Manuscript prepared for submission to J. Molecular Evolution

Title: **MEASURING SHIFTS IN FUNCTION AND EVOLUTIONARY OPPORTUNITY USING VARIABILITY PROFILES: A CASE STUDY OF THE GLOBINS**

Gavin J.P. Naylor*,
Department of Zoology and Genetics,
Iowa State University,
Ames,
Iowa 50011
tel: 515-294-1255
fax: 515-294-4257
email: gnaylor@iastate.edu

Mark Gerstein,
Molecular Biophysics and Biochemistry Department ,
Yale University
New Haven,
Connecticut 06520-8114
e-mail: Mark.Gerstein@yale.edu

*corresponding author

Abstract.

Variability profiles measured over a set of aligned sequences can be used to estimate evolutionary freedom to vary. Differences in variability profiles among clades relate evolutionary "shifts in function" to specific residues at the molecular level. We demonstrate such a "shift" across the alpha and beta sub-units of haemoglobin. We also show that the variability profiles for myoglobin are different between whales and primates and speculate that the differences between the two clades may reflect a shift associated with the novel oxygen storage demands in the lineage leading to whales. We discuss the relationship between sequence variability and "evolutionary opportunity" and explore the utility of Maynard Smith's multi-dimensional evolutionary opportunity space metaphor for exploring functional constraints, genetic redundancy, and the context dependency of the genotype-phenotype map. This work has useful implications for quantitatively defining and comparing protein function. Supplementary data is available from bioinfo.mbb.yale.edu/align .

Proteins evolve through amino acid substitution. Some substitutions are neutral, or nearly so, and have little effect on protein function. Others are deleterious and are removed by natural selection. The extent to which substitutions are tolerated varies from site to site and region to region within a protein, and reflects the degree of constraint. A region or site that is tightly constrained is less free to vary than one in which constraints are relaxed. Such differences in "freedom to vary" can be represented using Maynard-Smith's (1970) concept of a "protein sequence space" in which each site in an alignment is represented on its own axis and the number of axes required to represent all conceivable variants for a protein is equal to the number of sites in its sequence. Each sequence occupies a unique point in this space; variants differing at one site are adjacent (Hamming) neighbours. The collection of all viable sequence variants for a particular protein forms a localized interconnected 'neighbourhood' of points within the space. This representation has proved conceptually intuitive and analytically powerful (Vingron & Sibbald, 1993; Vingron & Waterman, 1994; Huynen et al.,1996; Fontana & Schuster, 1998; Bornberg-Bauer and Chan, 1999; Wuchty et al. 1999). In this paper we explore the relationships between sequence variation, protein function, shift in function and the process of evolutionary change within the context of the protein sequence space representation. A number of interesting insights emerge.

In protein sequence space, constraints are reflected in the multidimensional shape of the cluster of points that make up the "neighbourhood" of variants viable for a specific protein. The

boundary defining the edge of this neighbourhood is characteristic of the protein's function and can be thought of as its functional "signature". Any sequence combination falling outside the boundary will fail to function. Over the course of evolution, mutation pressure drives sites that are free to vary to explore the opportunity space available to them. Different lineages explore different parts of this space. Given enough time, a radiation of evolutionary lineages will collectively explore all of the opportunity space available for a particular protein function.

Sequence variants from different points within the opportunity space associated with a particular protein can be obtained by sequencing DNA for that protein from a variety of organisms. If we align a number of such sequence variants and estimate how many evolutionary changes have occurred at each residue in the alignment, we can determine a profile of variability over the alignment. This variability profile reflects the protein's constraints in much the same way as does the shape of its neighbourhood in protein sequence space (Benner et al. 1994). Indeed, the two are related. The variability profile reflects a mutationally directed walk through the neighbourhood and constitutes a historical sampling of the opportunity space. If such variability profiles are truly characteristic of protein function and can serve as identifiers or functional "signatures", then it follows that any shift in function should be reflected by a corresponding shift in signature. We explore this idea by analyzing sequences for the globin superfamily of proteins.

CASE STUDY - THE GLOBINS

Descriptions of the Molecules

Functional haemoglobin is a tetramer made up of 2 alpha and 2 beta globin sub-units. (Fig. 1a) The alpha and beta sub-units share a common ancestry due to ancient gene duplication. The two sub-units have a high degree of sequence similarity and almost identical tertiary structures. Both are box-like and consist of a two-layered sandwich of 8 alpha helices connected by turns (Fig 1b). Some of the functionally important residues are indicated in the figure. When the four sub-units are assembled into functional haemoglobin there is little contact between the 2 alpha chains or between the two beta chains, however there are several contacts between the pairs of unlike chains.

Myoglobin, like haemoglobin, is a member of the globin superfamily of proteins. It has a similar tertiary structure and exhibits a high degree of sequence similarity to the two haemoglobin sub-units. Like haemoglobin, myoglobin is involved in binding oxygen. However,

it differs in that it is used to store oxygen in muscle rather than to transport it through the blood. Furthermore, myoglobin is not allosterically regulated, as is haemoglobin.

Methods

We obtained alpha and beta amino acid sequences for 20 different carnivore species, 16 ungulates and 20 primates (Fig. 2) from public sequence data bases. All 112 (56 x 2) sequences were simultaneously aligned using a combined sequence and structure alignment approach (Gerstein et al., 1994; Gerstein & Altman, 1995; Gerstein & Levitt, 1996, 1998); that is, "key" structures representing fairly divergent sequences were first aligned based on their three-dimensional coordinates. Then sequences were aligned to the structure they were most homologous to. Six data subsets were isolated from the alignment: one Hb alpha and one Hb beta subset for each of the three mammalian groups. We used the computer programs PAUP*4.0 (Swofford 1999) and MacClade (Madison & Madison 1992) to infer using parsimony, the number of substitutions that had occurred at each site, for each of the three sets of taxa (Fig 2). This information was used to generate a profile of inferred variability for each of the 6 data sets (3 sets of taxa for 2 genes). The six resultant variability profiles were then compared using the coefficient of functional divergence Theta (Gu, 1999), which can be interpreted as the loss of rate correlation over sites between two homologous genes, or as the probability of a site being the state of functional constraint shift.

Results for Haemoglobin

Our results indicate that within the alpha sub-unit the variability profile is statistically similar for all three groups. The same is true for the beta sub-unit. However the variability profile for the alpha sub-unit is markedly distinct from that of the beta sub-unit. The coefficient of functional divergence theta (Gu, 1999) between the haemoglobin alpha and beta theta was 0.36, significantly larger than 0 ($p<0.01$). A structural model of haemoglobin coloured to reflect the degree of change at each site was used to visualize the data in its appropriate three-dimensional context (Fig. 3)

As might be expected, and as can be seen in the variability plots, the match among different clades of organisms for the same protein sub-unit is not perfect. This is likely a consequence of the stochasticity of the substitution process and the restricted sampling of taxa (and therefore of the evolutionary opportunity space) used for each data set. Given these draw backs, it is all the more remarkable that such clear-cut differences in variability profiles exist

between the alpha and beta sub-units (fig 2). It would appear from these results that variability signatures may indeed provide a powerful and sensitive way to represent the subtle but important differences in function that exist between closely related proteins - apparently even when they have highly similar tertiary structures as is the case for the globin genes presented.

