The Journal of Experimental Biology 201, 1099–1117 (1998)
Printed in Great Britain © The Company of Biologists Limited 1998
JEB1357

HEMOGLOBINS FROM BACTERIA TO MAN: EVOLUTION OF DIFFERENT PATTERNS OF GENE EXPRESSION

ROSS HARDISON*

Department of Biochemistry and Molecular Biology and the Center for Gene Regulation, The Pennsylvania State University, University Park, PA 16802, USA

*Address for correspondence: Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 206 Althouse Lab, University Park, PA 16802, USA (e-mail: rch8@psu.edu)

Accepted 15 October 1997; published on WWW 24 March 1998

Summary

The discovery of hemoglobins in virtually all kingdoms of organisms has shown (1) that the ancestral gene for hemoglobin is ancient, and (2) that hemoglobins can serve additional functions besides transport of oxygen between tissues, ranging from intracellular oxygen transport to catalysis of redox reactions. These different functions of the hemoglobins illustrate the acquisition of new roles by a pre-existing structural gene, which requires changes not only in the coding regions but also in the regulatory elements of the genes. The evolution of different regulated functions within an ancient gene family allows an examination of the types of biosequence data that are informative for various types of issues. Alignment of amino acid sequences is informative for the phylogenetic relationships among the hemoglobins in bacteria, fungi, protists, plants and animals. Although many of these diverse hemoglobins are induced by low oxygen concentrations, to date none of the molecular mechanisms for their hypoxic induction shows common regulatory proteins; hence, a search for matches in non-coding DNA sequences would not be expected to be fruitful. Indeed, alignments of non-coding DNA sequences do not reveal significant matches even between mammalian α- and β-globin gene clusters, which diverged approximately 450 million years ago and are still expressed in a coordinated and balanced manner. They are in very different genomic contexts that show pronounced differences in regulatory mechanisms. The α-globin gene is in constitutively active chromatin and is encompassed by a CpG island, which is a dominant determinant of its regulation, whereas the β-globin gene is in A+T-rich genomic DNA. Non-coding sequence matches are not seen between avian and mammalian β-globin gene clusters, which diverged approximately 250 million years ago, despite the fact that regulation of both gene clusters requires tissue-specific activation of a chromatin domain regulated by a locus control region. The *cis*-regulatory sequences needed for domain opening and enhancement do show common binding sites for transcription factors. In contrast, alignments of non-coding sequences from species representing multiple eutherian mammalian orders, some of which diverged as long as 135 million years ago, are reliable predictors of novel *cis*-regulatory elements, both proximal and distal to the genes. Examples include a potential target for the hematopoietic transcription factor TAL1.

Key words: hemoglobin, evolution, globin gene, promoter, locus control region, biosequence alignment, phylogenetic footprints.

Introduction

Many heme-binding proteins with diverse functions are known, including electron-transferring cytochromes, intracellular peroxidases and lignin-degrading extracellular peroxidases. Some of the most abundant hemoproteins, the hemoglobins, allow the reversible binding of oxygen to the heme. They are not commonly implicated in catalysis or electron transfer; instead, the heme-bound iron stays in its 2+ (FeII) oxidation state. Hemoglobins are usually thought of as the major proteins in erythrocytes circulating in the blood of vertebrates, carrying the oxygen inhaled by the lungs to the respiring tissues in the body. Hemoglobins were first found in blood simply because they are so abundant, with a concentration in normal human blood of 15 g per 100 ml. But it has become clear that hemoglobins are very widespread in the biosphere, are found in all groups of organisms, including prokaryotes, fungi, plants and animals, and carry out many different functions, including catalysis.

The widespread and diverse hemoglobins appear to be encoded by orthologous genes, i.e. the phylogenetic analysis indicates that the genes are descended from an ancient, common ancestral gene. Thus, the different functions of the hemoglobins illustrate the acquisition of new roles by a pre-existing structural gene, which requires changes not only in the coding regions but also in the regulatory elements of the genes. This paper will

review this broad distribution of hemoglobins and use that as a context in which to examine some aspects of the evolution of regulation in this gene family. The evolution of different regulated functions within an ancient gene family also allows an analysis of the types of biosequence comparisons that are informative for various kinds of biological questions and that will be a parallel theme of this paper.

Hemoglobin evolution from bacteria to man illustrates the differing roles of hemoglobin

Animal hemoglobins

The hemoglobin that can be readily isolated from the blood of any vertebrate is a heterotetramer of two α-globin and two β-globin polypeptides, with a heme tightly bound to a pocket in each globin monomer. The movements and interactions between the α- and β-globin subunits lead to the cooperative binding of oxygen to this hemoglobin, allowing it to pick up oxygen readily in the lungs and to unload it efficiently in the peripheral respiring tissues (Table 1). The amino acid sequences of the α- and β-globins are approximately 50% identical, regardless of which vertebrate species is the source, arguing that these two genes are descended from a common ancestor approximately 450 million years ago, in the ancestral jawed vertebrate (Goodman *et al.* 1987). Both α- and β-globins are about equally divergent from the monomeric myoglobin, an oxygen storage and delivery protein found in many tissues. It lacks the exquisite cooperativity of the blood hemoglobins, but its relationship to them is clear from both the primary sequence and the virtually identical three-dimensional structures, each containing the globin fold (Dickerson and Geis, 1983). Further studies have found hemoglobins in jawless vertebrates and in diverse invertebrates ranging from flies (arthropods) through earthworms (annelids) to nematodes (Riggs, 1991; Dixon *et al.* 1992; Sherman *et al.* 1992). The amino acid sequences of invertebrate hemoglobins can be aligned with those of vertebrate hemoglobins (examples are shown in Fig. 1). In a parsimony analysis of these aligned amino acid sequences, the vertebrate and invertebrate hemoglobins form separate, distinct, monophyletic clades within the overall tree for hemoglobins (Fig. 2). Thus, the primary structures of invertebrate hemoglobins are related, but somewhat distantly, to those of vertebrate globins. In some invertebrates, the large extracellular hemoglobins are fusion proteins composed of multiple copies of the familiar monomeric globins. As hemoglobins are found in more and more distantly related species, the estimated time for the last common ancestral hemoglobin gene moves further back, to at least 670 million years ago in the case of the

Table 1. *Selected hemoglobins illustrate the diversity of proposed functions and regulation*

Class	Exemplary genus	Hemoglobin	Regulation	Function (demonstrated and proposed)
Vertebrate	*Homo*	HbA	Hypoxia-induced increase in production of erythropoietin, which stimulates proliferation and differentiation of erythroid precursors, the progeny of which express Hb at a high level	Oxygen transport between tissues
Plant	*Glycine*	Lb	Nodulin-specific increase in transcription	May sequester oxygen away from nitrogenase May transport oxygen to electron transport chain in nodule
Plant	*Glycine*	Nonsymbiotic Hb	Induced by hypoxia?	Intracellular oxygen movement
Alga	*Chlamydomonas*	LI637 Hb	Light-inducible expression in chloroplast	Oxygen bound to LI637 Hb can be reduced. It may serve to accept electrons, sequester oxygen or deliver oxygen inside the organelle
Fungi	*Saccharomyces*	YHB (a flavo-hemoglobin)	Induced by high levels of oxygen or reactive oxygen species, or by blocking electron transport Repressed by hypoxia Induction is mediated by the transcription factors HAP1 and HAP2/3/4	Can transfer electrons from NADPH to heme iron May serve to protect from oxidative stress
Bacteria	*Alcaligenes*	FHP (a flavo-hemoglobin)	Induced anaerobically Promoter contains a potential binding site for NarL and FNR	Proposed electron transfer Possible role in anaerobic metabolism, perhaps gas metabolism during denitrification
Bacteria	*Vitreoscilla*	Hb	Induced by hypoxia Promoter contains binding sites for FNR	Can serve as terminal electron acceptor during respiration May scavenge oxygen

```
                A helix---------->B helix---------->C helix>       D helix>E helix--- ------------>
       1           15 16           30 31           45 46           60 61           75 76           90
                                                    P      F                        'H'
humhbb -------VHLTPEEK SAVTALWGKVN--VD EVGGEALGRLLVVYP WTQRFFESFGDLSTP DAVMGNPK----VKA HGKKVLGAFSDGLAH     77
humhbg -------GHFTEEDK ATITSLWFKVN--VE DAGGETLGRLLVVYP WTNRFFDSFGNLSSA SAIMGNPK----VLA HGKKVLTSLGDAIKH     77
humhba --------VLSPADK TNVKAAWGKVGAHAG EYGAEALERMFLSFP TTKTYFPHF------ DLSHGSAQ----VKG HGKKVADALTNAVAH     72
humhbz --------SLTKTER TIIVSMWAKISTQAD TIGTETLERLFLSHP QTKTYFPHF------ DLHPGSAQ----LRA HGSKVVAAVGDAVKS     72
soyhbn --TTTLERGFSEEQE ALVVKSWNVMKKNSG ELGLKFFLKIFEIAP SAQKLFSFL------ RDSTVPLEQNPKLKP HAVSVFVMTCDSAVQ     82
parhbn MSSSEVNKVFTEEQE ALVVKAWAVMKKNSA ELGLQFFLKIFEIAP SAKNLFSYL------ KDSPVPLEQNPKLKP HATTVFVMTCESAVQ     84
soylbc -------GAFTEKQE ALVSSSFEAFKANIP QYSVVFYNSILEKAP AAKDLFSFL------ ANGVDPTN--PKLTG HAEKLFALVRDSAGQ     75
pealbI --------GFTDKQE ALVNSSSE-FKQNLP GYSILFYTIVLEKAP AAKGLFSFL------ KDTAGVEDS-PKLQA HAEQVFGLVRDSAAQ     74
ytflHb  --------MLAEKTR SIIKATVPVLEQQGT VITRTFYKNMLTEHT ELLNIFNRT------ NQKVGAQP----N-A LATTVLAAAKNIDDL     71
bacfhb --------MLDNKTI EIIKSTVPVLQQHGE TITGRFYDRMFQDHP ELLNIFNQT------ NQKKKTQR----T-A LANAVIAAAANIDQL     71
vitrhb --------MLDQQTI NIIKATVPVLKEHGV TITTTFYKNLFAKHP EVRPLFDMG------ RQESLEQP----K-A LAMTVLAAAQNIENL     71
alcfhb --------MLTQKTK DIVKATAPVLAEHGY DIIKCFYQRMFEAHP ELKNVFNMA------ HQEQGQQQ----Q-A LARAVYAYAENIEDP     71
ascehb -----SANKTRELCM KSLEHAKVDTSNEAR QDGIDLYKHMFENYP PLRKYFKNR-----E EYTAEDVQNDPFFAK QGQKILLACHVLCAT     80
ptnohb --AIASASKTRELCM KSLEHAKVGTSKEAK QDGIDLYKHMFEHYP AMKKYFKHR-----E NYTPADVQKDPFFIK QGQNILLACHVLCAT     83
caehb1 -----NRQEISDLCV KSLEGRMVGTEAQNI ENGNAFYRYFFTNFP DLRVYFKGA-----E KYTADDVKKSERFDK QGQRILLACHLLANV     80
tetrhb ----------MNKPQ TIYEKLGG---ENAM KAAVPLFYKKVLADE RVKHFFKNT------ --DMDHQT------K QQTDFLTMLLGGPNH     63
chl637 ----------RKCPS SLFAKLGG---REAV EAAVDKFYNKIVADP TVSTYFSNT------ --DMKVQR------S KQFAFLAYALGGASE     63
nosthb ----------MS TLYDNIGG---QPAI EQVVDELHKRIATDS LLAPIFAGT------ --DMAKQR------N HLVAFLGQIFEGPKQ     60
humcyc ----------M GDVEKGKKIFIMKCS QCHTVEKGGKHKTGP NLHGLFGRK------ ---TGQAP------- -GYSYTAANKNKGII     59

          F helix- ------>     G helix------------ >       H helix ------------->
       91          105 106         120 121         135 136         150 151         165 166     176
                         H                                          'AW'
humhbb LDNLKGTFA------ TLSELHCDKLHVDPE -NFRLLGNVLVCVLA HHFGKE--FTPPVQA AYQKVVAGVANALAH KYH--------    146
humhbg LDDLKGTFA------ QLSELHCDKLHVDPE -NFKLLGNVLVTVLA IHFGKE--FTPEVQA SWQKMVTGVASALSS RYH--------    146
humhba VDDMPNALS------ ALSDLHAHKLRVDPV -NFKLLSHCLLVTLA AHLPAE--FTPAVHA SLDKFLASVSTVLTS KYR--------    141
humhbz IDDIGGALS------ KLSELHAYILRVDPV -NFKLLSHCLLVTLA ARFPAD--FTAEAHA AWDKFLSVVSSVLTE KYR--------    141
soyhbn LRKAGKVTVRESNLK KLGATHFRTGVANE- -HFEVTKFALLETIK EAVPEM--WSPAMKN AWGEAYDQLVDAIKS EMKPPSS----    160
parhbn LRKAGKVTVKESDLK RIGAIHFKTGVVNE- -HFEVTRFALLETIK EAVPEM--WSPEMKN AWGVAYDQLVAAIKF EMKPSST----    162
soylbc LKTNGTVVA----DA ALVSIHAQKAVTDP- -QFVVVKEALLKTIK EAVGGN--WSDELSS AWEVAYDELAAAIKK A----------    143
pealbI LRTKGEVVLG---NA TLGAIHVQKGVTNP- -HFVVVKEALLQTIK KASGNN--WSEELNT AWEVAYDGLATAIKK AMKTA------    147
ytflHb  SVLMDHVKQ------ -IGHKHRALQIKPE- -HYPIVGEYLLKAIK EVLGDA--ATPEIIN AWGEAYQAIADIFIT VEKK-------    139
bacfhb GNIIPVVKQ------ -IGHKHRSIGIKPE- -HYPIVGKYLLIAIK DVLGDA--ATPDIMQ AWEKAYGVIADAFIG IEKDM------    140
vitrhb PAILPAVKK------ -IAVKHCQAGVAAA- -HYPIVGQELLGAIK EVLGDA--ATDDILD AWGKAYGVIADVFIQ VEADLYAQAVE    146
alcfhb NSLMAVLKN------ -IANKHASLGVKPE- -QYPIVGEHLLAAIK EVLGNA--ATDDIIS AWAQAYGNLADVLMG MESEL------    140
ascehb YDDRETFNAYTR--- ELLDRHARDHVHMP- ---PEVWTDFWKLFE EYLGKKTTLDEPTKQ AWHEIGREFAKEINK HGRHA------    153
ptnohb YDDRETFDAYVG--- ELMARHERDHVKIP- ---NDVWNHFWEHFI EFLGSKTTLDEPTKH AWQEIGKEFSHEISH HGRHS------    156
caehb1 YTNEEVFKGYVR--- ETINRHRIYKMDPA- -----LWMAFFTVFT GYLESVGCLNDQQKA AWMALGKEFNAESQT HLKNS------    151
tetrhb YKGKNMTEA------ -----HKGMNLQNL- -HFDAIIENLAATLK ELG-----VTDAVIN EAAKVIEHTRKDMLG K----------    121
chl637 WKGKDMRTA------ -----HKDLVPHLSD VHFQAVARHLSDTLT ELGVPP---ED-ITD AMAVVASTRTEVLNM PQQ--------    126
nosthb YGGRPMDKT------ -----HAGLNLQQP- -HFDAIAKHLGEAMA VRGVS----AEDTKA ALDRVTNMKGAILNK ----------    118
humcyc WGEDTLMEY------ -----LENP------ ------KKYIPGTKM IFVGIK---KK---E ERADLIAYLKKATNE ----------    105
```

