



**HAL**  
open science

# The Symmetric Difference Distance: A New Way to Evaluate the Evolution of Interfaces along Molecular Dynamics Trajectories; Application to Influenza Hemagglutinin

Valentin Ozeel, Aurelie Perrier, Anne Vanet, Michel Petitjean

► **To cite this version:**

Valentin Ozeel, Aurelie Perrier, Anne Vanet, Michel Petitjean. The Symmetric Difference Distance: A New Way to Evaluate the Evolution of Interfaces along Molecular Dynamics Trajectories; Application to Influenza Hemagglutinin. *Symmetry*, 2019, 11 (5), pp.662. 10.3390/sym11050662 . hal-02322664

**HAL Id: hal-02322664**

**<https://hal.science/hal-02322664>**


Submitted on 13 Nov 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Article

# The Symmetric Difference Distance: A New Way to Evaluate the Evolution of Interfaces along Molecular Dynamics Trajectories; Application to Influenza Hemagglutinin

Valentin Ozeel <sup>1</sup>, Aurélie Perrier <sup>1,2</sup> , Anne Vanet <sup>1,3,4</sup>  and Michel Petitjean <sup>1,5,\*</sup> 

<sup>1</sup> E-pôle de Génoinformatique, CNRS UMR 7592, Institut Jacques Monod, F-75013 Paris, France; valentin.ozeel@gmail.com (V.O.); aurelie.perrier-pineau@univ-paris-diderot.fr (A.P.); anne.vanet@univ-paris-diderot.fr or vanet.anne@ijm.univ-paris-diderot.fr (A.V.)

<sup>2</sup> Chimie ParisTech, PSL Research University, CNRS, Institute of Chemistry for Life and Health Science (i-CLeHS, FRE 2027), F-75005 Paris, France

<sup>3</sup> Université Paris Diderot, Sorbonne Paris Cité, UFR Sciences du vivant, 5 rue Thomas Mann, F-75205 Paris CEDEX 13, France

<sup>4</sup> Pathologies de la réplication de l'ADN, Institut Jacques Monod, CNRS UMR 7592, CNRS, F-75013 Paris, France

<sup>5</sup> CMPLI, INSERM U1133 (BFA, CNRS UMR 8251), Université Paris Diderot, F-75205 Paris CEDEX 13, France

\* Correspondence: petitjean.chiral@gmail.com or michel.petitjean@univ-paris-diderot.fr

Received: 3 April 2019; Accepted: 8 May 2019; Published: 12 May 2019



**Abstract:** We propose a new and easy approach to evaluate structural dissimilarities between frames issued from molecular dynamics, and we test this methodology on human hemagglutinin. This protein is responsible for the entry of the influenza virus into the host cell by endocytosis, and this virus causes seasonal epidemics of infectious disease, which can be estimated to result in hundreds of thousands of deaths each year around the world. We computed the three interfaces between the three protomers of the hemagglutinin H1 homotrimer (PDB code: 1RU7) for each of its conformations generated from molecular dynamics simulation. For each conformation, we considered the set of residues involved in the union of these three interfaces. The dissimilarity between each pair of conformations was measured with our new methodology, the symmetric difference distance between the associated set of residues. The main advantages of the full procedure are: (i) it is parameter free; (ii) no spatial alignment is needed and (iii) it is simple enough so that it can be implemented by a beginner in programming. It is shown to be a relevant tool to follow the evolution of the conformation along the molecular dynamics trajectories.

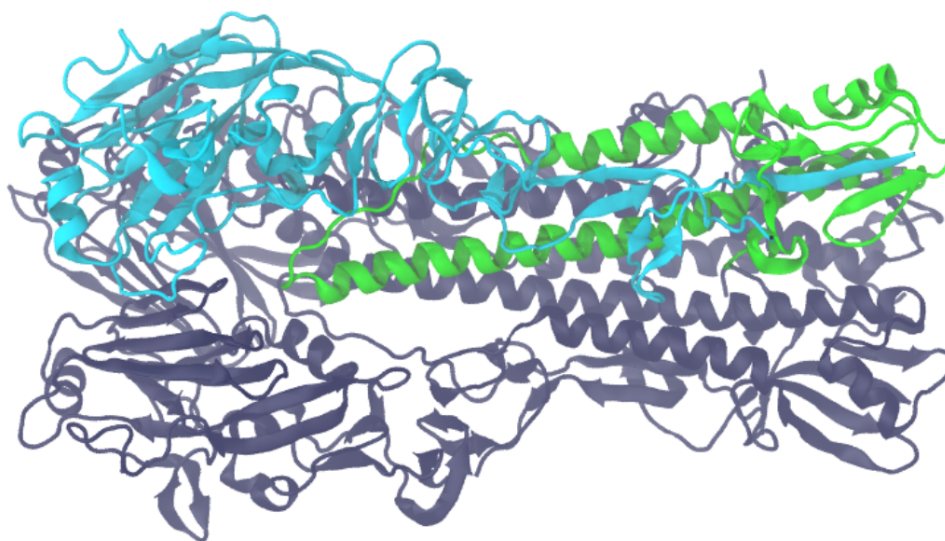
**Keywords:** macromolecular interfaces; PPI; symmetric difference distance; influenza hemagglutinin; molecular dynamics simulation

## 1. Introduction

Influenza virus causes seasonal epidemics of infectious disease and can even trigger pandemics. Inferring from U.S. data [1], these epidemics may cause hundreds of thousands of deaths each year around the world, mainly concerning the elderly, children, and immunosuppressed people, resulting in a real public health issue. Moreover, there were several severe influenza pandemics, such as the 1968 Hong Kong one [2] and the 1918 one, which resulted in about 20 millions deaths [3]. More recently, modeling studies attributed about 200,000 deaths to the 2009 influenza pandemic [4]. Initiation of virus infection involves multiple influenza Hemagglutinin (HA) binding [5] to sialic acids typically by  $\alpha 2,3$  or even  $\alpha 2,6$  linkages at the cell surface. The virus is then endocytosed through clathrin-dependent

pathways, and the fusion of the viral particle with the cell's membrane happens due to the acidification (about pH = 5) in the late endosomes. Indeed, this pH variation induces a folding change of HA homotrimers, resulting in the three fusion peptides' exposure to the cell's membrane [6]. HA is thus one of the two neutralizing antibody targets with neuraminidase, which is the second surface glycoprotein, allowing the virus to spread out from the cells [7].

In a recent modeling study to design new inhibitors against HA, two Molecular Dynamics (MD) simulations were performed, one at pH = 7 and one at pH = 5, this acidic pH being the one inducing a conformational change [8]. A motivation of this previous study was to compare the simulations at both pH values. Each protomer has two domains: HA1, responsible for the binding, and HA2 for the fusion, containing respectively 327 and 160 amino acid residues. The tridimensional structural data of HA were found in the Protein Data Bank (PDB), Code 1RU7, taken from [3] (see Figure 1). This structure is the one of A/Puerto Rico/8/1934 H1N1.



