Journal of Molecular Graphics and Modelling 27 (2009) 889–899

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Free-energy landscape, principal component analysis, and structural clustering to identify representative conformations from molecular dynamics simulations: The myoglobin case

Elena Papaleo [a,*], Paolo Mereghetti [a,b], Piercarlo Fantucci [a], Rita Grandori [a], Luca De Gioia [a]

[a] Department of Biotechnology and Bioscience, University of Milano-Bicocca, 20126 Milan, Italy
[b] Department of Chemistry, University of Sassari, 07100 Sassari, Italy

ARTICLE INFO

Article history:
Received 22 December 2008
Received in revised form 27 January 2009
Accepted 27 January 2009
Available online 6 February 2009

Keywords:
Molecular dynamics
Conformational sampling
Myoglobin
Structural clustering
Free-energy landscape

ABSTRACT

Several molecular dynamics (MD) simulations were used to sample conformations in the neighborhood of the native structure of holo-myoglobin (holo-Mb), collecting trajectories spanning 0.22 μs at 300 K. Principal component (PCA) and free-energy landscape (FEL) analyses, integrated by cluster analysis, which was performed considering the position and structures of the individual helices of the globin fold, were carried out. The coherence between the different structural clusters and the basins of the FEL, together with the convergence of parameters derived by PCA indicates that an accurate description of the Mb conformational space around the native state was achieved by multiple MD trajectories spanning at least 0.14 μs.

The integration of FEL, PCA, and structural clustering was shown to be a very useful approach to gain an overall view of the conformational landscape accessible to a protein and to identify representative protein substates. This method could be also used to investigate the conformational and dynamical properties of Mb apo-, mutant, or delete versions, in which greater conformational variability is expected and, therefore identification of representative substates from the simulations is relevant to disclose structure–function relationship.

1. Introduction

Myoglobin (Mb) is a small hemo-protein involved in oxygen transport and storage, which has been widely used for functional and structural studies [1–4]. Mb belongs to the globin family, which is characterized by a globular fold comprising eight α-helices (A to H, from the N- to the C-terminus) linked to each other by short loop regions [1]. Over the last decades, a number of new globins with different active-site structures has been discovered and characterized, providing interesting *variants on the theme* [5–7]. Moreover, the possibility of reversible heme removal and the large number of mutant and deleted versions of Mb available [8–15] has extended considerably the appeal of myoglobin for structure and dynamics investigation [2–3].

It is well known that Mb undergoes functionally relevant conformational transitions upon ligand binding, pH variation, but also in the native structure, which has been analyzed by IR spectroscopy [16,17], NMR [18,19], electrospray mass spectrometry [20,21], time-resolved crystallography [22,23] and, molecular dynamics simulations [24–28]. Most of these studies converge to the conclusion that full insight into conformational changes of Mb necessitates a detailed description of the conformational ensemble of the protein in the native state [29]. In this context, it is important to underline that proteins are very complex systems and that the energy surface describing the native state contains multiple minima corresponding to very similar, but slightly different conformations, or conformational substates [30–33].

Proper sampling of protein energy landscape requires the generation of a large sample of molecular conformations and computational methods can be applied to this task [33–35]. However, analysis of the conformational space, even for a relatively small molecule, is very demanding due to the extremely high dimensionality of the systems [34]. Moreover, a limitation common to all computational sampling procedures is that, although they generate large conformation ensembles, it is hard to assess the extent of sampling [34,36,37].

Multiple molecular dynamics (MD) simulations are well suited for generating ensembles of structures [38–42] at a fixed temperature, and, in fact, they have been successfully used to investigate the conformational landscape near a protein native state [38,39]. However, it has been observed that independent trajectories of the same system can sample the same local region

* Corresponding author.
E-mail address: elena.papaleo@unimib.it (E. Papaleo).

1093-3263/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2009.01.006

although following distinct paths, whereas in other cases some trajectories sample more than one of the major local regions for significant time periods. Thus, extending the trajectories for a limited time period does not guarantee more extensive sampling [38,40]. In addition, the ensemble average structure is typically taken as representative of the trajectory snapshots, even if, in the case of flexible structures, the average structure may not be adequately representative of the ensemble [36]. In fact, few studies have been devoted to a detailed analysis of the contribution of individual protein trajectories to the conformational sampling, and to the identification of representative structures from MD ensembles [34,43–45]. Therefore, care has to be used since the average structures resulting from independent trajectories could be very different, and therefore the identification of conformations belonging to low energy basins becomes crucial.

MD simulations can provide an atomic-level picture of protein motion and a representation of the free-energy landscape (FEL) [46,47] close to the protein native state [43,44]. However, since it is not possible to represent the FEL as a function of 3N-6 coordinates, it is essential to choose an appropriate set of reaction coordinates that allows to distinguish among different conformational substates. Reaction coordinates are usually obtained by Principal Component Analysis (PCA), which describes the largest amplitude protein motions during a simulation [34,35,43,48–50]. Once the reaction coordinates are chosen, the FEL can be derived from the probability density function [47,51].

In fact, since the FEL is projected on a low-dimensional space, it is relevant to verify that conformations that map to the same free-energy basin are also characterized by similar three-dimensional structures. Therefore a further approach for the analysis of the conformational sampling is offered by structural family clustering, a method that collects together conformations according to geometric similarity [34,51,52].

In order to explore how to achieve an adequate sampling of the native conformational space through MD simulations and to derive representative structures from the MD ensemble, we performed a set of eleven independent MD simulations at 300 K to sample the conformational space of holo-myoglobin (holo-Mb) in the neighborhood of the native structure, collecting trajectories spanning 0.22 μs. In particular, principal component and FEL analyses, integrated by several cluster analyses, were carried out to gain an overall view of the conformational space accessible to this globular protein. It turns out a high coherence among the different clusters and the basins of the FEL and convergence of parameters derived by PCA, which is not achieved by lower conformational sampling. In particular, our results suggest that multiple MD trajectories spanning 140–160 ns of simulations are sufficient to ensure an adequate sampling of a protein characterized by the globin fold, as myoglobin.

2. Materials and methods

2.1. MD simulations

MD simulations were performed using the 3.3 version of GROMACS software (www.gromacs.org), implemented on a parallel architecture, using GROMOS96 forcefield, which was used in previous MD studies of myoglobin [53] and neuroglobin [54] providing results in excellent agreement with the experimental data. The X-ray structure of the native holo-Mb (PDB entries 1A6N [55], and 1CQ2 [56]), were used as starting points for the MD simulations. Protein structures, including crystallographic water molecules and heme cofactor, were soaked in a dodecahedral box of SPC (Single Point Charge) water molecules [57] and simulated using periodic boundary conditions. All the protein atoms are at a distance equal or greater than 0.5 nm from the box edges. The ionization state of residues was set to be consistent with neutral pH and the tautomeric form of histidine residue was derived using GROMACS tools and confirmed by visual inspection and according to the experimental evidences by Bhattacharya et al. [58]. Further details on the MD setup and solvent equilibration, termalization and pressurization steps are reported in ref. [41].

Productive MD simulations were performed in the isothermal-isobaric (NPT) ensemble at 300 K, using an external bath with a coupling constant of 0.1 ps. Pressure was kept constant (1 bar) by modifying the box dimensions and the time-constant for pressure coupling was set to 1 ps. The LINCS [59] algorithm was used to constrain bond lengths, allowing the use of 2 fs time step. Electrostatic interactions were calculated using the Particle-mesh Ewald (PME) summation scheme [60]. Van der Waals and Coulomb interactions were truncated at 1.0 nm. The non-bonded pair list was updated every 10 steps and conformations were stored every 2 ps.

To improve the conformational sampling, eleven independent 20 ns simulations (*replicas*) were carried out initializing the MD runs with different initial atomic velocities or using different initial structures (Table 1). More in detail, an iterative procedure was pursued, starting from six *replicas* (*replicas 1–6*) for Holo-Mb (pdb entry, 1A6N), using PCA and FEL analysis of the corresponding concatenated trajectory (*total 1–6*) and increasing therefore the number of *replicas* by new independent MD simulations either starting from structures in the minimum energy pits obtained by FEL analysis (*total 1–8*) or starting from another crystallographic structure (pdb entry 1CQ2) (*total 1–11*).

2.2. Analysis of simulations

The mainchain root-mean-square-deviation (rmsd), which is a crucial parameter to evaluate trajectory stability, was computed using the starting structures of the MD simulations as a reference.

