SCIENTIFIC REPORTS

OPEN

Tracing whale myoglobin evolution by resurrecting ancient proteins

Yasuhiro Isogai[1], Hiroshi Imamura[2], Setsu Nakae[3], Tomonari Sumi[4], Ken-ichi Takahashi[3], Taro Nakagawa[3], Antonio Tsuneshige[5] & Tsuyoshi Shirai[3]

Received: 4 May 2018
Accepted: 29 October 2018
Published online: 15 November 2018

Extant cetaceans, such as sperm whale, acquired the great ability to dive into the ocean depths during the evolution from their terrestrial ancestor that lived about 50 million years ago. Myoglobin (Mb) is highly concentrated in the myocytes of diving animals, in comparison with those of land animals, and is thought to play a crucial role in their adaptation as the molecular aqualung. Here, we resurrected ancestral whale Mbs, which are from the common ancestor between toothed and baleen whales (*Basilosaurus*), and from a further common quadrupedal ancestor between whale and hippopotamus (*Pakicetus*). The experimental and theoretical analyses demonstrated that whale Mb adopted two distinguished strategies to increase the protein concentration *in vivo* along the evolutionary history of deep sea adaptation; gaining precipitant tolerance in the early phase of the evolution, and increase of folding stability in the late phase.

The analysis of ancient proteins, experimentally reproduced by means of bioinformatics and genetic engineering, is a powerful tool for elucidating the biological molecular evolution and the relationships in protein sequence-structure-function. For example, the reconstruction of ancestral alcohol dehydrogenases from yeast revealed the connection between the chemical behavior of enzymes and the global ecosystem changes[1]. Functional analyses of ancient mammalian uricases demonstrated the evolutionary history of the enzyme and provided new therapeutics for human diseases[2]. The recreation of ancient fluorescent proteins revealed their photochemistry, which was applied to expand the variations of useful biological probes[3]. The determination of the ancestral structures of fish galectin revealed the atomic details of the functional differentiation process of the proteins[4]. The analysis of Precambrian β-lactamase demonstrated the molecular mechanism of novel active site formation[5]. The information obtained from these studies can be utilized in protein engineering and biomedical sciences. In the present study, we investigated the molecular evolution of whale Mb by experimentally resurrecting ancient proteins. The analyses of their chemical properties and structures demonstrated how Mb molecules evolved to adapt ancient whales to deep-sea environments.

Extant whales, such as sperm whales, acquired the great ability to dive into the ocean depths during the evolution from their terrestrial ancestor, which has been dated back to about 50 million years ago[6,7]. Their adaptation to the deep sea is thought to involve various physiological changes at the anatomical, cellular, and molecular levels[8–11]. Hemoglobin (Hb) and myoglobin (Mb) are the key molecules in animal aerobic exercise, as they are responsible for molecular oxygen (O_2) transport in the bloodstream and its storage in the skeletal muscle, respectively. Thus, animal globins have been extensively studied and demonstrated to have evolved to adapt animals to their respective niches[12–18]. In the muscle tissues of deep diving animals, Mb is highly concentrated with its physiological function preserved, whereas the content is significantly lower in land animals, as shown in Table S1 and Fig. S1[8,11,19–22]. Thus, the diving capacity of mammals is thought to correlate with the Mb concentration in their myocytes. Recent studies suggested that the diving mammals have Mbs with more positive net surface charges (Z_{Mb}) than those of terrestrial mammals, and predicted that ancient Mbs had less positive Z_{Mb} values than the offspring diving animals[11,21–24]. These positive charges have been expected to cause electrostatic repulsion among the Mb molecules, to prevent their aggregation and maintain the high protein concentration[21,25,26].

[1]Department of Pharmaceutical Engineering, Toyama Prefectural University, Imizu, Toyama, 939-0398, Japan. [2]Department of Applied Chemistry, College of Life Sciences, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577, Japan. [3]Department of Computer Bioscience, Nagahama Institute of Bio-Science and Technology, 1266 Tamura-Cho, Nagahama, Shiga, 526-0829, Japan. [4]Research Institute for Interdisciplinary Science, Okayama University, 3-1-1 Tsushima-Naka, Kita-ku, Okayama, 700-8530, Japan. [5]Department of Frontier Bioscience and Research Center for Micro-Nano Technology, Hosei University, Koganei, Tokyo, Japan. Correspondence and requests for materials should be addressed to Y.I. (email: yisogai@pu-toyama.ac.jp) or T. Shirai (email: t_shirai@nagahama-i-bio.ac.jp)

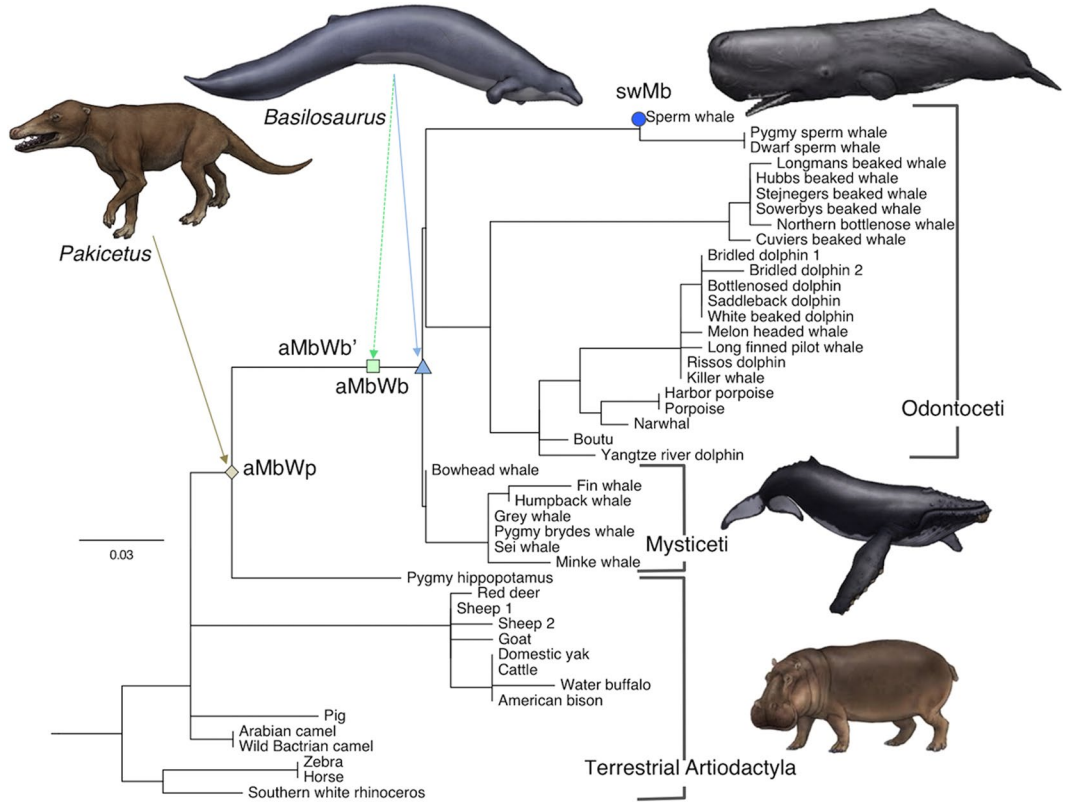


Figure 1. Molecular phylogenic tree of Mbs from whales, land animal relatives, and inferred ancestors. The presented tree is a part of the entire tree consisted of Hbs, Mbs, and other globins (Fig. S2a). Green, light blue, blue, and dark blue circles indicate the positions of aMbWp (land ancestor), aMbWb' (polyphyly whale ancestor), aMbWb (monophyly whale ancestor), and extant swMb. The illustrations of animals are not covered by the CC BY license. Credit to Satoshi Kawasaki. All rights reserved, used with permission.

This highly simplified and attractive adaptation mechanism, however, still remains to be verified in a straightforward manner. A direct comparison of the proteins before and after deep-sea adaptation would provide the most conclusive and convincing evidences. In the present study, the amino-acid sequences of Mbs from extinct genera of cetaceans dated back to the Eocene epoch were inferred, based on the molecular phylogeny of Mbs, Hbs, and other closely related globins known to date, in order to elucidate the adaptation mechanisms (Figs 1 and 2, also see Fig. S2 and Table S2). The three ancestral Mbs, which are keys to understanding the molecular evolution of Mbs in diving animals, were synthesized, and their structures and chemical properties were experimentally and theoretically investigated.

Results and Discussion

The amino-acid sequences of the three synthesized ancient Mbs; namely, aMbWb, aMbWb', and aMbWp, are compared with that of the extant sperm whale Mb (swMb) in Fig. 3a. The aMbWb and aMbWb' proteins are the common ancestors of the toothed (Odontoceti) and baleen (Mysticeti) whales, which would be closely related to the early whale *Basilosaurus*[27]. The aMbWb and aMbWb' are based on the two major conflicting hypotheses of cetacean evolution; namely, the monophyly versus polyphyly hypotheses for Odontoceti[28,29], respectively, and therefore are substitutable for each other. The aMbWp is from a further common ancestor of whales and hippopotamuses, which would be the quadruped terrestrial animal *Pakicetus*[30,31] or its closely related species. The seven residue replacements; namely, E27D, V13I, T34K, D53A, Q116H, K118R and N140K, were deduced to have occurred during the evolution from aMbWp (terrestrial ancestor) to aMbWb (aquatic ancestor). Two residues, G1V and G15A, are different between aMbWb (polyphyly ancestor) and aMbWb' (monophyly ancestor). In aMbWb', these residues are identical to those of aMbWp. Thus, under the currently dominating monophyly hypothesis, aMbWb' is positioned between aMbWp and aMbWb, and therefore regarded as a pseudo-ancestor in this study (Figs 1 and S2). Furthermore, the ten residue replacements, D4E, N12H, I13V, V28I, G35S, K45R, N66V, G74A, D109E and F151Y, occurred during the evolution from aMbWb to extant swMb. Consequently, a total of 17 residue replacements were deduced during the evolution from a terrestrial ancient whale to the existing sperm whale. The relative molecular masses (*Mr*), formal net charges, and formal p*I* values were calculated from the deduced amino acid sequences (Table S3).