Results for Myoglobin

We aligned myoglobin sequences for 15 primates and 15 cetaceans and subjected them to the same procedure described above. In contrast to the situation seen for haemoglobins, the variability profiles for myoglobin were noticeably different between the two orders of mammals. (Fig. 4). Fewer sites appear free to vary in the cetaceans than in the primates (29 variable sites for cetaceans, 38 for primates). Interestingly, though the cetaceans appear to have a reduced number of variable sites, those that are variable show a higher incidence of change than is the case for primates (mean for cetaceans: 1.7, std. dev 0.79; for primates 1.3 , std. dev 0.67). These differences are particularly pronounced in the region between sites 127-164 in the G and H helices (Fig 4) where 16 out of the 38 sites show variation in the primates but only 6 of the 38 show variation in the cetaceans. The fact that variable sites in cetaceans are fewer in number yet more prone to change indicates that myoglobin may be differently constrained in primates than it is in cetaceans. It might be argued that the whale and primate myoglobin variability plots differ, not as a result of differential constraints, but as a consequence of different taxon sampling schemes. Perhaps the 15 cetaceans radiated more recently than did the 15 primates? While this could account for the fact that fewer sites show variation in cetaceans than in primates, it cannot account for the increased amount of per site change for those sites that do show variation in the cetaceans. If differences in taxon sampling were the underlying cause for the differences we would expect the clade with the fewest number of variable sites to also show the lowest amounts of per site change (assuming a stochastic model of evolutionary change)

It is enticing to suggest that the difference in variability profiles reflect a shift in function associated with the novel oxygen storage demands of sustained deep diving in cetaceans. Unfortunately we cannot determine from the data whether it is the cetaceans or the primates (or both) that have shifted function. Myoglobin sequences from other orders of mammals will help to clarify this. At present there are insufficient myoglobin sequences in publically available data bases to determine this unequivocally· We emphasize that inferences about functional divergence based on variability profiles cannot substitute for careful comparative assays of the biochemical properties of the gene products such as those reviewed in Romero-Herrera et al. (1978) and Perutz (1983). We wish only to point out that the comparison of variability profiles between

paralogous sequences can provide a powerful, informative initial step toward understanding functional divergence. Information gleaned from such comparisons can be used to guide the choice of functionally important "candidate sites" for subsequent experimental verification using site directed mutagenesis, circumventing the need for random mutagenesis.

DISCUSSION

In the following sections we discuss how variability profiles relate to "evolutionary opportunity" within the protein sequence space representation . We speculate how drift and selection may interact with the underlying genetic architecture to shape molecular evolutionary change.

Exploring the immediately available opportunity space. We speculate that proteins serving different functions will, for the most part, occupy different parts of protein sequence space. Furthermore, we assert that the purely neutral sequence variants associated with a particular protein function will describe an opportunity space that is immediately available for local exploration through stochastic (passive diffusion) processes. In the haemoglobin examples presented, we see similar patterns of variability within each of the haemoglobin sub-units for three groups of mammals, but different patterns between the sub-units. This suggests that there is one opportunity space associated with the alpha sub-unit and another for the beta sub-unit.

Breaking into nearby opportunity spaces. We propose that groups of distinct but related neutral neighbourhoods, that correspond to alleles of different fitness for a particular protein function, are aggregated into clusters. Corridors of viability bring the different neutral neighbourhoods within a cluster into close proximity such that single mutational steps can occasionally provide entry points to alleles of different fitness (new phenotypes) (Huynen et al., 1996). If a particular phenotype confers a selective advantage its frequency in the population will increase.

Creating new opportunity spaces. The shape of a neutral space can change with context . As context shifts, part of the space can become "out of bounds" (no longer neutral) while new, previously "forbidden" space, can become neutral and available for exploration. In such a scenario, evolution would involve not only movement along pre-defined corridors, but also a change in the opportunity space available through context sensitive contraction and expansion of the corridors themselves. The result is a dynamically changing context-sensitive opportunity

space for evolutionary experimentation and a perpetually changing or "restless" genotype-phenotype map (Wagner & Altenburg, 1996).

Many factors can affect context. At one level, intrinsic changes in the freedom to vary of sites within a particular protein can be brought about by an influential substitution elsewhere in the protein as described in the Covarion model of Fitch (Fitch, 1971). At another level, interactions among proteins that either enhance function, or share the burden of a function, can "open up" the neutral space. For example, built in redundancies in metabolic pathways that foster architectural resilience could, in principal, render more of the protein space effectively neutral and thus available for exploration.

Complexity and Robustness. As systems become more complex, the number of ways to solve a task increases, which in turn leads to more evolutionary opportunity. The idea that evolutionary opportunity increases with complexity can seem superficially counter-intuitive because we tend to think of complex systems as sensitive to perturbation. This sentiment is reflected in the most recent edition of Futuma's "Evolutionary Biology" text:

"The greater the number and degree of functional integration of interacting parts, the more stringent constraints on evolution are likely to be, and the rarer will be evolutionary "breakthroughs" to new organismal designs" page 684. (Futuma, 1998).

This is certainly the case for mechanical devices whose sub-components are designed to work in an additive fashion without built-in redundancy. Such devices have no 'neutral space'. However, while small perturbations will quickly bring mechanical devices to a grinding halt, biological systems show considerable resilience. The resilience fostered by neutral space can be thought of as a means to hedge one's bets in a changing environment, while simultaneously providing a platform from which to explore the available space for improved fitness configurations. It is important to recognize that no intentional design is implied by such an architecture. The enhanced opportunity to explore new fitness configurations is merely a consequence of the resilience itself. Such 'robustness' has been documented in biochemical pathways (Barkai & Leibler, 1997; Glas et al., 1998) and has been suggested to occur in brain development (Fritsch, 1995) and also in ecosystems (Naeem & Li, 1997).

Innovation and adaptive landscapes. The protein sequence space representation can provide insight into the acquisition of new protein functions and by extension the evolutionary origin of novel phenotypes. As stated previously, proteins that carry out different functions are likely

centered in different parts of the sequence space. However, their context sensitive and dynamically changing neutral neighbourhoods may occasionally come into close proximity. If neighbourhoods representing different functions come into adjacency, the potential exists for mutational "jumps" that traverse functions (Huynen et al.1996). There is mounting empirical evidence that this may be more common than previously suspected (reviewied in Golding & Dean 1998). In general, gene duplication has been forwarded as the most plausible mechanism to account for the acquisition of new protein function (Walsh, 1995). It need not be the only mechanism. A protein could shift function without undergoing gene duplication if it were part of a resilient and robust network whose elements continued to function effectively in its absence (Barkai & Leibler, 1997). This sort of redundancy would "open up" neutral space (in the same way as would gene duplication), increasing the likelihood that neighbourhoods associated with different functions came into adjacency.

