Protein Science (1995), *4*:2179-2190. Cambridge University Press. Printed in the USA.

Alignment of 700 globin sequences: Extent of amino acid substitution and its correlation with variation in volume

OSCAR H. KAPP,[1] LUC MOENS,[2] JAAK VANFLETEREN,[3] CLIVE N.A. TROTMAN,[4] TOMOHIKO SUZUKI,[5] AND SERGE N. VINOGRADOV[6]

[1] Department of Radiology and Enrico Fermi Institute, University of Chicago, Chicago, Illinois 60637
[2] Department of Biochemistry, University of Antwerp, 2610 Wilrijk, Belgium
[3] Laboratory of Morphology and Systematics, University of Ghent, Ghent, Belgium
[4] Department of Biochemistry, University of Otago, Dunedin, New Zealand
[5] Department of Biology, Kochi University, Kochi 780, Japan
[6] Department of Biochemistry, Wayne State University School of Medicine, Detroit, Michigan 48201

(RECEIVED May 22, 1995; ACCEPTED July 13, 1995)

Abstract

Seven-hundred globin sequences, including 146 nonvertebrate sequences, were aligned on the basis of conservation of secondary structure and the avoidance of gap penalties. Of the 182 positions needed to accommodate all the globin sequences, only 84 are common to all, including the absolutely conserved PheCD1 and HisF8. The mean number of amino acid substitutions per position ranges from 8 to 13 for all globins and 5 to 9 for internal positions. Although the total sequence volumes have a variation ~2-3%, the variation in volume per position ranges from ~13% for the internal to ~21% for the surface positions. Plausible correlations exist between amino acid substitution and the variation in volume per position for the 84 common and the internal but not the surface positions. The amino acid substitution matrix derived from the 84 common positions was used to evaluate sequence similarity within the globins and between the globins and phycocyanins C and colicins A, via calculation of pairwise similarity scores. The scores for globin-globin comparisons over the 84 common positions overlap the globin-phycocyanin and globin-colicin scores, with the former being intermediate. For the subset of internal positions, overlap is minimal between the three groups of scores. These results imply a continuum of amino acid sequences able to assume the common three-on-three α-helical structure and suggest that the determinants of the latter include sites other than those inaccessible to solvent.

Keywords: amino acid substitution; globin; nonvertebrate; sequences; vertebrate

The sequences and crystal structures of V Hbs and Mbs obtained in the 1960s showed that the globin interior consisted of some 33 predominantly hydrophobic amino acids occupying an approximately constant volume (Perutz et al., 1965; Lim & Ptitsyn, 1970; Ptitsyn, 1974). The first NV globin structure, that of *Chironomus* Hb (Huber et al., 1971), was found to share with the V globins both the Mb-fold and many of the interior core residues (Huber et al., 1971). The first detailed alignment of V and NV globin structures and sequences (Lesk & Chothia, 1980) established that the Mb-fold is conserved even at levels of sequence identity as low as 16%, an unusual occurrence in view of the substantial divergences in secondary structure evinced by other protein families at comparably low sequence identities (Chothia & Lesk, 1986, 1987; Lesk & Chothia, 1986). The growth in the number of globin sequences and structures brought the realization that alignment of NV globins with the V globins using Mb-fold templates based on data sets consisting mostly of V globins could be uncertain (Bashford et al., 1987) and that, even when the crystal structures are known, alignment of distantly related globins is not straightforward (Pastore et al., 1988). The sequences of many NV globins obtained over the last 10 years have made it clear that the extent of variation in NV globin sequences far exceeds that observed in the numerically preponderant V globins (Vinogradov et al., 1993) and has led to the recognition of the existence of truncated globins (Takagi, 1993), several different kinds of chimeric globins (Riggs & Riggs, 1990; Trotman et al., 1991; Zhu & Riggs, 1992; Cramm

Reprint requests to: Oscar H. Kapp, Enrico Fermi Institute, University of Chicago, 5640 S. Ellis, Chicago, Illinois 60637; e-mail: bud@rainbow.uchicago.edu.

Abbreviations: V, vertebrate; NV, nonvertebrate; Hb, hemoglobin; Mb, myoglobin; σ, standard deviation; 3D, three-dimensional; RMSD, root mean square deviation; AAS, amino acid substitutions; SAS, solvent-accessible surface area.

et al., 1994), and of the possible occurrence of horizontal globin gene transfers (Moens et al., 1995).

We have used the secondary structure assignments from the known globin crystal structures to align 544 V AND 146 NV globin sequences, following the approach of Lesk and Chothia (1980) and of Bashford et al. (1987). Here we report the alignment, some unique features common to V and NV globins, the extent of AAS, a possible correlation between AAS and variation in volume, a globin-based amino acid substitution matrix, and its application to evaluate sequence similarity among the globins and between the globins, phycocyanins C, and colicins A.

Results

Alignment of globin sequences

Figure 1 shows the globin sequences that define the 182 positions necessary for the alignment of all 700 sequences. The N-terminal, pre-A helix region is defined by several NV sequences: we show the sequence of *Paracaudina* chain I (Suzuki, 1989). The AB interhelical region is defined by *Lumbricus* chain *c* (Fushitani et al., 1988) and *Lamellibrachia* chain BIV (Takagi et al., 1993). The CD region is defined by *Artemia* domain T4 (Trotman et al., 1991). The EF region is defined by several *Chironomus* sequences: we show *CTT X*. The FG region is defined by *Liolophura* Mb (Suzuki et al., 1993). The GH region is completely spanned by *Barbatia virescens* chain I (Suzuki et al., 1992), and the C-terminal end is determined by *Vitreoscilla* Hb (Wakabayashi et al., 1986).

The number of positions used in our alignment is greater than the 171 positions used earlier (Lesk & Chothia, 1980; Bashford et al., 1987), due to the 11 insertions found in the new NV sequences: one in the N-terminal, four in the AB, one each in the EF, FG, and GH interhelical regions, and two additional positions in the C-terminal region. The 554 V globins share 111 positions and the addition of 146 NV sequences lowers it to 84, reflecting the presence of truncated globin sequences from the protozoans *Paramecium* and *Tetrahymena* (Iwaasa et al., 1989, 1990; Takagi et al., 1993) and the cyanobacterium *Nostoc* (Potts et al., 1992) and the globins from the algae *Chlamydomonas* (Couture et al., 1994). The 84 common positions consist of Mb-fold positions A10–A15, B5–C7, CD1, CD2, E1–E20, F1–F9, G4–G17, and H6–H19. Our alignment of the sequences within the helical segments is in agreement with the alignments obtained previously (Bashford et al., 1987; Barton, 1990; Johnson et al., 1990a, 1990b, 1993; Sali & Blundell, 1990; Russell & Barton, 1992; Sippl & Weitckus, 1992; Aronson et al., 1994).

Gerstein et al. (1994) have identified 37 interior, solvent-inaccessible residues (SAS ≤15 Å^2) and 26 surface residues (SAS ≥50 Å^2). The surface position, A4, and internal positions A8, CD4, and FG4, are not always occupied in the NV sequences and are not included in the 84 positions common to all globins.

Amino acid substitution in globins

Table 1 provides quantitative estimates of the extent of amino acid substitution observed in the globin family. The mean num-

```
                          1   5   10   15        1   5   10   15 1   5
                          .   .    .    .        .   .    .    . .   .  CCCC
Mb fold^a                 AAAAAAAAAAAAAAAA       BBBBBBBBBBBBBBBBCCCCCCCDDDD
1mbn          **********VLSEGEWQLVLHVWAKVEA******DVAGHGQDILIRLFKSHPETLEKFDRF
2hbg          **********GLSAAQRQVIAATWKDIAGAD****NGAGVGKDCLIKFLSAHPQMAAVFG*F
1lh1          *********GALTESQAALVKSSWGGFNA******NIPKHTHRFFILVLEIAPAAKDLFSFL
1ecd          ***********LSADQISTVQASFDKVKG**********DPVGILYAVFKADPSIMAKFTQF
1mba          **********SLSAAEADLAGKSWAPVFA******NKNANGLDFLVALFEKFPDSANFFADF
1flp          **********SLEAAQKSNVTSSWAKASA******AWGTAGPEFFMALFDAHDDVFAKFSGL
1sdh          **PSVYDAAAQLTADVKKDLRDSWKVIGS******DKKGNGVALMTTLFADNQETIGYFKRL
1ith          **********GLTAAQIKAIQDHWFLNIKG*****CLQAAADSIFFKYLTAYPGDLAFFHKF
Para I        GGTLAIQSHGDLTLAQKKIVRKTWHQL****MR**NKTSFVTDLFIRIFAYDPAAQNKFPQM
Lumb c        *****DEHEHCCSEEDHRIVQKQWDILWRDTESSKIKIGFGRLLLTKLAKDIPDVNDLFK**
Liolo         **********GISADQAKALKDDIAVVAQ******NPNGCGKALFIKMFEMNPGWVEKFPAW
Bavir I       *PSVAAAVSAVTNKDVAQEIKDTWPFLKK******DMVANGISLMIKMFEKTPESLEFFKRM
Vitre         ***********LDQQTINIIKATVPVLK*****E*HGVTITTTFYKNLFAKHPEVRPLFD*M
Param         *********************SLFEQLGGQ****AAVQAVTAQFYANIQA*DATVATFFN**
1phca                         aaaaaaaaaaaaa      aa bbbbbbbbbbbbbcccccc
PCMLa         FLSNTELQAVNGRYQRAAASLEAARALTA******NAQRLIDGAAQAVYQKFPYLIQT****
1col                          -----α4-----            -----α5----
col A         MASLNKITANPAMKINKADRDALVNAWKH******V***DAQDMANKLGNLSKAFKVAD***
                       .         .         .         .         .         .
Position              10        20        30        40        50        60
```

