JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Quantifying Changes in Intrinsic Molecular Motion Using Support Vector Machines

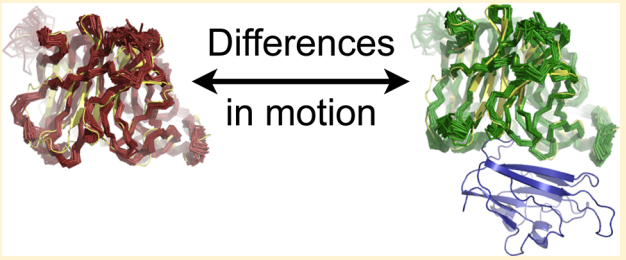
Ralph E. Leighty[†] and Sameer Varma*[,†,‡,§]

[†]Department of Cell Biology, Microbiology, and Molecular Biology, University of South Florida, Tampa, Florida 33620, United States

[‡]Department of Physics, University of South Florida, Tampa, Florida 33620, United States

[§]Institute of Pure and Applied Mathematics, University of California at Los Angeles, Los Angeles, California 90095, United States

ABSTRACT: The ensemble of three-dimensional (3-D) configurations exhibited by a molecule, that is, its intrinsic motion, can be altered by several environmental factors, and also by the binding of other molecules. Quantification of such induced changes in intrinsic motion is important because it provides a basis for relating thermodynamic changes to changes in molecular motion. This task is, however, challenging because it requires comparing two high-dimensional data sets. Traditionally, when analyzing molecular simulations, this problem is circumvented by first reducing the dimensions of the two ensembles separately, and then comparing summary statistics from the two ensembles against each other. However, since dimensionality reduction is carried out prior to ensemble comparison, such strategies are susceptible to artifactual biases from information loss. Here, we introduce a method based on support vector machines that yields a normalized quantitative estimate for the difference between two ensembles after comparing them directly against one another. While this method can be applied to any molecular system, including nonbiological molecules and crystals, here, we show how it can be applied to identify the specific regions of a paramyxovirus G protein that are affected by the binding of its preferred human receptor, Ephrin B2. This protein−protein interaction initiates the fusion of the virus with the host cell. Specifically, for every residue in the G protein, we obtain separately a quantitative difference between the ensemble of configurations they sample in the presence and in the absence of Ephrin B2. These ensembles were generated using molecular dynamics simulations. Rank-ordering and then mapping the residues that undergo the greatest change in motion onto the 3-D structure of the G protein reveals that they are clustered primarily on a single contiguous facet of the protein and include the set that is known experimentally to play a vital role in regulating viral fusion.



INTRODUCTION

The ensemble of three-dimensional (3-D) configurations exhibited by a molecule, that is, its intrinsic motion, is essential to its function and is correlated tightly with the properties of its environment.[1−12] Changes in intensive variables, such as temperature and ionic strength, modify the intrinsic motion of a molecule. Additionally, molecular motion also changes as a result of binding with other molecules, such as in ligand−substrate complexes or molecular assemblies. Moreover, the extent of the change depends upon multiple factors, including properties of the molecule and the nature of the perturbation or external potential. A quantitative characterization of such changes in molecular motion is important because it provides a basis to associate changes in thermodynamic properties directly with corresponding changes in molecular motion.

While quantifying the difference between two molecular configurations is now tractable,[13,14] obtaining a quantitative estimate for the difference between two ensembles of molecular configurations is challenging. For the motion of a n-particle molecule represented by m configurations, $\{a_n(\mathbf{x})\}_1^m$, the task of differentiating it from the molecule's reference state, $\{a_n(\mathbf{x}_0)\}_1^m$, involves comparing two $3n$-dimensional vector spaces. Traditionally, in a molecular simulation analysis, this problem is dealt with by first reducing the dimensions of the two ensembles separately, and then comparing the resulting summary statistics from the two ensembles against each other.[15] Dimensionality reduction is carried out by averaging over the n-space, or over the m-space, or over both n- and m-spaces. Averaging over the m-space yields time-averaged properties such as mean positions and fluctuations for the individual ensembles, which are then compared to obtain quantitative estimates for differences in the two ensembles (see, for example, refs 16 and 17). Averaging over the n-space involves some variation of particle clustering, such as a representation of a protein in terms of centers-of-masses of its constituent amino acids.

While such strategies for comparing ensembles have undoubtedly proven useful, they are prone to artifactual biases. This is because dimensionality reduction is carried out prior to comparison of ensembles; that is, information is left out or filtered even before ensembles are compared. Furthermore, the choice of an appropriate reduction scheme requires a priori

Received: August 7, 2012
Published: December 18, 2012

 © 2012 American Chemical Society

knowledge of the defining features of changes in molecular motion. For example, in cases where an external potential induces large structural changes in molecules, such as the folding or unfolding of protein domains, it can be assumed that the thermodynamic contribution from changes in vibrations are minor relative to contributions from molecular rearrangements and, in such cases, one may safely disregard changes in vibrations. Alternatively, for example, in studies where quantifying changes in dihedral distributions takes precedence over quantifying changes in molecular displacements, switching to frame-invariant dihedral internal coordinates reduces the dimensionality of the problem considerably, making quantification tractable. A recent study utilizes this dimension-reduction scheme in conjunction with the asymmetric Kullback–Leibler divergence in information theory to quantify changes in dihedral distributions.[18] However, when no a priori knowledge is available and if changes in no particular single mode of molecular motion, such as rotation, translation or fluctuation, are expected to dominate, a proper quantitative assessment of changes in molecular motion requires simultaneous consideration of all modes of motions. We encounter one such scenario when protein–protein interactions occur during the fusion of Nipah viruses with host cells.

