bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Characterizing molecular flexibility by combining lRMSD measures

F. Cazals*and R. Tetley†

July 29, 2018

Abstract

The root mean square deviation (RMSD) and the least RMSD are two widely used similarity measures in structural bioinformatics. Yet, they stem from global comparisons, possibly obliterating locally conserved motifs. We correct these limitations with the so-called *combined RMSD* , which mixes independent lRMSD measures, each computed with its own rigid motion. The combined RMSD can be used to compare (quaternary) structures based on motifs defined from the sequence (domains, SSE), or to compare structures based on structural motifs yielded by local structural alignment methods.

We illustrate the benefits of combined RMSD over the usual RMSD on three problems, namely (i) the analysis of conformational changes based on combined RMSD of rigid structural motifs (case study: a class II fusion protein), (ii) the calculation of structural phylogenies (case study: class II fusion proteins), and (iii) the assignment of quaternary structures for hemoglobin. Using these, we argue that the combined RMSD is a tool a choice to perform positive and negative discrimination of degree of freedom, with applications to the design of move sets and collective coordinates.

Combined RMSD are available within the Structural Bioinformatics Library (`http://sbl.inria.fr`).

Abbreviations: RMSD: root mean square deviation; lRMSD: least RMSD; c.c.: connected component.

1 Introduction

1.1 RMSD, lRMSD and their variants

The problem of geometrically comparing two point sets of the same cardinality, assuming a one-to-one correspondence between the points, has long been recognized as central in science and engineering. It is in general desired to perform a comparison oblivious to rigid motions, which prompts a solution computing concomitantly the geometric similarity measure and the associated optimal rigid motion. The most celebrated solution to this problem is the so-called least root mean square deviation (lRMSD)[1], namely the RMSD of positions upon applying the optimal rigid motion. This number, which is usually expressed in Å. (We note that the lRMSD is a coordinate RMSD, not the be confused with the $RMSD_d$, namely the RMSD of internal distances.) In the sequel, we review previous work, by restricting ourselves to structural bioinformatics.

Several strategies to compute the lRMSD [1, 2, 3] or its weighted variant [4, 5] were developed long ago, and it was also noted that the lRMSD induces a metric [6]. Owing to these properties, the lRMSD has been one of the most used similarity criteria in structural biology and bioinformatics. On the other hand, several limitations prompted developments from the design and computational perspectives.

On the design side, efforts were made to circumvent several limitations. The lRMSD is inherently hard to interpret, as medium values may stem from a fuzzy structural conservation contributed by all atoms, or from small regions that underwent large conformational changes while their complement is isometrically conserved.

*Inria, Université Côte d'Azur. 2004 route des Lucioles, F-06902 Sophia Antipolis.. Correspondence: Frederic.Cazals@inria.fr
†Inria, Université Côte d'Azur, France

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Possibly worse, the lRMSD involves the covariance matrix of centered atomic positions (see below), so that points far from the center of mass get more weight. This fact also has another consequence: by a packing argument, large proteins distribute atoms farther from the center of mass, so that the lRMSD depends on protein size. In order to weigh all points evenly and also to obtain a normalized measure, a variation of the RMSD obtained by restricting the calculation to unit vectors along the backbone and performing the optimization over rotations was developed [7]. Normalized alternatives were also proposed, respectively based on the radii of gyration of the molecules compared [8], and on normalization factors inferred from typical distributions of RMSD values [9, 10]. In a complementary line of attack, various superposition-free measures were proposed. To compare structures and models, one may use the lDDT which is an average value (computed over four thresholds) of fraction of distances within the chosen threshold [11]. Similarly, the contact area difference quantifies differences of contact areas between a model and a structure [12]. Finally, a normalized measure based on the Binet-Cauchy kernel, which inherently computes a scalar product between vectors defining the volume of tetrahedra was recently proposed [13].

Improvements were also reported to speed-up calculations. Efficient calculations targeting subsets of the aligned structures were investigated [14]. As an alternative to lRMSD calculations based on matrix decompositions–such as SVD, a fast determination of the optimal rotation matrix based on a Newton-Raphson quaternion-based method was reported [4]. Finally, recent work addressed the calculation of RMSD between flexible structures, the flexibility being modeled using collective coordinates obtained via normal mode analysis or PCA [15]. This latter work is especially interesting as it targets deformable structures modeled by means of collective motions.

To conclude, we also note that RMSD and lRMSD calculations are tightly related to the calculation of structural alignments between two structures. The intrication between scores and alignment methods was first exploited in [16], which performs an iterative alignment, guided by two scores, namely `GDT` (the fraction of residues (largest set, not sequence contiguous) that fit under a distance cutoff), and `LCS` (the fraction of amino acids defining the longest contiguous segment fitting under a given RMSD cutoff (positions of molecules fixed)). Since then, a variety of methods were proposed to detect and score structural motifs [17], including `DALI` [18], `TM-align` [19], `Apurva` [20], `LGA` [16], `Kpax` [21], or our persistence based method [22].

1.2 Contribution

In the sequel, we propose an elementary formula combining lRMSD measures, each associated with its own optimal rigid motion, so as in particular to assess molecular flexibility. With respect to the afore-discussed limitations of the lRMSD, our so-called *combined RMSD* has the following advantages:

- Flexibility is characterized by local structural alignments between structural motifs of the compared molecules, rather than with a global parameterization of motion.
- Flexibility can be assessed at multiple scales, by parameterizing the number of rigid motions used–this number is equal to the number of motifs mapped to one-another.
- The dependency of protein size is alleviated.

2 Combining independent lRMSD measures

2.1 Structures, alignments, and motifs

Let A and B represent either two distinct structures or two conformations of the same molecule or complex. Assuming these are proteins, we denote their number of amino acids n_A and n_B, respectively. Our structural comparisons are based on the coordinates of *particles*. In comparing two conformations of the same molecule, the particles may refer to all atoms, all heavy atoms only, or C_α atoms; for two distinct molecules, we assume the particles are C_α atoms.

