestrial animals.

Mb is a 17 kDa protein that binds $O_2$ with a 1:1 stoichiometry at its heme group, which is buried in a hydrophobic cavity within the protein (Phillips, 1980; see Fig. 1). The most important amino acid residues for this function are the distal and proximal His on each side of heme. $O_2$-binding involves forming a hydrogen bond to the $N_\varepsilon H$ of imidazole in His-64 (Perutz and Mathews, 1966). Studies of site-directed mutants have shown a 100-fold decrease in oxygen affinity by replacing His with apolar amino acids (Olson, 2008).

While most terrestrial animals have unlimited access to $O_2$, diving mammals must store oxygen in their blood, muscle and lungs to maintain aerobic metabolism while submerged. The observation that most diving mammals stay within their aerobic dive limit (ADL) during routine dives involving foraging, mating and migration is well-established (Kooyman et al., 1980; Ponganis et al., 1993; Costa et al., 2001). For Weddell seals, the ADL during routine dives has been calculated to be about 20 min but theoretically may range from ~5 to 30 min depending on the level of muscle metabolism, individual variation, and changes in convective oxygen transport associated with the dive response (Davis and Kanatous, 1999).

The role of Mb in prolonging the ADL can be understood from physiological models of the relationship between dive response and Mb concentration in seals (Wright and Davis, 2006). The oxy-Mb represents approximately one-third of the total $O_2$ store and half of the total $O_2$ used during aerobic dives (Wright and Davis, 2006) and thus contributes significantly to the ADL. Other adaptations such as increased blood volume and hematocrit and efficient modes of locomotion (stroke-and-glide swimming) (Costa et al., 1998; Williams et al., 2000; Costa et al., 2001; Williams, 2001) also enhance the ADL.

By applying thermochemical data to an integrated model of oxygen storage and transport, we recently showed that wild type (WT) Mb is more efficient in storing and transporting oxygen under severely hypoxic conditions in sperm whales (Dasmeh and Kepp, 2012). However and somewhat surprisingly, WT Mb is only more proficient

* Corresponding author. Tel.: +45 45 25 24 09.
*E-mail address:* kpj@kemi.dtu.dk (K.P. Kepp).

1095-6433/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cbpa.2012.10.010

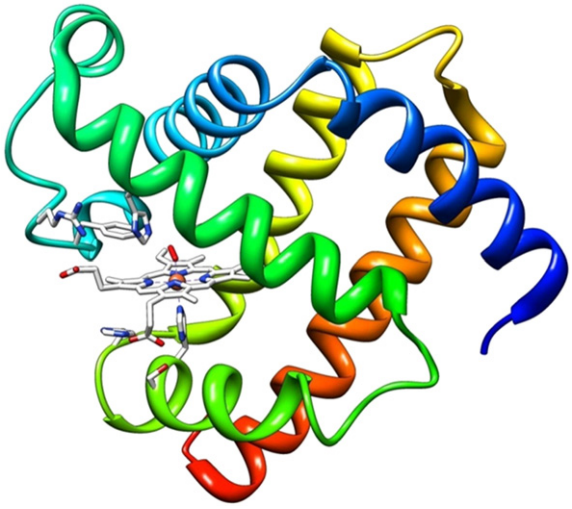


**Fig. 1.** Structure of oxy-Mb showing the heme and proximal and distal histidines. The image was produced using *Chimera*, University of California, San Francisco (Pettersen et al., 2004).

compared to low-affinity mutants under very hypoxic conditions, whereas low-affinity mutants are in fact *better* $O_2$-transporters under normoxic conditions, due to a steeper gradient of the saturation curve in this regime, affecting the diffusion rate (Dasmeh and Kepp, 2012). That WT Mb only shows its full importance at low oxygen partial pressure is consistent with conditions occurring at the end of routine dives, where hypoxia has been found necessary to efficiently release $O_2$ from Mb (Davis et al., 2004; Lin et al., 2007). This optimal use of oxy-Mb (i.e., full desaturation during dives) has also been confirmed in penguins (Williams et al., 2011).

To understand the potential evolutionary significance of Mb in marine mammals, we integrated thermochemical data of $O_2$-binding to Mb in sperm whales (Scott et al., 2001) with recently developed oxygen transport models (Dasmeh and Kepp, 2012) and ADL models for the Weddell seal (Davis and Kanatous, 1999). We found that the $O_2$-affinity of WT Mb contributes to a significantly higher ADL than e.g. His-64 impaired mutants (up to 14 min longer dives, depending on cardiac adjustments associated with the dive response and animal behavior), whereas many other mutations affect ADL only marginally. The model thus directly demonstrates and quantifies the effect of mutations in this single $O_2$-binding protein on the diving capacity, and by inference, the evolutionary fitness, of these animals.

## 2. Materials and methods

### 2.1. Experimental data

The values of $K_{O_2}$ for single point mutants of sperm whale (*Physeter macrocephalus*) Mb (Scott et al., 2001) were used to calculate the seal Mb mutant ADLs (see theory appendix and Supplementary Information, Table S1). From a Blast ClustalW alignment from the UniProt interface (UniProt Consortium, 2011) of harbor seal (*Phoca vitulina*), gray seal (*Halichoerus grypus*) and sperm whale Mb, all have identical sequence lengths (154), including 128 identical positions and 22 similar positions (Bradshaw and Gurd, 1969). Notably, all the sites studied in this work are identical in WT whales and seals.

The sperm whale $K_{O_2}$ values were first corrected from 20 °C to 37 °C by a factor of 0.2457 based on the temperature dependence of $P_{50}$ (Schenkman et al., 1997) and the oxygen solubility in the muscle tissue, $\alpha_{O_2}$ (Mahler et al., 1985). Then, using the $P_{50}$ values of 3.75 mm Hg and 3.45 mm Hg for sperm whale and Weddell seal WT Mb, all mutant seal Mb $K_{O_2}$-values were calculated by multiplying the sperm whale mutant $K_{O_2}$ at 37 °C by 1.085 (the ratio of WT seal $K_{O_2}$ to WT whale $K_{O_2}$) (see supporting information for uncorrected mutant ADLs, Figs. S1–S9). The applied, temperature-corrected $K_{O_2}$ values are found in Table S2 (Supporting Information). Given the small difference between $P_{50}$ for whale and seal, which is within the experimental uncertainty of thermochemically measured $K_{O_2}$, the use of whale mutant data is a good approximation and does not affect the significance of the conclusions, although the quantitative uncertainties in animal-specific mutant ADL are up to ~2 min (vide infra).

### 2.2. Computing the myoglobin saturation

The saturation $S$ of Mb by $O_2$ is computed with the model using the experimentally verified parameters and variables shown in Table 1. $S$ can be expressed in terms of the Hill equation (Hill, 1936):

$$S = \frac{\left(P_{O_2}\right)^n}{(P_{50})^n + \left(P_{O_2}\right)^n} \quad (1)$$

**Table 1**
List of parameters and variables used in this work.

| Parameter | Symbol | Values used in the present work |
|---|---|---|
| Cardiac output | $\dot{V}_b$ | 42.00 L $min^{-1}$ at rest |
| Brain blood flow rate | $\dot{Q}_B$ | 0.36 L $min^{-1}$ at rest |
| Heart blood flow rate | $\dot{Q}_H$ | 1.84 L $min^{-1}$ at rest |
| Skeletal muscle blood flow rate | $\dot{Q}_M$ | 7.90 L $min^{-1}$ at rest |
| Blood flow rate for splanchnic, renal, cutaneous, and other peripheral tissues | $\dot{Q}_{SRC}$ | 32.63 L $min^{-1}$ at rest |
| Brain oxygen consumption rate | $\dot{V}_{BO_2}$ | 13.3 mL $O_2$ $min^{-1}$ at rest |
| Heart oxygen consumption rate | $\dot{V}_{HO_2}$ | 112.5 mL $O_2$ $min^{-1}$ at rest |
| Skeletal muscle oxygen consumption rate | $\dot{V}_{MO_2}$ | 216.6 mL $O_2$ $min^{-1}$ at rest |
| $O_2$-consumption rate for splanchnic, renal, cutaneous, and other peripheral tissues | $\dot{V}_{SRCO_2}$ | 555 mL $O_2$ $min^{-1}$ at rest |
| Heart beat rate | $f_h$ | 51.5 beats $min^{-1}$ at rest |
| Arterial blood oxygen saturation | $S_a$ | 100–38% |
| Venous blood oxygen saturation | $S_v$ | 86–36% |
| Arterial blood $P_{O_2}$ | $P_a$ | 119–22 mm Hg |
| Venous blood $P_{O_2}$ | $P_v$ | 55–21 mm Hg |
| $P_{O_2}$ at mitochondria | $P_{mit}$ | ~0 mm Hg |
| $P_{O_2}$ at capillary | $P_c$ | 87–21 mm Hg |
| Average Mb saturation for mutants | $<S>$ | 99–27% |
| Average oxygen pressure in cell | $P_{O_2}$ | $P_{mit}<P_{O_2}<P_c$ |
| $P_{O_2}$ at which $S=0.5$ | $P_{50}$ | ~1–3 mm Hg for wild type Mb |
| Bimolecular Mb oxygenation constant | $K_{O_2}$ | ~1 $\mu M^{-1}$ |
| Oxygen solubility in the muscle tissue | $\alpha_{O_2}$ | $9.4\times10^{-7}$ mol $L^{-1}$ mm $Hg^{-1}$ |
| Mb concentration in Weddell seal muscle tissue | $C_{Mb}$ | 54 g $kg^{-1}$ muscle |

where $P_{O_2}$ is the oxygen partial pressure, $n$ is the oxygen binding cooperativity index, which is larger than unity for cooperative $O_2$-binding in multi-subunit proteins such as hemoglobin, and $P_{50}$ is the value of $P_{O_2}$ at which $S=0.5$. For Mb with $n=1$, $S$ can be derived using the bimolecular oxygenation equilibrium:

$$Mb + O_2 \Leftrightarrow \mathrm{MbO_2} \tag{2}$$

$$S = \frac{[MbO_2]}{C_{Mb}} = \frac{K_{O_2}[O_2]}{K_{O_2}[O_2]+1}. \tag{3}$$

Saturation is defined as $S=[MbO_2]/C_{Mb}$. $K_{O_2}=[MbO_2]/([O_2][Mb])$ is the bimolecular oxygenation equilibrium constant and $C_{Mb}$ is the total concentration of Mb within the cell that equals the sum of $[Mb]$ and $[MbO_2]$. Eq. (3) can be converted to Eq. (1) using $P_{50}=(K_{O_2}\alpha_{O_2})^{-1}$ and $[O_2]=\alpha_{O_2}P_{O_2}$, where $\alpha_{O_2}$ is the oxygen solubility constant within the sarcoplasm. The average saturation of Mb within the muscle cell can be calculated as:

$$\langle S\rangle = \frac{1}{P_c - P_{mit}}\int_{P_{mit}}^{P_c} S\,dP = \frac{1}{P_c - P_{mit}}\int_{P_{mit}}^{P_c} \frac{K_{O_2}\alpha_{O_2}P_{O_2}}{K_{O_2}\alpha_{O_2}P_{O_2}+1}\,dP \tag{4}$$

where $P_c$ and $P_{mit}$ are the partial pressures at the capillary and mitochondria, respectively. Assuming $P_{mit}\sim 0$, Eq. (4) gives:

$$\langle S\rangle = \frac{1}{P_c}\int_0^{P_c} \frac{K_{O_2}\alpha_{O_2}P_{O_2}}{K_{O_2}\alpha_{O_2}P_{O_2}+1}\,dP = 1 - \frac{\ln\left(K_{O_2}\alpha_{O_2}P_c+1\right)}{K_{O_2}\alpha_{O_2}P_c}. \tag{5}$$

Using $\langle S\rangle$ from Eq. (4) and the amount of available muscle $O_2$, the amount of available skeletal muscle oxygen in Weddell seal mutants can be calculated.

### 2.3. Physiological model to quantify ADL

During numerical computation, each iteration corresponded to one heart beat. During this period, convective oxygen transport ($t\times\dot{V}_b\times C_{VO_2}$) from the venous blood pool to the arterial pool was calculated, where $t$ (min) is the time of one heart beat, $\dot{V}_b$ (L min$^{-1}$) is the cardiac output, and $C_{VO_2}$ (mL $O_2$ L blood$^{-1}$) is the venous blood oxygen content. Convective transport of oxygen through the main organs and tissues ($t\times\dot{Q}_i\times C_{aO_2}$ mL $O_2$) and the amount of oxygen extracted ($t\times\dot{V}_i$ mL $O_2$) were calculated. Here, $\dot{Q}_i$ (L min$^{-1}$) and $\dot{V}_i$ (L min$^{-1}$) are the blood flow rate and oxygen consumption rate of each organ and tissue and $C_{aO_2}$ is the arterial blood oxygen content (Table 1). The extraction coefficient ($E_b$) of oxygen from the blood could not exceed 0.8 (i.e., maximum $E_b$ at critical oxygen delivery) during a single pass of the blood through an organ (Nelson et al., 1988; Samsel and Schumacker, 1994; Torrance and Wittnich, 1994). Except for the brain, where circulation is maintained at pre-dive levels, blood flow to the rest of the body decreased proportionately to $\dot{V}_b$ during a dive (Elsner et al., 1964; Blix et al., 1976).

Only oxygen stored in the blood and skeletal muscle was used to calculate whole body $O_2$ stores. $O_2$ stored as oxy-myoglobin in the heart was neglected as it constitutes less than 2% of the total muscle mass. Due to the complete functional pulmonary shunt in Weddell seal above 3–5 atm (~300–500 kPa), the lung oxygen is not available during the dive (Falke et al., 1985; Reed et al., 1994). A value of 96 L was used for the blood volume of a standard 450-kg Weddell seal (Kooyman et al., 1980) with 33% and 67% contribution of arterial and venous blood (Rowell, 1986; Hurford et al., 1996). The blood hemoglobin (Hb) concentration was 260 g/l and the oxygen-binding capacity of Hb was 1.34 mL $O_2$ per gram Hb (Kooyman et al., 1980; Qvist et al., 1986). Each liter of blood thus contained 348 mL $O_2$ (260 g/L blood × 1.34 mL $O_2$ per gram Hb).

The arterial blood was assumed to be 100% saturated with oxygen at the onset of diving because of pre-dive hyperventilation (Kooyman et al., 1980; Hurford et al., 1996). Venous blood was considered to be 86% saturated at the beginning of a dive with an initial $C_{aO_2}$ of 348 mL $O_2$ per liter blood (Kooyman et al., 1980; Ponganis et al., 1993). Arterial and venous muscle oxygen stores were calculated according to Davis et al. (Ponganis et al., 1993), assuming that 35% of the seal's body mass was skeletal muscle with $C_{Mb}=54$ g per kg muscle (Kooyman et al., 1980), and with an oxygen-binding capacity of 1.34 mL $O_2$ per gram Mb. The amount of oxygen stored in the muscle of an average adult seal was calculated as $450\times0.35\times54\times1.34\times\langle S\rangle = 11{,}397\langle S\rangle$, where $\langle S\rangle$ is the average saturation of Mb, which is calculated by Eq. (5). This factor depends not only on $P_{O_2}$, but also on the thermodynamic constant of oxygenation ($K_{O_2}$) which differs in various seal mutants.

### 2.4. Computational procedure

A modified model for the ADL of Weddell seals (Davis and Kanatous, 1999) was applied using the circulatory system in Fig. 2. Based on our previous integrated Krogh model (Dasmeh and Kepp, 2012), which integrates $O_2$ over the actual $P_{O_2}$ profile of the muscle cell to obtain the cell-integrated diffusion and storage as two distinct functions, we noted that diffusion will not affect the long time-average properties such as the total $O_2$ available for diving. Thus, whereas facilitated diffusion is important for short-term response to changes in $P_{O_2}$, for the differential ADLs of Mb mutant animals over longer time-averages (i.e. diving periods), we only need to consider the storage component of our previous model. This also implies that the choice of mitochondrial $P_{O_2}$, while obviously > 0 and dependent on mitochondrial $O_2$ consumption, will not affect the muscle integral of $O_2$-storage for normal dives where the muscles are saturated, and this is where diffusion is most important (Gros et al., 2010; Dasmeh and Kepp, 2012). To quantify this, the difference in average saturation of WT Mb and a His-64 impaired mutant is quite robust to changes in $P_{O_2}$ at the mitochondrial surface, $P_{mit}$, as shown in the supporting information.

Numerical integration procedures for evaluating $\langle S\rangle$ and ADL were used with MATLAB (Mathworks, vR2010a). As the percentage of cardiac output and muscle metabolic rate were varied in the simulation, the whole body $O_2$ decreased in each iteration. New values for $C_{aO_2}$ and $C_{VO_2}$ were calculated as $C_{aO_2}=348\times S_a$ and $C_{VO_2}=300\times S_v$, where $S_a$ and $S_v$ are the arterial and the venous blood saturation, calculated as the ratio of available $O_2$ to the initial total amounts in the arterial and venous blood. We use the reported $P_{50}=26.9$ mm Hg and Hill coefficient $n=2.39$ for an adult Weddell seal to calculate $P_a$ and $P_v$. Within the muscle, the value of $P_{O_2}$ at the mitochondria ($P_{mit}$) was assumed to be zero, and the value of $P_{O_2}$ at the capillaries ($P_c$) was calculated as the average of $P_a$ and $P_v$.

The simulated dive was terminated (i.e., the ADL was reached) when any organ or tissue did not receive sufficient oxygen through convective oxygen transport or $MbO_2$ to maintain aerobic metabolism in any organ or tissue or when the $P_a$ decreased below 22 mm Hg (Hurford et al., 1996).

To evaluate the effect on ADL of oxygen loading at the surface (i.e. forced dives occurring before full $O_2$-reloading due to e.g. sudden threats), we evaluated the ADL not only at the spectrum of cardiac and muscle outputs corresponding to variations in circulatory and muscle work, but also at different starting oxygen partial pressures as reflected by changing $P_c$ and re-computing impaired Mb saturations and new ADLs for both mutants and WT.

The uncertainties of the experimental oxygenation constants correspond to uncertainties of computed ADLs of ca. ±1 min. In addition, uncertainties in the initial parameters and the choice of $P_{O_2}$ in the mitochondria and capillary during various scenarios affected the saturation curve and ADL. A sensitivity analysis using ±10% change in the initial parameters showed that Mb concentration, muscle $O_2$

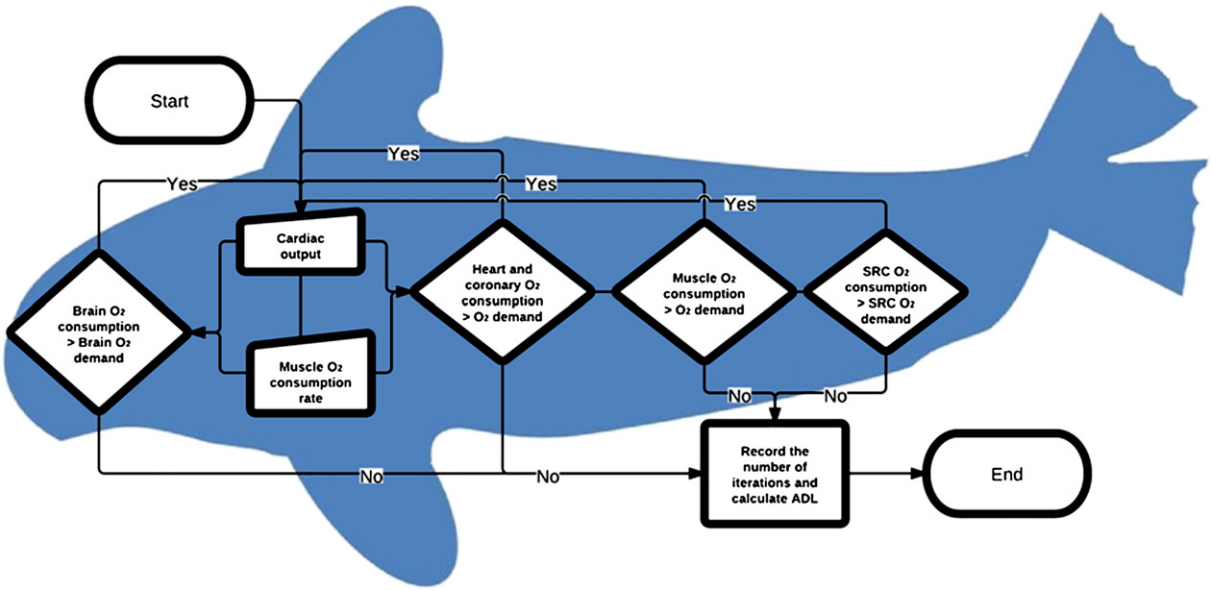



**Fig. 2.** The blood circulatory model used in this work simplified into two common venous 112 and arterial pools.

consumption rate, and blood volume were the main sources of potential errors, with ~15%, ~10%, and ~10% effect on ADL, respectively (see ADL data with error bars in Supplementary Information Figs. S23–S24). The effect of blood volume scaling linearly with mass on ADL has also been suggested by Costa et al. (1998). Uncertainties in other parameters had minor effects on the ADL. The model provided ADLs with uncertainties of approximately 2 min for the longer dives. Thus, our general conclusions are not affected by any reasonable change in input parameters arising from individual variation or inherent uncertainties in experimental data.

## 3. Results and discussion

### 3.1. The diving Weddell seal: effect of the oxygenation constant on ADL

To estimate the ADL of a diving seal as a function of Mb binding affinity, a range of $\dot{V}_{MO_2}$ and $\dot{V}_b$ were studied to reflect individual variation and to quantify different work rates. The routine dive conditions were estimated to be $\dot{V}_{MO_2}$ of 5 times resting level and a $\dot{V}_b = 0.27\,\dot{V}_b$ (rest), which is the muscle metabolic rate and the cardiac that maximizes the ADL in the model. As a critical test of the model, these values corresponded well with the routine range of muscle metabolic rate and cardiac output in Weddell seals (Davis and Kanatous, 1999; Davis et al., 2004).

Fig. 3 shows the effect of $K_{O_2}$ and $\dot{V}_{MO_2}$ on ADL at various multiples of resting $\dot{V}_{MO_2}$ for the Weddell seal at 27% of pre-dive, resting cardiac output (color plot), and pre-dive, resting cardiac output (grayscale plot). $K_{O_2}$ increased in the plot from 0.01 to 1.0 $\mu M^{-1}$, i.e., by two orders of magnitudes, which approximately corresponds to the decrease in oxygen affinity of the most impaired Mb mutants studied (e.g. H64V). Thus, Fig. 3 quantifies the impact of $K_{O_2}$ on ADL at different muscle metabolic rates, both at resting and at optimal routine dive cardiac output.

Both these diagrams display a distinct, sudden change in ADL but at different $\dot{V}_{MO_2}$ values. When cardiac output is 27% of resting (color

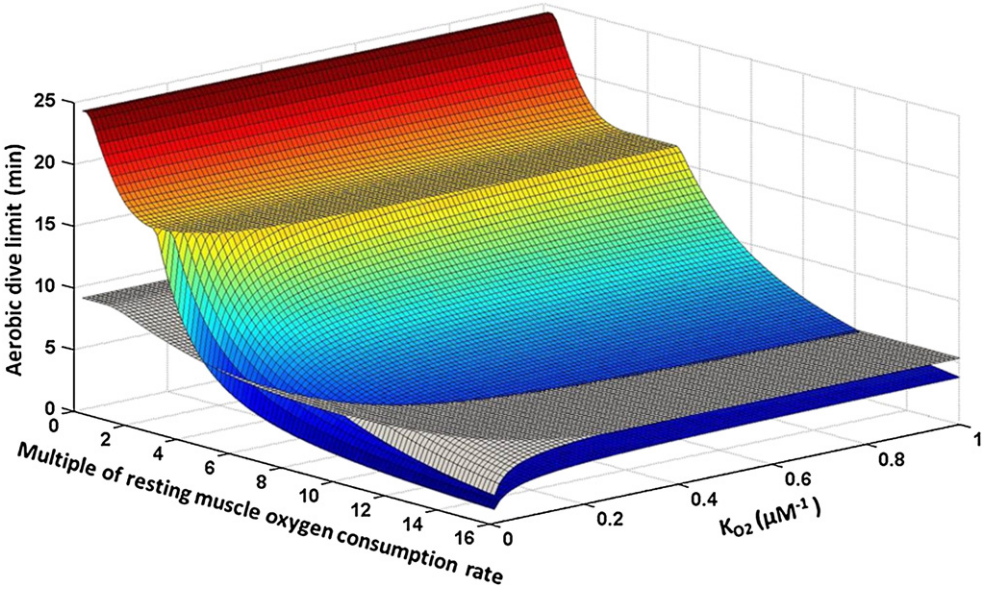



**Fig. 3.** The ADL (minutes) as a function of multiples of the resting muscle oxygen consumption, $\dot{V}_{MO_2}$, and oxygen binding affinity of Mb, $K_{O_2}$, using a cardiac output $\dot{V}_b = 0.27\,\dot{V}_b$ (*rest*) (color plot) or the resting cardiac output $\dot{V}_b = \dot{V}_b$ (*rest*) (grayscale). The effect of mutations can be assessed in the figure by their oxygen binding affinity, $K_{O_2}$, and thus the resulting ADL.

plot), this change occurs at five times resting $\dot{V}_{MO_2}$, i.e. under optimal routine dive conditions. Under these conditions, it can be seen that muscle $O_2$ becomes a limiting factor for the ADL. Prior to this point, the dive is terminated when $P_a$<22 mm Hg, illustrating that oxygen supply is critically reduced and the ADL is exceeded by definition in this model. Above $\dot{V}_{MO_2} = 5\dot{V}_{MO_2}(rest)$, all dives are eventually terminated due to the consumption of muscle $O_2$. The numerical data for different dive types are presented in Table 2.