Fig. 1. Alignment of the sequences of sperm whale Mb (*1mbn*), NV globins of known crystal structure, *Glycera* (*2hbg*), *Lupinus* (*1lh1*), *Chironomus* (*1ecd*), *Aplysia* (*1mba*), *Lucina* (*1flp*), *Scapharca* (*1sdh*), *Urechis* (*1ith*), and the NV globins that define the 182 positions used in this work: *Paracaudina* Hb chain I, *Lumbricus* Hb chain *c*, *Liolophura* Mb, *Barbatia virescens* Hb chain I, and *Vitreoscilla* Hb and the truncated Hb of *Paramecium*. The alignment of globins with phycocyanin A (PCMLa) and with colicin A (col A) are from Holm & Sander (1993a) and Orengo & Taylor (1993) and Pastore & Lesk (1990), respectively. The Mb-fold indicates the location in the secondary structure of sperm whale Mb (Lesk & Chothia, 1980); the 37 solvent-inaccessible positions with mean solvent accessible areas <15 Å^2 (Gerstein et al., 1994) are in bold. The 84 positions common to all globin sequences are underlined. Distal and proximal residues are indicated by D and P, respectively. (*Continues on facing page.*)

ber of different amino acid residues per position for all 700 globins ranges from 8 to 13, depending on the method of counting. The most conservative estimates for the 37 internal positions are 5.3 ± 2.6 for the 554 V globins (range 1–10) and 6.8 ± 2.9 for the 146 NV globins (range 1–13). If all the residues are counted at each position, the corresponding estimates are 6.9 ± 3.4 and 9.3 ± 3.7, respectively.

The NV globins exhibit a substantially greater variation than the V globins. Of the 15 internal positions with the smallest number of substitutions other than CD1 and F8, only 6 positions (less than 50%) are shared by the V and NV globins: Mb-fold locations B14, E7, E11, E15, FG4, and G5.

Variation in volume

The variation in the mean total volumes V_t, defined as $100\sigma_t/V_t$ of the 84 common and 37 internal residues is 2–3%, less than half the variation in the mean volumes of the 26 surface residues, in agreement with the earlier findings (Gerstein et al., 1994). The variation in the mean weighted volume per position

```
                   1   5  1   5   10   15   20              1   5   10
              CC CC.   .  .   . D  .    .    .E             .   .  P .FF
Mb fold       DD DDDDDDDDDEEEEEEEEEEEEEEEEEEEEF             FFFFFFFFFFGG
1mbn          KH*LKTEAEMKASEDLKKHGVTVLTALGAILKK*KGHHEAE*****LKPLAQSHATKH
2hbg          SG********ASDPGVAALGAKVLAQIGVAVSHL**GD*EGKMVAQMKAVGVRHKGYG
1lh1          KG**TSEV*PQNNPELQAHAGKVFKLVYEAAIQLEVTG*VVVTDATLKNLGSVHVSKG
1ecd          AGK**DLESIKGTAPFETHANRIVGFFSKIIGEL******PNIEADVNTFVASHKPRG
1mba          KGK**SVADIKASPKLRDVSSRIFTRLNEFVNN***AANAGKMSAMLSQFAKEHVGFG
1flp          FS*GAAKGTVKNTPEMAAQAQSFKGLVSNWVDNL***DNAGALEGQCKTFAANHKARG
1sdh          *GNV*SQG*M*ANDKLRGHSITLMYALQNFIDQL***DNPDDLVCVVEKFAVNHIT**
1ith          SS*V*PLYGLRSNPAYKAQTLTVINYLDKVVDALG*****GNAGALMKAKVPSHDAMG
Para I        AG*M*SASQLRSSRQMQAHAIRVSSIMSEYIEEL****DSDILPELLATLARTHDL*N
Lumb c        **RVD*I*EHAEGPKFSAHALRILNGLDLAINLL*DDP**PALDAALDHLAHQHEVR*
Liolo         KGK*S*LDEIKASDKITNHGGKVINELANWINNI******NSASGILKSQGTAHKGRS
Bavir I       *GDL*SNP**KANRKLRGHGSSLMYGIMAFVEQL***DSPSDLTDVAGKFAFNHVN**
Vitre         ***********GRQESLEQPKALAMTVLAAAQNI**ENLPAILPA*VKKIAVKHCQ**
Param         ************GIDMPNQTNKTAAFLCAALGGP******NAW**TGRNLKEVHAN**
1phca                  eeeeeeeeeeeeeeeeeeeeeeeee     ffffffff  FFFFFFFFF
PCMLa         *SGPNYAADARGKSKCARDIGHYLRIITYSLVAG****GTGPLDEYLIAGLNEINDAF
1col                         --------α6--------           -----α7-----
colA          ***************VVMKVEKVREKSIEGYETG*******NWGPLMLEVESWVLSGI
                     .         .         .         .         .         .
Position             70        80        90        100       110       120

                   1   5   10   15           1   5   10   15   20   25
              FFF F.   .    .    .          .   .    .    .    .    .
Mb fold       GGG*GGGGGGGGGGGGGGGGGGGG      HHHHHHHHHHHHHHHHHHHHHHHHHH
1mbn          **K*IPIKYLEFISEAIIHVLHSR*HPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG**
2hbg          *NKHIKAQYFEPLGASLLSAMEHRIGGK*MNAAAKDAWAAAYADISGALISGLQS*******
1lh1          ****VADAHFPVVKEAILKTIKEVVGAK*WSEELDSAWTIAYDELAIVIKKEMDDAA*****
1ecd          ****VTHDQLNNFRAGFVSYMKAHT***DFAG*AEAAWGATLDTFFGMIFSKM*********
1mba          ****VGSAQFENVRSMFPGFVASVAAPP***AGADAAWTKLFGLIIDALKAAGK********
1flp          ****ISAGQLEAAFKVLSGFMKSYGG*******DEGAWTAVAGALMGEIEPDM*********
1sdh          **K*ISAAEFGKINGPIKKVLASK**N**FGDKYANAWAKLVAVVQAAL*************
1ith          ****ITPKHFGQLLKLVGGVFQEEF*SA**DPTTVAAWGDAAGVLVAAMK************
Para I        **K*VGPAHYDLFAKVLMEALQAELGSD*FNQKTRDSWAKAFSIVQAVLLVKHG********
Lumb c        **EGVQKAHFKKFGEILATGLPQVL*****DDYDALAWKSCLKGILTKISSRLNA*******
Liolo         IGIEYFENVLPVIDATFAQQMGGAYTAA*MKDALKAAWTGVIVPGMKAGY************
Bavir I       **RNISSFEFQWALVPLLEVLRERLGRNRYRQETEEAWTKLVSVIQATLDDRKR********
Vitre         **AGVAAAHYPIVGQELLGAIKEVLGDAA*TDDILDAWGKAYGVIADVFIQVEADLYAQAVE
Param         MG**VSNAQFTTVIGHLRSALTGA*GVAAALVEQTVAVAETVRGDVVTV*************
1phca           F   gggggggggggggg             hhhhhhhhhhhhhhhhh
PCMLa         **E*LSPSWYINALKYIKANHGLS*******GQAANEANTYIDYVINALS************
1col            -------α8---------           --------α9-------
colA          *ASSVALGIFSATLGAYALSLG*******VPAIAVGIAGILLAAVVGALIDDKFADALNNEI
                        .         .         .         .         .         .
Position               130       140       150       160       170       180
```

Fig. 1. *Continued.*

Table 1. *Mean number of amino acid substitutions per position in the helical segments of 554 V globins and 146 NV globins*[a]

Helix	Positions	*n*	V	NV
A	16-27	12	8.0 ± 2.6	10.3 ± 3.4
B	40-51	12	8.5 ± 4.3	10.3 ± 4.2
C	52-58	7	5.9 ± 2.6	10.7 ± 2.7
E	75-94	20	8.0 ± 4.2	10.1 ± 3.5
F	109-118	10	7.5 ± 4.7	10.1 ± 4.1
G	129-142	14	7.1 ± 2.7	10.9 ± 3.6
H	155-169	15	8.9 ± 4.1	11.1 ± 3.5
All positions		—	8.3 ± 3.9[b]	10.3 ± 3.7[b]
		—	10.3 ± 4.0[c,d]	13.0 ± 3.5[c,d]
Internal positions		37	5.3 ± 2.6	6.8 ± 2.9
		37	6.9 ± 3.4[c]	9.3 ± 3.7[c]

[a] The mean (±SD) number of AAS per position (AAS). Only amino acids occurring more than once were counted, in order to avoid errors due to a single aberrant sequence.