Our comparisons shall use variants of the lRMSD, whose calculation requires an alignment of the two structures to be compared. We first recall:

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Definition. 1 *Consider two sequences of length n_A and n_B. An alignment of length $N \leq \min(n_A, n_B)$ is defined by two sets of indices $I = (i_1, i_2, \ldots, i_N)$ with $1 \leq i_1 < i_2 < \cdots < i_N \leq n_A$ and $J = (j_1, j_2, \ldots j_N)$ with $1 \leq j_1 < j_2 < \ldots j_N \leq n_B$. An* alignment *is specified by the perfect matching $\{(i_1, j_1), \ldots, (i_N, j_N)\}$.*

Once the two structures have been aligned, abusing notations, we may reduce them to the two ordered point sets $A = \{a_i\}_{i=1,\ldots,N}$ and $B = \{b_i\}_{i=1,\ldots,N}$.

Structural motifs. To *localize* structural comparisons, we assume that *structural motifs* have been identified. Practically, two types of motifs may be used: features of proteins (domains, SSE), or motifs yielded by local structural alignment methods–see Introduction. More formally, we define:

Definition. 2 *Consider two structures A and B. A* motif *is a pair of set of particles $M^{(A)} \subset A$ and $M^{(B)} \subset S_B$ of the same size, together with an alignment between them.*

The alignment allows computing lRMSD$(M^{(A)}, M^{(B)})$. The fact that motifs may overlap calls for the following processing.

Motif graph for overlapping motifs. When several motifs exist for two structures, an important question is to handle them coherently. Since motifs may overlap, we define (Fig. 1):

Definition. 3 *(Motif graph) The* motif graph *of a list of motifs $\{(M_i^{(A)}, M_i^{(B)})\}_{i=1,\ldots,p}$ is defined as follows: its node set is the union of the particles A and B; its edge set is the union of two types of edges:*

- matching edges*: the edges associated with the matchings defined by the motifs. NB: such edges are counted without multiplicity, that is, a matching edge present in several motifs is counted once.*
- motif edges*: edges defining a path connecting all amino acids in a motif.*

Consider a connected component (c.c.) of the motif graph. Restricting each c.c. to each structure yields two subgraphs. The set of all such subgraphs is denoted $\{C_i^{(A)}, C_i^{(B)}\}_{i=1,\ldots,m}$.

The following observations can be made (Fig. 1). A subgraph ($C_i^{(A)}$ or $C_i^{(B)}$) may not be connected. Also, the motif graph does not define, in general, a matching between the vertices associated with particles. More precisely, the *multiplicity* of a particle in a motif graph is defined as the number of edges incident to this particle. Despite these features, as we shall see below, the edges connecting the particles from $C_i^{(A)}$ and $C_i^{(B)}$ can be used to define a variant of the classical lRMSD.

2.2 Vertex weighted and edge weighted lRMSD

We introduce generalizations of the lRMSD and RMSD, using connected components of the motif graph. Before presenting these, we recall the construction of the weighted lRMSD (Def. 4).

Vertex weighted lRMSD : $\mathbf{lRMSD_{vw}}$. Consider two point sets $A = \{a_i\}$ and $B = \{b_i\}$ of size N. Also consider a set of positive weights $\{w_i\}_{i=1,\ldots,N}$, meant to stress the importance of certain particles. The weighted RMSD reads as:

$$\mathrm{RMSD_w}(A, B) = \sqrt{\frac{1}{\sum_i w_i} \sum_{i=1,\ldots,N} w_i \left\| a_i - b_i \right\|^2} \tag{1}$$

Let g a rigid motion from the the special Euclidean group $SE(3)$. To perform a comparison of A and B oblivious to rigid motions, we use the so-called *least RMSD* [1]:

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Definition. 4 *The* vertex weighted *lRMSD is defined by*

$$lRMSD_{vw}(A,B) = \min_{g \in SE(3)} RMSD_w(A, g(B)). \tag{2}$$

The rigid motion yielding the minimum is denoted g^{OPT} (A,B) or g^{OPT} for short.
The weight of the lRMSD$_{vw}$ is defined as $W_{vw}(A,B) = \sum_i w_i$.

Note that the celebrated lRMSD is the particular case of the previous with unit weights:

$$\text{lRMSD}(A,B) = \text{lRMSD}_{\text{vw}}(A,B) \text{ with } w_i \equiv 1. \tag{3}$$

Denote R the sought rotation matrix [2] and C the covariance matrix

$$C = \sum_i w_i b_i {a_i}^{\mathsf{T}}. \tag{4}$$

Upon centering the data, computing the lRMSD amounts to maximizing Trace(RC) [2, 4], a calculation which can be done with an SVD calculation.

Edge weighted lRMSD : lRMSD$_{\text{ew}}$. Consider now the case where motifs have been defined for the two structures A and B. We wish to compare A and B exploiting the information yielded by the connected components of the motif graph (Def. 3). Consider the i-th c.c. of the motif graph. Let e_i be the number of matching edges of this c.c. As usual, let $g(b_j)$ the position of atom b_j from $C_i^{(B)}$ matched with atom a_j from $C_i^{(A)}$, upon applying a rigid motion g. We define:

Definition. 5 *The* edge weighted lRMSD$_{\text{ew}}$ *of the i-th c.c. of the motif graph is defined by*

$$lRMSD_{ew}(C_i^{(A)}, C_i^{(B)}) = \min_{g \in SE(3)} \sqrt{\frac{1}{e_i} \sum_{j=1}^{e_i} \|a_j - g(b_j)\|^2} \tag{5}$$

The rigid motion yielding the minimum is denoted g_i^{OPT}.
The weight of the lRMSD$_{ew}$ is defined as $W_{ew}(C_i^{(A)}, C_i^{(B)}) = e_i$.

To compute this quantity, we proceed as for the lRMSD$_{\text{vw}}$, except that the covariance matrix from Eq. (4) is now obtained by summing over edges of the bipartite graph rather than on vertices. We also make the following

Observation. 1 *If the motif graph defines a perfect matching between the particles of a connected component, then lRMSD$_{vw}$ = lRMSD$_{ew}$ for that component.*

2.3 Combined RMSD : RMSD$_{\text{Comb.}}$

Since the lRMSD$_{\text{ew}}$ values are defined for each c.c. of the motif graph, we combine them to obtain a comparison of A and B.