Fig. 1. Aligned amino acid sequences of selected hemoglobins from mammals, invertebrates, plants, protists and bacteria. The positions in the alignment are given above the columns, and the positions of amino acids in each sequence are given at the end of the rows. Amino acids that are invariant (or have one mismatch in the case of proline at position 45) (P, F, H and AW) are indicated above the human β-globin sequence. The column with the distal histidine of vertebrate and plant hemoglobins is marked with an 'H'. The alignment was generated using ClustalW 1.7. humhbb, human β-globin; humhbg, human γ-globin; humhba, human α-globin; humhbz, human ζ-globin; soyhbn, soybean nonsymbiotic hemoglobin; parhbn, *Parasponia* nonsymbiotic hemoglobin; soylbc, soybean leghemoglobin C; pealbI, pea leghemoglobin I; ytflHb, yeast (*Saccharomyces cerevisiae*) flavohemoglobin YHB1; bacfhb, *Bacillus* flavohemoglobin; vitrhb, *Vitreoscilla* hemoglobin; alcfhb, *Alcaligenes* flavohemoglobin; ascehb, *Ascaris* hemoglobin; ptnohb, *Pseudoterranova* hemoglobin; caehb1, *Caenorhabditis elegans* hemoglobin; tetrhb, *Terahymena* hemoglobin; chl637, *Chlamydomonas* hemoglobin LI637; nosthb, *Nostoc* hemoglobin; humcyc, human cytochrome C.

invertebrate/vertebrate divergence (Goodman *et al.* 1988) (see Fig. 3).

Plant hemoglobins: symbiotic and nonsymbiotic

Plants not only make oxygen during photosynthesis, but they also use it for respiration *via* the electron transfer chain in mitochondria. Recent studies show that hemoglobins are widely used in plants to bind and transfer that oxygen. The first plant hemoglobins were discovered in the root nodules of legumes (reviewed in Appleby, 1984). These nodules are a symbiosis between rhizobial bacteria and the plant to allow fixation (reduction) of atmospheric nitrogen into a usable form, ammonia, which eventually appears in amino acids and other building blocks for the cells. Reduction of nitrogen consumes large amounts of energy, and the nodules have an abundant, plant-encoded hemoglobin, called leghemoglobin, that facilitates the diffusion of oxygen to the respiring bacteriods in the root nodule (Appleby, 1984). In addition, the binding of oxygen to leghemoglobin may help sequester the oxygen away from the nitrogen-fixing machinery, which is readily poisoned

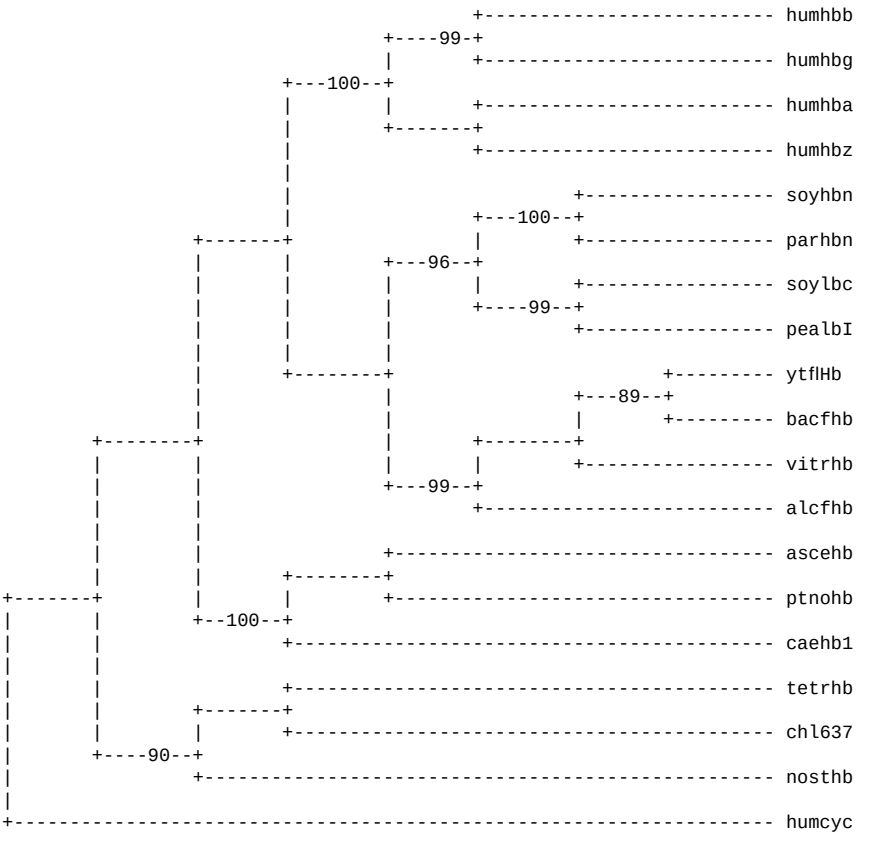

Fig. 2. Phylogenetic tree of hemoglobins. The results of a parsimony analysis of the alignment in Fig. 1, using the program PAUP, are shown. The number of times a node is supported in 100 bootstrap replicates is given to the left of the node. Abbreviations are the same as in Fig. 1.

by oxygen (Dickerson and Geis, 1983). Although the amino acid sequences of leghemoglobins differ from those of vertebrate globin genes at approximately 80 % of the positions, leghemoglobin folds into the same three-dimensional structure as the animal globins (Vainshtein *et al.* 1975). Thus, initial investigations clearly showed that some plants had hemoglobins, but it was thought that they were limited to legumes.

The recent discovery of hemoglobins in a large variety of plants, including many nonleguminous species, strongly argues that the leghemoglobins are a specialized product of divergence from an ancient plant hemoglobin gene – a gene that is itself descended from a hemoglobin gene in the last common ancestor to plants and animals. Hemoglobins distinct from leghemoglobin, called nonsymbiotic hemoglobins, have been discovered in root nodules of a nonleguminous plant (Appleby *et al.* 1983), in plants that do not form nodules (Bogusz *et al.* 1988) and in the monocotyledon cereals (Taylor *et al.* 1994). Recently, a nonsymbiotic hemoglobin gene was discovered in the legume soybean (*Glycine max*) that is distinct from the well-known leghemoglobin genes found in the same plant (Andersson *et al.* 1996). Thus, two different types of hemoglobin have been discovered in plants, a nonsymbiotic type that is widely distributed and perhaps ubiquitous among species and a symbiotic type that is induced upon nodulation.

The nonsymbiotic hemoglobins are synthesized in a wide range of tissues, including stems and young leaves of mature plants, seed cotyledons and young shoots. Although messenger RNA from the soybean gene is also present in root nodules, it is much less abundant than that for the leghemoglobins. The expression pattern indicates that this protein has a more generalized function in plants, such as facilitating oxygen diffusion to rapidly respiring cells (Andersson *et al.* 1996).

A cladistic analysis of the amino acid sequences of plant hemoglobins (Andersson *et al.* 1996) generates two distinct branches, one with the symbiotic hemoglobins (characterized by the leghemoglobins) and the other with the nonsymbiotic hemoglobins (Fig. 2). Since the latter nonsymbiotic hemoglobins have been found in a wide range of plant species, these observations strongly support the hypothesis that a gene encoding the nonsymbiotic hemoglobin was present in the ancestor to plants (Fig. 3). It is likely that the symbiotic hemoglobins arose *via* duplication of an ancestral gene followed by divergence to fulfil more specialized functions in root nodules.

A common ancestral gene for plant and animal hemoglobins

The hemoglobin gene inferred to be present in the ancestor to plants was probably related to the hemoglobin gene in the ancestor to mammals, i.e. there was a gene for hemoglobin in the last common ancestor to plants and animals. The evidence for this conclusion is based on the number and positions of introns in the contemporary genes. The plant hemoglobin genes (both symbiotic and nonsymbiotic) are separated into four exons by three introns (Jensen *et al.* 1981; Brisson and Verma, 1982; Andersson *et al.* 1996), as illustrated in Fig. 3.

The first and third introns are in positions homologous to those of the two introns found in vertebrate α- and β-globin and in myoglobin genes. The second plant intron interrupts the region coding for the E helix of hemoglobin. A similar intron/exon structure is found for the hemoglobin genes in the nematodes *Pseudoterranova* and *Ascaris* (Dixon *et al.* 1992; Sherman *et al.* 1992), which may represent an older structure than the two-intron form found in the annelid *Lumbricus* (Jhiang *et al.* 1988) or the intron-less form found in the insect *Chironomus* (Antoine and Niessing, 1984). Thus, one can propose that the ancestor to plants and animals had a hemoglobin gene with three introns (Fig. 3). This arrangement has been retained in all the plant hemoglobin genes, both symbiotic and nonsymbiotic, and also in certain nematodes. The central intron was lost prior to the divergence of annelids and arthropods and, hence, is absent in all vertebrate hemoglobin and myoglobin genes. Other nematode hemoglobin genes have lost one or more introns from the ancestral three-intron structure (reviewed in Goldberg, 1995).

Although this model is attractive in its simplicity, the assignment of the central intron as homologous between plant and nematode hemoglobins is not definitive. Alignment of plant and nematode hemoglobin sequences shows several matches on both sides of the E helix, but the region interrupted by the central intron is completely different (Fig. 1). Simply starting from the predicted beginning of the E helix in the published alignments (Sherman *et al.* 1992), the central intron in nematodes interrupts the eighth codon, whereas the central intron in plants falls between the fourteenth and fifteenth codons encoding this helix. Thus, the introns appear to be in slightly different places and in different phases. Whether this is the result of extensive divergence from a common ancestor,

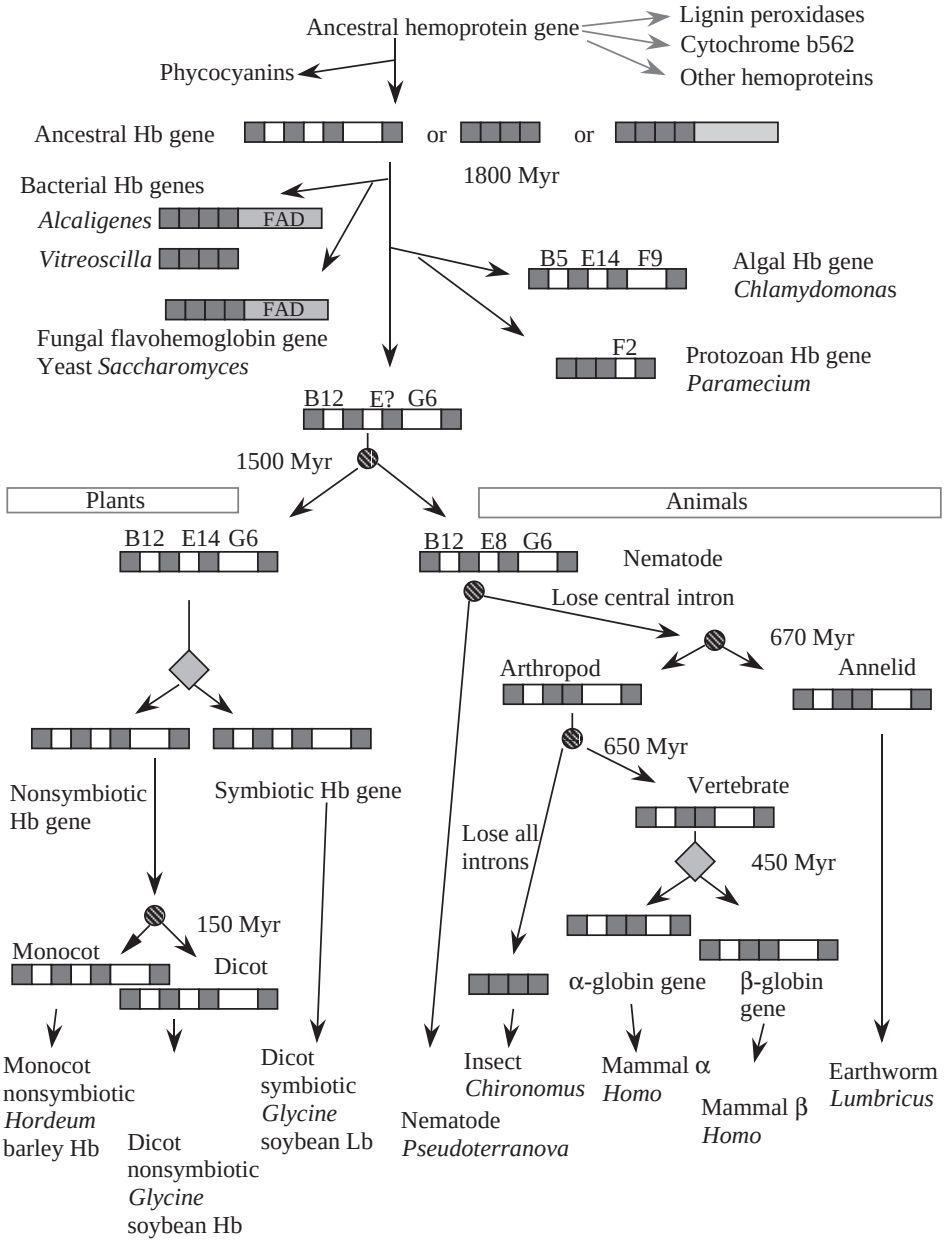

Fig. 3. Schematic overview of hemoglobin gene evolution from bacteria to man. Globin-coding exons and genes are shown as dark-filled boxes, portions of genes encoding a flavin-binding domain are shown as gray boxes, and introns are shown as open boxes in genes. Speciation events are depicted as diagonally striped circles, and gene duplications are shown as gray diamonds. Myr, millions of years ago; Monocot, monocotyledon; Dicot, dicotyledon.

resulting in intron 'sliding,' or independent insertions of introns (Dixon and Pohajdak, 1992) requires further study (discussed in Goldberg, 1995). Irrespective of whether the central intron was present early or was inserted later, the sequence relationships and overall gene structures argue strongly for an ancestral hemoglobin gene being present more than 1500 million years ago, prior to the divergence of plants and animals. It is also clear that the first and third introns are at least as old as the last common ancestor to plants and animals.

