PROTEINS: Structure, Function, and Genetics 31:201–213 (1998)

Accessibility to Internal Cavities and Ligand Binding Sites Monitored by Protein Crystallographic Thermal Factors

Oliviero Carugo[1,2]* and Patrick Argos[1]
[1]*European Molecular Biology Laboratory, Heidelberg, Germany*
[2]*Dipartimento di Chimica Generale, Universitá di Pavia, Pavia, Italy*

ABSTRACT **Protein structures are flexible both in solution and in the solid state. X-ray crystallographically determined thermal factors monitor the flexibility of protein atoms. A method utilizing such factors is proposed to delineate protein regions through which a ligand can exchange between binding site and bulk solvent. It is based on the assumption that thermally excited protein regions are excellent candidates for opening a ligand channel. Computationally simple and inexpensive, the method analyzes directions from which thermal factors can propagate within the protein, resulting in thermal motion paths (TMPs). Applications to engineered T4 lysozymes, where an artificial internal cavity can host hydrophobic molecules, and to sperm whale myoglobins, where the active site is completely buried, yielded results in agreement with other independent structural observations and with previous hypotheses. Further new features could also be suggested. The proposed TMP analysis could aid molecular dynamics simulation studies as well as time-resolved and site-directed mutagenesis experimental studies, especially given its modest computational expense and its direct roots in experimental results based on thermal factors determined in high-resolution crystallographic studies. Proteins 31:201–213,1998.** © 1998 Wiley-Liss, Inc.

Key words: accessibility to internal cavities; crystallographic thermal factors; ligand binding; protein dynamic; protein structure

INTRODUCTION

Protein flexibility (mobility or dynamics) is inherent to protein structural behavior.[1–6] Some random intramolecular motions appear not to have clear biological significance, although they could well influence thermodynamic state functions. However, many types of protein atom movements are clearly connected to some specific biochemical activity and have been differently classified.[5,7–9]

Experiments on protein flexibility are generally difficult to achieve. For example, solvent structure is often detected by crystallographic techniques but not by nuclear magnetic resonance (NMR) methods where the kinetics of exchange is too fast for relaxation experiments.[10] Nevertheless, experimental evidence for protein flexibility is extensive both in solution and in the solid state.[4,5,9,11] It has been often observed that no significant differences exist between protein reactivity in solution and in crystals, suggesting that crystal lattice constraints do not significantly modify the intrinsic flexibility of proteins. Even the buried protein core can be subjected to fast atomic movements, which allow various types of molecules to enter or exit buried cavities with highly effective kinetics.[2,3]

In the present work we have developed methods to detect possible pathways connecting the bulk solvent to buried cavities or ligand binding sites. We have assumed that protein regions of high thermal factors (or atomic displacement parameters[12]) indicate the preferred directions or paths for ligand exchange between protein core and solvent and that comparison of slightly different and independently determined molecular structures can simulate a reaction pathway.[13,14] Use of several structural determinations mitigates possible errors or misinterpretations and focuses the analysis of configurational modifications to stereochemical variances that point to regions with conformational freedom. A relatively early attempt to use thermal factors to infer information on structural deformation pathways was successful.[5] Confidence in the accuracy of thermal factors has been supported by studies in which thermal factors from similar (or identical) and independently refined protein crystal structures have been compared.[15,16]

We examined two typical cases, namely, the exchange of hydrophobic molecules between the bulk solvent and an internal engineered cavity in T4 lysozyme[3] and oxygen binding to myoglobin.[4,17] In

*Correspondence to: Dr. Oliviero Carugo, European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10.2209, D-69012 Heidelberg, Germany. E-mail: carugo@embl-heidelberg.de
Received 7 April 1997; Accepted 18 August 1997

both, no fully convincing experimental evidence has been provided to characterize channels connecting the bulk solvent and the core despite recent advances in time-resolved experiments on the myoglobin complexed with carbon monoxide.[18–20] The need for atomic motions to open a ligand exchange channel is obvious, since, if these proteins maintain their well-packed native and rigid structures as seen from crystallographic experiments, there are no energetically feasible paths for the ligands. In myoglobins, an early hypothesis based on molecular dynamics simulations[21] has found only indirect crystallographic confirmation.[17,22–24] Both for T4 lysozymes and sperm whale myoglobins we have found channels based on thermal factor analysis that connect the protein surface with buried cavities and ligand binding sites. Comparison of our results with other independent experimental data and with previous hypotheses showed agreement. We can also point to new features. Our approach should be a promising technique to interpret structural data from a dynamics perspective.

METHODS

Known three-dimensional protein crystal structures were taken from the Brookhaven Protein Data Bank.[25] Only structures with resolution better than or equal to 2.0 Å were retained to mitigate thermal factor errors.

Thermal factors generally accompany the positional coordinates of each atom in a crystal structure. Despite a sometimes substitution of "atomic thermal factor" by "atomic displacement parameter" as recommended by the International Union of Crystallography,[12] we prefer the former, which is more easily recognized by noncrystallographer structural biologists. The coordinates x, y, and z define the position at which the probability to find the atom is maximal, while the thermal factors describe how diffuse is the vibration of the atom around this equilibrium position. Thermal factors derive from the vibrational energy of an atom, which contributes to the observed diffracted X-ray intensities. Other factors, such as atomic positional disorder, can also influence their values. This may particularly apply to protein crystals where several regions of considerable conformational freedom are present. Solvent-accessible areas for protein atoms were computed with the program ASC[26] by using a probe radius of 1.25 Å as recommended by Hubbard and Argos.[27] Secondary structural assignments were effected with STRIDE.[28] Characteristics of the cavities were determined with SURFNET.[29] Atomic superpositions in equivalent structures were performed with the method of Kabsch and McLachan.[30,31] Hydrogen bonds were delineated with the algorithm HBPLUS[32] constrained by the stereochemical threshold values taken from Gandini.[33] Cutoffs at 3.5 Å and 2.5 Å were applied to the donor–acceptor and hydrogen–acceptor distances, respectively, while the angle at the donor was confined to the range 90°–180° and the angle at the acceptor to 90°–180° and 60°–180° for sp^2 and sp^3 hybridizations, respectively.

RESULTS

Thermal Motion Paths

All atom triplets are searched in a protein structure. A triplet is defined as a group of three atoms in contact amongst themselves (within 4.5 Å) and from different residues. The position of the triplet is defined as its center of mass and its thermal factor as the mean thermal factor of the three atoms. Triplets surrounding triplets are also searched (triplet centers within 7.1 Å) and that with highest thermal factor retained. This procedure, begun at all triplets, is applied itreratively moving from one triplet to the next with highest thermal factor in the surrounds, resulting in a thermal motion path (TMP) as illustrated in Figure 1. A TMP is characterized by its length (the number of triplets involved or procedural steps) and the mean value of the thermal factors of all triplets constituting it. The number of TMPs is equal to the number of starting triplets. The length of a TMP is typically between 2 and 7 for the protein examples examined here.

Such a procedure can, in principle, be extended to any number of atoms defining the basic unit. Considering only one or two instead of a triplet yields short and physically irrelevant TMPs, such as an isolated phenyl ring. A quartet of atoms yielded similar TMPs as those derived from triplets. The threshold of 7.1 Å between triplet mass centers to define their contact was selected as it corresponds to the distance between the center and the vertex of two equilateral triangles of side 4.5 Å as illustrated in Figure 1.

The atomic thermal factors taken from the Brookhaven Protein Data Bank[25] cannot be used directly, since they may be on different scales due to application of different refinement procedures.[34] For example, in the myoglobin structures examined here, the mean thermal factor (expressed in B units), calculated over all the protein atoms in each of the structures, ranges from 14 to 36 $Å^2$ (average 26 $Å^2$ and variance 5 $Å^2$) with corresponding standard deviations from 10 to 22 $Å^2$ (average 18 $Å^2$ and variance 3 $Å^2$). A similar trend was found for the lysozyme crystal structures examined here. We thus normalized them for each protein structure to have a distribution of zero mean and unit variance as

$$B \text{ normalized} = (\mathrm{B} - \langle \mathrm{B} \rangle)/\text{sigma-}B$$

where $\langle \mathrm{B} \rangle$ and sigma-B are, respectively, the mean value and the standard deviation of the distribution of observed thermal factors.