**Figure 1.** The tridimensional structure of the HA homotrimer (data found in the Protein Data Bank (PDB), Code 1RU7, taken from [3]). Each of the three protomers contains a subunit HA1 (in cyan in protomer 1) and a subunit HA2 (in green in protomer 1). HA1 domains are on the left. HA2  $\alpha$ -helices are on the right. The image was generated with PyMOL [9].

Starting from an initial conformation, 40 conformations were extracted every 0.5 ns for two performed simulations (see [8] for technical details about the protocol). The problem considered here is to evaluate, for each set of 41 conformations, the dissimilarities between pairs of conformations. For this purpose, we needed a simple procedure, easy to implement, and without spatial alignment because these latter generally involve unwanted contributions of meaningless parts of macromolecules, these parts being difficult to identify.

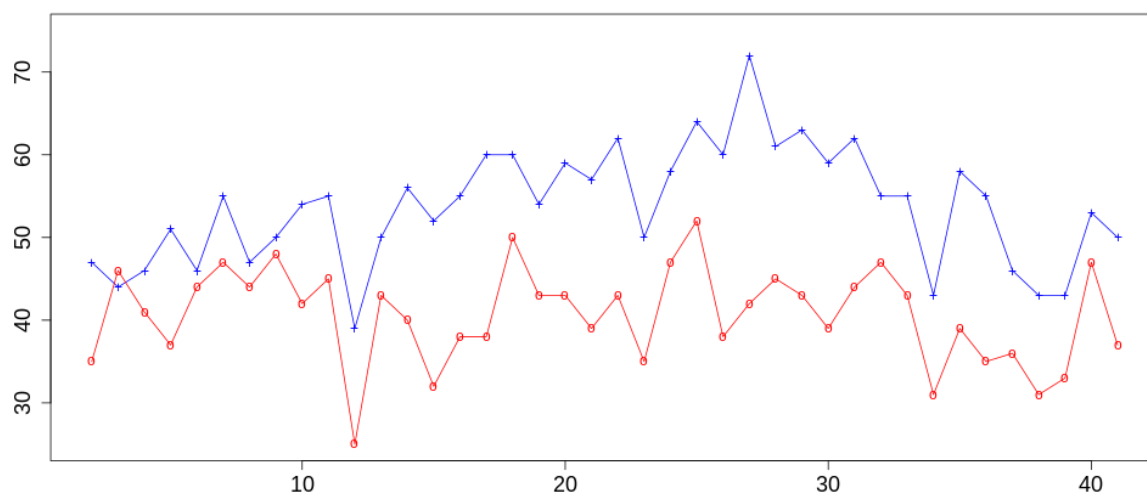
Usually, the dissimilarity between two conformers of the same macromolecule is measured through the computation of the RMSD (Root Mean Squared Deviation). For a macromolecule containing  $n$  heavy atoms, RMSD is the quadratic mean of the lengths of the  $n$  pairs of atoms ( $n$  atoms for the first conformer,  $n$  atoms for the second conformer), minimized for all rotations and translations of either of these two conformers. An analytical solution for this optimization problem is known: the optimal translation is such that the barycenter of the two conformers should coincide, and the optimal rotation is expressed with quaternions (see Appendix in [10] and Appendix A.5 in [11]). While the calculus of an RMSD is easy to compute, this quantity offers two drawbacks. The first one is the large numerical impact on the RMSD value from non-relevant parts of the macromolecule, such as the flexible ends of the backbone, or such as the irrelevant parts of the macromolecule far from the domain of interest for the experimentalist. For instance, in a protein–protein complex, the most relevant part is the protein–protein interface, because it is the area that can stand the weak bonds

responsible for the stability of the complex. This is also the case for a protein–ligand complex. In the case of HA, the relevant part of the trimer is the three interfaces between the pairs of protomers (each of the three pairs of protomers defines an interface), because they are the location of interprotomers' binding interactions. The second drawback of the RMSD is that, when it is computed over all heavy atoms, the side chains have an important numerical contribution to the RMSD value, although the location of the side chains is often considered to be meaningless due to their flexibility. For the reasons mentioned above, we restricted the evaluation of the dissimilarities between HA conformers generated by MD simulations to the  $C_\alpha$  atoms of the three polypeptidic chains' interfaces.

Thus, for each conformer, we have a set of  $C_\alpha$  atoms. The simplest way to evaluate the dissimilarity between two sets is to compute their Symmetric Difference Distance (SDD). The SDD between two sets having respectively  $n_1$  and  $n_2$  elements is  $n_1 + n_2 - 2n_{12}$ ,  $n_{12}$  being the number of elements of the intersection of the two sets [12]. The properties of a distance are recalled in Appendix A, including why they are useful. The computation of the SDD between the two sets of  $C_\alpha$  atoms associated with two conformations of our HA trimers (i.e., there is one set of  $C_\alpha$  atoms per trimer) is explained in Section 3. Here, the unit of this distance is the number of  $C_\alpha$  atoms. It is recalled that, opposite what is encountered during RMSD calculations, the sets of  $C_\alpha$  atoms defined by the interfaces are never pairwise associated, and in general, they have different cardinalities.