The presented sequence space construct provides a representation in which neutrality and robustness foster both architectural resilience for the stability of an existing function while also providing increased evolutionary opportunity for innovation. There are parallels between the presented model and the holey adaptive landscape forwarded by Gavrilets (1997). In Gavrilets' landscape there is an emphasis after Dobzhansky (1937) on "ridges of well fit genotypes that extend through the genotype space". These ridges connect clusters of viable genotypes in the genotype space. Evolution proceeds as a percolation through the nexus of connected components. One key difference between the two models is that in our representation, most of the evolutionary change occurs along neutral corridors where sequence differences in sequence do not result in corresponding differences in structure (i.e. phenotype) (Chothia & Gerstein, 1997). In Gavrilets' landscape, the ridges of viable genotypes have phenotypes that are intermediate in form. Gavrilets' landscape may be a better description of sequence evolution that is driven by selection rather than neutrality (Gillespie, 1991)

Phylogenetic inference. Just as the concept of a neutral neighbourhood can reconcile evolutionary stability (resilience) with evolutionary change (innovation), it is also consistent with the generally reliable phylogenetic performance of molecular sequence data and its occasional failures. Molecular data will tend to be reliable when a neutral neighbourhood is both large and stationary in protein sequence space. Under these conditions equally viable (neutral) variants of a protein arise at different points along an evolutionary trajectory and are passed on from ancestor to descendant lineages. The process is essentially one of passive diffusion, transparent to the distorting effects of natural selection that can cause character distributions to be phylogenetically misleading. The evolutionary branching pattern leaves an unbiased trace in the distribution of

different neutral variants among the terminal taxa. By contrast, when the neutral neighbourhood is non-stationary over a tree, sequence data can be phylogenetically misleading because sites run the risk of being free to vary in one lineage, but not in another. Such conditions promote among-lineage rate heterogeneity and highly skewed character distributions among terminal taxa, both of which are known to be problematic for phylogenetic inference (Pesole et al., 1995; Naylor & Brown. 1997, Sullivan & Swofford, 1997).

A Quantitative Measure of Protein Function. We believe that variability profiles may also provide a way to define protein function quantitatively and to measure the degree to which two protein families differ in function (Gu 1999). With the advent of whole-genome sequencing (Fleischman et al., 1995), the need to define and compare functions on a large scale has become a pressing issue (Riley, 1997, 1998; Hegyi & Gerstein, 1999; Jansen & Gerstein, 2000; Mewes et al., 1998). One would like to be able to compare the many gene families present in two organisms and describe numerically the degree to which they differ. This is not possible when function is described in terms of simple text phrases. However, when it is described in terms of a variability signature, one can envision a number of metrics that could be applied automatically on a large scale.

CONCLUSION

We have presented data for haemoglobin indicating that the respective functions of both its alpha and beta sub-units have remained static in 3 different groups of mammals for the past 40 million years. We have contrasted this with data that suggests myoglobin may have shifted its function in cetaceans over a similar time frame. We have expanded an idea originally introduced by Maynard Smith in 1970 and have speculated how it might shed light on topics as diverse as biological resilience, evolutionary opportunity, the origin of evolutionary novelty and the reasons for the generally success and occasional failures of molecular phylogenetic inference. However, the empirical evidence we present is based on a mere 140 globin sequences. There is a need for a broader and denser sampling of related protein sequences across the diversity of life to better test some of the presented ideas. We look forward to more complete answers to some of these questions as technologies for rapid DNA sequencing improve.

ACKNOWLEDGEMENTS

We are grateful to Dan Ashlock, Homayoun Bagheri, Mike Charleston, Xun Gu, Junhyong Kim, Günter Wagner, and Steve Willson for discussion and to Elizabeth Knurek for editorial improvement of the text. MG thanks the Donaghue and Keck Foundations for support.

REFERENCES

Barkai, N. and S. Leibler. 1997. Robustness in simple biochemical networks. Nature 387 913-917.

Benner, S. A., Badcoe, I., Cohen, M. A. and Gerloff, D. L. 1994. Bona fide prediction of aspects of protein conformation. Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. J Mol Biol 235, 926-58.

Bornberg-Bauer E. and H-S Chan, 1999. Modelling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. P.N.A.S. 96: 10698-10694

Chothia, C. and Gerstein, M. (1997). Protein evolution. How far can sequences diverge? Nature 385, 579-581.

Dobzhansky, T.H. 1937. Genetics and the Origin of Species. Columbia University Press.

Fitch, W. M. 1971. Rate of change of concomitantly variable codons. J. Mol. Evol. 1: 84-96

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995). Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science* **269**, 496-512

Fontana, W. And Schuster P. Continuity in Evolution:on the nature of transitions. 1998. Science 280:1451-1455.

Fritzsch, B. 1995. Evolution of the ancestral vertebrate brain. *In* The handbook of brain theory and neural networks.Ed. M.A. Arbib. MIT Press, Cambridge Mass.

Futuma, D. 1998. Evolutionary Biology, Third Edition. Sinauer Associates Inc. “The greater the number and degree of functional integration of interacting parts, the more stringent

constraints on evolution are likely to be, and the rarer will be evolutionary "breakthroughs" to new organismal designs" page 684.

Gavrilets S. 1997. Evolution and speciation on holey adaptive landscapes. TREE 12:8 307-312

Gerstein, M., Sonnhammer, E. & Chothia, C. (1994). Volume Changes on Protein Evolution. *J. Mol. Biol.* **236**, 1067-1078.

Gerstein, M. and Altman, R. (1995). Average core structures and variability measures for protein families: Application to the immunoglobulins. J. Mol. Biol. 251, 161-175.

Gerstein, M. and Levitt, M. (1996). Using Iterative Dynamic Programming to Obtain Accurate Pair-wise and Multiple Alignments of Protein Structures. In Proc. Fourth Int. Conf. on Intell. Sys. Mol. Biol., pp. 59-67, AAAI Press, Menlo Park, CA.

Gerstein, M. and Levitt, M. (1998). Comprehensive Assessment of Automatic Structural Alignment against a Manual Standard, the Scop Classification of Proteins. *Protein Science* **7**, 445-456.

Gillespie, J. 1991. The causes of Molecular Evolution. Oxford University Press.

Glas, R. Bogyo M., McMaster J., Gaczynska M., and Ploegh H. 1998. A proteolytic system that compensates for loss of function. Nature 392: 618-620.

Golding, G.B., & A.M. Dean, 1998. The structural basis of molecular adaptation. Mol. Biol. Evol. 15:355-369.

Gu X. 1999. Statistical methods for testing functional divergence after gene duplication Mol Biol Evol, 16(12):1664-1674

Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* **288**, 147-64

Huynen, M.A. Stadler, P.F. And Fontana W. 1996. Smoothness within ruggedness: the role of neutrality in adaptation P.N.A. S. U.S.A. 93:397-401

Jansen, R. & Gerstein, M. (2000). Analysis of the Yeast Transcriptome with Broad Structural and Functional Categories: Characterizing Highly Expressed Proteins. *Nuc. Acids Res.* **28**, 1481-1488

Maddison, W.P. and D.R. Maddison, 1992. MacClade: Analysis of phylogeny and character evolution.Version 3.0 Sinauer Associates, Sunderland, Massachusetts.

Maynard-Smith, J. 1970. Natural Selection and the concept of a protein space. Nature 225: 563-564

Mewes, H. W., Hani, J., Pfeiffer, F. & Frishman, D. (1998). MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Res* **26**, 33-7.

Naylor, G.J.P. and W.M. Brown. 1997. Structural Biology and Phylogenetic Estimation. Nature 388: 527-528.

Naeem, S.and S. Li. 1997 Biodiversity enhances ecosystem reliability. Nature 390: 507

Perutz, M.F. 1983. Species adaptation in a protein molecule. Mol Biol Evol. 1, 1-28

Pesole, G., G. Dellisanti, G. Preparata and C. Saccone. 1995. The importance of base composition in the correct assessment of genetic distance. J. Mol. Evol. 41:1124-1127.

Riley, M. (1997). Genes and proteins of Escherichia coli K-12 (GenProtEC). *Nucleic Acids Res* **25**, 51-2.

