J Mol Evol (1988) 27:236–249

Journal of
Molecular Evolution
© Springer-Verlag New York Inc. 1988

An Evolutionary Tree for Invertebrate Globin Sequences

Morris Goodman,[1] Janet Pedwaydon,[1] John Czelusniak,[1] Tomohiko Suzuki,[2] Toshio Gotoh,[3] Luc Moens,[4] Fumio Shishikura,[5]* Daniel Walz,[5] and Serge Vinogradov[6]

Departments of Anatomy,[1] Physiology,[5] and Biochemistry,[6] Wayne State University School of Medicine, Detroit, Michigan 48201, USA
[2] Department of Biology, Faculty of Science, Kochi University, Kochi 780, Japan
[3] Department of Biology, College of General Education, Tokushima University, Tokushima 770, Japan
[4] Department of Biochemistry, University of Antwerp, Antwerp, Belgium

Summary. A phylogenetic tree was constructed from 245 globin amino acid sequences. Of the six plant globins, five represented the Leguminosae and one the Ulmaceae. Among the invertebrate sequences, 7 represented the phylum Annelida, 13 represented Insecta and Crustacea of the phylum Arthropoda, and 6 represented the phylum Mollusca. Of the vertebrate globins, 4 represented the Agnatha and 209 represented the Gnathostomata. A common alignment was achieved for the 245 sequences using the parsimony principle, and a matrix of minimum mutational distances was constructed. The most parsimonious phylogenetic tree, i.e., the one having the lowest number of nucleotide substitutions that cause amino acid replacements, was obtained employing clustering and branch-swapping algorithms. Based on the available fossil record, the earliest split in the ancestral metazoan lineage was placed at 680 million years before present (Myr BP), the origin of vertebrates was placed at 510 Myr BP, and the separation of the Chondrichthyes and the Osteichthyes was placed at 425 Myr BP. Local "molecular clock" calculations were used to date the branch points on the descending branches of the various lineages within the plant and invertebrate portions of the tree. The tree divided the 245 sequences into five distinct clades that corresponded exactly to the five groups plants, annelids, arthropods, molluscs, and vertebrates. Furthermore, the maximum parsimony tree, in contrast to the unweighted pair group and distance Wagner trees, was consistent with the available fossil record and supported the hypotheses that the primitive hemoglobin of metazoans was monomeric and that the multisubunit extracellular hemoglobins found among the Annelida and the Arthropoda represent independently derived states.

Key words: Globin — Invertebrate — Phylogenetic tree — Maximum parsimony

Introduction

Compared to the wealth of structural information available for vertebrate globin chains, relatively little is known about the structure of invertebrate globin chains. Although globins are found uniformly and with few variations in quaternary structure throughout all vertebrate lineages, they are only sporadically found among and within the invertebrate phyla, where they exhibit great variety in their quaternary structures. The invertebrate intracellular hemoglobins are generally monomeric, dimeric, and tetrameric, although higher polymeric forms can also be found (Mangum 1976; Terwilliger 1980). The invertebrate extracellular hemoglobins display a broader variation in molecular size, ranging from monomeric molecules comparable in size to myoglobin chains to highly aggregated molecules that can be up to a hundred times larger than vertebrate hemoglobin. They can be classified into four groups based on their quaternary structure (Vinogradov 1985): (1) single-domain, single-subunit molecules

Offprint requests to: M. Goodman
* *Current address:* Department of Biology, Nihon University School of Medicine, Tokyo 173, Japan

consisting of a single polypeptide chain of ca. 16 kd and containing one heme group (a well-studied example of this group is the multiple hemoglobins of the dipteran *Chironomus*); (2) two-domain, multisubunit hemoglobins, ranging in size from 250 to 800 kd and consisting of 30–40-kd chains, each containing two heme-binding domains [these molecules are found predominantly in carapaced branchiopod crustaceans such as *Caenestheria* and *Daphnia* (Daniel 1983)]; (3) multidomain, multisubunit hemoglobins consisting of two or more long polypeptide chains each containing 8–20 linearly connected heme-binding domains [such molecules are found in carapaceless branchiopod crustaceans such as the brine shrimp *Artemia* (Moens 1982) and in some bivalve and gastropod molluscs (Terwilliger and Terwilliger 1985)]; (4) single-domain, multisubunit hemoglobins consisting of aggregates of several small subunits, some of which are disulfide-bonded and not all of which contain heme. These molecules, sometimes called erythrocruorins, possess a highly characteristic two-tiered hexagonal shape in electron micrographs (Vinogradov et al. 1980, 1982). The annelid chlorocruorins, which differ only in having a slightly altered heme group, must be considered part of this group.

The amino acid sequences of many invertebrate globin chains are known: among the intracellular hemoglobins, the monomeric hemoglobin of the polychaete annelid *Glycera dibranchiata* (Imamura et al. 1972), the dimeric hemoglobins of the bivalve mollusc *Anadara broughtonii* (Furuta and Kajita 1983), the two chains of the tetrameric hemoglobin of *Anadara trapezia* (Como and Thompson 1980; Gilbert and Thompson 1985), one dimeric globin of *Anadara trapezia* (Fisher et al. 1984), and the dimeric globin of *Scapharca inaequivalvis* (Petruzelli et al. 1985). Among the extracellular hemoglobins, the published amino acid sequences of the following globin chains are known: all 12 globins of *Chironomus thummi thummi* (Buse et al. 1979), one of the heme-binding domains (E_1) of *Artemia salina* (Moens et al. 1986), four chains of the polychaete *Tylorrhynchus heterochaetus* (Suzuki et al. 1982, 1985a,b; Suzuki and Gotoh 1986), and two chains of the oligochaete *Lumbricus terrestris* (Garlick and Riggs 1982; Shishikura et al. 1987). In addition, the primary structures of the monomeric myoglobins of the molluscs *Aplysia limacina, Aplysia kurodai,* and *Aplysia juliana* (Tentori et al. 1973; Suzuki et al. 1981; Takagi et al. 1984) and *Dolabella auricularia* (Suzuki 1986) and of the dimeric myoglobins of the gastropods *Busycon canaliculatum* (Bonner and Laursen 1977) and *Cerithidea rhizophorarum* (Takagi et al. 1983) are known.

The plant globins, the leghemoglobins, form a separate group of monomeric globins whose mode of evolution appears to resemble that of their animal counterparts (Brown et al. 1984). The finding that leghemoglobins are not confined solely to the Leguminosae but are also present in the Ulmaceae (Landsmann et al. 1986) raises the possibility that globin genes occur as widely in plants as in animals.

In the present communication we compare the invertebrate globin sequences to each other, to plant globins, and to vertebrate globin sequences. We have constructed a tentative evolutionary tree for known invertebrate globins and plant globins, and relate it to the phylogenetic tree for vertebrate globins (Goodman et al. 1975, 1984, 1987a,b; Goodman 1981).

Methods and Materials

Altogether, 245 globin amino acid sequences were employed in this study. Of these sequences, 6 were from angiosperm plants (5 representing the Leguminosae and 1 representing the Ulmaceae), 26 were from invertebrates, and 213 were from vertebrates. Among the invertebrate sequences, 7 represented the classes Polychaeta and Oligochaeta of the phylum Annelida, 13 represented the Insecta and the Crustacea of the phylum Arthropoda, and 6 represented three groups of the phylum Mollusca: subclasses Opisthobranchia and Prosobranchia of the class Gastropoda and the class Bivalvia. Among the vertebrate globins, 4 represented the Agnatha (Cyclostomata) and 209 represented the Gnathostomata, the latter being a selection of the more than 400 amino acid sequences known for myoglobins and the hemoglobins of the jawed vertebrates.

Most of these globin sequences have been previously catalogued and aligned against one another (Goodman 1981; Goodman et al. 1983, 1987a). In the present study, previous alignments were reexamined and extended to include the amino acid sequences from the dimeric and tetrameric clam hemoglobins, from a domain of the extracellular multidomain hemoglobin of the arthropod *Artemia*, from monomeric globins of the extracellular multisubunit hemoglobins of the annelids *Tylorrhynchus* and *Lumbricus*, and from the monomeric globin of the plant *Parasponia*. The principle of maximizing sequence matches or minimizing sequence differences, i.e., the parsimony principle, was followed in aligning the sequences. We first determined a series of pairwise alignments using the algorithm of Needleman and Wunsch (1970). The pairwise alignment scores, determined by computer, then served as a guide for aligning all 245 sequences against one another. This common alignment was achieved heuristically. It entailed evaluating by the maximum parsimony approach different genealogical (phylogenetic) arrangements and placing gaps that maximized sequence similarities that could be attributed to common ancestry while minimizing convergences.