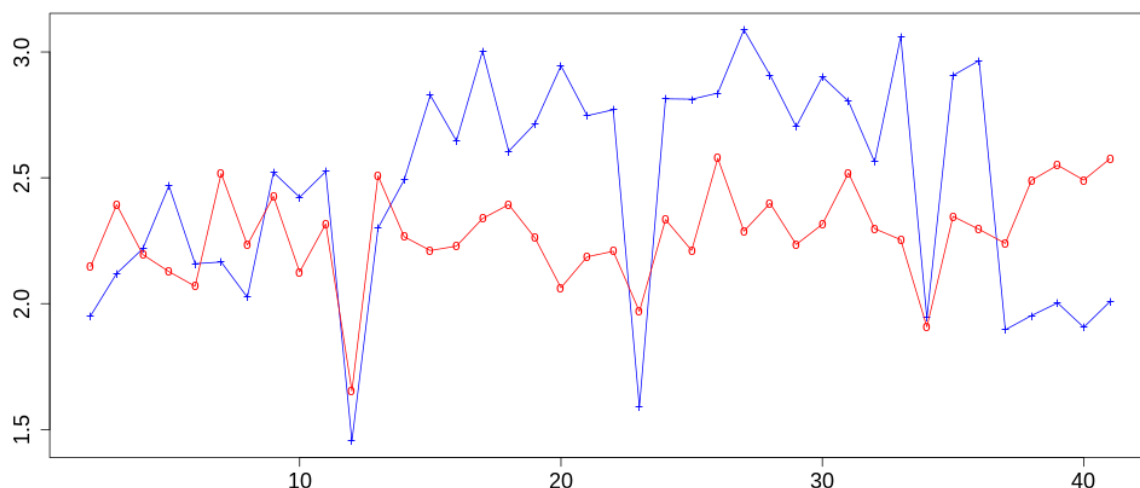
## 2. Results and Discussion

Starting from an initial conformation at time  $t_0$ , the MD simulation at pH = 7 generated 40 conformations, at times  $t_1, t_2, \dots, t_{40}$ , and similarly, the MD simulation at pH = 5 generated 40 conformations, the time step between two conformations being 0.5 ns [8]. Thus, there was a total of 41 conformations for each of these two MD simulations. We computed for the two pH values the 40 SDDs between the 40 generated conformations and their respective initial conformations at  $t_0$ . These distances values are plotted in Figure 2. The range of the distance values was 39–72 at pH = 7 (the unit of distances is the number of  $C_\alpha$  atoms). It was slightly smaller at pH = 5: 25–52. The maximal values were reached for the frames generated respectively at  $t_{26}$  and  $t_{24}$ . The maximal ratios of the SDDs values  $n_1 + n_2 - 2n_{12}$  to the total number of interface residues  $n_1 + n_2$  were 16.3% at pH = 7 and 17.2% at pH = 5. These moderately low values indicate that the generated conformations did not much differ from their respective initial ones at  $t_0$ . The correlation coefficient between the two sets of distances was 0.507. There was no reason to expect a high correlation coefficient since the SDD between the two initial conformations was 262, which was much higher than the range values at pH = 7 and at pH = 5.



**Figure 2.** The 40 SDDs, expressed as the number of  $C_\alpha$  atoms, between the initial conformation at time  $t_0$  and each of the 40 generated conformations by MD simulation, at pH = 7 (in blue) and at pH = 5 (in red). Time steps are in the abscissas. The image was generated with R [13].

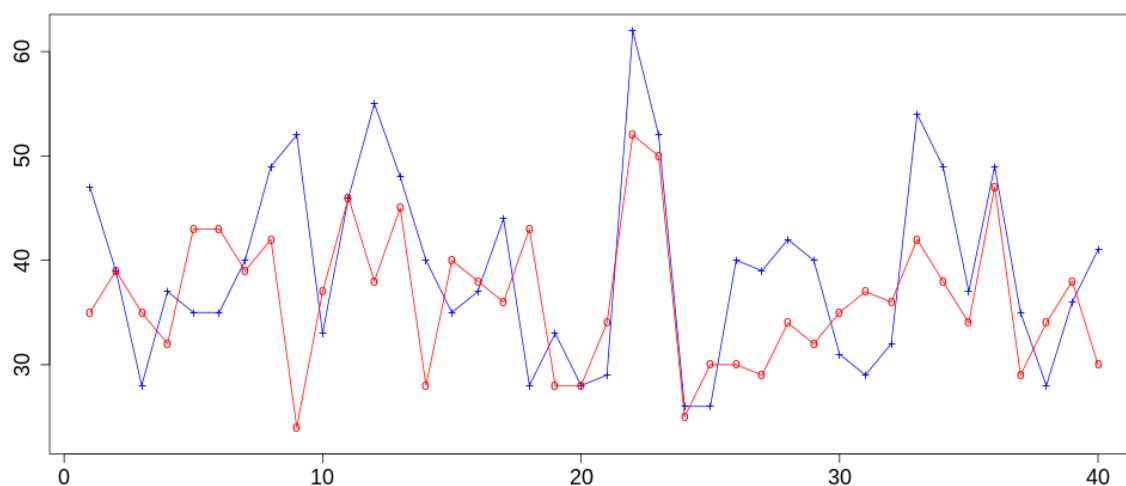
We also computed for the two pH values the 40 RMSD between these 40 conformations and their respective initial conformations at  $t_0$ : see Figure 3. The range of RMSD values was 1.45–3.09 Å at pH = 7. It was slightly smaller at pH = 5: 1.65–2.58 Å. The RMSD values were small relatively to the size of the HA: the radius of the smallest sphere enclosing the tridimensional structure of the 11,510 atoms of HA was 68.55 Å for 1RU7. These small RMSD values indicate that the generated conformations did not significantly differ from their respective initial ones at  $t_0$ . More importantly, the trend and the main peaks visible in Figure 3 are visible in Figure 2. This means that working on  $C_\alpha$  atoms at the interfaces rather than computing RMSD on all heavy atoms did not seem to cause any major information loss. Thus, our approach is pertinent.



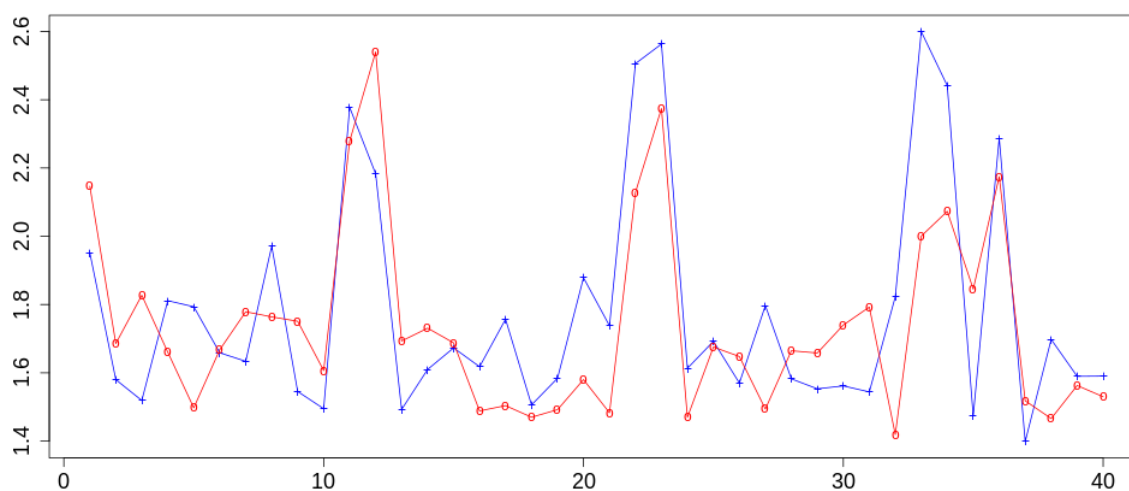
**Figure 3.** The 40 RMSD (in Å), taken over all heavy atoms, between the initial conformation at time  $t_0$  and each of the 40 generated conformations by MD simulation, at pH = 7 (in blue) and at pH = 5 (in red). Time steps are in the abscissas. The image was generated with R [13].

Then, we computed for the two pH values the SDDs between the 40 successive pairs of conformations. These distance values are plotted in Figure 4. The range of distance values was similar for both pH values: 26–62 at pH = 7 and 24–52 at pH = 5. The maximal values were reached between the frames generated at  $t_{22}$  for both simulations. The maximal ratios of the SDDs values  $n_1 + n_2 - 2n_{12}$  to the total number of interface residues  $n_1 + n_2$  were 13.9% at pH = 7 and 14.4% at pH = 5. These values were slightly lower than the ones of Figure 2 for pH = 7, and they were of the same magnitude as the ones of Figure 2 for pH = 5. For both simulations, the variations of the SDDs along time did not induce a trend to deviate more and more from the initial conformation, as shown in Figure 2. Thus, we can assume that the conformation is stable for the simulations at both pH values, at least for the interval of time considered (20 ns). The correlation coefficient between the two sets of distances was 0.468. There was no particular reason to expect a high correlation coefficient.