At $\dot{V}_b = \dot{V}_b(rest)$, the gray surface in Fig. 3, the transition in the limiting factor also depends on the Mb oxygen affinity, but since the cardiac output is almost four times higher, the muscle metabolism only becomes rate-limiting during extreme muscle work, e.g. >10 times resting muscle $O_2$ consumption rate, which is most likely beyond the aerobic capacity of seals, and only for Mbs with $K_{O_2}$<0.3 $\mu M^{-1}$. In higher-affinity Mbs, the transition never occurs, with the cardiac system always being a limiting factor to the ADL. WT seal Mb has $K_{O_2}$~0.3 $\mu M^{-1}$. Also, Fig. 3 reveals that the effectiveness of Mb as an $O_2$-storage protein depends substantially on the physiological conditions.

To better compare the effectiveness of low vs. high $O_2$-affinity Mbs (either from different organisms or e.g. for two mutants) at various physiological conditions, we computed the change in ADL when $K_{O_2}$ changes from 0.01 to 1 $\mu M^{-1}$ at different combinations of $\dot{V}_{MO_2}$ and $\dot{V}_b$, as shown in Fig. 4. The maximum effectiveness of a high-affinity Mb with $K_{O_2}$ ~1 $\mu M^{-1}$ occurs at $\dot{V}_{MO_2}$~3 $\dot{V}_{MO_2}(rest)$ and $\dot{V}_b$~0.15 $\dot{V}_b(rest)$. Under these conditions, and with fully oxygenated blood at the beginning of a dive, the most effective Mbs increase the ADL up to 14 min compared to impaired mutants with $K_{O_2}$~0.01 $\mu M^{-1}$. The single hydrogen bond between His-64 and $O_2$ now reveals its full physiological importance, since a 14-min reduction in ADL would have a substantial effect on foraging, mating, and predator evading abilities.

Fig. 4 can be thought of as a differential fitness landscape with only one selected phenotype (ADL) for the evolution of Mb applicable to diving mammals (the qualitative properties of the landscape will be the same for other diving mammals, but the values will change, vide infra). The vast majority of conditions in Fig. 4 can be seen to not effectively distinguish low-affinity Mb mutants from high-affinity mutants and WT. However, under very specific conditions, most pronounced at $\dot{V}_{MO_2}$~3–4 $\dot{V}_{MO_2}(rest)$ and $\dot{V}_b$~0.15 $\dot{V}_b(rest)$, high affinity Mbs are maximally proficient, compared to alternatives. Importantly, these conditions of optimal differential ADL are very similar to those prevailing under routine dives ($\dot{V}_{MO_2}$~5$\dot{V}_{MO_2}(rest)$). We might identify these conditions as those upon which the *selection pressure* operates mostly, because these physiological and environmental conditions optimize WT fitness relative to impaired mutants. Although the exact values of the parameters will change in various diving mammals, this qualitative observation is general and confirms previous explanations of why $C_{Mb}$ is so much higher in diving marine mammals (Kooyman and Ponganis, 1998) and how cardiac and muscle output has co-evolved not only with routine diving behavior (Davis et al., 2004) but also directly with the oxygen-storing capacity of WT Mb, quantified by experimental $K_{O_2}$.

### 3.2. Pre-dive oxygenation, hyperventilation, and pH effects

Until now, we have discussed the effect of impaired Mb mutants on ADL under optimal and routine dive conditions that include time to hyperventilate prior to the dive. However, this relies on the assumption that the selection pressure described above occurs only under such hyperventilated, routine dive conditions, which constitute the majority of dives. However, the model can also investigate the impact of animal behavior due to restricted pre-dive hyperventilation from e.g. sudden submergence on the ADL of both WT and mutant seals, to investigate the role of such behavior in the evolution of Mb.

Fig. 5 shows the effect of changes in $K_{O_2}$ on the ADL under somewhat less oxygenated starting conditions ($P_a$ being half of normal) obtained from direct computation of the initial saturation $S$ corresponding to this lower pre-dive $P_a$. From these results, it can be seen that when animal behavior is not invoked to maximize ADL, the difference between WT and impaired mutants becomes less pronounced. A maximum effect of ~10 min is observed at half fully hyperventilated $P_a$. This number both precludes hyperventilation and full oxygenation, and is not characteristic in itself, only as a measure of the sensitivity of relative ADL to different types of dives. The calculation however shows that not only does pre-dive hyperventilation increase the ADL of seals, it also contributes to the superiority of the WT vs. mutants, again indicating how natural selection causes co-evolution of behavior, protein structure, and muscular and cardiac physiology together. The detailed results for all mutants with and without hyperventilation and at various conditions are found in Figs. S10–S15 of the Supporting Information.

The absolute values of the WT ADL are shown in Fig. 6 (with hyperventilation) and Fig. 7 (without). From these two figures, the effect of hyperventilation can provide ~14±2 min more dive time under routine diving conditions for the WT seal. The error estimate comes from the sum of uncertainties in the $K_{O_2}$ data propagating to ADL and the use of the ADL model with hyperventilation and non-hyperventilation oxygen partial pressures for determining pre-dive saturations. Fig. 6 shows that while it is theoretically possible to have ADLs as high as ~30 min, this will only occur if the animal is resting while submerged. At the optimal muscle metabolic rate of routine dives (~5 times resting $\dot{V}_{MO_2}$), ADL of WT is ~17 min in this model. Also worth noting is the

**Table 2**
Dive termination conditions under different combinations of cardiac output and muscle metabolic rate.

| $K_{O_2}$ | % resting cardiac output | Multiple of resting muscle metabolic rate | ADL (min) | Dive terminator factor | End dive $P_a$ (mm Hg) | Total muscle oxygen (mL) consumed |
|---|---|---|---|---|---|---|
| 0.01 | 27 | 1 | 22.7 | Arterial $O_2$ | 22 | 8329 |
| | | 5 | 6.6 | Muscle $O_2$ | 40 | 11397 |
| | | 15 | 1.0 | Muscle $O_2$ | 73 | 11397 |
| | 100 | 1 | 9.5 | Arterial $O_2$ | 22 | 8329 |
| | | 5 | 6.9 | Arterial $O_2$ | 22 | 8332 |
| | | 15 | 2.8 | Muscle $O_2$ | 33 | 11397 |
| 0.1 | 27 | 1 | 22.7 | Arterial $O_2$ | 22 | 3090 |
| | | 5 | 14.7 | Muscle $O_2$ | 25 | 11397 |
| | | 15 | 3.2 | Muscle $O_2$ | 53 | 11397 |
| | 100 | 1 | 9.5 | Arterial $O_2$ | 22 | 3090 |
| | | 5 | 6.9 | Arterial $O_2$ | 22 | 3093 |
| | | 15 | 5.6 | Arterial $O_2$ | 22 | 11394 |
| 1 | 27 | 1 | 22.7 | Arterial $O_2$ | 22 | 615 |
| | | 5 | 17.3 | Arterial $O_2$ | 22 | 10922 |
| | | 15 | 4.2 | Muscle $O_2$ | 48 | 11397 |
| | 100 | 1 | 9.5 | Arterial $O_2$ | 22 | 615 |
| | | 5 | 6.9 | Arterial $O_2$ | 22 | 619 |
| | | 15 | 5.6 | Arterial $O_2$ | 22 | 8921 |

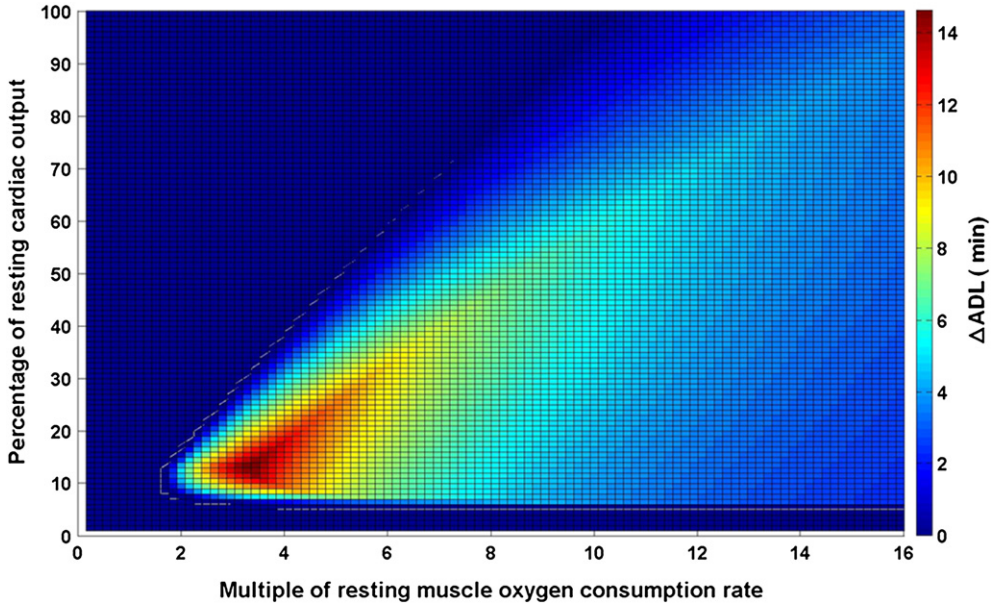



**Fig. 4.** Plot of the change in ADL (in minutes, color bar) going from a low-affinity to high-affinity mutant ($K_{O_2} = 0.01$ to 1 $\mu M^{-1}$), over possible combinations of muscle oxygen consumption, $\dot{V}_{MO_2}$ and cardiac output, $\dot{V}_b$, using temperature-corrected data. The maximum occurs at $\dot{V}_{MO_2} \sim 3\,\dot{V}_{MO_2}$ (*rest*) and $\dot{V}_b \sim 0.15\,\dot{V}_b$ (*rest*). Arterial and venous blood oxygen pressures are $P_a = 119$ mm Hg and $P_v = 55$ mm Hg respectively, corresponding to fully oxygenated, hyperventilated starting dive conditions. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

distinctive discontinuity running along the diagonal of the ADL landscapes in Fig. 6 and 7, which represents the co-optimum of cardiac and muscular outputs to maximize ADL under given dive conditions.

Until now, we have neglected pH-effects on the calculated ADL. Routine dives are aerobic and terminate in ~16–17 min for this size of seal. Under such conditions, nearly neutral pH is maintained (Kooyman et al., 1983). However, for an adult 450-kg Weddell seal, pH typically drops from ~7.4 to ~6.8 in dives exceeding 20 min, corresponding to ~2.7% of all dives (Kooyman et al., 1980). For young, smaller seals, the lactate builds up earlier (Kooyman et al., 1983). While this work concerns the vast majority of routine aerobic dives likely to determine fitness, it remains relevant to quantify possible pH-effects, in particular for Hb that has a larger pH-effect (the Redox Bohr Effect). To calculate the largest potential pH-effect on $P_{50}$, we assumed a pH-drop corresponding to that occurring after 20 min as a worst case scenario. Using the pH-dependence for horse Mb (Schenkman et al., 1997) corrected to seal Mb as $\frac{P_{50}(seal)}{P_{50}(horse)} = 1.55$, $P_{50}$ changes roughly from ~3.6 to ~3.8 mm Hg upon such a pH drop. For Hb, the change is much more substantial; approximately from ~26.9 to ~50.1 mm Hg at 37 °C, based on the relationship between $P_{50}$ and pH for Weddell seal Hb (Qvist et al., 1986).

Due to the reduced saturation and oxygen pools of Mb and Hb, ADL is reduced from 16.70 ± 2 to 16.65 ± 2 min due to Mb, which is negligible. More importantly, the decline in blood pH increases $P_{50}$ of Hb from 26.9 to 50.1 mm Hg. We calculated the impact of such a change occurring at 5 min, 10 min, and 15 min after diving.

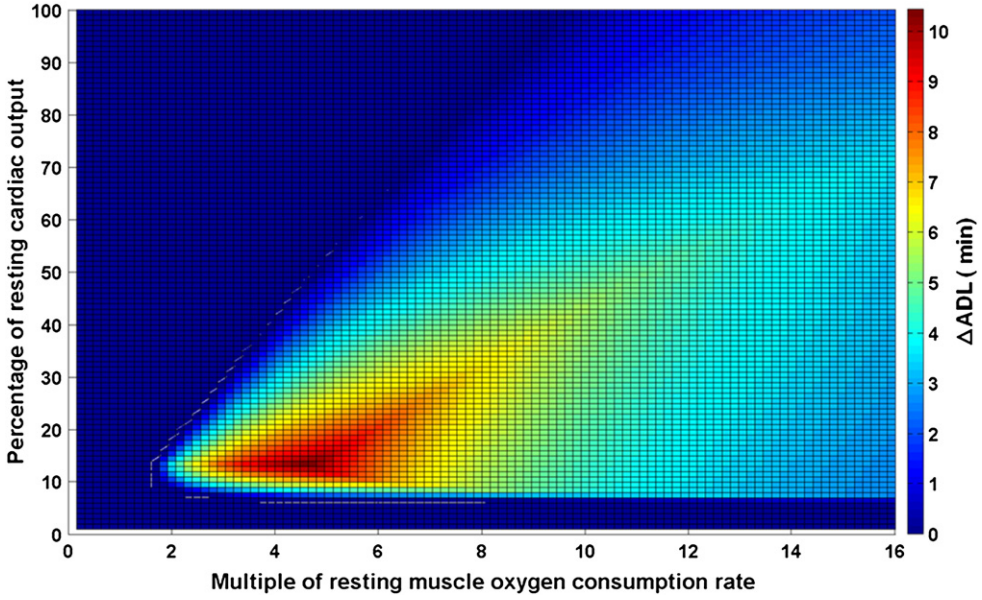



**Fig. 5.** Plot of change in ADL (in minutes) going from oxygen binding affinity $K_{O_2} = 0.01$ to 1 $\mu M^{-1}$, over possible combinations of muscle oxygen consumption, $\dot{V}_{MO_2}$ and cardiac output, $\dot{V}_b$, using temperature-corrected seal data. The maximum occurs at $\dot{V}_{MO_2} \sim 3\,\dot{V}_{MO_2}$ (*rest*) and $\dot{V}_b \sim 0.15\,\dot{V}_b$ (*rest*). The arterial and venous blood oxygen pressures are $P_a = 59.5$ mm Hg and $P_v = 27.5$ mm Hg respectively, corresponding to less oxygenated, i.e. not hyperventilated starting dive conditions.

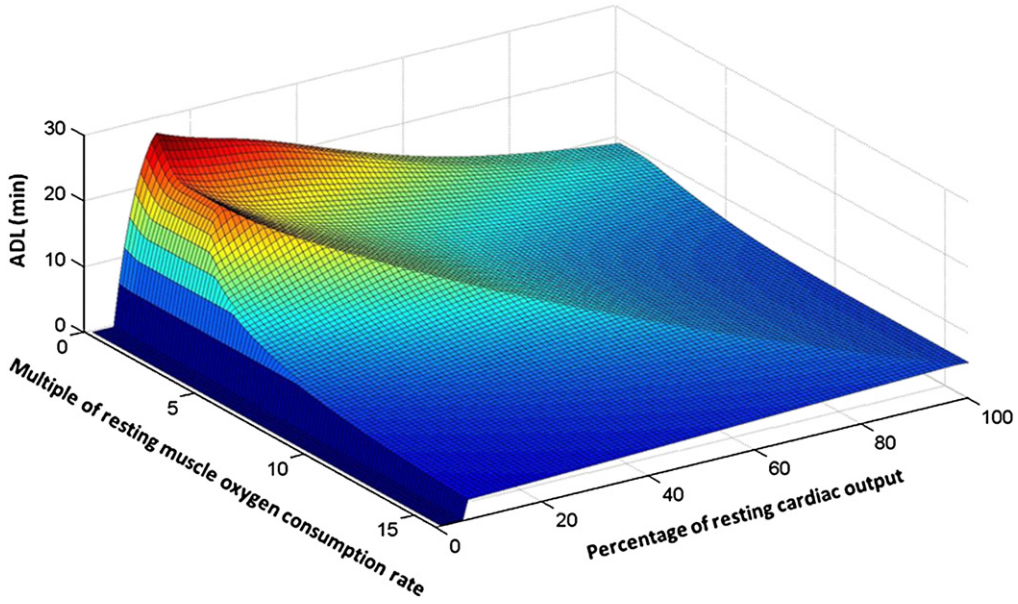



**Fig. 6.** Plot of ADL for WT seal Mb, over possible combinations of muscle oxygen consumption, $\dot{V}_{MO_2}$ and cardiac output, $\dot{V}_b$ after pre-dive hyperventilation ($P_a = 119$ mm Hg and $P_v = 55$ mm Hg). The maximum ADL for routine dives ($\dot{V}_{MO_2} \sim 5\dot{V}_{MO_2}(rest)$) occurs at $\dot{V}_b \sim 0.27\dot{V}_b(rest)$.

Importantly, even though the pH-effect of Hb is large, the ADL is unaltered, since the limiting factor for such routine dives remains muscle-$O_2$ rather than blood-$O_2$ content. For longer dives, e.g. due to predator evasion or uncharacteristically long hunts, the blood pool could be dive-limiting and hypoxia would lead to a pH-drop of the magnitude above, potentially affecting ADL. More seriously, dives initiated before restoration (i.e. before ventilation and normalization of blood pH) would for the same reasons terminate quickly, as both Mb and Hb saturation would be greatly impaired.

### 3.3. The ADL for specific, characterized Mb mutant seals

Using the model described above, it is possible to quantify the effect of single point mutations in Mb on the ADL for the Weddell seal, although the conclusions are valid for other animals with proper adjustment of physiological and thermochemical parameters. To calculate the average Mb saturation and ADL, we used a routine aerobic diving $\dot{V}_{MO_2}$ of five times resting and $\dot{V}_b \sim 0.27$ times that at rest.

ADLs for specific mutants are shown in Fig. 8. The error bars represent the uncertainty in ADL caused by 20% experimental errors in $K_{O_2}$-values. The total uncertainty of the ADL is approximately double this magnitude due to approximations in the ADL model and the choice of oxygen pressures $P_c$ and $P_{mit}$. It is important to note that the work rests on the description of one typical individual, a healthy 450-kg adult seal, as representative of the species: Other individuals with different sizes, ages, fat-to-muscle ratios, individual metabolic rates, etc. will display different absolute ADLs, but the trends in ADLs due to mutations will be similar. An example is given for a smaller 200-kg seal in the Supporting Information (Table S3) with shorter ADL (12.4 min vs. 16.7 min for the 450-kg seal) but correspondingly more deleterious effects of mutation H64V, reducing ADL to merely 2 min. Regional blood flow rates were assumed to scale linearly with

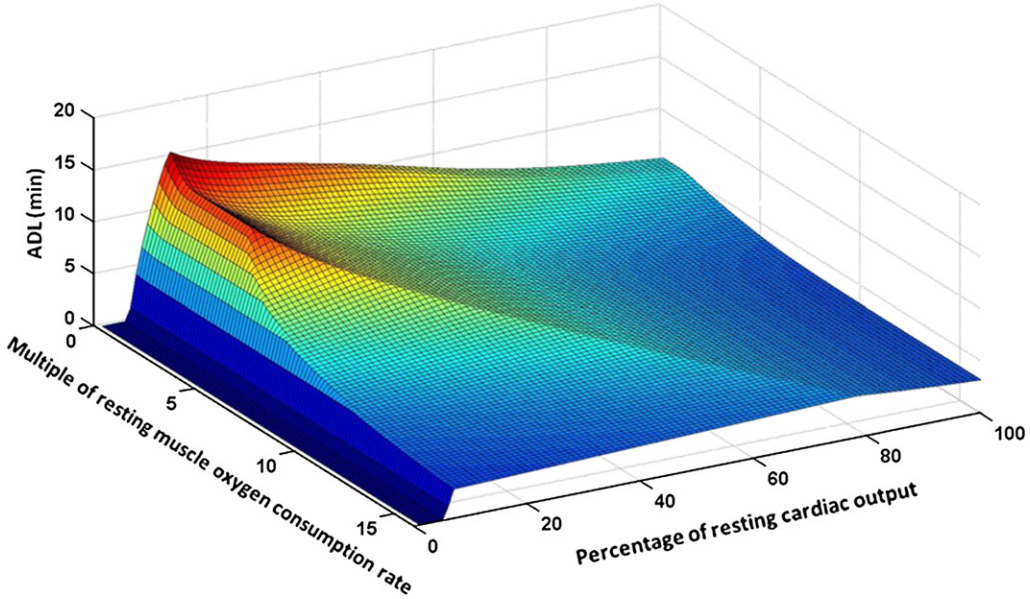



**Fig. 7.** Plot of ADL for WT seal Mb, over possible combinations of muscle oxygen consumption, $\dot{V}_{MO_2}$ and cardiac output, $\dot{V}_b$ without pre-dive hyperventilation ($P_a = 59.5$ mm Hg and $P_v = 27.5$ mm Hg). The maximum ADL for routine dives ($\dot{V}_{MO_2} \sim 5\dot{V}_{MO_2}(rest)$) occurs at $\dot{V}_b \sim 0.27\dot{V}_b(rest)$.

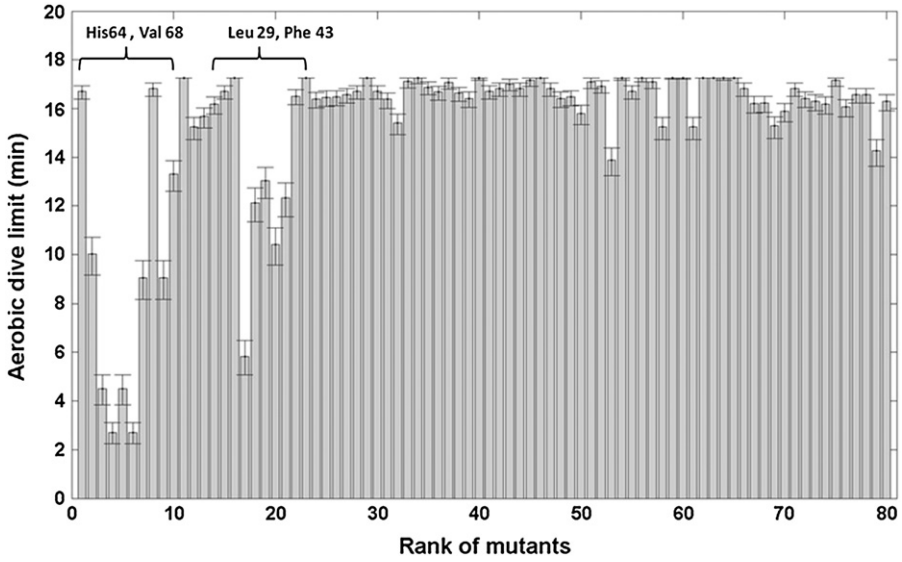



**Fig. 8.** The calculated ADL (min) of the simulated dive of a Weddell seal with $\dot{V}_{MO_2} \sim 5\dot{V}_{MO_2}$ (*rest*) and $\dot{V}_b \sim 0.27\dot{V}_b$ (*rest*) at 37° C for different Mb mutants, using converted seal mutant $K_{O_2}$ data. Error bars are calculated from 20% experimental error in $K_{O_2}$ values.

the mass ratio, and oxygen consumption rates were scaled with (mass ratio)$^{-0.25}$ (Davis and Kanatous, 1999).

