

## ORIGINAL ARTICLE

L. Banci · I. Bertini · A. Rosato · G. Varani

**Mitochondrial cytochromes *c*: a comparative analysis**

Received: 4 June 1999 / Accepted: 28 September 1999

**Abstract** The structures of 113 eukaryotic cytochrome *c* proteins of known sequence have been modeled in the oxidized state based on the existing crystallographic and NMR structures. The secondary structural elements and the overall three-dimensional structure were found to be maintained throughout the superfamily, despite variability in the sequence of individual proteins. The iron axial ligands and their reciprocal orientation were found to be nearly universally conserved. Residues constituting the hydrophobic core of the protein are also very highly conserved or conservatively substituted. Certain surface-exposed charged as well as hydrophobic groups have also been found to be conserved to the same degree as core residues. Patterns of conservation of exposed residues identify regions of the protein that are likely to be critical for its function in electron transfer.

**Key words** Cytochrome *c* · Electron transfer · Structure

**Introduction**

Cytochrome *c* (cyt *c*) is an intensively studied protein because of its central role in electron transfer in living organisms. Its diverse functional roles and the availability of high-resolution crystallographic data since the early 1970s have contributed to make this protein a paradigm in the study of electron transfer processes

[1, 2]. During aerobic respiration, cyt *c* shuttles electrons from the  $bc_1$  complex to cytochrome *c* oxidase (CCO) in the mitochondria. In chloroplasts, cyt *c* shuttles electrons from the cytochrome *bf* complex to photosystem I. In prokaryotes, cyt *c* is involved in both aerobic and anaerobic respiration. The recent findings that cyt *c* is a key factor in apoptosis [3], and the discovery that defects in cyt *c* biogenesis may induce pathogenic responses related to copper [4] and iron metabolism [5], and, in some prokaryotes, to heme biosynthesis [6], have provoked further interest in this protein.

In addition to providing a key paradigm in electron transfer studies, cyt *c* has also been a model system in the study of the evolution of protein sequence and structure [7–11]. Approximately 15 years ago, Chothia and Lesk [12] studied how the structure of *c*-type cytochromes had evolved during evolution to satisfy the requirement to provide a stable architectural framework for the heme pocket. The comparison of approximately 10 experimental structures of *c*-type cytochromes, determined by X-ray crystallography, allowed the authors to pinpoint the crucial structural features of this family of proteins that characterize its protein fold [12]. The explosion in information derived from bioinformatics and structural biology provides an opportunity to re-examine this classic study. Solution structures of cytochromes *c* are now available [13–18], to go with systematic studies of dynamic properties like NH exchange and backbone dynamics [17, 19–23] that can be related to local mobility. Most importantly, there are now very large numbers of primary sequences of cyt *c* and related heme proteins in the databases. The exploitation of the sequence and structural information through molecular modeling allowed us to compare structural features of all 113 eukaryotic cyt *c* sequences within current sequence banks.

The aim of the present study was to examine patterns of sequence variation and conservation among both buried and surface-exposed residues in eukaryotic cyt *c* proteins. Examination of sequence conser-

L. Banci · I. Bertini (✉) · A. Rosato  
Department of Chemistry and Centro Risonanze Magnetiche  
University of Florence, Via Luigi Sacconi 6  
I-50019 Sesto Fiorentino, Italy  
e-mail: bertini@cerm.unifi.it  
Tel.: +39-055-4209272  
Fax: +39-055-4209271

G. Varani  
Medical Research Council, Laboratory of Molecular Biology  
MRC Centre, Hills Road, Cambridge CB2 2QH, UK

vation of amino acids buried in the core of the proteins presented us with the opportunity to re-examine exhaustively the primary sequence requirements that are compatible with a functional cyt *c* protein. The existence of highly conserved surface-exposed residues revealed instead sites of potential importance for protein-protein interactions in electron transfer.

## Methods

Mitochondrial cyt *c* sequences were searched for in the GenBank (<http://www.ncbi.nlm.nih.gov/>) database using sequence similarity criteria. This was accomplished either by starting from known sequences (horse heart and *Saccharomyces cerevisiae* iso-1 cyt *c*) and performing a BLAST [24] search, or, later, by looking for cytochrome sequences from specific organisms in the genome database. Sequence alignments were done with CLUSTALW [25]. Cyt *c* structures were modeled with the program MODELLER v. 4.0 [26], using as reference structure the energy minimized solution structure of oxidized horse heart cyt *c* (PDB entry 1AKK) [16]. Amino acid residues have been numbered on the basis of the alignment of all 113 cytochromes. In the present work, residues are numbered according to the horse heart cyt *c* sequence, as commonly done. The alignment generates a 10-residue extension before the first residue and single amino acid insertions after residues 21 and 73 (horse heart numbering). The 10-residue extension was numbered from -10 to -1 (there is no residue numbered 0). The insertions after residues 21 and 73 were referred to with the numbers 21a and 73a, respectively.

The models derived from MODELLER v. 4.0 were energy minimized in vacuo, with the program AMBER 5.0 [27], using the force field parameters for the oxidized heme [28]. Sequences of subunit II of CCO were searched for, modeled, and energy minimized following the same approach as for cyt *c*. The X-ray structure of bovine CCO was used as the structural template (PDB entry 1OCC) [29]. Only the 19 CCO sequences for which the companion cyt *c* sequence was available were analyzed. Secondary structure analyses were performed with PROCHECK [30]. Solvent accessibility, surface potentials, and hydrogen bonding for individual residues were evaluated with MOLMOL v. 2.6 [31]. Two residues were assumed to be in contact if at least two atoms were closer than 4 Å.

## Results

### The structure of cyt *c*

There are structures of cyt *c* both in the solid state [32–41] and in solution [13–18]. They contain a heme prosthetic group, covalently bound to the polypeptide chain through a thioether linkage with two cysteines. The heme contains an iron ion. The latter is six-coordinated, the ligand atoms being the four pyrrole nitrogens of the heme moiety, the N $\epsilon$ 1 atom of the side chain of one histidine (His18), and the sulfur atom of the side chain of one methionine (Met80). The above cyt *c* structures show a N-terminal helix involving the residues up to roughly residue 15 (in the present numbering scheme). This helix is followed by a loop, containing the histidine ligand to the iron. A long stretch of residues without any regular secondary structure leads into the second helix, which involves residues 50–56. This helix is followed by a short loop,

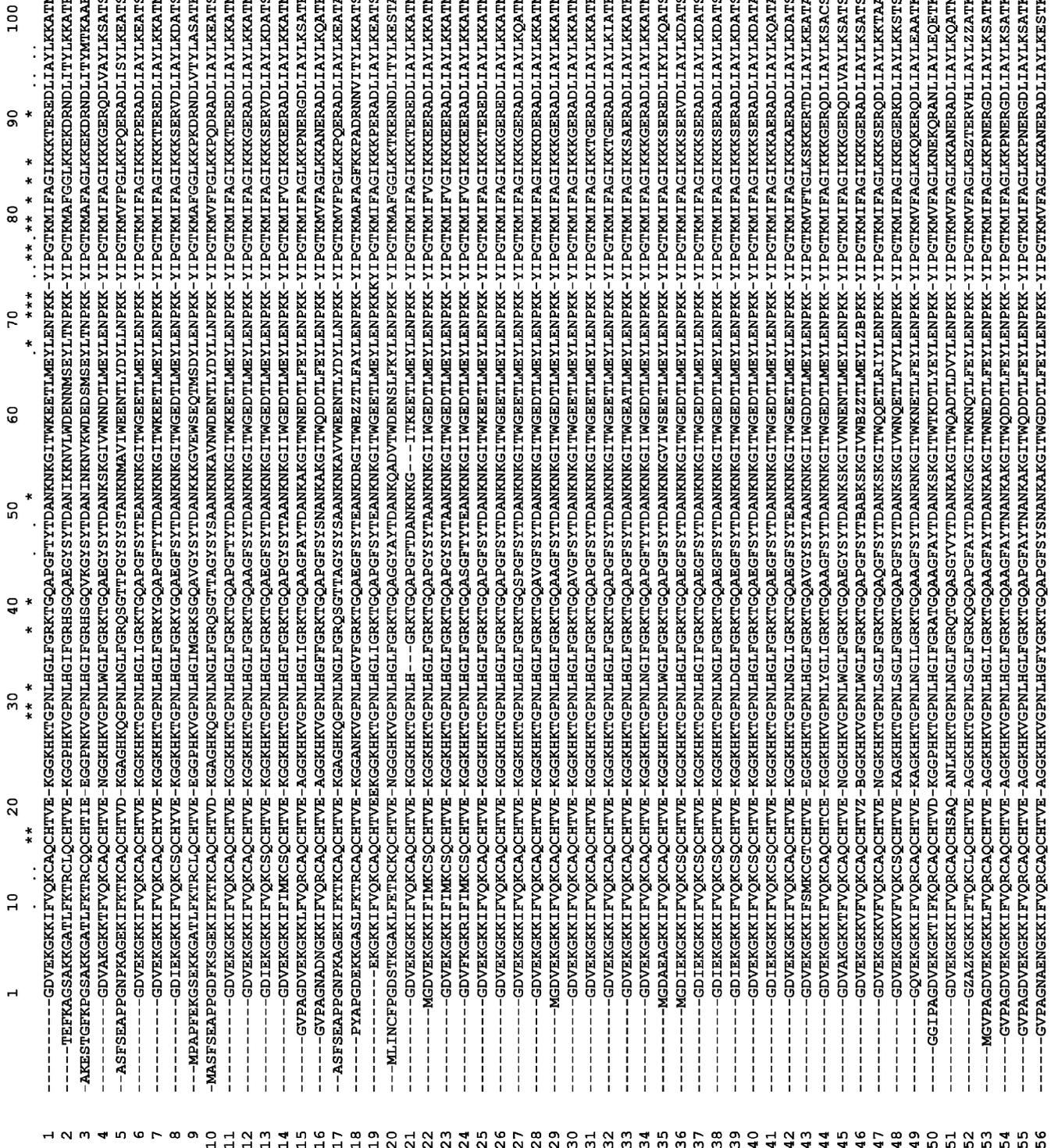
and then by the third helix, involving residues 61–69, and the fourth helix, involving residues 71–75. The loop following this helix contains the axial Met80 ligand. Finally, the C-terminal helix spans the residues from 90 to the end of the polypeptide chain. The structure used here as reference for modeling is the solution structure of horse heart cyt *c* [16]. This structure conforms very closely to the above schematic description, the most notable deviation being the fact that the N-terminal helix appears to be divided into two parts of roughly the same length (spanning residues 2–7 and 9–14, respectively).