Denote m the number of connected components of the motif graph and let $N_e = \sum_i e_i$. The edge weighted lRMSD can be combined into the following *edge-weighted combined RMSD*

$$\text{RMSD}_{\text{Comb.}}(A,B) = \sqrt{\frac{1}{N_e} \sum_{i=1}^{m} \sum_{j=1}^{e_i} \left\|a_j - g_i^{\text{OPT}}(b_j)\right\|^2} = \sqrt{\sum_{i=1}^{m} \frac{e_i}{N_e} \text{lRMSD}^2_{\text{ew}}(C_i^{(A)}, C_i^{(B)})} \tag{6}$$

It is easily checked that the previous is a particular case of the following combined RMSD, which mixes individual lRMSD, be they vertex weighted (lRMSD$_{\text{vw}}$) or edge weighted (lRMSD$_{\text{ew}}$):

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Definition. 6 *Consider two structures A and B for which non-overlapping regions* $\{C_i^{(A)}, C_i^{(B)}\}_{i=1,\dots,m}$ *have been identified – Def. 3. Assume that a lRMSD has been computed for each pair* $(C_i^{(A)}, C_i^{(B)})$*. Let* w_i *be the weights associated with an individual lRMSD. The* combined RMSD *is defined by*

$$RMSD_{Comb.}(A,B) = \sqrt{\sum_{i=1}^{m} \frac{w_i}{\sum_i w_i} lRMSD^2(C_i^{(A)}, C_i^{(B)})}. \tag{7}$$

The following bounds are straightforwards convexity inequalities (proof in SI Sec. 7.1):

Observation. 2 *The* combined RMSD *satisfies the following upper and lower bounds:*

$$RMSD_{Comb.}(A,B) \geq \sum_{i=1}^{m} \frac{w_i}{\sum_i w_i} lRMSD(C_i^{(A)}, C_i^{(B)}). \tag{8}$$

Let $l_{min} = \min_i lRMSD(C_i^{(A)}, C_i^{(B)})$ *and* $l_{max} = \max_i lRMSD(C_i^{(A)}, C_i^{(B)})$*. One has*

$$RMSD_{Comb.}(A,B) \leq \sum_{i=1}^{m} \frac{w_i}{\sum_i w_i} lRMSD(C_i^{(A)}, C_i^{(B)}) + 2\Big(\sqrt{\frac{l_{min}+l_{max}}{2}} - \frac{\sqrt{l_{min}}+\sqrt{l_{max}}}{2}\Big). \tag{9}$$

Remark 1 *Equation (7) defines a RMSD rather than a lRMSD. To see why, observe that Eq. (2) defines a number which is the minimum of a quadratic optimization problem involving Eq. (4). Instead, Eq. (7) defines a number obtained by mixing solutions of such problems for connected components of the motif graph.*

Remark 2 *Combined RMSD can be used iteratively. Consider two conformations of a a complex containing say two chains, each decomposed into motifs. In a first step, the combined RMSD exploiting the motifs can be used to compare the two instances of each chain across the two complexes. In a second steps, the combined RMSD of these combined RMSD can be computed. We illustrate this strategy to provide insights on novel conformations of quaternary complexes of hemoglobin in section 4.3.*

3 Implementation

All methods are available in the Structural Bioinformatics Library (`http://sbl.inria.fr`, [23]).

Three executables implementing the methods (`sbl-rmsd-flexible-proteins.exe`, `sbl-rmsd-flexible-conformations.exe`, `sbl-rmsd-flexible-motifs.exe`) are provided in the package Molecular_distances_flexible (`https://sbl.inria.fr/doc/Molecular_distances_flexible-user-manual.html`). Additional details are provided in SI Section 7.2.

Also of particular interest is the Structural_motifs package (see `https://sbl.inria.fr/doc/Structural_motifs-user-manual.html`), which makes available various methods to compute structural motifs [22].

4 Results

We illustrate insights yielded by the combined RMSD, which are out of reach for the classical lRMSD.

4.1 Assessing conformational changes: the example of a class II fusion protein

Biological context. Viruses replicate by hijacking the cellular apparatus of the host organism, with three main steps: entry into the host cell, multiplication, and egress of progeny virions from the host cell. For enveloped viruses, the entry requires the fusion of their envelope with the membrane of the host cell, a process triggered by a (low pH induced) conformational change of a membrane glycoprotein. Fusion proteins are ascribed to three classes denoted I, II and III, with class II fusion proteins typically structured three domains

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mainly composed of β strands. In their post-fusion conformation, these proteins act as trimers project out from the viral membrane. In the sequel, we use the combined RMSD to illustrate the conformation changes undergone by a prototypical class II fusion protein, from the tick-borne encephalitis virus. The ectodomain of this protein was crystallized both in soluble form (PDB: 1SVB, [24], 395 residues) and in postfusion conformation (PDB: 1URZ, [25], 400 residues) (Fig. 2). We use structural motifs computed using a method reported in a companion paper [22] to demonstrate the $\text{RMSD}_{\text{Comb.}}$ in such a setting (RMSD_MODE_MOTIF, SI Sec. 7.2.3).

Results. At first glance, a global lRMSD calculation yields lRMSD = 11.32. To further assess the presence of rigid motifs moving relatively to one another, we computed motifs (Def. 2) using the method presented in the companion paper [22]. In a nutshell, this methods identifies structurally conserved motifs, each such region being a connected domain whose connectivity is ensured by pairs of atoms whose distance is conserved between the structures studied. (We note in passing that this strategy departs from the classical approaches targeting quasi-isometric motifs via the calculation of cliques [26].)

For the tick-borne encephalitis virus, the method yields a total of $p = 31$ structural motifs distributed in $m = 2$ connected components ($|C_1| = 109$, $|C_2| = 51$; Fig. 2). Finally, in computing the combined RMSD between the two connected components (Def. 3), one obtains $\text{RMSD}_{\text{Comb.}} = 2.50$.