Once TMPs are collected, they can be clustered by grouping together those which share at least a given percentage of atoms. We applied overlap thresholds

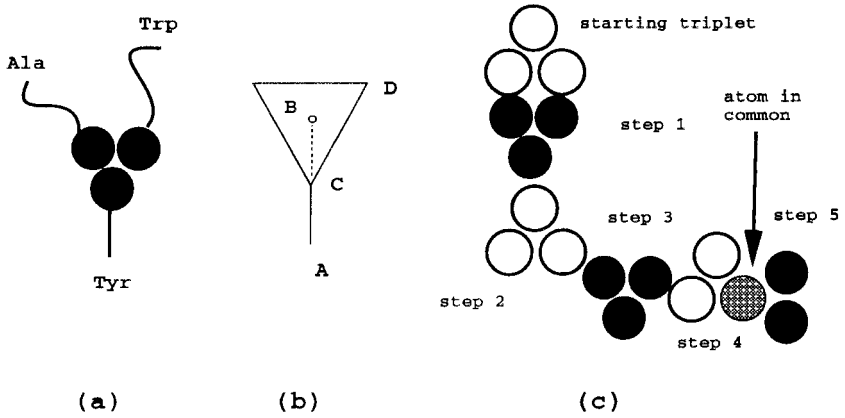


Fig. 1. **a:** Triplet of atoms in contact (within 4.5 Å) but from different residues. A triplet is characterized by its position (center of mass of constituent atoms) and by its thermal factor (mean of the thermal factors of defining atoms). **b:** Triplets are considered in contact if their centers of mass are closer than 7.1 Å, a threshold that corresponds to the distance *AB* or $[1 + (3)^{-1/2}]$ 4.5 Å, where $AC = CD = 4.5$ Å. Such a threshold allows superposition of triplets, i.e., concatenation of triplets that share one or two atoms. **c:** Starting at any triplet, the one that is in contact with it and with highest thermal factor is successively selected; one step involves a shared triplet atom. The consequent series of triplets forms a thermal motion path characterized by its length (the number of steps or triplets) and its thermal factor (mean of the thermal factors of its constituent triplets).

TABLE I. Tertiary Structures of T4 Lysozymes Examined in the Present Work*

PDB code	Mutations	Resolution	R factor	Ligand bound	Year of deposition	Refinement technique	Ref.
1811	C54T, C97A, L99A	1.80	0.158	Benzene	1995	TNT	39
1821	C54T, C97A, L99A	1.80	0.168	Benzofuran	1995	TNT	39
1831	C54T, C97A, L99A	1.80	0.177	Indene	1995	TNT	39
1841	C54T, C97A, L99A	1.80	0.162	Isobutylbenzene	1995	TNT	39
1851	C54T, C97A, L99A	1.80	0.169	Indole	1995	TNT	39
1861	C54T, C97A, L99A	1.80	0.164	*n*-Butylbenzene	1995	TNT	39
1871	C54T, C97A, L99A	1.80	0.170	*p*-Xylene	1995	TNT	39
1881	C54T, C97A, L99A	1.80	0.156	*o*-Xylene	1995	TNT	39
1183	C54T, C97A, L99A	1.70	0.152	Benzene	1992	TNT	37
1184	C54T, C97A, L99A, F153A	1.90	0.162	Benzene	1992	TNT	37
1nhb	C54T, C97A, L99A	1.80	0.165	Ethylbenzene	1995	TNT	39
1189	C54T, C97A, L99A, F153A	1.90	0.171		1989	TNT	54
1190	C54T, C97A, L99A	1.75	0.156		1990	TNT	54
11yd		2.00	0.191		1989	PROLSQ	55
21zm		1.70	0.193		1986	DERIV/EREF, TNT	56
31zm		1.70	0.157		1989	TNT	57
41zm		1.70	0.165		1991	TNT	58
51zm		1.80	0.159		1991	TNT	58
61zm		1.80	0.160		1991	TNT	58
71zm		1.80	0.155		1991	TNT	58

*All structures were crystallized in space group $P3_221$.

between 30% and 70% and found 50% appropriate. Each TMP cluster is characterized by the mean *B* normalized value of its constituent atoms.

Engineered T4 Lysozymes Complexed With Hydrophobic Molecules

We analyzed the TMPs of the T4 lysozyme crystal structures reported in Table I. There are 7 native forms, 11 mutants complexed with hydrophobic molecules, and 2 uncomplexed mutants for which complexed structures are also known. Only TMPs starting at the atoms bordering the internal engineered cavity were considered and, in the case of native forms, the cavity triplets in the benzene complex 182l were assumed as starting points. The cluster of TMPs of highest mean *B* normalized value of each crystal structures is shown in Figures 2 and 3. No other clusters of TMPs have been considered, since

TABLE II. Summary of the TMP Clusters Found in T4 Lysozymes

PDB code	Cluster of highest mean *B* normalized: No. of atoms	Cluster of highest mean *B* normalized: *B* Normalized	Cluster of second highest mean *B* normalized: No. of atoms	Cluster of second highest mean *B* normalized: *B* Normalized
1811	11	1.07	12	0.18
1821	18	1.45	12	0.21
1831	12	2.11	13	0.88
1841	11	1.60	13	0.12
1851	9	1.13	7	−0.15
1861	10	1.23	8	−0.12
1871	10	2.15	5	0.75
1881	11	1.41	10	−0.10
1183	9	1.35	12	0.06
1184	15	1.19	8	0.47
1nhb	11	2.20	16	0.49
1189	18	0.88	5	0.22
1190	18	0.99	6	0.10
11yd	18	0.66	11	0.24
21zm	13	1.00	8	−0.05
31zm	20	0.78	12	0.09
41zm	22	0.90	15	0.15
51zm	30	0.60	5	−0.07
61zm	19	1.14	6	−0.22
71zm	21	0.95	7	−0.13

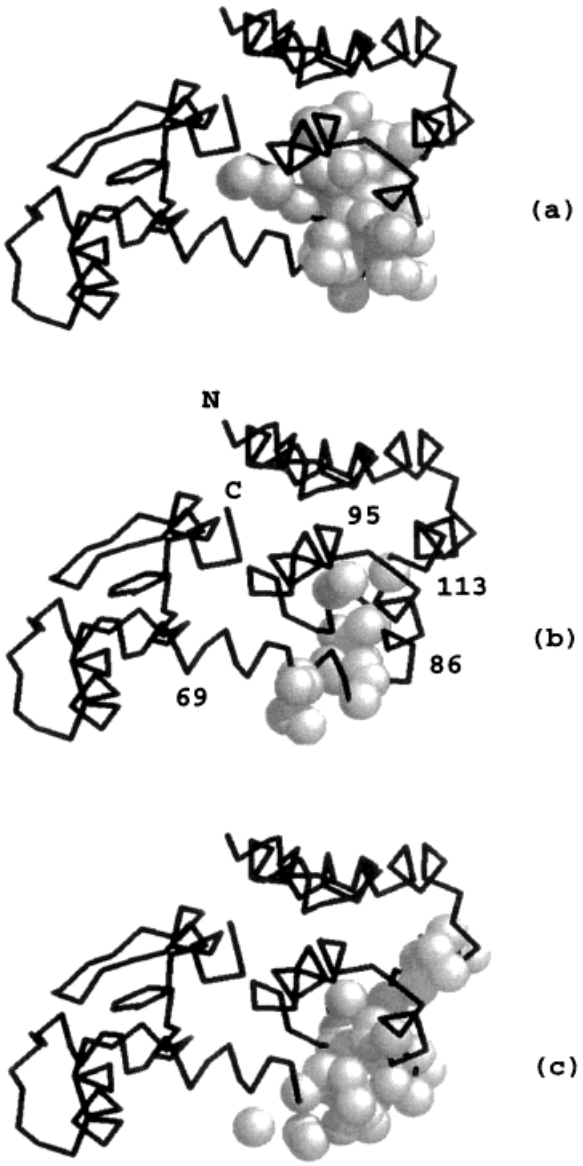


Fig. 2. Atoms constituting the clusters of TMPs of highest thermal factor in T4 lysozymes whose main-chain fold is indicated by lines connecting successive Cα atoms. **a:** Complexed mutants only. **b:** Uncomplexed mutants only. Approximate locations of Cα atoms for numbered residues are indicated along with the C terminus. **c:** Native T4 lysozymes. Each atom constituting TMPs is drawn as a sphere and some are constituents of more than one TMP and are therefore more noteworthy than others. For a quantitative evaluation see Figure 3.

they have much lower mean *B* normalized values (Table II). In most cases, the TMP cluster of second highest B normalized value involves the protein region between the location of the exogenous ligand and the N-terminal protein segment.

The TMP clusters of highest *B* normalized of the mutant complexed and uncomplexed forms involve mainly residues 75–85. The TMP clusters trend of the native T4 lysozymes is similar, albeit more atoms are involved. TMP lengths are also similar (mean values of 3.7, 3.0, and 3.6 for complexed mutants, uncomplexed mutants, and native forms, respectively), but the TMP thermal factors of the complexed mutants (Table II).

Inspection of the crystal structures suggests that a local unfolding of the protein segments 75–85 and 100–115 (Fig. 4) can be responsible for the formation of a path connecting the bulk solvent and the engineered internal cavity. It is also noteworthy that, if the protein atoms involved in at least one TMP of the complexed mutant T4 lysozymes are deleted to simulate roughly their large-scale movements, the buried hydrophobic molecules appear clearly solvent-accessible (Fig. 5).