The approach was iterative. After a tentative common alignment was achieved for all 245 sequences, a matrix of minimum mutation distances was constructed according to Jukes (1963) and Fitch and Margoliash (1967). Using this distance matrix, an unweighted pair-group tree of the 245 sequences was constructed by the clustering algorithm of Sokal and Michener (1958), and a distance Wagner tree constructed by the algorithm of Farris (1972). These two trees served as starting points in the search by branch-swapping algorithms (Goodman et al. 1979, 1984; Goodman 1981) for the most parsimonious tree, i.e., the tree of lowest NR length (NR = *N*ucleotide substitutions that cause amino acid *R*eplacements). After the search had revealed the lowest NR

```
                                    10        20        30        40        50
                                     |         |         |         |         |
Parasponia Lhb                SSSEVNKV FTEEQEALVVKAWA   VMKKNSAELGLQFFLKIF
Lupin Lhb                            GVLTDVQVALVKSSFE   EFNANIPKNTHRFFTLVL
Vicia Lhb                            G FTEKQEALVNSSSQ   LFKQNPSNYSVLFYTIIL
Phaseolus Lhb                        GAFTEKQEALVNSSWE   AFKGNIPQYSVVFYTSIL
Glycine C2 Lhb                       GAFTEKQEALVSSSFE   AFKTNIPQYSVVFYTSIL
Glycine Lhb                          VAFTEKQDALVSSSFE   AFKANIPQYSVVFYTSIL
Tyl Hb IIB                   DDCCSAADR HEVLDNWKGIWSAEFTGRRVAIGQAIFQELFALDPN
Tyl Hb IIC                   DTCCSIEDR REVQALWRSIWSAEDTGRRTLIGRLLFEELFEIDGA
Lumbricus HbI                  ECLVTEG LKVKLQWASAFGHAHQ  RVAFGLELWKGILREHPE
Lumbricus HbII               KKQCGVLEG LKVKSEWGRAYGSGHD  REAFSQAIWRATFAQVPE
Tyl Hb IIA                  SSDHCGPLQR LKVKQQWAKAYGVGHE  RVELGIALWKSMFAQDND
Tyl Hb I                      TDCGILQR IKVKQQWAQVYSVGES  RTDFAIDVFNNFFRTNPD
Glycera Hb                           G LSAAQRQVIAATWKDI AGNDNGAGVGKDCLIKHL
Artemia Hb                    ERVDPITG LSGLEKNAILDTWG   KVRGNL    QEVGKATFGKLF
CTT Hb IA                            GP SGDQIAAAKASWN   TVKNNQ    VDILYAVF
CTT Hb I                             GP SGDQIAAAKASWN   TVKNNQ    VDILYAVF
CTT Hb III alpha               VATPAMPSMTDAQVAAVKGDWE   KIKGSG    VEILYFFL
CTT Hb III                             LSADQISTVQASFD   KVKGDP    VGILYAVF
CTT Hb IV                              LTADQISTVQSSFA   GVKGDA    VGILYAVF
CTT Hb X                         DPEWHTLDAHEVEQVQATWK   AVSHDE    VEILYTVF
CTT Hb IX                            DPVSSDEANAIRASWA   GVKHNE    VDILAAVF
CTT Hb VIIA                          APLSADQASLVKSTWA   QVRNSE    VEILAAVF
CTT Hb II beta                       APLSADEASLVRGSWA   QVKHSE    VDILYYIF
CTT Hb VIIB                          SPLTADEASLVQSSWK   AVSHNE    VDILAAVF
CTT Hb VI                            AVLTTEQADLVKKTWS   TVKFNE    VDILYAVF
CTT Hb VIII                        AVTPMSADQLALFKSSWN   TVKHNE    VDILYAVF
Anadara b. Hb                PSVQGAAAQ LTADVKKDLRDSWKV   IGSDKKGNGVALMTTLF
Anadara t. Hb                VADAVAKVC GSEAIKGNLRRSWGVL  MSADIEATGLTYLANLF
Busycon Mb                           G LDGAQKTALKESWKVLGADGPTMMKNGSLLFGLLF
Cerithidea Mb                        S LQPASKSALASSWKTLAKDAATIQNNGATLFSLLF
Aplysia k. Mb                        S LSAAEADLVGKSWA   PVYANKDADGANFLLSLF
Aplysia l. Mb                        S LSAAEADLAGKSWA   PVFANKNANGADFLVALF
Myxine Hb                    PITDHGQPPTLSEGDKKAIRESWP   QIYKNFEQNSLAVLLEFL
Lampetra Hb                  PIVDSGSVAPLSAAEKTKIRSAWA   PVYSNYETSGVDILVKFF
Homo Mb                              G LSDGEWQLVLNVWG   KVEADIPGHGQEVLIRLF
Homo Hb beta                         VHLTPEEKSAVTALWG   KV  NVDEVGGEALGRLL
Homo Hb alpha                        V LSPADKTNVKAAWG   KVGAHAGEYGAEALERMF
```

Fig. 1. The alignment of the 6 plant, 26 invertebrate, and 5 of the 213 vertebrate globin amino acid sequences used in the construction of the phylogenetic tree

length trees, the alignment for the 245 sequences was reevaluated and realignments tested by the maximum parsimony method. That is, realignments that further lowered NR length on resuming the search for the most parsimonious tree were retained. In this iterative heuristic search procedure, alternative phylogenetic hypotheses on the relationships of the sequences served as bases for trying out possible realignments, the effect of each of which on NR length was then recorded.

The minimum number of NR needed to account for the branching arrangement of the sequences, the maximum parsimony score, was determined by two programs, MPALMX and MPAFEP, which use an algorithm that takes into account the genetic code. These procedures allow subtrees to be fixed: the set of codons corresponding to the parsimony solution for the ancestor of each subtree is computed and is used as a terminal taxon. The program MPALMX computes the scores of all possible trees with eight terminal taxa and the program MPAFEP iteratively tries to lower the score of an input tree by branch swapping.

Ancestral codons and branch lengths were calculated by the program TPAB, which determines these sequences and lengths by the parsimony method. Ambiguities in parsimony assignments of codons, different ancestral codons each giving the same NR score, were resolved by choosing codons that would minimize the sum of the distances on the tree for every pair of terminal taxa. The distance between terminal taxa on the tree is the sum of lengths of the branches connecting the two taxa. Numbers of nucleotide replacements on each link were corrected for superimposed mutations by the program TAVA. This algorithm propagates mutational information from pairs of nodes more populated by intervening links to those less populated (Moore 1977; Baba et al. 1981).

All of these programs were run on a Cray-2 supercomputer at the University of Minnesota. Time on this computer was obtained through the NSF supercomputer access program. These four programs (MPALMX, MPAFEP, TPAB, and TAVA) are written in FORTRAN and are available from the authors.

Results and Discussion

Figure 1 shows the alignment of the plant and invertebrate globin sequences with five of the vertebrate globin sequences used in the construction of our phylogenetic tree. A notable feature is that the sequence of the monomeric globin of *Glycera* and the six sequences of the extracellular, multisubunit annelid hemoglobins share three unique gaps (at alignment positions 61, 69–75, and 102–104). Similarly, all arthropod sequences, domain E_1 of the multidomain *Artemia* hemoglobin, and the 12 *Chironomus* sequences share a unique gap at positions 35–38.