[b] Counting positions with occupancy >98%; 140 for V and 97 for NV globins.

[c] Counting all amino acids per position.

[d] Number of positions: 140 for V and 139 for NV globins.

$100\sigma_p/V_p$, was found to vary from 9 to 29%, much higher than the variation in V_t, again in general agreement with the results of Gerstein et al. (1994).

Correlation between AAS and variation in volume per position

Figure 2A and B shows plots of AAS versus $100\sigma_p/V_p$ obtained for the 84 common and 37 internal positions, respectively, for all 700 globins. The resulting plots were fitted to a linear equation and to an asymptotic exponential function of the type,

$$Y = a(1 - \exp[-x/b]),$$

with two fitted constants a and b (Ratkowsky, 1990). The constants obtained from least-squares fits using the least-squares method and the Marquardt algorithm for the nonlinear case (Press et al., 1992) are summarized in Table 2. The Pearson correlation coefficients obtained for the linear model are significant at the 99.9% level (columns 4 and 5). The correlation coefficients for the nonlinear model are generally larger than for the linear fits. Although in this case no statistics are available to estimate the quality of the correlations (Draper & Smith,

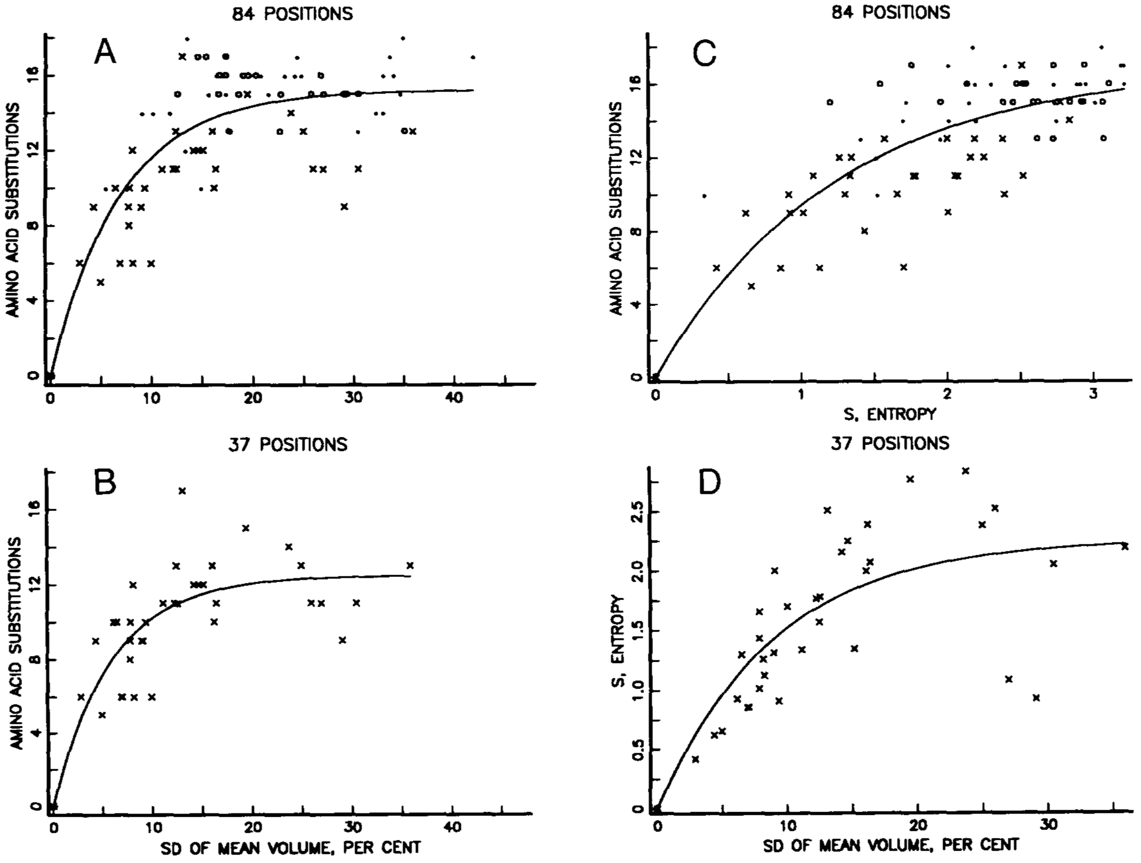

Fig. 2. Plots of AAS per position versus $100\sigma/V_P$ for 700 globin sequences (**A**) for the 84 common positions with internal positions indicated by crosses and surface positions indicated by circles; (**B**) for the 37 internal positions. **C**: Plot of AAS per position versus the information-theoretical entropy S (Shenkin et al., 1991) for 84 positions. **D**: Plot of entropy S versus $100\sigma/V_P$ for the 37 internal positions. The surface position A4 and internal positions A8, CD4, and FG4 are not always occupied in the NV sequences (Table 2) and are not included in the 84 positions common to all globins. The fits shown are to asymptotic exponentials, $Y = a(1 - \exp[-X/b])$ (see Table 2).

Table 2. *Correlations between number of amino acid substitutions, variation in volume, and accessible surface area per position*[a]

	Linear				Nonlinear				
Position	a	b	r	$r_{99.9}$[b]	a	b	r	Z[c]	$t_{99.9}$[d]
A. Number of amino acid substitutions versus variation in mean volume[e]									
All[f]	11.0	0.090	0.22	0.26	13.9	5.48	0.52	14.9	3.35
84	8.60	0.234	0.59	0.34	15.2	6.97	0.79	3.57	3.41
37	6.51	0.250	0.60	0.49	12.5	5.85	0.83	6.83	3.57
26	16.2	0.045	0.25	0.58	15.3	2.46	0.06	–	–
B. Number of amino acid substitutions versus accessible surface area[g]									
All[f]	6.37	0.051	0.34	0.26	9.55	11.5	0.48	6.81	3.35
84	5.10	0.100	0.61	0.34	11.0	15.1	0.69	1.25	3.4
37	3.89	0.120	0.39	0.49	6.51	5.14	0.21	–	–
26	14.3	0.060	0.38	0.58	10.7	0.50	0.00	–	–
C. Variation in mean volume versus accessible surface area[h]									
All[f]	15.0	0.140	0.34	0.26	23.4	11.7	0.36	0.92	3.35
84	13.8	0.164	0.41	0.34	22.7	9.97	0.38	–	–
37	12.8	0.057	0.06	0.49	13.7	1.01	0.04	–	–
26	27.3	0.065	0.16	0.58	23.7	11.3	0.13	–	–

[a] Least-squares fit to linear equations, of the form $Y = a \pm bX$, and to nonlinear equations, $Y = a[1 - \exp(-X/b)]$. r is the Pearson correlation coefficient, $r = \Sigma XY/(\Sigma X^2 \Sigma Y^2)^{1/2}$, where $X = (x_i - x)$ and $Y = (y_i - y)$ (Zar, 1984).

[b] Critical values for the correlation coefficient r at 99.9% probability (Powell, 1982).

[c] $Z = [z_l - z_{nl}]/\sigma_z$, where z_l and z_{nl} are the Fisher transforms of r for the linear and nonlinear models, respectively and σ_z, the SE in z is $(1/[N-3])^{1/2}$, N being the number of positions (Zar, 1984).

[d] Critical value of the t distribution at 99.9% confidence level (Zar, 1984).

[e] Y = AAS at each position and $X = 100\sigma_P/V_P$ where V_P is the weighted mean volume at each position and σ_P is the SD for the complete set of 700 globin sequences.

[f] All positions: 182 in case A and 166 for cases B and C.

[g] Y = AAS at each position for a selected set of 134 monomeric globin sequences and X = mean SAS at each position calculated from 9 monomeric globin crystal structures: 1mbs, 1mbc, 1myt, 1yma, 1mba, 1lh1, 2hbg, 1ecn, and 1flp (Abola et al., 1987).

[h] $Y = 100\sigma/V$ at each position for the set of 134 monomer globin sequences and X = mean ASA at each position from the 9 monomer globin crystal structures.

1981; Sachs, 1984), their Fisher transforms (Zar, 1984) in columns 9 and 10, which strictly speaking are only applicable to the straight line fits, do show a significant improvement over the linear correlations.

One of the reviewers pointed out that the Shannon information-theoretical entropy $S = -\Sigma p_i \log_2 p_i$ ($p_i = n_i/N$, n_i = number of times amino acid i appears at a given position in N sequences) proposed by Shenkin et al. (1991) should be a better measure of sequence variability per position than AAS. In fact, there appears to be a good correlation between S and AAS for the 84 common positions (Fig. 2C, $r = 0.80$). Furthermore, a plot of S versus $100\sigma_p/V_p$ for the 37 internal positions (Fig. 2D) shows a correlation similar to that found for AAS versus $100\sigma_p/V_p$ (Fig. 2B; $r = 0.80$). Similar results are obtained using the data set of immunoglobulin light chain sequences (over 123 positions) in Shenkin et al. (1991): AAS versus S ($r = 0.78$) and the nonlinear correlations between AAS and either S or $100\sigma_p/V_p$ appear to be significant, $r = 0.60$ and 0.69, respectively.