We also computed for the two pH values the RMSD values between the 40 successive pairs of conformations: see Figure 5. The range of RMSD values were similar for both pH values: 1.40–2.60 Å at pH = 7 and 1.42–2.54 Å at pH = 5. The RMSD values were small: we recall that the radius of the smallest sphere enclosing the 11,510 atoms of HA was 120.5 Å for 1RU7. Such a similarity between the range values at both pH is observed in Figure 4. The jumps observed in Figure 5 are also visible in Figure 4. Due to the relatively small ranges of RMSD values, the jumps seen Figure 5 are difficult to interpret. However, the maximal ratios of the SDDs values to the total number of interface residues that we already mentioned (around 14%) indicate that the interfaces may contribute significantly to the conformational changes along time. Again, working on  $C_\alpha$  atoms at the interfaces rather than computing RMSD values for all heavy atoms did not seem to cause any major information loss.



**Figure 4.** The SDDs, expressed as the number of  $C_{\alpha}$  atoms, between the 40 successive pairs of conformations generated by MD simulation, at pH = 7 (in blue) and at pH = 5 (in red). Time steps are in the abscissas. The image was generated with R [13].

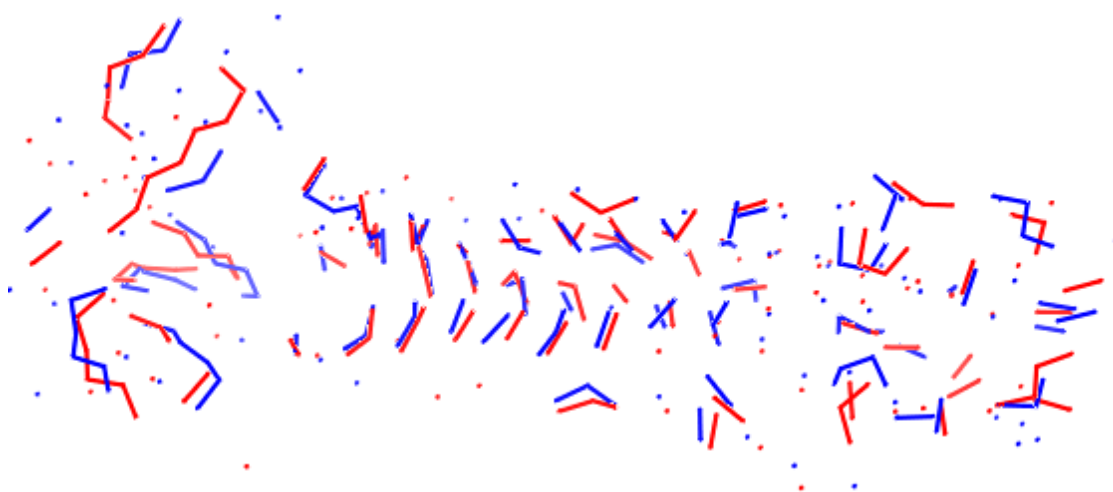


**Figure 5.** The RMSD values ( $\text{\AA}$ ), taken over all heavy atoms, between the 40 successive pairs of conformations generated by MD simulation, at pH = 7 (in blue) and at pH = 5 (in red). Time steps are in the abscissas. The image was generated with R [13].

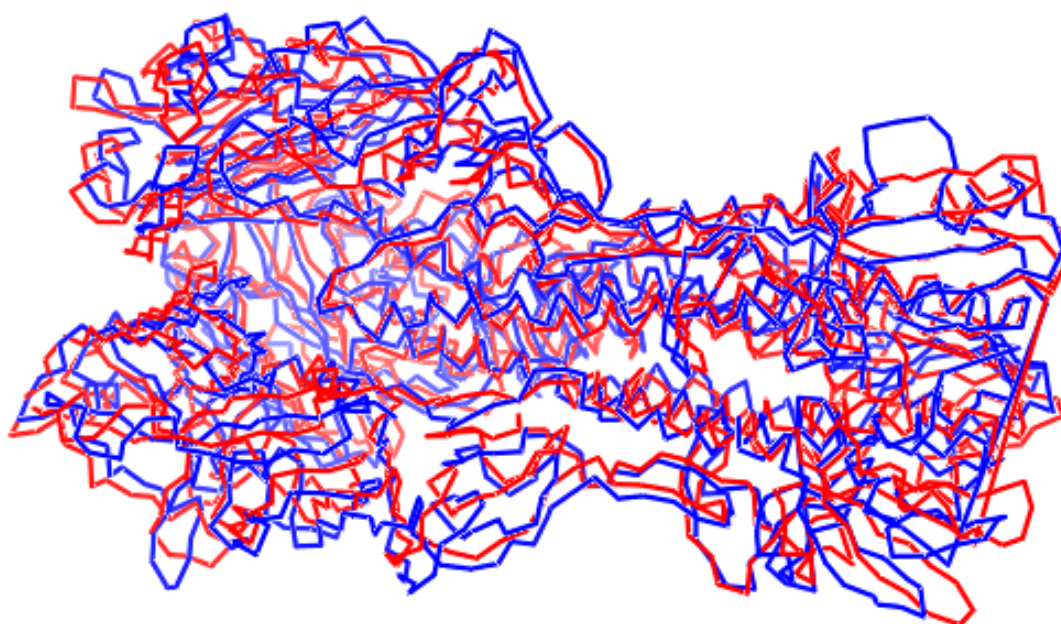
After having observed that the HA conformation remains stable over the 40-ns simulations, we needed to define a mean conformation for docking purposes. Indeed, considering such an average structure in the course of a docking simulation, rather than a structure determined from X-ray diffraction experiments, enables us to take into account the relaxation of the protein in the solvent, at  $T = 310 \text{ K}$  [8]. We defined this mean conformation with the algorithm described in [14]: it is the one that minimizes the sum of the squared distances to all other conformers. This mean conformation is the one of the frame generated at  $t_{30}$  for pH = 7, and it is the one of the frame generated at  $t_{15}$  for pH = 5. The residues at the interfaces of the mean conformers are given in Appendix B. The SDD between these two mean conformations was equal to 70  $C_{\alpha}$  atoms. This was few compared to the 1448 residues of each mean conformer, but it was larger compared to the respective 224 and 214 residues of the two interfaces, which had 184 residues in common. The RMSD between the 1448 pairs of  $C_{\alpha}$  atoms of these mean conformers was  $3.54 \text{ \AA}$ , while the largest atom-pair length was  $13.7 \text{ \AA}$  (obtained for the first serine of the HA2 domain). The optimal superposition of the two interfaces is shown in Figure 6. It was realized with the CSR freeware (<http://petitjeanmichel.free.fr/itoweb.petitjean.html>), which implements a non-parametric algorithm performing a spatial alignment without a cutoff distance and without any input pairwise correspondence [15].