Table 1 shows the matrix of minimum mutational distances in selected pairwise comparisons taken from the full set of pairwise comparisons among the 245 globin amino acid sequences. Each pairwise comparison value is presented as the minimum mutational difference (MMD) over the number of amino acid residue positions compared (the

```
                                60        70        80        90       100
                                 |         |         |         |         |
Parasponia Lhb          EIAPSAKNLFSYLKDSP VPLEQN PKLKPHATTVFVMTCESAVQLRKAG
Lupin Lhb               EIAPGAKDLFSFLKGSSEVPQNN  PDLQAHAGKVFKLTYEAAIQLE  V
Vicia Lhb               QKAPTAKAMFSFLKDSAGVVDS   PKLGAHAEKVFGMVRDSAVQLR  A
Phaseolus Lhb           EKAPAAKNLFSFLAN  GVDPTN  PKLTAHAESLFGLVRDSAAQLR  A
Glycine C2 Lhb          EKAPAVKDLFSFLAN  GVNPTN  PKLTGHAEKLFGLVRDSAGQLK  A
Glycine Lhb             EKAPAAKDLFSFLAN  GVDPTN  PKLTGHAEKLFALVRDSAGQLK  A
Tyl Hb IIB              A KGVFGRVN VDK PSE       ADWKAHVIRVINGLDLAVNLLEDPK
Tyl Hb IIC              T KGLFKRVN VDD THS       PEEFAHVLRVVNGLDTLIGVLGDSD
Lumbricus HbI           I KAPFSRVR GDN IYS       PQFGAHSQRVLSGLDITISMLDTPD
Lumbricus HbII          S RSLFKRVH GDH TSD       PAFIAHAERVLGGLDIAISTLDQPA
Tyl Hb IIA              A RDLFKRVH GED VHS       PAFEAHMARVFNGLDRVISSLTDEP
Tyl Hb I                  RSLFNRVN GDN VYS       PEFKAHMVRVFAGFDILISVLDDKP
Glycera Hb              SAHPQMAAVF GFSGASD       PAVADLGAKVLAZIGVAVSHLGDZG
Artemia Hb              AAHPEYQQMFRFFQG VQLAFLVQSPKFAAHTQRVVSALDQT  LLALNR
CTT Hb IA               KANPDIQTAFSQFAG KDLDSIKGTPDFSKHAGRVVGLFSEVMDLLGNDA
CTT Hb I                KANPDIQTAFSQFAG KDLDSIKGTPDFSKHAGRVVGLFSEVMDLLGNDA
CTT Hb III alpha        NKFPGNFPMFKKL G NDLAAAKGTAEFKDQADKIIAFLQGVIEKLGSD
CTT Hb III              KADPSIMAKFTQFAG KDLESIKGTAPFETHANRIVGFFSKII   GELP
CTT Hb IV               KADPSIQAKFTQFAG KDLDSIKGSADFSAHANKIVGFFSKII   GDLP
CTT Hb X                KAHPDIMAKFPKFAG KDLEAIKDTADFAVHASRIIGFFGEYVTLLGSSG
CTT Hb IX               SDHPDIQARFPQFAG KDLASIKDTGAFATHAGRIVGFISEIVALVGNES
CTT Hb VIIA             TAYPDIQARFPQFAG KDVASIKDTGAFATHAGRIVGFVSEIIALIGNES
CTT Hb II beta          KANPDIMAKFPQFAG KDLETLKGTGQFATHAGRIVGFVSEIVALMGNSA
CTT Hb VIIB             AAYPDIMAKFPQFAG KDLASIKDTGAFATHATRIVSFLSEVIALMGNAS
CTT Hb VI               KAYPDIMAKFPQFAG KDLDSIKDSAAFATHATRIVSFLSEVISLAGSDA
CTT Hb VIII             KANPDIQAKFPQFAG KDLDSIKDSADFAVHSGRIVGFFSEVIGLIGNPE
Anadara b. Hb           ADNQETIGYFKRLGN VSQG  MANDKLRGHSITLMYALQNFIDQLDNTD
Anadara t. Hb           TLRPDTKTYFTRLGD VQKG  KANSKLRGHAITLTYALDWFVDSLDDPS
Busycon Mb              KTYPDTKKHFKHFDD ATFAAMDTTGVGKAHGVAVFSGLGSMICSIDDDD
Cerithidea Mb           KQFPDTRNYFTHFGN MSDAEMKTTGVGKAHSMAVFAGIGSMIDSMDDAD
Aplysia k. Mb           EKFPNNANYFADFKG KSIADIKASPKLRDVSSRIFTRLNEFV   NNAA
Aplysia l. Mb           EKFPDSANFFADFKG KSVADIKASPKLRDVSSRIFTRLNEFV   NDAA
Myxine Hb               KKFPKAQDSFPKFSAKKSH  LEQDPAVKLQAEVIINAVNHTIGLMDKEA
Lampetra Hb             TSTPAAQEFFPKFKGMTSADELKKSADVRWHAERIINAVNDAVASMD  D
Homo Mb                 KGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHE
Homo Hb beta            VVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLK
Homo Hb alpha           LSFPTTKTYFPHF DLSH     GSAQVKGHGKKVADALTNAVAHVDDMP
```

Fig. 1. Continued.

n-alignment or numbers of sequence positions in which amino acid residues occur in both sequences). The selected comparisons involve all the sequences shown in Fig. 1. These comparisons reveal that the globin sequences from plants, *Glycera,* Arthropoda, Mollusca, agnathans, and gnathostomes are significantly related to one another. Also the sequences of the extracellular, multisubunit hemoglobins of *Tylorrhynchus* and *Lumbricus* are related to one another at high significance levels and at a lower, but still quite significant, level to the *Glycera* monomeric sequence. These judgments can be made easily by looking up each MMD value in Table 1, for that n-alignment, the critical values listed in table 3 of the paper by Moore and Goodman (1977); for the comparisons that we claim show significant homology, the observed MMD values are sufficiently small as to reject by this alignment statistic of Moore and Goodman (1977) the null hypothesis of no common ancestry. The *Tylorrhynchus* I sequence appears to be the most conserved of the annelid extracellular hemoglobin sequences in that its MMD values with nonannelid hemoglobin sequences consistently indicate significant homology. The *Tylorrhynchus* IIA, IIB, and IIC and the *Lumbricus* I and II sequences all tend to yield MMD values with nonannelid sequences that do not reject the null hypothesis of no common ancestry; however, their MMD values with *Aplysia* myoglobin sequences do reject the null hypothesis of no common ancestry, i.e., are indicative of significant sequence homology.

Just as the sequences from annelid extracellular, multisubunit hemoglobins show lower MMD values with the *Glycera* sequence than with any of the nonannelid sequences, the domain E_1 sequence from the extracellular, multisubunit hemoglobin of *Artemia* shows lower MMD values with *Chironomus* hemoglobin sequences than with any of the nonarthropod sequences. As judged by MMD values, the monomeric *Chironomus* hemoglobins CTT I and CTT Ia have the most conserved arthropod hemoglobin sequences. Similarly, the MMD values in Table 1 indicate that the most conserved mollusc globin sequences are those of the monomeric myoglobins of *Aplysia.* Of the mollusc globin sequences, those from the dimeric myoglobins of *Busycon* and *Cerithidea* and from the tetrameric hemoglobin of *Anadara* have higher MMD values than have the *Aplysia* monomeric sequences when compared to nonmollusc globins.

Figure 2 shows all the plant and invertebrate lineages and 5 of the 213 vertebrate lineages from the phylogenetic tree constructed for the 245 globin sequences on the basis of the maximum parsimony

```
                                      110       120       130       140       150
                                       |         |         |         |         |
Parasponia Lhb                K VTVKESDLKRIGAIHFK    TGVVNE  HF EVTRFALLETIKEAVP
Lupin Lhb                     N GAV ASDA TLKSVHVS    KGVVDA  HFPVVKE AILKTIKEVVG
Vicia Lhb                     T GEV VADG KDGSIHIQ    KGVLDP  HFVVVKE ALLKTIKEASG
Phaseolus Lhb                 N GAV VADA ALGSIHSQ    KGVSND  QFLVVKE ALLKTLKQAVG
Glycine C2 Lhb                   TV VADA ASGSIHAQ    KAITNP  EF VVKE ALLKTIKEAVG
Glycine Lhb                   S GTV VADA ALGSVHAQ    KAVTNP  EF VVKE ALLKTIKAAVG
Tyl Hb IIB                    A   L QEELKHLARQHRERSGVKAVYFD  EMEKALL KVLPQVSS H
Tyl Hb IIC                    T   L NSLIDHLAEQHKARAGFKTVYFK  EFGKALN HVLPEVAS C
Lumbricus HbI                 M   L AAQLAHLKVQHVER NLKPEFFD  IFLKHLL HVLGDRLGTH
Lumbricus HbII                T   L KEELDHLQVQHEGR KIPDNYFD  AFKTAIL HVVAAQLGDA
Tyl Hb IIA                    V   L NAQLEHLRQQHIKL GITGHMFN  LMRTGLA YVLPAQLGRC
Tyl Hb I                      V   L DQALAHYAAFHKQFGTIPFKAFGQTMFQTIAE HIHGAD
Glycera Hb                    K   M VAQMKAVGVRHKGY GNKHIKGQ  YFEPLGA SLLSAMEHRIG
Artemia Hb                    PSDQF VYMIKELGLDHIN    RG T         DR SFVEYLKESL
CTT Hb IA                     NTPTI LAKAKDFGKSHKS    RT SPA  QLDNFRK SLVVYLKGAT
CTT Hb I                      NTPTI LAKAKDFGKSHKS    RA SPA  QLDNFRK SLVVYLKGAT
CTT Hb III alpha               MGGA KALLNQLGTSHKA    MGITKD  QFDQFRQ ALTELL GNL
CTT Hb III                    NIEAD    VNTFVASHKP    RGVTHD  QLNNFRA GFVSYMKAHT
CTT Hb IV                     NIDGD    VTTFVASHTP    RGVTHD  QLNNFRA GFVSYMKAHT
CTT Hb X                      NQAAI RTLLHDLGVFHKT    RGITKA  QFGEFRE TMTAYLKGHN
CTT Hb IX                     NAPAM ATLINELSTSHHN    RGITKG  QFNEFRS SLVSYLSSHA
CTT Hb VIIA                   NAPAV QTLVGQLAASHKA    RGISQA  QFNEFRA GLVSYVSSNV
CTT Hb II beta                NMPAM ETLIKDMAANHKA    RGIPKA  QFNEFRA SLVSYLQSKV
CTT Hb VIIB                   NAAAV QGLLDKLGDDHKA    RGVSAA  QFGEFRT ALVAYLQAHV
CTT Hb VI                     NIPAI QNLAKELATSHKP    RGVSKD  QFTEFRT ALFTYLKAHI
CTT Hb VIII                   NRPAL KTLIDGLASSHKA    RGIEKA  QFEEFRA SLVDYLSHHL
Anadara b. Hb                 DLVCV VEKFA    VNHIT    RKISAA  EFGKING PIKKVL ASK
Anadara t. Hb                 RLKCV VEKFA    VNHIN    RKISGD  AFGSIIP EMKETLKARMG
Busycon Mb                    CVBGL AKKLS    RNHLA    RGVSAA  DFKLLE  AVFKZFLD EA
Cerithidea Mb                 CMNGL ALKLS    RNHIQ    RKIGAS  RFGEMR  QVFPNFLD EA
Aplysia k. Mb                 DAGKM SAMLSQFASEHVG    FGVGSA  QFENVR  SMFPAFVASLS
Aplysia l. Mb                 NAGKM SAMLSQFAKEHVG    FGVGSA  QFENVR  SMFPGFVASVA
Myxine Hb                     AMKKY    LKDLSTKHSTE   FQVNPD  MFKELSA VFVSTM
Lampetra Hb                   TEKMS    MKDLSGKHAKS   FQVDPQ  YFKVLA  VIADTV
Homo Mb                       AE       IKPLAQSHATK   HKIPVK  YLEFISE CIIQVLQSKHP
Homo Hb beta                  GT       FATLSELHCDK   LHVDPE  NFRLLGN VLVCVLAHHFG
Homo Hb alpha                 NA       LSALSDLHAHK   LRVDPV  NFKLLSH CLLVTLAAHLP
```