The three ancient Mbs and the extant swMb were synthesized in the holo-forms and purified to homogeneity (see Materials and Methods). The *E. coli* expression yield of swMb was higher than the ancient Mbs, and their yields increased along with the whale evolution (Fig. S3). The analyses with size-exclusion chromatography and

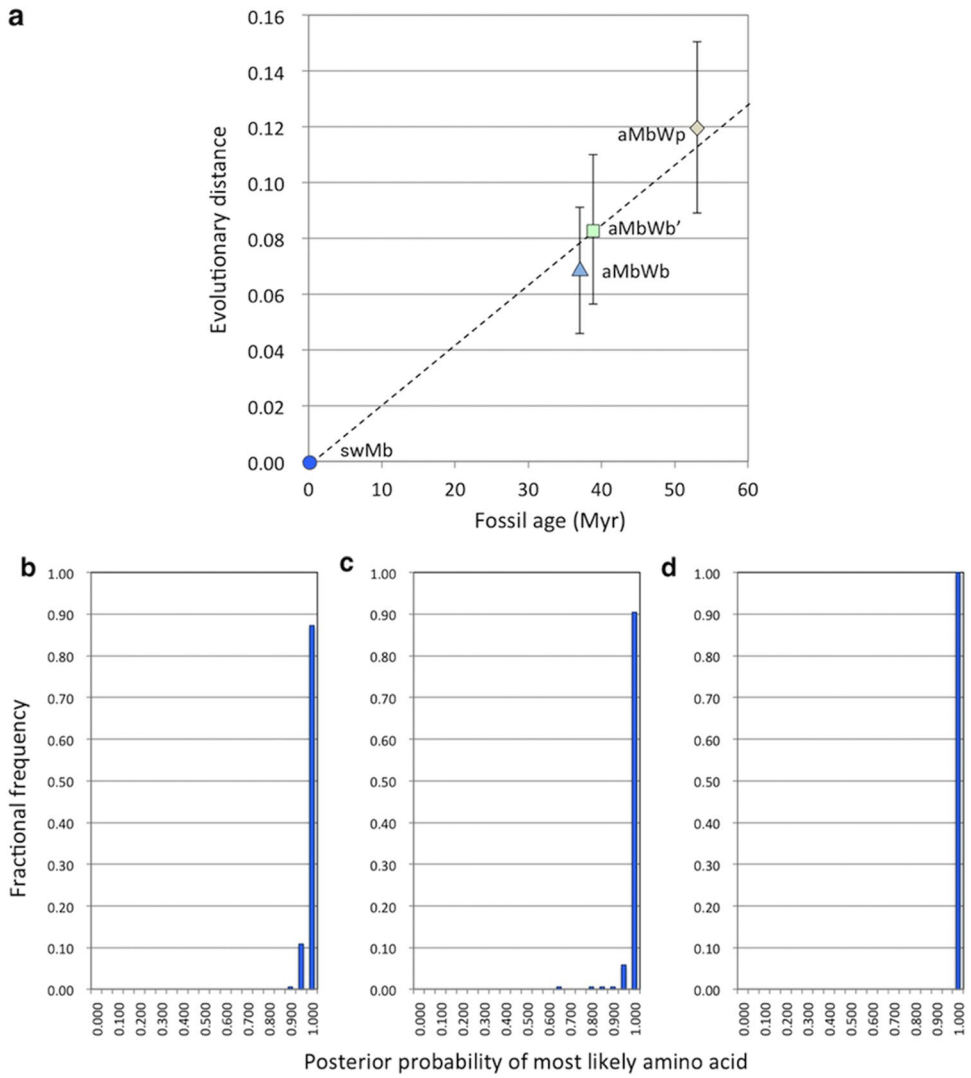


Figure 2. Evolutionary distance and posterior probability distribution of ancestral Mbs. (**a**) Correlation between fossil age of the whale ancestors (horizontal axis) and evolutionary distance of their inferred Mb sequences from that of extant sperm whale (vertical axis). Posterior probability distributions for the sites of aMbWp (**b**), aMbWb' (**c**), and aMbWb (**d**) sequences. Horizontal and vertical axes show posterior probability bins and fractional frequency of inferred sites, respectively.

small angle X-ray scattering (SAXS) indicated that all the Mb samples used here are almost monomeric under the broad range of Mb concentration (see below). The atomic structures of the synthesized aMbWp, aMbWb', aMbWb, and swMb were determined by X-ray crystallography to 2.4, 1.6, 1.4, and 0.8 Å resolutions, respectively (Fig. 3b and Table S4). The main chain structures were well conserved among the ancestral and extant Mbs (Fig. S4). The net surface charges Z_{Mb}, isoelectric points (pI), solvation free energies (ΔG_{solv}), and mutational folding energy changes ($\Delta\Delta G_{mut}$) were calculated, based on the crystal structures and the trajectories of molecular dynamics simulations (Table S3 and Fig. S5), and are discussed later along with the experimental data.

The synthesized Mbs were examined for their solubility, which is one of the most interesting properties of the ancestral Mbs. Generally, Mbs are highly soluble, and consequently it is difficult to obtain reproducible solubility values in normal buffer solutions, because highly concentrated protein solutions are apt to form gels and supersaturated solutions[32]. Thus, the quantitative solubility was evaluated by using polyethylene glycol (PEG) as a protein precipitant, according to the previous studies[33–35]. The solubility dependence on the PEG-6000 concentration in the protein solution was measured at room temperature, as shown in Fig. 4a. The relationship between protein solubility (S) and precipitant concentration can be described by a linear equation: $\log S = \log S_0 + \beta$[precipitant], where $\log S_0$ is the y-intercept of the plot and β is the slope. S_0 and β represent the solubility in water and the resistance to precipitant, respectively (Table S5). Unexpectedly, $\log S_0$ decreased during the evolution from aMbWp to aMbWb, and remained almost unchanged from aMbWb to swMb, indicating that the total solubility in pure water has not increased during evolution. This was also verified by the estimation of the solvation free energy ΔG_{solv} of the Mb molecules, based on a reference-modified density functional theory (Table S3)[36]. The consistent results between experimental and theoretical evaluations indicated that the observed solubility should be an

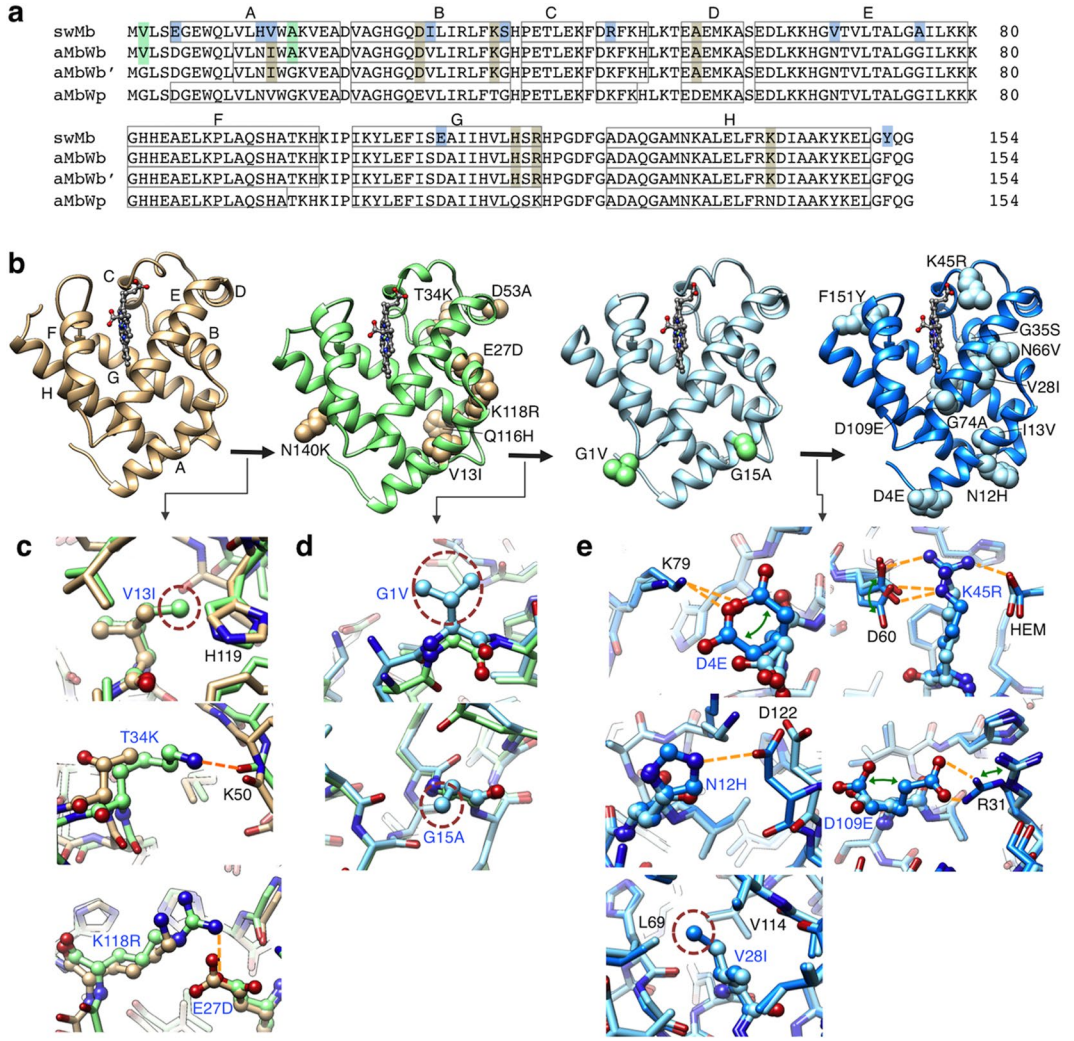


Figure 3. Residue replacements of whale myoglobin during the evolution from the terrestrial animal to sperm whale. (**a**) Amino-acid sequence alignment of ancestral and sperm whale Mbs. Amino acid replacements on aMbWp to aMbWb', aMbWb' to aMbWb, and aMbWb to swMb are meshed with light brown, light green, and light blue, respectively. The residues in the canonical helices A – H are boxed. (**b**) The replaced residues are shown on the crystal structures of aMbWp (PDB code 5YCG), aMbWb' (5YCI and 5YCJ), aMbWb (5YCH), and swMb (5YCE). The canonical helices A – H are indicated on the structure of aMbWp. (**c**) V13I, T34K, and K118R and E27D are replacements from aMbWp (light brown) to aMbWb' (light green). (**d**) G1V and G15A are those from aMbWb' to aMbWb (light blue). (**e**) D4E, V28I, N12H, K45R, and D109E are from aMbWb to swMb (blue). The electrostatic interactions/hydrogen bonds and cavity filling positions are indicted with the yellow dotted lines and red circles, respectively. The green arrows indicate alternative conformations.

intrinsic property of the Mb molecules determined by their structures. ΔG_{solv} increased during the evolution from aMbWp to aMbWb, whereas it minimally changed during the evolution from aMbWb to swMb, indicating that the single-molecule solubility in an aqueous solution rather decreased during the evolution. However, the resistance to the precipitant (β) significantly increased during the evolution from aMbWp to aMbWb', slightly increased from aMbWb' to aMbWb, and remained almost unchanged from aMbWb to swMb.