Hemoglobins in protists, fungi and bacteria

Recent studies show that the hemoglobin gene is truly ancient, preceding the divergence of prokaryotes and eukaryotes, and it now appears in a variety of exon/intron arrangements (Fig. 3). The subfamily of hemoglobins found in protists forms a distinct branch on cladograms (Zhu and Riggs, 1992), illustrated in Fig. 2. The hemoglobin gene in the protozoan *Tetrahymena* has no introns (Takagi *et al.* 1993), the homologous gene in *Paramecium* has a single intron that does not correspond to any other globin gene intron (Yamauchi *et al.* 1995), and one hemoglobin gene in the alga *Chlamydomonas* has three introns, at least two of which are in unique locations (Couture *et al.* 1994). Since these appear to be homologous genes, one must propose many introns in the ancestral gene followed by differential loss, substantial intron sliding or repeated insertions of introns to obtain the contemporary structures. Stoltzfus *et al.* (1994) have argued that this diversity in gene structure is incompatible with the idea that exons encode discrete units of protein structure.

The hemoglobins found in these unicellular organisms show more biochemical reactivities and hence have been implicated in a wider variety of potential functions that those traditionally associated with animal and plant hemoglobins (Table 1). The hemoglobins in *Chlamydomonas* are found in the chloroplast and are light-inducible. Unlike the case with many other oxyhemoglobins, O_2 bound to the *Chlamydomonas* hemoglobin can be reduced (Couture and Guertin, 1996), which raises the interesting possibility that this hemoglobin could serve as an electron acceptor, perhaps in the electron transport system or another redox system. Alternatively, it could serve to trap oxygen to protect oxygen-labile proteins or perhaps deliver oxygen to the cytochrome oxidase of the respiratory chain (Couture *et al.* 1994). A flavohemoglobin from the yeast *Saccharomyces* is a fusion of a heme-binding domain and an FAD-binding domain (Zhu and Riggs, 1992). In contrast to most other hemoglobins, its production is induced by *high* levels of oxygen (Zhu and Riggs, 1992; Crawford *et al.* 1995), and it may play a role in protecting the cell from oxidative stress (Zhao *et al.* 1996).

Hemoglobins have been found in many bacteria as well. Like the yeast protein, the flavohemoglobins from *Escherichia coli* (Vasudevan *et al.* 1991), *Bacillus subtilis* (LaCelle *et al.* 1996) and *Alcaligenes eutrophus* (Cramm *et al.* 1994) have two domains, one for binding heme and one for binding a flavin cofactor. The flavohemoglobins from yeast and bacteria form a distinct clade (Fig. 2). The three-dimensional structure has been determined for the *Alcaligenes* flavohemoglobin (Ermler *et al.* 1995). The structure corresponds to the classical globin fold, demonstrating the homology between the bacterial and eukaryotic hemoglobins. The *Alcaligenes* flavohemoglobin has been implicated in catalyzing a reduction reaction, transferring a hydride ion from NADH to FAD and then the two electrons, *via* the heme, to a still-unknown substrate (Ermler *et al.* 1995). The hemoglobin in the sliding bacterium *Vitreoscilla* is not fused with a flavoprotein domain (Wakabayashi *et al.* 1986), but it falls within the clade with the flavohemoglobins (Fig. 2). Its three-dimensional structure also conforms to that of the globin fold (Tarricone *et al.* 1997). Like the *Bacillus* flavohemoglobin and many other hemoglobins, and in contrast to the regulation of the yeast hemoglobin, it is induced when cells are grown in low-oxygen (hypoxic or anaerobic) conditions (Dikshit *et al.* 1990). Its ability to complement deficiencies of terminal cytochrome oxidases in *E. coli* suggests that this hemoglobin can receive electrons during respiration (Dikshit *et al.* 1992). A hemoglobin encoded within a *nif* operon in the cyanobacterium *Nostoc commune* (Potts *et al.* 1992) is similar to the hemoglobins found in the unicellular eukaryotes *Chlamydomonas*, *Tetrahymena* and *Paramecium*.

Thus, hemoglobins are found in virtually all kingdoms of organisms, including eubacteria, unicellular eukaryotes, plants and animals. No hemoglobins have been reported in the archaebacteria to date, but it would be surprising if they were truly absent. Although the hemoglobin genes have been diverging for an extremely long time, as much as 1800 million years, relationships among them can be analyzed by multiple alignment of amino acid sequences of the encoded proteins followed by construction of phylogenetic trees (Table 2). There is a consistent function associated with many of the hemoglobins – effective transport of oxygen. This can be achieved by making very large amounts of hemoglobin in particular cells (e.g. erythrocytes) for transport through the body or in specialized locations requiring intense respiration (e.g. nitrogen-fixing root nodules). Alternatively, smaller amounts of proteins, such as myoglobin or the nonsymbiotic hemoglobins in plants, can be made in virtually all cells, perhaps providing an efficient intracellular oxygen-delivery system to the respiring mitochondria and chloroplasts (Wittenberg and Wittenberg, 1987). This latter function appears to be very widespread in the biosphere, being found in plants, animals and algae. When the respiratory electron transport system is not in a separate intracellular compartment, as in bacteria, some species still utilize hemoglobin under hypoxic conditions, perhaps to provide oxygen as the terminal electron acceptor. Indeed, the flavohemoglobins appear to catalyze redox reactions, with the heme playing a direct role in electron transfers as it does with the cytochromes. Perhaps these latter proteins provide clues about the function of the ancestral hemoglobins.

Table 2. *Different types of biosequence or biostructure analysis are informative for various questions*

Relationships or phenomena studied	Phyla or time span covered	Biosequence/Biostructure
Hemoproteins in general	Biosphere since oxygen evolution	Three-dimensional structures
Hemoglobins	Bacteria, fungi, protists, plants, animals	Amino acid sequence, three-dimensional structure
Hemoglobin gene structure (intron/exon)	Bacteria, fungi, protists, plants, animals	Genomic DNA and cDNA sequences
Regulation of α- and β-globin genes	Vertebrate animals	General features of genomic DNA Patterns of transcription factor binding sites
Regulation of β-globin genes	Birds and mammals	Patterns of transcription factor binding sites
Regulation of β-globin genes	Mammals	Non-coding sequences in genomic DNA
Regulation of symbiotic and nonsymbiotic globin genes	Plants	Non-coding sequences in genomic DNA

Common origins for many hemoproteins

Is there an evolutionary connection between the hemoglobins, the cytochromes that pass electrons down an energy gradient in respiration and in photosynthesis, and other hemoproteins that catalyze oxidations? Keilin (1966) suggested some time ago that hemoglobins may have evolved from heme enzymes that utilize oxygen (discussed in Riggs, 1991). Irrespective of the relationships among the proteins that bind them, it is likely that metal-bound porphyrin rings or related compounds were present at the time that photosynthesis evolved; indeed, they may have been utilized then as now in capturing light energy. One can speculate on interesting scenarios for the use of hemoproteins once oxygen appeared *via* photosynthesis. Given the capacity of oxygen to damage various cellular components, oxygen-binding hemoproteins may have functioned initially to protect cells from this reactive species. Once the utility of oxygen as an electron acceptor was realized in the evolution of respiratory chains, hemoproteins could serve as electron-transfer agents (leading to contemporary cytochromes) and oxygen-bound hemoproteins could serve as the terminal electron acceptors. Further gene duplications and divergence would allow the capacity to catalyze other redox reactions to evolve. The intracellular oxygen-transport properties may have arisen from a need to scavenge scarce oxygen to provide it for the respiratory chain (leading to contemporary myoglobins and nonsymbiotic hemoglobins). In multicellular organisms, the oxygen-scavenging hemoglobins could evolve into the abundant hemoglobins now used to transport oxygen.

Primary sequence relationships may not be particularly useful in testing these proposed connections, since the ancestral amino acid sequence may have diverged beyond recognition after billions of years of evolution. A more useful guide will be the determination of three-dimensional structures by X-ray crystallography (Table 2). In this regard, it is notable that the light-harvesting biliprotein C-phycocyanin, from the cyanobacterium *Mastigocladus laminosus*, has a three-dimensional structure very similar to that of a globin (Schirmer *et al.* 1985), suggesting a common ancestry (Fig. 3). Although this is not a heme-binding protein *per se*, it does bind a linear tetrapyrrole pigment derived from heme. Heme binds between two α-helices, coordinated to histidine, in proteins as diverse as lignin peroxidase (Edwards *et al.* 1993) and cytochrome b562 (Mathews *et al.* 1979), but the topology of these helices differs from the globin fold. Is this divergent or convergent evolution? As more structures are determined, it will be highly instructive to see which distant relationships among hemoproteins will be confirmed and to determine how far the superfamily of hemoglobin genes reaches.

Mechanisms of regulating hemoglobin gene expression

Response to O_2

Given the differing roles of the hemoglobin proteins described above, it is not surprising that the regulation of the genes encoding them can differ dramatically. Since all these genes appear to be descended from a common ancestral gene (and hence are *orthologous*), the variations in regulatory mechanisms present examples of alterations of control sequences during evolution that allow pre-existing protein-coding genes to be adapted to different functions. In some cases, the regulatory changes and evolutionary distance may be so large that no remnant of the ancestral state is left to guide inferences from sequence alignments. For example, consider the regulation of various hemoglobin genes by O_2 levels. Production of many of the known hemoglobins is induced by low O_2 concentrations, as may be expected for proteins used for O_2 transport. However, the mechanism can be direct, as in several bacteria, or quite indirect, as it is in mammals (Fig. 4). Expression of the bacterial hemoglobin gene from *Vitreoscilla* (Dikshit *et al.* 1990) and the flavohemoglobin gene from *Bacillus subtilis* (LaCelle *et al.* 1996) is induced at low O_2 concentrations, but different proteins have been implicated in the regulation of these bacterial genes. The common anaerobic regulator FNR can be used to regulate positively expression of the *Vitreoscilla* hemoglobin gene (Joshi and Dikshit, 1994) and possibly of the *Alcaligenes* flavohemoglobin gene (Cramm *et al.* 1994). FNR, the fumarate nitrate reduction protein, induces expression of a large number of genes when O_2 concentration is low and electron transport switches to alternative electron acceptors such as fumarate and nitrate. The O_2 sensor in this case is well understood (Rouault and Klausner, 1996); under normal O_2 conditions, FNR is an apo-protein with no iron–sulfur (Fe–S) cluster but, when O_2 levels are low (i.e.

anaerobic conditions), a 4Fe–4S cluster is formed. The FNR protein with the 4Fe–4S cluster is an active transcriptional regulator that induces expression of many genes, including the *Vitreoscilla* hemoglobin gene. A two-component regulatory system involving the ResD and ResE proteins has been implicated in the anaerobic induction of the flavohemoglobin gene *hmp* from *Bacillus subtilis* (LaCelle *et al.* 1996). As is characteristic of two-component regulatory systems in bacteria, one protein, ResD, is the response regulator and the other, ResE, is the histidine protein kinase that transduces a signal. The ResD/ResE system was discovered recently (Sun *et al.* 1996), and currently both the nature of the oxygen sensor and the mechanism of induction of the target genes are unknown. The FNR protein is also involved in regulation of the *B. subtilis hmp* gene, but indirectly *via* its role in increasing levels of nitrite (LaCelle *et al.* 1996).

Like these bacterial genes, vertebrate hemoglobin production is also increased under conditions of hypoxia, but by the indirect mechanism of increased erythropoiesis (Fig. 4). The low O_2 concentration is actually sensed by cells in the kidney and liver, where it signals an increase in production of the hormone erythropoietin *via* response elements in a 3′ enhancer and the promoter of the gene (reviewed in Huang *et al.* 1997). Erythropoietin then acts on the erythroid progenitor cells in the bone marrow to increase proliferation, to stimulate further erythroid differentiation and to block apoptosis (Mason-Garcia and Beckman, 1991; Witthuhn *et al.* 1993; Migliaccio *et al.* 1996). This then leads to increased production of erythrocytes, each carrying abundant hemoglobin. Thus, the O_2-sensing system in vertebrates does not act directly on the hemoglobin genes, but rather acts on a hormone gene in a different tissue, eventually leading to an increase in the number of cells carrying hemoglobin. This may be viewed as an elaborate adaptation to the need for carrying oxygen to the many tissues in the body of a vertebrate, whereas hemoglobin regulation in bacteria only needs to serve the requirements of a single cell.

Much information has been gathered about the proteins and events that lead to the increased production of erythropoietin in the hepatoma cell line Hep3B. The protein HIF1 (hypoxia induction factor 1) plays a key role in increasing the production of both erythropoietin and other proteins that respond to

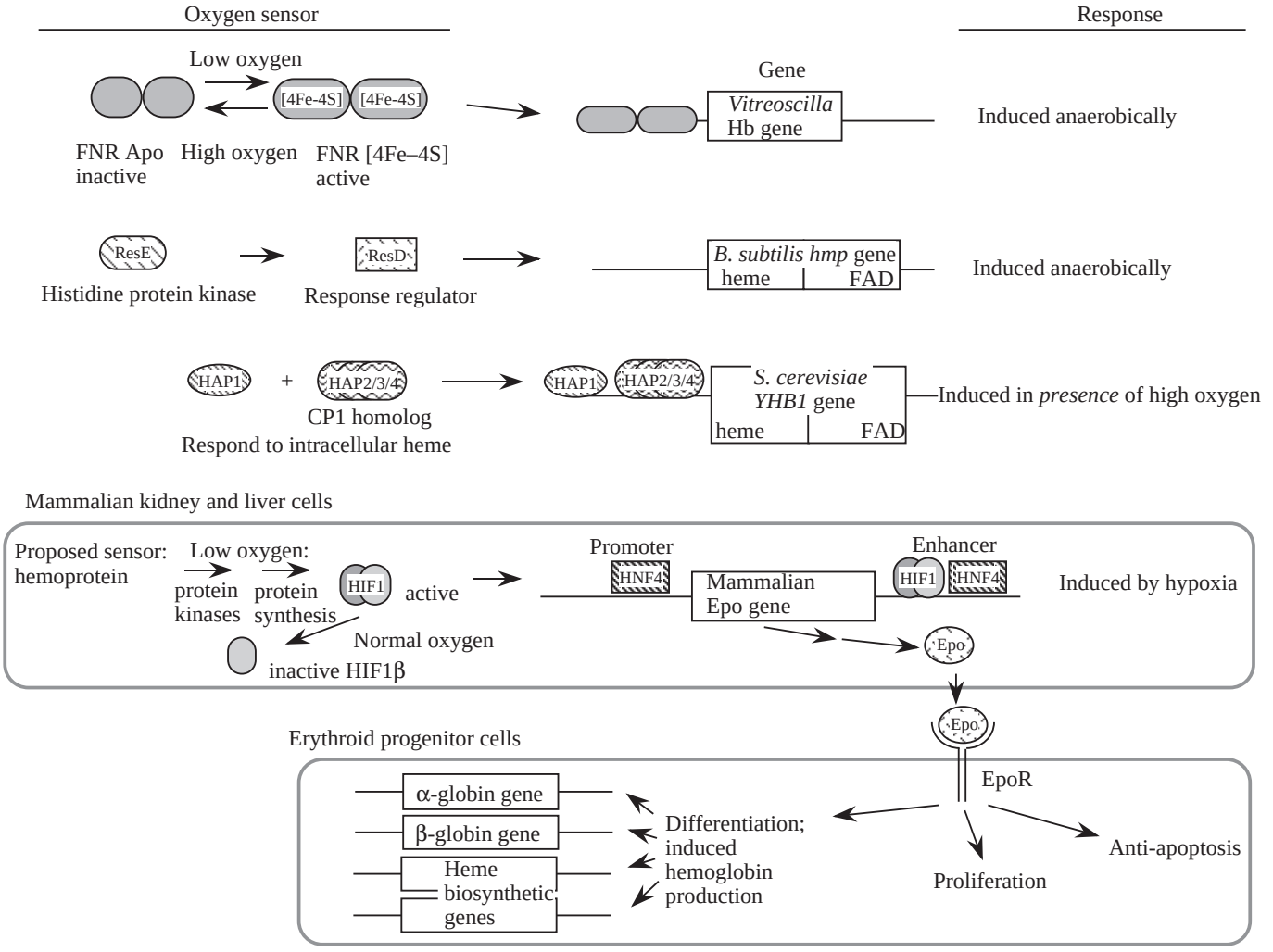

Fig. 4. Different types of oxygen sensors and their effects on regulation of hemoglobin genes in bacteria and mammals.