Fig. 1. Continued.

method. In a previous study (Goodman et al. 1987a) involving 218 globin sequences, 212 of the present 213 vertebrate globins had been employed, but only 6 nonvertebrate globins (an *Aplysia* sequence, the *Glycera* sequence, and 4 of the *Chironomus* sequences) served as outgroups of the vertebrate sequences. In this previous study as in earlier ones (Goodman et al. 1974, 1975; Goodman 1981), the gnathostome (jawed vertebrate) α- and β-hemoglobin branches, after grouping, were closest to the gnathostome myoglobin branch and next closest to the agnathan globin branch. In the present study, it proved slightly more parsimonious to group the agnathan globin branch first either with the gnathostome myoglobin branch or, as shown in Fig. 2, with the gnathostome hemoglobin branch. Otherwise the phylogenetic arrangements found for the more than 200 vertebrate globins were very similar in the present and previous studies. Thus, the previous study (Goodman et al. 1987a) may be consulted for details on the branching patterns within the vertebrate region of the globin tree. A finding that should be noted is that the most parsimonious branching arrangement for the plant and invertebrate regions of the globin tree, namely the one shown in Fig. 2, was not altered by placing the gnathostome myoglobin branch first either with the gnathostome hemoglobin branch, or alternatively, with the agnathan globin branch. Also, in the search for the most parsimonious globin tree we could choose either of the following two alternatives without altering the branching arrangements shown in Fig. 2. We could impose the constraint that species relationships found among eutherian mammals parallel each other in the myoglobin and α- and β-hemoglobin regions of the tree, as we did for the search that gave the results used in Fig. 2, or we could allow differing patterns of eutherian relationships to be depicted by the three types of globins when this lowered NR length, as we did on examining several hundred thousand alternative trees.

The genealogical trees found by these heuristic maximum parsimony search procedures divided the 245 eukaryotic globins into five major phylogenetic clades. Starting with the branch most distant from vertebrates and proceeding toward the vertebrates, all 6 plant globins group in the first clade, all 7 annelid globins in the second, all 13 arthropod globins in the third, all 6 mollusc globins in the fourth, and all 213 vertebrate globins in the fifth. This correspondence between the groups formed by the globin sequences and the groups expected from tradi-

```
                                    160       170       180       190       200
                                     |         |         |         |         |
Parasponia Lhb              EMWSPEMKNAWGVAYDQ  LVAAIKFEMKPSST
Lupin Lhb                   DKWSEELNTAWTIAYDE  LAIIIKKEMKDAA
Vicia Lhb                   DKWSEELSAAWEVAYDG  LATAIKAA
Phaseolus Lhb               DKWTDQLSTALELAYDE  LAAAIKKAYA
Glycine C2 Lhb              DKWSDELSSAWEVAYDE  LAAAIKKAF
Glycine Lhb                 DKWSDELSRAWEVAYDE  LAAAIKAK
Tyl Hb IIB                    FN   SGAWDRCFTRI AD VIKAELP
Tyl Hb IIC                    FN   PEAWNHCFDGL VD VISHRIDG
Lumbricus HbI                 FD   FGAWHDCVDQ  ID GIKDI
Lumbricus HbII                IA   CDGFARVLPQV LERGIKGHH
Tyl Hb IIA                    FD   KEAWAACWDEV IYPGIKHD
Tyl Hb I                           IGAWRACYAEQ IVTGITA
Glycera Hb                  GKMNAAAKDAWAAAYAD  ISGALISGLQS
Artemia Hb                    GDSVDEF   TVQSFGEVIVNFLNEGLRQA
CTT Hb IA                    KWDSAVESSWAPVLDF  VFSTLKNEL
CTT Hb I                     KWDSAVESSWAPVLDF  VFSTLKNEL
CTT Hb III alpha             GFGGNIG AWNATVDL  MFHVIFNALDGTPV
CTT Hb III                   DF AGAEAAWGATLDT  FFGMIFSKM
CTT Hb IV                    DF AGAEAAWGATLDA  FFGMVFAKM
CTT Hb X                     KWNADISHSWDDAFDK  AFSVIFEVLES
CTT Hb IX                    SWNDATADAWTHGLDN  IFGMIFAHL
CTT Hb VIIA                  AWNAAAESAWTAGLDN  IFGLLFAAL
CTT Hb II beta               SWNDSLGAAWTQGLDN  VFNMMFSYL
CTT Hb VIIB                  SWGNNVAAAWSKALDN  TFAIVVPRL
CTT Hb VI                    NFDGPTETAWTLALDT  TYAMLFSAMDS
CTT Hb VIII                  DWNDTMKSTWDLALNN  MFFYILHALEVAQ
Anadara b. Hb               N FGDKYANAWAKLVAV  VQAAL
Anadara t. Hb               S YSDDVGAAWVQAILG  MQNAVLSAL
Busycon Mb                  T  QRKATDAQKDADGA  LLTMLIKAHV
Cerithidea Mb               L  GGGASGDVKGAWDA  LLAYLQDNKQ
Aplysia k. Mb               A  PPA DDAWNKLFGL  IVAALKAAGK
Aplysia l. Mb               A  PPAGADAWTKLFGL  IIDALKAAGK
Myxine Hb                        GGKAAYEKLFSI  IATLLRSTYD
Lampetra Hb                     AAGDAGFEKLSMI  CILMLRSAY
Homo Mb                     GDFGADAQGAMNKALEL  FRKDMASNYKELGFQG
Homo Hb beta                KEFTPPVQAAYQKVVAG  VANALAHKYH
Homo Hb alpha               AEFTPAVHASLDKFLAS  VSTVLTSKYR
```

Fig. 1. Continued.

tional phylogenetic evidence on eukaryotic taxa is, in our opinion, a significant finding that speaks well for the validity of the maximum parsimony method in reconstructing phylogeny. The two trees, the unweighted pair group tree and the distance Wagner tree, produced from the MMD matrix by the distance clustering algorithm did not show such correspondence. Although the two trees agreed with the maximum parsimony trees in monophyletically grouping the plant globins together, all extracellular annelid globins together, all *Chironomus* globins together, the two *Anadara* globins together, the opisthobranch (*Aplysia*) globins together, the prosobranch (*Busycon* and *Cerithidea*) globins together, and all vertebrate globins together, they did not agree in other respects. They failed to group the opisthobranch and prosobranch clades together and then join this gastropod branch to the bivalve mollusc branch as did the maximum parsimony trees. Nor did the unweighted and distance Wagner trees depict a monophyletic Annelida and a monophyletic Arthropoda as the maximum parsimony trees did. The unweighted and distance Wagner trees had lengths of 6960 NR and 6873 NR, respectively, whereas the maximum parsimony tree found under the constraint that the same pattern of eutherian relationships be depicted by myoglobin and by α- and β-hemoglobin sequences had a length of 6843 NR. Without this constraint, i.e., when the three types of globin sequences were allowed to depict differing patterns of eutherian relationships as long as the branch swaps lowered NR length, the maximum parsimony tree had a length of 6777 NR. Although this tree showed more internal inconsistencies between the three globin regions than the constrained maximum parsimony tree 6843 NR long, it showed far fewer inconsistencies than the unweighted and the distance Wagner trees. Moreover, within the vertebrates as well as among the invertebrates, both the 6777-NR tree and the constrained maximum parsimony 6843-NR tree, agreed with the traditional phylogenetic evidence on the taxa represented by globin sequences much more so than the unweighted and distance Wagner trees.