Riley, M. (1998). Systems for categorizing functions of gene products. *Curr Opin Struct Biol* **8**, 388-92

Romero-Herrera AE, Lehmann H, Joysey KA, Friday AE (1978) Philos. Trans. R. Soc. Lond. B Biol. Sci. 283:995, 61-163

Sullivan, J. and D. L. Swofford. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mammal. Evol. 4 (2): 77-86.

Swofford, D. L. 1999. PAUP* 4.0b2: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, Massachusetts.

Vingron, M. and Sibbald, P. R. (1993). Weighting in sequence space: A comparison of methods in terms of generalized sequences. Proc. Natl. Acad. Sci. USA 90, 8777-8781.

Vingron, M. and Waterman, M. S. (1994). Sequence alignment and penalty choice. Review of concepts, case studies and implications. J. Mol. Biol. 235, 1-12.

Wagner G.P. and L.Altenburg, 1996. Complex adaptations and the evolution of evolvability. Evolution 50(3) 967-976

Walsh, J.B. 1995. How often do duplicated genes evolve new functions? Genetics 139:421-428.

Wuchty, S., W. Fontana, I. Hofacker, and P.Schuster. 1999. Complete Suboptimal Folding of RNA and the Stability of Secondary Structures Biopolymers 49(2): 145-165

FIGURE LEGENDS.

Figure 1. Schematic giving an overview of the globins.
(a) Schematic showing the orientation of the tetrameric form of haemoglobin. The two alpha sub-units contact each of the beta units but do not come into contact with each other. (b) The structural arrangement of the 8 alpha helices common to all globin monomers. The position of the heme is shown. Also, indicated are the approximate position of the three main exons in myoglobin: exon 1 extends from the N-terminus to B2, exon 2 from B3 to G6, exon 3 from G7 onward. (c) Some of the key resides in the globin fold are highlighted: F8 is the proximal His, which binds heme. E7 is the distal His, which contacts heme. Phe at CD1 and Leu at F4 are two other conserved heme contacts. Some inter sub-unit saltbridges switch between the R and T states. They are Tyr at alpha sub-unit C7 to Asp at beta sub-unit G1 in the T state. This becomes Asp at alpha sub-unit G1 to Asn at beta sub-unit G4 in the R state. In general the alpha sub-unit-1 to beta sub-unit-2 contacts are made by the FG and BC turns. The alpha sub-unit-1 to beta sub-unit-1 contacts are made by the G and H helices with the GH turn.

Figure 2. Variability profiles for Haemoglobin alpha and Haemoglobin beta for three mammalian groups (Carnivores, Ungulates and Primates). Profiles were determined using the tree topologies shown; the number of changes implied for each site being determined by parsimony. Plots from all six data sets are shown in register to facilitate direct comparison. The domains marked A to F correspond to the alpha helices A-F in Figure 1. Note that differences in variability profiles between alpha and beta sequences do not hinge on the phylogenies being correct as variability profiles for both alpha and beta sequences are based on identical tree topologies. Comparable numbers of taxa were used for each group in an effort to provide equivalent sampling of the neutral space for each of the different clades. Variability profiles are clearly different between the alpha and beta sub-units but are similar within sub-units for each of the three groups.

Figure 3. Ball and stick models of haemoglobin coloured to reflect the degree of change at each site. Variability profiles from the six data sets in figure 2 were plotted separately. Invariant residues are shown in grey those with 1 or two changes in green, those with 3 or 4 changes in orange and those with 5 or more changes in red. This figure gives a three dimensional structural

context to the variability plots shown in figure 2. Note that the evolutionary freedom to vary, while similar among the different mammalian groups for each Haemoglobin sub-unit, does not show perfect correspondence at the level of individual residues. Instead the patterns of variability suggest a correspondence at the level of particular sub-regions of the molecule (along particular surfaces of certain helices, for example)

Figure 4. Variability profiles for Myoglobin contrasting primates and cetaceans. Profiles were determined using the tree topologies shown. The relative position of each residue is depicted on the x axis. The domains marked A to F correspond to the alpha helices A -F in Figure 1. Myoglobins show distinctly different patterns of variability between primates and cetaceans suggesting that the constraint profiles differ between the two groups.

Figure 5. Ball and stick models of myoglobin coloured to reflect the degree of change at each site. Variability profiles from the data sets in figure 3 were plotted separately. Invariant residues are shown in grey those with 1 change in green, those with 2 changes in orange and those with 3 or more changes in red. (Note this is a different scale than used in Figure 3.) This figure gives a three dimensional structural context to the variability plots shown in figure 4.

Appendix.

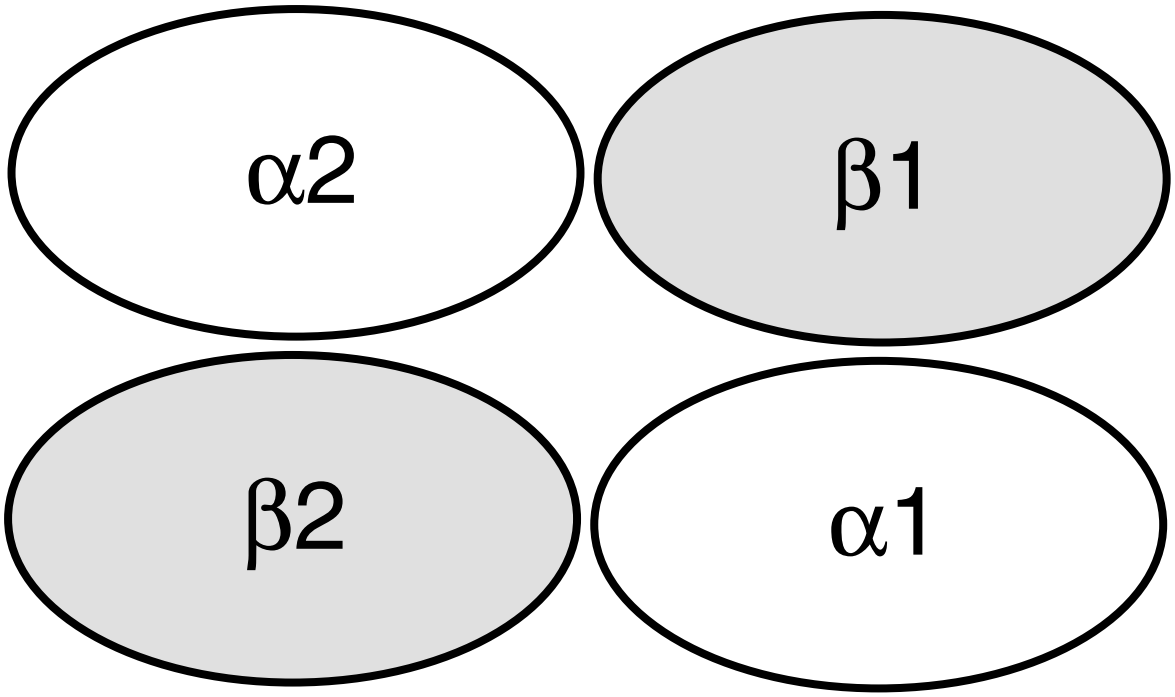
Globin sequences compared in this study can be found in SwissProt.