In the proposed hypothesis of Mb adaptation in diving animals, the Z_{Mb} increase was expected to enhance the protein solubility by preventing precipitation through positive charge repulsion among Mb molecules[11,21], although the present results seem to be inconsistent with that hypothesis. Therefore, we measured SAXS of the Mb solutions to analyze their self-interaction potentials, in order to monitor the repulsion between Mbs (Fig. S6). The second virial coefficients (A_2), which indicate either attractive ($A_2 < 0$) or repulsive ($A_2 > 0$) intermolecular interactions, were obtained from the analyses. The results demonstrated that the intermolecular repulsion has increased during evolution, not only from aMbWp to aMbWb but also from aMbWb to swMb (Fig. 4b,c).

Taken together, both Z_{Mb} and p*I* significantly increased during the evolution from aMbWp to aMbWb as previously hypothesized, whereas their increases are small during the evolution from aMbWb to swMb (Table S3 and Fig. S7). Contrary to the hypothesis, the solubility (log S_0) was shown to decrease, despite the increase in Z_{Mb} during the early stage of the evolution (Fig. 4a and Table S5), and the molecular repulsion increased even with no

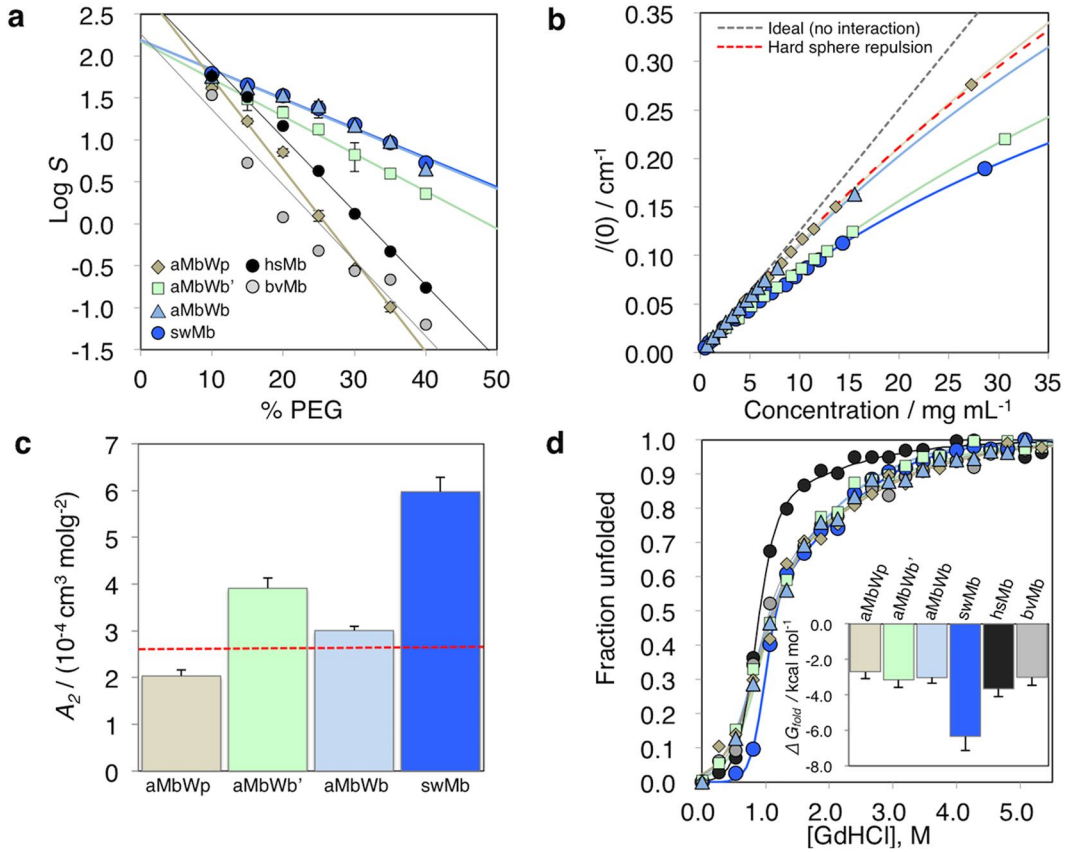


Figure 4. Experimental analyses of ancient and extant Mbs. The values of aMbWp, aMbWb', aMbWb, and swMb are indicated in diamonds (light brown), squares (light green), triangles (light blue), and circles (blue). The data for horse (hsMb) and cow (bvMb) apoMbs are indicated by black and gray circles, respectively, for comparison. (**a**) Solubility dependence of holo-forms of ancient and sperm whale Mbs on the concentration of PEG-6000. Log solubility values in mg/mL (S) are plotted against the precipitant concentration. (**b**) Dependence of the absolute scattering intensity at $q = 0$ on the protein concentration in small angle X-ray scattering experiments (see also Fig. S6). (**c**) Estimated second virial coefficients (A_2) indicating repulsive interaction between Mb molecules. A_2 calculated presuming that Mb is a hard sphere is shown as a red, dashed line. (**d**) Chemical denaturation profiles of apoMbs. The unfolded fractions estimated by the CD signal intensity at 222 nm were plotted against the Gd-HCl concentration. The inset shows the ΔG_{fold} (kcal mol^{-1}) of the proteins. The data for hsMb and bvMb are presented in black and gray, respectively, for comparison.

obvious increase in Z_{Mb} during the later stage of the evolution (Fig. 4c). It is remarkable that the positive charges on the protein surface provide only a small contribution toward increasing the solubility, and even decrease it in some cases[35,37–39]. The effect of repulsion between the Mb molecules due to the acquired positive charges appears to be largely compensated by the decrease in the solubility of single molecules.

However, the increase in the parameter β, demonstrated by the PEG sedimentation experiments, would significantly contribute toward increasing the Mb concentration (Fig. 4a), therefore suggesting that the Z_{Mb} increase might be a strategy to prevent sedimentation-induced interactions with precipitant molecules other than self-aggregation of cognate molecules. The inside of living cells, including myocytes, is crowded by high concentrations of metabolites and biomolecules[40–42]; thus, those molecular crowders are potential precipitants for Mbs and could be mimicked by PEG in the sedimentation experiments. Therefore, the present results require a revision of the physiological effects of the Z_{Mb} increase in the Mb evolution[11,21]. In general, the effect of polymer precipitants, such as PEG, has been thought to arise from the depletion force, which depends on the volume of a protein excluding the polymer precipitants, i.e., the excluded volume[43]. However, in the present case, the excluded volumes of the extant and ancient Mbs are almost identical, and thus the enhancement of the precipitant tolerance should not be attributed to the depletion force but to the changes in protein surface properties, including Z_{Mb}, during the evolution.

The analyses also revealed that the adaptation *via* the Z_{Mb}-increase strategy reached a plateau at the last common ancestor of whales (aMbWb' or aMbWb). However, not all of the offspring of the common whale ancestor have highly adapted to deep-sea environments. The maximum diving depth of Mysticeti species is 200~300 m[44]. Among the Odontoceti species, the sperm whale is an 'elite diver', which can dive to over 2000 m in depth, while dolphins do not dive so deep. Therefore, a considerable part of the molecular adaptation should also be observed between aMbWb' and swMb, which suggests that the Z_{Mb} increase was not the sole strategy in the whale Mb evolution.

It is also hypothesized that thermodynamic stabilization might contribute to maintaining the higher Mb concentration in the myocytes. Dasmeh *et al*. computationally predicted that the folding stabilities of extant whale Mbs were higher than those of ancient whale Mbs, based on 3D modeling with their inferred amino acid sequences and extant Mb structures[23]. Olson and coworkers demonstrated that the *in vivo* expression of mammalian Mb is governed by the apoMb stability, since the rate of aggregation of unfolded apoMb is significantly higher than that of holoMb, and thus a larger fraction of folded apoMb that can immediately bind heme is the key for the high-level expression of holoMb[24]. We performed chemical denaturation experiments of apoMbs with guanidine hydrochloride (GdHCl), as shown in Fig. 4d. The denaturation profiles were analyzed by assuming the three-state transition[45–47], and the thermodynamic parameters were estimated for the folding reactions (Table S6). The folding free energy changes from the intermediate state to the native state (ΔG_1) and the total free energy changes from the unfolded to the native (ΔG_{fold}) slightly decreased from aMbWp to aMbWb, whereas they significantly decreased from aMbWb to swMb. These results indicate that the fold stability mainly improved during the late phase of the evolution.

The stabilization mechanism was clearly observed in the Mb crystal structures. No residue insertion/deletion was anticipated during the evolution from aMbWp to swMb, and all of the changes in the molecular properties should arise from the side-chain replacements. On the crystal structures of aMbs, the residues replaced during the evolution were mostly localized on the molecular surface (Fig. 3b). The backbone structures, as well as the positions and conformations of the bound heme, did not substantially change. Most of the residues that were replaced between the ancestral and extant Mbs were suggested to increase the fold stability, and were found to have introduced additional interactions in the crystal structures (Fig. 3c–e). From aMbWp to aMbWb', V13I filled a cavity in the hydrophobic core, T34K added a hydrogen bond, and K118R (interacting with another substitution, E27D) introduced an electrostatic interaction on the molecular surface (Fig. 3c). Although the side chains of G1V and G15A from aMbWb' to aMbWb did not appear to be interacting with other residues, their introduction might increase the rigidity of the main chain conformation (Fig. 3d). V28I contributed to cavity filling, while D4E, N12H, K45R, and D109E enhanced the electrostatic interactions from aMbWb to swMb (Fig. 3e). Most of the additional interactions are formed by increasing the reach of the side chains. Consistently, the molecular weight (*Mr*) increased during the evolution from aMbWp to swMb. The structure stabilization with the replaced side chains was also evaluated with the structure-based computational method (Table S3 and Fig. S5). The contributions of each residue replacement to the total stability were consistent with the experimental results; *i.e*., V28I, G1V, V13I, K118R, and G15A were among the highest-contributing replacements to the fold stability.