### Amino acid sequences

Sequence databases currently (June 1999) contain 113 mitochondrial cyt *c* sequences from different eukaryotes. Each of these proteins contain 100–120 amino acids. Their alignment (Fig. 1) shows that the primary sequence is very highly conserved, with residue identity ranging between 45% and 100% across all eukaryotes. Figure 2 reports the number of amino acid types present at a given position in the 113 mitochondrial cyt *c* proteins. The N-terminal helix shows the highest sequence variability, followed by the C-terminal region. The most highly conserved region encompasses residues 67–87, which contains also the axial Met ligand (residue 80). The cyt *c* sequence is thus well preserved during evolution [7]. In particular, the region around the axial Met is one of the most strictly conserved, whereas that around the axial His is slightly more variable.

### The cyt *c* three-dimensional structure

The overall fold of cyt *c* is conserved well across different organisms. RMSD values between the modeled structures are in the range 0.5–1.5 Å. An example is shown in Fig. 3A, where we have reported a superposition of 10 different members of the cyt *c* superfamily, having a different degree of sequence identity. Figure 3 also reports a comparison between the experimental solution NMR structure of oxidized horse heart cyt *c* (Fig. 3B), which was taken as the starting structure for modeling, and that of *Crithidia oncopelti* (Fig. 3C). The cyt *c* protein from the latter organism has the largest difference in amino acid sequence with respect to the horse heart cyt *c* among the 113 sequences. The residue identity between these two sequences is 55.8% and the two structures have RMSD for the backbone atoms of 1.34 Å. As expected, larger deviations are found in the proximity of amino acid insertions or deletions. Experimental crystal or solution structures are available for some of the cyt *c* sequences investigated (*Albacore tuna*, PDB entry 5CYT; *Saccharomyces cerevisiae* iso-1, PDB entry 2YCC; *S. cerevisiae* iso-2, PDB entry 1YEA;



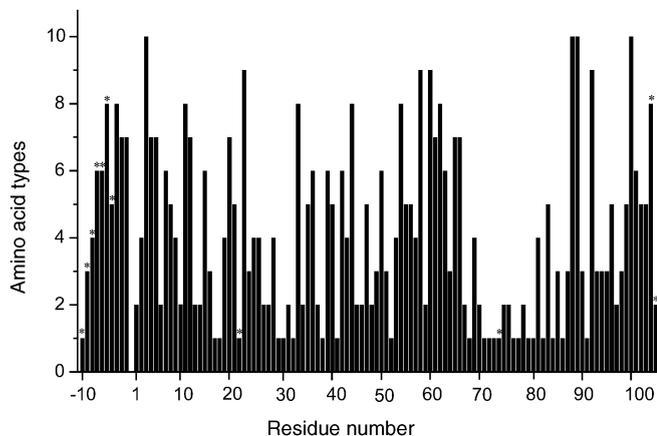
**Fig. 1** Alignment of the 113 mitochondrial cytochrome *c* protein sequences. Conserved residues are indicated by an asterisk; conservatively substituted residues are indicated by a dot. *B* and *Z* indicate that the experimental data are compatible with either aspartate or asparagine, and either glutamate or glutamine, respectively. In these cases, for the modeling it was assumed that the most common residue in all other sequences was at that position. The sequences are as follows: 1 horse heart; 2 *Saccharomyces cerevisiae* iso-1; 3 *S. cerevisiae* iso-2; 4 *Albacore tuna*; 5 rice; 6 American alligator; 7 *Equus grevyi*; 8 chicken; 9 *Pichia stipitis*; 10 *Fritillaria agrestis*; 11 *Equus hemionus*; 12 domestic

guinea pig; 13 American lobster; 14 chimpanzee; 15 house fly; 16 tobacco hornworm; 17 maize; 18 *Schizosaccharomyces pombe*; 19 China alligator; 20 *Emericella nidulans*; 21 *Ceratothorium simum*; 22 pigmy chimpanzee; 23 rhesus monkey; 24 spider monkey; 25 common zebra; 26 sheep; 27 hippopotamus; 28 *Oryctolagus cuniculus*; 29 Norway rat, somatic; 30 *Lama guanicoe*; 31 dog; 32 southern elephant seal; 33 Schreiber's long-fingered bat; 34 eastern gray kangaroo; 35 house mouse; 36 common turkey; 37 king penguin; 38 emu; 39 ostrich; 40 *Anas platyrhynchos*; 41 domestic pigeon; 42 snapping turtle; 43 eastern diamondback rattlesnake; 44 bullfrog; 45 skipjack tuna; 46 com-

57 -----GIPAGDPEKGGKIFVQKCAOCHTIE-SGGKHKVGNLYGVYGRKTKQAPGYSYTDANKGKIWNKEDTLFEYLENPKK-YIPGTKMVFAGLKKPOERADLIAYIEOASK  
 58 -----MGVPAGDVEKGGKIFVQKCAOCHTIE-AGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAKGIWNEDTLFEYLENPKK-YIPGTKMVFAGLKKPOERADLIAYIEOASK  
 59 -----GVPQDVEKGGKIFVQKCAOCHTIE-AGGKHKVGNLYGVYGRKTKQAPGYSYTDANKSGIWNEDTLFEYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLKESTK  
 60 -----PAPFKGSEKKGATLFTKRCLOCHTIE-KGGPHKVGPNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLANATK  
 61 -----PAPYKSGSKGANLFTKRCLOCHTIE-EGGPHKVGPNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 62 -----GDAERKGANLFTKRCLOCHTIE-KG-VNSTGCPALYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 63 -----AKGSEFPDAGKGANLFTKRCLOCHTIE-OGGANKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 64 -----MGFSAGDSKKGANLFTKRCLOCHTIE-EGGKNGKIPALHGFGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 65 -----GFEDDARKKARIFTKRCLOCHTIE-AGEPNKVGPNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 66 -----ASFDEAPPNGSAGEKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 67 -----ASFDEAPPNGSAGEKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 68 -----ASFZAPPBVKSGEIKFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 69 -----ASFZAPPBVKSGEIKFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 70 -----ASFQZAPPBBAKAGEKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 71 -----AFPSZAPPBZKAGQKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 72 -----ASFSEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 73 -----ASFSEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 74 -----ASFSEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 75 -----ATFSEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 76 -----MGSGDAENGGKIFVQKCAOCHTIE-VGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 77 -----MGDAEAGKIFVQKCAOCHTIE-VGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 78 -----GVPAGDVEKGGKIFVQKCAOCHTIE-AGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 79 -----MSFAEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 80 -----MPAPYKSGSKGANLFTKRCLOCHTIE-KGGPHKVGPNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 81 -----GDVEKGGKIFVQKCAOCHTIE-KGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 82 -----MSEKKGATLFTKRCLOCHTIE-KGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 83 -----MAKGGDSYSPGDAKAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 84 -----MPAPYKSGSKGANLFTKRCLOCHTIE-KGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 85 -----GKDSAPGDSAKGAKLFTKRCLOCHTIE-AGGPHKVGPNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 86 -----GDVEKGGKIFVQKCAOCHTIE-KGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 87 -----ASFAEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 88 -----GDVEKGGKIFVQKCAOCHTIE-KGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 89 -----GDVEKGGKIFVQKCAOCHTIE-KGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 90 -----GVPAGDVEKGGKIFVQKCAOCHTIE-AGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 91 -----GDVEKGGKIFVQKCAOCHTIE-KGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 92 -----MASFAEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 93 -----MASFAEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 94 -----MASFAEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 95 -----STFABAPPBBAKAGEKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 96 -----ASFZAPPBVKSGEIKFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 97 -----PKAREPPLPDDAAKAGEKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 98 -----ASFSEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 99 -----ATFSEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 100 -----ASFSEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 101 -----ASFSEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 102 -----ASFSEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 103 -----ASFZAPPBVKSGEIKFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 104 -----ASFQZAPPBBAKAGEKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 105 -----ASFSEAPPNGDAGKIFTKCAOCHTIVD-KGAGHKQGNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 106 -----GDVEKGGKIFVQKCAOCHTIE-KGGKHKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 107 -----PAPYKSGSKGANLFTKRCLOCHTIE-EGGPHKVGPNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 108 -----MSDIPAGDVEKGGKIFVQKCAOCHTIVD-STATKVGPNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 109 -----PKARAPLPPDAAKAGEKIFTKCAOCHTIVD-OGGANKVGNLYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 110 -----GDABERKGANLFTKRCLOCHTIE-KG-VNSTGCPALYGVYGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 111 -----PAPYKSGSKGANLFTKRCLOCHTIE-EGGPHKVGPNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 112 -----MPAPYKSGSKGANLFTKRCLOCHTIE-AGGPHKVGPNLHGFGRQSGAEGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK  
 113 -----MGFSEAGDSKKGANLFTKRCLOCHTIE-EGGKNGKIPALHGFGRKTKQAPGYSYTDANKKAVSEQMSDYLENPKK-YIPGTKMVFAGLKKPEERADLIAYLVRATK