Correlation between AAS and SAS

The AAS from a subset of 134 monomeric globin sequences representing the agnathan globins, V Mbs, and NV globins was correlated with the mean SAS from nine monomer globin crystal structures. The fitted constants for both the linear and nonlinear models and the corresponding correlation coefficients are given in Table 2. The correlation coefficients, although smaller than for AAS versus $100\sigma/V$, appear to be still significant over the 166 and 84 positions.

Finally, the results of fits to linear and nonlinear models of $100\sigma/V$ calculated for the 134 monomer globin sequences versus the corresponding mean SAS of the nine monomer globin crystal structures, over 84 common positions (Table 2C), suggest that there is little, if any, correlation between these two variables.

Globin-based block amino acid substitution matrix

The log-odds amino acid substitution matrix calculated from the observed frequencies over the block of 84 positions common to all 700 globin sequences, as described by Johnson and Overington (1993), was multiplied by 10 and scaled up by 18.65 to eliminate negative numbers. Table 3 shows the resulting matrix, which represents the first scoring matrix based only on one protein family.

Pairwise comparisons of globins, phycocyanins C, and colicins A

Figure 3A illustrates the results obtained from pairwise comparisons of individual V globin groups with other V and NV glo-

Table 3. *Amino acid substitution matrix derived from the 84 positions common to 700 globins*[a]

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	23.4																			
C	21.5	31.9																		
D	18.2	14.0	25.9																	
E	18.1	12.2	22.8	26.5																
F	11.9	13.5	9.04	9.24	27.1															
G	19.9	15.0	18.7	19.3	9.84	26.8														
H	11.5	13.5	17.2	15.7	10.7	12.2	30.8													
I	17.2	20.9	13.7	14.9	18.4	14.7	12.4	24.7												
K	17.1	9.83	18.2	19.2	9.68	15.6	15.5	14.2	26.4											
L	13.6	15.5	12.5	11.4	20.6	11.5	9.07	21.0	12.1	24.8										
M	16.2	19.8	15.5	15.2	20.2	15.0	11.4	20.9	15.4	22.4	25.9									
N	17.1	14.7	21.4	20.0	11.4	18.1	18.9	14.8	17.7	11.7	13.7	27.4								
P	17.7	5.43	19.4	19.1	3.14	17.0	11.7	9.35	16.8	7.28	11.6	14.6	32.2							
Q	17.2	13.8	22.2	20.4	11.5	16.3	21.2	15.8	22.0	13.3	17.8	18.6	18.0	26.5						
R	16.8	14.4	18.2	18.7	10.1	16.1	13.7	14.3	21.9	12.2	15.5	17.1	15.7	19.6	29.5					
S	20.4	17.7	19.1	17.7	9.65	20.2	13.8	14.6	16.6	12.2	14.7	20.9	17.9	17.6	17.5	25.4				
T	18.1	16.2	19.0	18.8	12.8	18.2	14.8	17.5	17.6	14.3	16.9	20.2	14.9	17.4	20.6	19.5	26.1			
V	16.7	21.4	15.1	15.1	17.2	14.5	12.9	22.2	13.5	18.9	19.1	14.8	11.1	15.2	14.4	15.1	17.9	25.0		
W	10.1	9.56	9.08	15.3	19.3	7.20	0.0	10.4	7.79	14.7	18.5	7.74	9.85	16.6	9.94	9.70	17.2	11.6	33.8	
Y	15.3	17.3	14.9	12.7	22.6	13.7	19.4	17.6	14.7	17.9	20.0	20.1	6.99	14.6	13.8	14.4	17.7	16.5	19.5	29.2
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y

[a] The log-odds matrix calculated as described in Johnson and Overington (1993) was multiplied by 10 and then scaled up by 18.65 to make all numbers positive.

bins, four phycocyanins C, and three colicins A (C), as a plot of percent identity versus the mean score over 84 positions, calculated using the Johnson and Overington (1993) matrix. Figure 3B and C shows representative results obtained for pairwise comparisons of NV globins with V globins (A + B + M), the phycocyanins, and the colicins over 84 positions, using the Henikoff matrix (Henikoff & Henikoff, 1992) and our matrix (Table 3), respectively. These plots illustrate the overlaps between the various groups of scores. Another way of presenting the results consists in the normalized similarity scores z (Schwartz & Dayhoff, 1978; Feng et al., 1987), calculated over the 84 positions using the five matrices and shown in Table 4.

Figure 3D, E, and F shows the results obtained for pairwise comparisons of NV globins with V globins, phycocyanins, and colicins using our matrix (Table 3), when the 84 common positions are broken down into 34 internal, 22 surface, and 28 remaining positions.

Discussion

Globin sequence alignment

We have used the available crystal structures of NV one-domain globins, which represent the majority of known one-domain globins (103 of 108 in our data set), to align them reliably with the 554 V globins. We were also able to include the remaining NV globins for which no crystal structures are available. Our alignment has several important advantages: (1) it was performed on a carefully selected group of sequences; (2) it was based on the alignment of clearly identified segments of secondary structure; and (3) the parsimonious allocation of positions in the interhelical regions, based solely on the need to accommodate the longest sequence, avoided the use of gap penalties. It should be noted that our alignment within the interhelical regions is heuristic at best; we have not sought to maximize it due to the notable variation in the exact location of the edges of the helical segments in going from one globin group to another. Many more NV globin crystal structures will be necessary before any reliable alignment of the interhelical regions can be attempted. The 84 common positions, reflecting the sum total of conserved helical segments, are underlined in Figure 1. These common positions are included in the 97-residue core from helical segments common to the structures of several V and NV globin structures (Aronson et al., 1994). Furthermore, Figure 1 clearly shows the impressive conservation of hydrophobic residues (marked in bold type) at the internal positions, which is probably responsible for the maintenance of the Mb-fold. Even *Paramecium* Hb, shown last as a representative truncated globin, aligns well with the other globins. The truncated globins of protozoans and the cyanobacterium *Nostoc* share with *Glycera* monomeric globin, *Vitreoscilla* Hb, and the chimeric globins of *Escherichia coli* and the yeasts *Candida* and *Saccharomyces* the feature of an absent D helix, in agreement with its lack of a well-defined role in V globins (Komiyama et al., 1991).

Conservation of 3D structure and sequence similarity

Chothia and Lesk (1986, 1987) were the first to demonstrate an exponential relationship between decrease in sequence identity and increase in the RMSD of Cα coordinates, which has been confirmed since (Orengo et al., 1992; Flores et al., 1993; Chelvanayagam et al., 1994; Laurents et al., 1994). Furthermore,