Nipah viruses (family Paramyxoviridae, genus *Henipavirus*) are emerging zoonotic pathogens that are capable of causing illness and fatality in domestic animals and humans.[19−24] Their fusion with host cells is initiated by the binding of their G proteins (NiV-G) to specific Ephrin receptors of host cells. This binding triggers changes in NiV-G that ultimately activate the viral fusion protein. X-ray crystallography suggests that the ectodomain of NiV-G undergoes minor backbone rearrangements when it binds to its preferred Ephrin B2 receptor on the host cell.[25] The root-mean-square deviation (RMSD) between the X-ray coordinates of NiV-G backbone atoms of the bound and unbound states is 1.9 Å, and the backbone rearrangements occur primarily within certain loops near the Ephrin B2 binding site.[25] Microsecond-time scale all-atom molecular dynamics simulations carried out under physiological conditions also suggest similar minor backbone rearrangements in the ectodomain of NiV-G due to Ephrin B2 (Figure 1). These findings suggests that the Ephrin-induced changes to other modes of motion in NiV-G, including backbone orientations, side-chain orientations, and amino acid fluctuations, may also contribute to the activation of the fusion protein. A proper quantitative assessment of the overall change in the intrinsic motion of NiV-G, which considers all modes of motion simultaneously, therefore, requires that the ensembles representing the motion of NiV-G in both its bound and unbound states are compared against each other directly, without dimensional reduction of the phase space.

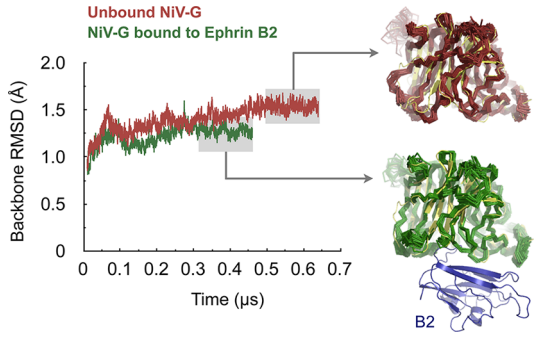


Figure 1. Time evolution of the RMSD of the backbone atoms of NiV-G calculated with respect to the X-ray structure of NiV-G cocrystallized with Ephrin-B2.[25] The two insets on the right show the backbone X-ray structures of NiV-G (drawn in yellow), and superimposed on them are twenty representative structures of NiV-G taken at regular intervals from the final 150 ns of the MD simulations.

Here, we introduce a method based on support vector machines[26−28] that compares directly two different ensembles of 3-D configurations and yields a normalized quantitative estimate for the difference between them. We then show how it can be applied to identify the region of NiV-G that undergoes the highest change in motion in response to Ephrin-binding. Specifically, we compare for each amino acid in NiV-G the ensemble of configurations it samples in the presence and in the absence of Ephrin and obtain a quantitative difference between them. The ensembles of configurations representing the motion of NiV-G were generated using molecular dynamics simulations. Rank-ordering and then mapping the set of amino acids that undergo the greatest change in motion onto the 3-D structure of NiV-G reveals that they are clustered primarily on a single contiguous facet of NiV-G that also interfaces with the Ephrin binding site. We expect this region of NiV-G to be involved in triggering viral fusion. This region also includes the set of amino acids that were identified recently using wet-lab experiments to be critical to viral fusion.[29]

METHODS

We describe first the molecular dynamics (MD) protocol used for generating the NiV-G configurational ensembles, and then the SVM-based method.

Molecular Dynamics. We carry out two separate all-atom MD simulations of the NiV-G ectodomain in ∼150 mM NaCl solution. While in one MD simulation the NiV-G ectodomain is bound to Ephrin-B2, NiV-G is ligand-free in the other. The starting configurations for both simulations are taken from the X-ray structure of the G-B2 complex (PDB ID: 2VSM).[25] The coordinates of the three amino acids of NiV-G, P208, V209, and V210, that were not resolved by X-ray crystallography were constructed using ModLoop.[30] The algorithm PDB2PQR[31] was used for adding and subsequently optimizing the positions of the protein hydrogen atoms in the gas phase. While the N-acetyl-D-glucosamine chains of NiV-G were removed, the water molecules that were resolved in the X-ray structure were retained. The MD unit cells corresponding to the B2-bound and unbound states of NiV-G contain a total of ∼31 000 and ∼41 000 water molecules, respectively.

Both MD simulations are carried out under isobaric–isothermal conditions, using Gromacs version 4.5.3.[32] Pressure is maintained at 1 bar using an extended-ensemble approach[33] and with a coupling constant of 1 ps and a compressibility of 4.5×10^{-5} bar^{-1}. An extended ensemble approach[34,35] is also used for maintaining temperature at 310 K, although a shorter coupling constant of 0.2 ps is employed. Electrostatic interactions are computed using the particle mesh Ewald scheme[36] with a Fourier grid spacing of 0.15 nm, a sixth-order interpolation, and a direct space cutoff of 10 Å. van der Waals interactions are computed explicitly for interatomic distance up to 10 Å. Charge neutrality of the two MD unit cells are maintained by selecting appropriate differences between the numbers of Na^+ and Cl^- ions. The bonds in proteins are constrained using the P-LINCS algorithm,[37] and the geometries of the water molecules are constrained using SETTLE.[38] These constraints permit the use of a large integration time step of 2

dx.doi.org/10.1021/ct300694e | J. Chem. Theory Comput. 2013, 9, 868−875

fs. The motion of the center of mass is reset every 0.2 ps. The protein and ions are described using OPLS-AA parameters,[39] and the water molecules are described using TIP4P parameters.[40]

We generate a 460 ns long trajectory of NiV-G in its bound state and a 640 ns long trajectory of NiV-G in its unbound state (Figure 1). The representative ensembles of configurations for the bound and unbound states of NiV-G are extracted from the final 150 ns of their respective trajectories.