Table 1
Summary information about the different *replicas* for holo-Mb simulations.

	Duration (ns)	Starting Structure
Replica 1	20	1A6N X-ray
Replica 2	20	1A6N X-ray
Replica 3	20	1A6N X-ray
Replica 4	20	1A6N X-ray
Replica 5	20	1A6N X-ray
Replica 6	20	1A6N X-ray
Replica 7	20	Structure from the minimum energy basin of *replica* 5
Replica 8	20	Structure from the minimum energy basin of *replica* 5
Replica 9	20	1CQ2 X-ray
Replica 10	20	1CQ2 X-ray
Replica 11	20	1CQ2 X-ray
Total_*replicas* 1–6	114	Concatenated trajectory including the first 6 *replicas* from 1A6N
Total_*replicas* 1–8	152	Concatenated trajectory including the 8 *replicas* from 1A6N
Total_*replicas* 1–11	209	Concatenated trajectory including the first 8 *replicas* from 1A6N and the 3 *replicas* from 1CQ2

For each protein, the equilibrated portions of each *replica* were joined together to obtain concatenated trajectories (Table 1), which are representative of different directions of sampling around the starting structure. The secondary structures were calculated using DSSP [61]. The visual analysis of protein structure was carried out using VMD (www.ks.uiuc.edu/Research/vmd) and Pymol (www.pymol.org).

2.3. Principal component analysis

PCA reveals high-amplitude concerted motion in MD trajectories, through the eigenvectors of the mass-weighted covariance matrix (C) of the atomic positional fluctuations [35], which was calculated on protein alpha carbons (Cα) using concatenated and single trajectories.

To define the dimensionality of the essential subspace, the fraction of total motion described by the reduced subspace was considered and computed as the sum of the eigenvalues relative to the included eigenvectors, describing the amount of variance retained by the reduced representation of the total space.

The cosine content (c_i) of the principal component (p_i) of C is an absolute measure that can be extracted from covariance analysis and ranges between 0 (no cosine) and 1 (perfect cosine):

$$c_i = \frac{2}{T}\left(\int \cos(i\pi t)\, p_i(t)\, \mathrm{d}t\right)^2 \left(\int p_i^2(t)\, \mathrm{d}t\right)^{-1}$$

where T is the total simulation time. It has been demonstrated that insufficient sampling could lead to high c_i values, representative of random motions [62]. The evaluation of cosine contribution for the first eigenvectors is sufficient to give a reliable idea of the protein behavior. When the cosine content of the first few PCs is close to 1, the largest scale motions in the protein dynamics resemble diffusion, and cannot be interpreted in terms of characteristic features of the energy landscape [50]. The cosine content was calculated on the first 20 principal components of each single *replica* and on the *total 1–11* trajectory (Fig. S2). Moreover, all the single *replica*s were randomly concatenated in several concatenated trajectories of different lengths. Therefore, the average cosine content for the first principal component (PC 1) was calculated as a function of the trajectory duration, allowing the evaluation of convergence of the conformational sampling (Fig. 1). The analysis of the sampling convergence can be performed computing the root mean square inner product (rmsip) as a measure of similarity between subspaces defined by their basis vectors [63]:

$$\mathrm{RMSIP} = \frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{D}(\eta_i^A \eta_j^B)$$

where η_i^A and η_i^B are the eigenvectors of the spaces to be compared and D the number of eigenvectors considered. Usually the rmsip is computed on the first 10 eigenvectors [63,64].

2.4. Free-energy landscape

The free-energy landscape of a protein can be obtained using a conformational sampling method that allows to explore the conformations near the native state structure. Here, the MD simulation technique was adopted as the sampling technique. To achieve a two-dimensional representation of the FEL, we define the probability of finding the system in a particular state characterized by a value q_α of some variables of interest (*reaction coordinate*) as proportional to ($e^{-G\alpha/kT}$) where $G\alpha$ is the free energy of the state. The FEL can be obtained from:

$$G\alpha = -kT \ln\left[\frac{P(q_\alpha)}{P_{\max}(q)}\right]$$

where k is the Boltzmann constant, T is the temperature of simulation, $P(q_\alpha)$ is an estimate of the probability density function obtained from a histogram of the MD data and $P_{\max}(q)$ is the probability of the most probable state. Considering two different reaction coordinates q_i and q_j. The two-dimensional free-energy landscapes were obtained from the joint probability distributions $P(q_i,q_j)$ of the system.

2.5. Root-mean-square-deviation (rmsd) matrices and cluster analysis

The rmsd matrices have been computed on the concatenated trajectory *total 1–11* (Table 1), by least square fitting on mainchain atoms. Several rmsd matrices were generated, which differ for the atom subsets used for the rmsd calculation. In particular, 9 rmsd matrices were created: one relative to the full-length protein chain

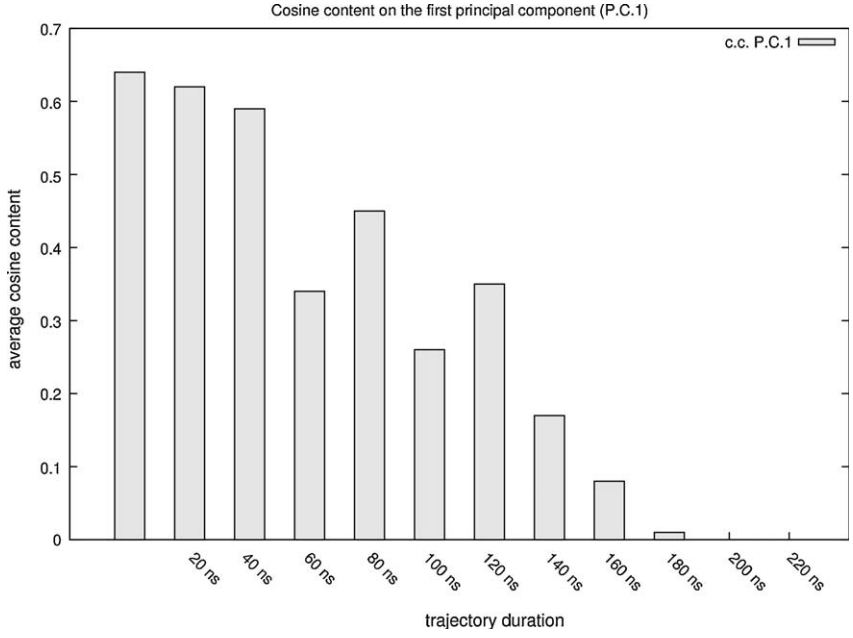

Fig. 1. Cosine content. Average cosine content of the first principal component calculated as a function of different concatenated trajectory lengths.

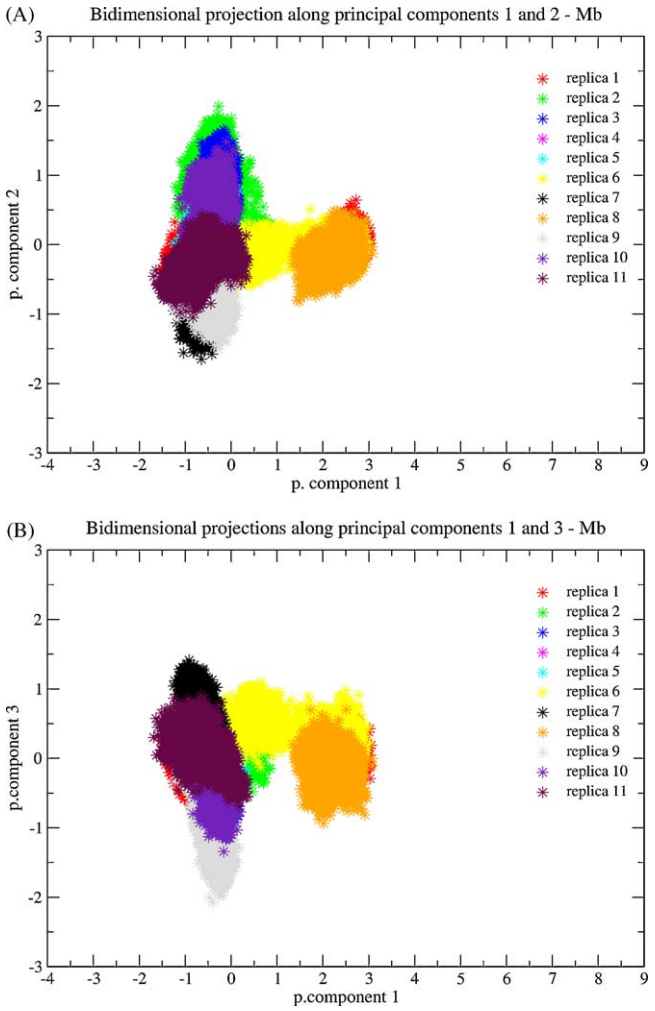

Fig. 2. Bidimensional projections of the principal components. (A) The projection of the concatenated trajectory *total 1–11* along the 1st and the 2nd principal components. (B) The projection of the concatenated trajectory *total 1–11* along the 1st and the 3rd principal components.

and eight relative to the individual helices of the globin fold. The amino acids belonging to each α-helix were defined according to DSSP definition applied to the holo-Mb X-ray structure (pdb entry 1A6N).