The phylogenetic hypotheses that we tested by the initial trees submitted to the branch-swapping algorithms were not limited to relationships suggested by the traditional evidence on plant and metazoan phylogeny, but also included hypotheses suggested by structural and functional features of the globins. For example, the sequence from the extracellular multisubunit hemoglobin of *Artemia*

Table 1. Matrix of minimum mutational distance values for selected pairwise comparisons of the globin sequences used in the present study

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 *Parasp*	110 147	99 141	106 142	104 139	106 140	187 136	194 137	175 133	180 136	180 135	162 127	158 140	174 138	160 138	159 138	179 146	158 132
2 *Lupin*	–	91 143	93 145	86 141	84 142	160 128	180 129	171 126	167 128	174 127	158 119	161 139	163 128	157 137	156 137	162 138	151 130
3 *Vicia*	–	–	60 141	58 138	62 140	170 126	168 126	169 126	171 127	174 127	149 119	151 136	153 122	150 134	149 134	151 131	150 128
4 *Phaseol*	–	–	–	30 141	34 142	169 126	171 126	167 124	162 126	171 125	149 117	158 136	155 124	142 136	141 136	149 134	151 129
5 *Glyc* C2	–	–	–	–	15 140	165 123	167 123	162 122	162 124	169 123	142 115	147 133	155 121	139 133	138 133	151 131	149 126
6 *Glyc*	–	–	–	–	–	160 123	165 123	162 123	163 124	173 124	145 116	149 133	156 122	147 134	146 134	151 131	146 127
7 *Tyl* IIB	–	–	–	–	–	–	109 148	139 140	141 144	144 143	154 136	177 136	166 122	164 123	163 123	166 127	163 117
8 *Tyl* IIC	–	–	–	–	–	–	–	135 140	155 144	140 143	138 136	177 137	178 123	162 123	163 123	167 128	157 117
9 *Lumb* I	–	–	–	–	–	–	–	–	108 141	118 141	119 133	166 134	163 118	168 122	168 122	163 125	160 116
10 *Lumb* II	–	–	–	–	–	–	–	–	–	123 145	133 136	167 136	166 122	167 124	166 124	172 127	162 118
11 *Tyl* IIA	–	–	–	–	–	–	–	–	–	–	113 136	162 135	159 121	167 123	166 123	170 126	157 117
12 *Tyl* I	–	–	–	–	–	–	–	–	–	–	–	150 127	159 113	160 115	160 115	164 119	146 110
13 *Glycera*	–	–	–	–	–	–	–	–	–	–	–	–	141 121	133 131	133 131	155 130	141 125
14 *Artemia* E1	–	–	–	–	–	–	–	–	–	–	–	–	–	131 129	130 129	168 135	140 123
15 CTT IA	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1 143	137 138	97 134
16 CTT I	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	136 138	96 134
17 CTT III a	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	118 131
18 CTT III	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
19 CTT IV	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
20 CTT X	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
21 CTT IX	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
22 CTT VIIA	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
23 CTT IIb	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
24 CTT VIIB	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
25 CTT VI	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
26 CTT VIII	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
27 *Anadara b.*	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 1. Continued

19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
159 132	174 146	167 140	162 140	167 140	161 140	166 142	180 146	184 140	175 146	178 140	181 140	145 140	152 141	174 140	170 139	187 146	171 140	160 136
147 130	163 141	170 139	161 139	166 139	156 139	163 141	166 143	174 131	173 137	161 138	172 138	155 138	157 139	176 132	167 133	183 143	178 139	168 134
146 128	155 136	168 136	157 136	156 136	157 136	162 136	161 136	162 130	163 135	161 135	170 135	152 135	158 136	162 128	161 130	183 137	168 135	161 132
146 129	151 139	156 138	151 138	148 138	151 138	147 139	151 139	152 130	161 136	157 137	169 137	156 137	158 138	158 130	161 131	181 138	160 137	158 132
142 126	148 135	157 135	150 135	150 135	155 135	155 135	151 135	147 127	157 133	161 133	166 133	152 133	156 134	154 126	158 128	177 135	168 134	162 129
138 127	150 136	158 136	155 136	153 136	152 136	155 136	158 136	151 129	161 134	165 134	164 134	147 134	149 135	156 127	160 129	173 135	164 134	155 129
160 117	171 130	173 125	167 125	171 125	163 125	175 126	172 128	171 129	170 135	167 128	170 128	163 125	159 125	183 132	166 128	171 128	166 126	164 127
153 117	167 131	163 125	157 125	156 125	153 125	167 127	158 129	169 129	183 135	169 128	174 128	172 125	166 125	189 132	178 128	179 129	159 126	173 127
156 116	182 128	174 124	170 124	171 124	170 124	179 124	175 126	166 128	183 133	164 124	160 124	161 123	159 123	176 127	171 124	172 126	155 124	169 125
163 118	183 130	168 126	171 126	174 126	164 126	180 126	180 128	175 131	171 137	165 126	163 126	172 125	167 125	174 131	160 128	173 128	157 126	157 127
155 117	184 129	171 125	169 125	180 125	164 125	174 125	179 127	184 131	190 136	169 125	166 125	168 124	164 124	181 130	186 127	175 127	172 125	181 126
141 110	169 121	164 117	155 117	163 117	165 117	173 117	167 119	166 124	171 127	153 119	153 119	153 117	156 117	181 127	182 124	169 119	150 117	161 118
139 125	139 135	136 133	137 133	144 133	143 133	148 135	143 135	160 129	161 136	163 135	175 135	149 133	149 134	156 129	143 126	160 139	144 136	155 137
142 123	163 136	143 130	147 130	153 130	145 130	152 132	164 136	150 126	160 131	155 125	170 125	159 126	158 127	152 126	151 125	152 128	148 123	140 119
100 134	114 143	107 143	101 143	105 143	109 143	101 143	93 143	150 130	161 135	162 135	164 135	144 135	140 136	155 130	144 130	164 136	154 135	147 130
99 134	113 143	106 143	100 143	104 143	180 143	100 143	92 143	151 130	162 135	161 135	165 135	143 135	139 136	155 130	144 130	165 136	154 135	148 130
121 131	130 146	136 140	124 140	123 140	127 140	124 142	129 146	154 134	164 138	157 133	167 133	151 133	156 134	171 135	172 134	162 138	158 133	157 129
24 136	108 136	101 136	88 136	94 136	98 136	86 136	106 136	151 124	155 129	162 129	160 129	161 132	163 133	147 127	155 127	154 133	156 131	133 127
—	113 136	102 136	89 136	104 136	101 136	90 136	105 136	146 124	150 129	158 129	156 129	157 132	159 133	151 127	150 127	153 133	149 131	134 127
—	—	102 145	105 145	100 145	94 145	106 147	99 149	155 136	167 141	172 138	179 138	159 138	162 139	179 137	164 136	168 140	181 138	162 133
—	—	—	57 145	67 145	81 145	91 145	91 145	161 132	170 137	160 137	167 137	149 137	146 138	153 132	149 132	171 138	164 137	149 132
—	—	—	—	64 145	70 145	81 145	79 145	162 132	168 137	157 137	161 137	142 137	138 138	148 132	141 132	160 138	160 137	145 132
—	—	—	—	—	80 145	88 145	84 145	157 132	160 137	166 137	166 137	149 137	151 138	150 132	144 132	156 138	165 137	150 132
—	—	—	—	—	—	82 145	91 145	140 132	152 137	153 137	154 137	136 137	134 138	157 132	150 132	165 138	158 137	146 132
—	—	—	—	—	—	—	91 147	161 132	172 137	163 138	174 138	147 138	149 139	158 133	155 132	167 140	172 138	156 133
—	—	—	—	—	—	—	—	165 134	172 139	169 138	166 138	154 138	160 139	166 135	150 134	170 142	164 138	156 133
—	—	—	—	—	—	—	—	—	115 146	144 135	144 135	147 131	148 132	161 134	161 133	143 131	143 129	145 127

Table 1. Continued

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
28 *Anadara t.*	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
29 *Busycon* Mb	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
30 *Cerith* Mb	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
31 *Apl k.* Mb	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
32 *Apl l.* Mb	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
33 *Myxine* Hb	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
34 *Lamp* Hb	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
35 *Homo* Mb	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
36 *Homo* beta	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
37 *Homo* alpha	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

The number of residue positions compared in each case is given underneath each MMD

was joined to the annelid extracellular multisubunit hemoglobin branch, while the monomeric *Glycera* sequence was joined to the monomeric–dimeric *Chironomus* branch and then the two resulting branches, *Artemia*–annelid and *Glycera–Chironomus,* were joined together. The length of this arrangement was 20 NR more than the length of the arrangement shown in Fig. 2, i.e., the program MPAFE after several branch swaps returned the tree to the branching pattern in Fig. 2 and in the process lowered the tree length by 20 NR.

Two globin phylogenetic trees were constructed recently by Feng et al. (1985) incorporating only *Lumbricus* chain II and *Tylorrhynchus* chain I data. They provide similar branching orders for the various globins except for the two annelid sequences relative to the kidney bean leghemoglobin. In the tree based on the empirical log-odds matrix the leghemoglobin diverges at the same time as the divergence of the annelid branch and the line leading to the vertebrates. In the other tree based on a unitary matrix program whereby only identities are scored, the leghemoglobin line diverges at a more distal point, in agreement with the tree derived by Goodman et al. (1974) by the maximum parsimony method.