To further substantiate the generality of such mutation effects, we also evaluated the ADL of a ~175 kg male and a ~87 kg female California sea lion with $C_{Mb}$ ~35 and ~50 g per kg muscle and blood volumes of ~20 and ~10 l, respectively (Weise and Costa, 2007). In both cases, ADL was reduced to less than 1 min by the deleterious H64V mutation in Mb (see the Supplementary Information, Table S3 and Figs. S25 and S26), making the H64 residue vital also to these diving mammals, which is the reason why it is conserved across the species.

It can be seen from Fig. 8 that the most deleterious single-point mutations in residues 64 or 29 decrease ADL from 17 min of the WT to e.g. ~3–4 min under routine diving conditions. With a routine swimming speed of 1.2 m s$^{-1}$ (Davis et al., 2003), during a 4-min dive, the mutant seal would be able to swim to a depth of about 144 m, which is shallower than the routine depth of its primary prey (*Pleuragramma antarcticum*) at depths greater than 160 m (Fuiman et al., 2007). In addition, it would have no time to pursue prey even if they were encountered. Such a reduction in ADL would therefore greatly reduce the foraging ability of the seal probably preventing it from reaching reproductive age and hence, it would be purged from the population.

The most impairing mutations occurring in sites 64, 68, 43, and 29 cause up to five-fold reductions in available muscle $O_2$ and ADL. Mutations with such substantial effects on ADL will most likely be quickly purged from the population, as explained above. On the other hand, the many mutations with minor effects on ADL (i.e. nearly neutral mutations) are expected to exist in the wild. While mutant prevalence data for mammals are generally unavailable, this expectation can be partly confirmed by the observation of natural variations in the nearly neutral sites 8, 28, 66, and 144 of WT sequences of various whales, seals and sea-lions (see Tables S4 and S5 in the supporting information). Future sequencing of many individuals of seals would be able to cast further light on nearly-neutral mutations prevailing because they do not impair ADL.

For high-affinity mutants with larger $K_{O_2}$ than WT (e.g. V68L and L29F), muscle $O_2$ is not a limiting factor for ADL under routine dive

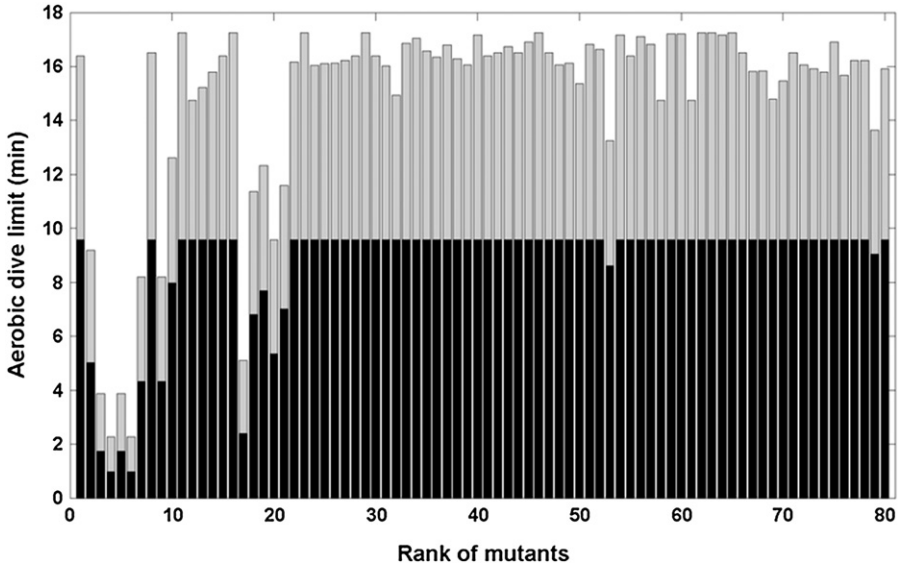



**Fig. 9.** The calculated ADL (min) of the simulated dive of a Weddell seal with $\dot{V}_{MO_2} \sim 5\dot{V}_{MO_2}$ (*rest*) and $\dot{V}_b \sim 0.27\dot{V}_b$ (*rest*) at 37° C for different Mb mutants, using converted seal mutant $K_{O_2}$ data, at $P_a = 119$ mm Hg and $P_v = 55$ mm Hg (gray bars) and $P_a = 59.5$ mm Hg and $P_v = 27.5$ mm Hg (black bars).

conditions, whereas cardiac output becomes a limiting factor, and thus ADL is not significantly higher than for the WT. This important result shows that the cardiac, muscular, and behavioral adaptations of seals are correlated with the WT proficiency, suggesting that all these phenotypes have co-evolved.

Fig. 9 shows the effect of pre-dive hyperventilation on the various mutants: the gray bars are for mutant seals at routine dive conditions with hyperventilation as in Fig. 8, whereas black bars illustrate ADLs without pre-dive hyperventilation. Without pre-dive hyperventilation, ADL is strictly limited by an upper bound of ~10 min due to cardiac output being rate-limiting when circulatory $O_2$ stores are not as high as upon hyperventilation.

## 4. Conclusions

The results discussed in this paper have implications for the interplay between the molecular evolution of proteins, in this case Mb, and the adaptations of the organism that relate to the proficiency of this and other proteins. First, the WT is most proficient under the routine dive conditions as seen in the differential fitness landscape, Fig. 4, suggesting co-evolution of animal behavior, cardiac adaptations, and molecular structure: high-affinity mutants ($K_{O_2}$~0.3 $\mu M^{-1}$) do not have significantly longer ADLs than WT at routine dive conditions because cardiac output becomes a limiting factor, showing that the WT is fit only under those conditions and adaptations. Further calculations of fitness landscapes are given in the Supporting Information under various stated conditions (Figs. S16–S22).

We have shown examples of deleterious mutants that have ADLs of only ~3 min whereas the WT has ADL ~17 min (Fig. 8) under typical dive conditions. This can be achieved by mutating one residue, the His-64 that changes $K_{O_2}$ by 100-fold. The effect of single mutations on mammalian fitness is not commonly explored, although fundamental to the concept of Darwinian molecular evolution, and it is our hope that the present models and computations may be helpful in understanding the direct correlation between protein structure, animal behavior and physiological adaptations in a common framework of evolution.

While the most deleterious mutations may be purged already in the first generation, the organism may protect itself against less severe mutations by a number of molecular and functional adaptations. For example, mice without Mb express more hypoxia-inducible factor and stress proteins, display elevated nitric oxide metabolism, and shift from fatty acid to glucose metabolism, probably to rescue the exercise capacity of the animal (Grange et al., 2001; Flogel et al., 2005). It remains open whether similar compensation mechanisms occur for marine mammals with slightly impaired mutant Mb so as to improve their fitness.

In this study, we have based most calculations on an individual, adult Weddell seal of 450 kg, and as such, the data only apply directly to this seal. However, the qualitative conclusions and all the semi-quantitative trends of the mutational impacts on ADL remain valid for other diving mammals, although the input parameters and exact values will of course change. Impairing $K_{O_2}$ in Mb will affect ADL in all these mammals, but how much depends on the physiological and compensatory mechanisms of each organism. We have shown here that the observed effects of single-mutations in Mb reducing ADL, with either deleterious or marginal effects on fitness, will be qualitatively and semi-quantitatively similar for other individuals and for sea lions. Thus, the prevalence of H64 and the natural variation in some sites that do not substantially change ADL can be explained from the present model.

A question then emerges whether the selection pressure on diving mammals occurs mostly on some age-range of the population, e.g. young seals with shorter ADLs that have just begun foraging on their own. Natural selection occurs on the whole population and is ideally an integration of fitness over all individuals of the population, and this integral will change depending on the properties of the population. However, the result that some mutations in just one protein dramatically impair ADL while other mutations do not, and the observed co-evolution and adaptation between optimal protein and diving conditions, appear to be general among diving mammals.

## Acknowledgments

This work was made possible by a grant from the Danish National Science Research Council, Project Case 272-08-0041.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.cbpa.2012.10.010.

## References

Blix, A.S., Kjekshus, J.K., Enge, I., Bergan, A., 1976. Myocardial blood flow in the diving seal. Acta Physiol. Scand. 96, 277–280.
Bradshaw, R.A., Gurd, F.R.N., 1969. Comparison of myoglobins from harbor seal, porpoise, and sperm whale. (V.) J. Biol. Chem. 244, 2167–2181.
Cossins, A., Berenbrink, M., 2008. Physiology: myoglobin's new clothes. Nature 454, 416–417.
Costa, D.P., Gales, N.J., Crocker, D.E., 1998. Blood volume and diving ability of the New Zealand seal lion, *Phocarctos hookeri*. Physiol. Zool. 71, 208–213.
Costa, D.P., Gales, N.J., Goebel, M.E., 2001. Aerobic dive limit: how often does it occur in nature? Comp. Biochem. Physiol. A 129, 771–783.
Dasmeh, P., Kepp, K.P., 2012. Bridging the gap between chemistry, physiology, and evolution: quantifying the functionality of sperm whale myoglobin mutants. Comp. Biochem. Physiol. A 161, 9–17.
Davis, R.W., Kanatous, S.B., 1999. Convective oxygen transport and tissue oxygen consumption in Weddell seals during aerobic dives. J. Exp. Biol. 202, 1091–1113.
Davis, R.W., Fuiman, L.A., Williams, T.M., Horning, M., Hagey, W., 2003. Classification of Weddell seal dives based on three-dimensional movements and video recorded observations. Mar. Ecol. Progr. Ser. 264, 109–122.
Davis, R.W., Polasek, L., Watson, R., Fuson, A., Williams, T.M., Kanatous, S.B., 2004. The diving paradox: new insights into the role of the dive response in air-breathing vertebrates. Comp. Biochem. Physiol. A 138, 263–268.
Elsner, R.W., Franklin, D.L., VanCritters, R.L., 1964. Cardiac output during diving in an unrestrained sea lion. Nature 202, 809–810.
Endeward, V., Gros, G., Jürgens, K.D., 2010. Significance of myoglobin as an oxygen store and oxygen transporter in the intermittently perfused human heart: a model study. Cardiovasc. Res. 87, 22–29.
Falke, K.J., Hill, R.D., Qvist, J., Schneider, R.C., Guppy, M., Liggins, G.C., Hochachka, P.W., Elliot, R.E., Zapol, W.M., 1985. Seal lungs collapse during free diving: evidence from arterial nitrogen tensions. Science 229, 556–558.
Flogel, U., Laussmann, T., Godecke, A., Abanador, N., Schafers, M., Fingas, C.D., Metzger, S., Levkau, B., Jacoby, C., Schrader, J., 2005. Lack of myoglobin causes a switch in cardiac substrate selection. Circ. Res. 96, 68–75.
Fuiman, L.A., Kiersten, M.M., Williams, T.M., Davis, R.W., 2007. Structure of foraging dives in the Antarctic fast-ice environment. Deep-Sea Res. II 54, 270–289.
Garry, D.J., Ordway, G.A., Lorenz, J.N., Radford, N.B., Chin, E.R., Grange, R.W., Bassel-Duby, R., Williams, R.S., 1998. Mice without myoglobin. Nature 395, 905–908.
Grange, R.W., Meeson, A., Chin, E., Lau, K.S., Stull, J.T., Shelton, J.M., Williams, R.S., Garry, D.J., 2001. Functional and molecular adaptations in skeletal muscle of myoglobin-mutant mice. Am. J. Physiol. Cell Physiol. 281, C1487–C1494.
Gros, G., Wittenberg, B.A., Jue, T., 2010. Myoglobin's old and new clothes: from molecular structure to function in living cells. J. Exp. Biol. 213, 2713–2725.
Helbo, S., Fago, A., 2012. Functional properties of myoglobins from five whale species with different diving capacities. J. Exp. Biol. http://dx.doi.org/10.1242/jeb.073726.
Hendgen-Cotta, U.B., Merx, M.W., Shiva, S., Schmitz, J., Becher, S., Klare, J.P., Steinhoff, H.J., Goedecke, A., Schrader, J., Gladwin, M.T., Kelm, M., Rassaf, T., 2008. Nitrite reductase activity of myoglobin regulates respiration and cellular viability in myocardial ischemia-reperfusion injury. Proc. Natl. Acad. Sci. U. S. A. 105, 10256–10261.
Hill, R., 1936. Oxygen dissociation curves of muscle haemoglobin. Proc. R. Soc. Lond. B 120, 472–483.
Hurford, W.E., Hochachka, P.W., Schneider, R.C., Guyton, G.P., Stanek, K.S., Zapol, D.G., Liggins, G.C., Zapol, W.M., 1996. Splenic contraction, catecholamine release and blood volume redistribution during diving in the Weddell seal. J. Appl. Physiol. 80, 298–306.
Kooyman, G.L., Ponganis, P.J., 1998. The physiological basis of diving to depth: birds and mammals. Annu. Rev. Physiol. 60, 19–32.
Kooyman, G.L., Wahrenbrock, E.A., Castellini, M.A., Davis, R.W., Sinnett, E.E., 1980. Aerobic and anaerobic metabolism during voluntary diving in Weddell seals: evidence of preferred pathways from blood chemistry and behavior. J. Comp. Physiol. B 138, 335–346.
Kooyman, G.L., Castellini, M.A., Davis, R.W., Maue, R.A., 1983. Aerobic dive limits in immature Weddell seals. J. Comp. Physiol. 151, 171–174.
Lin, P.C., Kreutzer, U., Jue, T., 2007. Myoglobin translational diffusion in rat myocardium and its implication on intracellular oxygen transport. J. Physiol. 578, 595–603.

Mahler, M., Louy, C., Homsher, E., Peskoff, A., 1985. Reappraisal of diffusion, solubility, and consumption of oxygen in frog skeletal muscle, with applications to muscle energy balance. J. Gen. Physiol. 86, 105–134.
Nelson, D.P., Samsel, R.W., Wood, L.D., Schumacker, P.T., 1988. Pathological supply dependence of systemic and intestinal $O_2$ uptake during endotoxemia. J. Appl. Physiol. 64, 2410–2419.
Noren, S.R., Williams, T.M., 2000. Body size and skeletal muscle myoglobin of cetaceans: adaptations for maximizing dive duration. Comp. Biochem. Physiol. A 126, 181–191.
Olson, J.S., 2008. Protein reviews: dioxygen binding and sensing proteins. Section 14: "From $O_2$ Binding Diffusion into Red Blood Cells to Ligand Pathways in Globins". Springer, Milan, Italy.
Perutz, M.F., Mathews, F.S., 1966. An X-ray study of azide methaemoglobin. J. Mol. Biol. 21, 99–207.
Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF chimera—a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612.
Phillips, S.E.V., 1980. Structure and refinement of oxymyoglobin at 1.6 Å resolutions. J. Mol. Biol. 142, 531–554.
Polasek, L.K., Dickson, K.A., Davis, R.W., 2006. Metabolic indicators in the skeletal muscles of harbor seals (*Phoca vitulina*). Am. J. Physiol. Regul. Integr. Comp. Physiol. 290, R1720–R1727.
Ponganis, P.J., Kooyman, G.L., Castellini, M.A., 1993. Determinants of the aerobic dive limit of Weddell seals: analysis of diving metabolic rates, post dive end tidal PO2's and blood and muscle oxygen stores. Physiol. Zool. 66, 732–749.
Ponganis, P.J., Kreutzer, U., Stockard, T.K., Lin, P.C., Sailasuta, N., Tran, T.K., Hurd, R., Jue, T., 2008. Blood flow and metabolic regulation in seal muscle during apnea. J. Exp. Biol. 211, 3323–3332.
Qvist, J., Hill, R.D., Schneider, R.C., Falke, K.J., Liggins, G.C., Guppy, M., Elliot, R.L., Hochachka, P.W., Zapol, W.M., 1986. Hemoglobin concentrations and blood gas tensions of freediving Weddell seals. J. Appl. Physiol. 61, 1560–1569.
Reed, J.Z., Chambers, C., Fedak, M.A., Butler, P.J., 1994. Gas exchange of freely diving grey seals (*Halichoerus grypus*). J. Exp. Biol. 191, 1–18.
Rowell, L.B., 1986. Human Circulation: Regulation during Physical Stress. Oxford University Press, Oxford, New York, p. 415.
Samsel, R.W., Schumacker, P.T., 1994. Systemic hemorrhage augments local $O_2$ extraction in canine intestine. J. Appl. Physiol. 77, 2291–2298.
Schenkman, K.A., Marble, D.R., Burns, D.H., Feigl, E.O., 1997. Myoglobin oxygen dissociation by multiwavelength spectroscopy. J. Appl. Physiol. 82, 86–92.
Scott, E.E., Gibson, Q.H., Olson, J.S., 2001. Mapping the pathways for O2 entry into and exit from myoglobin. J. Biol. Chem. 276, 5177–5188.
Torrance, S.M., Wittnich, C., 1994. Blood lactate and acid–base balance in graded neonatal hypoxia: evidence for oxygen-restricted metabolism. J. Appl. Physiol. 77, 2318–2324.
UniProt Consortium, 2011. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. 39, D214–D219.
Weber, R.E., Hemmingsen, E.A., Johansen, K., 1974. Functional and biochemical studies of penguin myoglobin. Comp. Biochem. Physiol. B 49, 197–214.
Weise, M.J., Costa, D.P., 2007. Total body oxygen stores and physiological diving capacity of California sea lions as a function of sex and age. J. Exp. Biol. 210, 278–289.
Williams, T.M., 2001. Intermittent swimming by mammals: a strategy for increasing energetic efficiency during diving. Am. Zool. 41, 166–176.
Williams, T.M., Davis, R.W., Fuiman, L.A., Francis, J., Le Boeuf, B., Horning, M., Calambokidis, J., Croll, D.A., 2000. Sink or swim: strategies for cost efficient diving by marine mammals. Science 288, 133–136.
Williams, C.L., Meir, J.U., Ponganis, P.J., 2011. What triggers the aerobic dive limit? Patterns of muscle oxygen depletion during dives of emperor penguins. J. Exp. Biol. 214, 1802–1812.
Wright, T.J., Davis, R.W., 2006. The effect of myoglobin concentration on aerobic dive limit in a Weddell seal. J. Exp. Biol. 209, 2576–2585.

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

# Positively Selected Sites in Cetacean Myoglobins Contribute to Protein Stability

**Pouria Dasmeh[1,2], Adrian W. R. Serohijos[2], Kasper P. Kepp[1]*, Eugene I. Shakhnovich[2]***

**1** Technical University of Denmark, DTU Chemistry, Kongens Lyngby, Denmark, **2** Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, United States of America

## Abstract

Since divergence ~50 Ma ago from their terrestrial ancestors, cetaceans underwent a series of adaptations such as a ~10–20 fold increase in myoglobin (Mb) concentration in skeletal muscle, critical for increasing oxygen storage capacity and prolonging dive time. Whereas the $O_2$-binding affinity of Mbs is not significantly different among mammals (with typical oxygenation constants of ~0.8–1.2 $\mu M^{-1}$), folding stabilities of cetacean Mbs are ~2–4 kcal/mol higher than for terrestrial Mbs. Using ancestral sequence reconstruction, maximum likelihood and Bayesian tests to describe the evolution of cetacean Mbs, and experimentally calibrated computation of stability effects of mutations, we observe accelerated evolution in cetaceans and identify seven positively selected sites in Mb. Overall, these sites contribute to Mb stabilization with a conditional probability of 0.8. We observe a correlation between Mb folding stability and protein abundance, suggesting that a selection pressure for stability acts proportionally to higher expression. We also identify a major divergence event leading to the common ancestor of whales, during which major stabilization occurred. Most of the positively selected sites that occur later act against other destabilizing mutations to maintain stability across the clade, except for the shallow divers, where late stability relaxation occurs, probably due to the shorter aerobic dive limits of these species. The three main positively selected sites 66, 5, and 35 undergo changes that favor hydrophobic folding, structural integrity, and intra-helical hydrogen bonds.

**Citation:** Dasmeh P, Serohijos AWR, Kepp KP, Shakhnovich EI (2013) Positively Selected Sites in Cetacean Myoglobins Contribute to Protein Stability. PLoS Comput Biol 9(3): e1002929. doi:10.1371/journal.pcbi.1002929

**Editor:** Nir Ben-Tal, Tel Aviv University, Israel

**Received** September 7, 2012; **Accepted** January 5, 2013; **Published** March 7, 2013

**Copyright:** © 2013 Dasmeh et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was made possible by a grant from the Danish National Science Research Council, Project Case 272-08-0041. PD acknowledges the Otto Moensted foundation for providing a travel grant for his stay at Harvard University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: kpj@kemi.dtu.dk (KPK); shakhnovich@chemistry.harvard.edu (EIS)

## Introduction

Upon adapting to the aquatic environment, marine mammals acquired features that improved their diving skills such as increased blood volume and hematocrit, efficient modes of locomotion (stroke-and-glide swimming) [1,2] and ~10–20 times higher myoglobin (Mb) concentration ($C_{Mb}$) in the skeletal muscles contributing substantially to total body oxygen stores and aerobic dive limits [3,4]. Using an integrated Krogh model of the muscle cell, models of convective oxygen transport and aerobic dive limit (ADL), and thermodynamics of $O_2$-binding, we recently showed that wild-type (WT) Mb is more efficient than mutants under severely hypoxic conditions, whereas low-affinity mutants are in fact better transporters at intermediate oxygen pressure [5]. Moreover, while many sites do not affect $O_2$-binding, conserved WT Mb traits are critical for prolonging the ADL of the animals: As the extreme example, mutating the distal His-64 residue can reduce the ADL by up to 14 minutes under routine dive conditions, and $C_{Mb}$ almost linearly extends the ADL *ceteris paribus*, explaining the extreme increase in $C_{Mb}$ occurring in the cetaceans [6].

Despite the intense research into the structure, function and physiological role of Mb [5–9], the evolution of Mb is not well understood [7]. Several studies have suggested that Mb is under a selection pressure for its function and structural integrity [7,10–11]. Based on amino acid chemical properties and comparative studies of known Mb sequences, some form of selection has been suggested in the evolution of mammalian Mb to favor retention of the conformational structure [10]. Moreover, it has been shown that variable sites in cetacean Mbs are fewer in number but more prone to change than primate Mbs suggesting a probable shift in the function of Mb in cetaceans [11]. However, it is still unclear what drives Mb evolution, as are the specific sites potentially under positive selection and the changes in phenotype they might introduce.

Mb is a relatively conserved protein in all mammals [12]. In a sequence alignment of Sperm whale, Pig, Bovine, Dog, Sheep, Horse and Human Mb, 107 out of 153 residues, including those essential for $O_2$ binding, are identical (See Text S1). Also, Mb oxygen affinity is nearly the same ($K_{O2} \approx 0.8–1.2\ \mu M^{-1}$) for mammalian species. This observation is probably due to the "reversible binding" requirement of molecular $O_2$ to Mb [13] at a given oxygen pressure, $P_{O2}$, which strongly constrains oxygen binding thermodynamics across mammalian cells [5]. Despite similar $K_{O2}$, another protein phenotype, the folding stability (i.e. the free energy of folding the protein, $\Delta G_{folding} = G_{folded} - G_{unfolded}$), is systematically higher in marine mammals compared to their terrestrial counterparts [14]. In a study of mammalian apoMbs, sperm whale apoMb was found to be ~2.5 kcal/mol more stable than horse apoMb [15]. The stability difference can reach up to ~4.5 kcal/mol when goose-beaked whale is compared to pig [16].

## Author Summary

In this work, we identify positive selection in cetacean myoglobins and an early, significant divergence event. While $O_2$-binding is nearly unchanged, positive selection acts to introduce and later maintain stability. Stability correlates with abundance across the species, supporting that selection for increased stability concurred with the known 10–20 fold increase in myoglobin abundance of cetaceans relative to terrestrial mammals, which itself resulted from speciation towards longer dive lengths of the animals. We suggest that this selection acted to keep constant the otherwise increasing number of unfolded Mb. Altogether, this work for the first time links protein phenotype (stability and abundance) in a specific, real protein to organism-level evolution and fitness of mammals.