The rmsd matrices have been then processed using both the Linkage [65] and Gromos [66] algorithms to extract clusters of similar conformations. Different rmsd cutoffs were tested and adopted for cluster analysis and, a value close to the average rmsd value derived from each rmsd matrix was selected. In particular, the cutoff adopted for Gromos clustering was 0.20 nm for all the rmsd matrices, with the exception of the rmsd matrix derived using the helix F atom subset (0.25 nm), whereas for the Linkage clustering a cutoff of 1.05 nm was adopted. The Gromos and Linkage algorithm gave similar results, differing only in the number of structures in the less populated clusters. Therefore only the results of Gromos clustering are reported. The average structure of the clusters is defined as the protein structure with the lowest average distance (rmsd) to all other structures belonging to the same cluster.

Moreover, different *ensembles* of structures were collected, comparing the results of cluster analysis based on the 9 rmsd matrices. In particular, the minimum consensus region in which structures which cluster together according to all the 9 atom subsets (i.e. according to the clustering carried out on the different rmsd matrices) was determined by comparison of cluster distribution along the concatenated *total 1–11* trajectory. Only clusters which collected at least 10% of the total structures were considered in order to filter out from the concatenated trajectory those structures which are not frequently sampled during the simulations.

3. Results and discussion

3.1. Stability of the simulations

We performed several independent simulations (*replicas*) for holo-Mb in order to increase the conformational sampling, using an iterative procedure as explained in Section 2. The analyses of the MD trajectories have been carried out discarding the portions of the *replicas* required by the system to reach stable values of

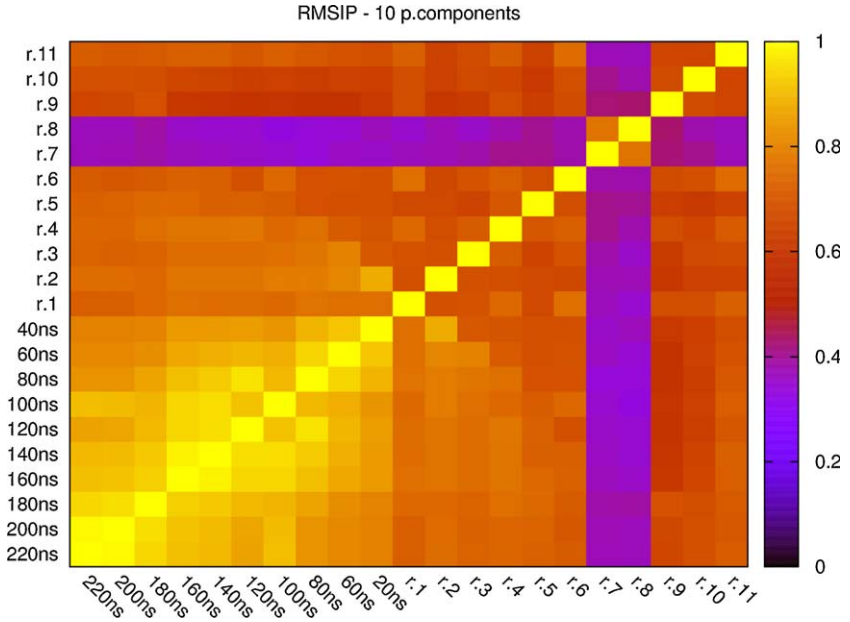

Fig. 3. Rmsip matrix. The matrix shows with different shade of colors the rmsip values calculated on holo-Mb simulations, considering the conformational subspace described by the first 10 principal components. Concatenated trajectories of different lengths (as in Fig. 1) and the single 20 ns-*replicas* ('r.') were used for the calculation.

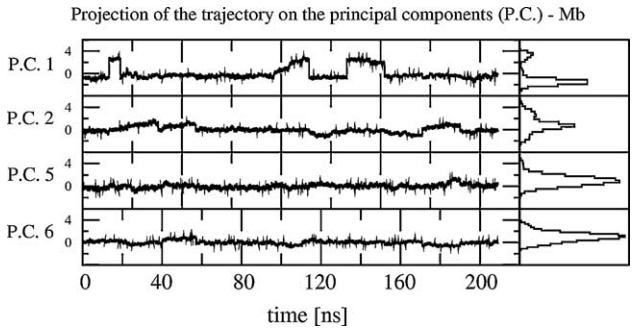

Fig. 4. Projection of the concatenated trajectory *total 1–11* of holo-Mb on the 1st, 2nd, 5th and 6th principal components (P.C.). For each projection, the corresponding time course along the simulation and the distribution of states along the PC are plotted.

mainchain rmsd and of gyration radius (about 1 ns), in order to ensure that calculated parameters reflect the intrinsic properties of the system (Fig. S1, Supplementary).

3.2. Evaluation of the conformational sampling

Since MD simulations are necessarily limited in time, conformations visited are only a subset of all the possible conformations that the protein can assume. In order to correlate the MD data with protein characteristics, one should ensure a sufficiently high sampling efficiency. To this aim, the PCA analysis is a suitable tool to provide information about conformational sampling.

When the sampling of MD trajectories is insufficient, protein motions along the principal components appear indistinguishable from the dynamics of random diffusion, not allowing an accurate description of the free-energy landscape [56]. In particular, the first principal components of random diffusion resemble a cosine function with a number of periods equal to half of the principal component index, and the sampling along these directions does not describe relevant motions [62]. We calculated the cosine content of the first 20 principal components obtained by PCA analysis of the single *replicas* and of the concatenated trajectory *total 1–11* (Fig. S2, Supplementary). The cosine content of the three first principal components is lower in the concatenated trajectory, confirming that a large number of MD simulations is an effective strategy to get a reliable conformational sampling. Moreover, the calculation of the cosine content as a function of trajectory duration (Fig. 1) points out that 140–160 ns, obtained concatenating 7–8 *replicas*, are required to get convergence of cosine content values.

PCA on single *replicas* of holo-Mb, shows that, in general, about 15 eigenvectors are required to describe more than 70% of the total variance [35], whereas, in the concatenated trajectories, this value is decreased by two-fold (data not shown). Most of the variance is generally related to the first three principal components, which, in the case of holo-Mb simulations, account for a major fraction of the total motion in the concatenated trajectories, but explain a