Support Vector Machines. A support vector machine (SVM) is used for binary classification.[26−28] It is trained on a set of instances for which their corresponding group identities, $y_i = \pm 1$, are known. In principle, the configurational ensembles of a n-particle molecule, $\{a_n(\mathbf{x}_-)\}_1^m$ and $\{a_n(\mathbf{x}_+)\}_1^m$, can serve as training data, which can produce a classification function for the prediction of the group identity of an unclassified configuration.[41] This, however, is not our goal. Here, we utilize the properties of the classification function generated during training of the SVM to obtain a physically meaningful quantitative estimate for the difference between the molecular ensembles.

The training of a SVM involves determining a set of two hyperplanes,

$$y_i(\mathbf{w}\cdot\mathbf{x} - b) = 1 \tag{1}$$

where $\mathbf{w}$ is a vector normal to the hyperplane and $b/\|\mathbf{w}\|$ is the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$. In our case, these hyperplanes separate the $2m$ instances ($\mathbf{x}_\pm$) of a given atom a into two groups

$$y_i(\mathbf{w}\cdot\mathbf{x} - b) \geq 1 \tag{2}$$

Therefore, for each particle in the ensemble, the optimization task comprises of maximizing the distance between its corresponding hyperplanes, that is, $2/\|\mathbf{w}\|$, or minimizing $\|\mathbf{w}\|$ subject to the condition given by eq 2. This constrained optimization problem can be cast in terms of Lagrange multipliers, $0 \leq \alpha_i \leq C$, as

$$L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{2m} \alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i - b) - 1] \tag{3}$$

The square on $\|\mathbf{w}\|$ permits quadratic programming optimization and the $^1/_2$ coefficient is introduced for mathematical convenience. The regularization parameter C, influences the complexities of the hyperplanes, and will, therefore, affect the quantitative estimate of the difference between the ensembles. Regardless, the auxiliary function L in equation eq 3 is minimized with respect to $\|\mathbf{w}\|$ and b and maximized with respect to α_i. This implies

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{2m} \alpha_i y_i \mathbf{x}_i \tag{4}$$

and

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{2m} \alpha_i y_i = 0 \tag{5}$$

Substituting these results into eq 3 yields

$$L = \sum_{i=1}^{2m} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{6}$$

where

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i\cdot\mathbf{x}_j \tag{7}$$

is referred to as a kernel function. The overall optimization problem, therefore, consists of maximizing the auxiliary function L given by eq 6 with respect to α_i, such that $0 \leq \alpha_i \leq C$ and $\sum \alpha_i y_i = 0\ \forall\ i$. We utilize the sequential minimal optimization (SMO) algorithm for this purpose.[42,43]

The optimization of eq 6 produces two distinct sets of Lagrange multipliers, $\{\alpha_i\} = 0$ and $\{\alpha_i\} > 0$, that define the binary classification function. The hyperplanes that we intend to seek are defined by the subset of $\mathbf{x}_i$, whose corresponding $\alpha_i > 0$. These $\mathbf{x}_i$ are referred to as support vectors. The number of support vectors, s, required to partition the set of $2m$ instances of a particle is, therefore, bounded; that is, $2 \leq s \leq 2m$. In addition, it has been shown that the fraction $s/2m$ serves as an upper bound to the classification bootstrap-error.[44] In general, it can be expected that the higher the similarity between the ensembles of a particle, the larger the classification error. This suggests that the number of support vectors generated during construction of the classifier can be used as a measure of similarity between the particle ensembles. Consequently, we define a normalized quantity called the particle discriminability index,

$$\eta = 1 - \frac{s}{2m} \tag{8}$$

which is bounded, that is, $0 \leq \eta < 1$, and takes up a value closer to 1 as the similarity between particle ensembles decreases. To quantify the difference between molecular ensembles, η can be averaged over the particles in the molecule.

Note that, in the optimization of eq 6, the feature used for classifying $\mathbf{x}_i$ is its linear projection on other $\mathbf{x}_i$. The vectors $\mathbf{x}_i$ may not, however, be linearly separable in the Euclidean space. To circumvent this issue, we replace the dot product in eq 7 with an alternative kernel that permits classification in a transformed feature space. The primary advantage of such a "kernel-trick" is that it bypasses the determination of the explicit form of the function that transforms the data from the Euclidean to the desired feature space.[27,28,45] Such a kernel-trick, however, requires that the substituting kernel is an inner product in the transformed feature space. We chose a Gaussian radial distribution function as the substituting kernel, that is,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2) \tag{9}$$

which satisfies the aforementioned condition.[46] This kernel is chosen due to its stationarity and its performance in classification as compared to linear, polynomial, or sigmoidal kernels.[46] This kernel has also been used for constructing classifiers for compound libraries[47] as well as for classifying molecular configurations in chemical reactions.[41] The parameter γ in the kernel, which has units of $1/\text{Å}^2$, controls the width of the kernel and thereby the smoothness of the underlying nonlinear classifier. It represents essentially the influence of a given instance on its local environment. Smaller γ correspond to larger Gaussian widths, which imply a larger contribution of a given instance to the classification. The specific choice of γ, therefore, influences the classification and the resulting discriminability.