As illustrated by this example, the ability to identify structurally conserved motifs and to combine their lRMSD makes a dramatic difference in the overall comparison of two conformations. This problem is well known e.g. for the analysis of molecular dynamics trajectories, where erroneous assignments of structures to the same clusters / meta-stable states may jeopardize free energy calculations [27].

4.2 Building a phylogeny from structural data: the example of class II fusion proteins

Biological context. Following on the previous example, we now illustrate the interest of combined RMSD to build phylogenies of class II fusion proteins. As a dataset, we use 6 monomers of class II fusion proteins in their post-fusion conformation (Fig. 3). In order to assess the merits of the usual lRMSD and that of the combined RMSD to build phylogenies via dendograms, we carry out the analysis at two levels, namely using whole domains and the 22 motifs defined by SSE (Fig. 3(b)). (This corresponds to the setting RMSD_MODE_SEQ for homologous proteins, see SI Section 7.2.2.)

Results. Out of the six structures and for each pairwise comparison–15 of them, we used the executables provided in the `SBL` to compute: (1) a structural alignment using the `Apurva` algorithm [20]. This yields a lRMSD (Table 1), (2) the $\text{RMSD}_{\text{Comb.}}$ upon processing the regions defined by the three domains (Table 2), (3) the $\text{RMSD}_{\text{Comb.}}$ upon processing the regions defined by the SSE labels (Table 3). From the distance matrices displayed in the aforementioned tables, we perform three complete linkage hierarchical clusterings (Fig. 4). Each level of comparison conveys different information.

- **Full structure.** This level of information is enough to classify the two flaviviruses together. Out of the 6 structures described in Fig. 3, we know that 2 (HRV-Hanta. and RBV-Rubi.) have a domain swap. This adds considerable noise to the clustering.
- **Domains.** Here the two flaviviruses as well as the two bunyaviruses are clustered together, regardless of the domain swap.
- **SSE.** The two flaviviruses as well as the two togoviruses are clustered together, regardless of the domain swap.

To conclude, while the global lRMSD falls short from providing a satisfactory classification, combined RMSD fixes this limitation. One can further use the individual lRMSD value on a per domain basis to illustrate the structural diversity of these domains (SI Fig. 8).

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4.3 Assigning quaternary structures: the example of hemoglobin

Biological context. Hemoglobin is the gas transporting metalloprotein in mammals. In humans, the predominant form of hemoglobin has a quaternary structure based on four subunits (SI Fig.9), namely two α chains (141 amino acids each), and two chains β (146 amino acids each). Each subunit contains a heme group consisting of a charged iron (Fe) ion held in an heterocyclic ring called porphyrin. It had long been believed that binding of O2 to one monomer triggered the transition of the tetramer from the tense (T) state to the relaxed (R) state, a mechanism at the core of cooperative binding [28]. Along this process, one pair of subunits rotates of an angle $\sim 15\,\text{deg}$ about the other ([29] and SI Fig.9). However, the mechanism is more complex, and a third composite quaternary state, usually denoted R2 or Y was discovered long ago [30]. Based on these reference structures, a combination of various biophysical experiments [31] and macroscopic analysis (using angles and distances to qualify the quaternary structures) [32] had provided insights on gas binding by hemoglobin. Recently, these models were questioned by crystal structures which revealed that each tetramer captured hemoglobin in three quaternary conformations [33]. (NB: the three crystal structures are: half-liganded with phosphate (HL+ , PDB: 4N7P), half-liganded without phosphate (HL- , PDB: 4N7O), and fully water-liganded met-hemoglobin with phosphate (FL+ , PDB: 4N7N).)

Importantly, each new crystal was found to contain 3 new quaternary conformations denoted A, B, C. An assignment procedure based on difference distance matrices of the $\alpha_1\beta_2$ subunits using R as a reference, (SI Sect. 7.4) supported the following: the A, B, C states respectively assumes states R2, R, and T. Moreover, visual inspection of the data [33, Fig. 2] shows that upon alignments, the tetramers tagged A or B superimpose almost perfectly, while those tagged C exhibit less coherent SSE (in particular helices from the FL+ crystal.

Results. As recalled above, assignment of quaternary states was mainly done so far using four rigid-body parameters related to the twofold symmetry [32], or using difference distance matrices with R as a reference state [33]. We revisit this problem, performing a complete hierarchical clustering of dimers using two combined RMSD : a combined RMSD mixing the lRMSD of the two chains; a combined RMSD mixing two combined RMSD –one for the 7 helices of chain α and one for the 8 helices of chain β. We use these combined RMSD to perform a hierarchical clustering (single linkage) of all combinations of α and β chains. Of particular interest are the clusterings of $\alpha_1\beta_1$ and $\alpha_1\beta_2$ dimers (Fig. 5, SI Fig. 10, SI Fig. 11), which evidence three interesting facts.

First, let us consider the ability to group coherently the states A, B, C, versus states $R2, R$ and T. The combined RMSD at the chain level is unable to separate the aforementioned conformations A, B, C (Fig. 5(Left), SI Fig. 10(Left)). On the other hand, the combined RMSD at the SSE level does retrieve the A, B, C conformations and groups them coherently with the $R2, R$ and T states (Fig. 5(Right), SI Fig. 10(Right)).

Second, it confirms the information yielded by visual inspection, according to which conformations C of the tetramer are less coherent. Indeed, the hierarchical clustering isolates conformation C from the FL+ crystal (PDB: 4N7N). It also singles out HL- [A], which is less coherent with the other two [A] but also with R and $R2$ (SI Tables 4 and 5). This somewhat mitigates the analysis from [33], where HL- [A] is reported as a relaxed state between R and R2.

Third, the $\text{RMSD}_{\text{Comb.}}$ values required to form the clusters are tighter than those corresponding to RMSD values of clusters formed using the aforementioned angles and distances [32]. While a much larger dataset was used in this latter paper, the mean values reported for lRMSD between R2, R and T states are 0.40, 0.43 and 0.36, against worst case values $\sim$ 0.27, 0.27 and 0.37 in our case.

Summarizing, the combined RMSD at the SSE level provides insights on the assignment of quaternary structures of hemoglobin, without using any reference state. Since SSE characterize quaternary structures, one can also assess the conformational changes undergone by helices in conjunction with the heme group (SI Fig. 6).