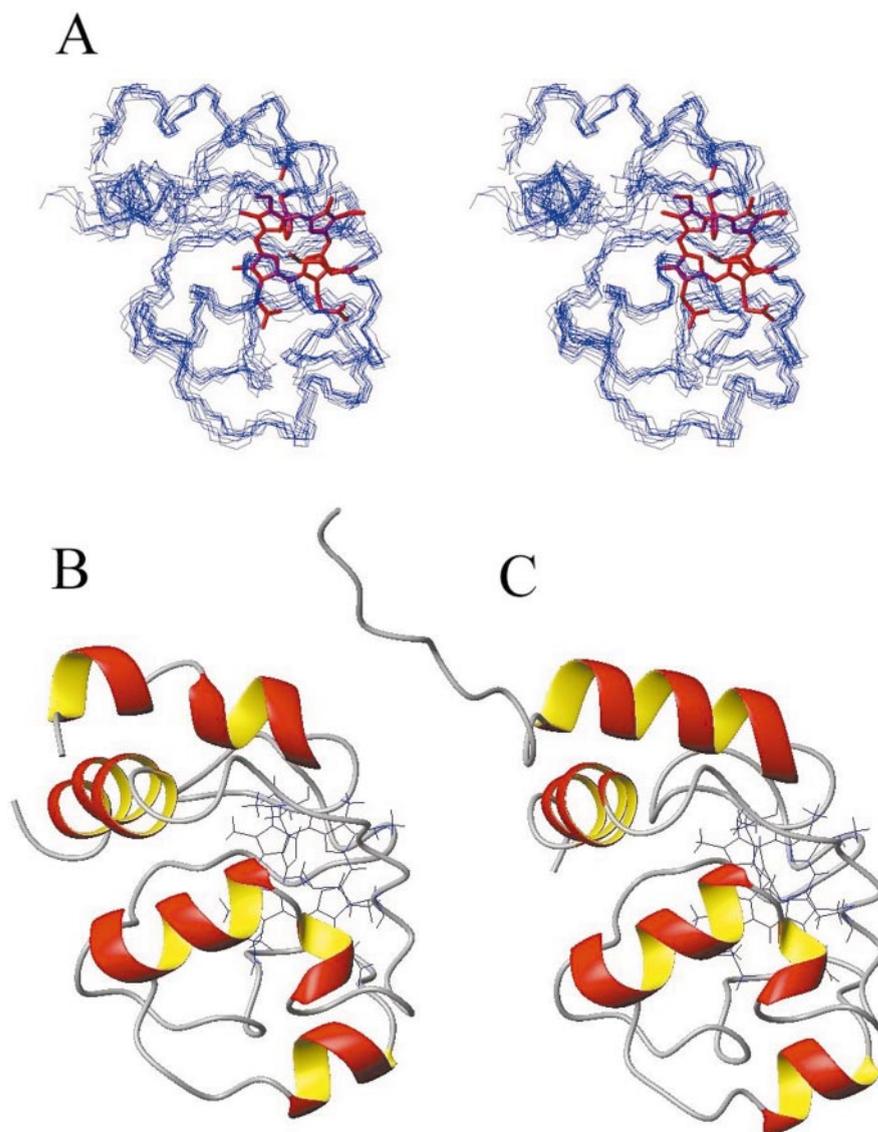
mon carp; 47 *Squalus sucklii*; 48 *Lampetra tridentata*; 49 European starfish; 50 common brandling worm; 51 Monsoos river-prawn; 52 brown garden snail; 53 Mediterranean fruit fly; 54 horn fly; 55 sheep blowfly; 56 *Samia cynthia*; 57 honeybee; 58 tobacco hornworm; 59 *Schistocerca gregaria*; 60 *Pichia anomala*; 61 *Debaryomyces hansenii*; 62 *Euglena viridis*; 63 *Thermomyces lanuginosus*; 64 *Neurospora crassa*; 65 *Ustilago spherogena*; 66 rape; 67 winter squash; 68 sesame; 69 castor bean; 70 China jute; 71 leek; 72 yellow lupine; 73 Niger seed; 74 common buckwheat; 75 maidenhair tree; 76 fruit fly; 77 Norway rat, testis-specific; 78 *Sarcophaga peregrina*; 79 *Chlamydomonas reinhardtii*; 80 *Deba-*

*rymomyces occidentalis*; 81 *Varanus varius*; 82 *Candida glabrata*; 83 *Emericella heterothallica*; 84 *Candida albicans*; 85 *Aspergillus niger*; 86 cow; 87 domestic pig; 88 California gray whale; 89 Arabian camel; 90 flesh fly; 91 man; 92 common sunflower; 93 Thale cress; 94 wheat; 95 hollow green seaweed; 96 hemp; 97 *Crithidia oncopelti*; 98 nasturtium; 99 spinach; 100 European elder; 101 mung bean; 102 parsnip; 103 love-in-a-mist; 104 sea island cotton; 105 donkey; 106 *Musca* sp.; 107 *Debaryomyces hansenii*; 108 *Caenorhabditis elegans*; 109 *Crithidia fasciculata*; 110 *Euglena gracilis*; 111 *Issatchenkia orientalis*; 112 *Kluyveromyces lactis*; 113 *Stellaria longipes*



**Fig. 2** Number of amino acid types present at a given position in the 113 mitochondrial cyt *c* protein sequences; a value of 1 corresponds to complete conservation. Positions marked with an *asterisk* correspond to a gap in the alignment in more than 50% of the sequences of the family. Positions 21a and 73a correspond to a gap in all the sequences but one

**Fig. 3** **A** Stereoview of the superposition of the backbone atoms of the horse heart cyt *c* structure with nine modeled structures, having 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, and 55% sequence identity to it, respectively (N-terminal extensions with respect to the horse heart protein are not shown). For clarity, only the heme moiety and the axial iron ligands of horse heart cyt *c* are shown. **B** Experimental solution structure of horse heart cyt *c* [16]. **C** Modeled structure of the cyt *c* sequence from *Crithidia oncopelti* (55.8% sequence identity to horse heart cyt *c*)



rice, PDB entry 1CCR). The backbone RMSD values between these and the corresponding modeled structures are around 1.3 Å, excluding the 21–25 loop which is disordered in the solution structure [16]. This value is comparable to the RMSD values found between experimental solution and crystal structures [42, 43] (for example, the backbone RMSD between the solution structure of horse heart cyt *c* and the corresponding X-ray structure is 1.10 Å) and therefore validates the modeling procedure adopted in this work.

Only bullfrog cyt *c* is known to have a disulfide bridge [44]. No effort has been made to identify possible disulfide bridges in the sequences investigated here.