In this work, using current Bayesian methods to detect selection and a physical force field to compute the stability of single-point mutations, we first identify specific residues under positive selection in the cetacean clade and find that the evolution rate is substantially higher in cetacean Mbs compared to terrestrials. Second, we find that mutations in positively selected sites overall contribute to maintaining stability. Third, using ancestral state reconstruction, we demonstrate that most stabilization occurred during the divergence of cetaceans from the terrestrials. Furthermore, we observe a correlation between Mb folding stability and its abundance across species, further confirming that Mb stabilization is selected for in proportion to protein abundance. Thus, the higher Mb abundance required by speciation of cetacean seem to be accompanied by a larger selection pressure to preserve stability, possibly to reduce the copy number of misfolded Mb in the cell, which is a suggested universal selection pressure for highly expressed proteins [17].

## Results/Discussion

### Phylogenetics

The available mammalian Mb sequences were divided into two datasets: 33 nucleotide sequences of mammalian Mbs were used to construct a phylogenetic tree used for evolutionary analysis with codon models (Figure 1A). To infer ancestral states with highest possible accuracy, a larger tree was also constructed from the substantially larger number (82) of available *amino acid* sequences of mammalian Mbs (Figure 1B). For both phylogenies, Zebra finch was the outgroup, cetaceans were divided into two major suborders, Mysticeti (minke whale and sei whale) and Odontoceti (sperm whales, beaked whales, dolphins, and porpoises), and all the branching patterns followed the known mammalian organism tree with order-specific patterns in primates, rodents, carnivore, cetardiodactylans, and cetaceans [18–26]. The accession numbers of all sequences used in this work, as well as full sequences of relevant ancestors are shown in Text S1.

The sequence of ancestral cetacean Mb was inferred from the available mammalian Mb sequences within all orders using the consensus mammalian species tree. Mb sequences from rodents and primates have minor effects on the most probable inferred ancestral sequence of cetacean Mb (see Text S1 for details).

### Detection of positive selection

To test for positive selection, we used codon-based models of nucleotide substitutions to estimate the rate of nonsynonymous to synonymous mutations, $dN/dS$, across different sites and branches of the mammalian phylogeny [27]. Also, all mutations were studied using the FoldX force field [28–30] to investigate whether the sites under selection in some way contribute to the stability phenotype of the Mbs (See Methods section for details).

Table 1 presents a comparison of the nested M0 (i.e. one $dN/dS$ for all lineages) and FR (i.e. one $dN/dS$ for each branch) models for both terrestrial and marine mammals. In the cetacean clade, the likelihood ratio test (LRT) gives a non-significant result of relatively similar $\omega$ ratios across the species. We also constrained $\omega$ to be the same in the whole cetacean clade ($\omega_1$) and different for the rest of the mammals ($\omega_0$). LRT is significant when it is compared with the one-ratio test with P-value$<10^{-16}$. For $\sim$26% of sites in Mb, $\omega_1 = 0.43$ and $\omega_0 = 0.19$, testifying to a significantly higher evolution rate in cetaceans. As a further support, a higher rate of evolution was also observed in the whole-gene $dN/dS$ comparison of cetaceans (Table 2) and primates (Table 3). The null hypothesis of two sets of $dN/dS$ in primate and cetacean Mbs being similar is strongly rejected with the P-value of $\sim 1.33\times10^{-16}$ using the two-sample t-test.

The higher rate of evolution in the cetacean clade could suggest accelerated evolution driven by positive selection of specific sites. To test this, we compared three site pair-models as M1–M2, M7–M8 and M8fix-M8 to identify sites under positive selection, as presented in Table 1 (see Methods section for details). From Table 1, the most stringent test (M8 vs. M8fix) indicated that seven sites (5, 22, 35, 51, 66, 121, and 129) are under positive selection with overall probabilities greater than 0.5 using the Bayes empirical Bayes (BEB) test [31]. Residue 21 was also detected to have a substantially high $dN/dS$, but its rate was not significantly greater than 1 and thus this residue was not detected by the BEB test. All eight sites are shown in Figure 2 with their posterior BEB probabilities using the M8 model, and with a mapping of sites onto the structure of sperm-whale Mb [32].

Table 1 also shows the results of a branch-site test of positive selection, model A, compared with M1a and the null model-A. Evolution rate (i.e. $\omega$) was left to vary (model A) or fixed to 1 (null-model A) on the foreground tree with the marked branch leading to cetaceans (Figure 1A). The LRT was in this case not significant when model A was compared with its null model, but significant compared to model M1a.

### Ancestral state reconstruction and the evolution of stability

To track the mutational pathways across different lineages of cetaceans, we constructed ancestral sequences as shown in Figure 3. Ancestral states were inferred using the large species tree in Figure 1B constructed from 82 Mb amino acid sequences, applying the Dayhoff substitution matrix allowing for among-site-rate-variation as explained in the Methods section. Overall probability of inference was 1 except in the sites 1, 13 and 28 where it is 0.5–0.9. In all of these sites, the alternative preferred amino acid is the initial mutated amino acid. Overall, our results did not encounter the problem of combinatorial ancestral characters that typically lead to non-unique reconstruction of ancestral sequences [33].

Using the FoldX algorithm, we computed the $\Delta\Delta G$ associated with the mutations in each branch of phylogeny as is shown in Figure 3. The overall stabilization or destabilization of each branch is depicted in red or blue, and the branch height is proportional to the absolute computed $\Delta\Delta G$ value of that specific branch. The overall stability increases in seven branches distributed from $-0.3$ to $-5.1$ kcal/mol.

Upon divergence of cetaceans from the rest of mammals, the most substantial increase of $\sim$5.1 kcal/mol was gained by

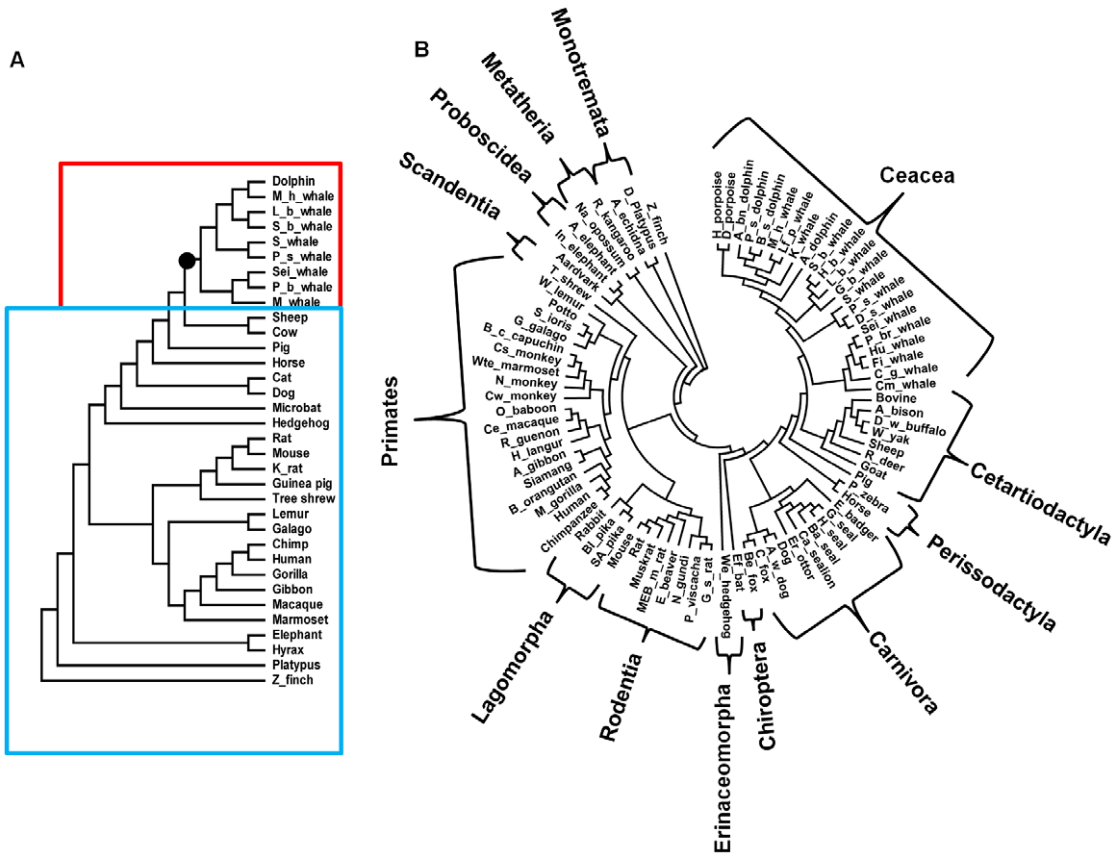



**Figure 1. The mammalian phylogenetic tree constructed from A) nucleotide sequences and B) amino acid sequences.** The smaller tree A was used in maximum likelihood tests for adaptive evolution while the tree B was explicitly used for ancestral state reconstruction. The best evolutionary model with the lowest BIC score was Tamura-Nei92 with transition/transversion bias, R = 1.66 in A and Dayhoff in B. Both models allow among-site-rate-variation sampled from a discrete gamma distribution with four categories and shape parameters 0.33 and 0.46 for nucleotide and amino acid sequences respectively. The phylogeny A is divided into two groups of cetaceans (shown in red) and terrestrial mammals (shown in blue) to test the non-uniformity of molecular clock across different lineages and sites. The branch leading to cetaceans is shown with a black circle in Figure 1A.
doi:10.1371/journal.pcbi.1002929.g001

mutations G15A, E27D, V28I, V101I, K118R, and G129A. From Table 1, the total ω is not significantly greater than 1, but this may be an unrealistically strict criterion for a small, highly constrained protein such as Mb, as evolutionary rate is strongly correlated to protein size due to the fraction of near-neutral sites increasing with size. Instead, LRT is significant when the branch-site test for positive selection (model A) is compared with the nearly neutral model (M1a), which indicates a higher ω in this first branch leading to cetaceans. In addition to positive selection under a new selection pressure (to be explained later, selection for a higher $C_{Mb}$ proportional to ADL, and additionally for folding stability), this might also be caused by relaxation of constraints (loss of selection pressure) [34]. Since the $O_2$-binding affinity of Mb is nearly the same in all mammalian species ($K_{O2}$ at 298 K and pH 7 of ~0.8–1.2 $\mu M^{-1}$), we conclude that the higher ω along this ancestral branch is consistent with positive selection under another arising selection pressure. As presented in Table 1, selection is further supported by the identified amino acid sites in the BEB test having high probabilities along this specific branch, and by the massive increase in the stability phenotype of ~5 kcal/mol occurring during this branching. Altogether, these results suggest that the common ancestor of whales already possessed the new stability phenotype that will later be shown to imply that this ancestor was most likely a deep-diver, although our terminal nodes contain both terrestrial, shallow-, and deep-diving mammals.

After this early divergence that presumably established the majority of the new Mb stability, throughout the cetacean lineages, folding stability is seen to be maintained by fixation of several stabilizing mutations. From Figure 3A, the key mutations preserving this tendency are G5A, V13I, V21I, V21L, E27D, G35S, S35H, N66V, N66H, N66I, G74A, D83E, K118R, G121S, and G129A mutations. Eight of these mutations occur in the five sites 5, 35, 66, 121, and 129 which were detected by to be under positive selection. Thus, the insight from pure sequence-based maximum likelihood methods, amino acid substitution probabilities, and changes in biophysical stability as detected by structure-based approaches converge to the same interpretation of positive selection to obtain and maintain a higher Mb stability for the whales. As a further support for the link, G5A, G35S, and G129A mutations have been observed in more stable Mbs in comparative studies [14].

Figure 3B shows $dN/dS$ values for the variable sites in the cetacean clade versus the inferred $\Delta\Delta G$ of the mutations. Four of the positively selected residues (i.e. residues 5, 35, 66, and 121) show an effect on folding stability >0.5 kcal/mol, with 5 and 66 being most significant, both towards stabilization (~0.7 and

**Table 1.** Log likelihood values and parameter estimates of the site models, and branch-site models.

| Clades | Model | ln L | Estimates of parameters | 2Δl | P-value | Positively selected sites (BEB: P(ω>1)>0.50)[a] |
|---|---|---|---|---|---|---|
| **Cetacea** | M0 (one ratio) | −1241.82 | $\omega_0 = 0.1980$ | | | |
| | Free ratio | −1236.39 | See Text S1 | (M0 vs. Free ratio) 10.86 | 0.69 | - |
| | **Site models** | | | | | |
| | M1a | −1251.18 | $p_0 = 0.83845$, $p_1 = 0.16155$, $\omega_0 = 0.02688$, $\omega_1 = 1$ | | | - |
| | M2a | −1248.47 | $p_0 = 0.84199$, $p_1 = 0.14878$, $p_2 = 0.00922$, $\omega_0 = 0.03212$, $\omega_1 = 1.00000$, $\omega_2 = 4.91963$ | (M1a vs M2a) 5.42 | 0.06 | **5**, 22, 35, **51**, **66**, 121, **129** |
| | M7 | −1251.39 | $p = 0.06085$ $q = 0.29213$ | | | - |
| | M8 | −1247.47 | $p_0 = 0.98777$, $p = 0.11682$, $q = 0.66881$, $p_1 = 0.01223$, $\omega = 4.33010$ | (M7 vs. M8) 7.84 | 0.019 | **5**, 22, 35, **51**, **66**, **121**, **129** |
| | M8fix | −1251.06 | $p_0 = 0.86441$, $p = 0.11615$, $q = 2.08136$, $p_1 = 0.13559$, $\omega = 1.00000$ | (M8fix vs M8) 7.18 | $7.37\times10^{-3}$ | - |
| **Terrestrial mammals** | M0 (one ratio) | −4499.91 | $\omega_0 = 0.1062$ | | | - |
| | Free ratio | −4469.29 | See Text S1 | (M0 vs. Free ratio) 61.24 | 0.065 | - |
| **Mammals** | M0 (one ratio) | −4926.63 | $\omega_0 = 0.08$ | | | - |
| | Free ratio | −4872.64 | See Text S1 | (M0 vs. Free ratio) 107.98 | $4.8\times10^{-4}$ | - |
| | **Site models** | | | | | |
| | M1a | −4646.77 | $p_0 = 0.88207$, $p_1 = 0.11793$, $\omega_0 = 0.05590$, $\omega_1 = 1$ | | | - |
| | Clade model (cetaceans) | −4594.72 | $p_0 = 0.68694$, $p_1 = 0.04973$, $p_2 = 0.26333$, branch type 0: $\omega_0 = 0.02043$, $\omega_1 = 1.00000$, $\omega_2 = 0.19272$, branch type 1: $\omega_0 = 0.02043$, $\omega_1 = 1.00000$, $\omega_2 = 0.43113$ | (M1a vs. Clade Model) 104.1 | $<10^{-16}$ | - |
| | **Branch-site models** | | | | | |
| | Model A | −4643.53 | $p_0 = 0.74119$, $p_1 = 0.09943$, $p_2 = 0.14053$, $p_3 = 0.01885$, $\omega_0 = 0.05388$, $\omega_1 = 1$, $\omega_2 = 1$ | (M1a vs Model A) 486 | $<10^{-16}$ | - |
| | Null model A (ω = 1) | −4643.53 | $p_0 = 0.62272$, $p_1 = 0.08364$, $p_2 = 0.25887$, $p_3 = 0.03477$, $\omega_0 = 0.05392$, $\omega_1 = 1$, $\omega_2 = 1$ | (model A vs Null model A) 2 | 1 | 15, 27, 28, 101, 118, 140 |

[a]: P(ω>1)>0.95 is shown in bold.

doi:10.1371/journal.pcbi.1002929.t001

**Table 2.** The pair-wise evolution rate (i.e. dN/dS) among cetacean Mbs using the maximum likelihood approach described in Methods section.

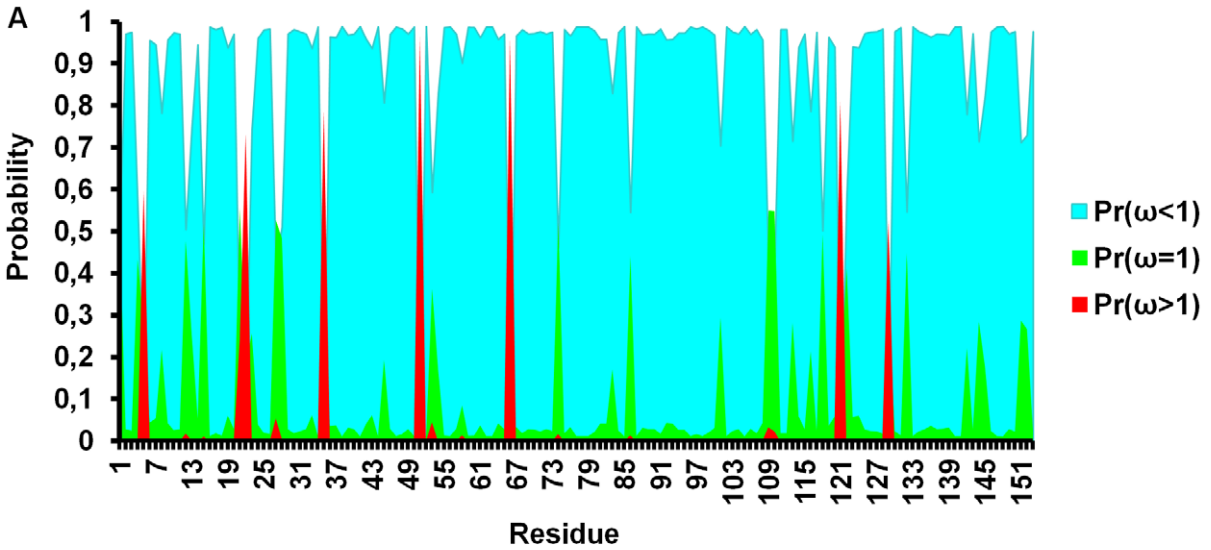
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| L_b_whale | | | | | | | | | |
| S_whale | 0.2761 | | | | | | | | |
| P_s_whale | 0.2259 | 0.2122 | | | | | | | |
| M_whale | 0.2647 | 0.2741 | 0.2735 | | | | | | |
| M_h_whale | 0.2433 | 0.1950 | 0.1890 | 0.1754 | | | | | |
| P_b_whale | 0.3057 | 0.2324 | 0.2636 | 0.1566 | 0.2386 | | | | |
| Sei_whale | 0.3469 | 0.2538 | 0.2832 | 0.1262 | 0.2173 | 0.001 | | | |
| S_b_whale | 0.1079 | 0.2641 | 0.2166 | 0.2796 | 0.2723 | 0.3261 | 0.3705 | | |
| Dolphin | 0.2805 | 0.2536 | 0.2374 | 0.2096 | 0.3328 | 0.2865 | 0.2592 | 0.3176 | |
| | L_b_whale | S_whale | P_s_whale | M_whale | M_h_whale | P_b_whale | Sei_whale | S_b_whale | Dolphin |

doi:10.1371/journal.pcbi.1002929.t002

**Table 3.** The pair-wise evolution rate (i.e. dN/dS) among primate Mbs using the maximum likelihood approach described in Methods section.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Human | | | | | | | | |
| Chimpanzee | 0.0312 | | | | | | | |
| Macaque | 0.0635 | 0.0860 | | | | | | |
| Gibbon | 0.0532 | 0.0774 | 0.0738 | | | | | |
| Marmoset | 0.1272 | 0.1480 | 0.0647 | 0.1101 | | | | |
| Gorilla | 0.0435 | 0.0941 | 0.0949 | 0.1053 | 0.1666 | | | |
| Lemur | 0.0487 | 0.0514 | 0.0566 | 0.0511 | 0.0537 | 0.0513 | | |
| Galago | 0.0964 | 0.0900 | 0.0742 | 0.1138 | 0.0753 | 0.0911 | 0.0905 | |
| | Human | Chimpanzee | Macaque | Gibbon | Marmoset | Gorilla | Lemur | Galago |

doi:10.1371/journal.pcbi.1002929.t003



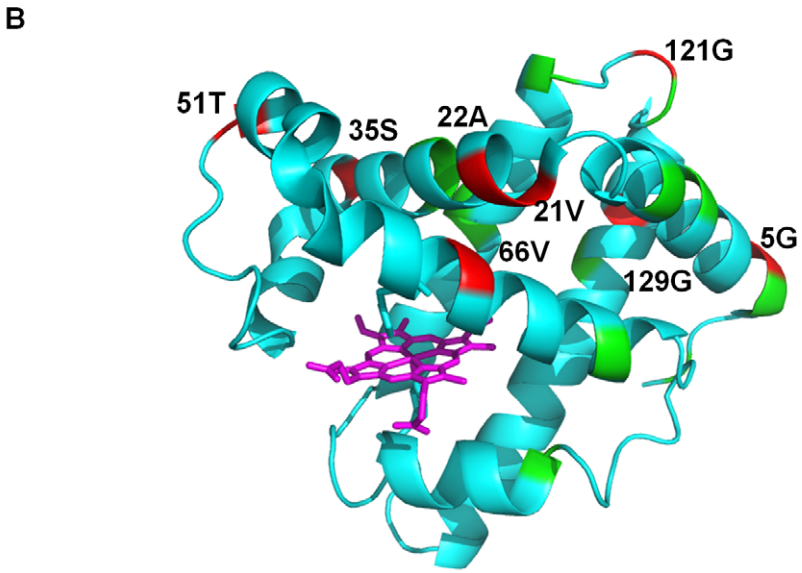




**Figure 2. The Bayes empirical Bayes predictions for ω values for each site in cetacean Mb.** A) For each residue p(ω<1), p(ω = 1) and p(ω>1) are shown in cyan, green and red respectively. Residues 5, 21, 22, 35, 51, 66, 121, and 129 have probabilities (ω>1)>0.5 with <ω> = 5.86 from the M8 model using the ML-estimated branch lengths under the M0 model. B) Crystal structure of sperm whale Mb taken from the protein data bank (ID = 1U7S) [32] with residues color coded by p(ω). The figure was created using PyMOL (http://www.pymol.org).
doi:10.1371/journal.pcbi.1002929.g002

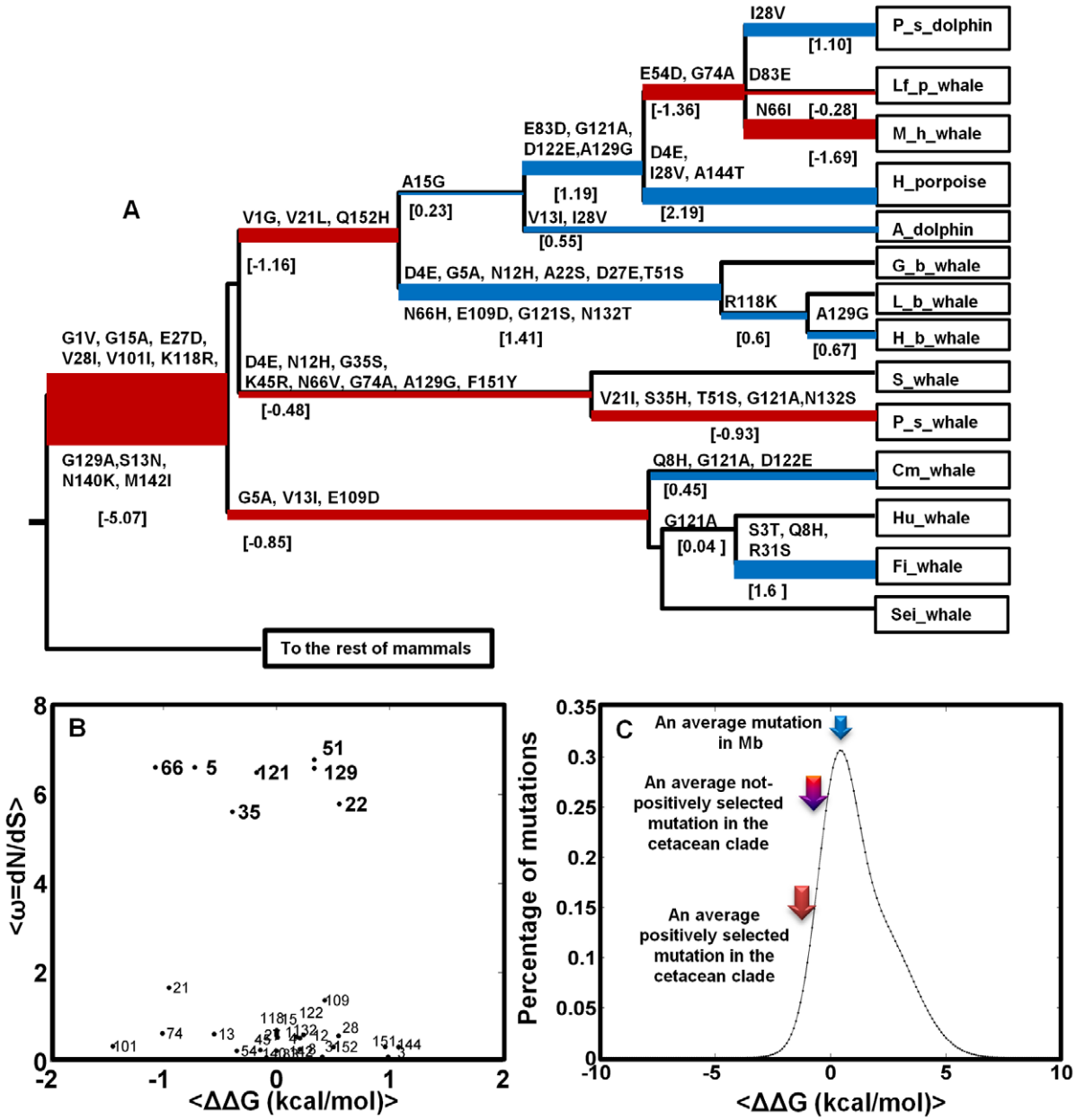



**Figure 3. A) The Phylogenetic tree of cetacean Mb upon the divergence from terrestrial counterparts.** Ancestral states were inferred using the maximum likelihood (ML) approach described in Methods [59]. Amino acid changes in each branch are shown with the respective changes in free energy of folding, ΔΔG in kcal/mol calculated from the FoldX force field [28]. Stabilization and destabilization is presented by red and blue colors respectively across the phylogeny, with branch height proportional to |ΔΔG| of that specific branch. B) The average ω = dN/dS for the variable sites in A from the M8 model is plotted versus the average ΔΔG of mutations in these sites. C) The distribution of mutational effects in Mb from [36] is shown with the solid black line where arrows show the average ΔΔG for an average mutation in Mb (~1.22 kcal/mol), in the cetacean clade among not-positively selected mutations (~0.06 kcal/mol) and, among the positively selected residues (~−0.26 kcal/mol). The probability of stabilization caused by positive selection is ~0.8.
doi:10.1371/journal.pcbi.1002929.g003

~1.0 kcal/mol). Although the G129A mutation, which is fixated in the first branch leading to cetaceans (see Figure 3), is stabilizing (i.e. $\Delta\Delta G = -0.69$ kcal/mol), it undergoes three inversions from Ala to Gly in the branches leading to sperm whales, beaked whales and the suborder of *Delphinidae*, which makes it net destabilizing when summing over occurrences, although this is less significant and could reflect a partial relaxation of stability selection. Insignificant destabilization is also observed in the residues 22 and 51 which will be discussed later.