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5 Discussion and outlook

The combined RMSD mixes independent lRMSD measures, each computed with its own rigid motion, therefore avoiding the global parameterization of conformational changes undergone by structures. Moreover, it can be computed at multiple scales, namely to compare secondary or tertiary structures based on structural motifs defined from the sequence or local structural alignments, or to compare quaternary structures from the same motifs. Finally, combined RMSD can be cascaded, so as for example to compare quaternary structures based on motifs defined from SSE elements. The notion of scale is in fact central to combined RMSD : on the one hand, the lRMSD computed for a whole structure typically yields an average signal shadowing the phenomenon scrutinized; on the other hand, computing a combined RMSD for too small structural elements would yield small values void of significance.

We illustrated the interplay between combined RMSD and pertinent scales on three non trivial examples, namely the analysis of large conformation changes, the design of phylogenies based on structural comparisons, and the identification of the quaternary structures of hemoglobin.

These examples may be discussed in the context of dynamical analysis of molecular machines, using the concepts of negative and positive discrimination of degrees of freedom. To articulate these notions, recall that two cornerstones of molecular simulations are move sets and collective coordinates. Move sets, on the one hand, are used in Monte Carlo methods and variants to generate conformations which are diverse and low in energy. Collective coordinates, on the other hand, are key to explore transition paths (and discover transient conformations), and compute free energy landscapes. Importantly, both concepts require understanding which degrees of freedom (dof) are key to account for the conformational changes studied.

To bridge the gap between move sets, collective coordinates, structural motifs, and combined RMSD, let us reconsider the scrutinized structures.

One the one hand, the example of the class II fusion protein undergoing a conformational change illustrates the notion of negative discrimination of dof. Indeed, the combined RMSD of the identified motifs being extremely small while the global lRMSD is large shows that in studying conformational changes between the two conformations studied, one can focus on those dof of atoms outside the motifs. In a sense, the combined RMSD rules out the dof of the motifs it qualifies.

On the other hand, the ability to cluster and qualify the quaternary structures of hemoglobin illustrates the notion of positive discrimination of dof. Indeed, while the global lRMSD does not yield any information, the restriction of lRMSD calculation to SSE, and the combination of the obtained values, yields valid biophysical classification. In other words, the combined RMSD positively discriminates the dof of the motifs it qualifies, calling for further studies to unveil the mechanism scrutinized. Such studies may focus on static analysis of crystal structures (detection of biophysical commonalities, formation/destruction of salt bridges, helix-to-coil transitions, etc). But they may also be dynamic, as the dof identified by may be targeted by complex move sets boosting the identification of structural intermediates.

Our methods to compute structural motifs and combined RMSD are made available within the Structural Bioinformatics Library (`http://sbl.inria.fr`). We anticipate that these tools will prove pivotal to conduct a wide array of structural analysis, both on static and dynamic structure of macro-molecules and their complexes.

References

[1] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976.

[2] K.S. Arun, T.S. Huang, and S.D. Blostein. Least-square fitting of two 3D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):698–700, 1987.

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[3] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 1991.

[4] P. Liu, D. Agrafiotis, and D. Theobald. Fast determination of the optimal rotational matrix for macromolecular superpositions. *Journal of computational chemistry*, 31(7):1561–1563, 2010.

[5] I. Kufareva and R. Abagyan. Methods of protein structure comparison. In *Homology Modeling*, pages 231–257. Springer, 2011.

[6] B. Steipe. A revised proof of the metric properties of optimally superimposed vector sets. *Acta Crystallographica Section A: Foundations of Crystallography*, 58(5):506–506, 2002.

[7] K.Kedem, P. Chew, and R. Elber. Unit-vector rms (urms) as a tool to analyze molecular dynamics trajectories. *Proteins: Structure, Function, and Bioinformatics*, 37(4):554–564, 1999.

[8] V. Maiorov and G. Crippen. Size-independent comparison of protein three-dimensional structures. *Proteins: Structure, Function, and Bioinformatics*, 22(3):273–283, 1995.

[9] Oliviero O. Carugo and S. Pongor. A normalized root-mean-spuare distance for comparing protein three-dimensional structures. *Protein science*, 10(7):1470–1473, 2001.

[10] M. Betancourt and J. Skolnick. Universal similarity measure for comparing protein structures. *Biopolymers*, 59(5):305–309, 2001.

[11] V. Mariani, M. Biasini, A. Barbato, and T. Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

[12] K. Olechnovič, E. Kulberkytė, and C. Venclovas. CAD-score: A new contact area difference-based function for evaluation of protein structural models. *Proteins: Structure, Function, and Bioinformatics*, 81(1):149–162, 2013.

[13] F. Guyon and P. Tufféry. Fast protein fragment similarity scoring using a Binet-Cauchy kernel. *Bioinformatics*, 30(6):784–791, 2014.

[14] T. Shibuya. Efficient substructure RMSD query algorithms. *Journal of Computational Biology*, 14(9):1201–1207, 2007.

[15] E. Neveu, P. Popov, A. Hoffmann, A. Migliosi, X. Besseron, G. Danoy, P. Bouvry, and S. Grudinin. RapidRMSD: Rapid determination of RMSDs corresponding to motions of flexible molecules. *Bioinformatics*, 2018.

[16] A. Zemla. LGA: a method for finding 3D similarities in protein structures. *Nucleic acids research*, 31(13):3370–3374, 2003.

[17] I. Wohlers, N. Malod-Dognin, R. Andonov, and G. Klau. CSA: comprehensive comparison of pairwise protein structure alignments. *Nucleic acids research*, 40(W1):W303–W309, 2012.

[18] L. Holm and C. Sander. Dali: a network tool for protein structure comparison. *Trends in biochemical sciences*, 20(11):478–480, 1995.

[19] Y. Zhang and J. Skolnick. TM-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

[20] R. Andonov, N. Malod-Dognin, and N. Yanev. Maximum Contact Map Overlap Revisited. *J. of Computational Biology*, 18(1):1–15, January 2011.

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[21] D. Ritchie, A. Ghoorah, L. Mavridis, and V. Venkatraman. Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics*, 28(24):3274–3281, 2012.