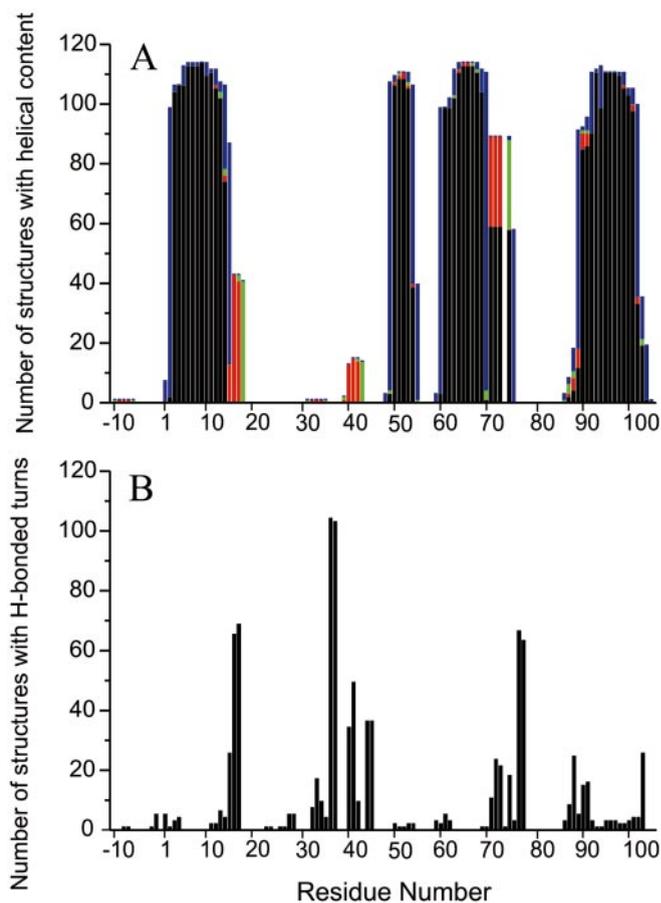
## Secondary structure

PROCHECK analysis of the modeled structures showed that secondary structure elements are well preserved, as expected. All structures comprise five helical elements, and these are found in corresponding positions in the sequence and in the three-dimensional models of the proteins. However, the length of individual helices varies from one model structure to the other. The positions of the various helices can be deduced from Fig. 4A, where the occurrence of helical structure in all 113 protein sequences is reported. The first 10 residues are almost never found in a helical conformation. This is in qualitative agreement with the observation that in the crystal structure of rice cyt *c* the eight-residue N-terminal extension does not form an extension of the first  $\alpha$ -helix, but shows only a short one-turn  $3_{10}$  helix from residue  $-5$  to residue  $-3$ . However, there is also an intrinsic difficulty in modeling the N-terminal extension of the polypeptide chain with respect to horse heart cyt *c*. All of the secondary structure elements described above are well conserved, but variability in the length of helix 3 and helix 4 leads to merging of these two helices into a single long helix in some models. From the comparison of Figs. 2 and 4, it is to be noted that elements of helical structure are very well maintained even in regions of the sequence where there is a quite high variability in the nature of the residues (see, for example, the first and last helices). Hydrogen-bonded turns are often found close to the N- or C-termini of the five helices, and around position 40 (Fig. 4B). Elements of  $\beta$  structure are also found in some of the modeled structures, but are much less conserved than the helices. Experimental cyt *c* structures nonetheless contain  $\beta$  bulges at the positions corresponding to those where elements of  $\beta$  structure occur more often in the present models.

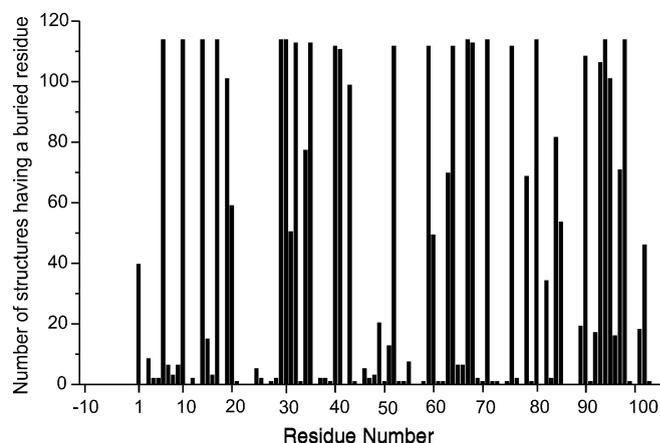
## The hydrophobic core

Buried hydrophobic residues are very highly conserved in all 113 cyt *c* sequences. Buried residues were defined as those having a solvent accessibility lower than  $25 \text{ \AA}^2$ . For most residues, this value indicates that 90% or more of the surface is buried. Although there is not a generally accepted threshold for the identification of buried residues, the value used here is close to others reported in the literature [45]. The number of structures for which a given position is buried is shown in Fig. 5. Notably, each residue position is either solvent accessible or buried in the majority of the structures (>80%); only 12 residues are buried in 30–70% of the structures.

Buried residues cluster in three different regions: (1) the first helix and the residues immediately following it, which are involved in hydrogen-bonded turns and interact with the heme (residues 5–20); (2) the



**Fig. 4A,B** Secondary structure elements in the 113 modeled structures of mitochondrial cyt *c* sequences, as deduced from PROCHECK analysis. Helical and irregularly helical elements (A; black helix; green distorted helix; red  $3_{10}$  helix; blue distorted  $3_{10}$  helix) and hydrogen-bonded turns (B) are shown



**Fig. 5** Number of structures having a buried residue at the reported residue position

region between residues 30 and 40, which is not consistently involved in any regular secondary structure; (3) residues 60–100. The last region can be further divided into three groups of residues, which are less

clearly separated than the two other regions as defined above. The three groups of this last region include the first part of helix 4 (residues 60–70), the residues between helix 4 and helix 5 (residues 75–83), and the N-terminal portion of helix 5 (residues 90–100). Sequence variations are lower for buried residues (the average number of different amino acids occurring at a given position is  $2.4 \pm 1.6$ ) than for the rest of the protein (where the analogous figure is  $4.7 \pm 2.7$ ). If we consider only the number of non-conservative substitutions occurring at each position, the above figures drop to  $1.6 \pm 0.9$  and  $2.8 \pm 1.4$ , respectively. Clearly, buried residues are better conserved than non-buried residues, as observed for other protein families [45].

Table 1 reports the contacts between buried residues which are most conserved among all 113 model structures. Nearly all the amino acids involved in the highly conserved contacts between buried amino acids described in Table 1 are conserved in 100 or more *cyt c* sequences. More variable residues are conservatively substituted in the majority of the protein sequences that we have analyzed.

#### The heme pocket

The amino acids coordinating the iron ion (His18 and Met80) are conserved in all 113 sequences. The CXX'C motif immediately preceding His18 is extremely highly conserved. The two cysteines form covalent bonds to the porphyrin ring. Cys14 is replaced by Ala in four sequences (sequences 97, 109, 110, and 113 in Fig. 1), while Cys17 is conserved in all *cyt c* sequences currently available. Residue 15 is an Ala in approximately 70% of the sequences; the second most frequent amino acid found at this position is Ser (13%), a more polar residue that occupies nonetheless a volume comparable to Ala. In nearly 11% of the sequences, however, residue 25 is a Leu, a hydrophobic residue with a much larger side-chain. Gly (2%), Asn (1%), and Lys (2%) are also found at position 15 in one or two sequences. Residue 16 is much better conserved, being Gln in 98% of the sequences. As discussed in the next section, this residue is also thought to play a role in *cyt c* interactions during electron transfer. The other amino acids found at this position are Thr (1%) and Glu (1%). The solvent-exposed residue preceding Cys14 is invariably Lys (72%) or Arg (28%). The residue following the first axial ligand, His 18, is Thr in 93% of all sequences. It is substituted either by Ser (3%), Tyr (2%), or Val (2%). Thus, in nearly all cases, the buried residue 19 has a hydroxyl group.

Residues surrounding Met80 (74–80, 82, and 84) are conserved in at least 95% of the structures. The heme propionate 7 is hydrogen bonded to the side chains of Trp59 and Asn52 in nearly 90% of the structural models we have generated, and these two resi-

**Table 1** Contacts between buried residues conserved in 113 mitochondrial *cyt c* model structures

First residue	Second residue	Occurrence (no. of structures)
6	10	103
6	93	92
6	94	96
10	14	103
10	19	78
10	32	92
10	94	96
10	98	79
14	17	100
17	29	100
19	30	84
19	32	90
29	30	102
30	32	102
30	43	83
32	35	96
32	98	85
35	59	98
35	64	96
35	98	92
40	41	92
40	52	97
40	59	96
41	52	94
52	59	96
52	75	96
59	64	97
59	67	96
59	75	96
64	67	97
64	68	100
64	95	90
67	68	97
67	71	99
67	75	96
67	80	97
68	71	81
68	80	87
68	94	97
68	95	90
71	75	96
71	80	95

dues are conserved in 99% and 98% of the structures, respectively. Propionate 6 is hydrogen bonded to the backbone amide protons of residue 49 and Lys79 in 69% and 86% of the structures, respectively. In addition, a hydrogen bond between propionate 6 and the hydroxyl proton of Thr78 is observed in 70% of the structures. Residue 49 is either a Thr (70%) or a Ser (29%), except for a single sequence where this residue