Figure 3B and 3C show an interesting feature of the evolutionary dynamics of protein stability. As was recently shown by relating protein stability (i.e. $\Delta G$) and evolution rate (i.e. $dN/dS$), proteins may evolve to a stability regime having a detailed balance between stabilizing and destabilizing mutations [35]. Without the stability effects of sites detected to be under positive selection, mutations are distributed nearly symmetrically in the $\Delta\Delta G$ vs. $dN/dS$ scatter plot with an average mutation having $\Delta\Delta G = 0.1$ kcal/mol. The average $\Delta\Delta G$ of an arising mutation in Mb is estimated

to be ~1.2 kcal/mol [36]. Together, these values suggest a balance between stabilizing and destabilizing mutations in the late branches of the cetacean clade.

Positive selection however shifts this balance by fixating stabilizing mutations such as G5A, G35S, S35H, N66V, N66H, N66I, G121S and G129A in the cetacean Mbs, providing a further stabilization of −1.7 kcal/mol for the whole clade and −4.4 kcal/mol when the branches leading to harbor porpoise and common minke whale are removed. These animals have $\Delta G$ similar to that of terrestrials both from experimental mutagenesis and stability measurements and from the FoldX computations. Also, they are shallow divers, consistent with their reduced $C_{Mb}$ (i.e. reduced need for a long ADL [6]), which might suggest that they are under less selection for stability (*vide infra*). Thus, after divergence towards the common deep-diving ancestor, positive selection still acted to maintain and purify Mb stability except in the mentioned case of apparent phenotype relaxation. The role of positive selection is also reflected in the probability of stabilization (i.e. $\Delta\Delta G<0$ kcal/mol) conditional of positive selection, $pr\ (\Delta\Delta G<0 \mid \omega>1)$, using the Bayes rule [37], being ~0.80 (see Text S1 for details). Moreover, the average $\Delta\Delta G$ of positively selected residues is significantly less than that of non-positively selected residues with P-values of 0.0382 and 0.0456 using the two-sample t-test assuming unequal and equal variances in the two datasets, respectively.

Among the seven positively selected sites, four sites display a mutation from Gly to Ala (1, 5, 121, and 129). Gly is known as a strong helix breaker and thus its replacement with Ala will strengthen the helix specifically in soluble proteins [38]. As is shown in Figure 4A and 4B, the G5A mutation is preferred in both Ziphidae (beaked whales) and Mysticeti (baleen whales) suborders of phylogeny. In position 66, a hydrophobic amino acid is stabilizing, confirmed by experimental measurements and most likely due to the hydrophobic effect (i.e. this mutation destabilizes the solvent exposed site in the unfolded protein relative to the folded protein). From Figure 4C, both Ser and His in position 35 can make a hydrogen bond to Arg31. The G35S/G35H mutations are selected in the two more stable physeter species (pygmy sperm whale and dwarf sperm whale) as is shown in Figure 4D. In position 51 which is a surface residue, a Thr to Ser mutation is preferred in two branches leading to beaked whales and to the more stable sperm whales. Both Thr and Ser have similar chemical properties and may form a hydrogen bond with αNH of residue 54 [14].

## Abundance and folding stability of cetacean Mbs correlate: Implications for fitness

So far we have shown that the systematic increase in folding stabilities of cetacean Mbs, partly known from experimental data and further elaborated by the FoldX calculations, is caused by positive selection in this clade of mammalian phylogeny. It is thus important to investigate the biological origin of the selection pressure driving this stabilization. Olson *et al.* has made the rationale for this increased stability as due to the sustained anaerobic and acidic conditions in the skeletal muscle of marine mammals [14,16]. Since whales and seals experience prolonged dives, their Mbs have been suggested to be under selective pressure for increased resistance to unfolding during acidosis [14,16].

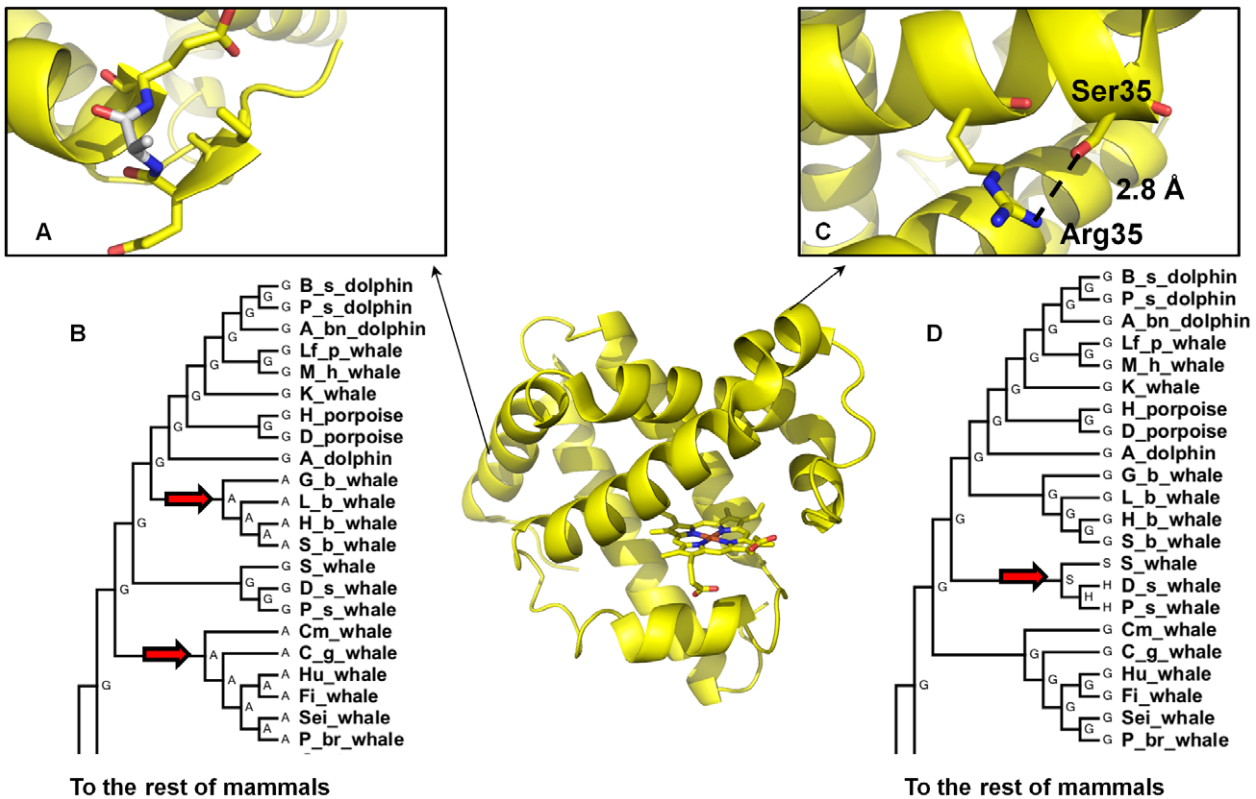



**Figure 4. A) Ala at position 5 shown in the crystal structure of sperm whale Mb as preferred over Gly in two lineages within the cetacean phylogeny B) leading to Baleen whales and Beaked whales.** C) Ser or His at position 35 is preferred over Gly for their ability to make a hydrogen bond with α-CO of Arg31 in the sperm whale clade of cetacean phylogeny D.
doi:10.1371/journal.pcbi.1002929.g004

This hypothesis is in contrast with several observations. First, marine mammals generally stay under aerobic metabolism due to the high cost of recovery after switch to anaerobic conditions [39]. The longest dives recorded for large whales such as blue and fin whales are much shorter than predicted the dive limits under aerobic conditions (ADL) [40]. In similar studies of sperm whales and seals, almost all the dives were found to not greatly exceed ADL [41,42]. Second, the pH-fall in muscle and blood of seals after the long dives is reported to be less than one unit from its physiological value (~7.5) which is too small to initiate unfolding [43]. These observations show that a switch to anaerobic metabolism and sustained acidosis in the muscle is less relevant for the diving patterns of marine mammals as observed in the wild [42].

As seen in Figure 5, upon divergence of marine mammals, a ~10–20 fold increase in Mb concentration ($C_{Mb}$) is experimentally observed, which has been shown to be critical for $O_2$ storage and diving capacity [6]. Moreover, the stability of Mb is also increased: For Pig, Horse, Sheep, Human, Bovine and Dog, $\Delta G$ of apoMb has been reported to be −4.4, −4.8, −4.9, −5.7, −5.8 and −6.3 kcal/mol [14], increasing to −5.1, −7.4, −7.5, −7.8, −8.4 and −8.7 in Dwarf sperm whale (*K. simus*), Pygmy sperm whale (*K. breviceps*), Sperm whale (*P. catadon*), Goose beak whale (*Z. cavirostris*), Dolphin (*Delphinus delphis*), and Minke whale (*B. acutorostrata*). The stability of holoMb is ~2.7 kcal/mol higher than that of apoMb and this difference is assumed to be a constant, since residues in the heme pocket are conserved across all cetaceans [12,14]. The average stability of holoMb is thus ~−7 to −8 kcal/mol for terrestrial mammals and ~−10 to −11 kcal/mol for cetaceans. More importantly, as shown in Figure 5, stability is highly correlated with the species-specific $C_{Mb}$ with a correlation coefficient $\rho = 0.88$ at the significance level $<0.01$. This correlation cannot be explained by adaptation to acidic conditions, because acidic robustness would not depend on protein abundance.

The $\Delta G - C_{Mb}$ correlation is sensitive to various factors: First, $C_{Mb}$ varies somewhat among different muscle types in mammals. Swimming muscles in dolphins contains ~82–86% of total Mb but constitute ~75–80% of total muscle mass, compared to non-swimming muscles [44]. In humans, it is generally known that slow oxidative type I muscles contain more Mb than fast twitch type II muscles [45]. Second, Mb concentration is also age-dependent. Several studies of marine mammals suggest that skeletal muscle of pups have approximately 30% less Mb compared to adults [46,47]. Despite these individual and tissue-wise variations in Mb expression, $C_{Mb}$ for marine mammals is still generally ~10 fold higher than for terrestrial mammals [48].

## Evolution against burden of protein misfolding as $C_{Mb}$ increases

The correlation between protein folding stability and its expression level in the cell was recently proposed to be a consequence of protein misfolding prevention [17]. This hypothesis could explain the universal, strong anti-correlation between protein expression level and evolution rate (ER) in proteins, known as ER anti-correlation, i.e. highly expressed proteins are under stronger selection for stability to reduce the copy number of misfolded proteins [49]. While there may be many other explanations for the ER anti-correlation (i.e. the fitness impact, and hence conservation, of a protein would be proportional to its abundance regardless of the property selected for), the observation of a correlation between protein folding stability in Mb, as one of the most highly expressed mammalian proteins, and its abundance level in different organisms is the first, specific indication that stability as a protein phenotype may be the main property under selection in a real mammalian protein.

We propose that selection against unfolded protein is the cause of both the observed increased evolution rate (Table 1 and 2/3) and the higher stability of the cetacean Mbs. The increased evolution rate of cetacean Mbs with higher expression level seems at first to be in contrast with the average tendency of highly abundant proteins to evolve slowly [50,51]. The explanation for this is most likely that highly expressed proteins that evolve slowly are normally close to equilibrium at their fitness optimum and under stronger selection for conserving stabilizing traits, whereas

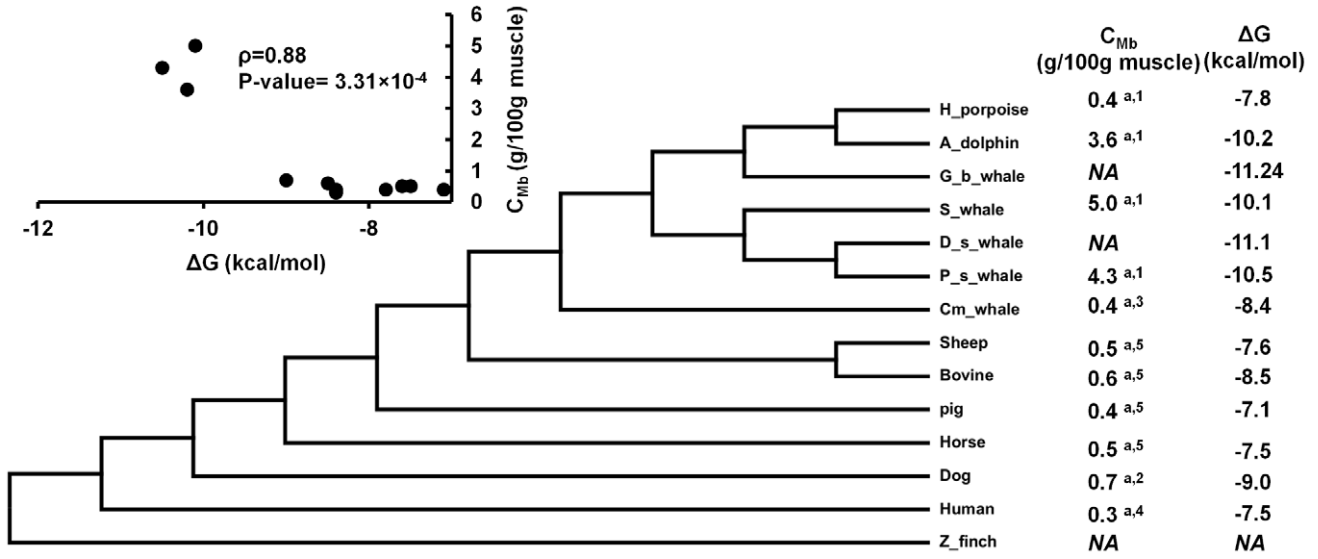



**Figure 5. Divergence of cetaceans and the increase in Mb concentration by ~10–20 fold.** The experimental folding stability of apoMb is added to the difference in stability of holo and apoMb reported for horse heart Mb (2.7 kcal/mol). Stability is highly correlated with Mb concentration with correlation coefficient $\rho = 0.88$ and p-value = 0.000331. The Mb concentration has been measured in [a]dorsi and in [b]psaos muscle types. Data are taken from 1: [57], 2: [75], 3: [44], 4: [7] and 5: [76]. All the folding stabilities are taken from [14].
doi:10.1371/journal.pcbi.1002929.g005

in the present specific evolutionary history, the increased evolutionary rate results from a divergence event where the higher abundance is established together with enhanced stability. This is fully consistent with our observed $C_{Mb}$-stability correlation using available experimental data, with the dive depths of the respective animals, and with the observation of highest evolutionary rate during the first branching event where stability (and presumably $C_{Mb}$) increased the most.

The present results thus also demonstrate how the evolution rate, dN/dS, of a single protein depends on a biophysical property such as in this case stability. Upon divergence to a new niche (deep-diving), the rate increased due to positive selection of new stabilizing mutations, but it is very conceivable that once the optimal stability has been obtained, fixation of new traits will also occur in cetacean Mbs, at least in so far as speciation is complete, which would reduce the rate of evolution as is partly seen in the latter part of the cetacean clade vs. the earlier part. Thus, our results are consistent with the general abundance-evolutionary rate anticorrelation but also suggest that the relation breaks down when highly expressed proteins undergo positive selection towards establishing new traits, leading to a speciation event of both higher evolutionary rate and higher abundance.

In this interpretation, upon the divergence of cetaceans from their terrestrial counterparts, the speciation towards deep divers quickly led to selection for higher $C_{Mb}$, which for deep divers is almost proportional to ADL and by inference, fitness [6]. This early speciation led to an increased selection pressure acting to increase Mb stability in order to minimize the burden of misfolded Mbs within the cell. With a typical 10-fold increase in $C_{Mb}$, an unchanged stability would increase the burden of unfolded Mb by 10-fold in cetaceans, but an average stability increase of ~2 kcal/mol would change the folding equilibrium constant to keep the total copy number of unfolded Mb almost constant across lineages, implying that the burden would be checked in this way.

### Evolution of sites with no significant effect on stability

Among the significantly stabilizing mutations, 5, 35, and 66 were detected to be under positive selection with high posterior probabilities (p ($\omega>1$)~0.80–0.95). The remaining detected sites under positive selection were not significantly affecting stability as seen in Figure 3B. However, they might affect the protein in various other ways that also relate to the increased need for Mb and the adaptation of Mb-enriched deep-divers such as increased signalling requirements or structure preservation beyond thermodynamic stability, e.g. kinetic denucleation/unfolding prevention.

Notably, sites 22 and 51 are predicted to be destabilizing by FoldX in an agreement with previous comparative mutagenesis experiments [16]. Since both these surface residues are substituted for Ser, they may be involved in post translational modifications such as phosphorylation, although a physiological role phosphorylation is unknown [52]. In fact, both residues 22 and 51 are predicted to be phosphorylation sites in whale Mbs using the NetPhos 2.0 server (available at http://www.cbs.dtu.dk/services/NetPhos/) with high scores of 0.82 and 0.97, respectively (See Text S1). Moreover, residue 117 is also detected here as a phosphorylation site as proposed relevant for Beluga whale (*Delphinapterus leucas*) Mb [52]. This observation is consistent with previous studies in enzymes that gain-of-function mutations are on average destabilizing [53], but overall, positive selection still contributes to stability despite these marginally destabilizing sites.

### Concluding remarks

This work suggests that in an important real case of protein evolution, folding stability could be selected for in response to speciation in a new habitat: Our results suggest that the evolution of cetacean Mbs concurred with a divergence of one phenotype – stability – while oxygenation properties remained similar. Folding stability increased significantly (~5.1 kcal/mol) due to the fixation of G15A, E27D, V28I, V101I, K118R, and G129A mutations. We have explained how and why increased Mb stability correlates with increased protein abundance during this evolutionary event, which probably involved substantial competition and speciation as niches were established in the diving regime.

The early, substantial increase in folding stability was accompanied by a significantly higher $dN/dS$ in the first branch leading to cetaceans as judged from the comparison between the nearly neutral model (M1a) and the branch-site model of positive selection on this specific branch. This initial gain of folding stability was then later maintained through the fixation of G5A, V13I, V21L, V21I, V28I, G35S, S35H, N66V, N66I, G74A, V101I, K118R, G121A, and G129A mutations which compensate the deleterious effects of various destabilizing mutations possibly having marginally beneficial fitness effects relating to e.g. regulation. The full picture of these other functionalities would be a relevant focus area in future work.

Later in the clade, we have observed relaxation of the selection for stability. Notably, the common minke whale (*Balaenopetra acutorostrata*) and harbor porpoise (*Phocoenoides phocoena*) display $\Delta G$ and $C_{Mb}$ similar to terrestrial mammals with −8.4 and −7.8 kcal/mol and 0.37 and 0.40 gram per 100 g muscle, respectively. Given the linear effect of $C_{Mb}$ on ADL and by inference the action radius and fitness of the marine mammals [3,6], This observation might be explained by the reduced oxygen consumption demands of both species during diving: Common minke whale is the smallest of the baleen whales with short dive times of ~5–10 minutes [54] compared to sperm whales with an average dive time of ~45 min [55]. Porpoises are also shallow divers (<50 m) with dive times less than two minutes [56]. Therefore, the selective pressure towards more (and more stable) Mb seems to be relaxed in these species if our mechanism is correct, explaining why shallow divers such as porpoises have reverted to less stable Mb. However, across the species, other factors, notably body mass reducing metabolic rate of the animal, also contribute to the total ADL [57], and future data on dive capacities vs. Mb stability would help to clarify the validity of the inferred mechanism.

While evolution is often interpreted as selection for new protein functionality [58], the evolution of cetacean Mbs described in this paper provides the first real example of protein stability being selected for as a consequence of protein abundance, using as control the terrestrials that have 10-fold less Mb. The mechanism by which evolution still acts on the cetacean Mbs, in addition to conservation of the heme pocket due to the reversible binding requirement [13], appears to be one of reducing the animal's burden of the more unfolded Mb copies in the muscle cells by increasing the selection for stability of the highly expressed protein. We suggest that this is the main explanation for the observed accelerated evolution in the cetacean clade.

## Methods

### Phylogenetic analysis and ancestral state reconstruction

The mammalian species tree was analyzed with the MEGA5 package [59] to select the best nucleotide/protein model with the lowest BIC scores, which was the Tamura-Nei92 and Dayhoff model allowing among-site-rate-variation (ASRV) sampled from a discrete gamma distribution with four categories (See Text S1 for details) [60–62]. To infer the ancestral sequences of the cetacean clade, branch lengths were first estimated using the Dayhoff model with

ASRV, and the Bayesian posterior probabilities were calculated for each possible ancestral state for each node [63]. To explore the ancestral sequences inferred, we then used the maximum likelihood method [64] instead of the maximum parsimony (MP) approach due to the limitations of MP in dealing with branch lengths and possible uncertainties in the phylogeny [65].

New Mbs of any member of Ancodonta such as Hippos *(Hippopotamus)*, Camelidae and more species from Cetardiodactyla order such as Alpaca *(Vicugna vicugna)* could possibly resolve better the branch leading to cetaceans and thus provide a finer tree for investigating the episodic nature of *dN/dS* with respect to protein stability.

### Estimating evolution rate and detecting adaptive evolution

The pair-wise comparisons of Mb sequences of cetaceans and primates shown in Table 1 were estimated by the Maximum likelihood approach with codon models in CODEML program implemented in the PAML suite [66]. The equilibrium codon frequencies were estimated from the products of the average observed nucleotide frequencies in the three codon positions (F3X4 model).