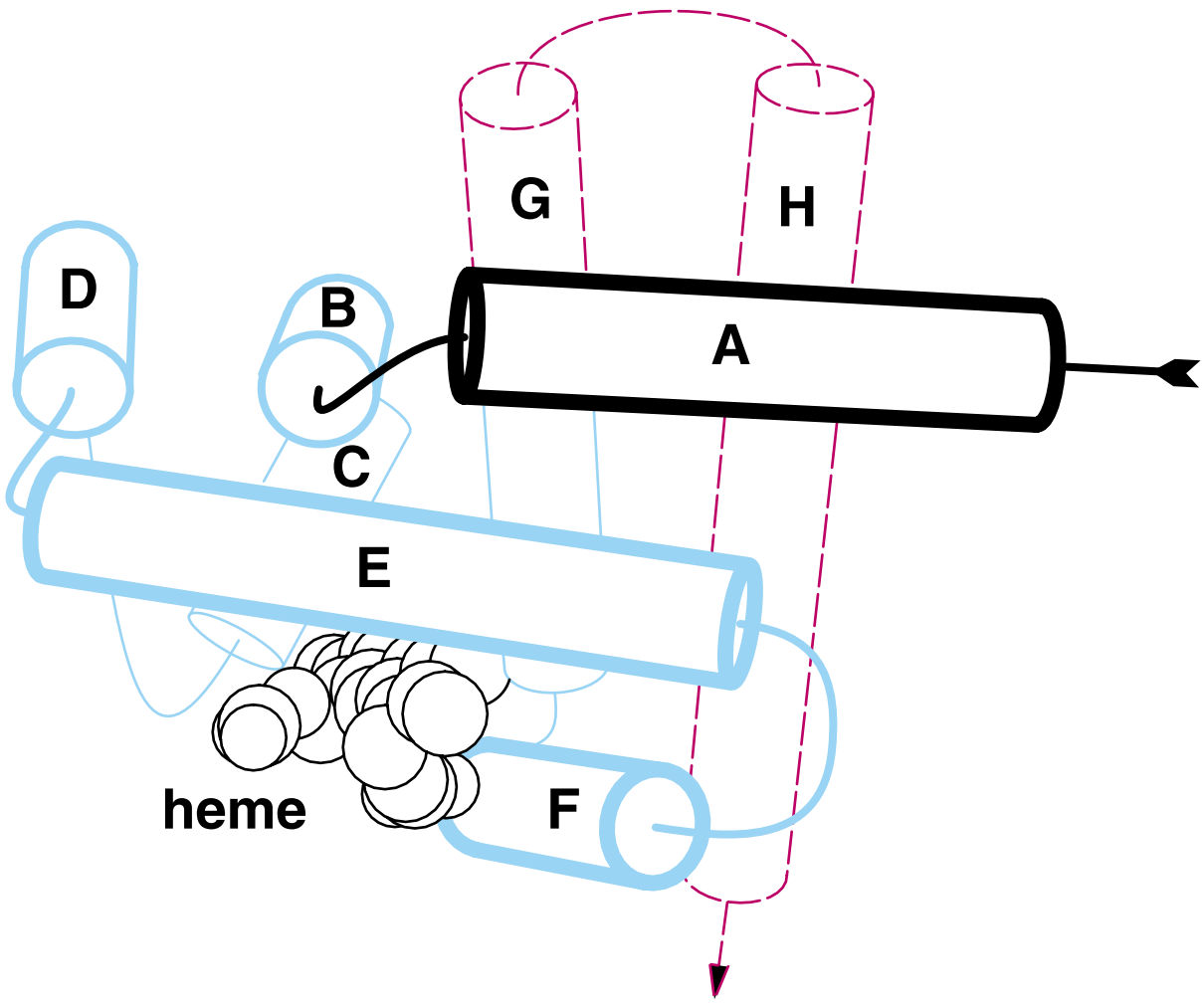
HBA_CERSI	*Ceratotherium simum*
HBA_EQUHE	*Equus hemionus*
HBA_HORSE	*Equus caballus*
HBA_RHIUN	*Rhinoceros unicornis*
HBA_ALCAA	*Alces alces*
HBA_BISBO	*Bison bonasus*
HBA_BOSGA	*Bos gaurus*
HBA_BOVIN	*Bos taurus*
HBA_CAMDR	*Camelus dromedarius*
HBA_CAPHI	*Capra hircus*
HBA_HIPAM	*Hippopotamus amphibius*
HBA_LAMGL	*Lama glama*
HBA_ODOVI	*Odocoileus virginianus*
HBA_PIG	*Sus scrofa*
HBA_RANTA	*Rangifer tarandus*
HBA_TRAST	*Tragelaphus strepsiceros*
HBB_CERSI	*Ceratotherium simum*
HBB_EQUHE	*Equus hemionus*
HBB_HORSE	*Equus caballus*
HBB_RHIUN	*Rhinoceros unicornis*
HBB_ALCAA	*Alces alces*
HBB_BISBO	*Bison bonasus*
HBB_BOSGA	*Bos gaurus*
HBB_BOVIN	*Bos taurus*
HBB_CAMDR	*Camelus dromedarius*
HBB_HIPAM	*Hippopotamus amphibius*
HBB_LAMGL	*Lama glama*
HBB_ODOVI	*Odocoileus virginianus*
HBB_PIG	*Sus scrofa*
HBB_RANTA	*Rangifer tarandus*
HBB_SHEEP	*Ovis aries*
HBB_TRAST	*Tragelaphus strepsiceros*
HBA_AILFU	*Ailurus fulgens*
HBA_AILME	*Ailuropoda melanoleuca*
HBA_CANFA	*Canis familiaris*
HBA_CROCR	*Crocuta crocuta*
HBA_FELCA	*Felis silvestris*
HBA_LEPWE	*Leptonychotes weddelli*
HBA_LUTLU	*Lutra lutra*
HBA_LYNLY	*Lynx lynx*