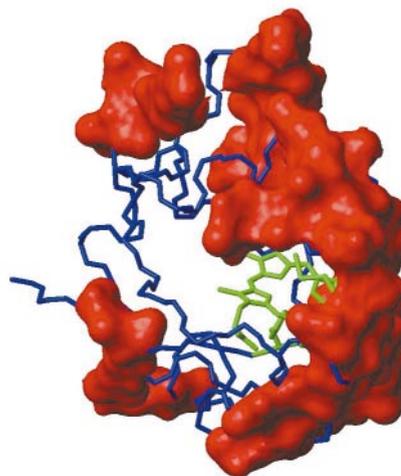
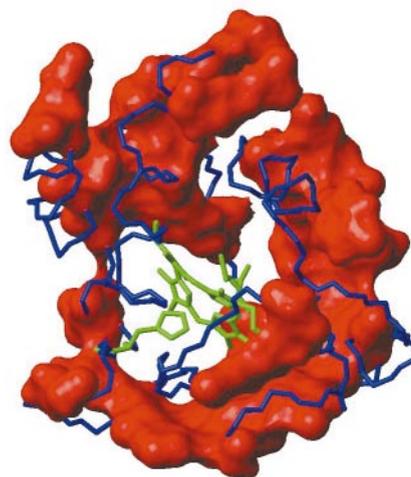
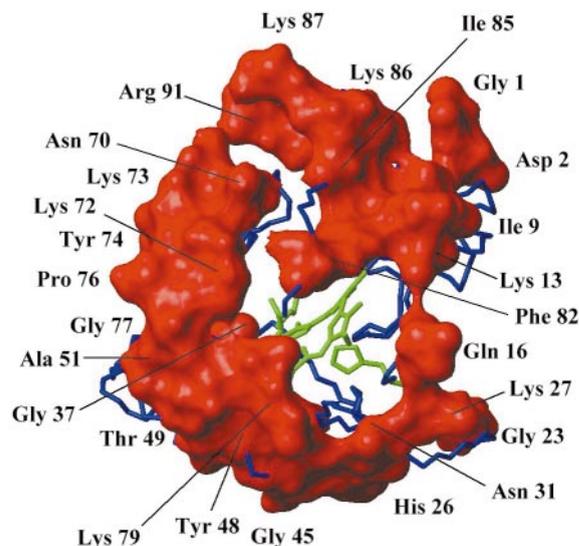
is Ala. Residue 79 is always Lys, whereas residue 78 is Thr in 99% of the structures.

### Protein-protein interaction sites

The recognition site for other proteins that interact with cyt *c* (cytochrome *c* oxidase, cytochrome *c* reductase, cytochrome *c* peroxidase) has been probed by a number of experimental approaches, including X-ray crystallography [46], chemical modification of exposed lysines [47–49], cross-linking experiments coupled to site-directed mutagenesis [50–52], and NMR spectroscopy [53–57]. Residues 9, 13, 16, 70, 72, 73, 81–83, 86, and 87 were found to be important for protein-protein recognition by one or more methods of analysis. It appears that the interaction site for different partners is the same (see [8, 58] and references therein). However, several studies utilized inorganic compounds or proteins that were not a true physiological partner of cyt *c* during electron transfer; the proposed interactions are therefore not necessarily physiologically significant. If an interaction is important, however, there will be genetic pressure for this residue to be conserved or conservatively substituted (e.g. Lys/Arg). We therefore examined the pattern of conservation and variability of exposed surface residues to determine patches of conservation that would highlight functionally important regions of cyt *c* proteins. The van der Waals surfaces of conserved and conservatively substituted exposed residues are shown in Fig. 6.

Residue 9 is conservatively substituted and solvent exposed in nearly all structures: only Ile (74%), Leu (21%), Val (3%), or Thr (2%) are found in this position. This amino acid had been proposed to take part in hydrophobic interactions with cytochrome *c* peroxidase (ccp); the conservative nature of substitutions at this position confirms its likely functional role. Residues 13 (either a Lys or an Arg) and 16 (nearly always Gln) are, respectively, the residues preceding the first and second heme-bound cysteines. They are solvent exposed in essentially all structures and conserved. Being in close proximity to the heme, these two residues could play a role in electron transfer, e.g. by altering the conformation of the heme pocket upon binding other members of the respiratory chain. Asn70 is exposed and conserved (98%). Lys72 and Lys73 are strictly conserved and generally solvent exposed in our structural models. Residue 81 is nearly

always conservatively substituted, being Ala (20%), Val (32%), Ile (46%), or Ser (2%). Residue 82 is always Phe, whereas residue 83 is somewhat variable, being Ala (59%), Val (4%), Pro (21%), Gly (15%),



**Fig. 6** Three views, rotated by 180° (*middle*) and 90° (*bottom*), of the van der Waals surface of the most conserved exposed residues in the cyt *c* superfamily, together with the backbone heavy atoms, are shown. The heme moiety and the iron axial ligands are displayed in *green*. The *top* orientation highlights the region of intermolecular contact, as deduced from mutagenesis/side-chain modification studies (see text). The amino acids reported in the labels are the ones most frequently found at each position

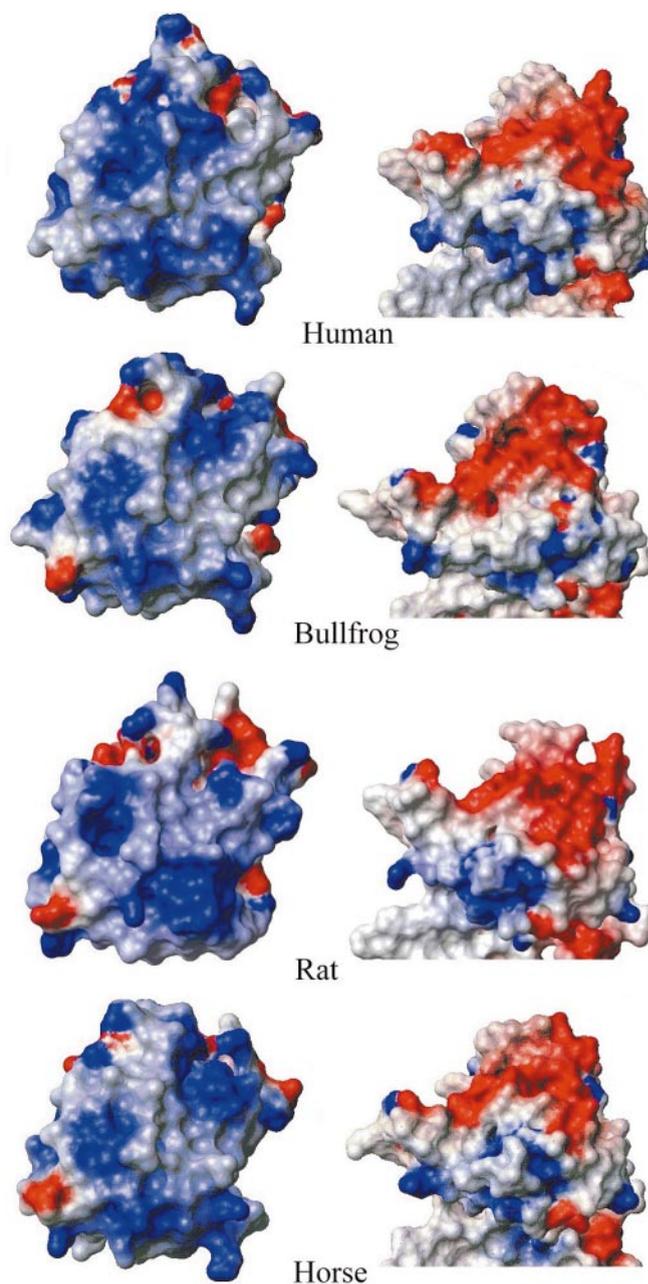
or Thr (1%). Residues 81–83 are involved in hydrophobic contacts in the crystallographic structure of the ccp/cyt *c* complex [46]. On the basis of absorption, CD, and MCD spectroscopic studies on the interaction of cyt *c* with CCO, it has been suggested that this region is also involved in protein-protein interactions in the cyt *c*-CCO complex [59]. Such interactions are possible following a structural rearrangement subsequent to partner binding (induced fit) [59]. Also note that Phe82 occupies a central position in the pattern of conserved exposed residues shown in the top panel of Fig. 6. Finally, residues 86 and 87 are very well conserved in all sequences examined here and are consistently solvent exposed.

Lys25 and Lys27 have also been proposed to play a role in the interactions of cyt *c* [8, 49], but were not found to take part in protein-protein contacts in the X-ray structure of the ccp/cyt *c* complex [46]. Both residues are consistently solvent exposed in our models, but only residue 27 is highly conserved (Lys 97% of the time). Position 25 is variable, being either Lys (55%), Gly (25%), Pro (13%), or Ala (6%). It is therefore unlikely that residue 25 plays an important role in functionally relevant protein-protein interactions of cyt *c*.

### Surface potentials

Electrostatic interactions have been proposed to play a crucial role in the interaction of cyt *c* with its biological partners in the respiratory chain. The electrostatic potential at the protein surface is determined primarily by the identity of solvent-exposed amino acids. Despite the conservation of several solvent-exposed residues described in the previous section, many other exposed amino acids are not conserved. It is therefore likely that the electrostatic features of the surface of cyt *c* are different between proteins from different organisms. If changes occur, however, these may have to be compensated by co-evolutionary changes on the surface of proteins that interact with cyt *c*. In order to investigate this possibility, we examined whether such differences were correlated with complementary variations in the surface potentials of known cyt *c* partners by modeling the structures of subunit II of CCO. A soluble fragment of subunit II of CCO containing the so-called copper A site recognizes cyt *c* in solution [60], and the rate and efficiency of the electron transfer reaction between the two are comparable to those observed with the complete protein enzyme.