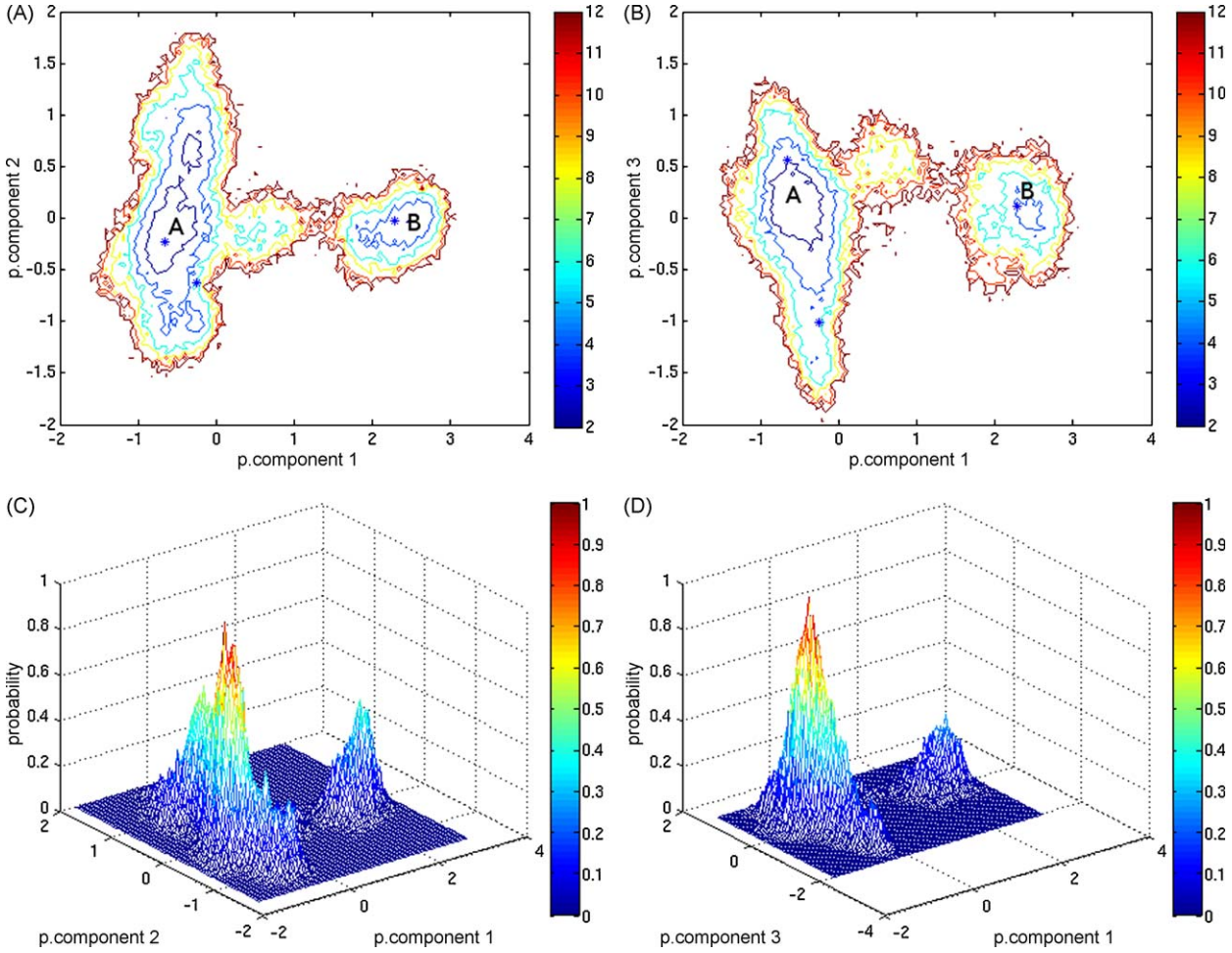

Fig. 5. Free-energy landscape. FEL (A,C) and probability (B,D) using as reaction coordinates the projection of the Mb concatenated trajectory *total 1–11* along the 1st–2nd (A,B) and 1st–3rd (C,D) principal components. Asterisks (*) indicate the localization on the FEL of the average structures of the different conformational *ensembles* (*A* and *B*) derived from the cluster analysis. The free energy is given in kJ/mol and indicated by color.

minimal part of the total motion if a single *replica* is considered (Fig. S3, Supplementary). A reduction in the number of the principal components required to describe a significant portion of the total variance is indicative of a wider sampling of the configurational space in the concatenated trajectories.

Generally, in order to get a reliable sampling, a wide region of the conformational space should be sampled and, at the same time, a partial overlap between different trajectories should be achieved [49]. Holo-Mb trajectories populate in particular two distinct regions, which are re-sampled by different *replicas* (Fig. 2).

The projection of the simulation frames in the 3D subspace defined by the first three principal components is only a qualitative index. To better investigate the subspace overlap (i.e. the overlap among sampled regions of the subspace) it is necessary to apply other analyses. In particular, a comparison of the motions described by the first 10 principal components of each simulation (see Section 2) can be carried out by calculating the rmsip. Rmsip ranges from 0 to 1: the value is 1 if the sampled subspaces are identical, and 0 if the sampled subspaces are completely orthogonal. This parameter was calculated for all pairwise comparisons of the single *replicas* and of the concatenated trajectories (Fig. 3), confirming the previous analyses and the existence of two main conformational regions of the 3D subspace defined by the first three eigenvectors (Fig. 2). In fact, it turned out that the rmsip value is greater than 0.6 (Fig. 3) when *replicas* mapping to the same region of the 3D subspace (Fig. 2) are compared, whereas rmsip values lower than 0.4 (Fig. 3) result from comparison of *replicas* sampling the two different subspaces (Fig. 2). Rmsip values confirm data from cosine content (Fig. 1): as indicated by Fig. 3, 160 ns are required in order to definitely ensure high values of overlap.

Moreover, the distribution of the motion projections along each of the first six principal components was calculated (Fig. 4), showing that the first three principal components are sufficient to describe the most relevant motions. In fact, the distribution of motion along the principal components tends to be anharmonic with two or more peaks on the first directions and a narrow Gaussian shape on the later ones. The projections along the first three principal components clearly gave a multimodal distribution for Mb *total 1–11* concatenated trajectory with two main peaks, whereas moving to the fourth or fifth principal components the distributions tend to be similar to a Gaussian curve (Fig. 4). Therefore, the subspace defined by the first three components can be used as the reference conformational subspace, in which to analyze protein dynamics if the concatenated trajectory is considered.

3.3. Free-energy landscape

After the evaluation of the extent of the configurational space explored by the simulations, it is possible to attempt an accurate and reliable description of the free-energy landscape. This calculation was based on the first three principal components: the bidimensional projections of principal component 1 *vs.* principal components 2 (Fig. 2A) or 3 (Fig. 2B) were analyzed.

The free-energy profiles in the two bidimensional projections are very similar and feature two main energy basins (Fig. 5) differing by about 2 kJ/mol. Moreover, the projection of the single