To choose appropriate values of γ as well as the regularization parameter C, we first construct model ensembles that represent the motions of the NiV-G atoms. We then

estimate η (using eq 8) between these ensembles for a range of γ and C values and select the set that minimizes the error with respect to analytical solutions. To construct model ensembles, we consider the probability distribution of distances $\|\mathbf{x}_i^a - \mathbf{x}_j^a\|$, where $\mathbf{x}_i^a$ and $\mathbf{x}_j^a$ are two position vectors that an atom a of NiV-G explores when NiV-G is simulated in the unbound state. We determine these probability distance distributions separately for all heavy atoms in NiV-G and then average them to obtain the probability distribution p_{ub}. We then determine separately another probability distribution p_b for the heavy atoms in NiV-G simulated in the bound state. We also determine a cross probability distribution p_{ub-b} of distances $(\|\mathbf{x}_i - \mathbf{x}_j\|)$ in which vectors $\mathbf{x}_i$ belong to the NiV-G atoms in the unbound state and vectors $\mathbf{x}_j$ belong to corresponding atoms of NiV-G simulated in the bound state. These three probability distributions are plotted in Figure 2. We find that, in the absence of the B2

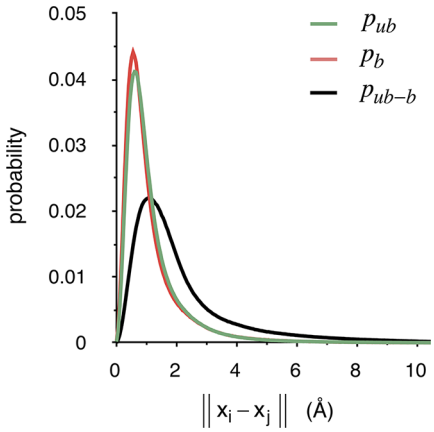


Figure 2. Probability distribution of distances $p(\|\mathbf{x}_i - \mathbf{x}_j\|)$ between the positions explored by the heavy atoms in NiV-G during MD simulations (see text). Molecular configurations from the final 150 ns of the MD trajectories were used for this analysis.

receptor, this distribution can be represented by a Gaussian radial distribution function with a standard deviation $\sigma_0 \sim 0.5$ Å. The nature of this distribution is perhaps a natural consequence of the central limit theorem. This distribution changes little when NiV-G binds B2. The cross probability distributions indicate that while the majority of the position displacements due to receptor binding are small, there exist a fraction of position displacements that are as large as 10 Å. The choice of γ and C parameters should, therefore, yield accurate η estimates for position displacements in the range 0−10 Å.

In light of the results above, we construct a single-particle ensemble $\{\mathbf{x}\}_1^m$ in which the $m = 2000$ coordinates are distributed normally, that is, $\mathbf{x} \in f(\mu_0, \sigma_0 = 0.5)$. We modify two properties of this Gaussian ensemble individually and construct two separate sets of ensembles. In one set of modified ensembles, $f(\mu_0 + \Delta\mu, \sigma_0)$, the mean of the Gaussian distribution is varied in unit increments of $\Delta\mu/\sigma_0 = \{1,2,...,20\}$. In the second set of modified ensembles, $f(\mu_0, \sigma)$, the standard deviation is varied in unit increments of the ratio $\sigma/\sigma_0 = \{2,3,...,15\}$. In the context of protein motion, these two modifications correspond, respectively, to changes in atomic mean-positions and fluctuations. We then estimate η between each pair of reference and modified ensembles as functions of $\gamma \in [10^{-3},10]$ and $C \in [1,10^8]$. We select the combination of γ and C that minimizes the mean absolute percent error (MAPE) with respect to analytical solutions. The analytical expressions for η between the reference and the modified ensemble are obtained by estimating the overlap between their Gaussian distributions. We find two combinations of γ and C that result in a MAPE $\leq$ 2.5. The combination $\gamma = 10^{-1}$ and $C = 10^2$ produces a MAPE of 2.50, and the combination $\gamma = 10^{-2}$ and $C = 10^4$ produces a slightly lower MAPE of 2.47. We choose the former combination over the latter due to its favorable run-time (smaller C). The computed η for this combination are compared against the analytical results in Figure 3.