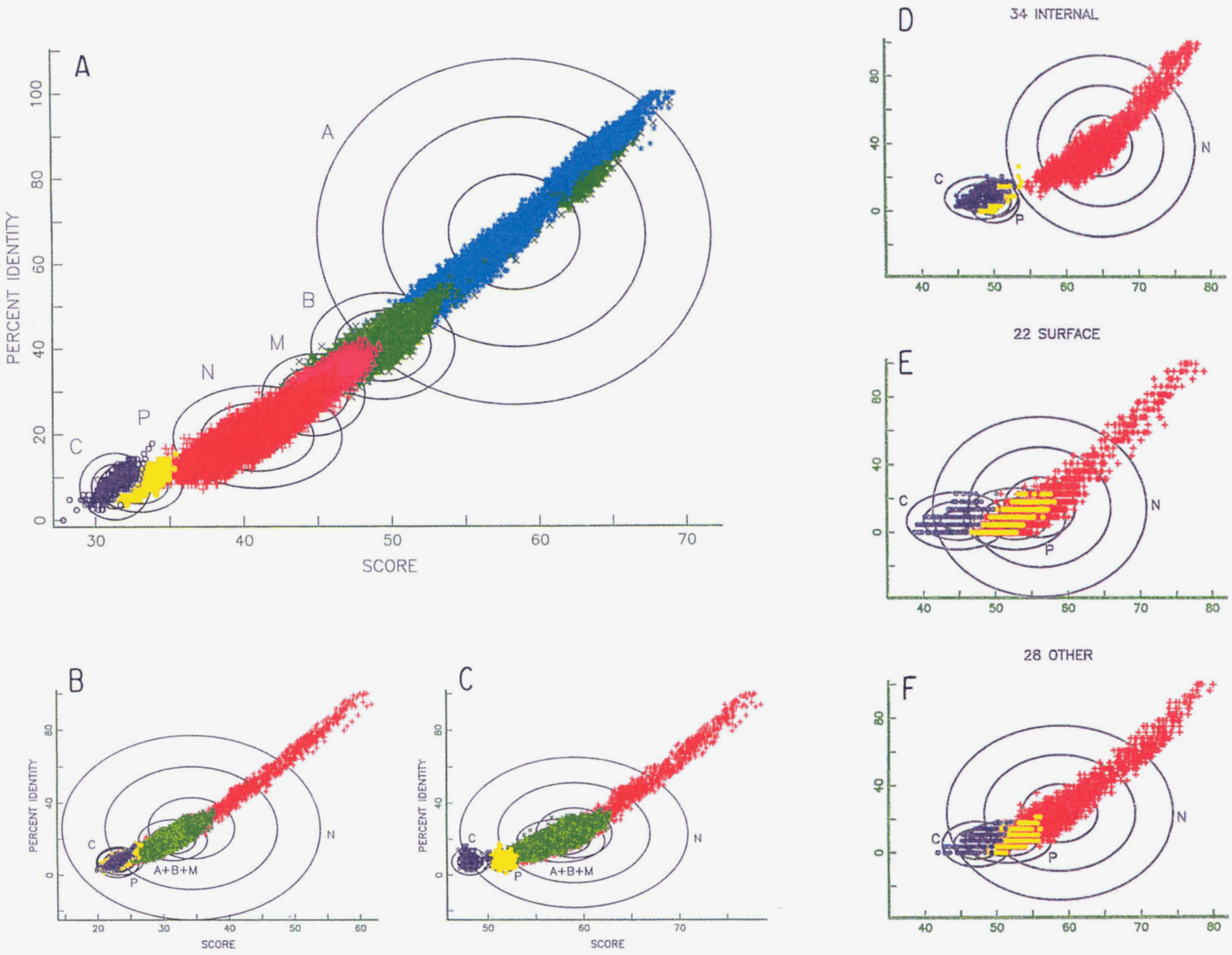

Fig. 3. Plots of percent identity versus similarity scores (with 1 SD, 2 SD, and 3 SD contours) (**A**) calculated over 84 common positions using the Johnson and Overington (1993) matrix for pairwise comparison of α-globins (A) with β-globins (B), Mbs (M), NV globins (N), four phycocyanins C (P), and three colicins (C). Pairwise comparison of NV globins with themselves (N), with V globins (A + B + M), four phycocyanins C (P), and three colicins A, using (**B**) the Henikoff matrix (Henikoff & Henikoff, 1992) and (**C**) our matrix (Table 4) over 84 common positions and (**D, E, F**) over subsets of the 84 common positions.

Table 4. *Normalized similarity scores z calculated from pairwise comparisons over 84 common positions, of V and NV globins, phycocyanins, and colicins using several amino acid substitution matrices*[a]

	α(A)						β(B)					MB(M)				Nonvert(N)			Phyco(P)		Col(C)
Matrix	A	B	M	N	P	C	B	M	N	P	C	M	N	P	C	N	P	C	P	C	C
DAY	16	12	8.6	6.9	2.0	0.6	17	8.5	5.6	1.4	0.7	17	6.8	2.0	0.2	8.1	1.4	0.5	13	0.1	12
DOO	14	11	7.2	5.5	1.7	0.7	15	7.5	5.4	1.1	0.8	15	5.9	1.6	0.3	6.3	1.5	0.6	15	0.4	11
HEN	19	14	9.7	6.9	1.9	1.0	20	9.6	7.0	1.3	1.2	21	7.1	1.7	0.1	8.6	1.7	0.7	18	0.3	17
J&O	14	7.6	5.1	2.1	0.4	0.2	15	4.9	2.6	0.6	0.2	18	3.3	0.2	0.5	4.6	0.4	0.6	17	0.3	9.0
TW	17	13	9.5	6.9	2.0	1.2	18	9.6	6.9	1.5	1.4	18	7.3	2.0	0.5	8.1	1.7	0.7	17	0.1	12

[a] The z score was calculated from the relation $z = (r - m)/\sigma$ (Schwartz & Dayhoff, 1978; Feng et al., 1985), where r is the mean of the scores from pairwise comparisons and m and σ are the mean score and SD, respectively, obtained from pairwise comparison of 32 statistically randomized sequences having the same amino acid compositions as the sequences compared. Amino acid substitution matrices were: DAY, Schwartz and Dayhoff (1978); DOO, Doolittle (Feng et al., 1985); HEN, Henikoff and Henikoff (1992); J&O, Johnson and Overington (1993); TW, this work (Table 3).

there appears to be a hierarchy in the correlations of various structural features with the RMSD of Cα coordinates: although the correlations of residue accessibility and secondary structure content are robust, the correlations with other parameters, such as the backbone conformational angles ν and ω and the side-chain conformational angle ω^1, are weaker and the correlations of finer grain properties, such as main-chain–main-chain and main-chain–side-chain hydrogen bonding and the side-chain conformational angles ω^2–ω^4, are poorer still (Chelvanayagam et al., 1994). The latter authors note that the globin family exhibits some of the highest conservation of main-chain–main chain hydrogen bonding in a data set of 175 tertiary structures in 34 different protein families, in agreement with their higher conservation than other protein families (>75%), of residue–residue contact maps based on 443 structurally aligned 3D structures (Rodionov & Johnson, 1994).

Several pairwise comparisons of the known protein crystal structures have demonstrated that they can be classified into a relatively limited number of 3D structures (Holm et al., 1992; Holm & Sander, 1993a, 1993b; Orengo & Taylor, 1993; Orengo et al., 1993). In particular, it was found that globins, phycocyanins C, and colicins A share a unique folding motif, the three-on-three α-helical sandwich (Holm & Sander, 1993a). The b-chain of phycocyanin C has also been aligned with *Aplysia* Mb (Jones et al., 1992), in accordance with the helix geometry of the latter being closest to phycocyanin (Pastore & Lesk, 1990). Eight of the 10 helical segments in the structure of colicin A (Parker et al., 1992) align well against the helical segments A through H of the Mb-fold (Orengo et al., 1993; Orengo & Taylor, 1993). In addition, Orengo and Taylor (1993) have also detected a similarity between the structures of diphtheria toxin (Choe et al., 1992), phycocyanin, and the globins. Recently, yet another structural similarity has been detected between the globins and the bacteriophage repressor proteins, with the five helices of the DNA-binding N-terminal domains superimposing remarkably well on the longer helices A, B, E, G, and H of the globins (Subbiah et al., 1993).

Amino acid substitutions in globin sequences

Not unexpectedly, the NV globins exhibit a substantially greater AAS than the V globins. Of 15 positions, other than CD1 and F8, with the smallest AAS, the two groups share only 6 positions: B14, E7, E11, E15, FG4, and G5 (with 4, 4, 4, 5, 3, 3 and 5, 5, 6, 7, 6, 5 substitutions in the V and NV groups, respectively). These positions all play a role in maintaining the integrity of the heme-binding cavity, except B14, which is involved in the B/E helix contact (Lesk & Chothia, 1980; Bashford et al., 1987).

Although many site-directed mutagenesis studies of several proteins have demonstrated clearly the ability of protein secondary and 3D structures to accommodate extensive AAS (Hurley, 1994; Johnson et al., 1994), the exact limits of the latter have remained a matter of conjecture. Our results for the globin family (Table 1) provide reliable estimates of the maximum AAS per position, 8–13 overall; furthermore, the extent of substitution remains relatively constant within the helical regions. Conservative estimates of AAS at the interior positions vary between 5 and 7, and the solvent-accessible positions appear to have very little if any restriction in substitution (AAS ~10–15).

In contrast to the extent of substitution at the vast majority of positions in the globin sequences, stands the absolute conservation of residues PheCD1 and HisF8. Although this feature of the globin family was recognized early on in the case of V globins (Perutz et al., 1965; Ptitsyn, 1974), its presence in all the NV globin sequences determined so far suggests that it is characteristic of globins.

The distal residue at position 81(E7) is overwhelmingly His in the V sequences. Gln is found in elephant Mb, the α-chain of opossum, and several amphibian, reptilian, and agnathan Hbs (Kleinschmidt & Sgouros, 1987). Tyr is found in only one sequence derived from the ϵ^2 β-globin gene of cow (Schimenti & Duncan, 1985). Gln is much more prevalent in the NV globins, together with hydrophobic residues such as Leu in the one-domain Hbs of *Glycera* (Arents & Love, 1990) and the nematodes *Trichostrongylus* (Frenkel et al., 1992) and *Nippostrongylus* (Blaxter, 1993), Val in *Aplysia* Mb (Bolognesi et al., 1989), and Ile in the monomeric Hb from the insect *Tokunagayusurika* (Fukuda et al., 1993).