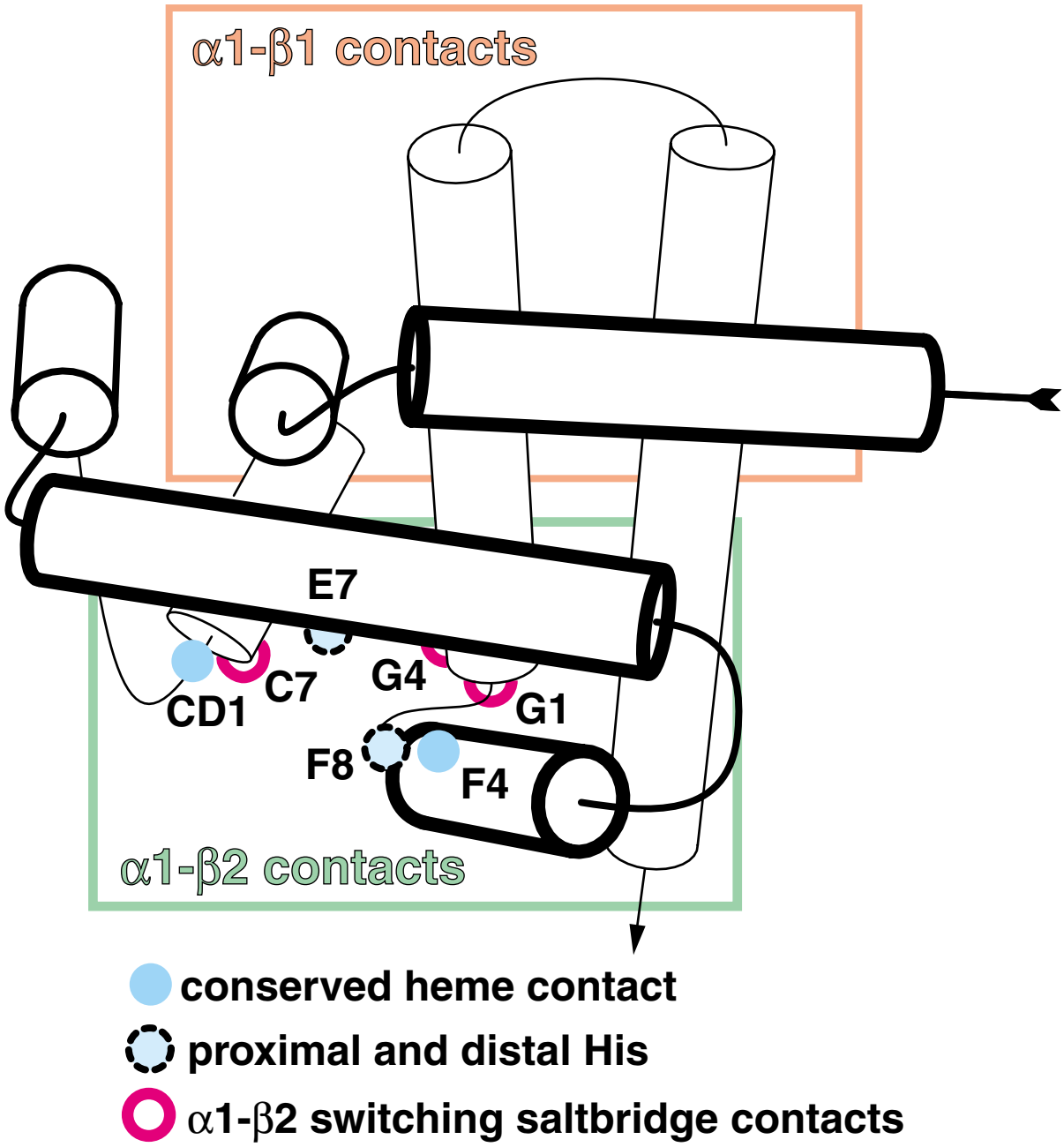
HBA_MELCA *Mellivora capensis*
HBA_MELME *Meles meles*
HBA_MUSLU *Mustela lutreola*
HBA_ODORO *Odobenus rosmarus*
HBA_PAGLA *Paguma larvata*
HBA_PANLE *Panthera leo*
HBA_PHOVI *Phoca vitulina*
HBA_PROCR *Proteles cristatus*
HBA_PROLO *Procyon lotor*
HBA_PTEBR *Pteronura brasiliensis*
HBA_URSMA *Thalarctos maritimus*
HBA_VULVV *Vulpes vulpes*
HBB_AILFU *Ailurus fulgens*
HBB_AILME *Ailuropoda melanoleuca*
HBB_CANFA *Canis familiaris*
HBB_CROCR *Crocuta crocuta*
HBB_FELCA *Felis silvestris*
HBB_LEPWE *Leptonychotes weddelli*
HBB_LUTLU *Lutra lutra*
HBB_LYNLY *Lynx lynx*
HBB_MELCA *Mellivora capensis*
HBB_MELME *Meles meles*
HBB_MUSLU *Mustela lutreola*
HBB_ODORO *Odobenus rosmarus*
HBB_PAGLA *Paguma larvata*
HBB_PANLE *Panthera leo*
HBB_PHOVI *Phoca vitulina*
HBB_PROCR *Proteles cristatus*
HBB_PROLO *Procyon lotor*
HBB_PTEBR *Pteronura brasiliensis*
HBB_URSMA *Thalarctos maritimus*
HBB_VULVV *Vulpes vulpes*
HBA_ATEGE *Ateles geoffroyi*
HBA_CALAR *Callithrix argentata*
HBA_CEBAP *Cebus apella*
HBA_CERAE *Cercopithecus aethiops*
HBA_CERTO *Cercocebus torquatus*
HBA_COLBA *Colobus badius*
HBA_EULFU *Eulemur fulvus*
HBA_GORGO *Gorilla gorilla*
HBA_HUMAN *Homo sapiens*
HBA_LORTA *Loris tardigradus*
HBA_LEMVA *Lemur varecia*
HBA_MACMU *Macaca mulatta*
HBA_MANSP *Mandrillus sphinx*
HBA_NYCCO *Nycticebus coucang*
HBA_PAPCY *Papio hamadryas*
HBA_TARSY *Tarsius syrichta*
HBA_PREEN *Presbytis entellus*
HBA_SAGFU *Saguinus fuscicollis*
HBA_THEGE *Theropithecus gelada*
HBA_GALCR *Galago crassicaudatus*
HBB_ATEGE *Ateles geoffroyi*
HBB_CALAR *Callithrix argentata*