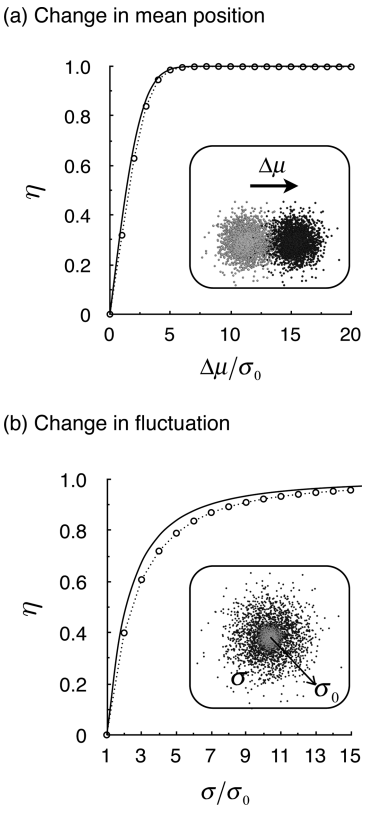


Figure 3. Comparison between computed and analytical estimates of discriminability. The computed estimates are depicted using circles connected by dashed lines, and the analytical estimates are drawn as solid lines (a) η is determined between two Gaussian ensembles that differ from each other only in their mean positions. The inset illustrates two such ensembles separated by $\Delta\mu/\sigma_0 = 4.0$. (b) η is determined between two Gaussian ensembles that differ from each other only in their fluctuation widths. The inset illustrates two such ensembles that have a fluctuation-width ratio $\sigma/\sigma_0 = 3$. The computed estimates in both cases are obtained using $\gamma = 0.1$ Å^{-2} and $C = 100$. The MAPE of the computed estimates with respect to analytical solutions is 2.50.

RESULTS AND DISCUSSION

While the method described in the previous section can be applied to any molecular system, here, we explore its sensitivity toward discriminating between configurational ensembles of amino acids. We then apply it to identify the specific regions of the NiV-G protein that are affected by the binding of its preferred human receptor, Ephrin B2.

Sensitivity Analysis. We explore how discriminability relates to modifications in three specific properties of configurational ensembles of amino acids: (a) the location of their geometric center, (b) their fluctuations, and (c) the orientation of their side-chains. We estimate the discriminability

dx.doi.org/10.1021/ct300694e | J. Chem. Theory Comput. 2013, 9, 868−875

of an amino acid by averaging over the discriminabilities of its n constituent atoms, that is, $\eta = 1 - \sum_1^n s/2mn$.

Figure 3 indicates that the method is sensitive to angstrom-level changes in particle mean positions and fluctuations. Since amino acid discriminability is estimated as an arithmetic average of particle discriminabilities, the method will also be sensitive to small changes in amino acid mean positions and fluctuations. However, η saturates over large changes, implying that its sensitivity toward detecting the difference between two separate large modifications will be small. Nevertheless, this, interferes little with our objective to filter out the amino acids undergoing large changes in motion.

To evaluate the sensitivity of the method toward changes in the side-chain orientation, we first construct model configurational ensembles of the amino acid side chains. For each amino acid, we replicate a representative configuration of its side chain over a 2-D Gaussian lattice. To modify the orientation, we rotate all the atoms about an axis perpendicular to the Gaussian lattice. The results of these calculations are plotted in Figure 4.

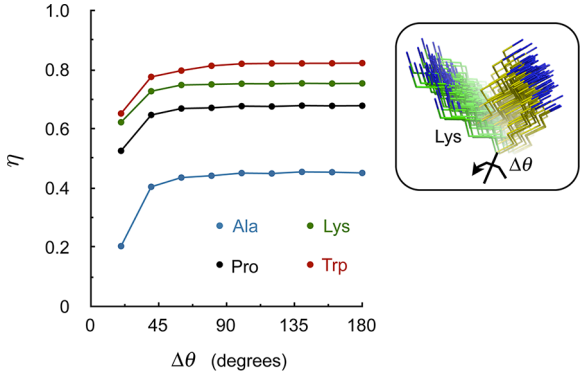


Figure 4. Relationship between amino acid discriminability and modifications in the relative orientation of side-chains. While tests on all natural amino acids were conducted, data belonging to only four representative amino acids are shown.

We find that, as expected, the computed discriminability depends on the size and the topology of the amino acid. We also find that the discriminability saturates for large rotations, suggesting that while the method is appropriate for detecting side chain rotations, it is less sensitive toward distinguishing between two different large side chain rotations.

Together, we find that the method is suitable to rank-order the amino acids of a given protein on the basis of their change in intrinsic motion, and determine subsequently the portions of a protein that undergo the highest change in motion in response to an external potential. In the next section, we apply this method to determine the specific portions of the NiV-G protein that are affected the highest by the binding of the Ephrin B2 receptor.

Application: G-Ephrin Interaction. To determine the effect of Ephrin-B2 on the intrinsic motion of NiV-G, we carry out two separate MD simulations of the NiV-G ectodomain in 150 mM NaCl solution: one in which it is complexed with Ephrin-B2 (bound state) and the other in which Ephrin-B2 is absent (unbound state). The representative ensembles of configurations for the bound and unbound states of NiV-G are extracted from the final 150 ns of their respective trajectories.

For each residue in NiV-G, we calculate the discriminability between its representative ensembles in the bound and unbound states, that is, between $\{a_n(\mathbf{x}_b)\}_1^m$ and $\{a_n(\mathbf{x}_{ub})\}_1^m$, where n denotes its non-hydrogen heavy atoms and m the number of its representative configurations. Prior to the calculation of the discriminability, the representative configurations of NiV-G are least-squares fitted on to the X-ray coordinates of NiV-G in the bound state. This is necessary to prevent the biasing of η against whole molecule rotation and center-of-mass motion. The results of these calculations are shown in Figure 5a. We find that residue discriminability ranges from 0.2574 to 0.9992 and appears to be distributed randomly over the primary sequence. Nevertheless, rank-ordering the residues on the basis of their discriminability and mapping those that fall within the top and bottom 25% categories on the X-ray structure brings out a discernible pattern in their distribution (Figure 5b). The residues that fall within the top and bottom 25% categories are clustered into distinct 3-D regions of the protein.