Variation in total volumes and mean volumes per position

Gerstein et al. (1994) have compared the variations in the calculated volumes of the sequences and of internal and surface residues in three different families of proteins: globins (568 sequences), plastocyanin (40 sequences), and dihydrofolate reductase (24 sequences). They found that the variation in the total volume of the residues occupying the internal positions is small, $\sigma/V \sim 2.5\%$, compared to the variation at individual sites, ~13% for the internal and ~24% for the surface sites, the latter being close to the variation found for random sequences. Our results are in complete agreement with theirs. Furthermore, Gerstein et al. (1994) suggest that the apparent constancy of the core volume need not be due to the imposition of any constraints on the evolution of protein sequences from a common precursor and that it is consistent simply with appropriately chosen random sequence changes, in agreement with an earlier conclusion (Ptitsyn & Volkenshtein, 1986). We find a statistically significant (at the 99.9% level) linear correlation between sequence variability, defined as AAS per position or the more sophisticated information-theoretical entropy S (Shenkin et al., 1991) and the variation in volume per position over 182, 84, and 37 positions (Table 2). The use of an asymptotic exponential rather than a linear model is intuitively much more appealing, due to the fulfillment of the obvious requirement that $100\sigma/V = 0$ when either AAS = 0 or $S = 0$ and the expectation that sequence variability should approach an asymptote given a finite number of different amino acids. Although statistical tests appropriate for a linear correlation cannot be used for a nonlinear one (Draper & Smith, 1981; Sachs, 1984), the fact that the r values obtained with the nonlinear fit are appreciably higher than for the linear fit (Table 2) allows us to conclude that a statistically significant correlation does exist between the two variables.

Of the 15 positions with the lowest $100\sigma/V100\sigma/V$ (<15%) and AAS (<10) per position, E7 is the position with the smallest variation in volume. Although there are at least four known substitutions of the distal His (Val, Gln, Leu, and Ile), volumes of the amino acid residues in question range from 139 to 164 Å^3, suggesting that the volume of the distal cavity may be a determinant of the nature of the distal residue side-chain group.

Pairwise comparison of globin, phycocyanin, and colicin sequences

We have carried out pairwise comparisons within and between the globin groups, between the globins on one hand and the phycocyanins and colicins on the other, and between the latter two groups using our amino acid substitution matrix (Table 3) and several other commonly used matrices over all positions, the 84 common positions, and subsets of the latter. The results shown in Table 4 and illustrated in Figure 3 can be summarized as follows. (1) The scores for V to V, V to NV, and NV to NV globin comparisons overlap completely with each other as shown by the respective range of z values (4.6–8.1 versus 2.1–7.3) in Table 4 and illustrated in Figure 3A. (2) The scores of comparisons of NV to V globins, NV to NV globins, NV globins to phycocyanins C, and NV globins to colicins A form a continuum of overlapping scores as illustrated in Figure 3B and C. However, both V and NV globins are more similar to the phycocyanins than the colicins, with z values two to three times higher and more overlap between the NV to NV globin scores and the NV to phycocyanin scores than with the NV globin–colicin scores (Table 4). (3) There is considerable overlap between the NV globin–phycocyanin and NV globin–colicin scores and between them and the randomly generated sequences (Table 4). In general, there appears to be essentially a continuum of scores ranging from scores representative of close to 100% identity to scores corresponding to ca. 7% identity or less found with randomly generated sequences.

In an effort to define the determinants of the three-on-three α-helical sandwich motif shared by the 3D structures of globins, phycocyanins, and colicins (Pastore & Lesk, 1990; Holm & Sander, 1993a, 1993b), we used our AAS matrix to compare the three groups based on scores calculated over subsets of the 84 common positions: the 34 internal, 22 surface, and 28 remaining positions (Fig. 3D,E,F). The NV globin–colicin and NV globin–phycocyanin scores overlap each other and are clearly separated from the globin–globin scores, when the calculation is performed over the 34 internal positions (Fig. 3D) but not for scores calculated over the other two subsets (Fig. 3E,F). This unexpected result implies that the common secondary structure is not solely determined by residues occupying internal positions.

What is a globin?

Russell and Barton (1994) have suggested a classification of protein pairs into three types based on 607 pairs of aligned 3D structures with variation in sequence identity from 1.1 to 86.2%. Type A proteins, having ≥20% sequence identity and sharing structural and functional similarity; type B, which show structural and functional similarity but have sequence identities <20%; and type C proteins, with only 3D structural similarity. The globin family encompasses type A and type B similarities and the globin superfamily would also include phycocyanins, colicins, diphtheria toxin, etc., representative of type C similarity.

The results presented here demonstrate that pairwise comparison using a globin-based AAS matrix is perhaps too crude a method to identify the determinants of the three-on-three α-helical sandwich structure common to globins, phycocyanins, and colicins. In particular, we would reasonably expect the conservation of amino acid residues at the internal positions to determine the conservation of that common structure. This expectation is not supported by the finding illustrated in Figure 3D and E, that the scores calculated for the subset of 34 internal positions differentiate between globins, phycocyanins, and colicins. Because the only distinguishing characteristic of the globin sequences is the presence of the two invariant residues, the Phe CD1 and the proximal His F8, it appears interesting to find out whether the appropriate conversions in phycocyanin C or colicin A would suffice to transform them into functional globins.

Although the 28 chimeric globins in our data set, which contain a heme-binding N-terminal domain or consist of two or more domains, appear to be bona fide members of the globin family, there exist several other chimeric proteins with internal heme-binding sequences and unclear relationship to globins, such as *Rhizobium* FixL hemeprotein (Gilles-Gonzalez et al., 1994, 1995), the Mb of the molluscs *Sulculus* and *Nordotis* (Suzuki & Takagi, 1992; Suzuki, 1994), and the glutamate racemase from the bacterium *Pediococcus pentosaceus* (Choi et al., 1994). Although the latter contains an internal sequence that exhibits a 21–27% identity with Mbs, the others cannot be readily aligned with globins without greatly increasing the number of positions. FixL appears to represent a new class of heme proteins with distinct ligand binding properties, whereas the abalone Mbs, because of a ~35% identity with vertebrate 2,3-indoleamine oxygenases, probably represent a case of convergent evolution. We explore elsewhere (Moens et al., 1995) the structural and functional relationships between the true globins and borderline globins.

Materials and methods

Globin sequences

Globin sequences were obtained from the Atlas of Protein and Genomic Sequences (National Biomedical Research Foundation), March 31, 1993 release of PIR1, PIR2, PIR3, and PATCHX databases (Barker et al., 1993). The 1,058 entries were evaluated using the following criteria. (1) All partial sequences, including those derived from pseudogenes, were rejected. In addition, globin sequences labeled tentative or which contained peptide(s) whose sequence(s) were not determined experimentally were also eliminated. (2) In cases where more than one version of the same sequence were found, either the most recent version or the genomic/cDNA sequence was preferred. (3) Sequences identical to the ones already determined, generally from closely related species, were not included.

The 554 selected unique V sequences included 56 Mbs, 247 α-like and 239 β-like chains, and 12 globins that were not identified as either α or β. A total of 146 NV globin sequences were used, comprising 113 from the PIR database, the remainder consisting of recently determined sequences. A list of sequences is available upon request.

Alignment of globin sequences

Pairwise alignment of globin sequences using the Needleman–Wunsch algorithm does not necessarily lead to the correct alignment of conserved secondary structures in globins. Lesk et al. (1986) have shown that the use of variable gap penalties in the alignment of human α and *Lupinus* Hb reduces but does not eliminate errors relative to the alignment based on 3D-structure.

It is evident that the alignments of Lesk and Chothia (1980) and of Bashford et al. (1987), based on the conservation of the secondary structure, provide the most effective approach toward the alignment of globin sequences. We employed the secondary structures identified in the following crystal structures (Abola et al., 1987): sperm whale Mb (1mbd), human α (4hhba), human β (4hhbb), human γ (1fdb), lamprey (2lhb), *Chironomus* (1eca), lupin (2lh4), *Glycera* (2hbg), *Aplysia* (1mba), *Scapharca* (1sdh), *Urechis* (1ith), and *Lucina* (1flp). Although the principal elements of the secondary structure, the Mb-fold, are conserved (Lesk & Chothia, 1980), the boundaries of the homologous α-helical segments can vary appreciably. The following guidelines were used in the alignment. (1) Globins belonging to a group in which one member has a known crystal structure, were assumed to have an identical secondary structure. (2) The number of positions assigned to interhelical regions were chosen parsimoniously, i.e., just enough to accommodate the longest known sequence(s): e.g., the AB interhelical region is defined by the sequences of *Lumbricus* chain *c* (Fushitani et al., 1988) and *Lamellibrachia* chain BIV (Takagi et al., 1993). (3) The homologous helical segments were aligned with each other and the Mb sequence. Because the D helix is present only in three of the seven NV globin structures, and because its presence or absence in V globins appears to be of no consequence (Komiyama et al., 1991), we considered it to be part of the CE interhelical region. (4) In the case of NV globin groups for which crystal structures are not available, we used the V Mb structure. (5) The interhelical sequences were aligned heuristically, following the preferred AAS for surface residues (Bordo & Argos, 1991).

Calculation of AAS, variation in volume, and SAS

The number of different amino acids per position was tabulated and the means calculated for the positions within each helical segment, all positions with occupancy >98%, and for internal and surface positions.

The Voronoi volumes of Gerstein et al. (1994) were used to calculate (1) the total volumes V_t of the individual 700 globin sequences, of 84 common, 37 internal, and 26 surface positions and (2) the mean weighted volume per position $V_p = (1/N)\Sigma n_i V_i$. The variations in V_t and V_p were defined as $100\sigma_t/V_t$ and $100\sigma_p/V_p$, respectively.