hypoxic stress (Wang and Semenza, 1993). HIF1 is a heterodimer (Wang *et al.* 1995), and the primary regulation is exerted on the synthesis and stability of the HIF1α subunit (Huang *et al.* 1996). The HIF1β subunit is identical to the ARNT protein, the aryl hydrocarbon receptor nuclear transport protein; its concentration does not change in response to O_2 (Wang *et al.* 1995). However, the HIF1α subunit is synthesized under low-O_2 conditions in a process that appears to be under translational control. Under normal-O_2 conditions, the concentration of the HIF1α subunit declines rapidly, leading to a decrease in erythropoietin production. The protein HNF4 also plays a critical positive role in the tissue-specific and hypoxia-inducible expression of the erythropoietin (Epo) gene (Galson *et al.* 1995). The nature of the O_2 sensor in kidney and liver cells is still not defined, although some studies have implicated a hemoprotein in this process (Goldberg *et al.* 1988; Huang *et al.* 1997). In particular, one model is that the oxidation state of the Fe determines the conformation of the heme and that the conformation with reduced Fe signals the pathway leading to increased erythropoietin production.

The flavohemoglobin gene *YHB1* from the yeast *Saccharomyces cerevisiae* is also regulated by oxygen, but in the opposite way – it is induced by high levels of O_2. The HAP1 and HAP2/3/4 proteins have been implicated in this aerobic induction (Crawford *et al.* 1995); these proteins respond to intracellular heme concentrations. The yeast HAP2 and HAP3 proteins are homologous to two subunits of the heteromeric CCAAT-binding protein CP1 (also known as NF-Y and CBF), which are implicated in activated expression of all the mammalian globin genes (discussed below). Thus, despite the dissimilarities in the oxygen response and the greater complexity of the mammalian mechanism, homologous transcription factors are implicated in the regulation of homologous genes in yeast and mammals.

Little commonality is obvious from these disparate regulatory systems, but in only one example has a protein with a demonstrated capacity to regulate gene expression by sensing O_2 levels, the FNR protein, been placed in the pathway. Proteins containing Fe–S clusters can respond reversibly to changes in oxidative conditions, whether by increasing the stability of a 4Fe–4S cluster in the case of FNR and the mammalian IRP1, or iron response protein (Rouault and Klausner, 1996), or by changing the conformation of a stable 2Fe–2S cluster in the SoxR protein. This latter protein binds to cognate sites in the promoter of the *soxS* gene under both anaerobic and aerobic conditions, but it changes its conformation (and apparently that of the promoter) under aerobic conditions to increase expression of the *soxS* gene (Hidalgo *et al.* 1997). The resulting increase in concentration of the SoxS protein induces expression of many genes involved in protection from oxidative stress. It is tantalizing to speculate that the initial monitors of O_2 levels (O_2 sensors) could be Fe–S proteins in many of the regulatory systems discussed here. The hemoproteins implicated in induction of erythropoeitin in mammals and *YHB1* in yeast could be acting downstream of the initial sensor. Further studies should test this possibility. It should be noted that *oxyR* regulation in *E. coli* responds to low oxygen concentrations without the involvment of an Fe–S cluster protein, so other types of oxygen sensor molecules are known (Storz *et al.* 1990). Even if Fe–S proteins are implicated more broadly in O_2 sensing, the large evolutionary distance between the species examined may preclude a clear determination of any ancestral relationships.

In contrast, evolutionary approaches provide a means to analyze the substantial amount of information available about the regulatory elements of hemoglobin genes in plants and animals. These show distinctive features that illustrate how DNA sequences have evolved to allow different homologous coding sequences (hemoglobin genes) to be expressed in different tissues, at different stages of development and at differing levels.

Plant hemoglobins

The leghemoglobin genes are expressed only in the nitrogen-fixing root nodules after the symbiotic bacteria have invaded, and they are expressed at high levels. In constrast, the nonsymbiotic hemoglobin genes are expressed in all tissues examined but at lower levels. The currently defined promoter elements for the two classes of gene are shown in Fig. 5. The promoters for leghemoglobin genes have 'nodulin boxes' that are critical for nodule-specific expression (Ramlov *et al.* 1993; Szczyglowski *et al.* 1994), but the promoters of genes for the nonsymbiotic plant hemoglobins lack this motif, having instead their own common conserved motifs (Andersson *et al.* 1996). One model for the role of these nodulin boxes is that specific activator proteins bind to these sequences in nodules, leading to high levels of expression of the leghemoglobin genes. Research into the regulation of the nonsymbiotic hemoglobin genes is at an early stage, and it will be most informative to investigate any similarities with the regulation in other species. The analysis of upstream promoter sequences has been useful in these plant hemoglobin genes (Table 2; Fig. 5).

Paradoxes in vertebrate globin evolution: α-globin versus *β-globin gene regulation*

An enormous amount of research has been devoted to understanding the regulation of vertebrate hemoglobin genes, including tissue- and developmental-stage-specificity and balanced production of the globin chains. Given the descent of α- and β-globin genes from a common ancestor (Figs 3, 6), one might have thought that their coordinated and balanced expression to produce the heterotypic tetramer $\alpha_2\beta_2$ in erythrocytes would be the easiest aspect of regulation to explain. Since the two genes would have been identical after the initial duplication, with identical regulatory elements, one might have expected selection to keep the regulatory elements very similar. However, much has changed between the α-like and β-like globin gene clusters since their duplication. They are now on separate chromosomes in birds and mammals, and in mammals they are in radically different genomic contexts (Fig. 7). The β-globin gene clusters have an A+T content

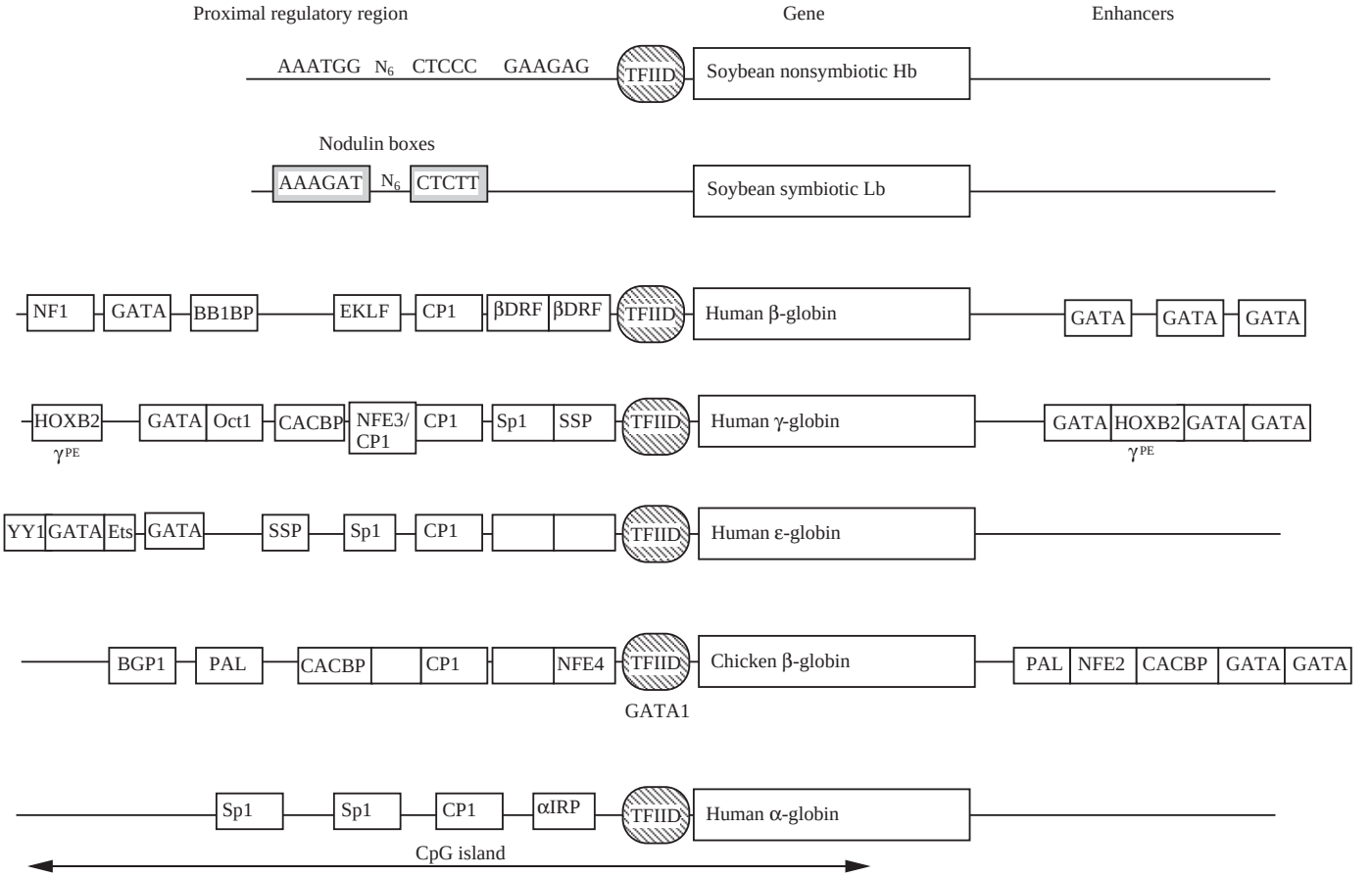

Fig. 5. Summary of protein-binding sites in globin gene promoters from plants and mammals. DNA sequences implicated in the regulation of the globin gene promoters in plants are given. For the human and chicken globin genes, boxes in the 5′ and 3′ flanking regions are labeled with the proteins that bind to these sites and are implicated in regulation. Unlabeled boxes are segments of DNA either conserved and/or to which proteins bind, but the identity of the proteins acting at these sites is unknown.

comparable to the bulk of mammalian DNA, they have no CpG islands, and the locus is in a DNAase-accessible, 'open' chromatin conformation only in erythroid cells, where the genes are expressed (reviewed in Collins and Weissman, 1984). In contrast, the α-like globin gene clusters are highly G+C-rich, they have a CpG island associated with each active gene, and the locus is in a constitutively 'open' chromatin conformation in all cells (Craddock *et al.* 1995). Tissue-specific gene expression is frequently correlated with an increased accessibility of the chromatin in a locus only in expressing cells, but this is not the case for the α-like globin genes of mammals.

Despite all these differences, expression of the α-globin and β-globin genes is appropriately balanced in erythroid cells, apparently by rather different mechanisms. Fig. 5 shows some of the better-characterized protein-binding sites in the proximal regulatory elements (roughly the 200 base pairs 5′ to the cap site of the genes). Comparing the human β- and α-globin genes (Efstratiadis *et al.* 1980), one sees the TATA motif, to which the general transcription factor TFIID binds, and the CCAAT motif, to which *trans*-activators such as CP1 can bind. However, other protein-binding sites are quite different. In particular, the 5′ flanking region and much of the α-globin gene are contained within a CpG island, which contains notable binding sites for Sp1, a relative of Sp1 called αIRP, and other less well-characterized proteins (Barnhart *et al.* 1988; Kim *et al.* 1988; Yost *et al.* 1993; Rombel *et al.* 1995). Aside from the TATA and CCAAT motifs, the protein-binding sites are completely different in the proximal regulatory region of the human β-globin gene. The CpG island encompassing the 5′ flanking region and much of the gene is a key component of the *cis*-regulatory elements for the α-globin gene of rabbits and humans, possibly through its effects on chromatin structure (Pondel *et al.* 1995; Shewchuk and Hardison, 1997), but no CpG island is found at any of the β-like globin genes.

As can be seen in Figs 2 and 3, the mammalian α- and β-globin genes are relatively close to each other on a phylogenetic scale that includes many taxa (animals, plants, fungi and bacteria), but this time frame is in fact quite long relative to the evolution of regulatory elements. Currently, the coordinated regulation of α- and β-globin genes is more paradoxical than clear. Differences are even seen between the distal elements that regulate the α-globin and β-globin gene clusters, called the locus control region (or LCR) for β-globin genes (reviewed in Grosveld *et al.* 1993)

and HS-40 for the α-globin genes (Higgs *et al.* 1990). As illustrated in Fig. 8, one powerful enhancing region of the β-globin LCR, called DNAase hypersensitive site 2, or HS2, and α-globin HS-40 are each composed of binding sites for transcription factors NFE2 (a member of the AP1 family), GATA1 or its relatives and a family of proteins that bind to the DNA sequences that include a CACC motif, generically referred to as CACBPs (Talbot *et al.* 1990; Jarman *et al.* 1991). In both cases, these distal regulatory sites cause a large increase in the level of expression of the target genes (Fig. 7). However, the similarities appear to end there. HS2 serves as an enhancer within the context of a much larger β-globin LCR, which also acts to open the chromatin over a discrete locus in erythroid cells, whereas HS-40 is currently the only characterized erythroid-specific, distal regulator, and it enhances globin gene expression within a locus that is part of a large block of constitutively active chromatin, which also includes several ubiquitously expressed genes. Thus, domain opening does not appear to play a role in regulation of α-globin genes (Craddock *et al.* 1995), but it is a key initial step in the regulation of β-globin genes (Groudine *et al.* 1983; Forrester *et al.* 1990).