Finally, the major biological function of Mbs, oxygen binding, was examined. The fractional occupancy of ferrous Mb by O_2 in a buffer solution was measured under equilibrium conditions with varied partial pressures of oxygen (pO_2), and the oxygen affinity was analyzed by a Hill plot (Fig. 5). The values of P_{50}, pO_2, at which 50% of Mb are filled, were 0.42, 0.46, 0.46, and 0.52 mmHg for aMbWp, aMbWb, aMbWb', and swMb, respectively. The slight increase in P_{50} values might contribute to an elevated supply of O_2 in the deep-sea environment, since Mbs with these P_{50} values are saturated with O_2 under atmospheric pressure at the sea surface, and Mbs with higher P_{50} values can be fully depleted of the bound O_2 under hypoxic conditions. Two of the mutations from aMbWb to swMb, namely V28I and K45R, were located close to the heme moiety (Fig. 3e). R45 formed a salt bridge with heme-6-propionate at the O_2 entrance, and I28 was located in the distal pocket for heme ligands. The increase in side chain volume at position 28, which eliminated the empty volume of the distal pocket as V28I mutation did, was shown to have remarkable effect in reducing ligand entry rate[48], and might explain the slight increase in P_{50} value from aMbWb to swMb.

The results, however, demonstrated that the oxygen affinity of Mb was not significantly changed during the adaptation process. The P_{50} values for extant animals were reported previously, and indicated that deep-sea and land animal Mbs had similar O_2-binding affinities[49], which was consistent with the present results. This seems to be reasonable because whales adsorb O_2 from the air on the sea surface, where pO_2 condition is similar to that for land animals, and enhancement of the O_2 affinity is not beneficial for the deep-sea adaptation. In order to increase a total supply of O_2 from Mbs, the most reasonable strategy should be increasing the Mb concentrations in myocytes. Thus, the Mb adaptation was incarnated by the enhancements of precipitant resistance and thermodynamic stability. This is strongly supported by the values of precipitant resistance (β) and thermodynamic stability (ΔG_{fold}) of extant land animal Mbs, which are similar to those of aMbWp, as shown in Fig. 4a,d, Tables S5 and S6 as follows.

In order to verify the significance of the observed molecular properties of ancestral Mbs to the deep-sea adaptation, the PEG sedimentation and chemical denaturation experiments were also performed for extant horse and cow Mbs, and compared with those for swMb and ancestral whale Mbs (Fig. 4a,d, Tables S5 and S6). The precipitant tolerance of these land animal Mbs were similar to the terrestrial ancestral whale Mb (aMbWp), whereas their stabilities were similar to the early ancestral whale Mbs (aMWb' and aMbWb). These results showed that the precipitant tolerance was not improved at all, and the stability was only slightly improved during the land animal evolution, indicating the relevance of the improved stability and precipitant tolerance to the deep-sea adaptation.

The theoretical and experimental results obtained in this study are summarized in Fig. 6. The present results demonstrated that the deep-sea adaptation of whale Mb should be divided into early and late phases. In the early phase, the Mb solubility in crowded intracellular conditions significantly improved with other property changes indicated in Fig. 6 to obtain the precipitant tolerance, but the single-molecule solvation free energy increased. The changes in the isoelectric point (p*I*), Z_{Mb}, log S_0, β, and the solvation free energy (ΔG_{solv}) are highly correlated (Fig. 6k). The *Mr*, the second virial coefficient (A_2), the mutational folding energy changes ($\Delta\Delta G_{mut}$), and the fold stability (ΔG_{fold}) form another high-correlation cluster. These findings implied that the short-range repulsive forces between Mbs were mainly acquired during the late phase of evolution to increase the total Mb concentration.

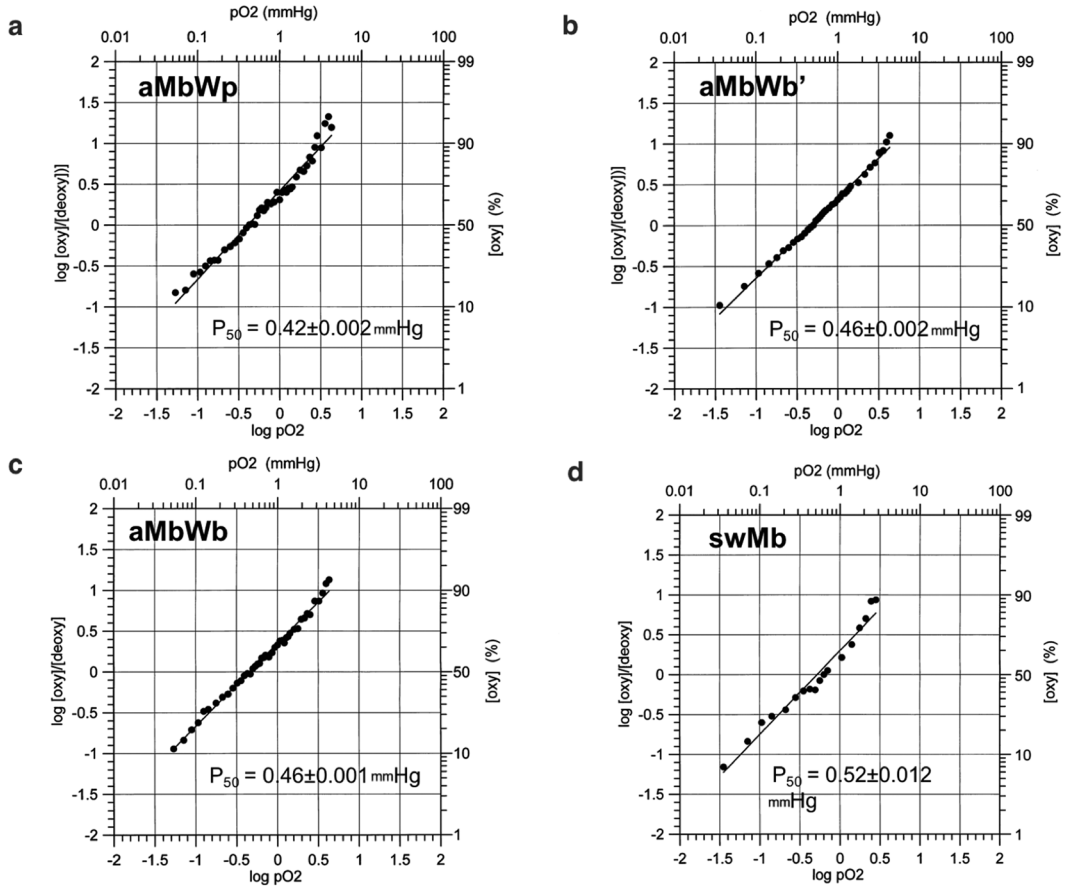
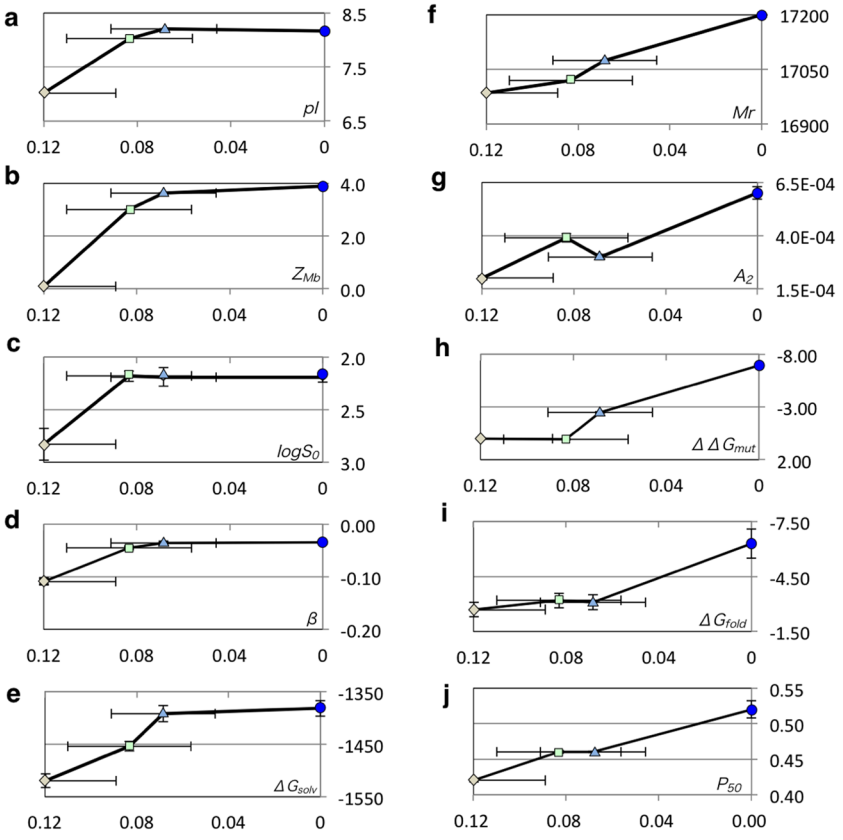


Figure 5. Oxygen equilibrium curves of ancient and sperm whale Mbs. The logarithmic ratios of oxy-myoglobin to deoxy-myoglobin are plotted against the logarithmic oxygen partial pressure (pO_2) for (**a**) aMbWp, (**b**) aMbWb', (**c**) aMbWb, and (**d**) swMb.

Since the enhancement of repulsion between Mb molecules is not correlated with the Z_{Mb} increase, the present results require another explanation for the molecular repulsion between Mbs, which might be related to the folding stability. Although the initiation mechanism of protein aggregation is not fully understood yet, the interactions between the hydrophobic interiors of proteins, which are exposed due to partial unfolding, are considered to play a crucial role[50,51]. Thus, the folding stability would contribute toward preventing occasional aggregation upon molecular collision, and reduce the apparent attractive force between molecules. Dasmeh *et al*. demonstrated that the stabilization strategy was utilized mainly in the early phase[23]. However, as the second virial coefficient and the fold stability showed higher correlations with the evolutionary distance (from extant swMb to ancestral Mbs), this strategy would not be limited to the early phase and would have been adopted in both the early and late phases. The increase of repulsive force during the late phase should be due to a decrease in attractive interaction between Mb molecules at a contact distance, which is caused by the stability increase and prevents the aggregation.