Calculation of the SASs were performed using the Connolly (1983) algorithm using the Biosym Inc. (San Diego, California) software. Because most distances between water O and protein O and N atoms are between 2.9 and 3.0 Å (Jeffrey & Saenger, 1991), we chose a probe radius of 1.5 Å. The coordinates of the surface-bound water molecules and bound inorganic ions were not included. The following nine high-resolution globin crystal structures from the Protein Data Bank were used to obtain mean SASs: 1mbs, 1mbc, 1myt, 1yma, 1mba, 1lh1, 1hbg, 1ecn, and 1flp (Abola et al., 1987). The mean SAS per position was calculated as $(1/N)\Sigma n_i(\text{ASA})_i$.

Construction of a globin-based amino acid substitution matrix

The derivation of a scoring matrix based on 84 positions common to all 700 globin sequences was carried out as described by Johnson and Overington (1993). The observed raw frequencies $G_{i,j}$ were converted to probabilities of replacement of residue i by residue j,

$$P_{i\to j} = G_{i,j} \Big/ \sum_{j=1}^{j=20} G_{i,j}.$$

The probability matrix was converted to the odds matrix $H_{i,j}$,

$$\Theta_{i,j} = P_{i\to j} \sum_{i=1}^{i=20} G_{i,j} \Big/ \sum_{i=1}^{i=20} \sum_{j=20}^{j=20} G_{i,j},$$

which in turn provided the log-odds scoring matrix L,

$$L_{i,j} = \log_{10} \Theta_{i,j}.$$

Substitutions involving cysteine and cystine were combined to give 20 amino acid types.

Evaluation of amino acid sequence similarity

Quantitative evaluation of sequence similarity was carried out via pairwise comparison and scoring using several different AAS matrices. In addition to the globin-based matrix, we used the PAM 250 matrix of Schwartz and Dayhoff (1978) and the matrices of Doolittle (Doolittle, 1979; Feng et al., 1985), Hennikoff (Hennikoff & Hennikoff, 1992, 1993), and Johnson and Overington (1993). In order to be able to compare the resulting scores, we scaled all the matrices, making the lowest AAS score zero and the highest 100. The scores obtained at each position in a pairwise comparison were added and divided by the number of positions compared to give a similarity score r. The pairwise comparisons were carried out for the V and NV globin sequences over the 84 common positions and the 37 internal positions. In addition, pairwise comparisons were carried between the globins and four phycocyanins C, the α and β chains of *Mastigocladus laminosus* and *Agmenellum quadruplicatum* and three colicin sequences (IKEBCA, IKECBB, S00867), using the alignments of Pastore and Lesk (1990) and Holm and Sander (1993a, 1993b), respectively. For each pairwise comparison, 32 random sequences were generated for each of the two globin sequences and also compared and scored. A normalized similarity score z (Dayhoff et al., 1983; Feng et al., 1985) was calculated from,

$$z = (r - m)/\sigma$$

where m is the mean of the r scores for a group of pairwise comparisons and σ is the standard deviation of the mean of the scores for the pairwise comparisons of the randomized sequences.

Acknowledgments

Supported in part by a grant from the Department of Energy to O.H.K. (DE-FG02-86-ER60437), a grant from the Belgian Fund for Joint Basic Research to L.M. (2.0023.94), a grant from the New Zealand Lottery Grants Board to C.N.A.T., and a grant from the National Institutes of Health to S.N.V. (DK38674).

References

Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein Data Bank. In: Allen FH, Bergerhoff G, Sievers R, eds. *Crystallographic databases—Information content, software systems, scientific, applications.* Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography. pp 107-132.

Arents G, Love WE. 1990. *Glycera dibranchiata* hemoglobin: Structure and refinement at 1.5 Å. *J Mol Biol 210*:149–161.

Aronson HEG, Royer WE Jr, Hendrickson WA. 1994. Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. *Protein Sci 3*:1706–1711.

Barker WC, George DG, Mewes HW, Pfeiffer F, Tsugita A. 1993. The PIR international databases. *Nucleic Acid Res 21*:2967–3092.

Barton GJ. 1990. Protein multiple alignment and flexible pattern matching. *Methods Enzymol 183*:403–425.

Bashford D, Chothia C, Lesk AA. 1987. Determinants of a protein fold: Unique features of the globin amino acid sequences. *J Mol Biol 196*: 199–216.

Blaxter ML. 1993. Nemoglobins: Divergent nematode globins. *Parasitol Today 9*:353–360.

Bolognesi M, Onesti S, Gatti G, Coda A, Ascenzi P, Brunori M. 1989. *Aplysia limacina* myoglobin: Crystallographic analysis at 1.6 Å resolution. *J Mol Biol 205*:529–544.

Bordo D, Argos P. 1991. Suggestions for "safe" residue substitutions in site-directed mutagenesis. *J Mol Biol 217*:721–729.

Chelvanayagam G, Roy G, Argos P. 1994. Easy adaptation of protein structure to sequence. *Protein Eng 7*:173–184.

Choe S, Bennett MJ, Fuji G, Curmi PMG, Kantardjieff KA, Collier RJ, Eisenberg D. 1992. The crystal structure of diphtheria toxin. *Nature 357*:216–222.

Choi SY, Esaki N, Ashiuchi M, Yoshimura T, Soda K. 1994. Bacterial glutamate racemase has high sequence similarity with myoglobins and forms an equimolar inactive complex with hemin. *Proc Natl Acad Sci USA 91*:10144–10147.

Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure of proteins. *EMBO J 5*:823–826.

Chothia C, Lesk AM. 1987. The evolution of protein structure. *Cold Spring Harbor Symp Quant Biol 52*:399–408.

Connolly ML. 1983. Solvent accessible surfaces of proteins and nucleic acids. *Science 221*:709–713.

Couture M, Chamberland H, St-Pierre B, Lafontaine J, Guertin M. 1994. Nuclear genes encoding chloroplast hemoglobins in the unicellular green alga *Chlamydomonas eugametos. Mol Gen Genet 243*:185–197.

Cramm R, Siddiqui RA, Friedrich B. 1994. Primary sequence and evidence for a physiological function of the flavohemoprotein of *Alcaligenes eutrophus. J Biol Chem 269*:7349–7354.

Dayhoff MO, Barker WC, Hunt LT. 1983. Establishing homologies in protein sequences. *Methods Enzymol 91*:524–545.

Doolittle RF. 1979. Protein evolution. In: Neurath H, Hill RL, eds. *The proteins, vol IV.* New York: Academic Press. pp 1–118.

Draper NR, Smith H. 1981. *Applied regression analysis.* New York: J. Wiley & Sons, Inc. pp 458–517.

Feng DF, Johnson MS, Doolittle RF. 1985. Aligning amino acid sequences: Comparison of commonly used methods. *J Mol Evol 21*:112–125.

Flores TP, Orengo CA, Moss DS, Thornton JM. 1993. Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci 2*:1811–1826.

Frenkel MJ, Dopheide TAA, Wagland BM, Ward CW. 1992. The isolation, characterization and cloning of a globin-like, host-protective antigen from the excretory secretory products of *Trichostrongylus colubriformis. Mol Biochem Parasitol 50*:27–36.

Fukuda M, Takagi T, Shikama K. 1993. Polymorphic hemoglobin from a midge larva (*Tokunagayusurika akamusi)* can be divided into two different groups. *Biochim Biophys Acta 1157*:185–191.

Fushitani K, Matsuura MSA, Riggs AF. 1988. The aminoacid sequences of chains a, b and c that form the trimer subunit of the extracellular hemoglobin of *Lumbricus terrestris. J Biol Chem 263*:65010–6517.

Gerstein M, Sonnhammer ELL, Chothia C. 1994. Volume changes in protein evolution. *J Mol Biol 236*:1067–1078.

Gilles-Gonzalez MA, Gonzalez G, Perutz MF. 1995. Kinase activity of oxygen sensor FixL depends on the spin state of its heme iron. *Biochemistry 34*:232–236.

Gilles-Gonzalez MA, Gonzalez G, Perutz MF, Kiger L, Marden M, Poyart C. 1994. Heme-based sensors, exemplified by the kinase FixL, are a new class of heme protein with distinctive ligand binding and autooxidation. *Biochemistry 33*:8067–8073.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA 89*:10915–10919.

Holm L, Ouzounis C, Sander C, Tupareve G, Vriend G. 1992. A database of protein structure families with common folding motifs. *Protein Sci 1*:1691–1698.

Holm L, Sander C. 1993a. Structural alignment of globins, phycocyanins and colicin A. *FEBS Lett 315*:301–305.

Holm L, Sander C. 1993b. Protein structure comparison by alignment of distance matrices. *J Mol Biol 233*:123–138.

Huber R, Epp O, Steigemann W, Formanek H. 1971. The atomic structure of erythrocruorin in the light of the chemical sequence and its comparison with myoglobin. *Eur J Biochem 19*:42–50.

Hurley JH. 1994. The role of interior side-chain packing in protein folding and stability. In: Merz K Jr, Le Grand S, eds. *The protein folding problem and tertiary structure prediction.* Boston: Birkhäuser. pp 549–578.