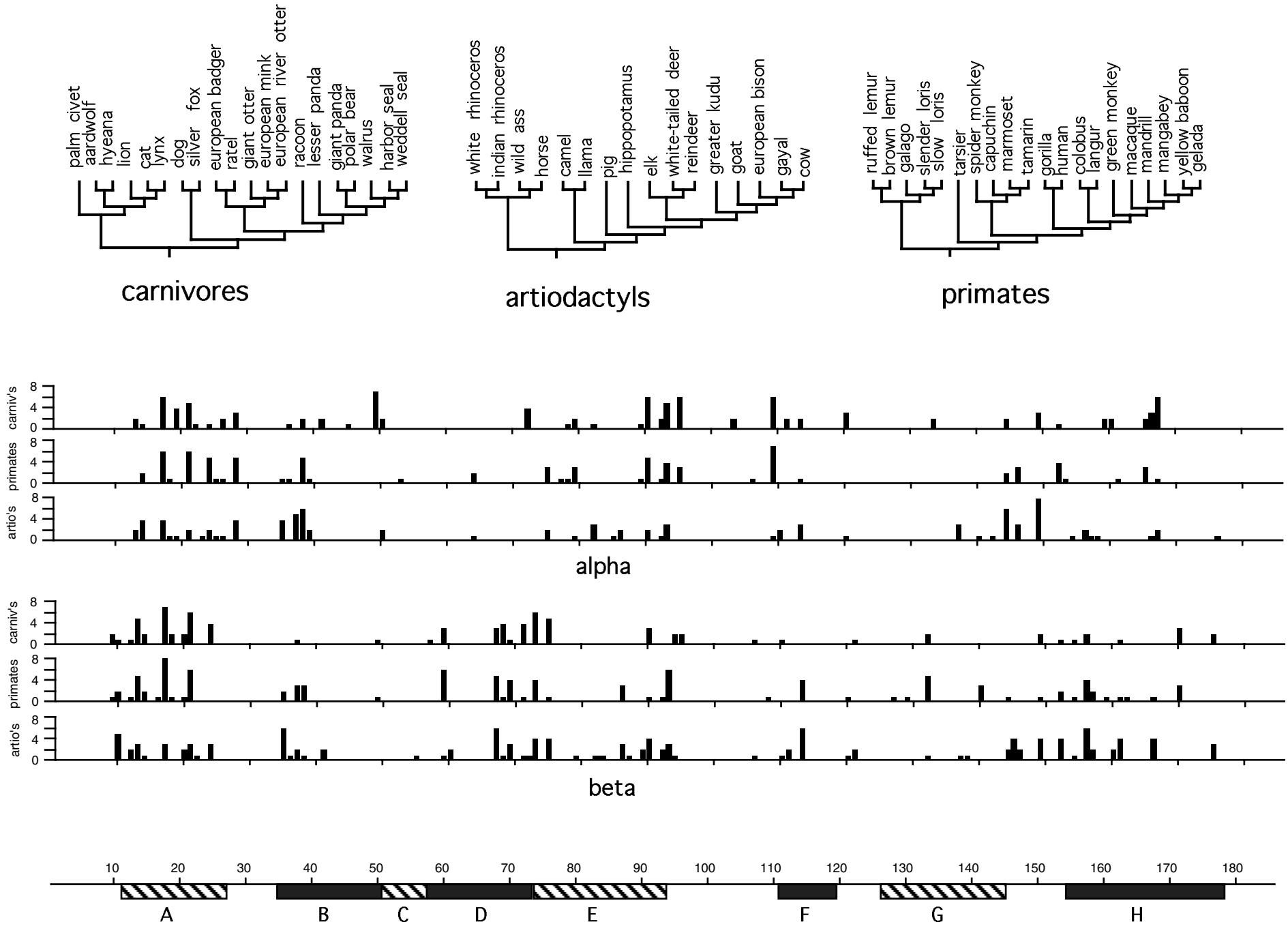
HBB_CEBAP *Cebus apella*
HBB_CERAE *Cercopithecus aethiops*
HBB_CERTO *Cercocebus torquatus*
HBB_COLBA *Colobus badius*
HBB_EULFU *Eulemur fulvus*
HBB_GORGO *Gorilla gorilla*
HBB_HUMAN *Homo sapiens*
HBB_LORTA *Loris tardigradus*
HBB_LEMVA *Lemur varecia*
HBB_MACMU *Macaca mulatta*
HBB_MANSP *Mandrillus sphinx*
HBB_NYCCO *Nycticebus coucang*
HBB_PAPCY *Papio hamadryas*
HBB_PREEN *Presbytis entellus*
HBB_SAGFU *Saguinus fuscicollis*
HBB_THEGE *Theropithecus gelada*
HBB_TARSY *Tarsius syrichta*
HBB_GALCR *Galago crassicaudatus*
MYG_BALAC *Balaenoptera acutorostrata*
MYG_BALPH *Balaenoptera physalus*
MYG_ESCGI *Eschrichtius robustus*
MYG_GLOME *Globicephala melas*
MYG_INIGE *Inia geoffrensis*
MYG_KOGSI *Kogia simus*
MYG_MEGNO *Megaptera novaeangliae*
MYG_MESCA *Mesoplodon carlhubbsi*
MYG_ORCOR *Orcinus orca*
MYG_PHYCA *Physeter catodon*
MYG_TURTR *Tursiops truncatus*
MYG_ZIPCA *Ziphius cavirostris*
MYG_PHOPH *Phocoena phocoena*
MYG_CALJA *Callithrix jacchus*
MYG_CEBAP *Cebus apella*
MYG_GALCR *Galago crassicaudatus*
MYG_GORBE *Gorilla gorilla*
MYG_HUMAN *Homo sapiens*
MYG_HYLAG *Hylobates agilis*
MYG_LAGLA *Lagothrix lagotricha*
MYG_LEPMU *Lepilemur mustelinus*
MYG_MACFA *Macaca fascicularis*
MYG_NYCCO *Nycticebus coucang*
MYG_PANTR *Pan troglodytes*
MYG_PAPAN *Papio hamadryas*
MYG_PERPO *Perodicticus potto*
MYG_PONPY *Pongo pygmaeus*
MYG_SAISC *Saimiri sciureus j*

Figure 1A
(1B and 1C follow on next two pages)



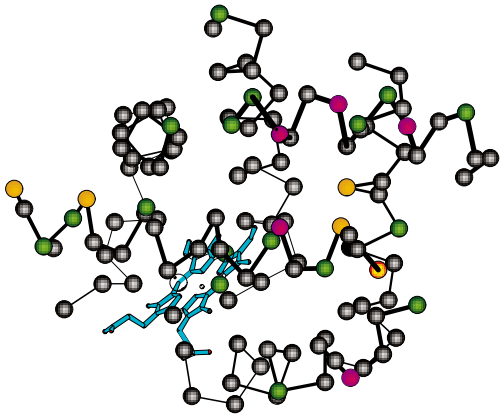

G
H
D
B
A
C
E
heme
F

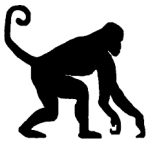

α1-β1 contacts
E7
C7
G4
CD1
G1
F8
F4
α1-β2 contacts
conserved heme contact
proximal and distal His
α1-β2 switching saltbridge contacts


palm civet
aardwolf
hyeana
lion
cat
lynx
dog
silver fox
european badger
ratel
giant otter
european mink
european river otter
racoon
lesser panda
giant panda
polar bear
walrus
harbor seal
weddell seal
carnivores
white rhinoceros
indian rhinoceros
wild ass
horse
camel
llama
pig
hippopotamus
elk
white-tailed deer
reindeer
greater kudu
goat
european bison
gayal
cow
artiodactyls
ruffed lemur
brown lemur
galago
slender loris
slow loris
tarsier
spider monkey
capuchin
marmoset
tamarin
gorilla
human
colobus
langur
green monkey
macaque
mandrill
mangabey
yellow baboon
gelada
primates
carniv's
primates
artio's
alpha
beta
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
A
B
C
D
E
F
G
H