Interestingly, the increase in molecular weight (*Mr*) also correlated with the second virial coefficient and the folding stability. Actually, the Mb structures revealed that most of the additional interactions are formed by replacing amino acids with larger ones. Another possible strategy that could correlate the *Mr* increase to the molecular repulsion might be the 'surface entropy increment'. Surface entropy reduction engineering is used to crystallize proteins, or to improve the quality of crystals, by replacing flexible amino acids on the protein surface, such as arginine or lysine, with amino acids with smaller side chains, such as alanine[52–54]. It is possible that the acquired positively charged residues, along with the others replaced during evolution, introduced an opposite effect that prevents the self-association of Mbs by increasing the flexibility of the residues on the molecular surface. Consistent with this hypothesis, some of the replaced side chains that were involved in the acquired interactions; namely, D4E (interacting with K79), K45R (interacting with D60), and D109E (interacting with R31), from aMbWb to swMb were found to adopt alternative conformations (Fig. 3e).

In conclusion, a total of 17 residues were replaced on or near the protein surface of Mb, during the evolution from the terrestrial ancestor to deep-diving extant whales through the intermediate ancestor. The time range of the early phase of evolution from the terrestrial (aMbWp) to the intermediate (aMbWb or aMbWb') is assumed to be ~10 M years, from the early Eocene to the middle Eocene, and that of the late phase from the intermediate to the sperm whale (swMb) is estimated to be approximately 40 M years from the middle Eocene to the present. Thus, the Z_{Mb} and the precipitant tolerance (β) had evolved first and rapidly by the nine residue replacements,



k

	pI	Z_{Mb}	$logS_0$	β	ΔG_{solv}	*d*	*Mr*	A_2	$\Delta\Delta G_{mut}$	ΔG_{fold}	P_{50}
Isoelectric point (*pI*)	1.00	0.99	-0.99	1.00	0.92	-0.73	0.65	0.67	-0.54	-0.48	0.16
Positive net charge (Z_{Mb})	0.99	1.00	-0.97	1.00	0.95	-0.80	0.74	0.74	-0.64	-0.57	0.28
Log of solubility ($logS_0$)	-0.99	-0.97	1.00	-0.99	-0.86	0.68	-0.59	-0.67	0.47	0.44	-0.07
Solubility slope (β)	1.00	1.00	-0.99	1.00	0.92	-0.75	0.68	0.70	-0.57	-0.51	0.19
Solvation free energy (ΔG_{solv})	0.92	0.95	-0.86	0.92	1.00	-0.86	0.85	0.71	-0.79	-0.65	0.51
Evolutionary distance (*d*)	-0.73	-0.80	0.68	-0.75	-0.86	1.00	-0.99	-0.94	0.95	0.94	-0.99
Relative molecular mass (*Mr*)	0.65	0.74	-0.59	0.68	0.85	-0.99	1.00	0.89	-0.99	-0.94	0.96
Second virial coefficient (A_2)	0.67	0.74	-0.67	0.70	0.71	-0.94	0.89	1.00	-0.83	-0.94	0.98
Fold stability changes ($\Delta\Delta G_{mut}$)	-0.54	-0.64	0.47	-0.57	-0.79	0.95	-0.99	-0.83	1.00	0.94	-0.91
Fold stability(ΔG_{fold})	-0.48	-0.57	0.44	-0.51	-0.65	0.94	-0.94	-0.94	0.94	1.00	-0.94
p50 mmHg (P_{50})	0.16	0.28	-0.07	0.19	0.51	-0.99	0.96	0.98	-0.91	-0.94	1.00

Figure 6. Molecular properties of ancestral and extant Mbs. The values of aMbWp, aMbWb', aMbWb, and swMb are indicated by diamonds (light brown), squares (light green), triangles (light blue), and circles (blue). The horizontal axis shows the evolutionary distance (*d*) of each Mb sequence from that of extant swMb, and the data points are those of aMbWp, aMbWb', aMbWb, and swMb from left to right. (**a**) Isoelectric point (p*I*), (**b**) positive net charge (Z_{Mb}), (**c**) log of solubility (log S_0), (**d**) solubility slope against precipitant (β), (**e**) solvation free energy (ΔG_{solv}), (**f**) relative molecular mass (*Mr*), (**g**) second virial coefficient (A_2), (**h**) mutational folding energy changes ($\Delta\Delta G_{mut}$), (**i**) folding free energy (ΔG_{fold}), and (**j**) half-saturation oxygen pressure (P_{50}). (**k**) Correlation coefficients between the values shown in panels (**a**–**j**) and evolutional distance. The highly positive (>0.9) and negative (<−0.9) coefficients are meshed in magenta and blue, respectively. The two clusters of molecular properties correspond to the early (from terrestrial to ancestral whale) and late (from ancestral whale to extant whale) evolutionary phases. The illustrations of animals are not covered by the CC BY license. Credit to Satoshi Kawasaki. All rights reserved, used with permission.

whereas the repulsive interaction had evolved subsequently and rather slowly by the ten residue replacements. The correlation between Z_{Mb} and β is evident, but still it is not clear whether the Z_{Mb} increase alone could sufficiently explain the β increase or not. The physicochemical mechanism for the β enhancement is an important subject to be tackled in the future.

Resurrections of ancient proteins enable the investigation into a process of protein evolution during a particular period. This is often difficult by simply comparing the proteins of existing species, because they independently accumulate both neutral and non-neutral mutations after the particular evolutionary process, which potentially disturb the analyses. In the present study, the deep-sea adaptation process of Mb was dissected into early and late phases by resurrecting ancestral whale Mbs. The present results, however, also raise the question: why was the Z_{Mb}-increasing strategy limited to the early phase, while the repulsion strategy was continued in the late phase. One possible explanation is that the number of positive charges that could be introduced on the relatively small Mb molecule without damaging its structural integrity is limited, and it already reached the upper-limit for whale Mbs. Another possibility is that the molecular adaptations in diving-capabilities to shallow and deep-sea depths are inevitably different. The Z_{Mb}-increasing strategy has been shown to be widely adopted among shallow-diving animals, including water shrew, beaver, and platypus[11,21]. Comparisons of the Mb adaptations in these animals with those of whales, by applying the ancient protein resurrection of the present study to their Mbs, will be interesting and required to answer this question.

Materials and Methods

Prediction of ancestral Mb sequences. The set of amino acid sequences, including 266 Mbs, 2,179 Hbs, and 31 other globins, was retrieved from the Genbank, Refseq, and UniProt databases with BLAST by using swMb sequence as a query[55–57]. The sequences were aligned by using ClustalW, and manually refined with the XCED program[58,59]. The topology of the phylogenetic tree was inferred with the neighbor-joining (NJ) method on the JTT matrix[60,61], and manually refined by referring to the literatures[21,62,63] (Fig. S2a).

The phylogeny and the alignment were used to infer the ancestral sequences with the PMAL application[64–66] (Figs 1, S2b and S2c). The last common ancestor of whales is thought to be a species of *Basilosaurus*, dated back to ~37 Mya[27]. The extant whales are classified into the morphologically and physiologically distinct Odontoceti (toothed whales) and Mysticeti (baleen whales) classes, and the monophyly of Odontoceti has been a longstanding problem[28,29,67–70]. Therefore, the whale ancestral sequences were inferred based on both the polyphyly and the monophyly hypotheses. The corresponding ancestral sequences in the former and the latter hypotheses were called aMbWb' and aMbWb, respectively (Fig. 1). The last four-footed land ancestor of the whales is assumed to be a *Pakicetus* species dated back to ~53 Mya, and the corresponding ancestral sequence was called aMbWp[71,72]. The correlation between fossil dates and the evolutionary distances, and the distributions of the posterior probabilities of the sites were verified for the ancestral Mbs (Fig. 2 and Table S2.).

Protein synthesis and purification. The Mbs (aMbWp, aMbWb', aMbWb, and swMb) were synthesized from artificial genes. *E. coli* strain BL21 (DE3) was transformed with the vector DNA harboring each Mb gene, and the recombinant protein was synthesized by expression. The proteins were purified from the extracts by Ni Sepharose 6 Fast Flow resin (GE Healthcare) chromatography. The His-tagged Mb was digested with thrombin, and the tag was removed by a His GraviTrap mini column (GE Healthcare). The protein was finally purified by size-exclusion chromatography. The protein identities were verified by matrix assisted laser-desorption time of flight (MALDI-TOF) mass spectrometry by using an AXIMA-CFR plus mass spectrometer (Shimadzu).

For the apoMbs, the genes were subcloned into the expression vector pRSET-C (Invitrogen), and the proteins were synthesized and harvested as mentioned above. The proteins were mainly expressed as inclusion bodies, and extracted from the insoluble fraction with 6 M urea by centrifugation. After dialysis against 0.1% trifluoroacetic acid, the proteins were purified from the extracts by reversed phase HPLC with an Inertsil WP300 C18 column (GL Science). The proteins were finally purified by size-exclusion chromatography.

Oxygen binding analyses. The purified ancestral Mbs (aMbWp, aMbWb', aMbWb) and the recombinant sperm whale Mb (swMb) were prepared in 0.1 M sodium phosphate buffer (pH 7.0), containing the mixture of reducing enzymes and substrates of the heme reduction system, which changed the proteins' heme state from ferric (met-form) to ferrous (oxy-form) overnight at 25 °C[73]. The heme concentrations were estimated from the absorption spectra between 700 and 400 nm, using the ratio of absorbance at 409 nm vs. 280 nm for the met-form, and the absorbances at 542 and 581 nm of the molar extinction coefficients of horse Mb for the oxy-form[74].

The deoxy Mbs (60 μM) were obtained by degasification of oxyMb samples placed in a tonometer (230 mL volume) attached to a glass cell and sealed with a high pressure rubber cap, until the absorbance at 562 nm became stable. Then, small amounts of air were accurately inserted into the tonometer by graduation from 50 to 500 μL, and gently mixed for equilibration for 30 s. The changes in spectra between 700 and 400 nm and the absorbances at 562 nm of the deoxy Mbs were monitored with an Agilent 8453 ultraviolet-visible spectrophotometer at 25 °C until convergence. The Hill plots of oxygen equilibrium curves (OEC) of the Mbs were obtained by plotting the logarithmic oxygen saturation ratio log([oxy]/[deoxy]) against logarithmic oxygen partial pressure pO_2. The P_{50} values of the Mbs were calculated with the cumulative standard deviation of the pO_2, due to the use of the tonometer (Fig. 5).