Figure 6 shows that by superposing the protein segments 60–90 (Cα atoms only) from all structures onto one, maximum variability is found in the positioning of segment 100–115. Similarly, superposing the segments 100–120 shows variability in the positioning of segments 75–85. The relative movement of these two protein segments is clearly observable only in the complexed mutants and not in the native T4 lysozymes.

Protein segments 75–85 and 100–115 have a common interface with a strong interaction observed in a hydrogen bond between ND1 of Asn 81 and OEX of Glu 108 (X = 1 or 2). The donor–acceptor distance is more variable in the complexed mutants (3.0 ± 0.3) than in the other cases (2.7 ± 0.2), suggesting a positional disorder of the residues surrounding the

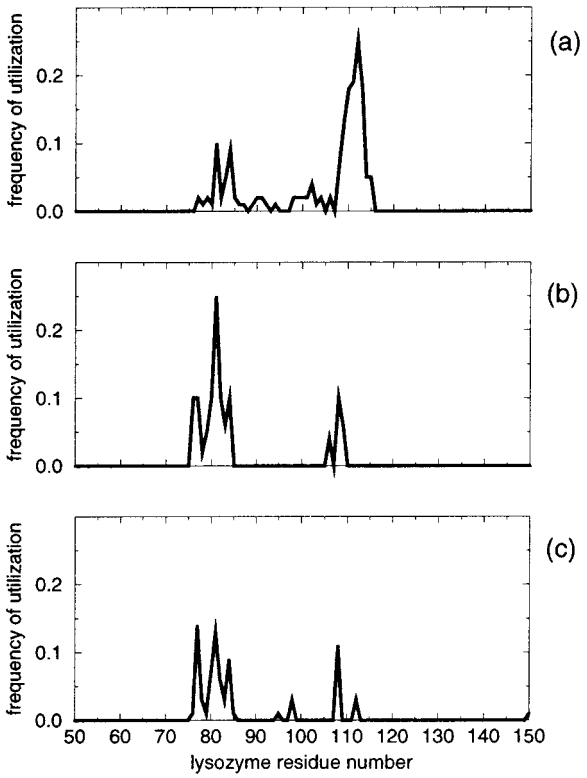


Fig. 3. Frequency of utilization of T4 lysozyme atoms from particular residues in the TMP cluster of the highest thermal factors. **a:** Complexed mutants. **b:** Uncomplexed mutants. **c:** Native structures. The frequency of utilization is defined as the percentage of atoms within each residue that are found in a TMP.

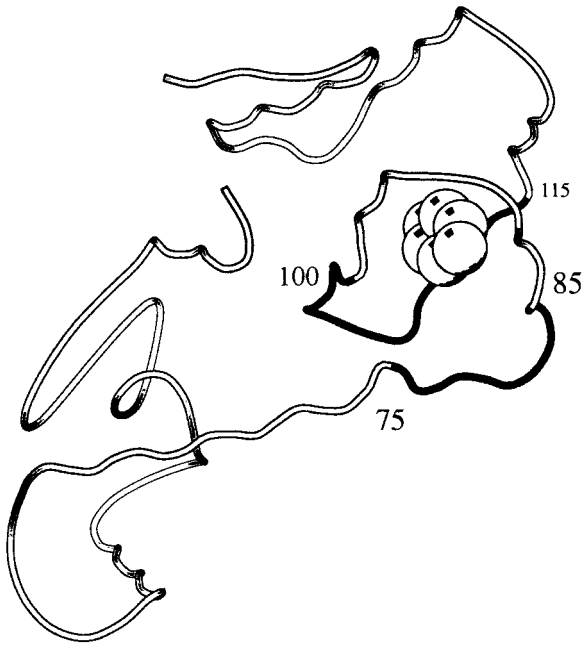


Fig. 4. Ribbon view of the T4 lysozyme backbone fold (structure 181l taken as an example). The protein segment 75–85 and 100–115 are delineated as well as the hydrophobic ligand (benzene) buried within the internal engineered cavity.

bond. It follows that the opening and closing of the path to the solvent is enhanced by the presence of the hydrophobic ligand.

The structure of the solvent at the protein surface patch corresponding to segments 75–85 and 100–115 is also indicative of their positional disorder. Figure 7 shows that no water molecules have ever been localized in the middle of the surface patch over the engineered cavity despite the presence of many polar groups in the vicinity. This is observed both in native and mutant lysozymes and can be interpreted as a consequence of disorder. Water molecules are crystallographically invisible if they dynamically occupy positions close to each other in a given region. It is also noteworthy that the shortest distance between the center of the engineered internal cavity and residue atoms at the protein surface involves residues 80–120 (data not shown).

Path of Molecular Oxygen to Iron in Myoglobins

We analyzed the tertiary myoglobin structures of space group $P2_1$ reported in Table III; they are reduced or ferric forms and are also uncomplexed or complexed to a series of variable ligands. Only TMPs starting at one of the protein atoms within 4.5 Å of the oxygen ligand in 1mbo (used as a canonical structure) were retained. Cluster analyses of TMPs always produced a unique big cluster on the heme distal side, with some little appendices in various directions. Since this is only slightly informative at a molecular level, only the TMP of highest thermal factor was considered for each structure. Similar results were nevertheless obtained by considering all TMPs.

Figure 8 shows that four protein segments are involved in TMPs (15–30, 40–50, 60–70, and 115–125). The last is less frequently observed, and it is not detectable in complexed iron(II) species where various ligands can actually enter and exit the protein core. Thus we will concentrate on the two protein segments 40–50 and 60–70 used in TMPs observed for all the types of myoglobins as well as on the protein segment 15–30 used in the reduced enzymes.

Superpositions of selected segments of backbone Cα atoms corresponding to various secondary structural elements did not show any significant positional variability as was the case for the lysozymes. We thus propose that no local unfolding takes place and exchange channels to the iron cation are opened by side chain movements, not surprising given the smallness of the natural oxygen ligand.

Superposition of the heme atoms excluding the propionate groups shows that a number of side chains have considerable positional variability. Given the periodicity of the torsion angle values, side chain

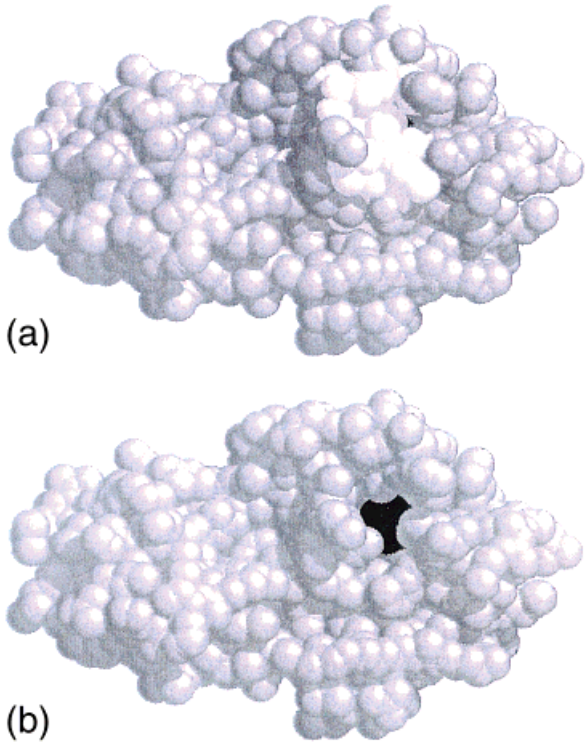


Fig. 5. Atomic space-filled view of the T4 lysozyme (181l). **a:** Benzene (*black*) is hidden from the solvent where the protein atoms involved in at least one TMP (only the TMP of highest thermal factor is considered for each structure) are shown in *white*. **b:** The protein atoms are removed such that the benzene molecule becomes exposed.