[22] F. Cazals and R. Tetley. Multiscale analysis of structurally conserved motifs. 2018. In preparation.

[23] F. Cazals and T. Dreyfus. The Structural Bioinformatics Library: modeling in biomolecular science and beyond. *Bioinformatics*, 7(33):1–8, 2017.

[24] F. Rey, F. Heinz, C. Mandl, C. Kunz, and S. Harrison. The envelope glycoprotein from tick-borne encephalitis virus at 2 å resolution. *Nature*, 375(6529):291, 1995.

[25] S. Bressanelli, K. Stiasny, S. Allison, E. Stura, S. Duquerroy, J. Lescar, F. Heinz, and F. Rey. Structure of a flavivirus envelope glycoprotein in its low-ph-induced membrane fusion conformation. *The EMBO journal*, 23(4):728–738, 2004.

[26] N. Malod-Dognin, R. Andonov, and N. Yanev. Maximum clique in protein structure comparison. In P. Festa, editor, *9th International Symposium on Experimental Algorithms*, pages 106–117, Ischia Island, Italy, 2010. Springer Berlin / Heidelberg.

[27] L. Nedialkova, M. Amat, I. Kevrekidis, and G. Hummer. Diffusion maps, clustering and fuzzy markov modeling in peptide folding transitions. *The Journal of chemical physics*, 141(11):09B611_1, 2014.

[28] M. Perutz. Stereochemistry of cooperative effects in haemoglobin1. In *From theoretical physics to biology*, pages 247–285. Karger Publishers, 1973.

[29] J. Baldwin and C. Chothia. Haemoglobin: the structural changes related to ligand binding and its allosteric mechanism. *JMB*, 129(2):175–220, 1979.

[30] F. Smith, E. Lattman, and C. Carter. The mutation β99 Asp-Tyr stabilizes Y—A new, composite quaternary state of human hemoglobin. *Proteins: Structure, Function, and Bioinformatics*, 10(2):81–91, 1991.

[31] W. Eaton, E. Henry, J. Hofrichter, and A. Mozzarelli. Is cooperative oxygen binding by hemoglobin really understood? *Rendiconti Lincei*, 17(1-2):147–162, 2006.

[32] S. Dey, P. Chakrabarti, and J. Janin. A survey of hemoglobin quaternary structures. *Proteins: Structure, Function, and Bioinformatics*, 79(10):2861–2870, 2011.

[33] N. Shibayama, K. Sugiyama, J. Tame, and S-Y. Park. Capturing the hemoglobin allosteric transition in a single crystal form. *Journal of the American Chemical Society*, 136(13):5097–5105, 2014.

[34] S. Simic. On a global upper bound for Jensen's inequality. *Journal of Mathematical Analysis and Applications*, 343(1):414–419, 2008.

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6 Artwork

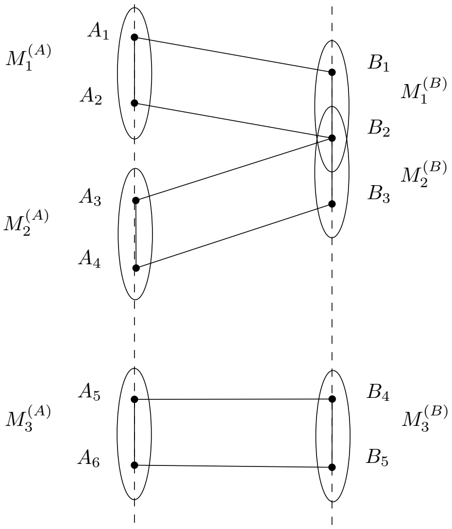
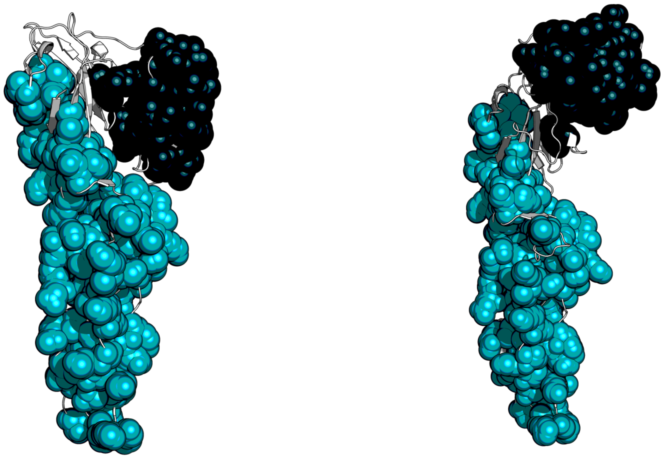


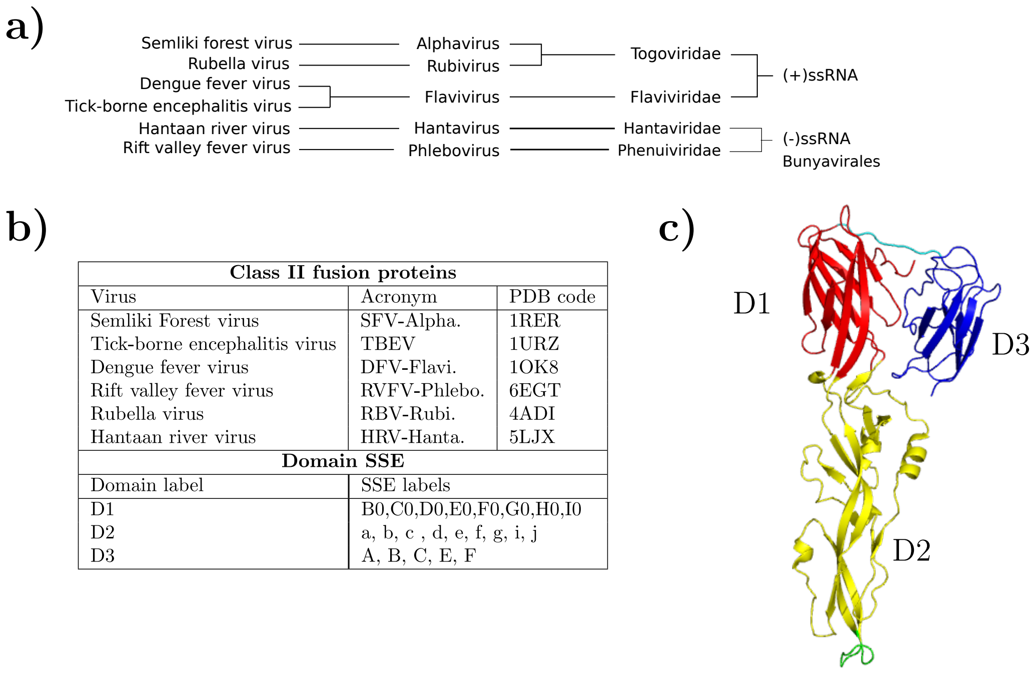
Figure 1: **A motif graph (Def. 3).** A toy system with two structures A and B involving 6 and 5 particles–say C_α-s, respectively. There are three motifs, namely $\{(M_1^{(A)}, M_1^{(B)}), (M_2^{(A)}, M_2^{(B)}), (M_3^{(A)}, M_3^{(B)})\}$. Motif edges are vertical edges connecting the particles; matching edges connect particles from the two structures. The three motifs induce two connected components, respectively containing 4 and 2 matching edges.