Crystal structure analyses. The crystal structures of the Mbs were determined by X-ray crystallography. The swMb and aMbWb crystals were grown by the batch method in a 76% saturated ammonium sulfate solution containing 6.5% (*w*/*v*) swMb or 2% (*w*/*v*) aMbWb[75]. The crystals of aMbWb' and aMbWp were obtained by the hanging drop vapor diffusion method, under initial conditions using 0.1 M MIB buffer (pH 9.0) containing 25% (*w*/*v*) PEG1500 for a reservoir for aMbWb', and 3.5 M ammonium sulfate solution for aMbWp. The hanging drops were prepared by mixing the 2 μL of the reservoir solution and 2 μL of the 2% (*w*/*v*) protein solution.

X-ray diffraction data were collected under cryogenic conditions, with a CCD detector Quantum315 (ADSC) at BL38B1 or MX225 (Rayonix) at BL26B2 in SPring-8 (Hyogo, Japan). The diffraction images were processed

with the MOSFLM program[76,77]. The crystal structures were solved by the molecular replacement method, using the Phaser-MR application of PHENIX or MOLREP of the CCP4 suites[78,79], and refined by using COOT and the phenix.refine application of PHENIX[78,80] (Table S4). The atomic coordinates and structure factors of aMbWp, aMbWb', imidazole-ligated aMbWb', aMbWb, and swMb have been deposited in the Protein Data Bank, with the accession codes 5YCG, 5YCI, 5YCJ, 5YCH, and 5YCE, respectively. The molecular graphics were prepared with CHIMERA[81].

Molecular dynamics simulations. The molecular dynamics (MD) simulations of the Mbs were performed with the AMBER12 package with the force field parameter for an O_2 ligand heme, along with the force field ff99SB for proteins[82–84]. The Mb crystal structures determined in this study and the same structures excluding the heme moieties were used as the starting structures for the holo and apo simulations, respectively. The solvent was explicitly considered with a truncated octahedral box of a TIP3P water model equilibrated at 298 K, with periodic boundary conditions based on the particle-mesh Ewald method[82,85]. A ligand O_2 molecule, which was absent from the crystal structures, was included in the system to make it compatible with the force field parameters used for the heme. The His residues 24, 81 and 93 in the Nδ1 (HID form), 12, 48, 64, 82, 97, 113, 116 and 119 in the Nε2 (HIE form), and 36 in the Nδ1 and Nε2 (positively charged, HIP form) were protonated. Na^+ or Cl^- ions were added to obtain a neutral simulation system.

The energy minimization was first performed for 1,000 cycles by restraining the Mb heavy atoms to the original positions, which was followed by additional 2,500 cycles without restraint. Then heating from 0 to 298 K was performed for 20-ps by using the heat bath coupling algorithm by constraining the atom positions[86]. The system was then subjected to an 80-ps NTP ensemble MD calculation with the same restraints at constant temperature (298 K) and constant pressure (1 bar). After releasing all of the restraints, a 60-ns NTP ensemble MD calculation with the same controls was performed, and the last 50-ns trajectories sampled every 5 ps were used for the data analyses. The radius of gyration, the root mean square deviation (RMSD) from the starting structure, the number of hydrogen bonds between Mb and water, and the conformational energy were calculated by using the tools in the AMBER package.

Solvation free energy calculation. The solvation free energy (SFE) calculations for the Mbs were performed with a reference-modified density functional theory (RMDFT)[36,87,88]. The site-density distribution functions of water around the Mbs were calculated by using the three-dimensional reference-interaction-site-model (3D-RISM) integral equation with the Kovalenko–Hirata (KH) closure[89]. The site-site direct correlation functions for bulk water and for the reference hard-sphere fluid were calculated using the one-dimensional (1D)-RISM integral equation with the KH closure[89] and the effective-density approximation (EDA)[90], respectively. For the 1D-RISM and EDA calculations, 0.00125 Å and 32,768 were employed as the grid spacing and the number of grids, respectively. The number density of water and the temperature were 0.033329 molecule/Å^3 and 298 K, respectively. The 3D-RISM integral equations were solved for a grid of 256^3 points in an 80 Å^3 cubic cell, using graphics processing units (GPUs)[91]. The SFE calculation was performed for 5,000 conformations of each Mb from the MD simulations based on the equation:

$$\Delta G_{slov} = \langle \Delta G_i \rangle - k_B T \ln\langle \exp[-(\Delta G_i - \langle \Delta G_i \rangle)/k_B T] \rangle, \quad (1)$$

where $\langle\ \rangle$ indicates the ensemble average over the conformations, ΔG_i is the SFE for each conformation, k_B is the Boltzmann constant, and T is the temperature. The first term in Eq. (1) provides the simple average of ΔG_i and the second term yields the fluctuation effect on ΔG_{slov} due to the conformation fluctuation.

PEG sedimentation analyses. The dependence of the Mb solubility on a precipitating agent was measured with PEG-6000. The 1.8~4.6 mM purified holo-Mb solutions in 100 mM HEPES-NaOH and 10~40% PEG-6000 were prepared. The solutions were incubated at approximately 25 °C for 2 hours, and were then centrifuged to remove the precipitates. The Mb concentration (measured solubility S) in the supernatant was determined with the Nano Drop spectrometer by using an ε_{409nm} of 157000 M^{-1} cm^{-1} [92]. The relationship between the protein solubility and the precipitant concentration was analyzed by assuming

$$\text{Log } S = \text{ Log } S_0 + \beta[\text{precipitant}] \quad (2)$$

where S_0 and β are the solubility in the absence of precipitant and the dependence of the solubility on the precipitant concentration (Fig. 4a)[35].

Small Angle X-ray Scattering. The small angle X-ray scattering (SAXS) experiments were performed at the beam line BL-10C, in the Photon Factory (PF) of the High Energy Accelerator Research Organization (KEK), Tsukuba, Japan. The purified Mb solutions were dialyzed against a 2 mM HEPES–NaOH buffer solution (pH 6.8) at 4 °C for one day. The dialyzed Mb solutions were concentrated to ~3–5 mM, and then centrifuged to remove the precipitate. The Mb solutions were diluted to the desired concentrations at a pH of 6.9 ± 0.1, and irradiated with X-ray wavelength $\lambda = 0.15$ nm (camera length of 1 m) for 2 s in a cell with quartz windows using a sample-flow system (~14.5 μL/min) at 20 ± 0.1 °C. X-ray intensities were recorded by a PILATUS3 2 M detector (DECTRIS Ltd., Switzerland). A total of 30 images were collected for each condition, and the circular 1D averaging of the images was performed with the program *Nika*[93].

The scattering parameter $q = |\boldsymbol{q}| = 4\pi\sin\theta/\lambda$, where $\boldsymbol{q}$ is the scattering vector and 2θ is the X-ray scattering angle, available in this experiment was 0.01–0.55 Å^{-1}. The scattering intensity was corrected by the intensity of the incident light and the transmittance of the X-rays. The absolute scattering intensity of the protein ($I(q)$) was determined as

$$I(q) = [I_S(q) - (1 - c_p v)I_B(q)]/f \quad (3)$$

where c_p is the protein concentration (g/cm^3), v is the specific volume of the solute (0.7425 cm^3/g), and f is the correction factor to convert the observed intensity in arbitrary units to the absolute intensity in units of cm^{-1}, respectively.

The absolute scattering intensity of the protein ($I(q)$) was extrapolated to the absolute scattering intensity $I(0)$ at $q = 0$. The $I(0)$ is related to the second virial coefficient A_2 as

$$I(0) = kMc_p/(1 + 2A_2Mc_p), \quad (4)$$

where M is the molecular weight of the protein and the k value is equal to $v^2(\rho_m - \rho_{solv})^2/N_A$. N_A is Avogadro's number, and $\rho_m - \rho_{solv}$ is the electron density difference between the protein and the solvent (2.8×10^{10} cm^{-2}, typically)[94,95].

Folding stability analyses. The stability of Mb was determined by chemical denaturation experiments with guanidine hydrochloride (Gd-HCl) by monitoring the circular-dichroism (CD) signal intensity at 222 nm for 5 μM proteins in a buffer solution containing 50 mM HEPES-NaOH (pH 7). The denaturation data were analyzed using a theoretical curve derived from the three state transition model[45]:

$$F \overset{K_1}{\Longleftrightarrow} I \overset{K_2}{\Longleftrightarrow} U \quad (5)$$

where F, I and U represent the folded, intermediate and unfolded states, respectively; and $K_1 = [F]/[I]$ and $K_2 = [I]/[U]$ are the equilibrium constants of F ⇔ I and I ⇔ U, respectively.

K_1 and K_2 give ΔG_1 and ΔG_2, the free energy of the folded state relative to the intermediate and that of the intermediate relative to the unfolded state, respectively, as follows:

$$\Delta G_1 = G_F - G_I = -RT \ln K_1 = \Delta G^\circ_1 + m_1 x \quad (6)$$

$$\Delta G_2 = G_I - G_U = -RT \ln K_2 = \Delta G^\circ_2 + m_2 x \quad (7)$$

where ΔG°_1 and ΔG°_2 are the ΔG_1 and ΔG_2 values in the absence of denaturant, respectively, and m_1 and m_2 are the dependences of ΔG_1 and ΔG_2 on x, the denaturant concentration, respectively. From these relationships, the following formulas were obtained:

$$\alpha = 1/\{1 + \exp A + \exp(-B)\} \quad (8)$$

$$\beta = \exp(-B)/\{1 + \exp A + \exp(-B)\} \quad (9)$$

where α and β are the fractions of the intermediate and the unfolded state, respectively, and $A = -(\Delta G^\circ_1 + m_1 x)/RT$; $B = -(\Delta G^\circ_2 + m_2 x)/RT$. Accordingly, the ratio of the helical content in the transition region per the total helical content of the folded form (y), is calculated as

$$y = 1 - \alpha - \beta + \gamma a = (\gamma + \exp A)/\{1 + \exp A + \exp(-B)\} \quad (10)$$

where γ is the ratio of the helical content of the intermediate per that of the folded state by assuming the helical content of the unfolded state to be zero[46,47]. The theoretical curves derived from Eq (10) were fitted to the observed denaturation data to obtain the thermodynamic parameters ΔG°_1, ΔG°_2, m_1 and m_2. The sum of $\Delta G_1{}^\circ$ and ΔG°_2 ($\Delta G^\circ_{1+2} = \Delta G^\circ_1 + \Delta G^\circ_2$) gives the free energy changes from the unfolded state to the folded state.