The geologic time scale shown on the right side of Fig. 2 is taken from Harland et al. (1982). Paleontological views concerning the ancestral separation of the species from which the globins came were used to place branch points on this geological time scale. Despite the antiquity of the Earth [ca. 4550 million years before present (Myr BP)], the fossil record indicates that metazoans only became abundant during the Ediacarian period of the late Precambrian, which lasted from 670 Myr BP to 550 Myr BP and in the early Cambrian (Cloud and Glassner 1982). The initial metazoan radiation spanned a period of approximately 150 Myr (ca. 680–530 Myr) and is best known from the relatively abrupt appearance of hard parts near the base of the Cambrian (ca. 570 Myr) (Morris 1985). Thus, the earliest split in the ancestral metazoan lineage is placed arbitrarily in our globin tree at 680 Myr BP. The origin of vertebrates and the divergence of agnathans from gnathostomes is usually placed near the Cambrian–Ordovician boundary, i.e., at ca. 510 Myr BP (Romer 1966; Løvtrup 1977). Thus we place the separation of the agnathan globin lineage from the stem to the gnathostome α- and β-hemoglobin lineages at 510 Myr BP. In addition, the separation of Chondrichthyes (sharks) from Osteichthyes (bony fishes), which paleontological evidence places at ca. 425 Myr BP, served to define the time zone within which the divergence between the gnathostome α- and β-hemoglobin lineages could be placed. By extrapolation from the NR values of the two internodal links, the date for the α/β divergence node was found to be about 455 Myr BP. Similarly, within the plant and invertebrate regions of the globin tree local "molecular clock" calculations were used to date branch points shown in Fig. 2.

Table 1. Continued

19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
–	–	–	–	–	–	–	–	–	–	161 142	152 142	154 137	157 138	171 138	157 137	163 137	161 135	144 133
–	–	–	–	–	–	–	–	–	–	–	109 147	160 140	157 141	158 132	159 132	170 138	159 136	149 132
–	–	–	–	–	–	–	–	–	–	–	–	156 140	156 141	157 132	161 132	158 138	168 136	158 132
–	–	–	–	–	–	–	–	–	–	–	–	–	22 144	146 131	139 131	154 138	162 136	136 132
–	–	–	–	–	–	–	–	–	–	–	–	–	–	152 132	136 132	157 139	165 137	142 133
–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	109 143	164 135	142 134	139 131
–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	143 134	143 133	113 128
–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	164 145	151 141
–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	96 139
–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

The branching pattern found for leghemoglobins and shown in Fig. 2 agrees with the view of botanists on the phylogenetic relationships of the species from which these plant globins were obtained. The divergence of Leguminosae and Ulmaceae (which contains *Parasponia*) is thought to have occurred in the middle Cretaceous, about 80–100 Myr BP (J. J. Doyle and K. Nixon, personal communication). The four genera of the Leguminosae, represented in Fig. 2 by *Lupinus, Vicia, Phaseolus,* and *Glycine,* all belong to the subfamily Papilionideae. The divergence of *Lupinus* first from the other genera agrees with fossil evidence for its earliest existence during the Eocene about 40 Myr BP; the divergence of *Vicia* (broad bean) next, then *Phaseolus* (kidney bean), and finally *Glycine* (soybean) in the legume portion of the globin tree is also in accordance with current views on legume systematics (Polhill 1981).

The annelid portion of the globin tree shown in Fig. 2 is mainly composed of paralogous lineages. The extracellular globins IIB and IIC of the polychaete *Tylorrhynchus* are closely related. This result is in agreement with the recent finding, based on the amino acid sequence determinations of the N-terminal portions (20–25 residues) of the four *Tylorrhynchus* chains and the corresponding four *Lumbricus* chains, that the eight globin chains fall into two groups: group A, consisting of *Lumbricus* chains I and II and *Tylorrhynchus* chains I and IIA, shares the invariant Lys-12 and Lys-14 and group B, consisting of *Lumbricus* chains III and IV and *Tylorrhynchus* chains IIB and IIC, shares the invariant Cys-4, Ser-6, and Asp-9 (Gotoh et al. 1987). The maximum parsimony results, however, indicate that the invariant Lys-12 and Lys-14 of *Lumbricus* chains I and II and *Tylorrhynchus* chains I and IIA are primitive retentions in these extracellular globins rather than shared derived characters. On the other hand, the sequences that fall into group B probably do constitute a monophyletic clade as judged by the close grouping of *Tylorrhynchus* chains IIB and IIC in the maximum parsimony tree in Fig. 2. The fossil record for the annelids is very incomplete, particularly for the oligochaetes. Probably all that can be said is that the known geologic range of the annelids is Proterozoic to recent (Tasch 1980).

The complementary DNA-derived amino acid sequence of one of the chains of the intracellular, tetrameric hemoglobin of the echiuran *Urechis caupo* has been determined recently (Garey and Riggs 1986). It shows maximum homology (ca. 20%) with the intracellular, monomeric globin of the polychaete *Glycera,* in agreement with the accepted relationship between the Echiura (a minor protostome phylum) and the annelids (Mettam 1985). Although this sequence was not included in the construction of the phylogenetic tree reported here, it probably should be placed close to *Glycera.* Interestingly, the nucleotide sequences of the 5S rRNA from an oligochaete, *Enchytraeus albidus,* two polychaetes, *Perinereis brevicirris* and *Sabellastarte japonica,* and of an echiuran, *Urechis unicinctus,* suggested that the latter was closer to the oligochaete than to the polychaetes and also closer to the oligochaete than were all three annelids to each other (Specht et al. 1986).

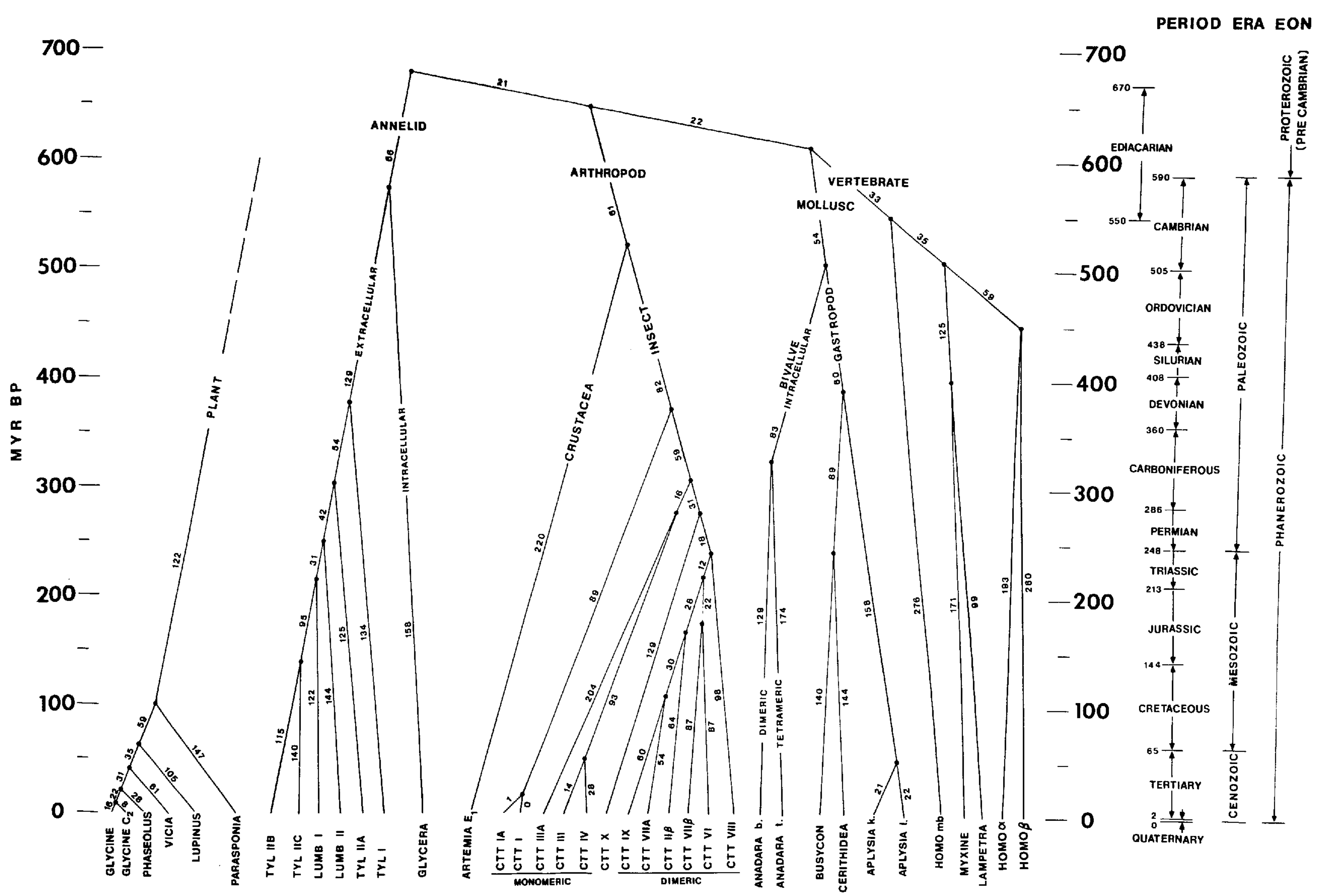

Fig. 2. The plant and invertebrate lineages and 5 of the 213 vertebrate lineages from the phylogenetic tree of globin amino acid sequences. The branching arrangement for this tree was found using parsimony programs MPALMX and MPAFP. The numbers shown on the lines of descent represent the nucleotide replacements between the ancestral and descendant codon sequences as determined by the maximum parsimony program TPAB and then corrected for superimposed mutations by the program TAVA (see Methods and Materials). The geological time scale given on the right is taken from Harland et al. (1982).