We have also performed an optimal superposition of the mean frames at  $t_{30}$  for pH = 7 and at  $t_{15}$  for pH = 5, on the basis of all pairs of 11,510 heavy atoms. The display of this optimal superposition in Figure 7 has been restricted to the two sets of 1448  $C_{\alpha}$  atoms because displaying both sets of 11,510 heavy atoms would have been much more confusing. Even the display of the two sets of 1448 pairs of  $C_{\alpha}$  atoms in Figure 7 remains confusing. This is a general problem: the display of optimally-superposed macromolecules produced overloaded images, and viewing the result remained confusing even with interactive tools on the screen of a workstation. Thus, comparing the contents of Figures 6 and 7 shows another reason why our approach based on interfaces is useful: it produces lightened images of optimally-superposed structural data, which would have been cumbersome to generate with existing visualization tools.



**Figure 6.** The optimal superposition of the interfaces of the mean frames, respectively at  $t_{30}$  for pH = 7 (224  $C_{\alpha}$  atoms, in blue) and at  $t_{15}$  for pH = 5 (214  $C_{\alpha}$  atoms, in red). The image was generated with PyMOL [9].



**Figure 7.** The optimal superposition computed for all heavy atoms of the mean frames, respectively at  $t_{30}$  for pH = 7 (224  $C_{\alpha}$  atoms, in blue) and at  $t_{15}$  for pH = 5 (214  $C_{\alpha}$  atoms, in red). Only  $C_{\alpha}$  atoms are displayed. Each straight line separates two successive  $C_{\alpha}$  atoms in the backbone. The image was generated with PyMOL [9].

### 3. Methods

#### 3.1. Computation of Interfaces within Macromolecular Complexes

The full procedure to compute a SDD between two macromolecular complexes relies on the computation of a protein–protein interface within each complex. According to [16], there are three main families of methods to compute interfaces:

1. The cutoff method.
2. The loss of accessible surface area upon binding.
3. The Voronoi tessellation method.

Machine learning methods [17,18] are not considered here because they are irrelevant in our context. The cutoff method requires an arbitrary cutoff distance value, which may have a strong impact on the resulting computed interface [19]. Computing accessible surface areas requires fixing the values of atomic radii. Significantly different sets of radii values are found in the literature, depending on how they are defined [20–22]. It was shown that these radii have a considerable impact on surface areas [23,24]. The existence of parameters external to the input data and having a strong numerical impact on the results are drawbacks of these two families of methods. Thus, a non-parametric method is desirable. The third method, based on the Voronoi tessellation, in its original variant was parameter-free [25,26]. It was implemented in PROVAT software [27]. The full mathematical description of the Voronoi tessellation can be found in [28]. It is out of the scope of this paper. To summarize, each atom lies inside a convex polyhedral cell having its polygonal faces located at mid-distance from its neighboring atoms. Thus, two atoms are neighbors if their Voronoi cells share a common face. However, computing interfaces with the Voronoi method generates large cells, which induce the existence of meaningless long distances between neighboring atoms.

This is why we needed to compute interfaces with another free parameter method: we retained the PPIC software. PPIC is publicly available with its documentation on a repository located at <http://petitjeanmichel.free.fr/itoweb.petitjean.html>. The method implemented in PPIC was introduced only in a very recent preprint [29]. It extends the approach of [30] used to compute the interfaces in protein–ligand complexes. Its input is the tridimensional structure of a protein–protein complex or of a protein–ligand complex. This complex has two parts (molecule or macromolecule), named A and B. The algorithm has two steps:

1. Generate the first part of the interface, constituted by the non-redundant set of all nearest neighbors of the atoms of A among the atoms of B.
2. Generate the second part of the interface, constituted by the non-redundant set of all nearest neighbors of the atoms of B among the atoms of A.

Thus, the interface has two parts, i.e., two half interfaces, one in A and one in B. The roles of A and B are symmetric in the algorithm. From [31], it is known that each of these two parts is a subset of the half interface that would be computed by the Voronoi tessellation method. A nice consequence is that we do not observe anymore meaningless long distance pairs of neighboring atoms (i.e., one atom in A and one atom in B).

As a by-product, PPIC outputs the list of the RNNs (Reciprocal Nearest Neighbors). It is recalled that, in the Euclidean space, the nearest neighbor  $y$  of some point  $x$  is not always such that the nearest neighbor of  $y$  is  $x$ , e.g., consider three aligned points with abscissas  $x = 0$ ,  $y = 2$ , and  $z = 3$ , for which the nearest neighbor of  $x$  is  $y$ , while the nearest neighbor of  $y$  is  $z$ , not  $x$ .

When all atoms (or all heavy atoms) of the complex are used as inputs in the calculation of the interface, the pairs of atoms defined by the list of the RNN are a rough estimate of the location of potential interacting atom pairs. Moreover, to estimate the location of PPIs (Protein–Protein Interactions), looking at interacting atom pairs among the nearest neighbors makes more sense than looking at interacting atom pairs at farther distances than the nearest neighbors.



PPIC was shown to be effective to compute interfaces on a database of 1050 protein homo- and hetero-dimers ([29]; data from [32]). It was used to compute HIV2 protease–ligand interfaces ([29]; data from [33]). For both datasets, it was observed that the best agreement between the interface calculations produced on heavy atoms by PPIC and those produced by the cutoff method occurred for a cutoff near 3.6–3.7 Å (see Figure 1 in [29]). This is in agreement with the maximal donor–acceptor distance established in [34] at 3.9 Å. It is significantly smaller than the 4.5 Å cutoff distance between non-H atoms used by other authors [33,35,36].

### 3.2. Computation of Interfaces in Macromolecular Polymers

For a macromolecule containing two partners, A and B, the interface is constituted by two sets of atoms, one in A and one in B. In the case of a complex containing  $k$  polypeptidic chains, there are  $(k(k - 1))/2$  interfaces to compute. This is the case of HA, for which  $k = 3$ : there are three subsets, A, B, and C. Since the atoms of all chains are in the same macromolecular unit, we defined the global interface as the intersection of these  $(k(k - 1))/2$  interfaces as the non-redundant list of atoms was extracted. Therefore, in the case of a macromolecular polymer, the global interface is constituted by only one list of atoms, not two. This is a difference from the case of macromolecular complexes.

### 3.3. Evaluation of the Dissimilarity between Two Interfaces

An interface is constituted by one or two sets of atoms. In the case of protein–ligand complexes containing several ligands, the interface may contain more than two sets of atoms. Considering two interfaces containing one set of atoms each, their dissimilarity can be evaluated by their SDD [12].

When each of the two interface contains  $K$  sets,  $K \geq 1$ , and when these two  $K$ -tuples of sets are pairwise associated, there are  $K$  SDDs to compute. A distance between these two  $K$ -tuples can be defined to be the sum of these  $K$  SDDs. It can be checked from the definition in Appendix A that it defines indeed a distance on the set of these  $K$ -tuples.