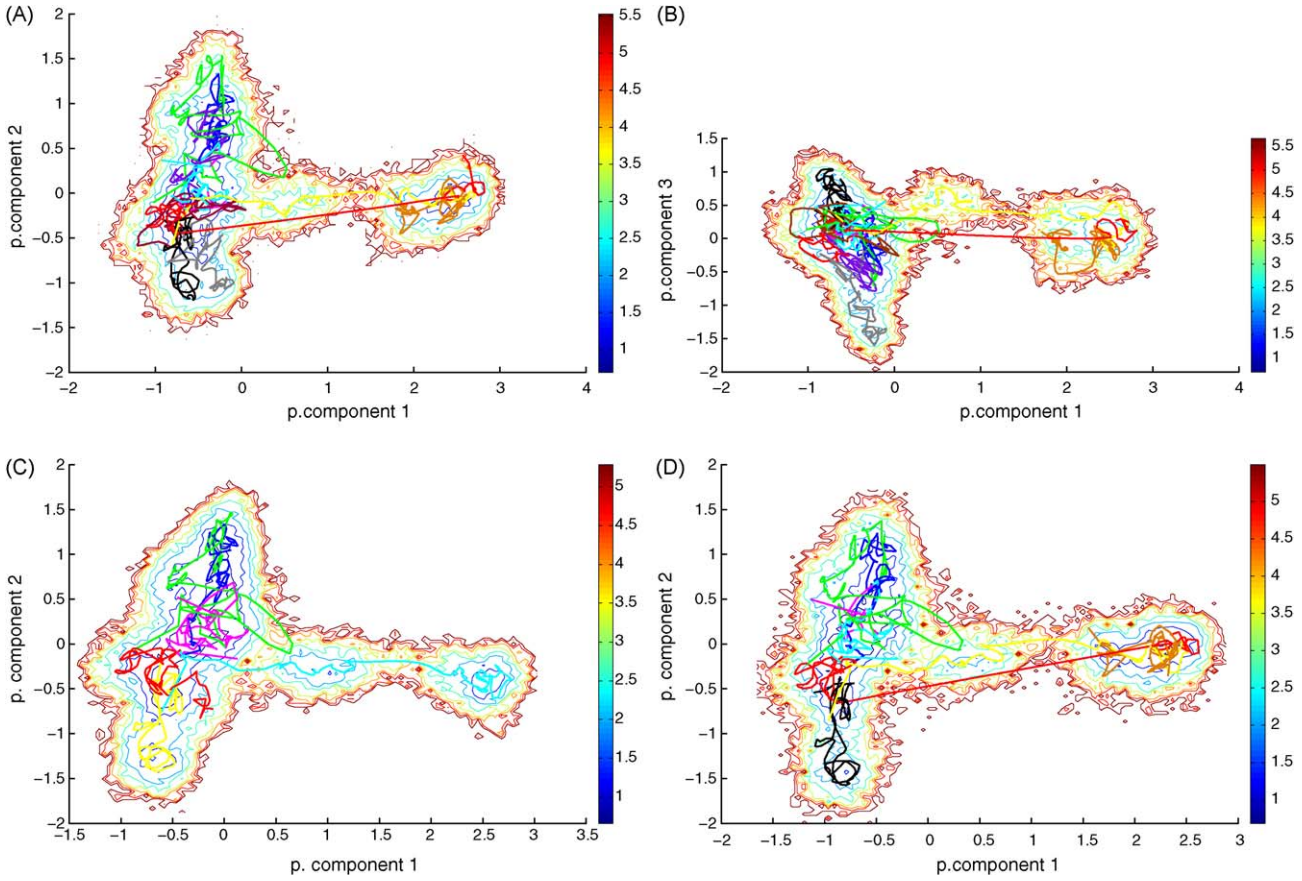

Fig. 6. Projection of the single *replicas* on the FEL. The trajectories explored by each single *replicas* are projected on the FEL using as reaction coordinates the projection of the Mb concatenated trajectory *total 1–11* along the 1st–2nd (A) and 1st–3rd (B) principal components. The projection of the Mb concatenated *total 1–6* (C) and *total 1–8* (D) trajectories along the 1st–2nd principal components are also shown. The frames belonging to each *replica* are indicated by different colors according to the color code of Fig. 2. The free energy is given in kJ/mol and indicated by color.

replica trajectories on the FEL allows to better highlight conformational transitions between the two basins (Fig. 6A and B). The regions of the conformational space corresponding to these two basins will be referred to as region A and B: structures from *replicas* 2, 3, 4, 5, 7, 9, 10, 11 populate the region A, whereas structures from *replica* 8 are localized in region B, and structures from *replicas* 1 and 6 span both regions.

Because of the large reduction in dimensionality when a two-dimensional free-energy profile is built from the multidimensional dynamics, the FEL plots present an incomplete picture of the protein conformational scenario. In this context, cluster analyses based on rmsd matrices, which were derived by the original MD trajectories, can potentially provide useful further information [44].

3.4. Rmsd matrices and cluster analysis

Measures of similarity among protein conformations to be used for cluster analysis are provided by rmsd matrices (Fig. 7), comparing structures of the concatenated trajectory *total 1–11*. Rmsd were calculated on the mainchain atoms of the full-length protein chain or of the individual helices (Section 2). The rmsd matrices show re-sampling of similar conformations for most of the helices, with main differences localized in helices F and H (Fig. 7).

The Mb rmsd matrices were processed to extract clusters of conformations (see Section 2) and the distribution of structures in the clusters along the concatenated trajectory *total 1–11* was then evaluated (Fig. S4, Supplementary). The analysis based on rmsd matrices of helices A, B, C, D, E and G (Fig. 7B, Fig. S4B) gave a single

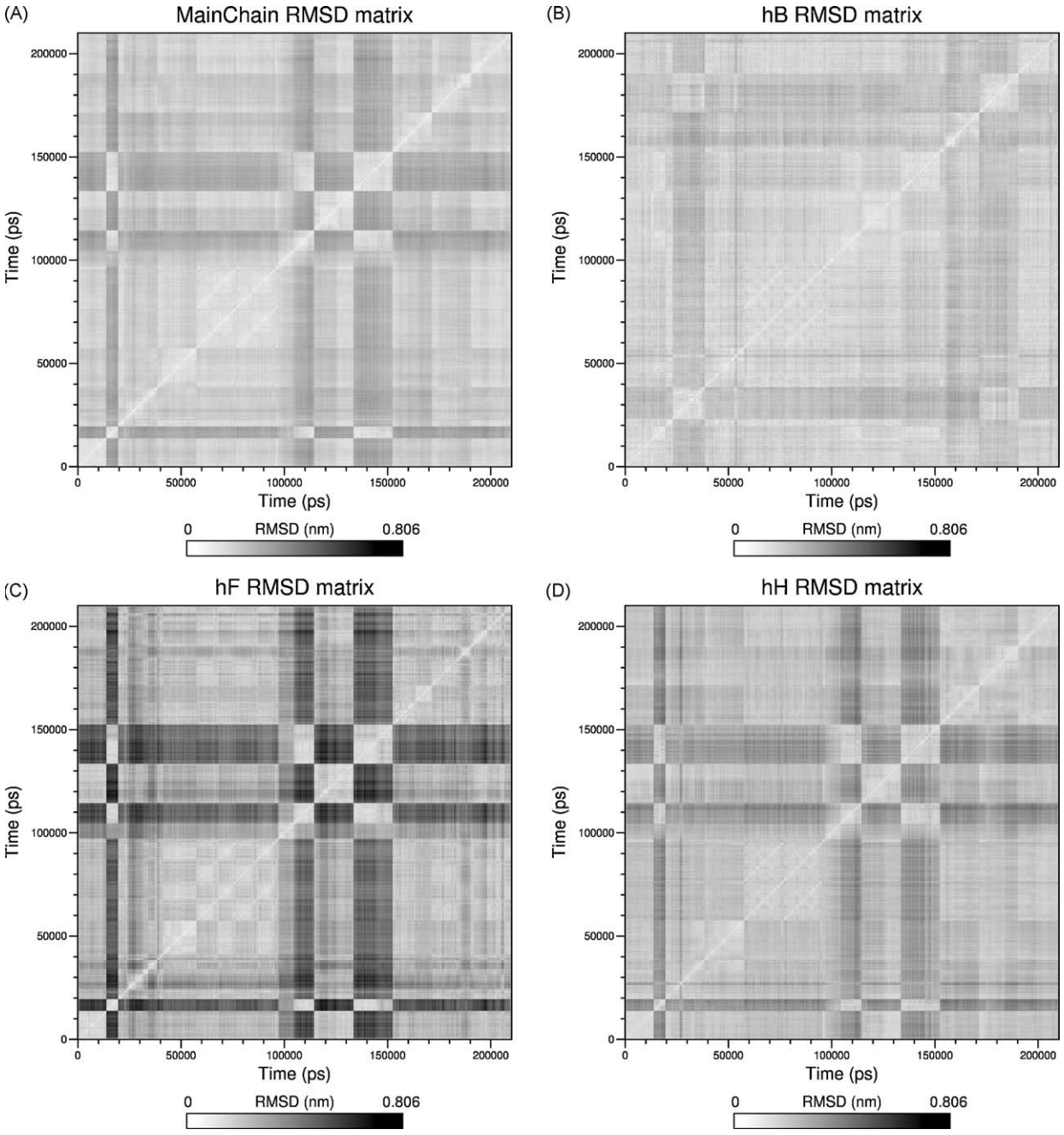

Fig. 7. Rmsd matrices relative to the comparison among the structures of the concatenated trajectory *total 1–11*. The rmsd values were calculated on the mainchain atoms of distinct residues substed in each matrix: (A) full-length protein chain, (B) helix B, (C) helix F, (D) helix H. Data for helices A, C, D, E, G were similar to those for helix B and therefore not reported in the figure for sake of clarity.

cluster, with only a few scattered configurations lying outside the cluster. On the contrary, cluster analysis on the full-length chain (Fig. 7A, Fig. S4A), helix F (Fig. 7C, Fig. S4C) and helix H (Fig. 7D, Fig. S4D) allowed to identify two main clusters: differences in helices F and H strictly reflect global structural differences.

The first main cluster comprises structures from the first part of *replicas 1* and *6* and *replicas* 2, 3, 4, 5, 7, 9, 10, and 11, which correspond to conformations localized in the region A identified by FEL analysis (Figs. 5 and 6). The second cluster is populated by conformations from the last part of *replicas 1* and *6* and structures from *replica 8*, which fit to region B (Figs. 5 and 6). The rmsd matrix analysis indicates that helix E, which includes the proximal histidine H64, and helix B have the most stable conformations during the Mb trajectories. Moreover, between the two conformational basins identified by PCA, structures vary mostly at the level of the helices H and F.

Cluster analysis combined to FEL description allowed to establish whether conformations that belong to the same energy basin are similar and, therefore, whether a correlation exists between structural similarity and energy basins in the sampled trajectories, enforcing evidences that an adequate sampling of the Mb native conformational space was achieved by our simulations.

3.5. Identification of representative structures in the trajectories

Once the quality of the conformational sampling was evaluated, representative structures could be extracted from the trajectories. In fact, the structures of the configurational space of Mb arise from the regular arrangement of relatively rigid α-helices [25]. Moreover, it was observed that even minute changes in the dynamical properties of the system alter the configurational space projection [25]. Therefore, we assigned to the same cluster those conformations that share most of their secondary structure motifs, i.e. that share the position and structure of the helices (see Section 2) (Fig. 7, Fig. S4).