lRMSD	$\text{RMSD}_{\text{Comb.}}$
11.32$\AA$	2.50$\AA$

Figure 2: **$\text{RMSD}_{\textbf{Comb.}}$ on overlapping structural motifs impervious to conformational changes: example on a class II fusion protein in soluble and post-fusion conformation.** We display the two connected components, composed by the 31 structural motifs found by our method [22]. Most of the motifs overlap, which justifies a definition for overlapping motifs (Def. 3).

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Class II fusion proteins		
Virus	Acronym	PDB code
Semliki Forest virus	SFV-Alpha.	1RER
Tick-borne encephalitis virus	TBEV	1URZ
Dengue fever virus	DFV-Flavi.	1OK8
Rift valley fever virus	RVFV-Phlebo.	6EGT
Rubella virus	RBV-Rubi.	4ADI
Hantaan river virus	HRV-Hanta.	5LJX

Domain SSE	
Domain label	SSE labels
D1	B0,C0,D0,E0,F0,G0,H0,I0
D2	a, b, c , d, e, f, g, i, j
D3	A, B, C, E, F

Figure 3: **Class II fusion structures.** **a)** Taxonomy of structures used in this study. **b)** Breakdown of the structures used in this study, as well as their domain and label labels. **c)** Domain decomposition of DFV-Flavi..

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

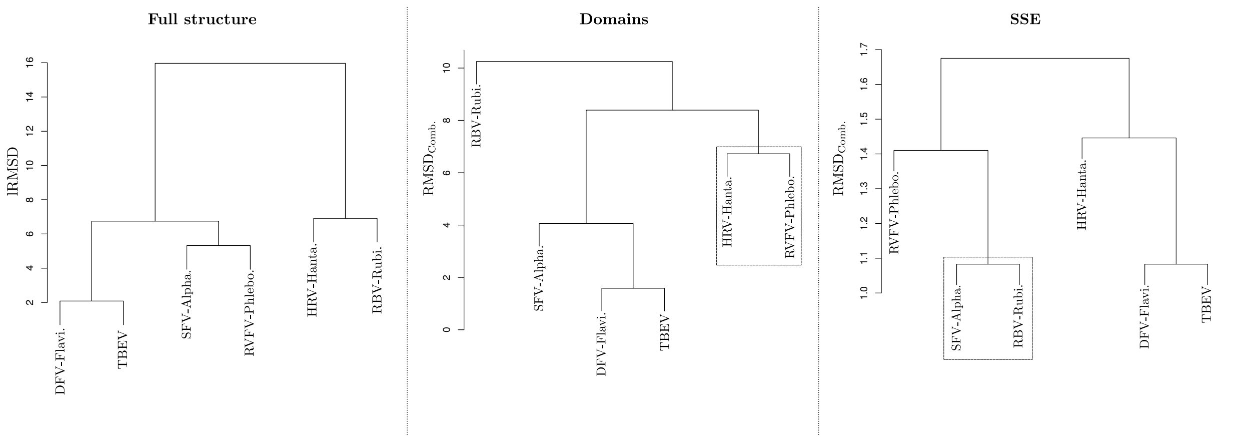


Figure 4: **RMSD$_{\mathbf{Comb.}}$ sharpens hierarchical clustering obtained for class II viral fusion proteins.** Complete linkage hierarchical clustering of the structures defined in Fig. 3. **(Left)** Clustering obtained upon processing distances from Tab. 1. Global lRMSD after aligning structures with the `Apurva` algorithm. **(Center)** Clustering obtained upon processing distances from Tab. 2. RMSD$_{\mathrm{Comb.}}$ using domains I, II and III. **(Right)** Clustering obtained upon processing distances from Tab. 3. RMSD$_{\mathrm{Comb.}}$ using motifs corresponding to SSE.