Note that, in these analyses, each of the two ensembles are represented by $m = 2500$ configurations. These configurations are extracted at regular intervals (60 ps) from the final 150 ns trajectory of each simulation. Increasing the number of representative configurations by a factor of 2 affects the discriminability rank-order minimally. We deduce this from the

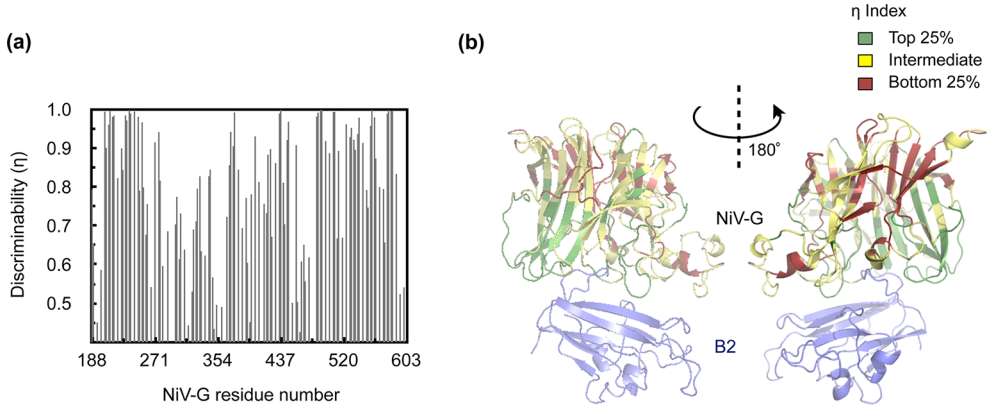


Figure 5. Effect of Ephrin-B2 binding on the intrinsic motion of NiV-G. (a) Residue-wise discriminability between the bound and unbound states of NiV-G. (b) The amino acids of NiV-G are rank-ordered on the basis of their discriminability shown in part a, and classified into three separate categories that are color-coded on the X-ray structure[25] of NiV-G. We find that the amino acids of NiV-G that undergo the highest change in intrinsic motion upon Ephrin binding are generally clustered on one facet.

dx.doi.org/10.1021/ct300694e | J. Chem. Theory Comput. 2013, 9, 868−875

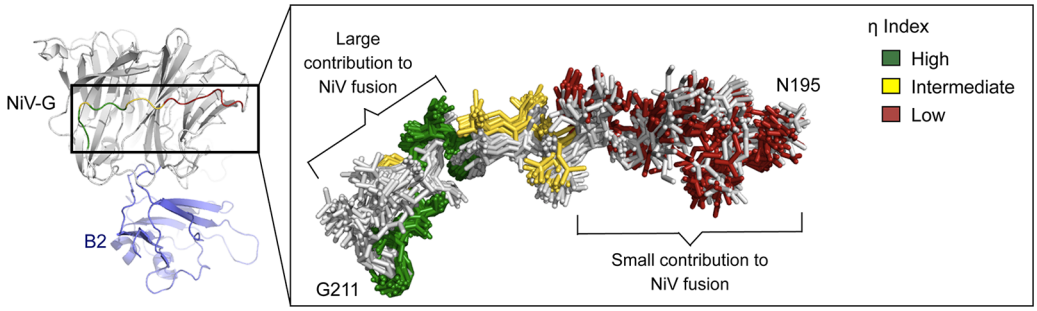


Figure 6. Effect of Ephrin-B2 binding on the intrinsic motion of a specific loop of NiV-G, NQILKPKLISYTLPVVG, and its relationship with alanine-scanning mutagenesis experiments.[29] Twenty representative configurations of the segment, ten each from the MD simulation of NiV-G in its Ephrin-bound and unbound states, are shown superimposed on each other. While the ten configurations from the simulation of NiV-G in its unbound states are colored gray, the ten configurations of NiV-G in its Ephrin-bound state are color-coded according to their discriminability index. We find an exact correspondence between the portions of the loop that have a high discriminability index, that is, those that undergo a high change in intrinsic motion, and those that were shown from experiments to contribute significantly to viral fusion.

large positive value of the Spearman correlation coefficient computed between the rank-orders of the two analyses, that is, $\rho(m,2m) = 0.9996$. The result stays the same when the number of representative configurations in each of the two ensembles is decreased by a factor of 2 ($\rho(m,m/2) = 0.9993$).

It is also interesting to note that the clustering pattern produced by the SVM-based method is different from that produced by traditional analyses schemes. While the SVM-based method compares high-dimensional data from the two simulations directly, the traditional analyses schemes compare summary statistics from the two simulations. One of the traditional ways to rank-order residues that are affected by Ephrin-binding is on the basis of backbone center-of-mass displacements. We compute for each residue its backbone center-of-mass displacement by calculating the distance between its corresponding mean positions in the two simulations, that is, $d = \|\langle \mathbf{x}_{ub} \rangle - \langle \mathbf{x}_{b} \rangle\|$. We find a moderate correlation of $r(\eta,d)$=0.7837 between the rank-orders associated with discriminability and center-of-mass displacements. This implies that while the core of the rank-ordering is due to backbone structural change, the two modes of motion excluded in the center-of-mass displacement analysis, that is, amino acid fluctuations and their orientations in Euclidian space, also contribute to rank-ordering.