Further insights into the evolution of coordinated expression between α- and β-globin genes may be gleaned by further analysis of the globin gene clusters in the amphibian *Xenopus* (Hosbach *et al.* 1983) or the zebrafish *Danio rerio* (Chan *et al.* 1997), in which the α-globin genes and β-globin genes are still closely linked. For instance, it would be very helpful to know the location and composition of the LCR in these cases.

Avian and mammalian β-globin genes

Since comparisons of mammalian α- and β-globin genes, whose ancestors diverged early in the vertebrate lineage approximately 450 million years ago, show more differences than similarities, one might expect comparisons over a shorter phylogenetic distance to reveal information about regulation, such as alignments within a gene lineage but between families of vertebrates (Fig. 6). Extensive sequences are available for both mammalian (e.g. human) and avian (chicken) β-globin gene clusters, and in both cases the genes are expressed only in erythroid cells in a developmentally regulated manner. One might anticipate common aspects of regulation, and in general this is true. Both human and chicken β-globin gene clusters are in an 'open' chromatin domain only in erythroid cells, and DNAase hypersensitive sites (HSs) appear in the promoters only at the developmental stage at which the gene is expressed (reviewed in Evans *et al.* 1990; Felsenfeld, 1993). Thus, alterations in both overall and specific chromatin structure are critical to the regulation of both gene clusters. In both species, the expression of the genes is controlled by both distal and proximal regulatory sequences. One may further anticipate that these common features of gene organization and regulation would be reflected in sequence comparisons, but that is largely not the case.

A comparison of the complete sequences of the chicken (Reitman *et al.* 1993) and human (Collins and Weissman, 1984) β-globin gene clusters shows a simple and somewhat disappointing pattern (Fig. 9, lower panel). The sequence

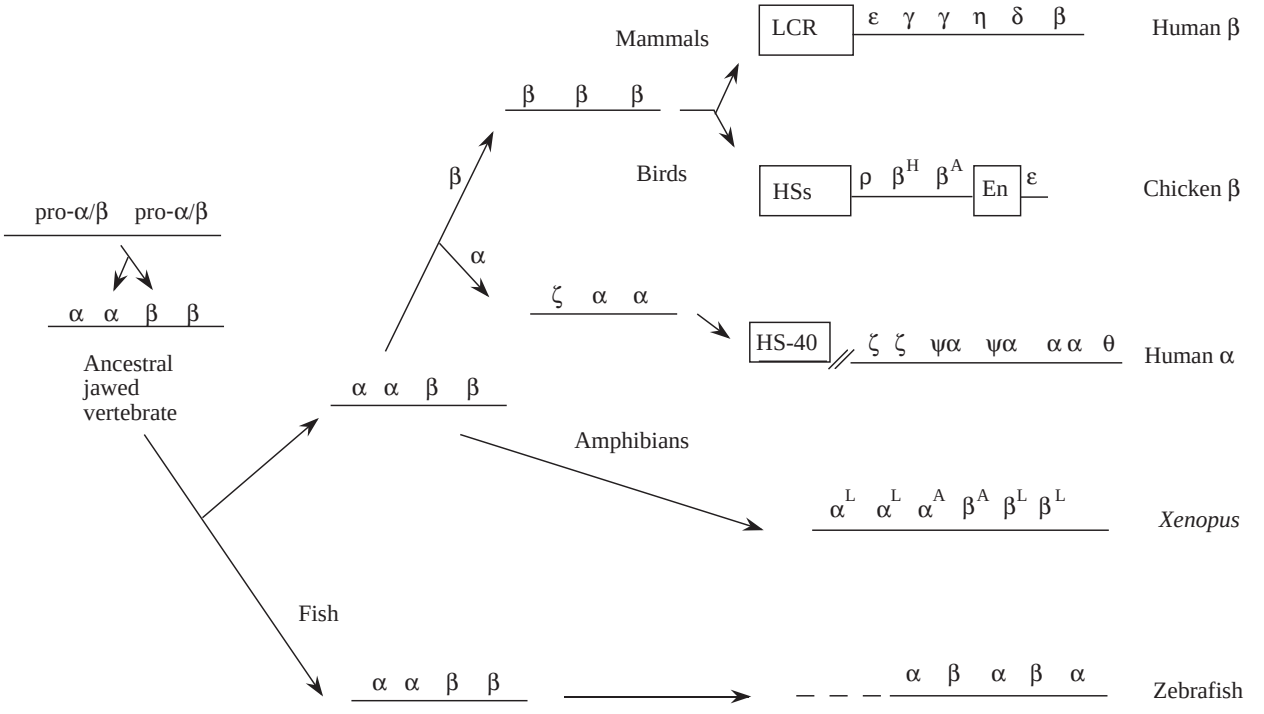

Fig. 6. Evolution of globin gene clusters in vertebrates. Each Greek letter represents a globin gene. Although all the globin genes in a contemporary gene cluster may be derived from a single gene in an ancestor, this does not preclude the possibility that the ancestor had multiple globin genes that may have been differentially regulated. Thus, a cluster of genes is shown in each ancestor.

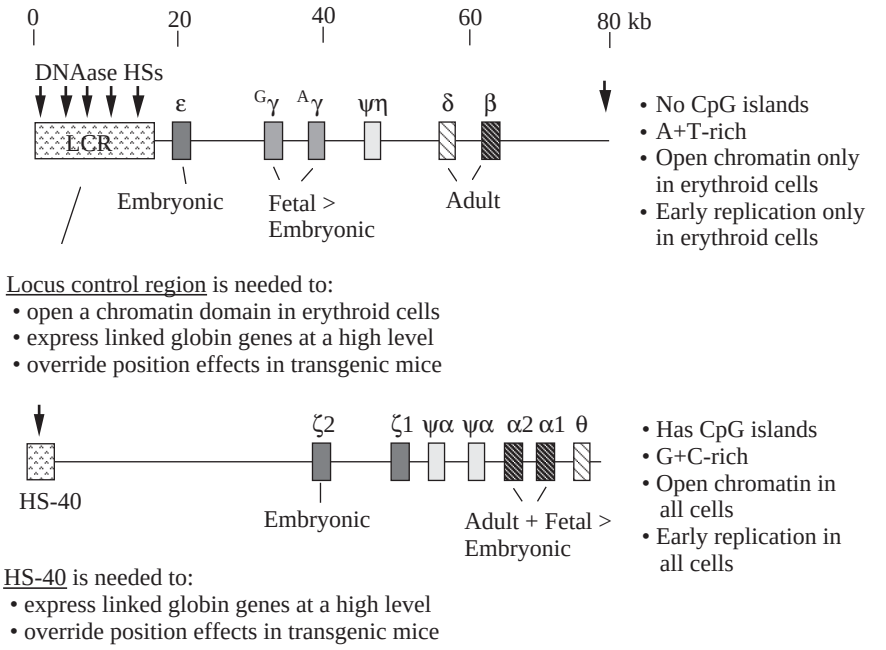

Fig. 7. Summary of the organization, genomic DNA context, chromatin structure and distal regulatory elements of human β- and α-globin gene clusters.

matches are restricted to portions of the protein-coding regions. Each β-related globin gene in humans is equally distant from each of the β-related globin genes in chickens, e.g. the human ε-globin gene is no more closely related to the chicken ε-globin gene than it is to the adult β^A-globin gene, despite having the same name. This suggests that the series of gene duplications and divergences that gave rise to the β-globin gene clusters occurred independently in the lineages to the ancestral mammal and to the ancestral bird, and that is consistent with the inferences drawn from phylogenetic reconstructions based on the amino acid sequences of the proteins (Fig. 6).

No statistically significant alignments are seen in the promoter regions or in the distal control elements. The chicken β/ε enhancer shows no striking matches to any portion of the human β-globin gene cluster, nor does the human β-globin LCR resemble any of the sequences around the 5′ HSs in the chicken gene cluster. At least for this particular gene cluster, the comparison between birds and mammals is too distant to discern candidates for *cis*-regulatory sequences by pairwise sequence alignments. The situation may be different for the α-globin gene clusters, since the human ζ-globin and avian π-globin genes are orthologous and restricted to embryonic erythroid expression (Proudfoot *et al.* 1982). A complete DNA sequence of the chicken α-globin gene cluster including the region comparable to HS-40 will be highly informative.

Even though the pairwise alignments did not reveal sufficiently long matching segments to be significant, a comparison of the protein-binding sites in known regulatory regions shows that some of the same proteins are used in both species. For instance, the chicken β/ε enhancer has binding sites for AP1/NFE2, a CACBP and GATA1 (Evans *et al.* 1990), strikingly similar to the array in HS2 of the human β-globin LCR and the α-globin HS-40 (Fig. 8). Alignments of long DNA sequences are not expected to be able to identify single, isolated bindings sites simply from the similarity score. A typical protein-binding site is usually 6–8 base pairs long, and some variation in the binding sites can occur without affecting binding affinity (and hence could be tolerated even in a region under strong selection). Thus, a functional binding site could comprise a string of as few as six nucleotides, only

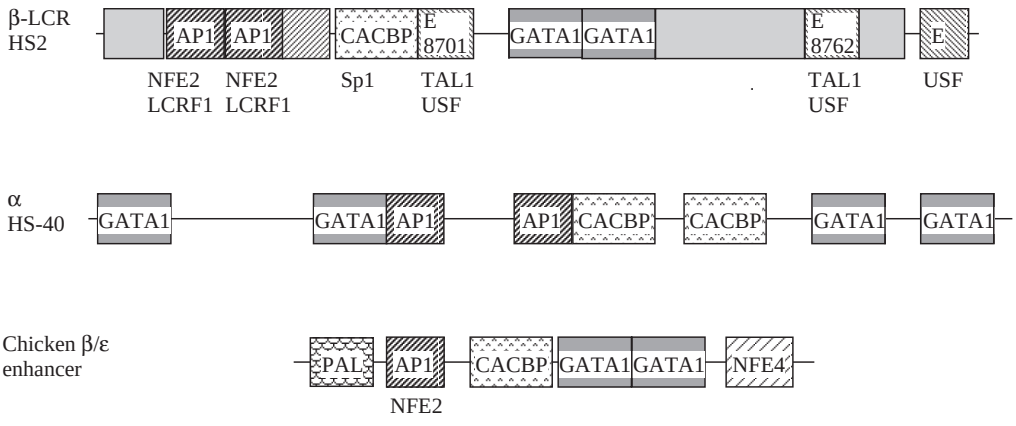

Fig. 8. Summary of protein-binding sites in enhancers and distal control elements of vertebrate globin genes. Boxes of distinctive shading are labeled with the proteins or classes of proteins that bind to these DNA sequences. For HS2 of the β-globin LCR, highly conserved segments for which no protein has been identified as binding are unlabeled.

four of which match in a pairwise comparison. Strings meeting that criterion occur randomly at too high a frequency to allow one to distinguish functional binding sites from random matches. However, a group of conserved binding sites could score as a significant alignment if the order and spacing are also conserved. In the three cases discussed here, the order NFE2/AP1–CACBP–GATA1 is the same, but the spacing differs. The inability to detect similar regulatory regions between chicken and human using nucleotide identities as the basis for a similarity score illustrates the need for the development of software that identifies all potential protein-binding sites and searches for similar patterns within these binding sites. This also has been espoused as a good approach for analyzing sequences between the pufferfish *Fugu* and humans (Aparicio *et al.* 1995).

The general result is that homologous proteins are playing important, and probably similar, roles in the regulation of the β-globin gene clusters in birds and mammals, even though the pairwise alignments do not reveal these as conserved elements. Hence, the comparison of transcription factor binding sites is more informative than the alignment of non-coding DNA sequences (Table 2).

Mammalian γ- and β-globins: phylogenetic footprinting and differential phylogenetic footprints

In contrast to the previous comparisons, the detailed study of globin gene clusters in many mammalian species has provided a rich resource of information from which to glean further insight not only into the evolution of the gene clusters but also into their regulation. The β-globin gene clusters have been extensively studied in human, the prosimian galago, the lagamorph rabbit, the artiodactyls goat and cow, and the rodent mouse. Diagrams of these gene clusters are shown in Fig. 10, and aspects of their evolution and regulation have been

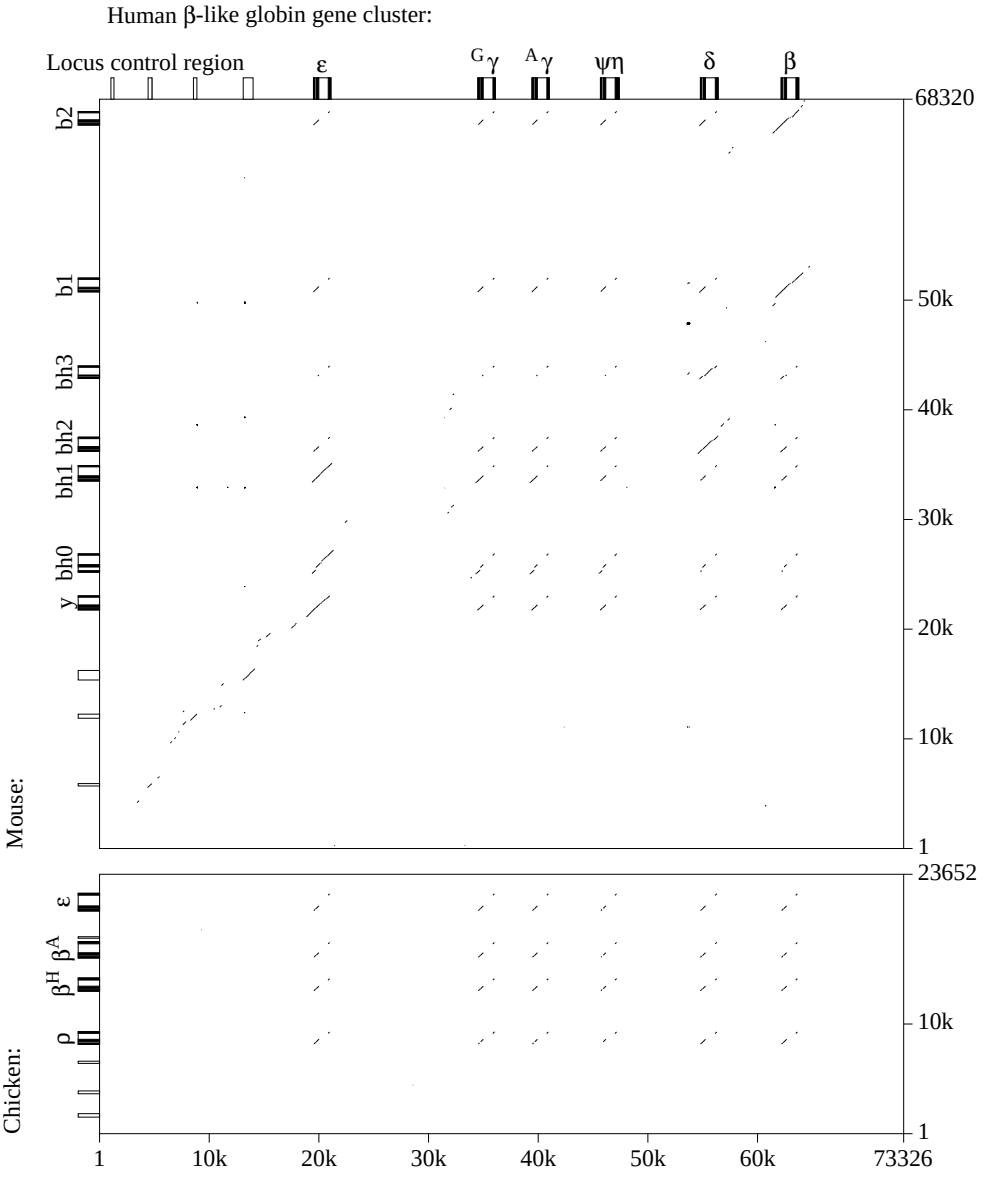

Fig. 9. Plot of positions of aligning sequences in comparisons between the β-globin gene clusters of human and mouse (top panel) and human and chicken (bottom panel). All local alignments between each pair of DNA sequences that score above a certain objective criterion were computed using the program SIM (Huang *et al.* 1990), those involving interspersed repeats were masked and the positions of the aligning segments are plotted. The axes are marked with genes (with three filled boxes for exons and two open boxes for introns) and DNAase hypersensitive sites (open boxes that are not juxtaposed to exons) associated with the locus control region in mammals (5′ to the human ε-globin gene and the mouse y gene encoding an ε-globin) and an enhancer (3′ to the β^A-globin gene) and upstream distal regulatory elements in chicken. This image was supplied by Dr W. Miller.