Therefore, two different *ensembles* corresponding to the main clusters and the FEL basins previously identified were generated and they are called *ensemble A* and *B*. In particular, the *ensemble A* contains structures relative to *replica 1* (structures in the interval from 2 to 13 ns), *replica 2* (2–20 ns), *replica 3* (3–19 ns), *replica 4* (5–16 ns), *replica 5* (5–18 ns), *replica 7* (4–20 ns), *replica 9* (2–20 ns) *replica 10* (1–19 ns) and *replica 11* (2–20 ns). The ensemble B collects structure from *replica 1* (14–20 ns), *replica 6* (11–19 ns) and *replica 8* (9–20 ns). Structures which are assigned to the same ensemble were linked together in order to create the corresponding *ensemble trajectories A* and *B*. An average structure was extracted from each *ensemble*.

Moreover, the average structures were evaluated considering their position on the FEL (Fig. 5A–C). The average structure from *ensemble A* (Fig. 9A) and B (Fig. 9B) corresponds to the global and second lower minimum of the FEL, respectively.

The average structures from the two ensembles are very similar, differing mainly for the conformations of the helices F and H, and the region between helices C and D where a helical motif appears in the *ensemble A* (Fig. 9). Interestingly, helix F is an important region of the Mb fold since it includes the proximal histidine residue and its folding is induced by the interaction with the prosthetic group [67]. The identification of two FEL basins characterized by slightly different conformations of helix F highlights the intrinsic flexibility of this region even when the prostetic group is bound. It has been previously suggested that protein motions involving helix F displacement play a crucial role in reversible oxygen binding by globins [68,69], as well as in the interaction with carbon monoxide [70].

3.6. Comparison between FEL and cluster analysis

In order to further compare the structural clustering and the clustering based on FEL analysis, some sample structures from the two *ensembles* (Fig. 8A) and some outlier structures from the clustering procedure (Fig.8A) were mapped on the FEL. Structures which belong to *ensemble trajectories A* and *B* mapped in the two main energy basins, whereas structures assigned to poorly populated clusters, and therefore not included in the representative *ensemble trajectories*, are characterized by relatively high free-energy values.

In fact, since the FEL is projected on the low-dimensional space described by the first principal components, the coherence between cluster analysis and FEL is relevant to ensure the establishment of convergence in the conformational sampling. If the same analyses presented for the *total 1–11* trajectory are applied to shorter concatenated trajectories in which PCA does not reach convergence, this correlation between cluster analysis and FEL does not exist since only low-resolution free-energy profiles are obtained. For instance, the structures from *replica 1* undergo conformational transitions between *ensemble*s A and B (Fig. 6A), due to a displacement of helix F (Figs. 6 and 9). If the concatenated *total 1–6* trajectory is analyzed by PCA and FEL is represented by the first three principal components, the conformations from *replica* 1 remain constrained in the same region of the low-defined 3D essential subspace (Fig. 6C). If the same analysis is carried out

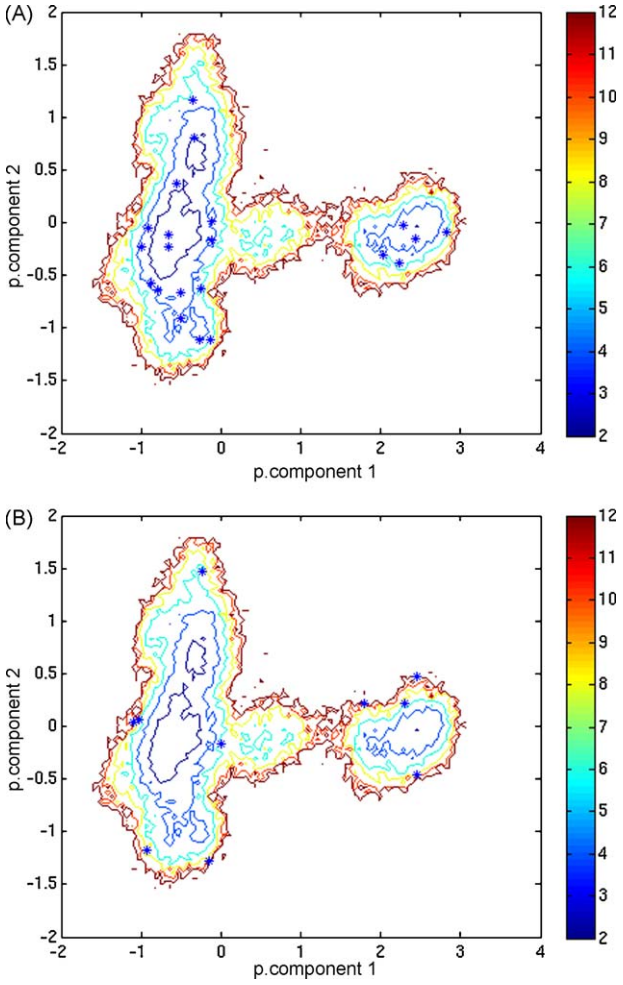

Fig. 8. FEL along projections onto the 1st and 2nd principal components of the holo-Mb concatenated trajectory *total 1–11*. Asterisks (*) indicate the localization on the FEL of some average structures extracted from the highly (A) or poorly (B) populated clusters. The free energy is given in kJ/mol and indicated by color.

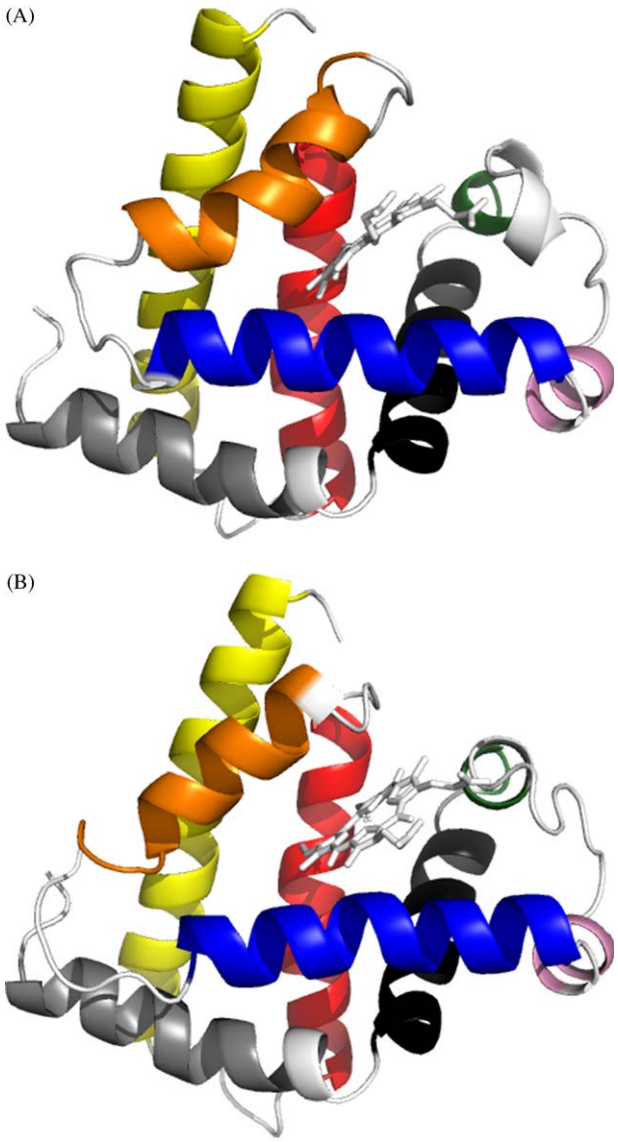

Fig. 9. 3D representation of the average structures from each cluster *ensemble*: A (A) and B (B). The eight helices are colored according to the following legend: Helix A (aa 4–17) in grey, Helix B (aa 21–35) in black, Helix C (aa 37–40) in green, Helix D (aa 52–56) in pink, Helix E (aa 59–76) in blue, Helix F (aa 83–95) in orange, Helix G (aa 101–118) in red, Helix H (aa 125–149) in yellow.

on the 160 ns *total 1–8* concatenated trajectory, which should assure an adequate conformational sampling according to PCA analysis (Figs. 1, 3 and 4), the FEL strongly resembles that obtained for the *total 1–11* trajectory and the same conformational transitions between regions A and B are present (Fig. 6D).