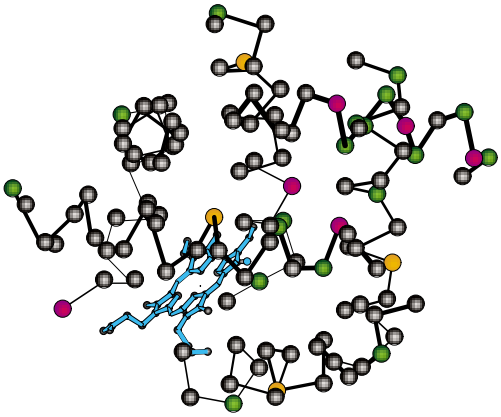
Hb α

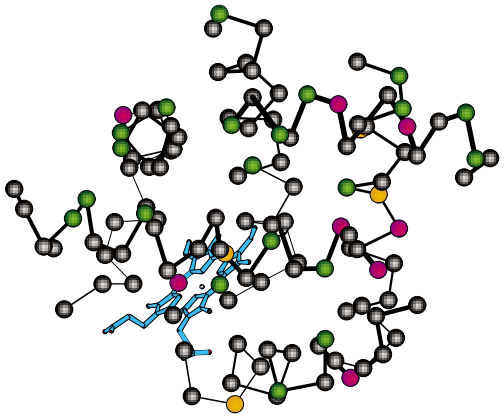
Hb β



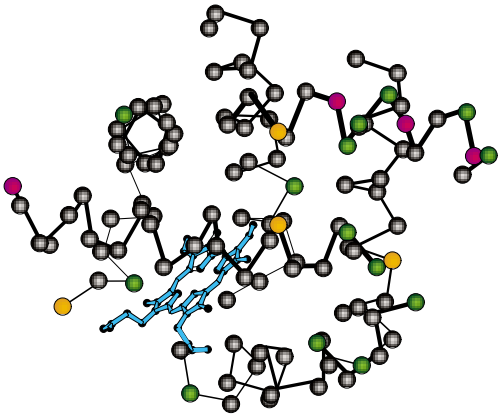


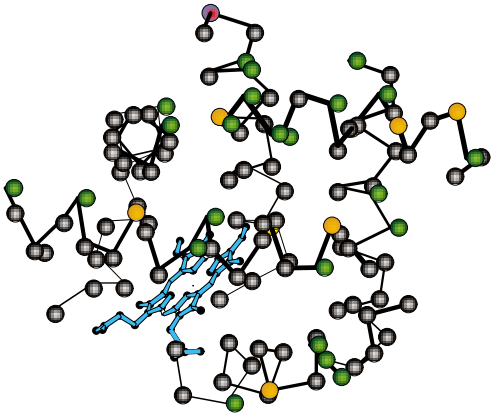
primates



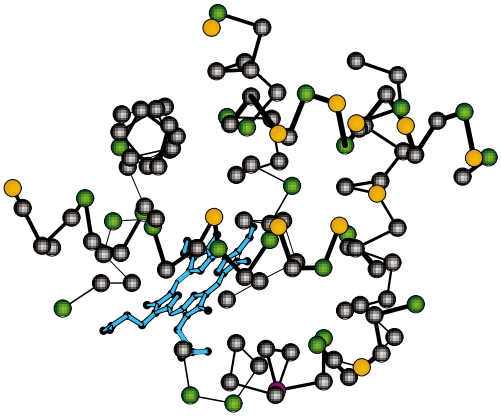


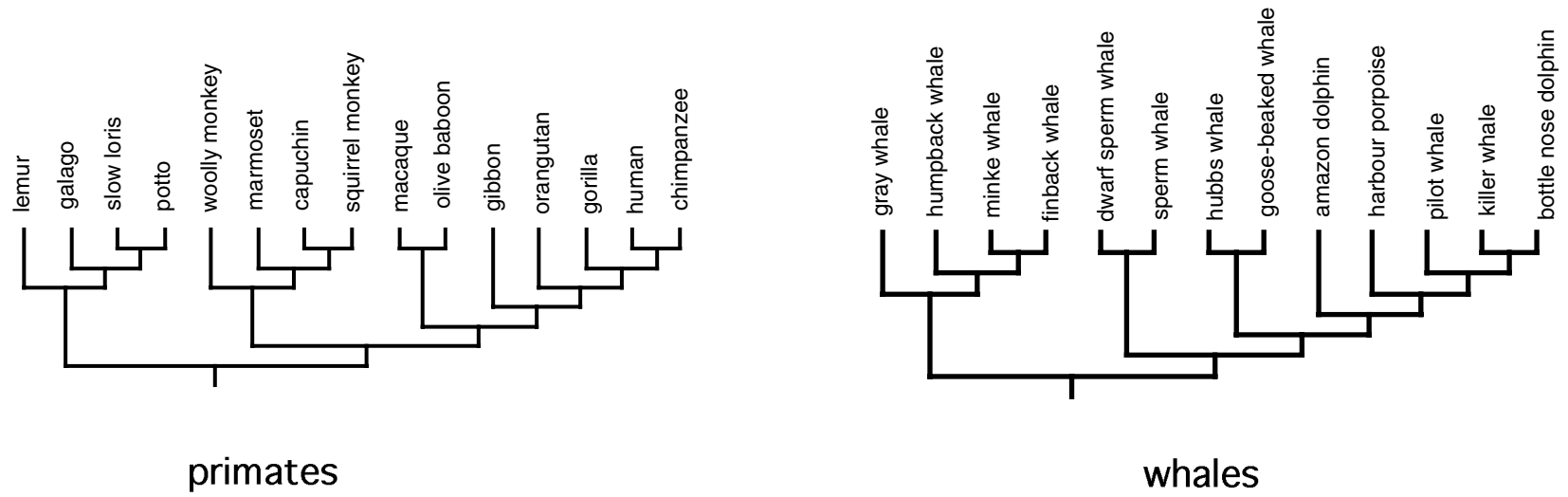
carnivores

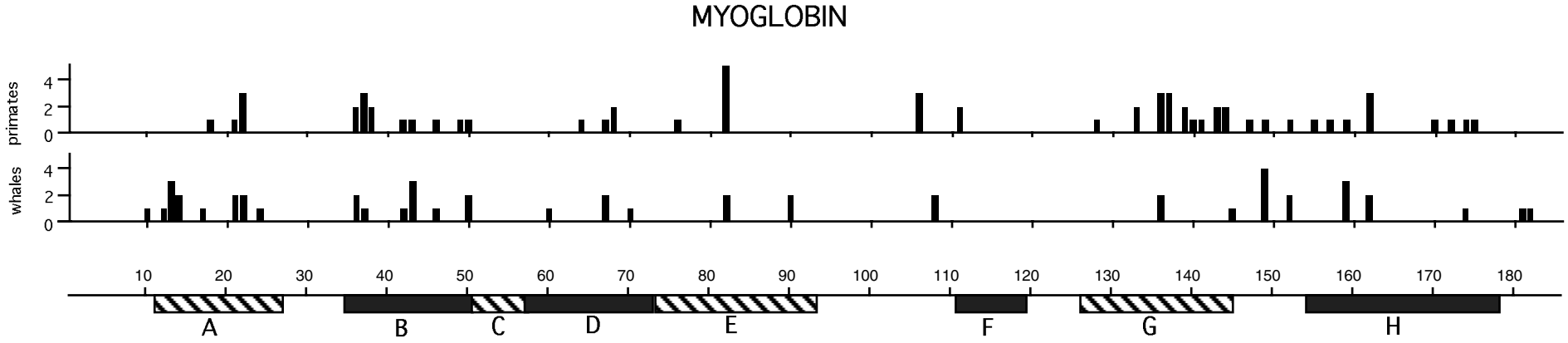


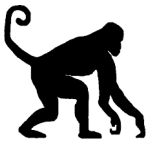


ungulates

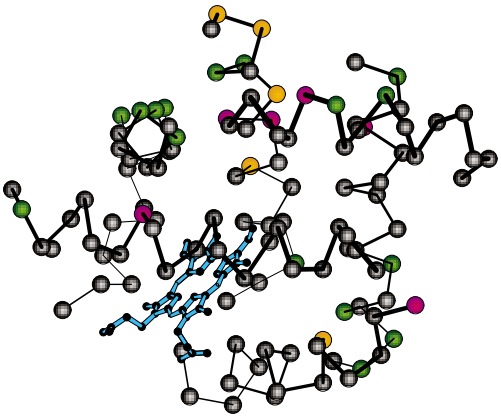


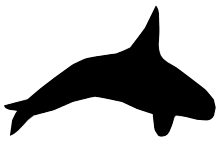

lemur
galago
slow loris
potto
woolly monkey
marmoset
capuchin
squirrel monkey
macaque
olive baboon
gibbon
orangutan
gorilla
human
chimpanzee
primates
gray whale
humpback whale
minke whale
finback whale
dwarf sperm whale
sperm whale
hubbs whale
goose-beaked whale
amazon dolphin
harbour porpoise
pilot whale
killer whale
bottle nose dolphin
whales


MYOGLOBIN
primates
whales
0
2
4
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
A
B
C
D
E
F
G
H



primates





cetaceans