When the two  $K$ -tuples of sets are not pairwise associated, which in fact means that we consider unordered  $K$ -tuples, a distance between these two  $K$ -tuples can be defined to be the minimal value of the sum of the  $K$  SDDs, this minimum being taken over all  $K!$  possible pairwise correspondences between the two  $K$ -tuples. It can be checked from the definition in Appendix A that this defines indeed a distance on the set of such  $K$ -tuples.

### 3.4. Comparison with Other Dissimilarity Measures

Several other structural dissimilarity criteria are encountered in the literature, which are qualified as metrics [37] (see also [38–40]), such as hydrogen bonds, distance from surface, the number of residues, or the number of heavy or polar atoms, or the number of waters in the vicinity of a specific region, RMSF (Root Mean Square Fluctuation), SASA (Solvent Accessible Surface Area), and gyration radius.

We outline here an ambiguity about the word “metric”: in many papers, it is used to name a structural dissimilarity criterion, which does not have the mathematical meaning mentioned in Appendix A. Even the dissimilarity criteria based on surfaces (e.g., SASA) cannot lead to defining a metric, except in the case of convex sets [41]. As mentioned at the end of Section 1, the SDD can be qualified as a metric [12]. The RMSD can be qualified as a metric when it is viewed as a distance induced by the Frobenius matrix norm, and this is why we compared the SDD with the RMSD. This happens in structural biology when we consider two matrices  $A_1$  and  $A_2$  of  $n$  lines and three columns, each matrix containing the spatial coordinates of a set of  $n$  points (the two sets of  $n$  points are thus pairwise associated). Setting  $A = A_1 - A_2$ ,  $\sqrt{n}\text{RMSD} = \sqrt{\text{trace}(A^t A)}$ , where  $A^t$  is the transpose of  $A$ .

### 3.5. Steps of the Methodology

The steps of the methodology to evaluate the evolution of interfaces along MD trajectories are summarized below. We assumed that the macromolecule of interest contained at least two partners, such as a protein and a ligand, two proteins or two protomers, etc.

1. Generate the frames of the MD simulation.
2. For each frame, generate the global interface with the procedure described in Section 3.2.
3. For each couple of successive frames, evaluate the dissimilarity between the interfaces with the SDD, as described in Section 3.3.
4. Follow the evolution of the interface along the trajectory using the SDD as a coordinate varying as a function of the time.

#### 4. Conclusions

Our approach to evaluate structural dissimilarities between frames issued from MD calculations has several advantages over the traditional ones involving either RMSD calculations or interface calculations:

- It is parameter free.
- No spatial alignment is needed, thus no non-trivial numerical solver is needed.
- The problem of molecular graph symmetries occurring in some contexts for residues Val, Leu, Arg, Phe, Tyr, Glu, and Asp, which is almost always neglected when computing RMSD values, does not exist in our approach.
- All the steps of our algorithm can be coded by a beginner in programming.
- The dissimilarity between interfaces is measured with a distance (see Appendix A).
- Unwanted contributions of meaningless parts of macromolecules can be discarded (e.g., disordered parts in macromolecules, etc.).
- Images of optimal superpositions of full macromolecules are too overloaded compared to those of optimal superpositions of interfaces.

The software PPIC implementing our non-parametric algorithm of interfaces calculations is publicly available for free at <http://petitjeanmichel.free.fr/itoweb.petitjean.html>. The user can optionally run the cutoff method or the Voronoi tessellation method.

As an example, we presented results about two MD simulations of influenza HA (PDB Code 1RU7, 1448 residues). Knowing from MD simulations at pH = 7 and pH = 5 that HA is stable, the magnitude of numerical values that we observed can serve as a first reference basis to discuss the results of MD simulations of other macromolecules or complexes. Our approach is neither claimed to overcome the previous ones, nor is it devoted to replacing them: it is just an additional tool devoted to helping structural biologists, which is simple to use and which can be easily reprogrammed by beginners.

**Author Contributions:** Methodology, M.P.; software, M.P.; computations, M.P. and V.O.; validation, A.P. and A.V.; writing, original draft preparation, M.P.; writing, and editing, A.P., A.V., and V.O.; supervision and project administration, A.P. and A.V.

**Funding:** This research received no external funding.

**Acknowledgments:** We are grateful to the reviewer who brought to us references [37–40].

**Conflicts of Interest:** The authors declare no conflict of interest.

#### Abbreviations

The following abbreviations are used in this manuscript:

HA	Hemagglutinin
MD	Molecular Dynamics
PDB	Protein Data Bank
PPI	Protein–Protein Interaction
RMSD	Root Mean Squared Deviation
RMSF	Root Mean Squared Fluctuation

RNN Reciprocal Nearest Neighbors  
 SASA Solvent Accessible Surface Area  
 SDD Symmetric Difference Distance

## Appendix A. Definition and Properties of Distances

For convenience, we recall the definition of a distance (see any textbook on analysis or topology).

**Definition A1.** A metric  $d$  (or distance function) on a set  $E$  is a function from the Cartesian product  $E \times E$  on  $\mathbb{R}$  satisfying the following four conditions:

1.  $d(x, y) = d(y, x)$  (symmetry)
2.  $d(x, x) = 0 \quad \forall x \in E$
3.  $d(x, y) = 0 \implies x = y \quad \forall (x \in E, y \in E)$
4.  $d(x, y) \leq d(x, z) + d(z, y) \quad \forall (x \in E, y \in E, z \in E)$  (triangle inequality)

Setting  $y = x$  in the triangle inequality and using the symmetry condition, we deduce a property that is flagged as an additional condition in some books:  $d(x, y) \geq 0 \quad \forall (x \in E, y \in E)$ .

A value taken by the distance function is called a distance. As long as no confusion can be made between the function and the value it takes, a distance function can be itself called a distance.

To understand why all four conditions of Definition A1 are of practical importance, imagine that one of them is missing and try to interpret numerical results from experiments:

- Removing the symmetry condition would mean that there exist two elements  $x$  and  $y$  such that  $d(x, y) \neq d(y, x)$ : understanding this result may be difficult.
- Removing the condition  $d(x, x) = 0$  would mean that there exists an element  $x$  such that  $d(x, x) > 0$ : what should one think about such an element?
- Many authors define dissimilarities between objects, although the third condition does not stand: nothing can be deduced when a null distance  $d(x, y) = 0$  is observed between two distinct element  $x$  and  $y$ , a really embarrassing situation.
- The triangle inequality is useful, as well: it would be difficult to understand a situation where three distinct elements  $x$ ,  $y$ , and  $z$  would be such that  $d(x, y) > d(x, z) + d(z, y)$ .

This is why, when possible, working with the distance appears to be a better choice than working with some other dissimilarity criterion.

## Appendix B. Interfaces Residues of the Mean Frames

It is recalled that, for each mean frame (one at pH = 7 and one at pH = 5), we computed the three interfaces between the three HA protomers, then we defined the final list of interface residues as the non-redundant set of residues involved in the union of these three interfaces.