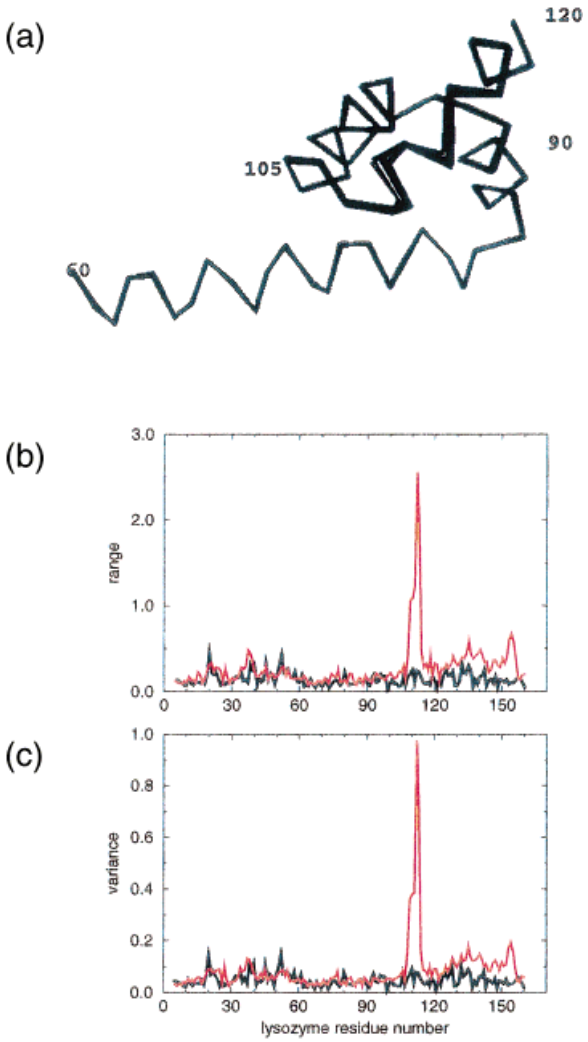


Fig. 6. Selected Cα atoms of all T4 lysozymes when those in segments 60–90 are superposed onto the corresponding Cα atoms in the structure 1lyd. **a:** The variable position of the protein segment 100–115 is clearly shown to be very variable. **b:** Maximal distances found between equivalenced Cα atoms for the complexed mutants (*red*) and the native forms (*black*). The highest variability is in the protein segment 100–115. **c:** Variance of the distances between equivalenced Cα atoms with the complexed mutant results shown in *red* and those from the native forms in *black*. The highest variability is in the protein segment 100–115. In Figures b and c the reciprocal variability in the positions of the protein segments 75–85 and 100–115 is much higher in complexed mutants than in native T4 lysozymes.

variability can be effectively measured with the circular dispersion.[13] All the single torsions are vectorially summed, and the length of the resulting vector is inversely proportional to their variability; by normalizing the sum vector length over the number of contributors, it follows that a circular dispersion close to 1.0 indicates little variability, while 0.0 points to a randomly distributed sample. The mean circular dispersion of residues in the range 20–70 are shown in Figure 9; the dispersion is inversely correlated with the conformational variability of the side chain atoms. The smallest values are found in the three segments of interest (20–30, 40–50, 60–70) and those <0.7 correspond to Gln 26, Phe 43, Arg 45, and His 64. The only completely buried side chains with circular dispersion <0.9 are His 64 and Leu 61; structure inspection clearly shows that they can be involved in two different paths. It is unlikely that they identify the same ligand exchange channel, since we did not find any significant correlation between their positions. Possible paths nearly parallel to the heme plane or nearly perpendicular to the heme plane are depicted in Figures 10 and 11. The first path (Fig. 10) principally involves the side chains of His 64 and Arg 45 and the D propionate of the heme, while the second (Fig. 11) utilizes the side chains of Leu 61 and Gln 25. Their random movements, probably coupled with minor perturbations of the surrounding residues, could well open a corridor to the heme iron.

The solvent structure is different at the protein surface patches relevant to the two proposed exchange corridors. The protein surface patch near the D propionate and the Arg 45 side chain is obviously polar, and a number of water molecules have been crystallographically localized in their vicinity. Figure 12 shows that no solvent molecules have been positioned on the surface patch involving the path perpendicular to the heme plane despite the pres-

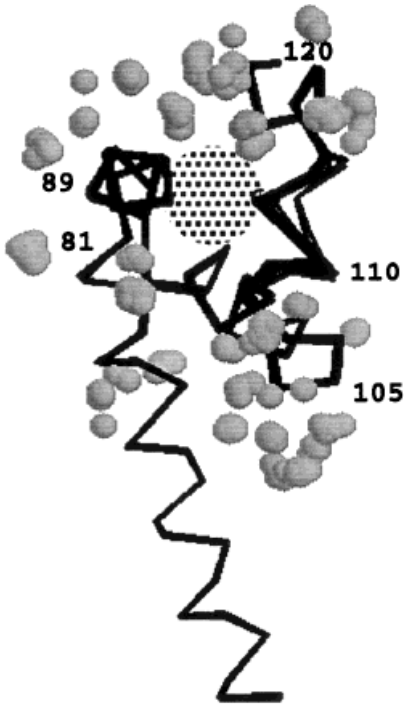


Fig. 7. Selected Cα atoms of all T4 lysozymes when all segments 60–90 are superposed onto that of 1lyd. Besides the emphasized positional variability of the segment 100–115, a protein surface patch over the cavity (*dotted*) is not covered by water molecules as observed in all the various structures.

ence of many hydrophilic residues. The path nearly parallel to the heme plane involves residues that correspond to several of the shortest possible distances between the metal center and the protein surface, while the exchange corridor nearly perpendicular to the heme plane does not.

To confirm the above results, we also analyzed tertiary myoglobin structures crystallized in other space groups (Table III); namely, 1abs (space group P6) [35] and 2cmm (space group $P2_12_12_1$).[36] For the former we obtained results closely similar to those reported above for the structures in space group $P2_1$ (Fig. 13). The TMP cluster of highest mean B normalized value is on the heme distal side, close to His 64, Leu 61, Arg 45, and Gln 26. For 2cmm most TMPs were nearly perpendicular to the heme plane and clearly indicate the ligand exchange path involving Leu 61 and Gln 26 (Fig. 14). No TMPs are found near His 64 and Arg 45 corresponding to the exchange channel parallel to the heme plane. In 2cmm the porphyrin is not the native one and lacks the two propionate groups likely resulting in the reorientation of the Arg 45 side chain toward the protein exterior.[36] This is also a likely explanation for the lack of TMPs along a ligand exchange corridor parallel to the heme plane. Moreover, the reorientation of the Arg 45 side chain could cause the unique space group such that the protein surface patch involved in the path through His 64 and Arg 45 is partially involved in crystal packing contacts (Fig. 14).

DISCUSSION

TMPs in Lysozyme and Globin

Atomic thermal factors in proteins can be very informative regarding molecular and submolecular movements, since they reflect not only atomic thermal motion (the vibration of an atom around its equilibrium position) but also the static and/or dynamic disorder of the atomic position.[15,16] They span a large range of values, from a few to many tens of square angstroms, when expressed as *B* values. The mean value of the *B*s in reported protein crystal structures (generally about 15–20 $Å^2$) indicates a very high root-mean-square amplitude *u* of the atomic displacement from the equilibrium position [about 0.4–0.5 Å, since $u = (B/8\pi^2)^{1/2}$].

The method we propose to find passages for ligands from solvent to buried cavities or binding sites is very simple and easy to use. We assume that the protein regions most likely involved in opening corridors can be identified by contacting-atom triplets as well as contacting triplets, all with high thermal factors. The efficacy of the method was shown by application to T4 lysozymes and sperm whale myoglobins.

It has been found that, by replacing Leu 99 and/or Phe 153 with Ala in T4 lysozyme, the protein structure changes little except in the formation of a large internal cavity which can host a hydrophobic molecule such as benzene or a derivative with a resultant increase in protein stability.[37–39] Later studies suggested that the hydrophobic ligands can enter the cavity rapidly (bimolecular rate constants of about 10^6–10^7 $M^{-1}s^{-2}$) and with a low activation energy barrier (2–5 kcal/mol).[3] Since the cavity is completely buried by protein atoms, these results are surprising and suggest that relatively large (1–2 Å) atomic movements occur quickly (at the microsecond time scale).[3]

Protein segment 108–113 has been suggested the protein region most likely to allow the hydrophobic molecule to reach the cavity [3]; our results based on TMPs in several structures (Table I) support this hypothesis. However, we propose in addition to the 100–115 segment a second region involving residues 75–85 such that through a local unfolding of these two segments, a direct channel is opened to the buried cavity. This hypothesis is supported by a number of independent observations. There is high variability in the positions of the one of the fragments upon superposition of the other segment taken from several structures. The positional spread is greater for T4 lysozymes complexed with hydrophobic molecules, suggesting that they strongly perturb these regions by entering and exiting the buried cavity. The strongest discrete interaction between the two protein segments involves a hydrogen bond between Asn 81 and Glu 108 which is strongly perturbed in complexed lysozymes. No ordered solvent molecules have ever been observed at the

TABLE III. Tertiary Structures of Sperm Whale Myoglobin Examined in the Present Work*