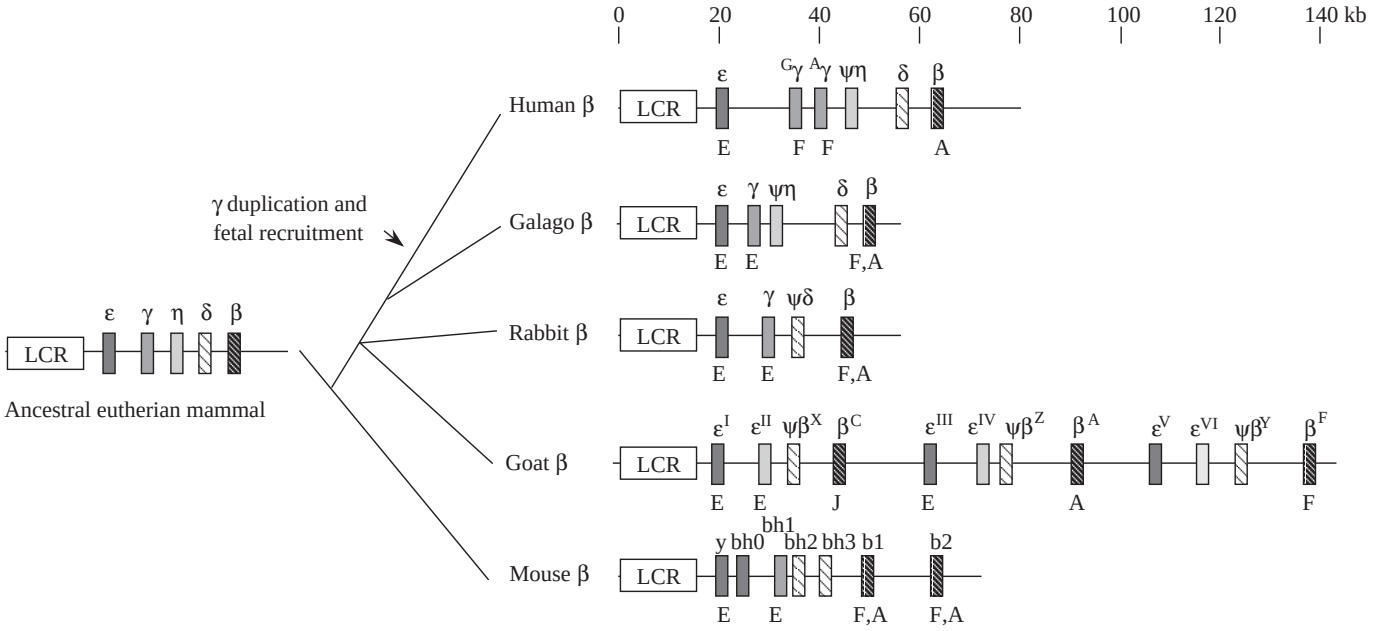

Fig. 10. Proposed evolution of β-globin gene clusters in eutherian mammals. The genes are shown as boxes, with their names above and their time of expression during development indicated below each box. Orthologous genes have the same distinctive shading in each box. E, embryonic; F, fetal; A, adult; J, juvenile.

reviewed (Collins and Weissman, 1984; Goodman *et al.* 1987; Hardison, 1991; Hardison and Miller, 1993). The ε-globin gene is at the 5′ end of all the mammalian globin gene clusters and is expressed only in embryonic red cells. In most species, expression of the γ-globin gene is also limited to embryonic red cells, but in anthropoid mammals its expression continues and predominates in fetal red cells. The appearance of this new pattern of fetal expression of the γ-globin genes coincides roughly with the duplication of the genes in primate evolution, which leads to the hypothesis that the duplication allowed the changes that caused the fetal recruitment (Fitch *et al.* 1991). The β-globin gene is expressed after birth in all mammals, but in galago, mouse and rabbit its expression initiates and predominates in the fetal liver (arguing that fetal expression of the β-globin gene is the ancestral state). The recruitment of γ-globin genes for fetal expression in anthropoid primates is accompanied by a corresponding delay in expression of the β-globin gene.

Both the invariant patterns in gene regulation in mammals as well as the changes in expression pattern of the anthropoid γ- and β-globin genes should be reflected in changes in *cis*-acting DNA sequences regulating the expression of the genes. The invariant patterns should be reflected in DNA sequences in the control regions that change very slowly over evolutionary time, which are recognizable as conserved sequence blocks or phylogenetic footprints. Comparisons of the DNA sequences of entire globin gene clusters reveal regions of high similarity extending for over 1000 base pairs 5′ to the orthologous genes and in long regions throughout the distal LCR (Li *et al.* 1990; Hardison and Miller, 1993; Hardison *et al.* 1997*b*; Slightom *et al.* 1997), as illustrated for the human–mouse sequence comparison in Fig. 9 (top panel). The LCR has been implicated in opening the chromatin in the β-globin gene domain to allow or stimulate high-level gene expression (Fig. 7), and the 5′ flanking regions contain the proximal promoters (to approximately −100) plus upstream regulators involved in induction and silencing (reviewed in Stamatoyannopoulos and Nienhuis, 1994). Thus, the overall pattern of sequence conservation outside the coding regions includes most of the currently characterized regulatory regions, although in many cases the sequence conservation extends beyond them, indicating the possibility that more will be discovered. This general pattern of sequence conservation between eutherian mammals (human and mouse in particular) localized to known or testable regulatory regions is seen for several other mammalian loci, demonstrating the utility of this general approach (Hardison *et al.* 1997*a*).

Some of the protein-binding sites flanking the globin genes are shown in more detail in Fig. 5. The TATA box (the binding site for TFIID), the CCAAT box (the binding site for CP1 and other families of proteins) and the CACC box (the binding site for EKLF) were initially recognized as conserved regions (Efstratiadis *et al.* 1980; Lacy and Maniatis, 1980), as was the DRE in mammalian β-globin genes (Stuve and Myers, 1990). Important sequence motifs at approximately −160 were noted as conserved (Hardison, 1983) prior to the discovery of the proteins binding to them, such as the GATA1-binding site in the promoters for ε- and γ-globin genes (Tsai *et al.* 1989; Gong *et al.* 1991) and the BB1-binding site in the promoter for β-globin genes (Antoniou *et al.* 1988; Macleod and Plumb, 1991). In other cases, such as the AP1/NFE2 binding sites and the GATA1 sites in the locus control regions and enhancers of globin genes, specific binding and evidence of conservation were discovered at approximately the same time (deBoer *et al.*

1988; Ney *et al.* 1990; Talbot *et al.* 1990). Even in extensively studied regions, such as HS2 of the LCR, a conserved E box sequence was the initial observation (Hardison *et al.* 1993) that led to the recent discovery of the importance of this region in full enhancement by this element, *via* the action of basic helix–loop–helix proteins such as TAL1 and USF and/or a novel factor called HS2NFE5 (Lam and Bresnick, 1996; Elnitski *et al.* 1997). Many of the protein-binding sites in the proximal regulatory regions of the human globin genes are conserved in the orthologous genes in all mammals examined, in keeping with important roles in regulation and illustrative of the power of the phylogenetic approach.

Differences in the patterns of expression can be analyzed by a differential phylogenetic footprinting approach (Gumucio *et al.* 1994). One striking example is a region in the γ-globin gene promoter that is conserved in anthropoid primates, but is different in other mammals. This is a binding site for a factor called the stage-selector protein, or SSP (Fig. 5), that has been implicated in the differential expression of γ-globin and β-globin genes (Jane *et al.* 1992). SSP is a heterodimer (Jane *et al.* 1995) between CP2 (Lim *et al.* 1992) and some other protein. Interestingly, the NFE4 protein implicated in the stage-specific expression of the chicken β-globin gene (Foley and Engel, 1992) also contains CP2 as part of a heterodimer (Jane *et al.* 1995). Thus, this approach of looking for patterns of conservation consistent with differential gene regulation in eutherian mammals has indeed led to the discovery of a protein that is probably involved in that differential expression.

Another level of differential analysis compares the proximal regulatory regions of genes expressed at different times of development, i.e. mammalian ε-, γ- and β-globin genes. As illustrated in Fig. 5, most of the protein-binding sites are different in these promoters. For instance, βDRF, EKLF and BB1BP have been implicated only in the regulation of the β-globin gene (Evans *et al.* 1990). A similar but distinctive CACC motif is found in a comparable position in the 5′ flank of all three genes, but EKLF is active only at the β-globin CACC box (Donze *et al.* 1995; Perkins *et al.* 1995), leaving open the important possibility that other CACBPs, perhaps active only at one developmental stage, are regulating γ- and ε-globin genes. A GATA-binding site is conserved at approximately the same position in both γ- and ε-globin gene 5′ flanking regions, but the comparable region for β-globin does not have a conserved GATA site (Hardison *et al.* 1994). Even conserved DNA sequence motifs with very similar sequences may serve as binding sites for different proteins. A CCAAT motif located at approximately −80 in all the vertebrate globin gene promoters can be bound by a heteromeric complex called CP1, NF-Y or CBF (Hooft van Huijsduijnen *et al.* 1990, and references therein). However, preparations of CP1 bind much more strongly to the CCAAT box in the α-globin gene promoter than in the β-globin gene promoter (Cohen *et al.* 1986). Also, multiple additional proteins bind to the CCAAT box, some of which have been implicated in the activation of β-globin gene expression (deBoer *et al.* 1988; Delvoye *et al.* 1993).

Thus, non-coding DNA sequence alignments of these groups of orthologous genes in different eutherian mammals reveal protein-binding sites important for regulated expression (Table 2). The differences in the arrays of proteins functioning at ε-, γ- and β-globin genes indicate that a distinctive battery of proteins functions in the promoter for each type of gene. Indeed, this implication is consistent with the observation that *cis*-acting sequences needed for stage-specific regulation of expression map close to the genes (Trudel and Costantini, 1987, and references therein). In the context of the evolutionary tree shown in Fig. 6, these results show that, for the globin gene clusters, comparisons among orthologous genes that share a common ancestor early in the eutherian lineage are useful for revealing conserved *cis*-regulatory sites (mostly protein-binding sites). However, comparisons between paralogous genes resulting from gene duplications as recently as the divergence between the ancestor to both β-globin and ε-globin genes in the mammalian lineage do not reveal common regulatory elements. Hence, it is not surprising that sequence comparisons between genes whose ancestor diverged even earlier, e.g. mammalian *versus* avian β-globin genes, or mammalian α- *versus* β-globin genes, do not reveal matches in regulatory elements.

These conclusions apply equally well to distal regulatory elements such as the LCR. Comparisons of orthologous sequences among eutherian mammals show highly conserved sequences throughout the β-globin LCR, and these correlate precisely with regulatory elements (reviewed in Hardison *et al.* 1997*b*). Some parts of the LCR, in particular those homologous to DNAase hypersensitive sites 1, 2 and 3, are conserved in the marsupial and monotreme mammals (R. Hope, R. Baird, J. Kulibawa and M. Goodman, personal communication). As the conservation of the LCR is mapped even more deeply on an evolutionary tree, the issue of its origin comes into tighter focus. This important element has been implicated in initiating and maintaining an open chromatin domain in the otherwise highly repressed nucleus of erythroid cells. Thus, is may have arisen around the time that vertebrate erythrocytes evolved to carry large amounts of hemoglobin. One might expect similar sequences to be detectable in many vertebrates but, in fact, homologous sequences were not seen even in the avian–mammalian comparison, despite the fact that functionally analogous DNA sequences are known. Since homologous LCRs are seen early in the mammalian lineage (prior to the eutherian–metatherian split), but they are not detected in avians, the question arises as to whether the avian (or mammalian) LCRs have been rearranged to the point where they are no longer detectable in these comparisons or whether the distance is just too great. Analysis of more intermediate species would help answer this question.

Identification of conserved and differentially conserved sequences is an important guide to identifying *cis*-regulatory sequences and is helpful in finding the many proteins involved in regulation of the genes. This approach, plus many studies on the function of these sequences, has revealed a highly complex array of regulatory sequences and proteins. However,

sequence analysis provides little insight into the important issues of how these DNA sequences, with proteins bound, work together to accomplish the several levels of regulation. Fundamental questions about the mechanism of action of the LCR in domain opening, the identification and mechanisms of proteins required for developmental control and the possibility of interaction between the promoters and the LCR (and hence possible competition as a mechanism for regulation) remain unresolved (Tuan *et al.* 1992; Martin *et al.* 1996; Wijgerde *et al.* 1996) and require further study.