The 224 residues at the interface of the mean frame for pH = 7 are:

V19, L20, E100, E103, K159, K161, S184, E194, N195, S199, V201, S203, N204, N206, R207, R208, T210, E212, I213, A214, E215, R216, P217, L232, K234, G236, T238, I240, E242; G324, L325, F326, G327, G331, F332, K362, S363, N366, G370, N373, K374, S377, K381, N383, Q385, K395, L396, K398, R399, M400, N402, L403, N404, K406, V407, G410, F411, D413, I414, W415, Y417, N418, L421, L422, L425, E426, R429, F433, S436, N440, E443, K444, S447, K450, N451, E455, I456, G457, G478, P483; V502, L503, E504, D577, E583, E586, K642, N678, S682, V684, S686, N687, Y688, N689, R690, R691, E695, I696, A697, E698, R699, P700; G807, L808, F809, G810, G814, F815, K845, N849, G853, N856, K857, S860, K864, M865, N866, K878, L879, K881, R882, M883, N885, L886, N887, K889, V890, D892, G893, F894, D896, I897, Y900, L904, L905, L908, E909, R912, F916, S919, N923, E926, K927, K933, E938, G940, P966; V985, L986, D1060, E1066, E1069, N1161, S1165, V1167, S1169, N1170, Y1171, N1172, R1173, R1174, T1176, E1178, A1180, E1181, R1182, P1183, K1200, F1226; G1290, L1291, F1292, G1293, A1296, G1297,

K1328, N1332, G1336, N1339, K1340, S1343, K1347, M1348, N1349, F1352, N1360, K1361, L1362, E1363, K1364, R1365, M1366, N1368, L1369, N1370, K1372, V1373, D1375, G1376, F1377, D1379, I1380, Y1383, N1384, L1387, L1388, L1391, E1392, R1395, F1399, S1402, N1406, E1409, K1410, S1413, K1416, N1417, K1420, E1421, I1422, G1423, G1444, D1447, Y1448.

The 214 residues at the interface of the mean frame for pH = 5 are:

V19, L20, G93, E100, N195, S199, V201, S203, N204, N206, R207, R208, T210, E212, I213, A214, E215, R216, P217, K234, E242; L325, F326, G331, F332, N366, G370, N373, K374, S377, K381, M382, N383, F386, A388, K391, K395, L396, K398, R399, M400, N402, L403, N404, K406, V407, D409, G410, F411, D413, I414, W415, Y417, N418, L421, L422, L425, R429, F433, S436, N437, N440, E443, K444, S447, K450, N451, E455, G457, N458;

V502, L503, E583, E586, N678, Y680, S682, V684, S686, N687, Y688, N689, R690, R691, T693, E695, I696, A697, E698, R699, P700, K717, E725, F743; G807, L808, F809, G810, A813, G814, F815, N849, G853, N856, S860, V861, K864, M865, N866, Q868, K874, K878, L879, K881, R882, M883, N885, L886, N887, K889, V890, G893, F894, D896, I897, W898, Y900, N901, L904, L905, L908, R912, F916, S919, N923, E926, K927, S930, K933, N934, E938, I939, G940, Y965, P966;

V985, L986, D1060, E1066, E1069, K1125, E1160, N1161, Y1163, S1165, V1167, S1169, N1170, Y1171, N1172, R1173, R1174, F1175, T1176, E1178, I1179, A1180, E1181, R1182, P1183; G1290, L1291, F1292, F1298, K1328, N1332, G1336, S1343, V1344, K1347, M1348, N1349, K1361, L1362, K1364, R1365, M1366, N1368, L1369, N1370, K1372, V1373, G1376, F1377, D1379, I1380, Y1383, N1384, L1387, L1391, R1395, F1399, S1402, N1406, E1409, K1410, S1413, K1416, N1417, K1420, E1421, I1422, G1423, Y1448.

## References

1. Thompson, W.W.; Weintraub, E.; Dhankhar, P.; Cheng, P.Y.; Brammer, L.; Meltzer, M.I.; Bresee, J.S.; Shay, D.K. Estimates of US influenza-associated deaths made using four different methods. *Influenza Other Respir. Viruses* **2009**, *3*, 37–49. [[CrossRef](#)] [[PubMed](#)]
2. Viboud, C.; Grais, R.F.; Lafont, B.A.P.; Miller, M.A.; Simonsen, L. Multinational Impact of the 1968 Hong Kong Influenza pandemic: Evidence for a smoldering pandemic. *J. Infect. Dis.* **2005**, *192*, 233–248. [[CrossRef](#)]
3. Gamblin, S.J.; Haire, L.F.; Russell, R.J.; Stevens, D.J.; Xiao, B.; Ha, Y.; Vasisht, N.; Steinhauer, D.A.; Daniels, R.S.; Elliot, A.; et al. The structure and receptor binding properties of the 1918 Influenza hemagglutinin. *Science* **2004**, *303*, 1838–1842. [[CrossRef](#)] [[PubMed](#)]
4. Simonsen, L.; Spreeuwenberg, P.; Lustig, R.; Taylor, R.J.; Fleming, D.M.; Kroneman, M.; Van Kerkhove, M.D.; Mounts, A.W.; Paget, W.J. GLaMOR Collaborating Teams. Global mortality estimates for the 2009 Influenza pandemic from the GLaMOR project: A modeling study. *PLoS Med.* **2013**, *10*, e1001558. [[CrossRef](#)]
5. Gamblin, S.J.; Skehel, J.J. Influenza hemagglutinin and neuraminidase membrane glycoproteins. *J. Biol. Chem.* **2010**, *285*, 28403–28409. [[CrossRef](#)]
6. Smrt, S.T.; Lorieau, J.L. Membrane fusion and infection of the Influenza hemagglutinin. In *Protein Reviews (Advances in Experimental Medicine and Biology, 966)*; Atassi, M.Z., Ed.; Springer: Singapore, 2017; Volume 18, pp. 37–54. [[CrossRef](#)]
7. Skehel J.J.; Wiley, D.C. Receptor binding and membrane fusion in virus entry: The Influenza hemagglutinin. *Annu. Rev. Biochem.* **2000**, *69*, 531–569. [[CrossRef](#)]
8. Perrier, A.; Eluard, M.; Petitjean, M.; Vanet, A. Design of new inhibitors against hemagglutinin of Influenza. *J. Phys. Chem. B* **2019**, *123*, 582–592. [[CrossRef](#)]
9. *The PyMOL Molecular Graphics System, Version 1.8.4.0*; Schrödinger, LLC: New York, NY, USA, 2016. Available online: <http://www.pymol.org> (accessed on 18 April 2019).
10. Petitjean, M. On the root mean square quantitative chirality and quantitative symmetry measures. *J. Math. Phys.* **1999**, *40*, 4587–4595. [[CrossRef](#)]
11. Petitjean, M. Chiral mixtures. *J. Math. Phys.* **2002**, *43*, 4147–4157. [[CrossRef](#)]
12. Deza, M.M.; Deza E. *Encyclopedia of Distances*; Springer: Berlin, Germany, 2009; p. 46. [[CrossRef](#)]
13. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018. Available online: <https://www.R-project.org/> (accessed on 18 April 2019).