Accession Numbers. The atomic coordinates and structure factors have been deposited in the Protein Data Bank, www.wwpdb.org (PDB codes 5YCG, 5YCI, 5YCJ, 5YCH, and 5YCE).

References

1. Thomson, J. M. *et al*. Resurrecting ancestral alcohol dehydrogenases from yeast. *Nat Genet* **37**, 630–635, https://doi.org/10.1038/ng1553 (2005).
2. Kratzer, J. T. *et al*. Evolutionary history and metabolic insights of ancient mammalian uricases. *Proc Natl Acad Sci USA* **111**, 3763–3768, https://doi.org/10.1073/pnas.1320393111 (2014).
3. Ugalde, J. A., Chang, B. S. & Matz, M. V. Evolution of coral pigments recreated. *Science* **305**, 1433, https://doi.org/10.1126/science.1099597 (2004).
4. Konno, A., Kitagawa, A., Watanabe, M., Ogawa, T. & Shirai, T. Tracing protein evolution through ancestral structures of fish galectin. *Structure* **19**, 711–721, https://doi.org/10.1016/j.str.2011.02.014 (2011).
5. Risso, V. A. *et al*. *De novo* active sites for resurrected Precambrian enzymes. *Nat Commun* **8**, 16113, https://doi.org/10.1038/ncomms16113 (2017).
6. Gingerich, P. D. Land-to-sea transition in early whales: evolution of Eocene Archaeoceti (Cetacea) in relation to skeletal proportions and locomotion of living semiaquatic mammals. *Paleobiology* **29**, 429–454, https://doi.org/10.1666/0094-8373 (2003).
7. Bajpai, S., Thewissen, J. G. & Sahni, A. The origin and early evolution of whales: macroevolution documented on the Indian subcontinent. *J Biosci* **34**, 673–686, https://doi.org/10.1007/s12038-009-0060-0 (2009).
8. Snyder, G. K. Respiratory adaptations in diving mammals. *Respir Physiol* **54**, 269–294, https://doi.org/10.1016/0034-5687(83)90072-5 (1983).
9. Kooyman, G. L. *Diverse Divers: Physiology and behavior*. (Springer 1989).
10. Davis, R. W. A review of the multi-level adaptations for maximizing aerobic dive duration in marine mammals: from biochemistry to behavior. *J Comp Physiol B* **184**, 23–53, https://doi.org/10.1007/s00360-013-0782-z (2014).

11. McGowen, M. R., Gatesy, J. & Wildman, D. E. Molecular evolution tracks macroevolutionary transitions in Cetacea. *Trends Ecol Evol* **29**, 336–346, https://doi.org/10.1016/j.tree.2014.04.001 (2014).
12. Lesk, A. M. & Chothia, C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol* **136**, 225–270, https://doi.org/10.1016/0022-2836(80)90373-3 (1980).
13. Fermi, G. & Perutz, M. F. *Atlas of Molecular Structures in Biology: Vol. 2, Haemoglobin and myoglobin*. (Clarendon Press, 1981).
14. Wittenberg, B. A. & Wittenberg, J. B. Transport of oxygen in muscle. *Annu Rev Physiol* **51**, 857–878, https://doi.org/10.1146/annurev.ph.51.030189.004233 (1989).
15. McIntyre, I. W., Campbell, K. L. & MacArthur, R. A. Body oxygen stores, aerobic dive limits and diving behaviour of the star-nosed mole (Condylura cristata) and comparisons with non-aquatic talpids. *J Exp Biol* **205**, 45–54 (2002).
16. Rezende, E. L. *et al*. Maximal oxygen consumption in relation to subordinate traits in lines of house mice selectively bred for high voluntary wheel running. *J Appl Physiol (1985)* **101**, 477–485, https://doi.org/10.1152/japplphysiol.00042.2006 (2006).
17. Natarajan, C. *et al*. Epistasis among adaptive mutations in deer mouse hemoglobin. *Science* **340**, 1324–1327, https://doi.org/10.1126/science.1236862 (2013).
18. Rummer, J. L., McKenzie, D. J., Innocenti, A., Supuran, C. T. & Brauner, C. J. Root effect hemoglobin may have evolved to enhance general tissue oxygen delivery. *Science* **340**, 1327–1329, https://doi.org/10.1126/science.1233692 (2013).
19. Noren, S. R. & Williams, T. M. Body size and skeletal muscle myoglobin of cetaceans: adaptations for maximizing dive duration. *Comp Biochem Physiol A Mol Integr Physiol* **126**, 181–191, https://doi.org/10.1016/S1095-6433(00)00182-3 (2000).
20. Nery, M. F., Arroyo, J. I. & Opazo, J. C. Accelerated evolutionary rate of the myoglobin gene in long-diving whales. *J Mol Evol* **76**, 380–387, https://doi.org/10.1007/s00239-013-9572-1 (2013).
21. Mirceta, S. *et al*. Evolution of mammalian diving capacity traced by myoglobin net surface charge. *Science* **340**, 1234192, https://doi.org/10.1126/science.1234192 (2013).
22. Helbo, S. & Fago, A. Functional properties of myoglobins from five whale species with different diving capacities. *J Exp Biol* **215**, 3403–3410, https://doi.org/10.1242/jeb.073726 (2012).
23. Dasmeh, P., Serohijos, A. W., Kepp, K. P. & Shakhnovich, E. I. Positively selected sites in cetacean myoglobins contribute to protein stability. *PLoS Comput Biol* **9**, e1002929, https://doi.org/10.1371/journal.pcbi.1002929 (2013).
24. Samuel, P. P., Smith, L. P., Phillips, G. N. Jr & Olson, J. S. Apoglobin stability Is the major factor governing both cell-free and *in vivo* expression of holomyoglobin. *J Biol Chem* **290**, 23479–23495, https://doi.org/10.1074/jbc.M115.672204 (2015).
25. Wilkinson, D. L. & Harrison, R. G. Predicting the solubility of recombinant proteins in Escherichia coli. *Biotechnology (N Y)* **9**, 443–448 (1991).
26. Kuroda, Y., Suenaga, A., Sato, Y., Kosuda, S. & Taiji, M. All-atom molecular dynamics analysis of multi-peptide systems reproduces peptide solubility in line with experimental observations. *Sci Rep* **6**, 19479, https://doi.org/10.1038/srep19479 (2016).
27. Gingerich, P. D., Smith, B. H. & Simons, E. L. Hind limbs of Eocene basilosaurus: evidence of feet in whales. *Science* **249**, 154–157, https://doi.org/10.1126/science.249.4965.154 (1990).
28. Nikaido, M. *et al*. Retroposon analysis of major cetacean lineages: the monophyly of toothed whales and the paraphyly of river dolphins. *Proc Natl Acad Sci USA* **98**, 7384–7389, https://doi.org/10.1073/pnas.121139198 (2001).
29. Nikaido, M., Piskurek, O. & Okada, N. Toothed whale monophyly reassessed by SINE insertion analysis: the absence of lineage sorting effects suggests a small population of a common ancestral species. *Mol Phylogenet Evol* **43**, 216–224, https://doi.org/10.1016/j.ympev.2006.08.005 (2007).
30. Gingerich, P. D., Wells, N. A., Russell, D. E. & Shah, S. M. Origin of whales in epicontinental remnant seas: new evidence from the early Eocene of Pakistan. *Science* **220**, 403–406, https://doi.org/10.1126/science.220.4595.403 (1983).
31. Thewissen, J. G., Williams, E. M., Roe, L. J. & Hussain, S. T. Skeletons of terrestrial cetaceans and the relationship of whales to artiodactyls. *Nature* **413**, 277–281, https://doi.org/10.1038/35095005 (2001).
32. Khan, M. A., Islam, M. M. & Kuroda, Y. Analysis of protein aggregation kinetics using short amino acid peptide tags. *Biochim Biophys Acta* **1834**, 2107–2115, https://doi.org/10.1016/j.bbapap.2013.06.013 (2013).
33. Middaugh, C. R., Tisel, W. A., Haire, R. N. & Rosenberg, A. Determination of the apparent thermodynamic activities of saturated protein solutions. *J Biol Chem* **254**, 367–370 (1979).
34. Atha, D. H. & Ingham, K. C. Mechanism of precipitation of proteins by polyethylene glycols. Analysis in terms of excluded volume. *J Biol Chem* **256**, 12108–12117 (1981).
35. Kramer, R. M., Shende, V. R., Motl, N., Pace, C. N. & Scholtz, J. M. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophys J* **102**, 1907–1915, https://doi.org/10.1016/j.bpj.2012.01.060 (2012).
36. Sumi, T., Mitsutake, A. & Maruyama, Y. A solvation-free-energy functional: a reference-modified density functional formulation. *J Comput Chem* **36**, 1359–1369, https://doi.org/10.1002/jcc.23942 (2015).
37. Shaw, K. L., Grimsley, G. R., Yakovlev, G. I., Makarov, A. A. & Pace, C. N. The effect of net charge on the solubility, activity, and stability of ribonuclease Sa. *Protein Sci* **10**, 1206–1215, https://doi.org/10.1110/ps.440101 (2001).
38. Chan, P., Curtis, R. A. & Warwicker, J. Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci Rep* **3**, 3333, https://doi.org/10.1038/srep03333 (2013).
39. Hojgaard, C. *et al*. A soluble, folded protein without charged amino acid residues. *Biochemistry* **55**, 3949–3956, https://doi.org/10.1021/acs.biochem.6b00269 (2016).
40. Ellis, R. J. & Minton, A. P. Protein aggregation in crowded environments. *Biol Chem* **387**, 485–497, https://doi.org/10.1515/BC.2006.064 (2006).
41. Rivas, G. & Minton, A. P. Macromolecular Crowding *In Vitro*, *In Vivo*, and In Between. *Trends Biochem Sci* **41**, 970–981, https://doi.org/10.1016/j.tibs.2016.08.013 (2016).
42. van den Berg, B., Wain, R., Dobson, C. M. & Ellis, R. J. Macromolecular crowding perturbs protein refolding kinetics: implications for folding inside the cell. *EMBO J* **19**, 3870–3875, https://doi.org/10.1093/emboj/19.15.3870 (2000).
43. Wang, Y. *et al*. Quantitative evaluation of colloidal stability of antibody solutions using PEG-induced liquid-liquid phase separation. *Mol Pharm* **11**, 1391–1402, https://doi.org/10.1021/mp400521b (2014).
44. Schreer, J. F. & Kovacs, K. M. Allometry of diving capacity in air-breathing vertebrates. *Can J Zool* **75**, 339–358, https://doi.org/10.1139/z97-044 (1997).
45. Barrick, D. & Baldwin, R. L. Three-state analysis of sperm whale apomyoglobin folding. *Biochemistry* **32**, 3790–3796, https://doi.org/10.1021/bi00065a035 (1993).
46. Isogai, Y., Ishii, A., Fujisawa, T., Ota, M. & Nishikawa, K. Redesign of artificial globins: effects of residue replacements at hydrophobic sites on the structural properties. *Biochemistry* **39**, 5683–5690, https://doi.org/10.1021/bi992687 (2000).
47. Isogai, Y. Native protein sequences are designed to destabilize folding intermediates. *Biochemistry* **45**, 2488–2492, https://doi.org/10.1021/bi0523714 (2006).
48. Scott, E. E., Gibson, Q. H. & Olson, J. S. Mapping the pathways for O2 entry into and exit from myoglobin. *J Biol Chem* **276**, 5177–5188, https://doi.org/10.1074/jbc.M008282200 (2001).
49. Wright, T. J. & Davis, R. W. Myoglobin oxygen affinity in aquatic and terrestrial birds and mammals. *J Exp Biol* **218**, 2180–2189, https://doi.org/10.1242/jeb.119321 (2015).
50. Kelly, J. W. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr Opin Struct Biol* **8**, 101–106 (1998).