PDB code	Resolution	R factor	Ligand bound	Year of deposition	Refinement technique	Ref.
1mbd	1.40			1981	EREF	59
1vxa	2.00	0.177		1996	X-PLOR	59
1vxb	2.00	0.200		1996	X-PLOR	60
1vxc	1.70	0.155		1996	X-PLOR	60
1vxd	1.70	0.153		1996	X-PLOR	60
1vxe	1.70	0.156		1996	X-PLOR	60
1vxf	1.70	0.146		1996	X-PLOR	60
1vxg	1.70	0.157		1996	X-PLOR	60
1vxh	1.70	0.140		1996	X-PLOR	60
5mbn	2.00	0.179		1988	EREF	61
1mbo	1.60		O_2	1981	EREF	62
1mbc	1.50	0.171	CO	1988	PROLSQ	41
1spe	2.00	0.173	CO	1995	X-PLOR	60
1swm	1.80	0.149	Azide	1992	TNT	63
2mya	2.00	0.178	Ethyl isocyanide	1993	X-PLOR	24
2myb	1.90	0.152	Methyl isocyanide	1993	X-PLOR	24
2myc	1.80	0.167	*n*-Butyl isocyanide	1993	X-PLOR	24
2myd	1.80	0.148	*n*-Propyl isocyanide	1993	X-PLOR	24
2mye	1.68	0.172	Ethyl isocyanide	1993	X-PLOR	64
4mbn	2.00	0.172	Water	1988	EREF	61
1mbi	2.00	0.148	Imidazole	1990	TNT	17
1abs	1.50	0.207	CO	1997	X-PLOR	35
2cmm	1.80	0.187	Cyanide	1993	X-PLOR	36

*All structures were crystallized in space group $P2_1$ with the exception of labs (P6) and 2cmm ($P2_12_12_1$).

exposed protein surface patch associated with these fragments despite significant hydrophilicity in the region. This is probably a result of local unfolding on a time scale that does not allow the formation of crystallographically detectable solvent structure. The entrance/exit path is the shortest connecting the buried engineered cavity with the bulk solvent.

Muscle myoglobins bind oxygen tightly through coordination to a heme–iron(II) metal center and thus function as oxygen storage proteins.[40] It is not completely understood how oxygen enters the distal cavity nor how it exits toward the bulk solvent. If myoglobin were to maintain its rigid native structure known from crystallographic analysis, there is no discernible, energetically feasible path.

An early study involving molecular dynamics[21] suggested that two paths were probable, one approximately parallel to the heme group through the E helix and the other directed perpendicular to the heme through the B helix. The hypothesis of the first path was supported by theoretical calculations[41] with an energy barrier of approximately 5 kcal/mol. Quite a few experiments supported the existence of the first path. Bolognesi[22] also showed that in ferric sperm whale myoglobin few residues are extremely mobile; when imidazole binds the heme, His 64 swings up from the iron cation and toward the protein surface while Arg 45 moves into the solvent leaving a channel connecting the distal cavity and the bulk solvent.[17,22] Similar features were observed in the reduced globin complexed with alkyl isocya-

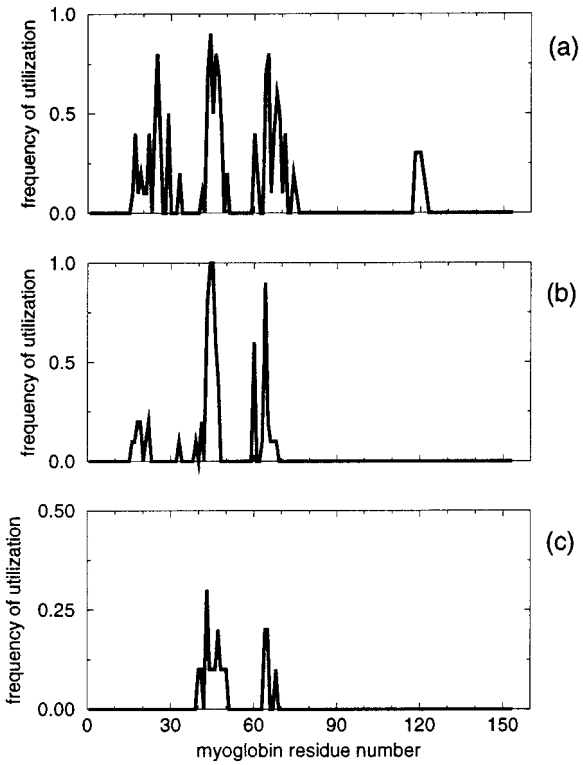


Fig. 8. Frequency of utilization of residues by the TMPs of the highest thermal factors. **a:** Uncomplexed myoglobins. **b:** Complexed myoglobins. **c:** Ferric myoglobins. See Figure 3 legend for definitions.

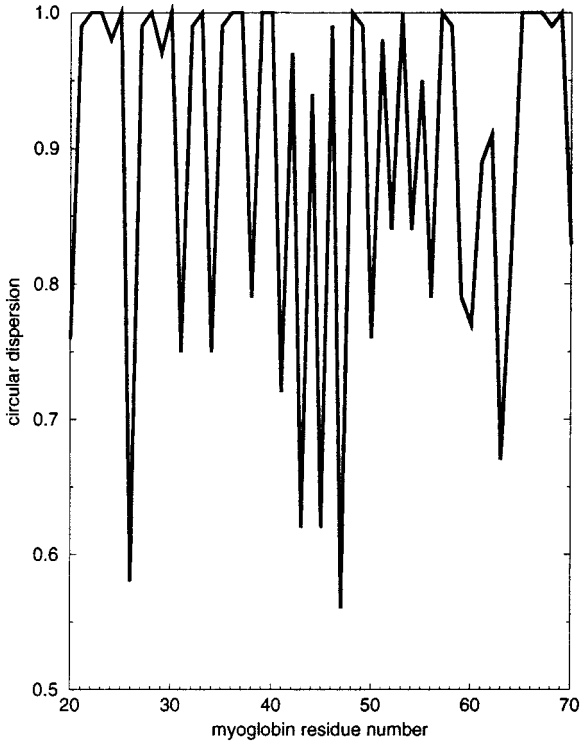


Fig. 9. Variability of the side-chain conformations in the myoglobin protein segment 20–70 after superposition of the heme atoms. Variability is evaluated with circular statistic techniques on the torsions, since the latter are periodic (see text). The circular dispersion values range between 0.0 and 1.0 and values close to 1.0 indicate little variance in conformation, while those equal to 0.0 indicate a randomly distributed sample.

nides[24] and phenylhydrazine.[23] More recently time-resolved experiments have further confirmed previous hypotheses and have highlighted some more detailed molecular features in the closely related case of myoglobin complexed with carbon monoxide.[18,19]

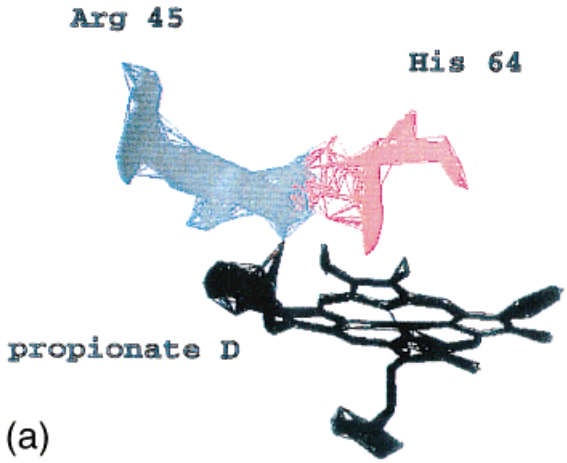
By analyzing the structures of sperm whale myoglobin reported in Table III, we found that the TMPs of highest thermal factor involved protein segments 15–30, 40–50, and 60–70. By further examining residues and heme components with greatest positional variability and by assuming that they indicate the direction of the channels connecting the iron cation to the bulk solvent, we have suggested two possible pathways: one parallel to the heme plane utilizing the side chains of His 64 and Arg 45 and the D propionate of the heme and the other perpendicular to the heme plane mainly involving Gln 26 and Leu 61. Since no particular positional variability of secondary structural elements relative to one another was observed, we propose that these corridors can be opened by movements of the side chains alone, a hypothesis supported by the small dimension of oxygen. Our results resemble those of the early molecular dynamics simulations [21] which favored the channel parallel to the heme plane. The channel perpendicular to the heme plane has not been extensively investigated experimentally.

It is possible that both globin channels are active depending on the oxidation state of the iron cation or on the ligand, which can be either molecular oxygen or water. The second path, detectable only in iron(II) myoglobins, likely involves oxygen while the first could prefer water. The two corridor hypothesis is also not in conflict with the well-known theory of conformational substates applied to myoglobins.[4] Observations that rebinding of the ligand to the heme active center is nonexponential in time below 200 K[42] could also be explained by the presence of two exchange channels with molecular features differently influenced by temperature.