14. Meslamani, J.E.; André, F.; Petitjean, M. Assessing the geometric diversity of cytochrome P450 ligand conformers by hierarchical clustering with a stop criterion. *J. Chem. Inf. Model.* **2009**, *49*, 330–337. [[CrossRef](#)] [[PubMed](#)]
15. Petitjean, M. Interactive maximal common 3D substructure searching with the combined SDM/RMS algorithm. *Comput. Chem.* **1998**, *22*, 463–465. [[CrossRef](#)]
16. Dequeker, C.; Laine, E.; Carbone, A. INTerface Builder: A fast protein–protein interface reconstruction tool. *J. Chem. Inf. Model.* **2017**, *57*, 2613–2617. [[CrossRef](#)]
17. Xue, L.C.; Dobbs, D.; Bonvin, A.M.J.J.; Honavar, V. Computational prediction of protein interfaces: A review of data driven methods. *FEBS Lett.* **2015**, *23*, 3516–3526. [[CrossRef](#)]
18. Wang, W.; Yang, Y.; Yin, J.; Gong, X. Different protein–protein interface patterns predicted by different machine learning methods. *Sci. Rep.* **2017**, *7*, 16023. [[CrossRef](#)]
19. de Vries, S.J.; Bonvin, A.M.J.J. How proteins get in touch: Interface prediction in the study of bio-molecular complexes. *Curr. Prot. Peptide Sci.* **2008**, *9*, 394–406. [[CrossRef](#)]
20. Bondi, A. Van der Waals volumes and radii. *J. Phys. Chem.* **1964**, *68*, 441–451. [[CrossRef](#)]
21. Allinger, N.; Zhou, X.; Bergsma, J. Molecular mechanics parameters. *J. Mol. Struct.* **1994**, *312*, 69–83. [[CrossRef](#)]
22. Gavezzotti, A. The calculation of molecular volumes and the use of volume analysis in the investigation of structured media and of solid-state organic reactivity. *J. Am. Chem. Soc.* **1983**, *105*, 5220–5225. [[CrossRef](#)]
23. Petitjean, M. On the analytical calculation of van der Waals surfaces and volumes: Some numerical aspects. *J. Comput. Chem.* **1994**, *15*, 507–523. [[CrossRef](#)]
24. Petitjean, M. Spheres unions and intersections and some of their applications in molecular modeling. In *Distance Geometry: Theory, Methods, and Applications*; Mucherino, A., Lavor, C., Liberti, L., Maculan, N., Eds.; Springer: New York, NY, USA, 2013; Chapter 4, pp. 61–83. [[CrossRef](#)]
25. Cazals, F.; Proust, F.; Bahadur, R.P.; Janin, J. Revisiting the Voronoi description of protein–protein interfaces. *Prot. Sci.* **2006**, *15*, 2082–2092. [[CrossRef](#)] [[PubMed](#)]
26. Bouvier, B.; Grünberg, R.; Nilges, M.; Cazals, F. Shelling the Voronoi interface of protein–protein complexes reveals patterns of residue conservation, dynamics, and composition. *Proteins* **2009**, *76*, 677–692. [[CrossRef](#)]
27. Gore, S.P.; Burke, D.F.; Blundell, T.L. PROVAT: A tool for Voronoi tessellation analysis of protein structures and complexes. *Bioinformatics* **2005**, *21*, 3316–3317. [[CrossRef](#)]
28. Edelsbrunner, H. Voronoi Diagrams. In *Algorithms in Combinatorial Geometry*; Brauer, W., Rozenberg, G., Salomaa, A., Eds.; Springer: Berlin, Germany, 1987; Chapter 13, pp. 293–334.
29. Laville, P.; Martin, J.; Launay, G.; Regad, L.; Camproux, A.-C.; de Vries, S.; Petitjean, M. A non-parametric method to compute protein–protein and protein–ligands interfaces. Application to HIV-2 protease-inhibitors complexes. *bioRxiv* **2018**, 498923. [[CrossRef](#)]
30. Cerisier, N.; Regad, L.; Triki, D.; Camproux, A.-C.; Petitjean, M. Cavity versus ligand shape descriptors: Application to urokinase binding pockets. *J. Comput. Biol.* **2017**, *24*, 1134–1137. [[CrossRef](#)]
31. Eppstein, D.; Paterson, M.S.; Yao, F.F. On nearest-neighbor graphs. *Discrete Comput. Geom.* **1997**, *17*, 263–282. [[CrossRef](#)]
32. Martin, J.; Regad, L.; Etchebest, C.; Camproux, A.-C. Taking advantage of local structure descriptors to analyze interresidue contacts in protein structures and protein complexes. *Proteins* **2008**, *73*, 672–689. [[CrossRef](#)]
33. Triki, D.; Cano Contreras, M.E.; Flatters, D.; Visseaux, B.; Descamps, D.; Camproux, A.-C.; Regad, L. Analysis of the HIV-2 protease’s adaptation to various ligands: Characterization of backbone asymmetry using a structural alphabet. *Sci. Rep.* **2018**, *8*, 710. [[CrossRef](#)]
34. McDonald, I.K.; Thornton, J.M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **1994**, *238*, 777–793. [[CrossRef](#)]
35. Gao, M.; Skolnick, J. iAlign: A method for the structural comparison of protein–protein interfaces. *Bioinformatics* **2010**, *26*, 2259–2265. [[CrossRef](#)]
36. Esque, J.; Oguey, C.; de Brevern, A.G. Comparative analysis of threshold and tessellation methods for determining protein contacts. *J. Chem. Inf. Model.* **2011**, *51*, 493–507. [[CrossRef](#)]
37. Mohammadiarani, H.; Shaw, V.S.; Neubig, R.R.; Vashisth, H. Interpreting hydrogen–deuterium exchange events in proteins using atomistic simulations: Case studies on regulators of G-protein signaling proteins. *J. Phys. Chem. B* **2018**, *122*, 9314–9323. [[CrossRef](#)] [[PubMed](#)]



38. Mohammadi, M.; Mohammadiarani, H.; Shaw, V.S.; Richard R. Neubig, R.R.; Vashisth, H. Interplay of cysteine exposure and global protein dynamics in small-molecule recognition by a regulator of G-protein signaling protein. *Proteins* **2019**, *87*, 146–156. [[CrossRef](#)] [[PubMed](#)]
39. Shaw, V.S.; Mohammadiarani, H.; Vashisth, H.; Neubig, R.R. Differential protein dynamics of regulators of G-protein signaling: Role in specificity of small-molecule inhibitors. *J. Am. Chem. Soc.* **2018**, *140*, 3454–3460. [[CrossRef](#)] [[PubMed](#)]
40. Mohammadiarani, H.; Vashisth, H. Insulin mimetic peptide S371 folds into a helical structure. *J. Comput. Chem.* **2017**, *38*, 1158–1166. [[CrossRef](#)] [[PubMed](#)]
41. Petitjean, M. Geometric molecular similarity from volume-based distance minimization: Application to saxitoxin and tetrodotoxin. *J. Comput. Chem.* **1995**, *16*, 80–90. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).