51. Dobson, C. M. Protein folding and disease: a view from the first Horizon Symposium. *Nat Rev Drug Discov* **2**, 154–160, https://doi.org/10.1038/nrd1013 (2003).
52. Cooper, D. R. *et al*. Protein crystallization by surface entropy reduction: optimization of the SER strategy. *Acta Crystallogr D Biol Crystallogr* **63**, 636–645, https://doi.org/10.1107/S0907444907010931 (2007).
53. Derewenda, Z. S. & Vekilov, P. G. Entropy and surface engineering in protein crystallization. *Acta Crystallogr D Biol Crystallogr* **62**, 116–124, https://doi.org/10.1107/S0907444905035237 (2006).
54. Derewenda, Z. S. Rational protein crystallization by mutational surface engineering. *Structure* **12**, 529–535, https://doi.org/10.1016/j.str.2004.03.008 (2004).
55. Kaminuma, E. *et al*. DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Res* **38**, D33–38, https://doi.org/10.1093/nar/gkp847 (2010).
56. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2 (1990).
57. O'Leary, N. A. *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–745, https://doi.org/10.1093/nar/gkv1189 (2016).
58. Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* Chapter 2, Unit 2 3, https://doi.org/10.1002/0471250953.bi0203s00 (2002).
59. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066, https://doi.org/10.1093/nar/gkf436 (2002).
60. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425, https://doi.org/10.1093/oxfordjournals.molbev.a040454 (1987).
61. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275–282, https://doi.org/10.1093/bioinformatics/8.3.275 (1992).
62. Meredith, R. W. *et al*. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* **334**, 521–524, https://doi.org/10.1126/science.1211028 (2011).
63. Spaulding, M., O'Leary, M. A. & Gatesy, J. Relationships of Cetacea (Artiodactyla) among mammals: increased taxon sampling alters interpretations of key fossils and character evolution. *PLoS One* **4**, e7062, https://doi.org/10.1371/journal.pone.0007062 (2009).
64. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586–1591, https://doi.org/10.1093/molbev/msm088 (2007).
65. Yang, Z. & Nielsen, R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol* **46**, 409–418, https://doi.org/10.1007/PL00006320 (1998).
66. Yang, Z. & Rannala, B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol* **23**, 212–226, https://doi.org/10.1093/molbev/msj024 (2006).
67. Milinkovitch, M. C., Orti, G. & Meyer, A. Revised phylogeny of whales suggested by mitochondrial ribosomal DNA sequences. *Nature* **361**, 346–348, https://doi.org/10.1038/361346a0 (1993).
68. Milinkovitch, M. C., Orti, G. & Meyer, A. Novel phylogeny of whales revisited but not revised. *Mol Biol Evol* **12**, 518–520, https://doi.org/10.1093/oxfordjournals.molbev.a040226 (1995).
69. Cassens, I. *et al*. Independent adaptation to riverine habitats allowed survival of ancient cetacean lineages. *Proc Natl Acad Sci USA* **97**, 11343–11347, https://doi.org/10.1073/pnas.97.21.11343 (2000).
70. Marx, F. G. & Fordyce, R. E. Baleen boom and bust: a synthesis of mysticete phylogeny, diversity and disparity. *R Soc Open Sci* **2**, 140434, https://doi.org/10.1098/rsos.140434 (2015).
71. Thewissen, J. G., Hussain, S. T. & Arif, M. Fossil evidence for the origin of aquatic locomotion in archaeocete whales. *Science* **263**, 210–212, https://doi.org/10.1126/science.263.5144.210 (1994).
72. Bajpai, S. & Gingerich, P. D. A new Eocene archaeocete (Mammalia, Cetacea) from India and the time of origin of whales. *Proc Natl Acad Sci USA* **95**, 15464–15468, https://doi.org/10.1073/pnas.95.26.15464 (1998).
73. Hayashi, A., Suzuki, T. & Shin, M. An enzymic reduction system for metmyoglobin and methemoglobin, and its application to functional studies of oxygen carriers. *Biochim Biophys Acta* **310**, 309–316, https://doi.org/10.1016/0005-2795(73)90110-4 (1973).
74. Bowen, W. J. The absorption spectra and extinction coefficients of myoglobin. *J Biol Chem* **179**, 235–245 (1949).
75. Watson, H. C. The stereochemistry of the protein myoglobin. *Prog Stereochem* **4**, 299 (1969).
76. Otwinowski, Z. & Minor, W. In *Methods in Enzymology*, *Volume 276: Macromolecular Crystallography*, *partA* Vol. 276 (eds Carter, C. W. Jr. & Sweet, R. M.) 307–326 (1997).
77. Battye, T. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D Biol Crystallogr* **67**, 271–281, https://doi.org/10.1107/S0907444910048675 (2011).
78. Adams, P. D. *et al*. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213–221, https://doi.org/10.1107/S0907444909052925 (2010).
79. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr D Biol Crystallogr* **66**, 22–25, https://doi.org/10.1107/S0907444909052925 (2010).
80. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501, https://doi.org/10.1107/S0907444910007493 (2004).
81. Pettersen, E. F. *et al*. UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612, https://doi.org/10.1002/jcc.20084 (2004).
82. Hornak, V. *et al*. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712–725, https://doi.org/10.1002/prot.21123 (2006).
83. Case, D. A. *et al*. *AMBER 12*. (University of California, 2012).
84. Arcon, J. P., Rosi, P., Petruk, A. A., Marti, M. A. & Estrin, D. A. Molecular mechanism of myoglobin autoxidation: insights from computer simulations. *J Phys Chem B* **119**, 1802–1813, https://doi.org/10.1021/jp5093948 (2015).
85. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* **98**, 10089–10092, https://doi.org/10.1063/1.464397 (1993).
86. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J Chem Phys* **81**, 3684–3690, https://doi.org/10.1063/1.448118 (1984).
87. Sumi, T., Maruyama, Y., Mitsutake, A. & Koga, K. A reference-modified density functional theory: An application to solvation free-energy calculations for a Lennard-Jones solution. *J Chem Phys* **144**, 224104, https://doi.org/10.1063/1.4953191 (2016).
88. Sumi, T., Maruyama, Y., Mitsutake, A., Mochizuki, K. & Koga, K. Application of reference-modified density functional theory: Temperature and pressure dependences of solvation free energy. *J Comput Chem* **39**, 202–217, https://doi.org/10.1002/jcc.25101 (2018).
89. Kovalenko, A. & Hirata, F. Self-consistent description of a metal-water interface by the Kohn-Sham density functional theory and the three-dimensional reference interaction site model. *J Chem Phys* **110**, 10095–10112, https://doi.org/10.1063/1.478883 (1999).
90. Sumi, T. & Sekino, H. A Self-Consistent Density-Functional Approach for Homogeneous and Inhomogeneous Classical Fluids. *Journal of the Physical Society of Japan* **77**, 034605–034605, https://doi.org/10.1143/jpsj.77.034605 (2008).
91. Maruyama, Y. & Hirata, F. Modified anderson method for accelerating 3D-RISM calculations using graphics processing unit. *J Chem Theory Comput* **8**, 3015–3021, https://doi.org/10.1021/ct300355r (2012).
92. Walenta, E. Small angle X-ray scattering. *Acta Polymerica* **36**, 296–296, https://doi.org/10.1002/actp.1985.010360520 (1985).

93. Ilavsky, J. Nika: software for two-dimensional data reduction. *J Appl Crystallogr* **45**, 324–328, https://doi.org/10.1107/S0021889812004037 (2012).
94. Zimm, B. H. The scattering of light and the radial distribution function of high polymer solutions. *J Chem Phys* **16**, 1093–1099, https://doi.org/10.1063/1.1746738 (1948).
95. Goldenberg, D. P. & Argyle, B. Self crowding of globular proteins studied by small-angle x-ray scattering. *Biophys J* **106**, 895–904, https://doi.org/10.1016/j.bpj.2013.12.004 (2014).

Acknowledgements

We thank Drs. Manabu Ishida and Kiyohiro Imai for critical discussions, and Satoshi Kawasaki (http://www.geocities.co.jp/NatureLand/5218/) for kind permission to use his illustrations of animals. X-ray diffraction and SAXS experiments were conducted at SPring-8 and KEK, Japan, respectively, under the approval of the Photon Factory Program Advisory Committee: Proposal No. 2016G032. We thank Dr. Satoshi Shibuta (Chiba Institute of Technology) for his contributions to the SAXS experiment. This work was partly supported by Grants-in-Aid for scientific research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (JP17H01818), and Platform Project for Supporting Drug Discovery and Life Science Research (BINDS) from Japan Agency for Medical Research and Development (AMED) (17am0101111j0001).

Author Contributions

Y.I. and T. Shirai designed research. K.T. and T. Shirai predicted the ancient protein sequences. Y.I. synthesized proteins and performed biochemical analyses. T.N. and A.T. performed O_2-binding analyses. S.N. and T. Shirai performed protein crystallization and 3D structure determination. Y.I., T. Sumi, K.T., and T. Shirai conducted theoretical calculation. H.I. performed SAXS analyses. Y.I. and T. Shirai wrote the main manuscript text. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-34984-6.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.