4. Conclusions

Multiple MD simulations with different initial atomic velocities were used to sample conformations in the proximity of the native structure of holo-myoglobin using an iterative procedure. Eleven independent 20 ns trajectories (*replicas*), which sample only a fraction of the conformational distribution, were concatenated, allowing to collect more than 0.22 μs of simulation at 300 K.

The coordinate space evolution of the different *replicas* was examined through low-dimensional projections (via principal component analysis) of the conformational space explored by the trajectories. In the sample of eleven room temperature trajectories, the results from principal component analysis and from the FEL description showed that two major regions of the conformational space play an important role. It is likely that these regions only represent a portion, albeit an important one, of the total conformational space accessible to room-temperature trajectories of myoglobin. In general, in our simulations, a trajectory only samples one region, but some transitions between the two main regions are observed. Interestingly, the substates which populate the two distinct energy basins differ in particular for the orientation of the helices F and H.

Therefore, our analysis highlights that the available conformational space for holo-Mb is quite constrained, in agreement with previous results from pioneering and recent MD simulations and from experimental studies [24–29]. In fact, Mb folds as a compact monomer and does not undergo large conformational changes in comparison with multidomain proteins, but localized changes are observed. The available evidence suggests that the conformational ensemble of Mb is characterized by subtle conformational changes, which require a detailed description of the conformations in the neighborhood of the native state. The analysis of rmsd matrices referring to distinct structural elements, instead of a single global rmsd matrix, has proven useful for structural clustering and for discrimination of the structural substates of the globin fold. This approach could be of general utility in the analysis of simulations of proteins characterized by other folds.

The convergence of parameters as cosine content and rmsip, the coherence among the identified clusters and the free-energy basins indicate that an accurate description of the distribution of conformations in the configurational space has been achieved and that the concatenation of 7/8 *replicas* spanning 140/160 ns of simulations is sufficient to ensure an adequate sampling of the native conformational space of Mb. If a smaller sampling were considered, conformational transitions due to displacement of single helices, might not be captured by free energy or principal component analysis.

In the present contribution, results for Mb were presented as an exploratory study since it is a small system representative of one of the main protein folds (the globin fold) and since it could be interesting to extend this MD approach to investigate the conformational and dynamic properties of apo- or mutant versions of this protein, as well as deletions aimed to the identification of the minimal fragment which retains a native-like conformation [8–11]. These are variants in which greater conformational variability is expected and the identification of representative substates from MD trajectories will be crucial.

Acknowledgements

The authors thank CINECA (Project 696 – 2008) for the use of computational facilities and Marco Pasi for helpful discussions and critical reading of the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2009.01.006.

References

[1] H. Frauenfelder, B.H. McMahon, P.W. Fenimore, Myoglobin: the hydrogen atom of biology and a paradigm of complexity, Proc. Natl. Acad. Sci. USA 100 (2003) 8615–8617.
[2] F.G. Parak, G.U. Nienhaus, Myoglobin, a paradigm in the study of protein dynamics, Chemphyschem 3 (2002) 249–254.