To detect adaptive evolution, three codon-based models of nucleotide substitutions for the data [67] with the maximum likelihood inference were employed, first via "branch models" that allow the ω ratio (i.e. dN/dS) to vary among branches in the phylogeny [68]; M0 (one ω ratio for all lineages) and FR (one ω ratio for each branch), and second, via "site models" that allow the ω ratio to vary among codon sites within the sequence [69]. We used five different models referred to as M1 (nearly neutral), M2 (positive selection), M7 (beta), M8 (beta and ω), and M8fix (M8 with ω fixed at 1) [31]. The tree branch lengths were first estimated with the M0 model and were used in the more advanced codon models. We also used the site-models by estimating the branch lengths rather than taking their ML estimated values from the M0 model. With both approaches, the same sites were detected to be under positive selection with significant results in LRTs (see Table S3 in Text S1 for details). Positive selection in the specified residues was also robust to the use of gene tree instead of the organism tree (see Table S4 in Text S1 for details). Synonymous estimates in both marine and terrestrial mammals were less than 1.5 with the exception of one branch having $\omega = 1.56$, and could thus be considered reliable. We ran the CODEML program several times with different initial values to prevent local optima in the Bayesian identification.

To compare the fit of nested models, classified as null and alternative models, the Likelihood Ratio Tests (LRT) was used [70]. Within a LRT test, twice the log-likelihood difference between two nested models has a chi-square distribution with a number of degrees of freedom equal to the free-parameter differences [71]. Different nested pairs of models were compared using the LRT such as branch models M0 versus FR, and Site models M1 versus M2, M7 versus M8, and M8fix versus M8. In cases where the LRT was significant, the Bayes empirical Bayes (BEB) method implemented for models M2 and M8 was employed to calculate the posterior probabilities for codon classes. A third class of LRT tests known as "branch-site" model that allow the ω ratio to vary among both sites and lineages [34] was also employed to infer positively selected sites in the ancestral branch leading to cetaceans. This branch-site test of positive selection was only used on the first branch leading to cetaceans to test the importance of this branching event in the overall divergence of cetaceans from terrestrials (shown with a black circle in Figure 1A). Any further statistical inference in the cetacean clade by detecting branches with high dN/dS values based on the free-ratio model should be corrected by the multiple-hypothesis corrections [72].

### Estimating effects of point mutations on folding stability

The initial 3D-structures used for calculating the stability of single point mutations were taken from the PDB structures of sperm whale Mb at 1.6 Å [32] and 1.4 Å resolution [73]. These structures were subject to the standard protocol of FoldX [28]. We validated the FoldX predicted $\Delta\Delta G$ values for both PDB structures against a set of experimentally reported Mb mutants. We then finally used the repaired PDB structure at 1.4 Å [73] which gave the strongest correlation between calculated and experimental $\Delta\Delta G$s, for computing stabilities within the phylogeny. Individual mutations in the cetacean clade (Figure 3A) were built using "Build Model" command, and $\Delta\Delta G$ values were extracted from the FoldX output files. For both the validation set and mutations in Figure 3A, we repeated each mutation five times and took the average $\Delta\Delta G$ to reduce internal uncertainties of FoldX in estimating the stability effects of mutations, as recently recommended [74] (see Text S1 for details).

## Supporting Information

**Text S1** Text S1 contains the following information: **Table S1:** Experimental and computed FoldX ΔΔG for a range of Mb mutations. The FoldX results (last two columns) are reported using two PDB structures: 1MBO and 1U7S. **Figure S1: ΔΔG** values predicted by FoldX versus experimental ΔΔGs (kcal/mol) for the validation set (pdb = 1MBO). **Figure S2:** ΔΔG values predicted by FoldX versus experimental ΔΔGs (kcal/mol) for the validation set (pdb = 1U7S). **Table S2:** FoldX calculations for all mutations in the Cetacean clade using PDB structure 1U7S. Mutations in the sites detected to be under positive selection are shown in grey. **Table S3:** The best nucleotide and amino acid substitution models fitted to the data. **Table S4:** Results of amino acid substitution models for the whale clade. **Table S5:** Results of nucleotide substitution models for the whale clade. **Table S6:** Likelihood ratio tests for site models when branch lengths are estimated for each model rather than taking the ML-estimated branch lengths from the M0 model. LRT values are shown for M7 vs. M8 and M8 vs. M8fix. **Scheme S1:** Alignment for sperm whale, pig, bovine, dog, sheep, horse and human myoglobin (Mb) sequences. **Scheme S2:** The most probable cetacean ancestor with the complete phylogenetic tree (Figure 1-B), primate-rodent truncated tree, and only the cetacean clade. **Table S7:** LRT values for M7 vs. M8 and M8 vs. M8fix for the gene tree of cetaceans rather than using the species tree. **Table S8:** Species name and accession number of Mb sequences used in this study. The end of Text S1 contains CODEML and NetPhos Output.
(PDF)

## Author Contributions

Conceived and designed the experiments: PD KPK EIS. Performed the experiments: PD AWRS. Analyzed the data: PD AWRS KPK EIS. Wrote the paper: PD AWRS KPK EIS.

## References

1. Williams TM, Davis RW, Fuiman LA, Francis J, Le Boeuf B, et al. (2000) Sink or Swim: strategies for cost efficient diving by marine mammals. Science 288: 133–136.
2. Williams TM (2001) Intermittent swimming by mammals: a strategy for increasing energetic efficiency during diving. Am Zool 41: 166–176.

3. Kooyman GL, Ponganis PJ (1998) The physiological basis of diving to depth: birds and mammals. Annu Rev Physiol 60: 19–32.
4. Davis RW, Kanatous SB (1999) Convective oxygen transport and tissue oxygen consumption in Weddell seals during aerobic dives. J Exp Biol 202: 1091–1113.
5. Dasmeh P, Kepp KP (2012) Bridging the gap between chemistry, physiology, and evolution: Quantifying the functionality of sperm whale myoglobin mutants. Compar Biochem Physiol Part A 161: 9–17.
6. Dasmeh P, Kepp KP, Davis RW (2013) Aerobic dive limits of seals with mutant myoglobin using combined thermochemical and physiologivcal data. Compar Biochem Physiol Part A 164: 119–128.
7. Gros G, Wittenberg BA, Jue T (2010) Myoglobin's old and new clothes: from molecular structure to function in living cells. J Exp Biol 213: 2713–2725.
8. Ho BK, Dill KA (2006) Folding very short peptides using molecular dynamics. PLoS Comput Biol 2: e27.
9. Beard DA (2006) Modeling of oxygen transport and cellular energetics explains observations on in vivo cardiac energy metabolism. PLoS Comput Biol 2: e107.
10. Bogardt RA, Jones BN, Dwulet FE, Garner WH, Lehman LD, et al. (1980) Evolution of the amino acid substitution in the mammalian myoglobin gene. J Mol Evol 15: 197–218.
11. Naylor GJP, Gerstein M (2000) Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins. J Mol Evol 51: 223–233.
12. Suzuki T, Imai K (1998) Evolution of Myoglobin. Cell Mol Life Sci 54: 979–1004.
13. Jensen KP, Ryde U (2004) How heme binds $O_2$: reasons for reversible binding and spin inversion. J Biol Chem 279: 14561–14569.
14. Scott EE, Paster EV, Olson JS (2000) the stabilities of mammalian apomyoglobin vary over a 600–fold range and can be enhanced by comparative mutagenesis. J Biol Chem 275: 27129–27136.
15. Regis WCB, Fattori J, Santoro MM, Jamin M, Ramos CHI (2005) On the difference in stability between horse and sperm whale myoglobins. Arch Biochem Biophys 436: 168–177.
16. Scott EE (1998) Apoglobin Stability and Ligand Movements in Mammalian Myoglobins. Ph.D. dissertation. Rice University, Houston, TX.
17. Drummond DA, Wilke CO (2008) Mistranslation–induced protein misfolding as a dominant constraint on coding–sequence evolution. Cell 134: 341–352.
18. Prasad AB, Allard MW, Green ED (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol Biol Evol 25: 1795–1808.
19. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, et al. (2011) A molecular phylogeny of living primates. PLoS Genetics 7: e1001342.
20. Blanga–Kanfi S, Miranda H, Penn O, Pupko T, DeBry RW, et al. (2009) Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. BMC Evol Biol 9: 71.
21. Bininda–Emonds ORP, Gittleman JL, Purvis A (1999) Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). Biol Rev 74: 143–175.
22. Price SA, Bininda–Emonds ORP, Gittleman JL (2005) A complete phylogeny of the whales, dolphins and even–toed hoofed mammals (Cetartiodactyla). Biol Rev Comb Philos Soc 80: 445–473.
23. Dornburg A, Brandley MC, McGowen MR, Near TJ (2011) Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia: Cetacea). Mol Biol Evol 29: 721–736.
24. Price SA, Bininda–Emonds ORP, Gittleman JL (2005) A complete phylogeny of the whales, dolphins and even–toed hoofed mammals (Cetartiodactyla). Biol Rev Comb Philos Soc 80: 445–473.
25. Hassanin A, Delsuc F, Ropiquet A, Hammer C, Jansen van Vuuren B, et al. (2012) Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. C R Biol 335: 32–50.
26. McGowen MR, Spaulding M, Gatesy J (2009) Divergence date estimation and a comprehensive molecular tree of extant cetaceans. Mol Phylogenet Evol 53: 891–906.
27. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon–substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431–449.
28. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, et al. (2005) The FoldX web server: an online force field. Nucl Acids Res 33: W382–W388.
29. Sánchez IE, Beltrao P, Stricher F, Schymkowitz J, Ferkinghoff–Borg J, et al. (2008) Genome–wide prediction of SH2 domain targets using structural information and the FoldX algorithm. PLoS Comput Biol 4: e1000052.
30. Kiel C, Aydin D, Serrano L (2008) Association rate constants of ras–effector interactions are evolutionarily conserved. PLoS Computational Biology 4: e1000245.
31. Yang Z, Wong WS, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22: 1107–1118.
32. Phillips SEV (1980) Structure and refinement of oxymyoglobin at 1.6 Å resolutions. J Mol Biol 142: 531–554.
33. Gaucher EA (2007) in Ancestral Sequence Reconstruction (ed. Liberles DA). Oxford: Oxford University Press.
34. Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch–site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472–2479.
35. Serohijos AWR, Rimas Z, Shakhnovich EI (2010) Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly. Cell report 2: 249–256.
36. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. J Mol Biol 369: 1318–1332.
37. Bayes T (1763) An essay toward solving a problem in the doctrine of chances. Philos Trans R Soc Lond 53: 370–418.
38. O'Neil KT, DeGrado WF (1990) A thermodynamic scale for the helix–forming tendencies of the commonly occurring amino acids. Science 250: 646–651.
39. Ponganis PJ (2011) Diving mammals. In: Terjung R, editor. Comprehensive Physiology. Hoboken, NJ: John Wiley & Sons, Inc. pp 448–464.
40. Croll DA, Acevedo–Gutiérrez A, Tershy BR, Urbán–Ramírez J (2001) The diving behavior of blue and fin whales: is dive duration shorter than expected based on oxygen stores? Comp Biochem Physiol A 129: 797–809.
41. Watwood SL, Miller PJO, Johnson M, Madsen PT, Tyack PL (2006) Deep–diving foraging behaviour of sperm whales (Physeter macrocephalus). Journal of Animal Ecology 75: 814–825.
42. Kooyman GL, Wahrenbrock EA, Castellini MA, Davis RW, Sinnett EE (1980) Aerobic and anaerobic metabolism during voluntary diving in Weddell seals: evidence of preferred pathways from blood chemistry and behavior. J Comp Physiol 138: 335–346.
43. Hughson FM, Baldwin RL (1989) Using of site–directed mutagenesis to destabilize native apomyoglobin relative to folding intermediates. Biochemistry 28: 4415–4422.
44. Dolar ML, Suarez P, Ponganis PJ, Kooyman GL (1999) Myoglobin in pelagic small cetaceans. J Exp Biol 202: 227–236.
45. Nemeth P, Lowry O (1984) Myoglobin levels in individual human skeletal muscle fibers of different types. J Histochem Cytochem 32: 1211–1216.
46. Clark CA, Burns JM, Schreer JF, Hammill MO (2007) A longitudinal and cross–sectional analysis of total body oxygen store development in nursing harbor seals (Phoca vitulina). J Comp Physiol B 177: 217–227.
47. Kanatous SB, Hawke TJ, Trumble SJ, Pearson LE, Watson RR, et al. (2008) The ontogeny of aerobic and diving capacity in the skeletal muscles of Weddell seals. J Exp Biol 211: 2559–2565.
48. Kanatous SB, Mammen PPA (2010) Regulation of myoglobin expression. J Exp Biol 213: 2741–2747.
49. Yang JR, Zhuang SM, Zhang J (2010) Impact of translational error–induced and error–free misfolding on the rate of protein evolution. Mol Syst Biol 6: 421.
50. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 102: 14338–14343.
51. Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158, 927–931.
52. Stewart JM, Blakely JA, Karpowicz PA, Kalanxhi E, Thatcher BJ, et al. (2004) Unusually weak oxygen binding, physical properties, partial sequence, autoxidation rate and potential phosphorylation sites of beluga what (Delphinapterus leucas) myoglobin. Comp Biochem Physiol B 137: 401–412.
53. Tokuriki N, Stricher F, Serrano L, Tawfik DS (2008) How protein stability and new functions trade off. PLoS Comput Biol 4: e1000002.
54. Stern JS (1992) Surfacing rates and surfacing patterns of minke whales (Balaenoptera acutorostrata) off central California, and the probability of a whale surfacing within visual range. Reports of the International Whaling Commission 42: 379–385.
55. Watwood SL, Miller PJO, Johnson M, Madsen PT, Tyack PL (2006) Deep–diving foraging behavior of sperm whales (Physeter macrocephalus). J Anim Ecol 75: 814–825.
56. Westgate AJ, Read AJ, Berggren P, Koopman HN, Gaskin DE (1995) Diving behaviour of harbour porpoise, Phocoena phocoena. Can Fish Aquat Sci 52: 1064–1073.
57. Noren SR, Williams EE (2000) Body size and skeletal muscle myoglobin of cetaceans: adaptations for maximum dive duration. Comp Biochem Physiol A 126: 181–191.
58. Biswas S, Akey JM (2006) Genomic insights into positive selection. Trends Genet 22: 437–446.
59. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731–2739.
60. Yang Z (1996) Among–site rate variation and its impact on phylogenetic analyses. Trends In Ecology & Evolution 11: 367–372.
61. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512–526.
62. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Natl Biomedical Research pp. 345–352.
63. Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141: 1641–1650.
64. Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics. New York: Oxford University Press.
65. Bollback JP (2006) SIMMAP: stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics 7: 88.
66. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556.

67. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon–substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431–449.
68. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568–573.
69. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon–substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431–449.
70. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568–573.
71. Whelan S, Goldman N (1999) Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol Biol Evol 16: 1292–1299.
72. Anisimova M, Yang Z (2007) Multiple hypotheses testing to detect adaptive protein evolution affecting individual branches and sites. Mol Biol Evol 24: 1219–1228.
73. Kondrashov DA, Zhang W, Aranda IV R, Stec B, Phillips GN (2008) Sampling of the native conformational ensemble of myoglobin via structures in different crystalline environments. Proteins Struct Funct Bioinf 70: 353–362.
74. Christensen NJ, Kepp KP (2012) Accurate Stabilities of Laccase Mutants Predicted with a Modified FoldX Protocol. J Chem Inf Model 52: 3028–42.
75. Reynafarje B (1963) Simplified method for the determination of myoglobin. J Lab Clin Med 6: 138–145.
76. Lawrie RA (1953) The activity of the cytochrome system in muscle and its relation to myoglobin. Biochem J 55: 298–305.

**ARTICLE**[1]

# Influence of protein biophysics on *dN/dS* variations in phylogenetic tress and on overdispersion of the molecular clock

Pouria Dasmeh[1,2,a], Adrian W.R. Serohijos[2,a], Kasper P. Kepp[1], Eugene I. Shakhnovich[2*]

[1]*Technical University of Denmark, DTU Chemistry, DK 2800 Kongens Lyngby, Denmark.*

[2] *Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA USA 02139;*

[a]*Equal contribution.*

**Correspondence: shakhnovich@chemistry.harvard.edu*

[1] **Submitted to *Molecular Biology and Evolution*.**

**Abstract:**

Understanding the relative strength and contribution of the various evolutionary processes—selection, drift, and adaptation—is fundamental to evolutionary biology. An ubiquitous approach is to estimate the ratio of nonsynonymous to synonymous substitutions (i.e., *dN/dS*) in phylogenetic trees. Over the past decades, there has been a great interest in relating *dN/dS*-variations to biophysical properties of proteins. In particular, how does a general protein property such as folding stability affect the rate of protein evolution? In this work, we investigated protein evolution along phylogenetic trees using simulations that combine explicit protein sequences and folding stability changes. First, we tested the correspondence between evolutionary rates from simulation and those estimated using maximum likelihood (ML) method. We found that the *dN/dS* computed by ML methods ($\omega_{ML}$) is highly predictive of the *dN/dS* inferred from the simulated phylogenies. This agreement is strongest in the regime of high stability where proteins are mostly evolving neutrally. Second, we observed molecular clock dispersion, which was especially strong among phylogenetic branches with proteins of low stability. Altogether, our work shows the important contribution of protein biophysics to evolutionary rates and dispersion of the molecular clock.

**Introduction:**

As proposed by Zuckerkandl and Pauling (Zuckerkandl and Pauling 1962) (and subsequently by Margoliash (Margoliash 1963)), the number of amino acid differences between two orthologous proteins is apparently proportional to the elapsed time since the organisms carrying them diverged from a common ancestor. This conjecture initiated the hypothesis that protein evolution exhibits a clock-like behavior, ticking when a single amino acid is fixed in the evolving population (Zuckerkandl and Pauling 1965). Over the last five decades the molecular clock was central to some

debates in evolutionary biology (i.e., selectionism vs. neutralism) and provided the basis for estimating the divergence time of populations and species, detecting natural selection at genomic scales and understanding the origin of sequence variations (Ranala and Yang 2003; Kumar 2005).

Traditionally, the ratio of the rate of nonsynonymous substitutions to synonymous substitutions (*dN/dS*) is used to detect patterns of selection in molecular evolution (Kimura 1977; Yang and Bielawski 2000). The protein is often considered to be under positive selection when the rate of nonsynonymous substitutions (*dN*) exceeds the rate of synonymous substitutions (*dS*). Conversely, $dN/dS < 1$ usually implies that the protein is evolving slowly under negative (purifying) selection (i.e. is more conserved), because most of the non-synonymous substitutions are detrimental to fitness. When $dN/dS \sim 1$, the protein is considered to evolve neutrally.

Estimating *dN/dS* in practice usually requires statistical models of sequence evolution, such as Markov chains (Lio and Goldman 1998; Felsenstein and Churchill 1996; Holder and Lewis 2003). Specifically, maximum likelihood (ML) and Bayesian methods determine the probabilities of substitutions between orthologous sequences using different nucleotide/amino acid substitution models (Whelan 1999; Anisimova et al. 2001; Yang et al. 2005). It is likewise possible to test several biological hypotheses with regards to *dN/dS*-variation across different sites in a protein and along branches and clades of phylogenetic trees and distinguish between them using the likelihood ratio tests (LRT) (Yang 1998).

Nonetheless, despite the prevalence and utility of these statistical tools, it is unclear how their results are dictated or influenced by the biophysical properties of proteins. From a molecular biophysics perspective, the folding stability (folding free energy, i.e., $\Delta G$) is one of the major determinants of sequence evolution (Dokholyan and Shakhnovich 2001; Pal et al. 2006; Zeldovich et al. 2007; Goldstein 2008; Chen and Dokholyan 2008; Taverna and Goldstein a2002;

Taverna and Goldstein b2002). Proteins should fold to their native structure to function, and thus should be stable (except intrinsically disordered proteins (Dyson and Wright 2005). Misfolding caused by accumulation of destabilizing mutations is the main source of loss of function in proteins and the etiological basis of major diseases (Chiti and Dobson 2006; Serohijos et al. 2008). Selection for protein folding, including in some cases selection against detrimental effects of protein aggregation is an emerging global constraint to the evolution of protein coding sequences (Mirny et al. 1998; Mirny and Shakhnovich 1999; Drummond et al. 2005; Zeldovich et al. 2007; Drummond and Wilke 2008; Serohijos et al. 2012; Lobkovsky et al. 2010; Cherry 2010).

To systematically investigate the influence of protein folding stability on *dN/dS* in phylogenetic trees, we have evaluated the impact of $\Delta G$ on the evolution of Myoglobin (Mb) sequences within simulated phylogenetic trees. To that end we simulated populations of cells encoding Mb in their genomes under explicit genotype-phenotype assumptions linking cell fitness to folding free energy $\Delta G$ of Mb proteins in model cells. This ''numerically exact model'' approach enabled us to record complete evolutionary histories and compare *dN/dS* from simulations (explicit count of mutations fixed) with rates estimated from standard approaches such as maximum likelihood. We particularly chose myoglobin because its main functional phenotype (i.e., $O_2$-binding) is constant as measured by $P_{50}$ (i.e., $O_2$ pressure at half Mb saturation) which is also reflected in the conservation of the important functional residues (Suzoki 1998; Scott et al. 2000). Thus, other substitutions in Mb could be due to neutral evolution or to maintenance of biophysical properties such as folding stability.

We found a strong correlation between ML-estimated *dN/dS* and the computed *dN/dS* from simulations in the regime where the evolving protein is very stable, which largely validates the maximum likelihood approach. However, in the regime where proteins are less stable, to maintain folding stability, *dN/dS* was elevated due to selection for stabilizing mutations, causing per site

*dN/dS* ~1.5 across the Mb sequences. Furthermore, the resolution of the phylogenetic tree affected the probability of observing positive selection, specifically, *dN/dS>1* was observed more frequently at higher resolution (shorter branch lengths).

We next explored if the clock was overdispersed in the simulated phylogenies. The neutral theory of molecular evolution predicts that substitutions follow a Poisson process and the variance divided by the mean of substitutions along the branches of a phylogeny (i.e., index of dispersion, *R*) should be 1 (Ohta and Kimura 1971). However, in practice, *R* is found to be larger than 1, suggesting that the molecular clock is overdispersed (Ohta and Kimura 1971; Gillespie 1984; Takahata 1987; Cutler 2000; Hartl 2008). Several models attempted at providing a mechanistic explanation for this overdispersion (Cutler 2000; Hartl 2008). In our simulated phylogenies, we found that the estimated clock from ML was indeed overdispersed depending on the number of branches sampled from the phylogeny (i.e., the size of phylogenetic trees). We discuss how these results have implications for the inference of *dN/dS* and provide a biophysical explanation for the apparent irregularity of the molecular clock.

**Results:**

**Biophysical model for protein evolution**

To investigate the *dN/dS* of a protein evolving under a selection pressure to maintain folding stability, fitness *F* is proportional to the fraction of folded proteins in the cell defined as $F = P_{nat}$ where $P_{nat}$ is the probability that a sequence is in the native state at equilibrium given the two-state model for protein unfolding (Privalov and Khechinashvili 1974; Shakhnovich and Finkelstein 1989;):

$$P_{nat} = \frac{\exp(-\beta\Delta G)}{1+\exp(-\beta\Delta G)} \quad (1)$$

Here, $\Delta G$ is the free energy of folding and $\beta=1/RT$. The effect of mutations on folding stability is modeled as:

$$\Delta G_{after} = \Delta G_{before} + \Delta\Delta \mathrm{G}_{\mathrm{mutation}} \quad (2)$$

The arising mutation would have a selection coefficient *s* defined as (Wylie and Shakhnovich 2011):

$$s = \frac{F_{after}-F_{before}}{F_{before}} \sim e^{\beta\Delta G_{before}}(1 - e^{\beta\Delta\Delta \mathrm{G}_{\mathrm{mutation}}}) \quad (3)$$

which can be positive, negative or zero depending on the nature of mutations to be beneficial, deleterious or neutral. For each mutation, the probability of fixation is defined as:

$$P_{fix} = \frac{1-\exp(-2s)}{1-\exp(-2s\times N_{eff})} \quad (4)$$

where $N_{eff}$ is the effective population size which is ~$10^4$–$10^5$ for mammals (Lynch and Conery 2003; Mailund et al. 2011). The effect of all single point mutations on folding stability is assumed to be additive (Fersht et al. 1992). We thus assume that epistasis comes from the non-linear form of fitness function defined in equation 1. This assumption, in practice, significantly reduces the computational cost (See Methods section for details).