Iwaasa H, Takagi T, Shikama K. 1989. Protozoan myoglobin from *Paramecium caudatum. J Mol Biol 208*:355–358.

Iwaasa H, Takagi T, Shikama K. 1990. Protozoan hemoglobin from *Tetrahymena pyriformis. J Biol Chem 265*:8603–8609.

Jeffrey GA, Saenger W. 1991. *Hydrogen bonding in biological structures.* Berlin/Heidelberg: Springer Verlag.

Johnson MS, Overington JP. 1993. A structural basis for sequence comparisons; an evaluation of scoring methodologies. *J Mol Biol 233*:716–738.

Johnson MS, Overington JP, Blundell TL. 1993. Alignment and searching for common protein folds using a data bank of structural templates. *J Mol Biol 231*:735–752.

Johnson MS, Sali A, Blundell TL. 1990a. Phylogenetic relationships from three-dimensional protein structures, *Methods Enzymol 183*:670–690.

Johnson MS, Srinivasan N, Sowdhamini T, Blundell TL. 1994. Knowledge-based protein modeling. *Crit Rev Biochem Mol Biol 29*:1–69.

Johnson MS, Sutcliffe MJ, Blundell TL. 1990b. Molecular anatomy: Phyletic relationships from three-dimensional protein structures. *J Mol Evol 30*:43–59.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature 358*:86–89.

Kleinschmidt T, Sgouros JG. 1987. Hemoglobin sequences. *Biol Chem Hoppe-Seyler 368*:579–615.

Komiyama NH, Shih DT, Looker D, Tame J, Nagai K. 1991. Was the loss of the D helix in a globin a functionally neutral mutation? *Nature 352*:349–351.

Laurents DV, Subbiah S, Levitt M. 1994. Different protein sequences can give rise to highly similar folds through different stabilizing interactions. *Protein Sci 3*:1938–1944.

Lesk AM, Chothia C. 1980. How different amino acids sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J Mol Biol 136*:225–270.

Lesk AM, Chothia CH. 1986. The response of protein structures to amino acid sequence changes. *Phil Trans R Soc Lond A317*:345–356.

Lesk AM, Levitt M, Chothia C. 1986. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Protein Eng 1*:77–78.

Lim VI, Ptitsyn OB. 1970. On the constancy of the hydrophobic nucleus volume in myoglobins and hemoglobins. *Mol Biol (USSR) 4*:372–382.

Moens L, Vanfleteren Blaxter MLJ, van de Peer Y, Peeters K, Kapp OH, Goodman M, Vinogradov SN. 1995. Globins in nonvertebrate species: Dispersal by horizontal gene transfer and evolution of the structure-function relationships. *Mol Biochem Evol.* Forthcoming.

Orengo CA, Brown NP, Taylor WR. 1992. Fast structure alignment for protein databank searching. *Proteins Struct Funct Genet 14*:139–167.

Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993. Identification and classification of protein fold families. *Protein Eng 6*:485–500.

Orengo CA, Taylor WR. 1993. A local alignment method for protein structure motifs. *J Mol Biol 233*:488–497.

Parente A, Verde C, Malorni A, Montecucchi P, Aniello F, Geraci G. 1993. Amino acid sequence of the cooperative dimeric myoglobin from the radular muscle of the mollusc *Nassa mutabilis. Biochim Biophys Acta 1162*:1–9.

Parker MW, Postma JPM, Pattus F, Tucker AD, Tsernoglou D. 1992. Refined structure of the pore forming domain of colicin A at 2.4 Å. *J Mol Biol 224*:639–657.

Pastore A, Lesk AM. 1990. Comparison of the structures of globins and phycocyanins: Evidence for evolutionary relationship. *Proteins Struct Funct Genet 8*:133–155.

Pastore A, Lesk AM, Bolognesi M, Onesti S. 1988. Structural alignment and analysis of two distantly related proteins: *Aplysia limacina* myoglobin and sea lamprey globin. *Proteins Struct Funct Genet 4*:240–250.

Perutz MF, Kendrew JC, Watson HC. 1965. Structure and function of haemoglobin. II. Some relations between polypeptide chain configuration and amino acid sequence. *J Mol Biol 13*:669–678.

Potts M, Angeloni SV, Ebel RE, Bassam D. 1992. Myoglobin in a cyanobacterium. *Science 256*:1690–1692.

Powell FC. 1982. *Statistical tables for the social, biological and physical sciences.* Cambridge, UK: Cambridge University Press. p 78.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. *Numerical recipes in C.* New York: Cambridge University Press. pp 656–706.

Ptitsyn OB. 1974. Invariant features of globin primary structure and coding of their secondary structure. *J Mol Biol 88*:287–300.

Ptitsyn OB, Volkenshtein MV. 1986. Protein structures and the neutral theory of evolution. *J Biol Struct Dynam 4*:137–156.

Ratkowsky DA. 1990. *Handbook of nonlinear regression models*. New York: Marcel Dekker Inc. pp 75–122.

Riggs CK, Riggs AF. 1990. cDNA-derived amino acid sequences of single and two-domain globins from the clam *Barbatia reeveana*. In: Preaux G, Lontie R, eds. *Invertebrate dioxygen carriers*. Leuven: Leuven University Press. pp 57–60.

Rodionov MA, Johnson MS. 1994. Residue–residue contact substitution probabilities derived from aligned three-dimensional structures and the identification of common folds. *Protein Sci 3*:2366–2377.

Russell RB, Barton GJ. 1992. Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins Struct Funct Genet 14*:309–323.

Russell RB, Barton GJ. 1994. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts, secondary structure and accessibility. *J Mol Biol 244*:332–350.

Sachs L. 1984. *Applied statistics, a handbook of techniques*. New York: Springer Verlag. pp 447–453.

Sali A, Blundell TL. 1990. Definition of general topological equivalence in protein structures. *J Mol Biol 212*:403–428.

Schimenti JC, Duncan CH. 1985. Concerted evolution of the cow ϵ^2 and ϵ^4 β-globin genes. *Mol Biol Evol 2*:505–513.

Schwartz RM, Dayhoff MO. 1978. Matrices for detecting distant relationships. In: Dayhoff MO, ed. *Atlas of protein sequence and structure, vol 5, suppl 3*. Washington, DC: National Biomedical Research Foundation. pp 353–358.

Shenkin PS, Erman B, Mastrandrea LD. 1991. Information-theoretical entropy as a measure of sequence variability. *Proteins Struct Funct Genet 11*:297–313.

Sippl MJ, Weitckus S. 1992. Detection of native-like models for amino acid sequence of unknown three-dimensional structure in a data base of known protein conformations. *Proteins Struct Funct Genet 13*:258–271.

Subbiah S, Laurents DV, Levitt M. 1993. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr Biol 3*:141–148.

Suzuki T. 1989. Amino acid sequence of a major globin from the sea cucumber *Paracaudina chilensis*. *Biochim Biophys Acta 998*:292–296.

Suzuki T. 1994. Abalone myoglobins derived from indoleamine dioxygenase: The cDNA-derived amino acid sequence of myoglobin from *Nordotis madaka*. *J Protein Chem 14*:9–13.

Suzuki T, Furukohri T, Okamoto S. 1993. Amino acid sequence of myoglobin from the chiton *Liolophura japonica* and a phylogenetic tree for molluscan globins. *J Protein Chem 12*:45–50.

Suzuki T, Nakamura A, Satoh Y, Inai C, Furukohri T, Arita T. 1992. Primary structure of chain I of the heterodimeric hemoglobin from the blood clam *Barbatia virescens*. *J Protein Chem 11*:629–633.

Suzuki T, Takagi T. 1992. A myoglobin evolved from indoleamine-2,3-dioxygenase. *J Mol Biol 228*:698–700.

Takagi T. 1993. Hemoglobins from single-celled organisms. *Curr Biol 3*: 413–418.

Takagi T, Iwaasa H, Yuasa H, Shikama K, Takemasa T, Watanabe Y. 1993. Primary structure of *Tetrahymena* hemoglobin. *Biochim Biophys Acta 1173*:75–78.

Trotman CNA, Manning AM, Moens L, Guise KJ, Tate WP. 1991. Translation of the cDNA sequence for the polymeric hemoglobin of *Artemia*. In: Vinogradov SN, Kapp OH, eds. *Structure and function of invertebrate oxygen carriers*. New York: Springer Verlag. pp 207–216.

Vinogradov SN, Walz DA, Pohajdak B, Moens L, Kapp OH, Suzuki T, Trotman CNA. 1993. Adventitious variability? The amino acid sequences of nonvertebrate globins. *Comp Biochem Physiol 106B*:1–26.

Wakabayashi S, Matsubara H, Webster DA. 1986. Primary sequence of a dimeric bacterial haemoglobin from *Vitreoscilla*. *Nature 322*:481–483.

Zar JH. 1984. *Biostatistical analysis, 2nd ed.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc. pp 306–327.

Zhu H, Riggs AF. 1992. Yeast flavohemoglobin is an ancient protein related to globins and a reductase family. *Proc Natl Acad Sci USA 89*:5015–5019.