The structures of CCO from organisms for which the cyt *c* structure had been already derived were modeled as described in the Methods section. Figure 7 shows the horse heart, rat, bullfrog, and human cyt *c* and CCO represented with a surface electrostatic potential and oriented in such a way that the contact region of the proteins is observable. The regions of contact on the two proteins have been identified



**Fig. 7** Surface potentials of cyt *c* (left) and of subunit II of cyt *c* oxidase (right) from man, bullfrog, rat, and horse heart (from top to bottom). Surface electrostatic potentials were generated using MOLMOL [31]. The orientation of the two proteins highlights the regions of intermolecular contact, as deduced from mutagenesis/side-chain modification studies (see text)

through several different techniques, and described in a number of studies [48, 59–62]. Some variation in surface potential can be observed between proteins from different species. However, the interaction between the two proteins is most likely to happen with essentially the same reciprocal orientation in all species, since a change in the reciprocal orientation would greatly affect the electron transfer rate. It can be clearly seen from Fig. 7 that the gross features of the

surface potential and shape of the proteins in and around the recognition site are maintained in the four sets of structural models.

---

## Discussion

### The hydrophobic core and protein folding

Protein structures are determined by the identity of relatively few hydrophobic residues and their respective interactions [63]. In this study, we have systematically examined the nature and character of interactions conserved in eukaryotic cyt *c* proteins by modeling the structures of 113 eukaryotic cyt *c* sequences. The comparison of the 113 model structures shows that the backbone of all model structures can be superimposed within small RMS deviations. In a previous comparison of tuna cyt *c* and *Pseudomonas aeruginosa* cytochrome *c*<sub>551</sub> structures [12], different arrangement of the helices were observed after structures had been superposed by best fitting the heme coordinates. In contrast, no large differences are observed in the present set of models. The angles between corresponding helical axes are always smaller than 10° and the largest relative displacements between helix axes are of only 1.5 Å. These values are comparable to those previously reported for the comparison of tuna cyt *c* and *Rhodospirillum rubrum* cytochrome *c*<sub>2</sub>, and smaller than reported in the comparison of cyt *c* and cytochrome *c*<sub>551</sub> [12].

Conserved contacts between consistently buried residues are summarized in Table 1. Many of these contacts involve residues within a single secondary structure element, and their role is most likely to contribute to the local stability of the structure. This is likely to ensure a proper conformation and orientation of the backbone and side chain of the amino acids involved in these interactions and of those immediately adjacent. Among the contacts listed in Table 1 there are also long-range interactions between different regions of the sequence. These include interactions between helix 1 and helix 5; helix 1 and the 29–32 region; helix 2 and helix 4; helix 3 and helix 4; and the 40–43 region with helix 2. These long-range contacts constrain and determine the structure of cyt *c*. The pattern of helix-helix contacts pointed out by Chothia and Lesk [12] appears to be conserved across all mitochondrial cyt *c*. The packing of helices 1 and 2 against helix 5 is a conserved structural feature in different *c*-type cytochromes. Among the helix 1-helix 5 contacts, interactions between residue 6 and residues 93 and 97 and those of residue 97 with amino acids 7 and 10 were maintained in both tuna cyt *c* and *Ps. aeruginosa* cytochrome *c*<sub>551</sub>. Note that of the residues involved in these interactions, only 6 and 10 are buried in all structures. Residue 97 is buried in 62% of structures, whereas 7 and 93 are solvent exposed. Table 1 lists the interactions between consistently

buried residues, and thus does not report the above-mentioned ones. Examination of our models indicates that all of the above contacts are maintained in at least 90% of the model structures of cyt *c* examined here. Residue 7 is variable; Ala, Glu, and Lys are all found with high frequency (20%, 55%, and 22%, respectively) at this position. However, positions 6, 10, 93, and 97 are essentially invariant or conservatively substituted. The importance of the chemical nature of the side chains at positions 6, 10, 94, and 97 for biological functionality of cyt *c* was probed by mutagenesis [9, 64–67]. It appears that more, different, amino acids can be accommodated in the above positions and still yield a functional (at least partly) cyt *c* with respect to what is suggested by the frequency of amino acid replacements observed in the 113 naturally occurring cyt *c* sequences [64, 65]. Only loose restraints for amino acid substitutions are observed (e.g. no functional cyt *c* has a positively charged amino acid in position 6) [64, 65, 67]. This suggests that even the interactions which are key for cyt *c* folding have a certain degree of flexibility. Measurements of the thermodynamics of unfolding of several cyt *c* mutants suggest that nature selects protein sequences which yield structures with a similar free energy of unfolding [66]. Mutant sequences yielding a functional cyt *c* characterized by a much higher or lower free energy of unfolding can be realized in the laboratory, but they do not occur naturally [66]. A second set of long-range interactions highlighted by Chothia and Lesk [12] included the helix 3-helix 5 contacts common to both cyt *c* and cytochrome *c*<sub>551</sub>. These interactions are well conserved in the present set of model structures, with the exception of the contact between residues 65 and 92, which is present in less than 50% of the models generated in this study. The lack of conservation of this interaction could be a reflection of the variability observed at these positions. Residue 65 is mostly Met (42%) or Phe/Tyr (48%), whereas residue 92 can be Ala/Val (61%), Asn/Gln (21%), or Gly (7%).

Leu94 plays an important role in cyt *c* folding [68]. This residue is conserved in 94% of the sequences examined here, and is replaced only conservatively by Val (3%) or Ile (3%). The substitution of Leu94 with either Val or Ile does not seem to affect the energetics and kinetics of horse heart cyt *c* folding [68]. As mentioned above, mutant cyt *c*'s where residue 94 is substituted by other aliphatic residues (Cys, Thr, Ala) are functional, but do not occur naturally [65]. His33 had been proposed to coordinate the iron ion under denaturing conditions above pH 6 [69]. However, NMR data for guanidinium-denatured cyt *c* at pH 8.5 are inconsistent with this hypothesis (unpublished data). His33 is not strongly conserved (His 62%, Asn 23%), suggesting that this residue is not likely to play an essential role in binding iron early during protein folding. Recent data for a mutant cyt *c* where all histidines had been removed show that ligation of the

N-terminal amino group to the iron ion occurs upon denaturation of cyt *c* under the same conditions where His ligation would be expected [70]. Finally, measurements of the protection of amide protons against H/D exchange in refolding experiments have identified regions of the protein that fold first [71]. These include the C-terminal part of the first  $\alpha$ -helix and the residues immediately following this helix, residues 65–70, and residues 90–100. On the other hand, in the so-called A (i.e. molten globule) state of the protein (low pH, high ionic strength), there are native-like contacts between the terminal helices, while the 60's helix is not completely formed [72–74]. The above-mentioned residues are consistently buried from the solvent in all 113 models of eukaryotic cyt *c* proteins studied here, suggesting that hydrophobicity of these regions may be the driving force to initiate cyt *c* folding.

Backbone amide hydrogen exchange properties [17, 19–21, 23] and backbone mobility [22] of cyt *c* have been intensively investigated. It appears that the patterns of protection from hydrogen exchange correlate quite well with the pattern of solvent accessibility shown in Fig. 5. However, other factors besides solvent accessibility play an important role in determining the rates of solvent exchange [21]. By comparing the above properties for the oxidized and the reduced forms of cyt *c*, it has been shown that in the oxidized species the exchangeability of the protein backbone is, globally, somewhat increased and that the mobility of the protein backbone is higher in the oxidized than in the reduced form [17, 19, 23]. There is no clear-cut correlation between the occurrence of differences in the amide hydrogen exchange regime or in the backbone mobility upon change of oxidation state and other structural or evolutionary parameters, such as occurrence of secondary structure elements, residue burial, and amino acid variability in the cyt *c* superfamily.

### The heme environment

The residues between the two heme-bound cysteines (Cys14 and Cys17) in *c*-type cytochromes are among the most important amino acids in determining the extent and the type of distortion of the porphyrin macrocycle from planarity [75]. As outlined in the Results section, these two Cys are nearly always separated by two residues. These intervening residues are conserved in 75% of all sequences in the database, the consensus sequence being CAQC. This observation suggests that heme distortions are similar in the large majority of eukaryotic cyt *c* proteins. Replacement of either Cys or of the axial His in *S. cerevisiae* cyt *c* hampers the growth of this organism [9]. This defect has been attributed to the fact that heme is not inserted into the polypeptide chain [9].