The difference between the thermal factors of the two possible ligand exchange channels is not large. Side chains of His 64 and Arg 45 have slightly higher mean *B* normalized values (0.40) than side chains of Leu 61 and Gln 26 (0.19), though both values are considerably higher than the mean *B* normalized of the heme (−0.55). For the typical case with mean B factor 20 Å^2 and standard deviation 8 Å^2, the root-mean-square deviation from the equilibrium position, given by $(8\pi^2 B)^{1/2}$, is comparable for the path parallel (0.54 Å) and perpendicular (0.52 Å) to the heme plane, making any preference difficult to ascertain. Further, the dimensions of the thermal factors cannot in general yield reliable information regarding the kinetics of atomic movements, and it is therefore not possible to estimate systematically activation energy barriers from B factors alone, which reflect both fast events such as thermal vibrations as well as slower movements such as conformational transitions. No information can be systematically obtained from crystallographic methods regarding the correlation of the motions of different atoms,[43] especially given that only isotropic thermal factors are determined[44,45] for most protein crystal structures due to insufficient diffraction data to refine six anisotropic parameters for each atom. We reasonably assume that both vibrations and conformational transitions are randomly directed in space as a result of protein elasticity and that both can contribute to the opening of ligand channels.

Analyzing TMPs

The calculation of TMPs allows proposals regarding protein regions involved in ligand entrance to and exit from buried cavities and binding sites. Further structural information can be used to characterize atomic paths and channels. Time-resolved experimental techniques, theoretical dynamics simulations, and landscape energy calculations can suggest various individual molecular details. However, to a first level of approximation, simple and readily available structural data can help in focusing the molecular features of atomic paths and channels pointed out by TMP calculations.



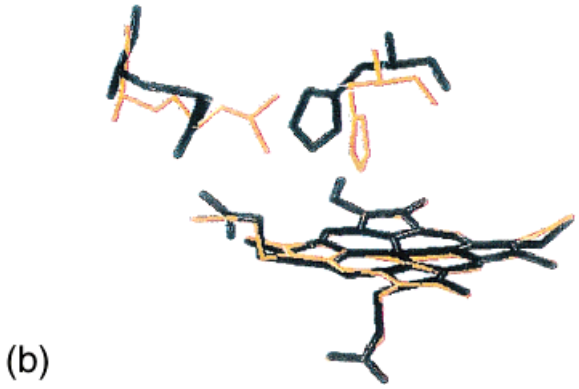
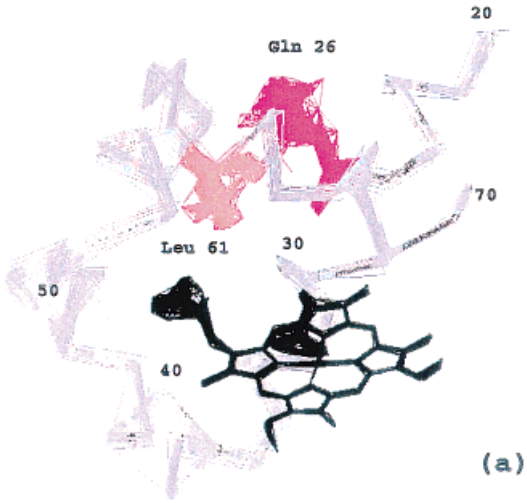


Figure 10.



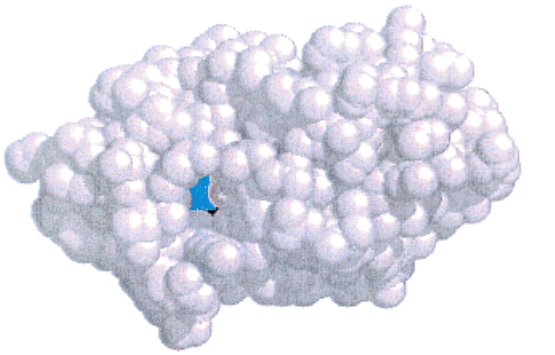


Figure 11.

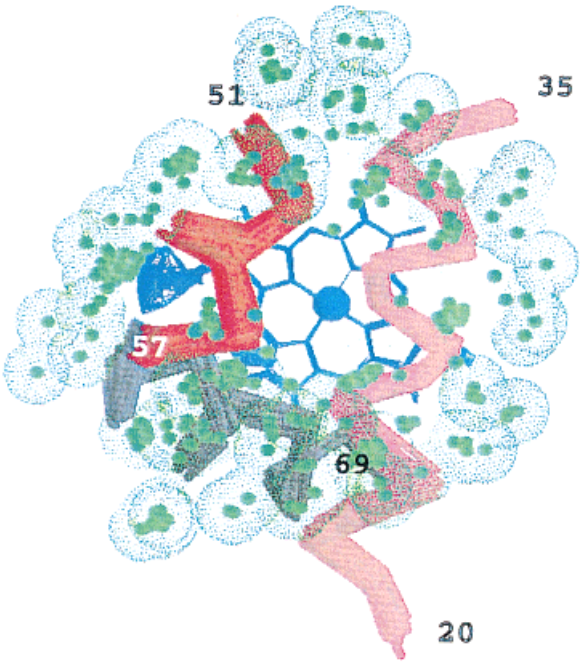


Figure 12.

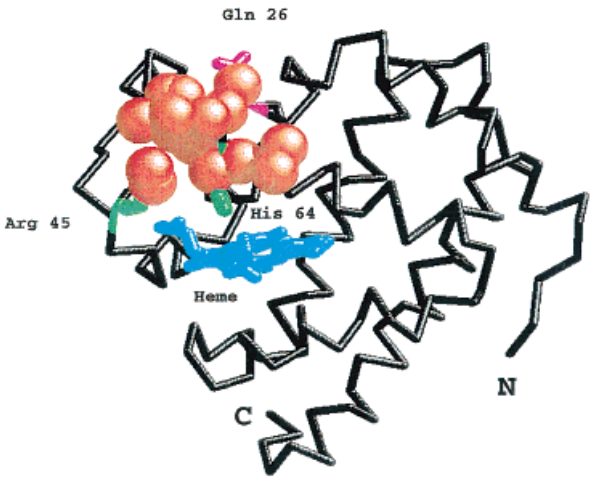


Figure 13.