A population of $10^4$ cells ($N_{eff}$) each having one Mb gene was evolved under the simplifying assumption of a monoclonal regime (Materials and Methods). At each generation, a mutation (which changes the folding stability by $\Delta\Delta G$) was attempted, then the probability that the mutation became fixed in the population was determined by equations 3 and 4. We simulated

phylogenetic trees by bifurcating the population every $\lambda$ arising mutations (branch length) (Figure 1B). Each "daughter" population was further bifurcated after $\lambda$ arising mutations. In Figure 1B, we show an exact phylogenetic tree with 1024 external nodes. The simulated phylogenies have different branch lengths due to the stochastic nature of fixation of mutations in the population. These trees were analyzed using the ML and Bayesian tests (i.e., M1-M2, M7-M8 and M8-M8fix analysis).

**Statistical estimation of *dN/dS* is accurate when proteins are stable**

Because we know the full history of the population, we have the explicit count for *dS*, *dN*, and consequently, *dN/dS*. We wanted to investigate whether ML methods could accurately estimate the rates obtained from simulation. We used the codon models and ML estimation implemented in CODEML (Yang 2007) to compute *dN/dS* (i.e., $\omega_{ML}$) for the sequences resulting from simulation. We defined the *dN/dS* from simulations as $\omega_{pop}$ by counting the number of synonymous and nonsynonymous substitutions and normalizing by the number of synonymous and nonsynonymous sites using the sequence information in the simulation trajectories.

The top and right panels in Figure 2A show the histograms of $\omega_{ML}$ and $\Delta G$ for 20,887 branches of 12 independent phylogenetic trees all having the same initial ancestral Mb sequence with $\Delta G$ = -6.84 kcal/mol (i.e., the expected value of $\Delta G$ after fitness equilibration, see Methods section for details). All the branches where *dN* and *dS* are higher than 1.5 were neglected. Bifurcations occurred after $10^5$ mutational attempts (i.e., the $\lambda$ parameter) in the Mb sequence, which corresponds to ~5 fixed amino acid mutations. From the figure, most branches were under purifying selection, having $\omega_{ML} < 1$ with an average of 0.55 and a standard deviation of 0.51. However, there was a finite probability of observing $\omega_{ML} > 1$. Overall, 3035 out of 20887 branches had an elevated rate of nonsynonymous vs. synonymous substitutions. $\Delta G$ spanned from ~ -4

kcal/mol to ~ -10 kcal/mol with an average of -6.34 kcal/mol and a standard deviation of 0.83 kcal/mol. The obtained skewed distribution of $\Delta G$ is in agreement with the empirical distribution of folding stabilities of proteins from the Protherm database (Figure 2A) (Kumar et al. 2006).

The relation between protein folding stability and *dN/dS* was recently generalized as a molecular clock surface (Serohijos et al. 2012). For an evolving protein under selection for being stable, there are three conceivable regimes for *dN/dS*. First, at high stabilities, most mutations do not have a selective advantage/disadvantage. For a protein with $\Delta G$ = -10 kcal/mol, an average mutation with $\Delta\Delta G$ = 1 kcal/mol has a fixation probability of $\sim 10^{-4}$, similar to a neutral mutation (i.e., $P_{fix} = 1/N_{eff}$). At this regime of high stability, most mutations are neutral with *dN/dS* ~ 1. Second, at low stabilities, there is strong selection for stabilizing mutations to avoid misfolding. For a protein with $\Delta G$ = -4 kcal/mol at this regime, even a slightly beneficial mutation with $\Delta\Delta G$ = -0.5 kcal/mol is fixed with a probability of 0.0017, which is ~10 times higher than a neutral mutation.

Figure 2A shows the scatter plot of folding stabilities, $\Delta G$ vs. $\omega_{ML}$. Each point corresponds to an ancestral $\Delta G$ with the $\omega_{ML}$ value calculated to its closest extant sequence in 12 simulated phylogenetic trees all having the same initial ancestral Mb sequence with $\Delta G$ = -6.84 kcal/mol. Bifurcations occur every $\lambda = 10^5$ mutations. There is a higher probability of observing $\omega_{ML} > 1$ at lower stabilities, in agreement with the theoretical prediction (Serohijos et al. 2012). Within the phylogeny, Mb spends much of its time under purifying selection (i.e., $\omega_{ML} < 1$) while traversing to very high and low stabilities with lower probabilities as reflected in $\omega_{ML} \sim 1$ and $\omega_{ML} > 1$ respectively. Compared to very stable regimes where proteins evolve neutrally, the probability of observing $\omega_{ML} > 1$ is increased at intermediate stabilities up to its maximum at $\Delta G \sim$ -6 kcal/mol where $\Delta G$ has its most probable value (Figure 2B). While the molecular clock is expected to tick fastest at the least stable regime (Serohijos et al. 2012), the probability of observing $\omega_{ML} > 1$

decreases because the probability density (i.e., distribution function of $\Delta G$) approaches to 0 at $\Delta G$ = 0 kcal/mol (Goldstein 2011; Bloom et al. 2007; Bloom et al. 2005; Zeldovich et al. 2007).

In the ML estimation of ER, the nonsynonymous to synonymous mutation rate ratio, i.e., the acceptance rate, is treated as a variable of the transition rate matrix of chain states (here 61 states, equal to the number of codons) in a Markov model. This variable along with the branch length and transition/transversion ratio are estimated by maximizing the likelihood function later used in the evaluation of *dN/dS* as $\omega_{ML}$ (Yang 2006). In a pairwise sequence comparison, the nonsynonymous to synonymous substitution rate ratio is then used in the instantaneous rate of all non-synonymous substitutions irrespective of their selection coefficient and thus, fitness effect. In the biophysical model, the relationship between the mutation, the mutational effect on the protein $\Delta G$, the mutational effect on fitness (selection coefficient), and the rate is explicit. Thus, it is interesting to investigate the overall accuracy of the codon-based ML approach by comparing the rates from the biophysical model.

Figure 3A shows the distribution of the ratio $\omega_{pop}/\omega_{ML}$ with a peak at $\omega_{pop}/\omega_{ML}$ = 1; specifically, more than 90% of all comparisons show $\omega_{pop}/\omega_{ML}$ ~1. However, there are deviants in the maximum likelihood inference of $\omega_{pop}$ (i.e., $\omega_{ML}$) that are more frequently observed at lower folding stabilities. The null hypothesis of $\omega_{pop}$ and $\omega_{ML}$ being independent random samples from normal distributions with equal means and equal but unknown variances is strongly rejected when $\Delta G$ > -6 kcal/mol. This indicates a systematic deviation of $\omega_{ML}$ from $\omega_{pop}$ in these regimes (Figure S3 in the supplementary information).

At higher folding stabilities, most mutations do not affect the instantaneous *dN/dS* significantly. For Mb with $\Delta G$ = -9 kcal/mol, *dN/dS* is only altered by mutations having $\Delta\Delta G$ > 4 kcal/mol which have a probability of occurrence < 0.04 (Figure S2 in the supplementary

information). For proteins having stabilities close to the average observed stabilities in the simulated phylogenies (i.e., $\Delta G$ = -6.34 kcal/mol), *dN/dS* fluctuates between high and low values due to the more frequent mutations with marginal effects on stability (~ ± 1 kcal/mol). There is thus a stronger agreement between ML estimation of *dN/dS* and the *dN/dS* from simulations at higher stabilities when nonsynonymous mutations have similar fitness effects. Altogether, we see that the ML estimates of *dN/dS* using codon models is largely accurate in estimating the rate of protein evolution especially in the regime when proteins evolve neutrally (i.e., very stable $\Delta G$).

For an evolving Mb sequence under mutation-selection balance, about five amino acid substitutions occur per $10^5$ mutational attempts in the sequence. This amount of substitution is probabilistically sufficient to bring the protein to the precipice of misfolding and thus forces $\omega_{ML}$ to be > 1 to fix beneficial stabilizing mutations. It is thus interesting to explore whether P($\omega_{ML}$ > 1) is affected by the number of amino acid substitutions along the branches of simulated phylogenies (i.e., the resolution of phylogenetic tree). We chose five different values for $\lambda$ as $10^5$, $1.5\times10^4$, $2\times10^5$, $3\times10^5$ and $5\times10^5$ in 12 phylogenetic trees with each resolution having the same initial ancestral Mb sequence with $\Delta G$ = -6.84 kcal/mol.

Figure 4 shows that P($\omega_{ML}$ > 1) increases at higher resolutions (i.e., smaller $\lambda$ value or fewer amino acid substitutions). For an evolving Mb sequence with $\lambda$ =$5\times10^5$, $3\times10^5$, $2\times10^5$, $1.5\times10^5$ and $10^5$ mutations, P($\omega_{ML}$ > 1) is ~ 0.012, 0.044, 0.080, 0.103, and 0.107, respectively. Moreover, the coefficient of variation (i.e., the standard deviation divided by the mean) of $\omega_{ML}$, as a measure of the dispersion of the distribution, is ~ 0.94, 0.93, 0.89, 0.76, and 0.51 for an evolving Mb sequence with $\lambda$=$10^5$, $1.5\times10^5$, $2\times10^5$, $3\times10^5$, and $5\times10^5$ mutations, respectively. Higher resolutions (i.e., smaller $\lambda$ values) thus capture the stochastic nature of protein evolution with respect to folding stability. For a Mb sequence with a low folding stability ($\Delta G$ ~ -4 kcal/mol), given the infrequent arising stabilizing mutations, there is a higher chance to fix stabilizing

mutations when more amino acid substitutions are allowed (i.e., in smaller $\lambda$ values) prior to the bifurcations in the phylogeny.

To explore the sensitivity of our method to larger population sizes, we simulated also a phylogenetic tree with 1024 extant sequences and $N_{eff} = 10^5$. The average and the variance of *dN/dS* was 0.51 and 0.22, respectively, with the larger population size (i.e., $N_{eff} = 10^5$), significantly smaller than 0.55 and 0.26 at $N_{eff} = 10^4$ (two sample t-test at the significance level of 0.05). Furthermore, $P(\omega_{ML} > 1)$ was slightly higher at the smaller population size with 0.14 and 0.13 for $N_{eff} = 10^4$ and $N_{eff} = 10^5$, respectively. With the larger population size, the average $\Delta G$ decreased to ~ -7.66 kcal/mol, consistent with previous studies on the relation between population size and the strength of selection for folding stability (Goldstein 2011; Wylie and Shakhnovich 2011). This effect is mainly caused by the fact that in smaller populations, deleterious mutations have a higher chance of fixation. Therefore, on average, proteins have lower $\Delta G$ at $N_{eff} = 10^5$. Since proteins are more stable at this condition, we observed a lower probability of $\omega_{ML} > 1$.

**Observation of positive selection**

The statistical significance of the observation of $dN/dS > 1$ is usually inferred by using the likelihood ratio test (LRT) (Yang 1998). We evaluated the statistical significance of *dN/dS* in the simulated phylogenetic tree by comparing two models when $\omega_{ML}$ is set to 1 and is left to vary (Yang 1998). For the simulated sequences, the cases where $\omega_{ML} > 1$ are not statistically significant as judged by the LRT (Figure S4 in the supplementary information). However, it has been shown that proteins with the whole gene-*dN/dS* values in the range of ~0.25 still have signatures of positive selection in specific segments of the gene sequence (Swanson et al. 2004; Sawyer and Malik 2006).

To test if the observation of positive selection in specific gene sequence fragments can be reconciled with the simulated sequences, we used the codon-based models of nucleotide

substitutions to estimate the rate of nonsynonymous to synonymous mutations, *dN/dS*, across different sites (i.e., site models). For an evolving Mb sequence with $\lambda = 10^5$ mutational attempts, three pair-models as M1-M2, M7-M8 and M8fix-M8 were employed to identify sites under positive selection as presented in Table 1 (see Materials and Methods for details). As shown in Table 1, the LRT gave a significant result, with six sites detected to be under positive selection having high posterior probabilities using the Bayes Emprical Bayes test (BEB) (Yang et al. 2005)

**Table 1**. Log likelihood values of the site models with detected sites under positive selection.

| Phylogeny | Models (number of parameters) | ln L | 2Δl | P value | Positively selected sites (BEB: $Pr(\omega > 1) > 0.5)^a$ [$\omega_{ML}$] |
|---|---|---|---|---|---|
| Simulated ($\lambda=10^5$) | M1a (2) | -65183.82 | - | - | - |
| | M2a (4) | -65141.86 | (M1a vs. M2a) 83.92 | $<10^{-16}$ | 34 [1.47], **48** [1.49], **59** [1.50], **119** [1.49], **133** [1.50], **139** [1.50] |
| | M7 (2) | -64591.18 | - | - | - |
| | M8 (4) | -64563.17 | (M7 vs. M8) 56.02 | $6.84\times10^{-13}$ | 48[1.32], **59** [1.50], **119** [1.48], **133** [1.50], **139** [1.50] |
| | M8fix (3) | -64586.49 | (M8 vs. M8fix) 46.64 | $8.53\times10^{-12}$ | - |

a: $Pr(\omega_{ML} > 1) > 0.95$ is shown in bold.

To investigate the reproducibility of the results, we analyzed ten different phylogenetic trees with evolving Mb sequences with $\lambda = 10^5$. LRT was significant in all cases, and different sites were detected to be under positive selection (See the supplementary information). As presented in Table 1, the maximum $\omega_{ML}$ for the sites under positive selection was 1.5, pointing to a weak yet significantly elevated rate of evolution in these positions. This condition imposes a lower bound for the observation of positive selection of $dN/dS \sim 1.5$. Any signal of positive selection, judged by a significant LRT at specific residues with $dN/dS \sim 1.5$ may thus be due to maintaining stability rather than gaining new functions, in particular if mutations occur far from the active site.

**Overdispersion of the molecular clock is observed in simulated phylogenies**

Since selection for maintaining the protein's folding stability (equation 3) can produce *dN/dS* variations among different branches of simulated phylogenies, we then asked how much of this variation could account for the observed overdispersion of the molecular clock in nature. To this end, we calculated the index of dispersion *R* (see Methods). In principle the calculation of *R* could depend on the number of nodes in the phylogeny, and thus we randomly sampled subsets of the phylogenetic tree with increasing number of external nodes, specifically 4, 16, 64, 256, and 1024 (Figure 5A).

Figure 5B shows the probability distribution function of *R* for all possible sub-phylogenies with different sizes obtained from a total of 10 phylogenetic trees with $\lambda = 10^5$ and 1024 extant sequences. For sub-phylogenies with four extant sequences, *R* spans two orders of magnitude from the minimum of ~ 0.07 to the maximum of ~ 4. However, as the sizes of the phylogenetic trees increase, the probability of observing large values of *R* decreases. For example, the probability of observing $R > 1.5$ is ~ 0.31, 0.16, 0.07, 0.03 and <0.01 for sub-phylogenies with 4, 16, 32, 64 and > 64 extant sequences. We also observe that the magnitude of the dispersion for simulated sequences is smaller than that observed in nature (Bedford and Hartl 2008; Gillespie 1984; Gillespie 1986). This difference could be due to the assumption of monoclonality in our simulations, or more likely, to other selective constraints on protein evolution in the real phylogenies apart from selection for folding stability.

We likewise observe that the index of dispersion is slightly higher among sub-phylogenies with ancestral protein sequences that are less stable (Figure 5C). In the stable regime, a protein would evolve at a nearly neutral rate (most mutations have almost neutral effects). In the

regime of low folding stability, a larger fraction of the arising mutations would have deleterious effects (in fact lethal) because the protein is prone to misfolding. Indeed, prior work systematically showed that deleterious mutations raise *R* while beneficial mutations lower it (Cutler 2000).

**Discussion**

Maintenance of stability imposes a selection pressure for many proteins (Mirny and Shakhnovich 1999; Dokholyan and Shakhnovich 2001; Pal et al. 2006; Zeldovich et al. 2007; Goldstein 2008; Heo et al. 2011; Serohijos et al 2012; Serohijos et al. 2013; Soskine and Tawfik 2010). This selection pressure, as we showed in this work, is strong enough to influence the rate of protein evolution, i.e., the "molecular clock".

First, at higher folding stabilities, most arising mutations are neutral as they do not impose tangible effects on fitness (i.e., $P_{nat}$): A highly stable protein (e.g. $\Delta G$ < -9 kcal/mol) is still "stable enough" within ±1 kcal/mol. This stems from the sigmoidal relation between the fraction of folded proteins and folding free energy (Chen and Shakhnovich 2009). In the process of calculating *dN/dS* by ML methods, the nonsynonymous to synonymous substitution rate ratio is assumed to be unchanged for all nonsynonymous substitutions, which is most likely the case at higher folding stabilities. For proteins in this regime, *dN/dS* inferred from ML methods, $\omega_{ML}$, shows a better correlation with the *dN/dS* from simulations calculated by explicitly tracking the number of synonymous and nonsynonymous substitutions and normalizing by the number of synonymous and nonsynonymous sites, $\omega_{POP}$.

Second, for proteins prone to unfolding, stabilizing mutations are fixed at a higher rate, leaving an imprint on the coding sequences through a significantly elevated rate of nonsynonymous substitutions compared to synonymous ones. For Mb sequences evolved under a realistic selection pressure for maintaining folding stability, *dN/dS* was ~1.5. This observation is

justified by the selection of approximately 1−2 nonsynonymous substitutions to bring the folding stability of Mbs back to its average value shown in Figure 2A. The probability of observing $\omega_{ML} > 1$, $pr(\omega_{ML} > 1)$ is also resolution-dependent. Lastly, the observation of positive selection is in line with the view that beneficial mutations must accelerate in order to compensate for deleterious mutations (Mustonen and Lässig 2009; Fisher 1930; Sawyer et al. 2007).

Furthermore, our work provides a natural explanation for the episodic nature of the molecular clock (Gillespie 1984; Gillespie 1986) based on protein biophysics. As shown in Figure 4, for a protein under purifying selection for stabilization, there is a higher chance of observing instantaneous fixation of deleterious destabilizing mutations when branch lengths have higher resolutions (i.e., fewer nonsynonymous mutations). Consequently, the *dN/dS* is elevated due to the fixation of compensatory stabilizing mutations which brings the protein back to the purifying selection regime where $dN/dS < 1$. We thus naturally expect to observe short bursts of rapid evolution to fix stabilizing mutations, followed by long periods of slow evolution where proteins are in the purifying selection regime with $dN/dS < 1$.

One of the advantages of the current approach over sequence-implicit models is the use of more realistic estimates of mutational effects on $\Delta G$. However, it is desirable to compute $\Delta G$ of proteins after each substitution to capture possible non-linear epistatic effects in substitution processes. As was shown recently by Pollock et al., these considerations can influence the propensity of each residue to the nature of close-by residues in the structure (Pollock et al. 2012). Proteins thus adjust to the existence of one amino acid at a residue by coevolution of other residues to keep the amino acid at the initial position (Pollock et al. 2012). This effect, known as the evolutionary Stokes-shift (Pollock et al. 2012), can increase the residence time of proteins in the nearly neutral regimes as the preference for similar amino acids in a position is more likely to maintain the achieved stability of the protein.

As we have shown in this work, the molecular clock will become overdispersed under a selection pressure for stability, and the extent of overdispersion depends on the size of phylogenetic trees. In smaller phylogenies, there is a higher probability for violation of the mutation-selection balance and thus, the index of dispersion becomes larger or smaller than one depending on the excess of fixed deleterious or beneficial mutations. However, in larger phylogenies, $R$ is ~1 and the nearly-neutral theory generally holds.

**Materials and Methods:**

**Estimating the effect of point mutations on protein folding stability**

We used the structure of sperm whale Mb (PDB code=1MBO) (Phillips 1980) as our model protein. For each site in the protein, we exhaustively mutated to all possible amino acid type and estimated the $\Delta\Delta G$ using the mutational free energy change estimator ERIS (Yin et al. 2007a; Yin et al. 2007b). To track the effect of mutations on folding stability within simulations, a 154×20 matrix of $\Delta\Delta G$ values was constructed where each row corresponded to a specific residue in sperm whale Mb and each column corresponds to a possible mutated amino acid (See Table S1 in the supplementary information). For mutations in the residues important for heme and $O_2$ binding (i.e., residues 29, 43, 63, 64, 65, 68, 91, 92, and 93), we did not calculate the $\Delta\Delta G$, but *a priori* assigned $P_{fix} = 0$ to mimic loss of function. The obtained $\Delta\Delta G$ distribution is consistent with experimental $\Delta\Delta G$ values in the ProTherm database (Kumar et al. 2006) and with data from exhaustive computational mutagenesis (Tokuriki et al. 2007).

When the attempted substitutionis non-synonymous, we estimated the change in folding stability as

$$\Delta G = \Delta G_0 + \sum_{i=1}^{n} \Delta\Delta G_i$$

where $\Delta G_0$ is the stability of the protein at time = 0 and $\Delta\Delta G_i$ is the estimated change in folding stability due to a single-point mutation (tabulated in the matrix mentioned above), and $n$ is the total number of amino acid differences of the current sequence with respect to the primordial sequence (i.e., the sequence at time = 0). The effect on folding stability of multiple mutations is simply the additive effect of all mutations acting independently.

**Protein evolution model and the simulated phylogeny**

The Mb sequence was evolved using the nucleotide sequences in a population of $N_{eff} = 10^4$ cells, reasonable effective population size for mammals (Lynch and Conery 2003; Charlesworth 2009; Mesnick et al. 1999). As the mutation rate of mammalian globins, $\mu$, is $\sim 10^{-9}$ amino acid per year (Harris and Hey 1999) and thus $N_{eff} \times \mu << 1$, the evolution of Mb was assumed to be under monoclonal conditions, i.e., at each generation, the population was represented by a single Mb gene. This gene was then let to evolve under selection for folding stability (i.e., equations 3 and 4) until the fitness reached to a steady value, i.e., mutation-selection balance. This fitness-equilibrated sequence which has ~32% identity to sperm whale Mb (See the supplementary information for details) was then used for making the simulated phylogeny as is shown in Figure 1. The initial population was bifurcated after the specified arising mutations (i.e., the $\lambda$-parameter) defined here as multiples of population size (e.g. $\lambda = 10N_{eff} = 10^5$ arising mutations). Bifurcations were then continued to make a simulated phylogeny with 1024 external nodes. For the sake of tractability, more phylogenetic trees were made from the same initial population up to this limit.

**Bioinformatics**

The CODEML program within the PAML suite (Yang 2007) was employed to compute ML-based$dN/dS$, $\omega_{ML}$, for the pairwise comparison of Mb sequences from simulations. The equilibrium

codon frequencies were estimated from the products of the average observed nucleotide frequencies in the three codon positions (F3X4 model). There was no codon preference in the model.

To detect positively selected sites, a multiple sequence alignment of 1024 Mb sequences of the external nodes in the simulated bifurcating phylogenetic tree was used along with the tree in the newick format. The tree branch lengths were first estimated with the M0 model and were used in the more advanced codon models. We then investigated five different models described as M1 (nearly neutral), M2 (positive selection), M7 (beta), M8 (beta and ω), and M8fix (M8 with ω fixed at 1) (Nielsen and Yang 1998; Yang et al. 2005). In this way, we only used the available external sequences (similar to the extant sequences in the real phylogenies) in the process of detecting the sites to be under positive selection.

LRT was applied to compare the fit of nested models, represented as null and alternative models (Yang 1998). Within a LRT test, twice the log-likelihood difference between two nested models should be a $\chi^2$ distribution with a number of degrees of freedom equal to the differences of free parameters in two models. Different nested pairs of site models were compared using the LRT as M1 versus M2, M7 versus M8, and M8fix versus M8. The Bayes empirical Bayes (BEB) method is employed in cases where the LRT was significant and the posterior probabilities of sites putatively under positive selection were recorded. Simulations were performed on the Odyssey cluster at Harvard University for the branches of phylogenetic trees with 5 different $\lambda$ values and also for site models implemented in the CODEML package. The index of dispersion, $R$, was calculated by dividing the variance by the mean for both nonsynonymous and synonymous substitutions in the sub-phylogenies sampled from 10 phylogenetic trees with 1024 extant sequences.

**Acknowledgments**

PD acknowledges the Otto Moensted foundation for providing a travel grant for his stay at Harvard.

**References:**

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol*. 18:1585–1592.

Bedford T, Hartl DL. 2008. Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in Drosophila. *Mol Biol Evol*. 25: 1631–1638.

Bininda-Emonds ORP, Gittleman JL, Purvis A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biol Rev.* 74: 143–75.

Blanga-Kanfi S, Miranda H, Penn O, Pupko T, DeBry RW, Huchon D. 2009. Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. *BMC Evol Biol.* 9: 71.