The primitive hemoglobin of metazoans was probably monomeric. If so, the multisubunit hemoglobins represent independently derived states in Annelida and Arthropoda. Lending support to this hypothesis, the ancestral sequence that the parsimony method found for annelid globins diverges much less from the monomeric globin of *Glycera* than from any extracellular annelid globins (see Fig. 2). Similarly, the ancestral sequence found for the arthropod globins diverges less from some of the monomeric globins of *Chironomus* than from the multisubunit hemoglobin of *Artemia.*

The portion of the tree representing the relationships among the 12 globins of *Chironomus* is essentially identical to the tree presented earlier (Goodman et al. 1983), in which the 6 globins with a tendency to dimerize group separately from the 5 that are monomeric in solution. CTT X, which exists both as monomer and dimer, occupies an intermediate position between the two groups. The relationship between the insect globins, which are single-chain, single-domain globins, and one of the domains of the multidomain, multisubunit hemoglobin of the branchiopod crustacean *Artemia* is very distant, as could be expected. Although the fossil record of the crustaceans extends from the late Cambrian to Recent, the fossil record of the Branchiopoda starts with the Late Devonian (Schram, 1982). The first insects are observed in the middle of Early Devonian, about 390 Myr BP.

Among the intracellular mollusc hemoglobins, in addition to the sequences representing the Bivalvia used in the present study, namely the dimeric globin of *Anadara broughtonii* (Furuta and Kajita 1983) and the alpha chain of the tetrameric hemoglobin of *Anadara trapezia* (Como and Thompson 1980), the sequences of the beta chain of *Anadara trapezia* (Gilbert and Thompson 1985) and of its dimeric globin IIB (Fisher et al. 1984) and of the dimeric globin of *Scapharca inaequivalvis* (Petruzelli et al. 1985) have become available since the start of this study. All three dimeric globins share a 90% identity, and hence should be grouped closely together. In contrast, the dimeric globins have only 45% identity with the two chains of tetrameric hemoglobin of *Anadara trapezia.* As already noted, in the maximum parsimony globin tree shown in Fig. 2, the clam hemoglobin branch joins the gastropod myoglobin branch and this branch, in turn, divides into the dimeric prosobranch sequences of *Busycon* and *Cerithidea* and the monomeric opisthobranch sequences of *Aplysia.* The fact that among the gastropod myoglobins, the monomeric sequence diverges less from the ancestral state than the dimeric sequence (see link lengths in Fig. 2) further supports the hypothesis that the primitive globins in metazoans were probably monomeric. The sequence of the myoglobin of the opisthobranch *Dolabella auricularia* has been determined recently (Suzuki 1986); since this gastropod mollusc belongs to the family Aplysiidae, it is not surprising that its sequence shows a very strong similarity ranging from 72 to 77%, with the two *Aplysia* sequences used in this study and the recent *Aplysia juliana* sequence. The earliest fossil record for a bivalve is from the Late Cambrian, about 540–570 Myr BP (Pojeta et al. 1973). Although the fossil record of the prosobranch gastropods is from the Late Cambrian to Recent, the opisthobranch gastropods occur in the Devonian to Recent (Cox 1960).

The crystal structures of the intracellular monomeric hemoglobin of *Glycera dibranchiata* (Padlan and Love 1974), the leghemoglobin of *Lupinus* (Vainshtein 1981), the extracellular monomeric hemoglobin of *Chironomus thummi thummi* (CTT III) (Steigemann and Weber 1979), the myoglobin of *Aplysia limacina* (Bolognesi et al. 1985), and the intracellular dimeric and tetrameric hemoglobins of the arcid clam *Scapharca inaequivalvis* (Royer et al. 1985) have been determined. The structures all have the usual eight helices (A through H) arranged in the typical "globin fold" (Perutz 1979). The clam molecules appear to have an additional alpha helix at the N terminus and a subunit arrangement that is different from that of the vertebrates: although in the latter the E and F helices are external while the G and H helices are largely internal, in the clam tetramer it is the E and F helices that are largely internal while the G and H helices are external. Comparison of sequences of hemoglobins and myoglobins from many species has led to the conclusion that the split of the genes for the alpha and beta chains occurred after the emergence of the vertebrates (Goodman et al. 1975). Since both the alpha and beta chains are essential for the allosteric cooperativity in vertebrate hemoglobins (Benesch and Benesch 1974), it appears that the development of cooperativity in the clam hemoglobins was evolutionarily independent of that which occurred in vertebrate hemoglobins.

It is evident that many more globin sequences are necessary before any detailed phylogenetic tree can be constructed for the three major groups of invertebrates represented in the present study.

Acknowledgments. This work was supported in part by grants NIH HL 33940 and NSF BSF 86-07202 and the Alfred P. Sloan Foundation (M.G.) and grants NIH AM/DK 38376 (S.N.V.) and NIH AM/DK 30382 (D.A.W.). We thank Drs. J. Doyle, K. Nixon, J. Farmer, and W. Schopf for valuable discussions concerning the paleontology and systematics of plants and metazoans.

References

Baba ML, Darga LL, Goodman M, Czelusniak J (1981) Evolution of cytochrome c investigated by the maximum parsimony method. J Mol Evol 17:197–203

Benesch R, Benesch RE (1974) Homos and heteros among the hemos. Science 185:905–908

Bolognesi M, Coda A, Gatti G, Ascenzi P, Brunori M (1985) Crystal structure of ferric *Aplysia limacina* myoglobin at 2.0Å resolution. J Mol Biol 183:113–115

Bonner AG, Laursen RA (1977) The amino acid sequence of a dimeric myoglobin from the gastropod mollusc *Busycon canaliculatum.* FEBS Lett 73:201–203

Brown GG, Lee JS, Brisson N, Verma DPS (1984) The evolution of a plant globin gene family. J Mol Evol 21:19–32.

Buse G, Stettens GJ, Braunitzer G, Steer W (1979) Hamoglobine. XXV Hamoglobin (Erythrocruorin) CTTIII aus *Chironomus thummi thummi*: Primarstruktur und Beziehung zu anderer Hemproteine. Hoppe Seyler's Z Physiol Chem 360: 89–97

Cloud P, Glassner MF (1982) The Ediacaran period and system: Metazoa inherit the earth. Science 217:783–788

Como PF, Thompson EOP (1980) Amino acid sequence of the alpha chain of the tetrameric haemoglobin of the bivalve mollusc *Anadara trapezia.* Aust J Biol Sci 33:653–664

Cox LR (1960) Gastropoda: general characteristics. In: Moore RC (ed) Treatise on invertebrate paleontology, part I. University of Kansas Press, Lawrence, pp 85–169

Daniel E (1983) Subunit structure of arthropod erythrocruorin. Life Chem Rep, Suppl 1, pp 157–166

Farris JS (1972) Estimating phylogenetic trees from distance matrices. Am Nat 106:645–668

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

Fisher WK, Gilbert AT, Thompson EOP (1984) Amino acid sequence of the globin IIB chain of the dimeric haemoglobin of the bivalve mollusc *Anadara trapezia.* Austr J Biol Sci 37: 191–203

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155:279–284

Furuta H, Kajita A (1983) Dimeric hemoglobin of the bivalve mollusc *Anadara broughtonii*: complete amino acid sequence of the globin chain. Biochemistry 22:917–922

Garey JR, Riggs AF (1986) The hemoglobin of *Urechis capo.* J Biol Chem 261:16446–16450

Garlick RL, Riggs A (1982) The amino acid sequence of a major polypeptide chain of earthworm hemoglobin. J Biol Chem 257:9005–9015

Gilbert AT, Thompson EOF (1985) Amino acid sequence of the beta chain of the tetrameric hemoglobin of the bivalve mollusc *Anadara trapezia.* Austr J Biol Sci 38:221–236

Goodman M (1981) Decoding the pattern of protein evolution. Prog Biophys Mol Biol 37:105–164