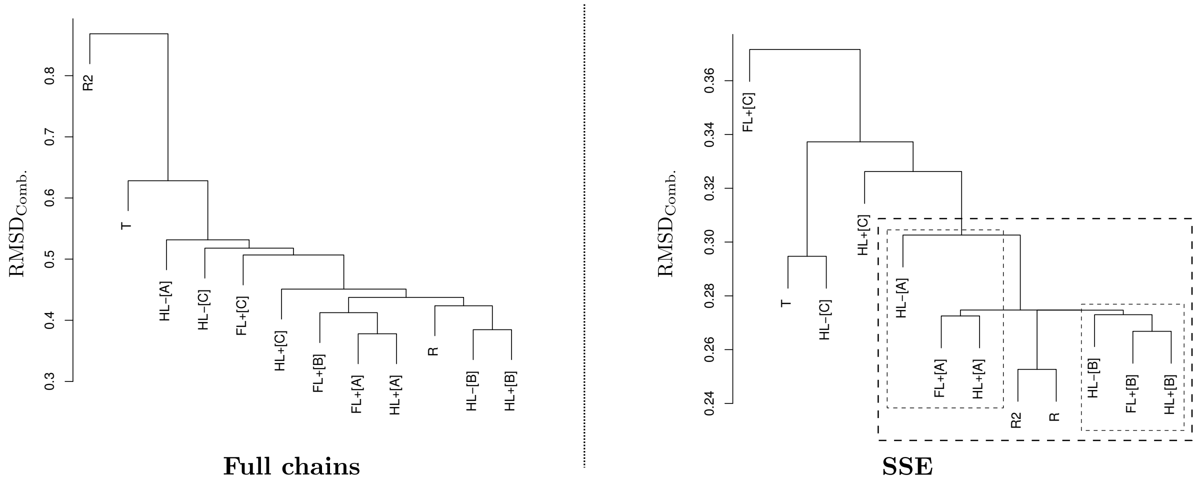


Figure 5: **Assigning quaternary structures of hemoglobin using $\alpha_1\beta_1$ dimers.** The goal is to check which similarity measures allow one to cluster coherently the newly reported conformations A, B, C of hemoglobin tetramers ([33] and Sec. 4.3.), assumed to adopt quaternary structures corresponding to the R2, R and T states. The displayed hierarchical clusterings were built using the single linkage scheme. **(Left)** Using RMSD$_{\mathrm{Comb.}}$ combining the lRMSD of the two chains α_1 and β_1. The hierarchical clustering obtained does not cluster coherently states A, B, C, and does not provide a coherent clustering with states R2, R and T either. **(Right)** Using RMSD$_{\mathrm{Comb.}}$ combining the RMSD$_{\mathrm{Comb.}}$ of the two chains α_1 and β_1, the former (resp. latter) based on the 7 (resp. 8) lRMSD between its helices. The clusters of conformations A, B and to a lesser extent C are well formed and coherent with the R2, R and T states.

bioRxiv preprint doi: https://doi.org/10.1101/379784; this version posted July 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

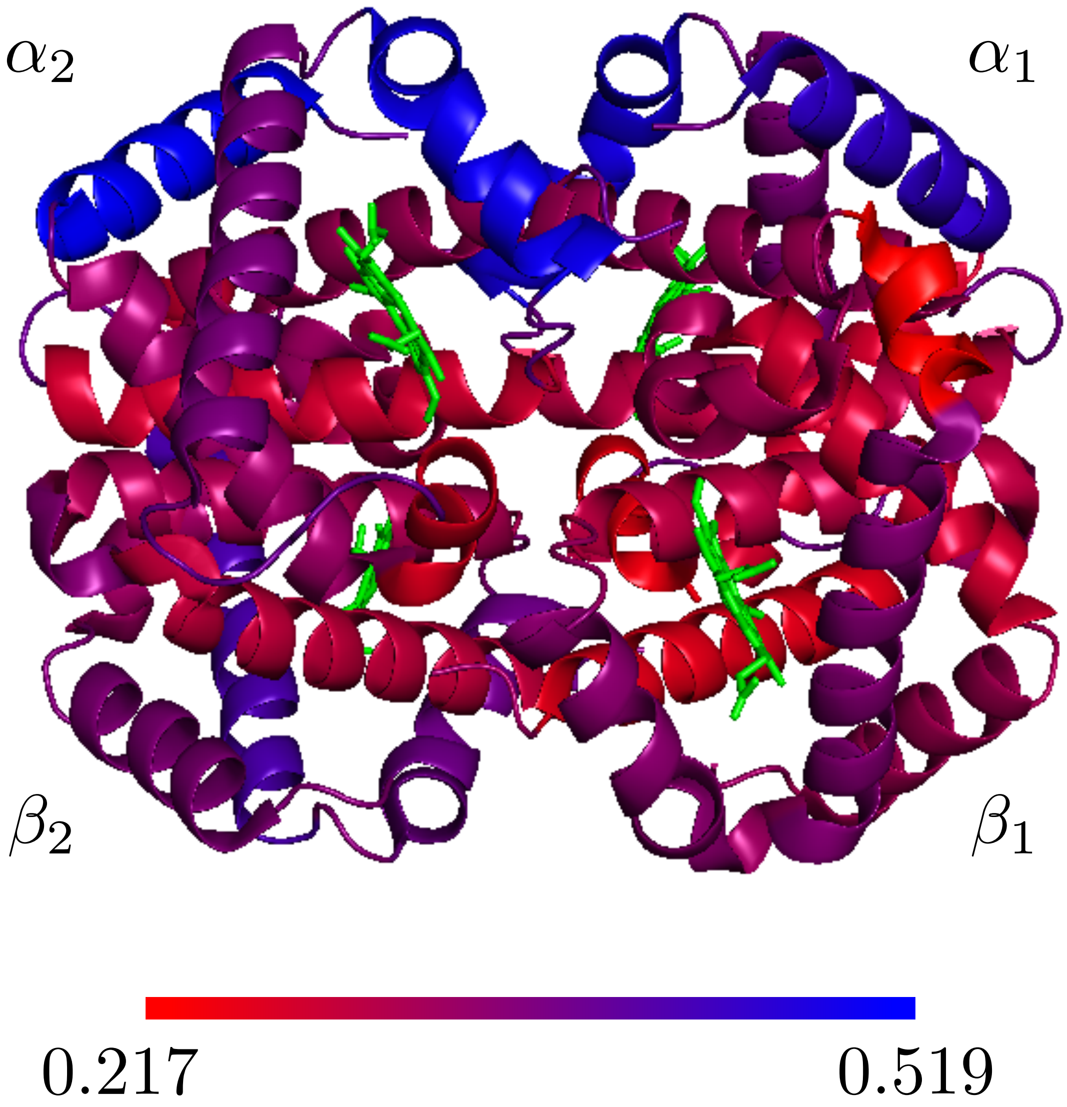


Figure 6: **Structural conservation of hemoglobin.** The α and β chains were respectively decomposed into 7 and 8 helices (Main text). For each helix, all pairwise lRMSD were computed using the 12 structures. Each helix was then color coded according to the gradient indicated. Visualization done with T conformation (pdbid: 2dn2).