Bloom JD, et al. 2005. Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci USA.* 102:606–611.

Bloom JD, Raval A, Wilke CO. 2007. Thermodynamics of neutral protein evolution. Genetics. 175:255-266.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet.* 10:195-205.

Chen Y, Dokholyan NV. 2008. Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. *Mol Biol Evol.* 25:1530–1533.

Chen P, Shakhnovich EI. 2009. Lethal mutagenesis in viruses and bacteria. *Genetics.* 183: 639–650.

Cherry JL. 2010. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol Biol Evol.* 27:735–741.

Chiti F, Dobson CM. 2006. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem. 75*:333–366.

Cutler DJ. 2000. Understanding the over-dispersed molecular clock. *Genetics*. 154: 1403–1417.

Dokholyan NV, Shakhnovich EI. 2001. Understanding hierarchical protein evolution from first principles. *J Mol Biol*. 312: 289–307.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA.* 102:14338–14343.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell.* 134:341–352.

Dyson HJ, Wright PE. 2005. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol.* 6:197–208.

Felsenstein J, Churchill GA. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol Biol Evol.* 13: 93–104.

Fersht AR, Matouschek A, Serrano L. 1992. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol.* 224:771–782.

Fisher RA. 1930. The genetical theory of natural selection. Clarendon Press, Oxford.

Gillespie JH. 1984. The molecular clock may be an episodic clock. *Proc Natl Acad Sci USA*. 81: 8009–8013.

Gillespie JH. 1986. Natural selection and the molecular clock. *Mol Biol Evol*. 3:138–155.

Goldstein RA. 2008. The structure of protein evolution and the evolution of protein structure. *Curr Opin Struct Biol.* 18: 170–177.

Goldstein RA. 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins*. 79: 1396–1407.

Harris EE, Hey J. 1999. X chromosome evidence for ancient human histories. *Proc Natl Acad Sci USA*. 96:3320–24.

Heo M, Maslov S, Shakhnovich EI. 2011. Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proc Natl Acad Sci USA*. 108:4258–4263.

Holder M, Lewis PO. 2003. Phylogeny estimation: Traditional and Bayesian approaches. *Nat Rev Genet.* 4:275–284.

Kimura M. 1983. The Neutral Theory of Molecular Evolution (Cambridge Univ. Press, Cambridge, 1983).

Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267: 275–276.

Kumar S. 2005. Molecular clocks: four decades of evolution. *Nat Rev Genet* 6:654–662.

Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A. 2006. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* 34:D204–206.

Lio` P, Goldman N. 1998. Models of molecular evolution and phylogeny. *Genome Res.* 8:1223–1244.

Lobkovsky AE, Wolf YI, Koonin EV. 2010. Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc Natl Acad Sci USA.* 107:2983–2988.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science.* 302:1401–1404.

Mailund T, Dutheil JY, Hobolth A, Lunter G, Schierup MH. 2011. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS Genet* 7: e1001319.

Mirny LA, Abkevich VI, Shakhnovich EI. 1998. *Proc Natl Acad Sci USA.* 95:4976–4981.

Mirny LA, Shakhnovich EI. 1999. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol* 291:177–196.

Margoliash E. 1963. Primary structure and evolution of cytochrome c. *Proc Natl Acad Sci USA.* 50:672–679.

Mesnick SL, Taylor BL, Duc RGL, Trevino SE, O'Corry-Crowe GM, Dizon AE. 1999. Culture and genetic evolution in whales. *Science.* 284:2055a.

Mustonen V, Lassig M. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends Genet. 25:111–119.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics.* 148:929–936.

Nielsen R, Yang Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol.* 20:1231–1239.

Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst*. 23:263–286.

Ohta T, Kimura M. 1971. On the constancy of the evolutionary rate of cistrons. *J Mol Evol*. **1:** 18–25.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet.* 7:337–348.

Phillips SEV. 1980. Structure and refinement of oxymyoglobin at 1.6 Å resolutions. *J Mol Biol.* 142:531–554.

Pollock DD, Thiltgen G, Goldstein RA. 2012. Amino acid coevolution induces an evolutionary stokes shift. *Proc Natl Acad Sci USA*. 109: E1352–9.

Privalov PL, Khechinashvili NN. 1974. A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *J Mol Biol.* 86:665–684.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*.164:1645–1656.

Sawyer SL, Malik HS. 2006. Positive selection of yeast nonhomologous endjoining genes and a retrotransposon conflict hypothesis. *Proc Natl Acad Sci USA.* 103:17614–17619.

Sawyer SA, Parsch J, Zhang Z, Hartl DL. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc Natl Acad Sci USA.* 104: 6504–6510.

Scott EE, Paster EV, Olson JS. 2000. The stabilities of mammalian apomyoglobin vary over a 600-fold range and can be enhanced by comparative mutagenesis. *J Biol Chem.* 275: 27129–27136.

Serohijos AWR, Hegedus T, Aleksandrov AA, He L, Cui L, Dokholyan NV, Riordan JR. 2008. Phenylalanine-508 mediates a cytoplasmic-membrane domain contact in the CFTR 3D structure crucial to assembly and channel function. *Proc Natl Acad Sci USA.* 105:3256–3261.

Serohijos AWR, Rimas Z and Shakhnovich EI. 2012. Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly. *Cell report.* 2:249–256.

Serohijos AWR, Lee SY and Shakhnovich EI. 2013. Highly Abundant Proteins Favor More Stable 3D Structures in Yeast. *Biophysical Journal*, 104: L1–L3.

Shakhnovich EI, Finkelstein AV. 1989. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers.* 28:1667–1680.

Suzuki T, Imai K. 1998. Evolution of myoglobin. *CMLS Cell Mol Life Sci*. 54, 979–1004.

Soskine M, Tawfik DS. 2010. Mutational effects and the evolution of new protein functions. *Nat Rev Genet*. 11:572–582.

Swanson WJ, Wong A, Wolfner MF, Aquadro CF. 2004. Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. *Genetics.* 168:1457–1465.

Takahata N. 1987. On the overdispersed molecular clock. *Genetics*.116:169–179.

Tamuri AU, Dos Reis M, Goldstein RA. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics.* 190:1101–1115.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28: 2731–2739.

Taverna DM, Goldstein RA. a2002. Why are proteins marginally stable? *Proteins.* 46:105-109.

Taverna DM, Goldstein RA. b2002. Why are proteins so robust to site mutations? *Journal of Molecular Biology*. 315:479–484.

Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. 2007. The stability effects of protein mutations appear to be universally distributed. *J Mol Biol.* 369:1318–1332.

Tokuriki N, Tawfik DS. 2009. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol.* 19:596–604.

Whelan S, Goldman N. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol Biol Evol.* 16:1292–1299.

Wylie CS, Shakhnovich EI. 2011. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc Natl Acad Sci USA.* 108: 9916–9921.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol.* 15:568–573.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496–503.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 22:1107–1118.

Yang Z. 2006. Computational Molecular Evolution. (Oxford University Press, Oxford).

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.

Yin S, Ding F, Dokholyan NV. 2007a. Eris: an automated estimator of protein stability. *Nat Methods*. 4:466–467.

Yin S, Ding F, Dokholyan NV. 2007b. Modeling Backbone Flexibility Improves Protein Stability Estimation. *Structure.* 15:1567–1576.

Zeldovich KB, Chen P, Shakhnovich EI. 2007. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci USA*. 104:16152–16157.

Zuckerkandl E, Pauling L. 1962. in Horizons in Biochemistry. eds Kasha M and Pullman B (Academic Press, New York), pp 189–225.

Zuckerkandl E, Pauling L. 1965. in Evolving Genes and Proteins. eds Bryson V and Vogel HJ. (Academic Press, New York), pp 97–166.

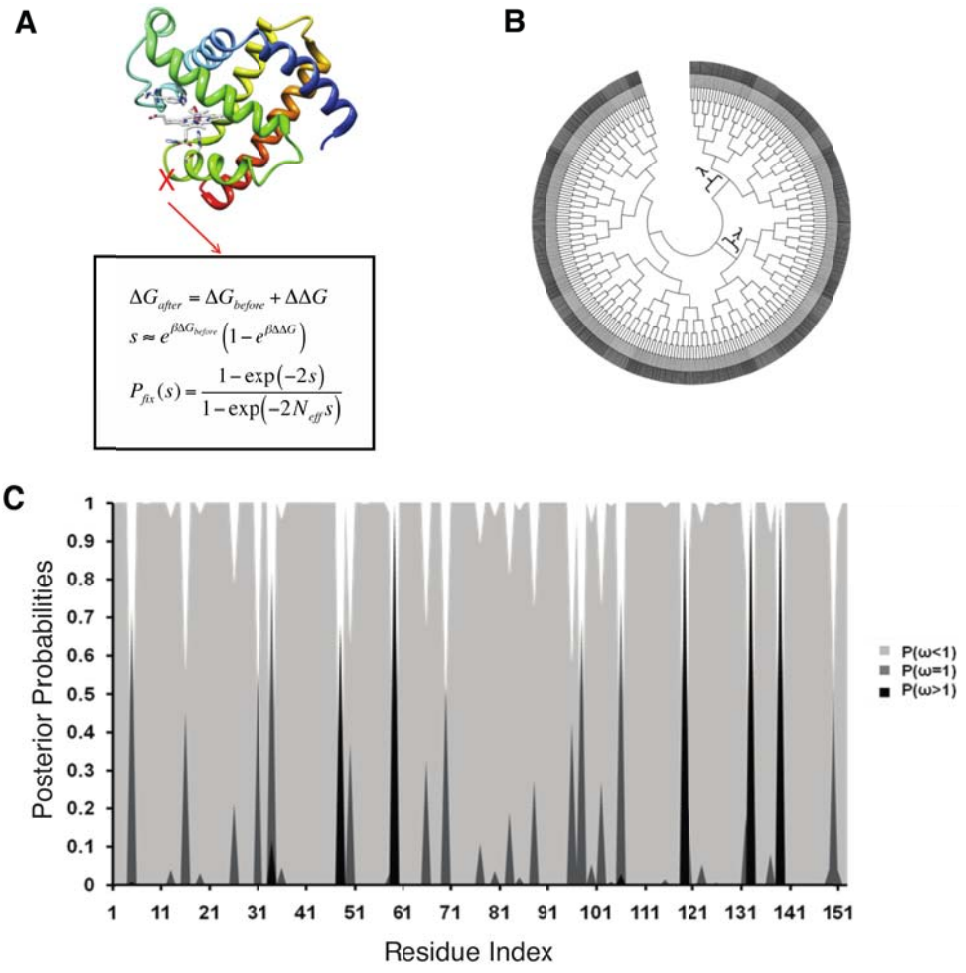



**Figure1:** A) The Mb sequences were evolved in a population of $N_{eff}$ = $10^4$ cells under monoclonal conditions with selection for folding (equation 3 and 4). B) A bifurcating simulated phylogeny with 1024 external nodes was constructed from an initial Mb sequence with $\Delta G$ = -6.84 kcal/mol. Each bifurcation happens after $\lambda$ arising mutations in the ancestral sequence. C) The resulted phylogenetic tree is then used for further evolutionary rate analysis using the ML and Bayesian methods.

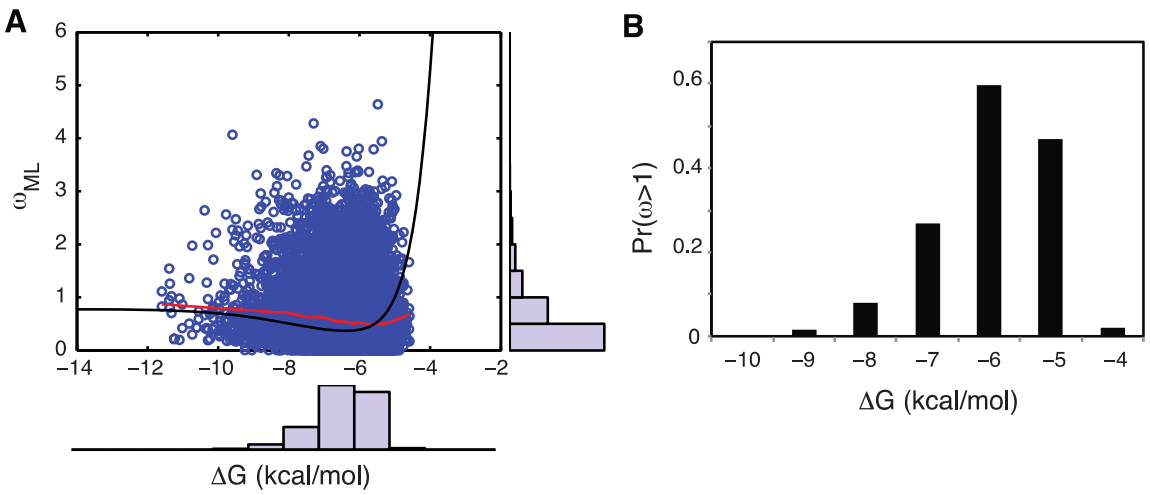



**Figure 2.** A) Top panel: distribution of ω inferred by maximum likelihood estimation for 20887 branches of 12 independent phylogenetic trees. Branches with (*dN* and *dS*) > 1.5 were removed from the total number of branches. Right panel: Folding stabilities of internal branches of the phylogenies. Main figure: The analytical molecular clock curve (in black) (Serohijos et al. 2012) with the scattered $\Delta G$ and $\omega_{ML}$ from the internal nodes of simulations (in red). The locally weighted scatterplot smoothing (LOWESS) line is shown in blue. B) The probability of observing $\omega_{ML} > 1$, $pr(\omega_{ML} > 1)$, at different folding stabilities.

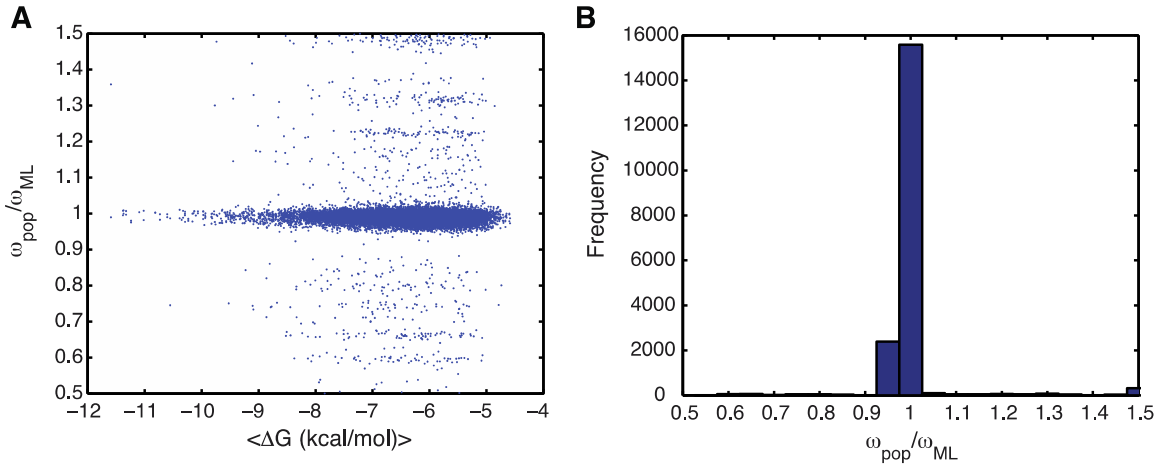



**Figure 3.** A) The ratio between *dN/dS* from simulations, $\omega_{pop}$, and the maximum likelihood estimation of *dN/dS*, $\omega_{ML}$ versus $\Delta G$. The spearman rank correlation between $\omega_{ML}$ and $\omega_{pop}$ shows a correlation coefficient of $\rho = 0.96$ and the *P*-value of ~0. B) Frequency distribution of the ratio $\omega_{pop}/\omega_{ML}$ indicate the overall accuracy the Bayesian methods. Deviations between $\omega_{pop}$ and $\omega_{ML}$ are predominantly in the regime when proteins are less stable (panel A). All analyses were performed on branches of 12 simulated bifurcating phylogenetic trees each having 1024 external nodes.

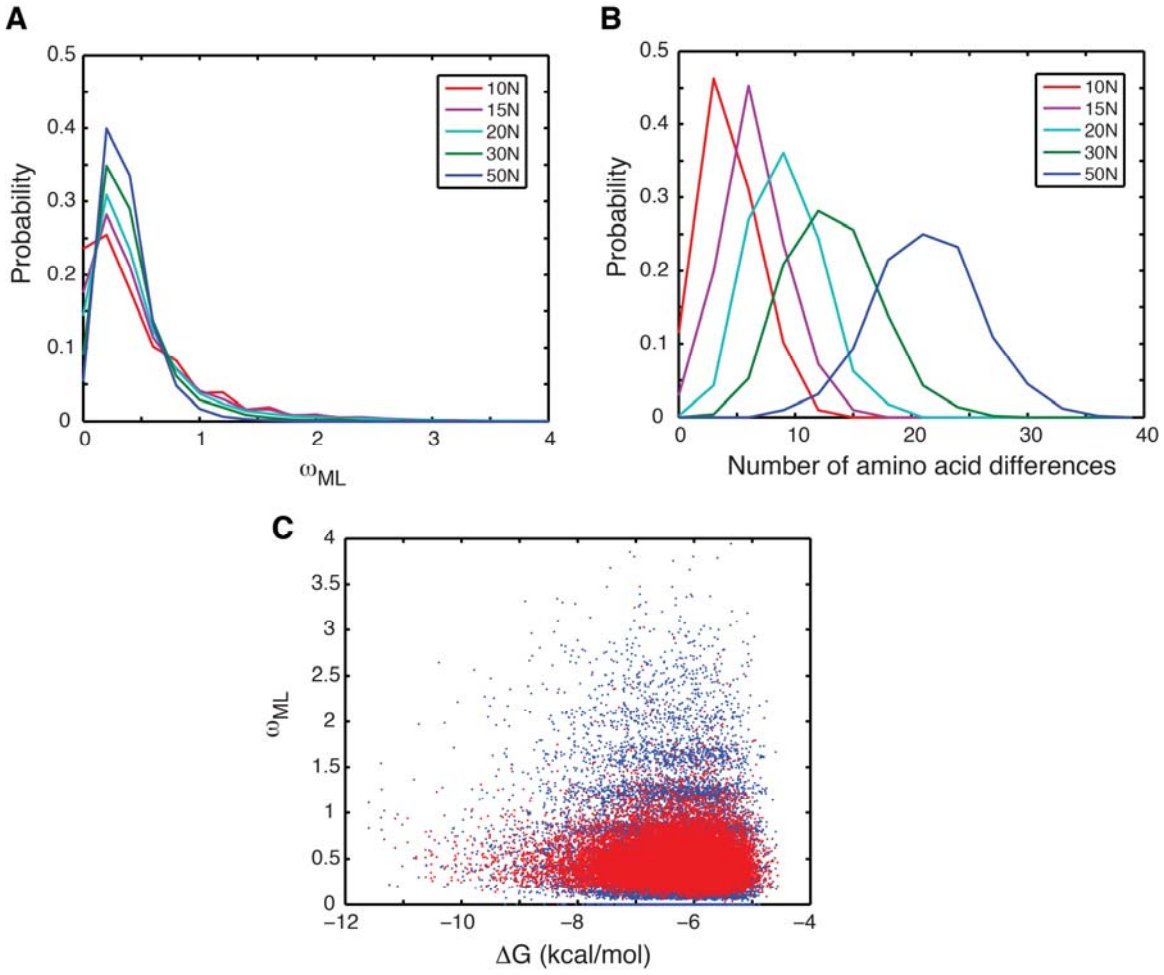



**Figure 4.** Histograms of A) *dN/dS* inferred from the maximum likelihood method, $\omega_{ML}$, and B) the number of amino acid differences among branches of 12 simulated bifurcating simulated phylogenies with 1024 external nodes. Histograms are colored according to the $\lambda$ values defined as the multiples of population size ($N_{eff}$=$10^4$). C) $\omega_{ML}$ versus $\Delta G$ for Mb sequences evolved with $\lambda=10^5$ (in blue) and $5\times10^5$ (in red) mutations having the coefficient of variation of ~ 0.94 and ~0.54 respectively.

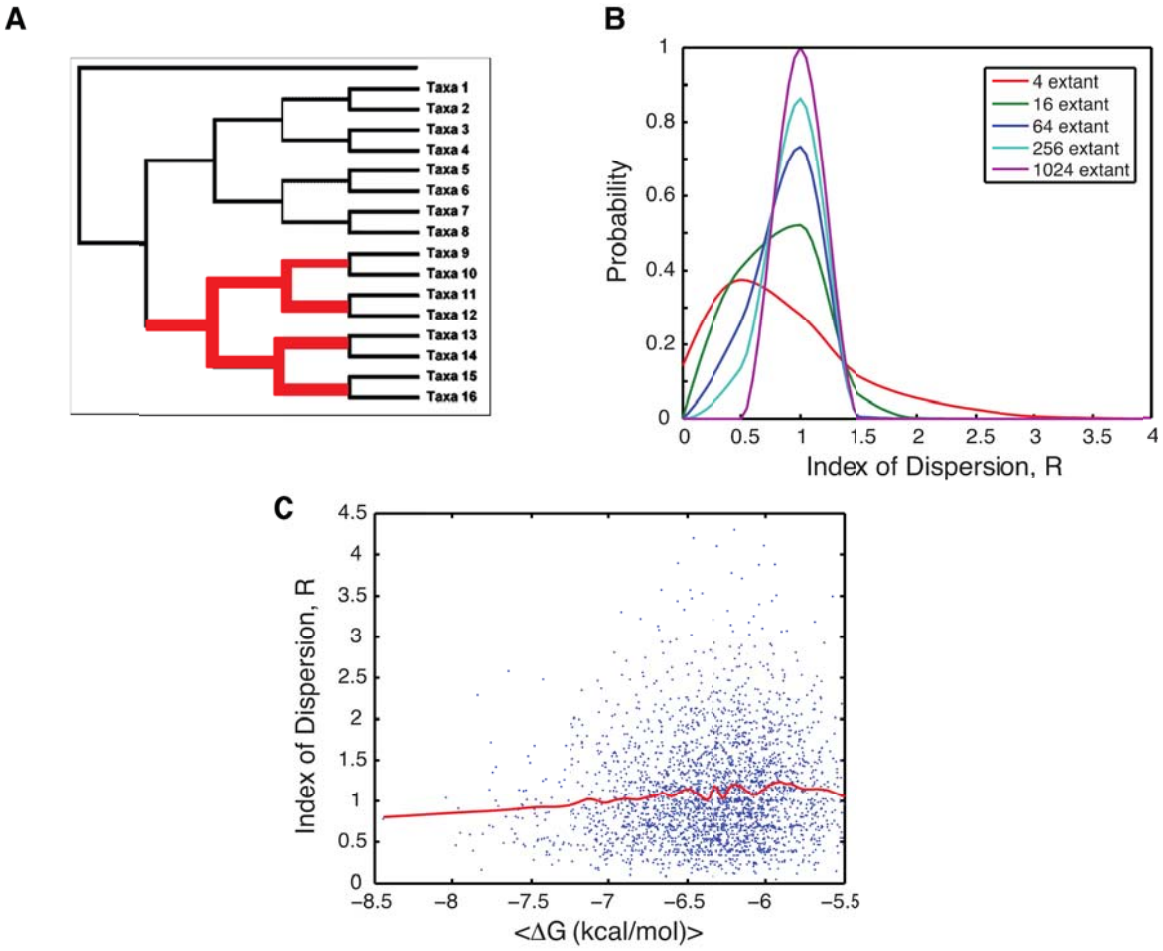



**Figure 5.** A) Schematic of a phylogenetic tree with 16 extant sequences with a sub-phylogeny with 4 extant sequences shown in red. The phylogenetic trees employed for the calculation of the index of dispersion, *R*, have 1024 extant sequences with varying size of sub-phylogenies from 4 to 1024 extant sequences (i.e., the phylogenetic tree itself). B) The probability distribution of R as a function of the size of sub-phylogenies. The sizes are shown in a color spectrum. The bin size is 0.5 so the distribution of R for phylogenetic trees with 1024 extant sequences is a delta function. C) Scatter plot of dispersion index, R, for nonsynonymous mutations for 2560 sub-phylogenies with 4 extant sequences versus the average folding stability, $\langle \Delta G \rangle$, of all ancestral proteins in each sub-phylogeny. Red line is the locally weighted scatterplot smoothing (LOWESS).