Goodman M, Moore GW, Barnabas J (1974) The phylogeny of human globin genes investigated by the maximum parsimony method. J Mol Evol 3:1–48

Goodman M, Moore GW, Matsuda G (1975) Darwinian evolution in the genealogy of hemoglobin. Nature 253:603–608

Goodman M, Czelusniak J, Moore GW, Romero-Herrera A, Matsuda G (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst Zool 28:132–163

Goodman M, Braunitzer G, Kleinschmidt I, Aschauer H (1983) The analysis of a protein polymorphism. Evolution of monomeric and dimeric hemoglobins of *Chironomus thummi thummi.* Hoppe Seyler's Z Physiol Chem 364:205–217

Goodman M, Koop BF, Czelusniak J, Wiess ML, Slightom JL (1984) The η-globin gene: its long evolutionary history in the β-globin gene family of mammals. J Mol Biol 180:803–823

Goodman M, Miyamoto MM, Czelusniak J (1987a) Pattern and process in vertebrate phylogeny revealed by coevolution of molecules and morphologies. In: Patterson C (ed) Molecules and morphology in evolution: conflict or compromise? Cambridge University Press, pp 141–176

Goodman M, Czelusniak J, Koop BF, Tagle DA, Slightom JL (1987b) Globins: a case study in molecular phylogeny. Cold Spring Harbor Symp Quant Biol 52 (in press)

Gotoh T, Kamada Y (1980) Subunit structure of erythrocruorin from the polychaete *Tylorrhynchus heterochaetus.* Biochem J (Tokyo) 87:557–562

Gotoh T, Shishikura F, Snow JS, Ereifej KI, Vinogradov SN, Walz DA (1987) Two globin strains in the giant annelid extracellular haemoglobins. Biochem J 241:441–445.

Harland WB, Cox AV, Llewellyn PG, Pickton CAG, Smith AG, Walters R (1982) A geologic time scale. Cambridge University Press, pp 7–55

Imamura T, Baldwin TO, Riggs A (1972) The amino acid sequence of the monomer hemoglobin component from the bloodworm *Glycera dibranchiata.* J Biol Chem 247:2785–2797

Jukes TH (1963) Some recent advances in studies of the transcription of the genetic message. Adv Biol Med Phys 9:1–41

Landsmann J, Dennis ES, Higgins TJV, Appleby CA, Kortt AA, Peacock WJ (1986) Common evolutionary origin of legume and non-legume plant haemoglobins. Nature 324:166–168

Løvtrup S (1977) The phylogeny of Vertebrata. Wiley, London

Mangum M (1976) Primitive respiratory adaptations. In: Newell PC (ed) Adaptation to environment: physiology of marine animals. Butterworth's, London, pp 191–278

Mettam C (1985) Functional constraints in the evolution of the Annelida. In: Morris SC, George JD, Gibson R, Platt HM (eds) The origins and relationships of lower invertebrates. Clarendon Press, Oxford, pp 297–309

Moens L (1982) The extracellular hemoglobin of *Artemia salina.* A biochemical and ontogenetical study. Acad Anal 44: 1–21

Moens L, Van Hauwaeert ML, Geelen D, Verproten G, Van Beeumen J (1986) The amino acid sequence of a structural unit isolated from the high molecular weight globin chains of *Artemia* sp. In: Linzen B (ed) Invertebrate oxygen carriers. Springer, Berlin, pp 81–84

Moore GW (1977) Proof of the populous path algorithm for missing mutations in parsimony trees. J Theor Biol 66:95–101

Moore GW, Goodman M (1977) Alignment statistic for identifying related protein sequences. J Mol Evol 9:121–130

Morris SC (1985) Non-skeletalized lower invertebrate fossils: a review. In: Morris SC, George JD, Gibson R, Platt HM (eds) The origins and relationships of lower invertebrates. Clarendon Press, Oxford, pp 343–359

Needleman SB, Wunsch CB (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 98:443–453

Padlan EA, Love WE (1974) Three-dimensional structure of the hemoglobin from the polychaete annelid *Glycera dibranchiata* at 2.5Å resolution. J Biol Chem 249:309–338

Perutz M (1979) Regulation of oxygen affinity of hemoglobin: influence of structure of the globin on the heme iron. Annu Rev Biochem 48:327–386

Petruzelli R, Goffredo BM, Barra D, Bossa F, Boffi A, Verzili D, Ascoli F, Chiancone E (1985) Amino acid sequence of the cooperative homodimeric hemoglobin from the mollusc *Scapharca inaequivalvis* and topology of intersubunit contacts. FEBS Lett 184:328–332

Pojeta J, Runnegar B, Kriz J (1973) *Fordilla troyensis* Barrande: the oldest known pelecypod. Science 180:866–868

Polhill RM (1981) Papilionideae. In: Polhill RM, Raven PH (eds) Advances in legume systematics, part I. Royal Botanic Gardens, Kew, pp 191–208

Romer AS (1966) Vertebrate paleontology, ed 3. University of Chicago Press, Chicago

Royer WE, Love WE, Fenderson FF (1985) The cooperative dimeric and tetrameric chain hemoglobins are novel assemblages of myoglobin folds. Nature 316:277–280

Schram FR (1982) The fossil record and evolution of Crustacea. In: Abele LG (ed) The biology of the Crustacea, vol 1, pp 94–147

Shishikura F, Snow JS, Gotoh T, Vinogradov SN, Walz DA (1987) The amino acid sequence of the monomer subunit of the extracellular hemoglobin of *Lumbricus terrestris*. J Biol Chem 262:3123–3131

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kans Sci Bull 38:1409–1438

Specht T, Ulbrich N, Erdmann VA (1986) Nucleotide sequence of the 5S rRNA from the Annelida species *Enchytraeus albidus*. Nucleic Acids Res 14:4372

Steigemann W, Weber E (1979) Structure of erythrocruorin in different ligand states refined at 1.4Å resolution. J Mol Biol 127:309–338

Suzuki T (1986) Amino acid sequence of myoglobin from the mollusc *Dolabella auricularia*. J Biol Chem 261:3692–2699

Suzuki T, Gotoh T (1986) The complete amino acid sequence of giant multisubunit hemoglobin from the polychaete *Tylorrhynchus heterochaetus*. J Biol Chem 261:9257–9267

Suzuki T, Takagi T, Shikama K (1981) Amino acid sequence of myoglobin from *Aplysia kurodai*. Biochim Biophys Acta 669:79–83

Suzuki T, Takagi T, Gotoh T (1982) Amino acid sequence of the smallest polypeptide chain containing heme of extracellular hemoglobin from the polychaete *Tylorrhynchus heterochaetus*. Biochim Biophys Acta 708:253–258

Suzuki T, Furukohri T, Gotoh T (1985a) Subunit structure of extracellular hemoglobin from the polychaete *Tylorrhynchus heterochaetus* and amino acid sequence of the constituent polypeptide chain (IIC). J Biol Chem 260:3145–3154

Suzuki T, Yasunaga H, Furukohri T, Nakamura K, Gotoh T (1985b) Amino acid sequence of polypeptide chain IIB of extracellular hemoglobin from the polychaete *Tylorrhynchus heterochaetus*. J Biol Chem 260:11481–11487

Takagi T, Tobita M, Shikama K (1983) Amino acid sequence of dimeric myoglobin from *Cerithidea rhizophorarum*. Biochim Biophys Acta 745:32–36

Takagi T, Iida S, Matsuoka A, Shikama K (1984) *Aplysia* myoglobins with an unusual amino acid sequence. J Mol Biol 180: 1179–1184

Tasch P (1980) Paleobiology of the invertebrates. Wiley, New York, pp 441–470

Tentori L, Vivaldi G, Carta S, Marinucci M, Massa A, Antonini E, Brunori M (1973) The amino acid sequence of myoglobin from the mollusc *Aplysia limacina*. Int J Pept Protein Res 5: 182–200

Terwilliger RC (1980) Structure of invertebrate hemoglobins. Am Zool 20:53–67

Terwilliger RC, Terwilliger NB (1985) Molluscan hemoglobins. Comp Biochem Physiol B Comp Biochem 81B:255–261

Vainshtein BK (1981) The structure of leghemoglobin. In: Dodson G, Glusker CJP, Sayre D (eds) Structural studies of molecular biological interest. Oxford University Press, pp 39–43

Vinogradov SN (1985) The structure of invertebrate extracellular hemoglobins (erythrocruorins and chlorocruorins). Comp Biochem Physiol B Comp Biochem 82B:1–15

Vinogradov SN, Shlom JM, Kapp OH, Frossard P (1980) The dissociation of annelid extracellular hemoglobins and their quaternary structure. Comp Biochem Physiol B Comp Biochem 67B:1–16

Vinogradov SN, Kapp OH, Ohtsuki M (1982) The extracellular haemoglobins and chlorocruorins of annelids In: Harris J (ed) Electron microscopy of proteins, vol 3. Academic Press, London, pp 135–164

Received September 6, 1987/Revised November 16, 1987