Another important property affecting the iron ion coordination sphere is the orientation of the imidazole plane of the axial ligand His18. The orientation is essentially the same in all the modeled structures. It is interesting to observe that the H $\delta$ 1 proton of the imidazole ring is hydrogen bonded to the carbonyl oxygen of Pro30 in ca. 90% of the structures. Pro30 is absolutely conserved, and the region around this residue is quite well conserved, although not involved in any defined secondary structure elements (see Figs. 1, 2, and 4). These observations highlight the importance of the Pro30-His18 interaction in determining the orientation of the imidazole ring of His18. Notably, *S. cerevisiae* mutant strains in which Pro30 has been replaced by Leu grow to a level much lower than the non-mutated control strain [9]. The present data do not allow us to assess whether this interaction has just a structural role or it is also directly relevant to the functional properties of the iron ion. There is some evidence that the strength of the iron-histidine interaction may modulate the iron redox potential. The latter interaction is tuned by the strength of the hydrogen bond involving the H $\delta$ 1 proton, which regulates the imidazole/imidazolate character of the axial ligand [76].

The side chain of the residue following His18 contains an OH group, which, in the majority of the 113 model structures, is hydrogen bonded to the carbonyl oxygen of residue 22 or 24. Since residue 19 is buried in most model structures, the hydroxyl group is not accessible to solvent. Buried polar residues are more conserved than buried non-polar residues, probably because their side chains form structurally important hydrogen bonds [45]. The low accessibility to solvent implies that these hydrogen bond interactions cannot be replaced by interaction with water molecules. The hydrogen bond between the hydroxyl of residue 19 and the carbonyls of residues 22 or 24 is likely to play an important role in defining the orientation of the polypeptide chain at and around the heme-ligand amino acids.

The conformation of the heme propionates, and particularly of propionate 7, is redox-state dependent [16, 17, 77]. This finding is important given its proposed role in the electron transfer pathway [78]. In the experimental structures of the oxidized form of *S. cerevisiae* iso-1 cyt *c*, propionate 7 is hydrogen bonded to the backbone amide proton of Gly41 and to the side chains of Asn52 and Trp59 [17, 35]. The hydrogen bond network of propionate 7 in horse heart cyt *c* could not be firmly established in the solution structure, owing to the disorder of the carboxylate moiety of the propionate [16]. The interactions of propionate 7 with the side chains of Asn52 and Trp59 are maintained in the models of the 113 eukaryotic cyt *c* structures, but the interaction with Gly41 is not. Propionate 6 is disordered in solution [16, 17, 77], but in the present set of models it forms a well-conserved hydrogen bond with the backbone amide proton of

Lys79. Cyt *c* proteins undergo a conformational rearrangement above pH 9, where the axial Met80 is replaced by another strong-field ligand to maintain a six-coordinate low-spin heme. It is believed that the sixth ligand is provided by the N $\zeta$  atom of a Lys side chain [79–82]. This is also in agreement with electrochemical data, which show that the reduction potential of the alkaline form is much lower than that of the native form, in agreement with that expected for the replacement of a Met ligand with a Lys ligand [83, 84]. <sup>1</sup>H NMR data on wild-type and mutated *S. cerevisiae* cyt *c* indicate that two conformers of cyt *c* exist under alkaline conditions [81, 82, 85]. Interestingly, at neutral pH and temperatures above 42 °C, cyt *c* adopts a new form remarkably similar to the alkaline form [80]. As in the alkaline form, the axial Met detaches from the iron ion at high temperature as well, although the iron ion remains six-coordinated [86]. The binding of a lysine to the iron at high pH or high temperature is also supported by the similarity of spectroscopic data between the corresponding forms of cyt *c* and the complex of cyt *c* with ammonia (which replaces Met80 as the sixth iron ligand) [87]. Lys72 and Lys79 are the two most plausible candidates for replacing the axial Met in the high-pH and high-temperature forms [85]. The present analysis indicates that both residues are invariant in eukaryotic cyt *c*, supporting their important structural and functional role.

#### Interaction of cyt *c* with other electron transfer proteins

The recognition sites on the surface of cyt *c* for other proteins in the respiratory chain have been found to be essentially the same between different physiological and non-physiological interacting proteins [2, 8, 46, 48, 58]. However, the relative importance of different amino acids in tuning these interactions was found to depend on the specific interacting protein. Exposed residues are significantly more variable than buried residues, which are constrained by the requirements to provide a stable folded protein structure. Therefore, a high degree of conservation of exposed amino acids implies an important functional role in intermolecular recognition and electron transfer. Figure 6 shows that the majority of conserved (or conservatively substituted) exposed residues is located in the region of the protein which is involved in intermolecular interactions. Moreover, all residues found to be important for the interaction of cyt *c* with other proteins on the basis of X-ray crystallography and chemical modification experiments were conserved (or conservatively substituted) in the 113 sequences examined in the present work, with the exception of residues 25 and 83. A role in protein-protein interaction for residue 25 was proposed on the basis on lysine chemical modification experiments, but this interaction was not

observed in the X-ray structure of the ccp/cyt *c* complex [46].

Residue 13, immediately preceding the first heme-bound cysteine, is either Arg or Lys in every cyt *c* sequence in the database. This conservatively substituted basic residue may be involved in tuning the iron properties upon interaction with cyt *c* partners, for example by modulating the protein conformation near the heme. A role for this residue in protein-protein interactions had been proposed from the crystal structure of the ccp/cyt *c* complex [46], and in experiments where lysine side chains were chemically modified [8, 47]. Residue 79, which precedes the axial Met ligand, is also conserved in all 113 eukaryotic cyt *c* sequences. As suggested for Lys/Arg13, this residue may constitute part of a communication channel modulating the heme environment in response to protein binding. However, there is no compelling experimental evidence at this time for the involvement of Lys79 in protein recognition or electron transfer [8].

Protein-protein interactions mediated by cyt *c* are likely to be transient and electrostatic in nature [2]. Therefore, surface electrostatic potential may be more important than the identity of individual residues in modulating and determining molecular orientation in protein interactions and electron transfer. In order to test whether the complementarity between surface potentials of cyt *c* and its partners is maintained in different species, we have modeled the structure of subunit II of CCO as well. The surface potentials of cyt *c* and CCO for the horse heart, rat, bullfrog, and human proteins are shown in Fig. 7. The general features of the cyt *c* and CCO surfaces are well maintained in all the species. In particular, the recognition site on cyt *c* is always positively charged, whereas that on CCO is negatively charged. Despite variations in the shape of the surfaces and in the details of the electrostatic potential at the surface of the two, the overall patterns are strikingly similar among different species. This observation lends support to the suggestion that the interaction between the two proteins occurs with essentially the same reciprocal orientation in all species. The comparison of the crystal structures of the complexes of *S. cerevisiae* cyt *c* peroxidase with *S. cerevisiae* cyt *c* and with horse heart cyt *c* shows that the two cyt *c*'s interact slightly differently with the same peroxidase [46]. However, the importance of this finding with respect to *in vivo* recognition is unclear, as the two crystals were obtained in somewhat different conditions [46], and the interaction between *S. cerevisiae* cyt *c* peroxidase and horse heart cyt *c* is non-physiological. Indeed, since any change in orientation would greatly affect electron transfer rates and efficiency, the optimal relative arrangement of the chromophores of the two proteins should be conserved regardless of the specific protein environment. The similarities in surface electrostatic potentials and shape in cyt *c* and CCO from different organisms highlighted in Fig. 7 support this concept.

## Conclusions

We have aligned the 113 existing sequences of eukaryotic mitochondrial cytochrome *c* proteins, and modeled their three-dimensional structure. By comparing these model structures, we have been able to single out interactions between buried residues which are responsible for the overall structure of cyt *c* proteins. Conserved surface-exposed residues identify instead sites that are important for the cyt *c* function in electron transfer. The analysis of putative protein-protein interaction sites and of the surface electrostatic potential suggests that residue-specific interactions within a framework of electrostatic complementarity are important for the function of cyt *c* in electron transfer.

**Acknowledgements** We thank Andrea Barzanti for performing the modeling of the structures and for some preliminary analyses. Financial support from the EU (contract ERBFMRX CT980218) and the Italian CNR (contract CT CNR 97.01027.49) is gratefully acknowledged.