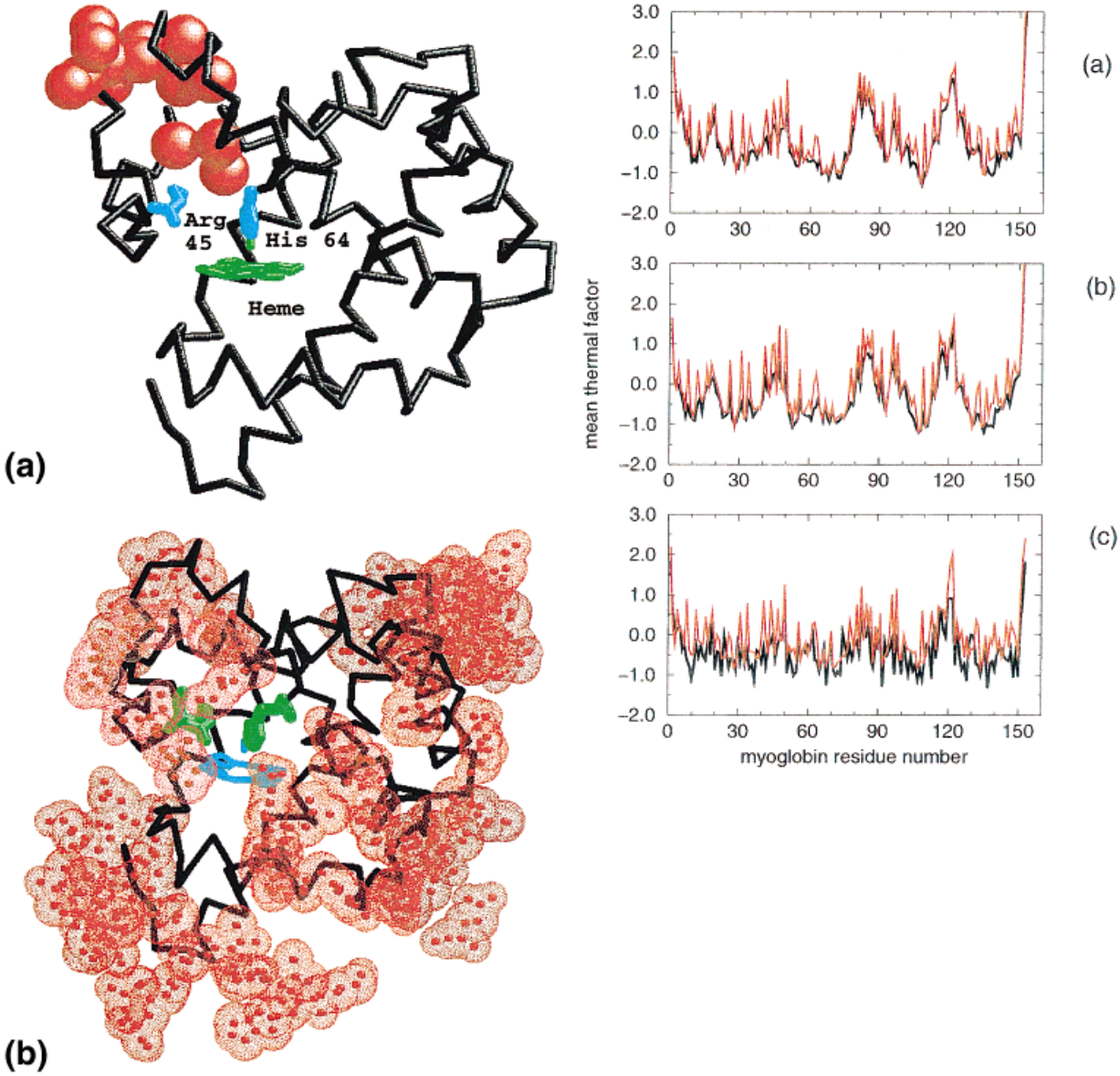


Fig. 14. **a:** TMP analysis of the crystal structure 2cmm. The protein atoms constituting the TMP cluster of highest mean B normalized value are shown as *red* spheres, which are on the heme distal side and consistent with a ligand exchange channel perpendicular to the heme plane. Residues involved in a ligand exchange channel parallel to the heme plane (His 64 and Arg 45) are untouched by the TMPs. **b:** Atoms of symmetry-related protein molecules in contact with the central molecule of 2cmm are delineated as *red dotted* spheres. The protein surface patch close to His 64 and Arg 45 is partially involved in crystal lattice contacts.

Fig. 15. (above, right) Mean thermal factors of versus the residue number. **a:** Uncomplexed myoglobins. **b:** Complexed myoglobins. **c:** Ferric myoglobins. Thermal factors for Cα atoms are *black*, while average thermal factors for the side-chain atoms are *red*.

Fig. 10. Possible molecular interpretation of the opening of a ligand exchange channel parallel to the heme plane with views from the solvent side. **a:** The heme atoms (excluding the two propionates) of all the structures were superposed to those of 1mbd. The side chains of His 64 and Arg 45 and the D propionate group of the heme show the largest conformational variability. **b:** Superposition of the heme atoms (excluding the two propionates) of 1mbd and 1mbi. The deoxymyoglobin (1mbd, *orange*) does not allow access of the heme iron ion from the solvent. If the His 64 side chain rotates about 90° around its Cα—Cβ bond, Arg 45 moves its side chain toward the solvent, and the D propionate of the heme is reoriented (1mbi, ferric myoglobin complexed with imidazole, *black*), the iron cation becomes solvent-accessible.

Fig. 11. Possible molecular interpretation of the opening of a ligand exchange channel perpendicular to the heme plane. The heme atoms (excluding the two propionates) of all the structures were superposed to those of 1mbd. Random movements of the side chains of Leu 61 and Gln 26 can make the heme solvent-accessible. **a:** View from the proximal side. **b:** If Leu 61 and Gln 26 are removed, the heme (*blue*) and iron cation (*black*) become solvent-accessible.

Fig. 12. Protein surface patch near the TMP leading to the iron cation perpendicular to the heme plane. **a:** View from the solvent after superposition of the heme atoms (with the exception of the two propionates) of all the structures to corresponding heme atoms of 1mbd (Table I). No water molecules (depicted as *green* solid spheres) have been crystallographically located at the entrance of the corridor in any of the structures.

Fig. 13. TMP analysis of the crystal structure 1abs. The protein atoms constituting the TMP cluster of highest mean B normalized value are shown as *red* spheres. They are on the heme distal side and involve residues whose side chains can open ligand exchange channels.

The backbone segments flanking a TMP can be analyzed either by superposing selected protein segments and/or by examining the stereochemistry of discrete local interactions. The observation of a reciprocal positional variability through successive superpositions of each of the segments suggests local unfolding of the main chain, while the contrary points to side chain movements as the major determinant of channel formation. The ligand size can also be used to discriminate among mechanisms. Local backbone unfolding is a reasonable prerequisite for large exogenous ligands, while smaller ligands might enter the protein core by simply perturbing the side chain stereochemistry. Static changes in the backbone structure caused by the steric requirements of large exogenous ligands can be observed, but they may not be accompanied by high thermal motion.

Another important feature is the solvent accessibility of the possible channel. It is of course essential that any hypothesized channel involve a protein surface patch accessible to the solvent.[11] In many cases, protein crystals have been soaked with the appropriate exogenous ligand without perturbing the intrinsic crystal stability and quality. This clearly indicates that the exchange channel is open to the solvent. It is unlikely that a TMP exits at a site of protein–protein crystal packing contact, since the associated atomic thermal motion is expected to be lower than that in solvent-accessible regions. The fraction of protein surface involved in crystal contacts is quite variable,[46,47] and it is possible that in specific cases the crystal structural model is not representative of the in vivo solution structure.

The absence of ordered solvent structure around the surface patch near the ligand entrance of exit can depend on its hydrophobicity[48,49] or on its thermal excitation and/or disorder. If no water molecules are crystallographically detected near a polar surface patch, it is reasonable to suppose the latter explanation, though such observations clearly depend on the degree of disorder and on the time scale of the atomic movements and cannot automatically discriminate local unfolding from side chain movements. Nevertheless, TMP channels along directions corresponding to the shortest paths between the internal site and bulk solvent should be preferred.

TMP analysis cannot be reduced to simple observations of thermal factor variation along the protein sequence. Figure 15 shows that the TMP protein segments suggested here (15–30, 40–50, and 60–70) display thermal factors that are not particularly distinguishable from other regions in complexed, uncomplexed, or ferric myoglobin forms. Similar results can be shown for the lysozymes. The protein three-dimensional structure is necessary to use the information provided by thermal factors. Caution should also be taken in protein regions involving crystal packing contacts where particular atomic thermal factors may be artificially lowered and thus some possible TMPs missed.

TMP analysis could also aid molecular dynamics simulation studies as well as time-resolved and point-mutation experimental studies, especially given its modest computational expense and its direct roots in experimental results based on thermal factors determined in high resolution structures from X-ray crystallography and from NMR with analogous positional dispersion parameters.[50] By TMP analysis it is possible to delineate the protein region(s) to be analyzed by more extensive and expensive methods to detect local interaction affinities and mechanisms at the atomic level. Intriguing problems that could be investigated by TMP analysis include oxygen and proton pumps in cytochrome c oxidases,[51,52] buried water exchanges,[2] and amide proton exchanges.[53] TMPs should also prove useful in designing and engineering ligand binding sites.

REFERENCES

1. Ogata, K., Kanei-Ishii, C., Sasaki, M., et al. The cavity in the hydrophobic core of Myb DNA-binding domain is reserved for DMA recognition and trans-activation. Nature Struct. Biol. 3:178–187, 1996.
2. Denisov, V.P., Peters, J., Horlein, H.D., Halle, B. Using buried water molecules to explore the energy landscape of proteins. Nature Struct. Biol. 3:505–509, 1996.
3. Feher, V.A., Baldwin, E.P., Dahlquist, W. Access of ligand to cavities within the core of a protein is rapid. Nature Struct. Biol. 3:516–521, 1996
4. Frauenfelder, H., Sligar, S.G., Wolynes, P.G. The energy landscapes and motions of proteins. Science 254:1598–1603, 1991.
5. Ringe, D., Petsko, G.A. Study of protein dynamics by X-ray diffraction. Methods Enzymol. 131:389–433, 1986.
6. Karplus, M. Internal dynamics of proteins. Methods Enzymol. 131:283–307, 1986.
7. Chothia, C., Lesk, A., Dodson, G.G., Hodgkin, D.C. Transmission of conformational change in insulin. Nature 302: 500–505, 1983.
8. Lesk, A., Chothia, C. Mechanisms of domain closure in proteins. J. Mol. Biol. 174:175–191, 1984.
9. Gerstein, M., Lesk, A., Chothia, C. Structural mechanism for domain movements in proteins. Biochemistry 22:6739–6749, 1991.
10. Bryant, R.G. The dynamics of water-protein interactions. Annu. Rev. Biophys. Biomol. Struct. 25:29–53, 1996.
11. Mozzarelli, A., Rossi, G.L. Protein function in the crystal. Annu. Rev. Biophys. Biomol. Struct. 25:343–365, 1996.
12. Trueblood, K.N., Bürgi, H.B., Burzlaff, H., et al. Atomic displacement parameter nomenclature: Report of a subcommittee on atomic displacement parameter nomenclature. Acta Crystallogr. A52:770–781, 1996.
13. Bürgi, H.B., Dunitz, J.D. "Structure Correlation." Weinheim, Germany: VCH Verlagsgesellschaft, 1994.
14. Domenicano, A., Hargittai, I. "Accurate Molecular Structures." Oxford, UK: Oxford University Press, 1992.
15. Giacovazzo, C. "Fundamentals of Crystallography." Oxford, UK: Oxford University Press, 1992.
16. Glusker, J.P., Lewis, M., Rossi, M. "Crystal Structure Analysis for Chemists and Biologists." New York: VCH Publishers, 1994.
17. Lionetti, C., Guanziroli, M.G., Frigerio, F., Ascenzi, P., Bolognesi, M. X-ray crystal structure of the ferric sperm whale myoglobin: Imidazole complex at 2.0 Å resolution. J. Mol. Biol. 217:409–412, 1991.
18. Lim, M., Jacksin, T.A., Anfinrud, P.A. Ultrafast rotation and trapping of carbon monoxide dissociated from myoglobin. Nature Struct. Biol. 4:209–214, 1997.
19. Srajer, V., Teng, T., Ursby, T., et al. Photolysis of the carbon monoxide complex of myoglobin: Nanosecond time-resolved crystallography. Science 274:1726–1729, 1996.
20. Vitkup, D., Petsko, G.A., Karplus, M. A comparison be-