Another scheme to rank-order residues is on the basis of differences in root-mean-square fluctuations, ΔRMSF. We compute ΔRMSF for each residue by first estimating its RMSF in the bound state and then subtracting it from its corresponding value in the unbound state, that is, $\Delta\text{RMSF} = |\text{RMSF}_{ub} - \text{RMSF}_{b}|$. The correlation between the rank-orders associated with residue-wise discriminability and ΔRMSF is $r(\eta,\Delta\text{RMSF}) = 0.2445$. We find that these rank-orders are positively correlated, which suggests that while ΔRMSF certainly contributes to rank-ordering residues, rank-ordering changes in intrinsic motion based simply on ΔRMSF can be misleading.

In the SVM-based analysis, we find that the residues that fall within the top 25% category are not just restricted to the portion of NiV-G that interfaces with Ephrin directly (Figure 5b). They span the protein surface contiguously from the Ephrin-interface to regions over 2 nm away from the interface. These residues could, therefore, participate directly in signal transduction and, in fact, could belong to the allosteric pathway that NiV-G utilizes to transmit the Ephrin-binding signal to the viral fusion protein, NiV−F.[19,22,23,48,49] Recent alanine-scanning mutagenesis experiments[29] identify a stretch of amino acids I203−G211 that is crucial to viral fusion (Figure 6). These experiments also identify an adjacent stretch of amino acids N195−L202 that belongs to the same solvent-exposed loop but, which, curiously, plays a minor role in fusion. We find from our analysis that while the former stretch of amino acids undergoes a large change in intrinsic motion due to Ephrin binding, the intrinsic motions of the latter stretch are affected minimally. Notice from the superimposed snapshots in Figure 6 that the changes in intrinsic motion do not just comprise of backbone displacements but also include changes in fluctuations and side-chain orientations. This strong correspondence between our analysis and experiments is promising and provides direct biophysical insight underlying experimental findings.

CONCLUSIONS

Here, we present a SVM-based method to estimate quantitatively the differences between two ensembles of molecular configurations. Its primary advantage over traditional approaches is that it does not require a priori reduction in phase space but rather compares high-dimensional data directly. While this method can be applied to any molecular system, we test sensitivity against model configurational ensembles of amino acids, where we find that it is sensitive to angstrom-level differences in the relative positions, fluctuations, and orientations of amino acids. In addition, the method does not bias discriminability with respect to displacements or fluctuation-changes toward any particular amino acid. However, we find that its discriminability saturates over large perturbations in ensemble properties, which implies that the method is less sensitive toward discriminating between two separate large perturbations. Nevertheless, it is suitable to rank-order the amino acids of a given protein on the basis of their relative change in intrinsic motion. We apply the method to rank-order the amino acids of a viral protein NiV-G on the basis of their change in intrinsic motion due to the binding of its preferred receptor, Ephrin-B2. We identify distinct clusters in NiV-G whose intrinsic motions respond strongly to Ephrin binding, which also serve to explain the observations from recent wet-lab experiments.

AUTHOR INFORMATION

Corresponding Author

*E-mail: svarma@usf.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We sincerely thank Klaus Müller, Sagar Pandit, David Rabson, and Bradley Cramer for stimulating discussions. We also acknowledge the use of the services provided by Research Computing at USF. This research was supported by a startup grant from USF and by the National Science Foundation through XSEDE resources provided by the XSEDE Science Gateways program.