## References

- Pascher T, Chesick JP, Winkler JR, Gray HB (1996) *Science* 271:1558–1560
- Scott RA, Mauk AG (eds) (1996) *Cytochrome c. A multidisciplinary approach*. University Science Books, Sausalito, Calif
- Kluck RM, Bossy-Wetzel E, Green DR, Newmeyer DD (1997) *Science* 275:1132–1136
- Yang C-H, Azad HR, Cooksey DA (1996) *Proc Natl Acad Sci USA* 93:7315–7320
- Yeoman KH, Delgado MJ, Wexler M, Downie JA, Johnston AWB (1997) *Microbiology* 143:127–134
- Biel SW, Biel AJ (1990) *J Bacteriol* 172:1321–1326
- Fitch WM (1976) *J Mol Evol* 8:13–40
- Mathews FS (1985) *Prog Biophys Mol Biol* 45:1–56
- Hampsey DM, Das G, Sherman F (1988) *FEBS Lett* 231:275–283
- Moore GR, Pettigrew GW (1990) *Cytochromes c: evolutionary, structural and physicochemical aspects*. Springer, Berlin Heidelberg New York
- Brayer GD, Murphy MEP (1996) In: Scott RA, Mauk AG (eds) *Cytochrome c. A multidisciplinary approach*. University Science Books, Sausalito, Calif, pp 103–166
- Chothia C, Lesk AM (1985) *J Mol Biol* 182:151–158
- Qi PX, Di Stefano DL, Wand AJ (1994) *Biochemistry* 33:6408–6417
- Qi PX, Beckman RA, Wand AJ (1996) *Biochemistry* 35:12275–12286
- Baistrocchi P, Banci L, Bertini I, Turano P, Bren KL, Gray HB (1996) *Biochemistry* 35:13788–13796
- Banci L, Bertini I, Gray HB, Luchinat C, Reddig T, Rosato A, Turano P (1997) *Biochemistry* 36:9867–9877
- Banci L, Bertini I, Bren KL, Gray HB, Sompornpisut P, Turano P (1997) *Biochemistry* 36:8992–9001
- Banci L, Bertini I, Huber JG, Spyroulias GA, Turano P (1999) *JBIC* 4:21–31
- Marmorino JL, Auld DS, Betz DS, Doyle DF, Young GB, Pielak GJ (1993) *Protein Sci* 2:1966–1974
- Bai YW, Sosnick TR, Mayne L, Englander SW (1995) *Science* 269:192–197
- Milne JS, Mayne L, Roder H, Wand AJ, Englander SW (1998) *Protein Sci* 7:739–745
- Baxter SM, Fetrow JS (1999) *Biochemistry* 38:4480–4492
- Baxter SM, Fetrow JS (1999) *Biochemistry* 38:4493–4503
- Altschul FS, Madden TL, Schaeffer A, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) *Nucleic Acids Res* 25:3389–3402
- Thompson JD, Higgins DG, Gibson TJ (1994) *Nucleic Acids Res* 22:4673–4680
- Sali A, Blundell TL (1993) *J Mol Biol* 234:779–815
- Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, Ferguson DM, Seibel GL, Singh UC, Weiner PK, Kollman PA (1997) *AMBER 5.0*. University of California, San Francisco
- Banci L, Gori Savellini G, Turano P (1997) *Eur J Biochem* 249:716–723
- Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S (1996) *Science* 272:1136–1144
- Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) *J Appl Crystallogr* 26:283–291
- Koradi R, Billeter M, Wüthrich K (1996) *J Mol Graphics* 14:51–55
- Dickerson RE, Takano T, Eisenberg D, Kallai OB, Samson L, Cooper A, Margoliash E (1971) *J Biol Chem* 246:1511–1535
- Ashida T, Yamane T, Tsukihara T, Kakudo M (1973) *J Biochem (Tokyo)* 73:463
- Takano T, Dickerson RE (1981) *J Mol Biol* 153:79–94
- Takano T, Dickerson RE (1981) *J Mol Biol* 153:95–155
- Ochi H, Hata Y, Tanaka N, Kakudo M, Sakuri T, Achara S, Morita Y (1983) *J Mol Biol* 166:407–418
- Louie GV, Hutcheon WLB, Brayer GD (1988) *J Mol Biol* 199:295–314
- Bushnell GW, Louie GV, Brayer GD (1990) *J Mol Biol* 214:585–595
- Louie GV, Brayer GD (1990) *J Mol Biol* 214:527–555
- Berghuis AM, Brayer GD (1992) *J Mol Biol* 223:959–976
- Sanishvili R, Volz KW, Westbrook EM, Margoliash E (1995) *Structure* 3:707–716
- Bertini I, Luchinat C, Rosato A (1996) *Prog Biophys Mol Biol* 66:43–80
- Banci L, Bertini I, Luchinat C, Turano P (1999) In: Kadish KM, Smith KM, Guillard R (eds) *The porphyrin handbook*. Academic Press, Burlington, Mass (in press)
- Brems DN, Cass R, Stellwagen E (1982) *Biochemistry* 21:1488–1493
- Rodionov MA, Blundell TL (1998) *Proteins Struct Funct Genet* 33:358–366
- Pelletier H, Kraut J (1992) *Science* 258:1748–1755
- Butler J, Davies DM, Sykes AG (1980) *J Am Chem Soc* 103:469–471
- Rieder R, Bosshard HR (1980) *J Biol Chem* 255:4732–4739
- Armstrong GD, Chapman SK, Sisley MJ, Sykes AG, Aitken A, Osheroff N, Margoliash E (1986) *Biochemistry* 25:6947–6951
- Moench S, Satterlee JD, Erman JE (1987) *Biochemistry* 26:3821–3826
- Pappa HS, Poulos TL (1995) *Biochemistry* 34:6573–6580
- Pappa HS, Tajbaksh S, Saunders AJ, Pielak GJ, Poulos TL (1996) *Biochemistry* 35:4837–4845
- Yi Q, Erman JE, Satterlee JD (1994) *Biochemistry* 33:12032–12041
- Guiles RD, Sarma S, DiGate RJ, Banville D, Basus VJ, Kuntz ID, Waskell L (1996) *Nat Struct Biol* 3:333–339
- Sukits SF, Erman JE, Satterlee JD (1997) *Biochemistry* 36:5251–5259
- Ubbink M, Bendall DS (1997) *Biochemistry* 36:6326–6335
- Moore GR, Cox MC, Crowe D, Osborne MJ, Rosell FI, Bujons J, Barker PD, Mauk MR, Mauk AG (1998) *Biochem J* 332:449
- Gray HB, Ellis WR Jr (1994) In: Bertini I, Gray HB, Lipard SJ, Valentine JS (eds) *Bioinorganic chemistry*. University Science Books, Mill Valley, Calif, pp 315–364

59. Michel B, Proudfoot AEI, Wallace CJA, Bosshard HR (1989) *Biochemistry* 28:456–462
60. Lappalainen P, Watmough NJ, Greenwood C, Saraste M (1995) *Biochemistry* 34:5824–5830
61. Millett F, de Jong C, Paulson L, Capaldi RA (1983) *Biochemistry* 22:546–552
62. Witt H, Malatesta F, Nicoletti F, Brunori M, Ludwig B (1998) *Eur J Biochem* 251:367–373
63. Dill KA (1990) *Biochemistry* 29:7133–7155
64. Auld DS, Pielak GJ (1991) *Biochemistry* 30:8684–8690
65. Fredericks ZL, Pielak GJ (1993) *Biochemistry* 32:929–936
66. Pielak GJ, Auld DS, Beasley JR, Betz SF, Cohen DS, Doyle DF, Finger SA, Fredericks ZL, Hilgen-Willis S, Saunders AJ (1995) *Biochemistry* 34:3268–3276
67. Beasley JR, Pielak GJ (1996) *Proteins* 26:95–107
68. Colón W, Elöve GA, Wakem LP, Sherman F, Roder H (1996) *Biochemistry* 35:5538–5549
69. Colón W, Wakem LP, Sherman F, Roder H (1997) *Biochemistry* 36:12535–12541
70. Hammack B, Godbole S, Bowler BE (1998) *J Mol Biol* 275:719–724
71. Sauder JM, Roder H (1998) *Fold Des* 3:293–301
72. Marmorino JL, Pielak GJ (1995) *Biochemistry* 34:3140–3143
73. Marmorino JL, Lehti M, Pielak GJ (1998) *J Mol Biol* 275:379–388
74. Hostetter DR, Weatherly GT, Beasley JR, Bortone K, Cohen DS, Finger SA, Hardwidge P, Kakouras DS, Saunders AJ, Trojak SK, Waldner JC, Pielak GJ (1999) *J Mol Biol* 289:639–644
75. Jentzen W, Ma J-G, Shelnett JA (1998) *Biophys J* 74:753–763
76. Banci L, Rosato A, Turano P (1996) *JBIC* 1:364–367
77. Banci L, Bertini I, De la Rosa MA, Koulougliotis D, Navarro JA, Walter O (1998) *Biochemistry* 37:4831–4843
78. Winkler JR, Gray HB (1992) *Chem Rev* 92:369–379
79. Brautigam DL, Feinberg BA, Hoffman BM, Margoliash E, Peisach J, Blumberg WE (1977) *J Biol Chem* 252:574–582
80. Angström J, Moore GM, Williams RJP (1982) *Biochim Biophys Acta* 703:87–94
81. Hong X, Dixon DW (1989) *FEBS Lett* 246:105–108
82. Pearce LL, Gartner AL, Smith M, Mauk AG (1989) *Biochemistry* 28:3156
83. Battistuzzi G, Borsari M, Sola M, Francia F (1997) *Biochemistry* 36:16247–16258
84. Battistuzzi G, Borsari M, Cowan JA, Eicken C, Loschi L, Sola M (1999) *Biochemistry* 38:5553–5562
85. Ferrer JC, Guillemette JG, Bogumil R, Inglis SC, Smith M, Mauk AG (1993) *J Am Chem Soc* 115:7507–7508
86. Taler G, Schejter A, Navon G, Vig I, Margoliash E (1995) *Biochemistry* 34:14209–14212
87. Banci L, Bertini I, Spyroulias GA, Turano P (1998) *Eur J Inorg Chem* 1:583–591