Concluding remarks

DNA and protein sequence comparisons and alignments allow one to apply the principles of evolutionary biology to learn much about genes and their regulation, but one needs to compare sequences from species or genes separated by an appropriate distance to obtain useful information. In some cases, the amino acid sequences may be so different that comparisons of three-dimensional structures are needed to deduce truly ancient relationships, e.g. among different hemoproteins such as a hemoglobin, ligninases, cytochromes, etc. Comparisons of the amino acid sequences of proteins are highly informative within a family, with members ranging from bacteria to mammals. Comparisons of gene structure are clearly informative for globin genes only from the ancestor to plants and animals, although the information in gene structures from protists needs more analysis. Deductions on gene regulation based on sequence analysis between different vertebrate families, such as birds and mammals, may need the development of new software analyzing protein-binding sites. Alignments of non-coding DNA sequences in a group of mammals are highly informative about regulatory elements, but similar analyses between birds and mammals have been uninformative, at least for the β-globin gene cluster. Thus, depending on the type of question being asked, sequence or structural comparisons will be informative, but the appropriate phylogenetic distance needs to be employed. The choice of species will probably need to be varied for different loci, given the differences in evolutionary rates for various loci, but for many mammalian loci, comparisons between human and mouse are highly informative for studies of regulatory regions.

Work from this laboratory was supported by PHS grants 1RO1 DK27635, 1RO1 LM05773 and 1RO1 LM05110. I thank Dr W. Miller for Fig. 9.

References

ANDERSSON, C. R., JENSEN, E. O., LLEWELLYN, D. J., DENNIS, E. S. AND PEACOCK, W. J. (1996). A new hemoglobin gene from soybean: A role for hemoglobin in all plants. *Proc. natn. Acad. Sci. U.S.A.* **93**, 5682–5687.

ANTOINE, M. AND NIESSING, J. (1984). Intron-less globin genes in the insect *Chironomus thummi*. *Nature* **310**, 795–798.

ANTONIOU, M., DEBOER, E., HABETS, G. AND GROSVELD, F. (1988). The human β-globin gene contains multiple regulatory regions: Identification of one promoter and two downstream enhancers. *EMBO J.* **7**, 377–384.

APARICIO, S., MORRISON, A., GOULD, A., GILTHORPE, J., CHAUDHURI, C., RIGBY, P., KRUMLAUF, R. AND BRENNER, S. (1995). Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc. natn. Acad. Sci. U.S.A.* **92**, 1684–1688.

APPLEBY, C. A. (1984). Leghemoglobin and *Rhizobium* respiration. *A. Rev. Plant Physiol.* **35**, 443–478.

APPLEBY, C. A., TJEPKEMA, J. D. AND TRINICK, M. J. (1983). Hemoglobin in a nonleguminous plant, *Parasponia*: Possible genetic origin and function in nitrogen fixation. *Science* **220**, 951–953.

BARNHART, K., KIM, C., BANERJI, S. AND SHEFFERY, M. (1988). Identification and characterization of multiple erythroid cell proteins that interact with the promoter of the murine α-globin gene. *Molec. cell. Biol.* **9**, 3215–3226.

BOGUSZ, D., APPLEBY, C. A., LANDSMANN, J., DENNIS, E. S., TRINICK, M. J. AND PEACOCK, W. J. (1988). Functioning haemoglobin genes in non-nodulating plants. *Nature* **331**, 178–180.

BRISSON, N. AND VERMA, D. P. (1982). Soybean leghemoglobin gene family: normal, pseudo and truncated genes. *Proc. natn. Acad. Sci. U.S.A.* **79**, 4055–4059.

CHAN, F. Y., ROBINSON, J., BROWNLIE, A., SHIVDASANI, R. A., DONOVAN, A., BRUGNARA, C., KIM, J., LAU, B. C., WITKOWSKA, H. E. AND ZON, L. I. (1997). Characterization of adult alpha- and beta-globin genes in the zebrafish. *Blood* **89**, 688–700.

COHEN, R. B., SHEFFERY, M. AND KIM, C. G. (1986). Partial purification of a nuclear protein that binds to the CCAAT box of the mouse α1-globin gene. *Molec. cell. Biol.* **6**, 821–832.

COLLINS, F. S. AND WEISSMAN, S. M. (1984). The molecular genetics of human hemoglobin. *Prog. Nucleic Acids Res. molec. Biol.* **31**, 315–462.

COUTURE, M., CHAMBERLAND, H., ST-PIERRE, B., LAFONTAINE, J. AND GUERTIN, M. (1994). Nuclear genes encoding chloroplast hemoglobins in the unicellular green alga *Chlamydomonas eugametos*. *Molec. gen. Genet.* **243**, 185–197.

COUTURE, M. AND GUERTIN, M. (1996). Purification and spectroscopic characterization of a recombinant chloroplastic hemoglobin from the green unicellular alga *Chlamydomonas eugametos*. *Eur. J. Biochem.* **242**, 779–787.

CRADDOCK, C. F., VYAS, P., SHARPE, J. A., AYYUB, H., WOOD, W. G. AND HIGGS, D. R. (1995). Contrasting effects of alpha and beta globin regulatory elements on chromatin structure may be related to their different chromosomal environments. *EMBO J.* **14**, 1718–1726.

CRAMM, R., SIDDIQUI, R. A. AND FRIEDRICH, B. (1994). Primary structure and evidence for a physiological function of the flavohemoprotien of *Alcaligenes eutrophus*. *J. biol. Chem.* **269**, 7349–7354.

CRAWFORD, M. J., SHERMAN, D. R. AND GOLDBERG, D. E. (1995). Regulation of the *Saccharomyces cerevisiae* flavohemoglobin gene. *J. biol. Chem.* **270**, 6991–6996.

DEBOER, E., ANTONIOU, M., MIGNOTTE, V., WALL, L. AND GROSVELD, F. (1988). The human β-globin promoter; nuclear protein factors and erythroid specific induction of transcription. *EMBO J.* **7**, 4203–4212.

DELVOYE, N. L., DESTROISMAISONS, N. M. AND WALL, L. A. (1993). Activation of the β-globin gene promoter by the locus control region correlates with binding of a novel factor to the CCAAT bos

in murine erythroleukemia cells but not in K562 cells. *Molec. cell. Biol.* **13**, 6969–6983.

DICKERSON, R. E. AND GEIS, I. (1983). *Hemoglobin: Structure, Function, Evolution and Pathology*. Menlo Park, CA: The Benjamin/Cummings Publishing Co., Inc.

DIKSHIT, K. L., DIKSHIT, R. P. AND WEBSTER, D. A. (1990). Study of *Vitreoscilla* globin (*vgb*) gene expression and promoter activity in *E. coli* through transcriptional fusion. *Nucleic Acids Res.* **18**, 4149–4155.

DIKSHIT, R. P., DIKSHIT, K. L., LIU, Y. AND WEBSTER, D. A. (1992). The bacterial hemoglobin from *Vitreoscilla* can support aerobic growth of *Escherichia coli* lacking terminal oxidases. *Archs Biochem. Biophys.* **293**, 241–245.

DIXON, B. AND POHAJDAK, B. (1992). Did the ancestral globin gene of plants and animals contain only two introns? *Trends biochem. Sci.* **17**, 486–488.

DIXON, B., WALKER, B., KIMMINS, W. AND POHAJDAK, B. (1992). A nematode hemoglobin gene contains an intron previously thought to be unique to plants. *J. molec. Evol.* **35**, 131–136.

DONZE, D., TOWNES, M. M. AND BIEKER, J. J. (1995). Role of erythroid kruppel-like factor in human γ- to β-globin gene switching. *J. biol. Chem.* **270**, 1955–1959.

EDWARDS, S. L., RAAG, R., WARIISHI, H., GOLD, M. H. AND POULO, T. L. (1993). Crystal structure of lignin peroxidase. *Proc. natn. Acad. Sci. U.S.A.* **90**, 750–754.

EFSTRATIADIS, A., POSAKONY, J. W., MANIATIS, T., LAWN, R. M., O'CONNELL, C., SPRITZ, R. A., DERIEL, J. K., FORGET, B. G., WEISSMAN, S. M., SLIGHTOM, J. L., BLECHL, A. E., SMITHIES, O., BARALLE, F. E., SHOULDERS, C. C. AND PROUDFOOT, N. J. (1980). The structure and evolution of the human β-globin gene family. *Cell* **21**, 653–668.

ELNITSKI, L., MILLER, W. AND HARDISON, R. (1997). Conserved E boxes function as part of the enhancer in hypersensitive site 2 of the β-globin locus control region: Role of basic helix–loop–helix proteins. *J. biol. Chem.* **272**, 369–378.

ERMLER, U., SIDDIQUI, R. A., CRAMM, R. AND FRIEDRICH, B. (1995). Crystal structure of the flavohemoglobin from *Alcaligenes eutrophus* at 1.5 Angstrom resolution. *EMBO J.* **14**, 6067–6077.

EVANS, T., FELSENFELD, G. AND REITMAN, M. (1990). Control of globin gene transcription. *A. Rev. Cell Biol.* **6**, 95–124.

FELSENFELD, G. (1993). Chromatin structure and the expression of globin-encoding genes. *Gene* **135**, 119–124.

FITCH, D. H., BAILEY, W. J., TAGLE, D. A., GOODMAN, M., SIEU, L. AND SLIGHTOM, J. L. (1991). Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. *Proc. natn. Acad. Sci. U.S.A.* **88**, 7396–7400.

FOLEY, K. P. AND ENGEL, J. D. (1992). Individual stage selector element mutations lead to reciprocal changes in β- vs. ε-globin gene transcription: genetic confirmation of promoter competition during globin gene switching. *Genes Dev.* **6**, 730–744.

FORRESTER, W. C., EPNER, E., DRISCOLL, M. C., ENVER, T., BRICE, M., PAPAYANNOPOULOU, T. AND GROUDINE, M. (1990). A deletion of the human β-globin locus activation region causes a major alteration in chromatin structure and replication across the entire β-globin locus. *Genes Dev.* **4**, 1637–1649.

GALSON, D. L., TSUCHIYA, T., TENDLER, D. S., HUANG, L. E., REN, Y., OGURA, T. AND BUNN, H. F. (1995). The orphan receptor hepatic nuclear factor 4 functions as a transcriptional activator for tissue-specific and hypoxia-specific erythropoietin gene expression and is antagonized by EAR3/COUP-TF1. *Molec. cell. Biol.* **15**, 2135–2144.

GOLDBERG, D. E. (1995). The enigmatic oxygen-avid hemoglobin of *Ascaris*. *BioEssays* **17**, 177–182.

GOLDBERG, M. A., DUNNING, S. P. AND BUNN, H. F. (1988). Regulation of the erythropoietin gene: evidence that the oxygen sensor is a heme protein. *Science* **242**, 1412–1415.

GONG, Q.-H., STERN, J. AND DEAN, A. (1991). Transcriptional role of a conserved GATA-1 site in the human ε-globin gene promoter. *Molec. cell. Biol.* **11**, 2558–2566.

GOODMAN, M., CZELUSNIAK, J., KOOP, B., TAGLE, D. AND SLIGHTOM, J. (1987). Globins: A case study in molecular phylogeny. *Cold Spring Harbor Symp. quant. Biol.* **52**, 875–890.

GOODMAN, M., PEDWAYDON, J., CZELUSNIAK, J., SUZUKI, T., GOTOH, T., MOENS, L., SHISHIKURA, F., WALZ, D. AND VINOGRADOV, S. (1988). An evolutionary tree for invertebrate globin sequences. *J. molec. Evol.* **27**, 236–249.

GROSVELD, F., ANTONIOU, M., BERRY, M., DE BOER, E., DILLON, N., ELLIS, J., FRASER, P., HANSCOMBE, O., HURST, J., IMAM, A., LINDENBAUM, M., PHILIPSEN, S., PRUZINA, S., STROUBOULIS, J., RAGUZ-BOLOGNESI, S. AND TALBOT, D. (1993). The regulation of human globin gene switching. *Phil. Trans. R. Soc. Lond.* **339**, 183–191.

GROUDINE, J., KOHWI-SHIGEMATSU, T., GELINAS, R., STAMATOYANNOPOYLOS, G. AND PAPYANNOPOULOU, T. (1983). Human fetal to adult hemoglobin switching: Changes in chromatin structure of the β-globin gene locus. *Proc. natn. Acad. Sci. U.S.A.* **80**, 7551–7555.

GUMUCIO, D. L., SHELTON, D. A., BLANCHARD-MCQUATE, K., GRAY, T. A., TARLE, S. A., HEILSTEDT-WILLIAMSON, H., SLIGHTOM, J., COLLINS, F. S. AND GOODMAN, M. (1994). Differential phylogenetic footprinting as a means to identify base changes responsible for recruitment of the anthropoid γ gene to a fetal expression pattern. *J. biol. Chem.* **269**, 15371–15380.

HARDISON, R. C. (1983). The nucleotide sequence of the rabbit embryonic globin gene β4. *J. biol. Chem.* **258**, 8739–8744.

HARDISON, R. C. (1991). Evolution of globin gene families. In *Evolution at the Molecular Level* (ed. R. K. Selander, T. S. Whittam and A. G. Clark), pp. 272–289. Sunderland, MA: Sinauer Associates, Inc.

HARDISON, R., CHAO, K.-M., SCHWARTZ, S., STOJANOVIC, N., GANETSKY, M. AND MILLER, W. (1994). Globin gene server: A prototype E-mail database server featuring extensive multiple alignments and data compilation. *Genomics* **21**, 344–353.

HARDISON, R. AND MILLER, W. (1993). Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. *Molec. Biol. Evol.* **10**, 73–102.

HARDISON, R., OELTJEN, J. AND MILLER, W. (1997*a*). Long human–mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7**, 959–966.

HARDISON, R., SLIGHTOM, J. L., GUMUCIO, D. L., GOODMAN, M., STOJANOVIC, N. AND MILLER, W. (1997*b*). Locus control regions of mammalian β-globin gene clusters: Combining phylogenetic analyses and experimental results to gain functional insights. *Gene* **205**, 73–94.

HARDISON, R., XU, J., JACKSON, J., MANSBERGER, J., SELIFONOVA, O., GROTCH, B., BIESECKER, J., PETRYKOWSKA, H. AND MILLER, W. (1993). Comparative analysis of the locus control region of the rabbit β-like globin gene cluster: HS3 increases transient

expression of an embryonic ε-globin gene. *Nucleic Acids Res.* **21**, 1265–1272.

HIDALGO, E., DING, H. AND DEMPLE, B. (1997). Redox signal transduction *via* iron–sulfur clusters in the SoxR transcription activator. *Trends biochem. Sci.* **22**, 207–210.

HIGGS, D., WOOD, W., JARMAN, A., SHARPE, J., LIDA, J., PRETORIOUS, I. M. AND AYYUB, H. (1990). A major positive regulatory region located far upstream of the human α-globin gene locus. *Genes Dev.* **4**, 1588–1601.

HOOFT VAN HUIJSDUIJNEN, R., LI, X. Y., BLACK, D., MATTHES, H., BENOIST, C. AND MATHIS, D. (1990). Co-evolution from yeast to mouse: cDNA cloning of the two NF-Y (CP-1/CBF) subunits. *EMBO J.* **9**, 3119–3127.

HOSBACH, H. A., WYLER, T. AND WEBER, R. (1983). The *Xenopus laevis* globin gene family: Chromosomal arrangement and gene structure. *Cell* **32**, 45–53.