■ REFERENCES

(1) Koshland, D. E. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **1958**, *44*, 98−104.
(2) Huber, R.; Bennett, W. S., Jr. Functional significance of flexibility in proteins. *Biopolymers* **1983**, *22*, 261Ð279.
(3) Kumar, S.; Ma, B.; Tsai, C. J.; Sinha, N.; Nussinov, R. Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Sci.* **2000**, *9*, 10−19.
(4) Gutteridge, A.; Thornton, J. M. Conformational changes observed in enzyme crystal structures upon substrate binding. *J. Mol. Biol.* **2005**, *346*, 21−28.
(5) Dyson, H. J.; Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell. Biol.* **2005**, *6*, 197−208.
(6) Tobi, D.; Bahar, I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 18908−18913.
(7) Henzler-Wildman, K. A.; Lei, M.; Thai, V.; Kerns, S. J.; Karplus, M.; et al. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* **2007**, *450*, 913−916.
(8) Watt, E. D.; Shimada, H.; Kovrigin, E. L.; Loria, J. P. The mechanism of rate-limiting motions in enzyme function. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 11981−11986.
(9) Henzler-Wildman, K.; Kern, D. Dynamic personalities of proteins. *Nature* **2007**, *450*, 964−972.
(10) Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L. Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.* **2008**, *18*, 756−764.
(11) Bakan, A.; Bahar, I. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 14349−14354.
(12) Bahar, I.; Lezon, T. R.; Yang, L. W.; Eyal, E. Global Dynamics of proteins: Bridging between structure and function. *Annu. Rev. Biophys.* **2010**, *39*, 23−42.
(13) Damm, K. L.; Carlson, H. A. Gaussian-weighted RMSD superposition of proteins: A Structural comparison for flexible proteins and predicted protein structures. *Biophys. J.* **2006**, *90*, 4558−4573.
(14) Wolfe, K. C.; Chirikjian, G. S. Quantitative comparison of conformational ensembles. *Entropy* **2012**, *14*, 213−232.
(15) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Academic Press: New York, 2007.
(16) Varma, S.; Chiu, S. W.; Jakobsson, E. The influence of amino acid protonation states on molecular dynamics simulations of a bacterial porin OmpF. *Biophys. J.* **2006**, *90*, 112−123.
(17) Varma, S.; Teng, M.; Scott, H. L. Non-intercalating contact points create asymmetry between bilayer leaflets. *Langmuir* **2012**, *28*, 2842−2848.
(18) McClendon, C. L.; Hua, L.; Barreiro, G.; Jacobson, M. P. Comparing conformational ensembles using the Kullback−Leibler divergence expansion. *J. Chem. Theory Comput.* **2012**, *8*, 2115−2126.
(19) Smith, E. C.; Popa, A.; Chang, A.; Masante, C.; Dutch, R. E. Viral entry mechanisms: The increasing diversity of paramyxovirus entry. *FEBS J.* **2009**, *276*, 7217−7227.
(20) Ksiazek, T. G.; Rota, P. A.; Rollin, P. E. A review of Nipah and Hendra viruses with an historical aside. *Virus Res.* **2011**, *162*, 173−183.
(21) Bowden, T. A.; Jones, E. Y.; Stuart, D. I. Cells under siege: Viral glycoprotein interactions at the cell surface. *J. Struct. Biol.* **2011**, *175*, 120−126.
(22) Lee, B.; Ataman, Z. A. Modes of paramyxovirus fusion: a Henipavirus perspective. *Trends Microbiol.* **2011**, *19*, 389−399.
(23) Steffen, D. L.; Xu, K.; Nikolov, D. B.; Broder, C. C. Henipavirus mediated membrane fusion, virus entry, and targeted therapeutics. *Viruses* **2012**, *4*, 280−309.
(24) Talekar, A.; Sengupta, U.; Moscona, A.; Glickman, F.; Briese, T.; et al. Rapid development of treatments and screening systems for emerging viral diseases. *PLoS One [Online]* **2012**, *7*, e30538.
(25) Bowden, T. A.; Ariscescu, A. R.; Gilbert, R. J. C.; Grimes, J. M.; Jones, E. Y.; et al. Structural basis of Nipah and Hendra virus attachment to their cell-surface receptor ephrin-B2. *Nat. Struct. Mol. Biol.* **2008**, *15*, 567−572.
(26) Cortes, C.; Vapnik, V. Support vector networks. *Mach. Learn.* **1995**, *20*, 273−297.
(27) Schölkopf, B.; Burges, C.; Smola, A. *Advances in Kernel Methods: Support Vector Learning*; MIT Press: Cambridge, MA, 1999.
(28) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, 2000.
(29) Aguilar, H. C.; Ataman, Z. A.; Aspericueta, V.; Fang, A. Q.; Stroud, M.; et al. A novel receptor-induced activation site in the Nipah virus attachment glycoprotein (G) involved in triggering the fusion glycoprotein (F). *J. Biol. Chem.* **2009**, *284*, 1628−1635.
(30) Fiser, A.; Do, R. K. G.; Sali, A. Modeling of loops in protein structures. *Protein Sci.* **2000**, *9*, 1753−1773.
(31) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; et al. PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **2007**, *35*, W522−W525.
(32) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(33) Parrinello, M.; Rahman, A. Polymorphic transition in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7128−7190.
(34) Nose, S.; Molecular-dynamics, A. Method for simulations in the canonical ensemble. *Mol. Phys.* **1984**, *52* (2), 255−268.
(35) Hoover, W. G. Canonical dynamics—Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695−1697.
(36) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An NLog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(37) Hess, B. P-LINCS: A parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 116−122.
(38) Miyamoto, S.; Kollman, P. SETTLE: An analytical version of the SHAKE and RATTLE algorithms for rigid water molecules. *J. Comput. Chem.* **1992**, *13*, 952−962.
(39) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparameterization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474.
(40) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(41) Pozun, Z. D.; Hansen, K.; Sheppard, D.; Rupp, M.; Müller, K. R.; et al. Optimizing transition states via kernel-based machine learning. *J. Chem. Phys.* **2012**, *136*, 174101.
(42) Platt, J. C. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*, Report MSR-TR-98-14; Microsoft Research: Redmond, WA, 1998.
(43) Joachims, T. *Making Large-Scale SVM learning Practical*, Report LS-8 Report 24; Computer Science Department, University of Dortmund: Dortmund, Germany, 1998.
(44) Vapnik, V.; Chapelle, O. Bounds on error expectation for support vector machines. *Neural Comput.* **2000**, *12*, 2013−2036.

(45) Aizerman, M.; Braverman, E.; Rozonoer, L. Theoretical foundations of the potential function method in pattern recognition learning. *Autom. Remote Control* **1964**, *25*, 821−837.
(46) Smola, A. J.; Schölkopf, B.; Müller, K. R. The connection between regularization operators and support vector kernels. *Neural Networks* **1998**, *11*, 637−649.
(47) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549−561.
(48) Porotto, M.; Yi, F.; Moscona, A.; LaVan, D. A. Synthetic protocells interact with viral nanomachinery and inactivate pathogenic human virus. *PLoS One [Online]* **2011**, *6*, e16874.
(49) Aguilar, H. C.; Aspericueta, V.; Robinson, L. R.; Aanensen, K. E.; Lee, B. A quantitative and kinetic fusion protein-triggering assay can discern distinct steps in the Nipah virus membrane fusion cascade. *J. Virol.* **2011**, *84*, 8033−8041.