tween molecular dynamics and X-ray results for dissociated CO in myoglobin. Nature Struct. Biol. 4:202–208, 1997.
21. Case, D.A., Karplus, M. Dynamics of ligand binding to heme proteins. J. Mol. Biol. 132:343–368, 1979.
22. Bolognesi, M., Cannillo, E., Ascenzi, P., Giacometti, G.M., Merli, A., Brunori, M. Reactivity of ferric *Aplysia* and sperm whale myoglobins towards imidazole: X-ray and binding study. J. Mol. Biol. 65:305–315, 1982.
23. Ringe, D., Petsko, G.A., Kerr, D.E., Ortiz de Montellano, P.R. Reaction of myoglobin with phenylhydrazine: A molecular doorstop. Biochemistry 23:2–4, 1984.
24. Johnson, K.A. "High-Resolution X-Ray Structures of Myoglobin and Hemoglobin-Alkyl Isocyanide Complexes." Doctoral thesis, Rice University, Houston, TX, USA, 1993.
25. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., et al. The protein databank: Computer-based archival file for macromolecular structure. J. Mol. Biol. 112:535–542, 1977.
26. Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C., Scharf, M. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J. Comput. Chem. 16:273–284, 1995.
27. Hubbard, S.J., Argos, P. Detection of internal cavities in globular proteins. Protein Eng. 10:1011–1015, 1996.
28. Frishman, D., Argos, P. Knowledge-based secondary structure assignment. Proteins 23:566–579, 1995.
29. Laskowski, R.A. SURFNET: A program for visualizing molecular surfaces, cavities and intermolecular interactions. J. Mol. Graph. 13:323–330, 1995.
30. Kabsch, W. A discussion of the best solution of the best rotation to relate two sets of vectors. Acta Crystallogr. A34:828–828, 1978.
31. McLachan, A.D. Gene duplication in the structural evolution of chymotrypsin. J. Mol. Biol. 128:48–67, 1979.
32. McDonald, I.K., Thornton, J.M. Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. 238:777–793, 1994.
33. Gandini, D., Gogioso, L., Bolognesi, M., Bordo, D. Patterns in ionizable side chain interactions in protein structures. Proteins 24:439–449, 1996.
34. Tronrud, D.E. Knowledge-based B-factor restraints for the refinement of proteins. J. Appl. Crystallogr. 29:100–104, 1996.
35. Schlichtig, I., Berendzen, J., Phillips, G.N. Jr, Sweet, R.M. Crystal structure of photolysed carbomonoxy myoglobin. Nature 371:808–810, 1994.
36. Sato, T., Tanaka, N., Moryiama, H., et al. Authors of the PDB file. 1993.
37. Eriksson, A.E., Baase, W.A., Wozniak, J.A., Matthews, B.W. The cavity-containing mutant of T4 lysozyme is stabilized by buried benzene. Nature 355:371–373, 1992.
38. Morton, A., Baase, W.A., Matthews, B.W. Energetic origins of specificity of ligand binding in an internal nonpolar cavity of T4 lysozyme. Biochemistry 34:8564–8575, 1995.
39. Morton, A., Matthews, B.W. Specificity of ligand binding in a burial nonpolar cavity of T4 lysozyme: Linkage of dynamics and structural plasticity. Biochemistry 34:8576–8588, 1995.
40. Antonini, E., Brunori, M. "Hemoglobin and Myoglobin in Their Reactions With Ligands."Amsterdam: Elsevier/North-Holland, 1971.
41. Kuriyan, J., Wilz, S., Karplus, M., Petsko, G.A. X-ray structure and refinement of carbo-monoxy (Fe(II))-myoglobin at 1.5 angstroms resolution. J. Mol. Biol. 192:133–154, 1986.
42. Austin, R.H., Beeson, K.W., Eisenstein, L., Frauenfelder, H., Gonsalus, I.C. Dynamics of ligand binding to myoglobins. Biochemistry 14:5355–5373, 1975.
43. Trueblood, K.N. Diffraction studies of molecular motion in crystals. In: "Accurate Molecular Structures." Domenicano, A., Hargittai, I. (eds.). Oxford, U.K.: Oxford University Press, 1992:199–219.
44. Cruickshank, D.W.J. The analysis of the anisotropic thermal motion of molecules in crystals. Acta Crystallogr. 9:754–756, 1959.
45. Schomaker, V., Trueblood, K.N. On the rigid–body motion of molecules in crystals. Acta Crystallogr. B24:63–76, 1968.
46. Janin, J., Rodier, F. Protein–protein interaction at crystal contacts. Proteins 23:580–587, 1995.
47. Carugo, O., Argos, P. Protein crystal packing contacts. Protein Sci. 6:2261–2263, 1997.
48. Walshaw, J., Goodfellow, J.M. Distribution of solvent molecules around apolar side-chains in protein crystals. J. Mol. Biol. 231:392–414, 1993.
49. Levitt, M., Park, B.H. Water: Now you see it, now you don't. Structure 1:223–226, 1993.
50. Wilmanns, M., Nilges, M. Molecular replacement with NMR models using distance-derived pseudo B factors. Acta Crystallogr. D52:973–982, 1996.
51. Iwata, S., Ostermeier, C., Ludwig, B., Michel, H. Structure at 2.8 Å resolution of cytochrome c oxidase from *Paracoccus denitrificans*. Nature 376:660–669, 1995.
52. Tsukihara, T., Aoyama, H., Yamashita, E., et al. The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å. Science 272:1136–1144, 1996.
53. Clarke, J., Fersht, A.R. An evaluation of the use of hydrogen exchange at equilibrium to probe intermediates on the protein folding pathway. Folding Design 1:243–254, 1996.
54. Eriksson, A.E., Baase, W.A., Matthews, B.W. Similar hydrophobic replacements of leu 99 and phe 153 within the core of T4 lysozyme have different structural and thermodynamic consequences. J. Mol. Biol. 229:747–769, 1993
55. Rose, D.R., Phipps, J., Michmiewicz, J., et al. Crystal structure of T4-lysozyme generated from synthetic coding DNA expressed in *Escherichia coli*. Protein Eng. 2:277–282, 1988.
56. Weaver, L.H., Matthews, B.W. Structure of bacteriophage T4 lysozyme refined at 1.7 angstroms resolution. J. Mol. Biol. 193:189–199, 1987.
57. Wilson, K., Fabe, R., Dao-Pin, S., Matthews, B.W. Authors of the PDB file, 1989.
58. Bell, J.A., Wilson, K., Zhang, X.J., Faber, H.R., Nicholson, H., Matthews, B.W. Comparison of the crystal structure of bacteriophage T4 lysozyme at low, medium, and high ionic strengths. Proteins 10:10–21, 1991.
59. Phillips, S.E.V., Schoenborn, B.P. Neutron diffraction reveals oxygen-histidine hydrogen bond in oxymyoglobin. Nature 292:81–82, 1981.
60. Yang, F., Phillips, G.N. Jr. Crystal structure of CO-, deoxy-and Met-myoglobins at various pH values. J. Mol. Biol. 256:762–774, 1996.
61. Takano. T. Refinement of myoglobin and cytochrome C. In: "Methods and Applications in Crystallographic Computing." Hall S.R., Ashida T. (eds.). Oxford, U.K.: Oxford University Press, 1984:262–289.
62. Phillips, S.E.V. Structure and refinement of oxymyoglobin at 1.6 angstroms resolution. J. Mol. Biol. 142:531–554, 1980.
63. Rizzi, M., Ascenzi, P., Coda, A., Brunori, M., Bolognesi, M. Molecular bases for heme:ligand recognition in sperm whale (*Physeter catodon*) and *Aplysia limacina* myoglobin. Rend. Fis. Ac. Lincei S9:65–73, 1993.
64. Johnson, K.A, Olson, J.S., Phillips, G.N. Jr. Structure of myoglobin-ethyl isocyanide: Histidine as a swinging door for ligand entry. J. Mol. Biol. 207:459–463, 1989.