HUANG, L. E., ARANY, Z., LIVINGSTON, D. M. AND BUNN, H. F. (1996). Activation of hypoxia-inducible transcription factor depends primarily upon redox-sensitive stabilization of its alpha subunit. *J. biol. Chem.* **271**, 32253–32259.

HUANG, L. E., HO, V., ARANY, Z., KRAINC, D., GALSON, D., TENDLER, D., LIVINGSTON, D. M. AND BUNN, H. F. (1997). Erythropoietin gene regulation depends on heme-dependent oxygen sensing and assembly of interacting transcription factors. *Kidney Int.* **51**, 548–552.

HUANG, X., HARDISON, R. AND MILLER, W. (1990). A space-efficient algorithm for local similarities. *Computer appl. Biosci.* **6**, 373–381.

JANE, S., NIENHUIS, A. AND CUNNINGHAM, J. (1995). Hemoglobin switching in man and chicken is mediated by a heteromeric complex between the ubiquitous transcription factor CP2 and a developmentally specific protein. *EMBO J.* **14**, 97–105.

JANE, S. M., NEY, P. A., VANIN, E. F., GUMUCIO, D. L. AND NIENHUIS, A. W. (1992). Identification of a stage selector element in the human γ-globin gene promoter that fosters preferential interaction with the 5′ HS2 enhancer when in competition with the β-promoter. *EMBO J.* **11**, 2961–2969.

JARMAN, A., WOOD, W., SHARPE, J., GOURDON, G., AYYUB, H. AND HIGGS, D. (1991). Characterization of the major regulatory element upstream of the human α-globin gene cluster. *Molec. cell. Biol.* **11**, 4679–4689.

JENSEN, E. O., PALUDAN, K., HYLDIG-NIELSEN, J. J., JORGENSEN, P. AND MARCKER, K. A. (1981). The structure of a chromosomal leghaemoglobin gene from soybean. *Nature* **291**, 677–679.

JHIANG, S. M., GAREY, J. R. AND RIGGS, A. F. (1988). Exon–intron organization in genes of earthworm and vertebrate globins. *Science* **240**, 334–336.

JOSHI, M. AND DIKSHIT, K. L. (1994). Oxygen dependent regulation of *Vitreoscilla* globin gene: evidence for positive regulation by FNR. *Biochem. biophys. Res. Commun.* **202**, 535–542.

KEILIN, D. (1966). *The History of Cell Respiration and Cytochrome*. Cambridge, UK: Cambridge University Press.

KIM, C., BARNHART, K. AND SHEFFERY, M. (1988). Purification of multiple erythroid cell proteins that bind the promoter of the α-globin gene. *Molec. cell. Biol.* **8**, 4270–4281.

LACELLE, M., KUMANO, M., KURITA, K., YAMANE, K., ZUBER, P. AND NAKANO, M. M. (1996). Oxygen-controlled regulation of the flavohemoglobin gene in *Bacillus subtilis*. *J. Bacteriol.* **178**, 3803–3808.

LACY, E. AND MANIATIS, T. (1980). Nucleotide sequence of the rabbit pseudogene ψβ1. *Cell* **21**, 545–553.

LAM, L. AND BRESNICK, E. H. (1996). A novel DNA binding protein, HS2NF5, interacts with a functionally important sequence of the human β-globin locus control region. *J. biol. Chem.* **271**, 32421–32429.

LI, Q., ZHOU, B., POWERS, P., ENVER, T. AND STAMATOYANNOPOULOS, G. (1990). β-Globin locus activation regions: Conservation of organization, structure and function. *Proc. natn. Acad. Sci. U.S.A.* **87**, 8207–8211.

LIM, L. C., SWENDEMAN, S. L. AND SHEFFERY, M. (1992). Molecular cloning of the alpha-globin transcription factor CP2. *Molec. cell. Biol.* **12**, 828–835.

MACLEOD, K. AND PLUMB, M. (1991). Derepression of mouse β-major-globin gene transcription during erythroid differentiation. *Molec. cell. Biol.* **11**, 4324–4332.

MARTIN, D. I. K., FIERING, S. AND GROUDINE, M. (1996). Regulation of β-globin gene expression: Straightening out the locus. *Curr. Opinions Genetics Dev.* **6**, 488–495.

MASON-GARCIA, M. AND BECKMAN, B. A. (1991). Signal transduction in erythropoiesis. *FASEB J.* **5**, 2958–2964.

MATHEWS, F. S., BETHGE, P. H. AND CZERWINSKI, E. W. (1979). The structure of cytochrome b562 from *Escherichia coli* at 2. A resolution. *J. biol. Chem.* **254**, 1699–1706.

MIGLIACCIO, A. R., VANNUCCHI, A. M. AND MIGLIACCIO, G. (1996). Molecular control of erythroid differentiation. *Int. J. Hematol.* **64**, 1–29.

NEY, P., SORRENTINO, B., MCDONAGH, K. AND NIENHUIS, A. (1990). Tandem AP-1-binding sites within the human β-globin dominant control region function as an inducible enhancer in erythroid cells. *Genes Dev.* **4**, 993–1006.

PERKINS, A. C., SHARPE, A. H. AND ORKIN, S. H. (1995). Lethal β-thalassaemia in mice lacking the erythroid CACCC-transcription factor EKLF. *Nature* **375**, 318–322.

PONDEL, M., MURPHY, S., PEARSON, L., CRADDOCK, C. AND PROUDFOOT, N. (1995). Sp1 functions in a chromatin-dependent manner to augment human alpha-globin promoter activity. *Proc. natn. Acad. Sci. U.S.A.* **92**, 7237–7241.

POTTS, M., ANGELONI, S. V., EBEL, R. E. AND BASSAM, D. (1992). Myoglobin in a cyanobacterium. *Science* **256**, 1690–1691.

PROUDFOOT, N. J., GIL, A. AND MANIATIS, T. (1982). The structure of the human zeta-globin gene and a closely linked, nearly identical pseudogene. *Cell* **31**, 553–563.

RAMLOV, K. B., LAURSEN, N. B., STOUGAARD, J. AND MARCKER, K. A. (1993). Site-directed mutagenesis of the organ-specific element in the soybean leghemoglobin *lbc3* gene promoter. *Plant J.* **4**, 577–580.

REITMAN, M., GRASSO, J. A., BLUMENTAHL, R. AND LEWIT, P. (1993). Primary sequence, evolution and repetitive elements of the *G. gallus* (chicken) β-globin cluster. *Genomics* **18**, 616–626.

RIGGS, A. F. (1991). Aspects of the origin and evolution of non-vertebrate hemoglobins. *Am. Zool.* **31**, 535–545.

ROMBEL, I., HU, K.-Y., ZHANG, Q., PAPYANNOPOULOU, T., STAMATOYANNOPOULOS, G. AND SHEN, C.-K. J. (1995). Transcriptional activation of human α-globin gene by hypersensitive site-40 enhancers: function of nuclear factor-binding motifs occupied in erythroid cells. *Proc. natn. Acad. Sci. U.S.A.* **92**, 6454–6458.

ROUAULT, T. A. AND KLAUSNER, R. D. (1996). Iron–sulfur clusters as biosensors of oxidants and iron. *Trends biochem. Sci.* **21**, 174–177.

SCHIRMER, T., BODE, W., HUBER, R., SIDLER, W. AND ZUBER, H. (1985). X-ray crystallographic structure of the light harvesting biliprotein C-phycocyanin from the thermophilic cyanobacterium

Mastigocladus laminosus and its resemblance to globin structures. *J. molec. Biol.* **184**, 257–277.

SHERMAN, D. R., KLOEK, A. P., KRISHNAN, B. R., GUINN, B. AND GOLDBERG, D. E. (1992). *Ascaris* hemoglobin gene: Plant-like structure reflects the ancestral globin gene. *Proc. natn. Acad. Sci. U.S.A.* **89**, 11696–11700.

SHEWCHUK, B. M. AND HARDISON, R. C. (1997). CpG islands from the α-globin gene cluster increase gene expression in an integration-dependent manner. *Molec. cell. Biol.* (in press).

SLIGHTOM, J., BOCK, J., TAGLE, D., GUMUCIO, D., GOODMAN, M., STOJANOVIC, N., JACKSON, J., MILLER, W. AND HARDISON, R. (1997). The complete sequences of the galago and rabbit β-globin locus control regions: Extended sequence and functional conservation outside the cores of DNase hypersensitive sites. *Genomics* **39**, 90–94.

STAMATOYANNOPOULOS, G. AND NIENHUIS, A. W. (1994). Hemoglobin switching. *The Molecular Basis of Blood Diseases* (ed. G. Stamatoyannopoulos, A. W. Nienhuis, P. W. Majerus and H. Varmus), pp. 107–155. Philadelphia: W. B. Saunders Co.

STOLTZFUS, A., SPENCER, D. F., ZUKER, M., LOGSDON, J. M. AND DOOLITTLE, W. F. (1994). Testing the exon theory of genes: The evidence from protein structure. *Science* **265**, 202–207.

STORZ, G., TARTAGLIA, L. A. AND AMES, B. N. (1990). Transcriptional regulator of oxidative stress-inducible genes: direct activation by oxidation. *Science* **248**, 189–194.

STUVE, L. L. AND MYERS, R. M. (1990). A directly repeated sequence in the β-globin promoter regulates transcription in murine erythroleukemia cells. *Molec. cell. Biol.* **10**, 972–981.

SUN, G., SHARKOVA, E., CHESNUT, R., BIRKEY, S., DUGGAN, M. F., SOROKIN, A., PUJIC, P., EHRLICH, S. D. AND HULETT, F. M. (1996). Regulators of aerobic and anaerobic respiration in *Bacillus subtilis*. *J. Bacteriol.* **178**, 1374–1385.

SZCZYGLOWSKI, K., SZABADOS, L., FUJIMOTO, S. Y., SILVER, D. AND DE BRUIJN, F. J. (1994). Site-specific mutagenesis of the nodule-infected cell expression (NICE) element and the AT-rich element ATRE-BS2× of the *Sesbania rostrata* leghemoglobin *glb3* promoter. *Plant Cell* **6**, 317–332.

TAKAGI, T., IWAASA, H., YUASA, H., SHIKAMA, K., TAKEMASA, T. AND WATANABE, Y. (1993). Primary structure of *Tetrahymena* hemoglobins. *Biochim. biophys. Acta* **1173**, 75–78.

TALBOT, D., PHILIPSEN, S., FRASER, P. AND GROSVELD, F. (1990). Detailed analysis of the site 3 region of the human β-globin dominant control region. *EMBO J.* **9**, 2169–2178.

TARRICONE, C., GALIZZI, A., CODA, A., ASCENZI, P. AND BOLOGNESI, M. (1997). Unusual structure of the oxygen-binding site in the dimeric bacterial hemoglobin from *Vitreoscilla* sp. *Structure* **5**, 497–507.

TAYLOR, E. R., NIE, X. Z., MACGREGOR, A. W. AND HILL, R. D. (1994). A cereal haemoglobin gene is expressed in seed and root tissues under anaerobic conditions. *Plant molec. Biol.* **24**, 853–862.

TRUDEL, M. AND COSTANTINI, F. (1987). A 3′ enhancer contributes to the stage-specific expression of the human β-globin gene. *Genes Dev.* **1**, 954–961.

TSAI, S. F., MARTIN, D. I., ZON, L. I., D'ANDREA, A. D., WONG, G. G. AND ORKIN, S. H. (1989). Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. *Nature* **339**, 446–451.

TUAN, D., KONG, S. AND HU, K. (1992). Transcription of the hypersensitive site HS2 enhancer in erythroid cells. *Proc. natn. Acad. Sci. U.S.A.* **89**, 11219–11223.

VAINSHTEIN, B. K., HARUTYUNYAN, E. H., KURANOVA, I. P., BORISOV, V. V., SOSFENOV, N. I., PAVLOVSKY, A. G., GREBENKO, A. I. AND KONAREVA, N. V. (1975). Structure of leghaemoglobin from lupin root nodules at 5 angstrom resolution. *Nature* **254**, 163–164.

VASUDEVAN, S. G., ARMAREGO, W. L., SHAW, D. C., LILLEY, P. E., DIXON, N. E. AND POOLE, R. K. (1991). Isolation and nucleotide sequence of the *hmp* gene that encodes a haemoglobin-like protein in *Escherichia coli* K-12. *Molec. gen. Genet.* **226**, 49–58.

WAKABAYASHI, S., MATSUBARA, H. AND WEBSTER, D. A. (1986). Primary sequence of a dimeric bacterial haemoglobin from *Vitreoscilla*. *Nature* **322**, 481–483.

WANG, G. L., JIANG, B. H., RUE, E. A. AND SEMENZA, G. L. (1995). Hypoxia-inducible factor 1 is a basic helix–loop–helix PAS heterodimer regulated by cellular O_2 tension. *Proc. natn. Acad. Sci. U.S.A.* **92**, 5510–5514.

WANG, G. L. AND SEMENZA, G. L. (1993). General involvement of hypoxia-inducible factor 1 in transcriptional response to hypoxia. *Proc. natn. Acad. Sci. U.S.A.* **90**, 4304–4308.

WIJGERDE, M., GRIBNAU, J., TRIMBORN, T., NUEZ, B., PHILIPSEN, S., GROSVELD, F. AND FRASER, P. (1996). The role of EKLF in human β-globin gene competition. *Genes Dev.* **10**, 2894–2902.

WITTENBERG, B. A. AND WITTENBERG, J. B. (1987). Myoglobin-mediated oxygen delivery to mitochondria of isolated cardiac myocytes. *Proc. natn. Acad. Sci. U.S.A.* **84**, 7503–7507.

WITTHUHN, B. A., QUELLE, F. W., SILVENNOINEN, O., YI, T., TANG, B., MIURA, O. AND IHLE, J. N. (1993). JAK2 associates with the erythropoietin receptor and is tyrosine phosphorylated and activated following stimulation with erythropoietin. *Cell* **74**, 227–236.

YAMAUCHI, K., TADA, H. AND USUKI, I. (1995). Structure and evolution of *Paramecium* hemoglobin genes. *Biochim. biophys. Acta* **1264**, 53–62.

YOST, S. E., SHEWCHUK, B. AND HARDISON, R. (1993). Nuclear protein binding sites in a transcriptional control region of the rabbit α-globin gene. *Molec. cell. Biol.* **13**, 5439–5449.

ZHAO, X. J., RAITT, D., BURKE, P., CLEWELL, A. S., KWAST, K. E. AND POYTON, R. O. (1996). Function and expression of flavohemoglobin in *Saccharomyces cerevisiae*: Evidence for a role in the oxidative stress response. *J. biol. Chem.* **271**, 25131–25138.

ZHU, H. AND RIGGS, A. F. (1992). Yeast flavohemoglobin is an ancient protein related to globins and a reductase family. *Proc. natn. Acad. Sci. U.S.A.* **89**, 5015–5019.