[3] M. Brunori, D. Bourgeois, B. Vallone, The structural dynamics of myoglobin, J. Struct. Biol. 147 (2004) 223–234.
[4] D.E. Bikiel, L. Boechi, L. Capece, A. Crespo, P.M.D. Biase, S.D. Lella, M.C.G. Lebrero, M.A. Marti, A.D. Nadra, L.L. Perissinotti, D.A. Scherlis, D.A. Estrin, Modeling heme proteins using atomistic simulations, Phys. Chem. Chem. Phys. 8 (2006) 5611–5628.
[5] M. Bolognesi, D. Bordo, M. Rizzi, C. Tarricone, P. Ascenzi, Nonvertebrate hemoglobins: structural bases for reactivity, Prog. Biophys. Mol. Biol. 68 (1997) 29–68.
[6] K. Shikama, A. Matsuoka, Structure–function relationships in unusual nonvertebrate globins, Crit. Rev. Biochem. Mol. Biol. 39 (2004) 217–259.
[7] J.A. Hoy, H. Robinson, J.T. Trent, S. Kakar, B.J. Smagghe, M.S. Hargrove, Plant hemoglobins: a molecular fossil record for the evolution of oxygen transport, J. Mol. Biol. 371 (2008) 168–179.
[8] M.T. Reymond, G. Merutka, H.J. Dyson, P.E. Wright, Folding propensities of peptide fragments of myoglobin, Protein Sci. 6 (1997) 706–716.
[9] R. Grandori, S. Schwarzinger, N. Muller, Cloning, overexpression and characterization of micro-myoglobin, a minimal heme-binding fragment, Eur. J. Biochem. 267 (2002) 1168–1172.
[10] E.A. Ribeiro, C.H.I. Ramos, Circular permutation and deletion studies of myoglobin indicate that the correct position of its N-terminus is required for native stability and solubility but not for native-like heme binding and folding, Biochemistry 44 (2005) 4699–4709.
[11] H. Ji, L. Shen, R. Grandori, N. Muller, The effect of heme on the conformational stability of micro-myoglobin, FEBS J. 275 (2008) 89–96.
[12] M. Jamin, The folding process of apomyoglobin, Protein Pept. Lett. 12 (2005) 229–234.
[13] C. Nishimura, H.J. Dyson, P.E. Wright, Identification of native and non-native structure in kinetic folding intermediates of apomyoglobin, J. Mol. Biol. 355 (2005) 139–156.
[14] C. Nishimura, P.E. Wright, H.J. Dyson, Role of the B helix in early folding events in apomyoglobin: evidence from site-directed mutagenesis for native-like long range interactions, J. Mol. Biol. 334 (2003) 293–307.
[15] H. Dyson, P.E. Wright, H.A. Scheraga, The role of hydrophobic interactions in initiation and propagation of protein folding, Proc. Natl. Acad. Sci. USA 103 (2006) 13057–13061.
[16] A. Ansari, C.M. Jones, E.R. Henry, J. Hofrichter, W.A. Eaton, Conformational relaxation and ligand binding in myoglobin, Biochemistry 17 (1994) 5128–5145.
[17] H.H. de Jongh, E. Goormaghtigh, J.M. Ruysschaert, Amide-proton exchange of water-soluble proteins of different structural classes studied at the submolecular level by infrared spectroscopy, Biochemistry 44 (1997) 13603–13610.
[18] Y. Yamamoto, T. Nakashima, E. Kawano, R. Chujo, 1H-NMR investigation of the influence of the heme orientation on functional properties of myoglobin, Biochim. Biophys. Acta 1388 (1998) 349–362.
[19] S. Cavagnero, Y. Theriault, S.S. Narula, H.J. Dyson, P.E. Wright, Amide proton hydrogen exchange rates for sperm whale myoglobin obtained from 15N-1H NMR spectra, Protein Sci. 9 (2000) 186–193.
[20] D.A. Simmons, S.D. Dunn, L. Konermann, Conformational dynamics of partially denatured myoglobin studied by time-resolved electrospray mass spectrometry with online hydrogen-deuterium exchange, Biochemistry 42 (2003) 5896–5905.
[21] D.A. Simmons, L. Konermann, Characterization of transient protein folding intermediates during myoglobin reconstitution by time-resolved electrospray mass spectrometry with on-line isotopic pulse labeling, Biochemistry 41 (2002) 1906–1914.
[22] R. Aranda, E.J. Levin, F. Schotte, P.A. Anfinrud, G.N. Phillips, Time-dependent atomic coordinates for the dissociation of carbon monoxide from myoglobin, Acta Crystallogr. D Biol. Crystallogr. 62 (2006) 776–783.
[23] F. Schotte, J. Soman, J.S. Olson, M. Wulff, P.A. Anfinrud, Picosecond time-resolved X-ray crystallography: probing protein function in real time, J. Struct. Biol. 147 (2004) 235–246.
[24] R. Elber, M. Karplus, Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin, Science 235 (1987) 318–321.
[25] B.K. Andrews, T. Romo, J.B. Clarage, B.M. Pettitt, G.N. Phillips, Characterizing global substates of myoglobin, Structure 6 (1998) 587–594.
[26] M. Aschi, C. Zazza, R. Spezia, C. Bossa, A. Di Nola, M. Paci, A. Amadei, Conformational fluctuations and electronic properties in mioglobin, J. Comput. Chem. 25 (2004) 974–984.
[27] J.Z. Ruscio, D. Kumar, M. Shukla, M.G. Prisant, T.M. Murali, A.V. Onufriev, Atomic level computational identification of ligand migration pathways between solvent and binding site in myoglobin, Proc. Natl. Acad. Sci. USA 105 (2008) 9204–9209.
[28] C. Bossa, A. Amadei, I. Daidone, M. Anselmi, B. Vallone, M. Brunori, A. Di Nola, Molecular dynamics simulation of sperm whale mioglobina: effects of mutations and trapped CO on the structure and dynamics of cavities, Biophys. J. 89 (2005) 465–474.
[29] D.A. Kondrashov, W. Zhang, R. Aranda, B. Stec, G.N. Phillips, Sampling of the native conformational ensemble of myoglobin via structures in different crystalline environments, Proteins 70 (2008) 353–362.
[30] A. Cooper, Conformational fluctuation and change in biological macromolecules, Sci. Prog. 66 (1980) 473–497.
[31] H. Frauenfelder, S.G. Sligar, P.G. Wolynes, The energy landscapes and motions of protein, Science 254 (1991) 1598–1603.
[32] K. Henzler-Wildman, D. Kern, Dynamic personalities of proteins, Nature 450 (2007) 964–972.
[33] D.J. Wales, The energy landscape as a unifying theme in molecular science, Philos. Trans. A Math. Phys. Eng. Sci. 363 (2005) 357–377.
[34] O.M. Becker, Geometric versus topological clustering: an insight into conformation mapping, Proteins 27 (1997) 213–226.
[35] A. Amadei, A.B. Linssen, H.J. Berendsen, Essential dynamics of proteins, Proteins 17 (1993) 412–425.
[36] B. Zagrovic, V.S. Pande, How does averaging affect protein structure comparison on the ensemble level? Biophys. J. 87 (2005) 2240–2246.
[37] S.B. Dixit, S.Y. Ponomarev, D.L. Beveridge, Root mean square deviation probability analysis of molecular dynamics trajectories on DNA, J. Chem. Inf. Model. 46 (2006) 1084–1093.
[38] L.S. Caves, J.D. Evanseck, M. Karplus, Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin, Protein Sci. 7 (1998) 649–666.
[39] G. Moraitakis, A. Purkiss, J. Goodfellow, Simulated dynamics and biological macromolecules, Rep. Prog. Phys. 66 (2003) 383–406.
[40] J. Straub, A. Rashkin, D. Thirumalai, Dynamics in rugged energy landscapes with applications to the S-peptide and ribonuclease-A, J. Am. Chem. Soc. 116 (1994) 2049–2063.
[41] E. Papaleo, M. Pasi, L. Riccardi, I. Sambi, P. Fantucci, L. De Gioia, Protein flexibility in psychrophilic and mesophilic trypsins. Evidence of evolutionary conservation of protein dynamics in trypsin-like serine-proteases, FEBS Lett. 582 (2008) 1008–1018.
[42] D.M. Zuckerman, E. Lyman, A second look at canonical sampling of biomolecules using replica exchange simulation, J. Chem. Theory Comput. 2 (2006) 1200–1202.
[43] I. Tavernelli, S. Cotesta, E.E.D. Iorio, Protein dynamics, thermal stability, and free-energy landscapes: a molecular dynamics investigation, Biophys. J. 85 (2003) 2641–2649.
[44] H. Lei, C. Wu, H. Liu, Y. Duan, Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations, Proc. Natl. Acad. Sci. USA 104 (2007) 4925–4930.
[45] G. Colombo, G. Morra, M. Meli, G. Verkhivker, Understanding ligand-based modulation of the Hsp90 molecular chaperone dynamics at atomic resolution, Proc. Natl. Acad. Sci. USA 105 (2008) 7976–7981.
[46] J.N. Onuchic, Z. Luthey-Schulten, P.G. Wolynes, Theory of protein folding: the energy landscape perspective, Annu. Rev. Phys. Chem. 48 (1997) 545–600.
[47] M. Gruebele, Downhill protein folding: evolution meets physics, C. R. Biol. 328 (2005) 701–712.
[48] Y. Mu, P.H. Nguyen, G. Stock, Energy landscape of a small peptide revealed by dihedral angle principal component analysis, Proteins 58 (2005) 45–52.
[49] B. Hess, Convergence of sampling in protein simulations, Phys. Rev. E Stat. Nonlin. Soft Matter. Phys. 65 (2002) 031910.
[50] G.G. Maisuradze, D.M. Leitner, Free energy landscape of a biomolecule in dihedral principal component space: sampling convergence and correspondence between structures and minima, Proteins 67 (2007) 569–578.
[51] F. Hamprecht, C. Peter, X. Daura, W. Thiel, W. van Gusteren, A strategy for analysis of (molecular) equilibrium simulations: configuration space density estimation, clustering, and visualization, J. Chem. Phys. 114 (2001) 2079–2089.
[52] D. Shortle, K.T. Simons, D. Baker, Clustering of low-energy conformations near the native structures of small proteins, Proc. Natl. Acad. Sci. USA 95 (1998) 11158–11162.
[53] M. Anselmi, A. Di Nola, A. Amadei, The kinetics of ligand migration in crystallized mioglobin as revealed by molecular dynamics simulations, Biophys. J. 94 (2008) 4277–4281.
[54] M. Anselmi, M. Brunori, B. Vallone, A. Di Nola, Molecular dynamics simulation of the neuroglobin crystal: comparison with the simulation in solution, Biophys. J. 95 (9) (2008) 4157–4162.
[55] J. Vojtechovsky, K. Chu, J. Berendzen, R.M. Sweet, I. Schlichting, Crystal structures of myoglobin-ligand complexes at near-atomic resolution, Biophys. J. 77 (1999) 2153–2174.
[56] F. Shu, V. Ramakrishnan, B.P. Schoenborn, Enhanced visibility of hydrogen atoms by neutron crystallography on fully deuterated myoglobin, Proc. Natl. Acad. Sci. USA 97 (2000) 3872–3877.
[57] H. Berendsen, J. Postma, W. van Gusteren, J. Hermans, Interaction models for water in relation to protein hydratation, Reidel, Dordrecht (1981) 331–342.
[58] S. Bhattacharya, S.F. Sukits, K.L. MacLaughlin, J.T. Lecomte, The tautomeric state of histidines in myoglobin, Biophys. J. 73 (1997) 3230–3240.
[59] B. Hess, H. Bekker, H. Berendsen, J. Fraahje, LINCS: a linear constraint solver for molecular interactions, J. Comp. Chem. 18 (1997) 1463–1472.
[60] T. Darden, D. York, L. Pedersen, Particle mesh Ewald. An N.log(N) method for Ewald sums in large systems, J. Chem. Phys. 98 (1993) 10089–10092.
[61] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[62] B. Hess, Similarities between principal components of protein dynamics and random diffusion, Phys. Rev. E Stat. Phys. Plasmas. Fluids Relat. Interdiscip. Topics 62 (2000) 8438–8448.
[63] A. Amadei, M.A. Ceruso, A. Di Nola, On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations, Proteins 36 (1999) 419–424.
[64] F. Pontiggia, G. Colombo, C. Micheletti, H. Orland, Anharmonicity and self-similarity of the free energy landscape of protein G, Phys. Rev. Lett. 98 (2007) 048102.
[65] G. Ross, Algorithm AS 15: single Linkage cluster analysis, Appl. Stat. 18 (1969) 106–111.

[66] X. Daura, K. Gademann, B. Jaun, W. van Gusteren, A. Mark, Peptide folding: when simulation meets experiment, Angew. Chem. Int. Ed. 38 (1999) 236–240.
[67] J.T. Lecomte, S.F. Sukits, S. Bhattacharya, C.J. Falzone, Conformational properties of native sperm whale apomyoglobin in solution, Protein Sci. 8 (1999) 1484–1491.
[68] M. Laberge, T. Yonetani, Common dynamics of globin family proteins, IUBMB Life 59 (2007) 528–534.
[69] S. Maguid, S. Fernandez-Alberti, L. Ferrelli, J. Echave, Exploring the common dynamics of homologous proteins. Application to the globin family, Biophys. J. 89 (2005) 3–13.
[70] V. Guallar, A.A. Jarzecky, R.A. Friesner, T.G. Spiro, Modeling of ligation-induced helix/loop displacements in myoglobin: toward an understanding of hemeglobin allostery, J. Am. Chem. Soc. 128 (2006) 5427–5435.