

Characterizing and Predicting Protein Hinges for Mechanistic Insight

Pranav M. Khade, Ambuj Kumar, Robert L. Jernigan*

Bioinformatics and Computational Biology Program, Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA, 50011, USA

Abstract

The functioning of proteins requires highly specific dynamics, which depend critically on the details of how amino acids are packed. Hinge motions are the most common type of large motion, typified by the opening and closing of enzymes around their substrates. The packing and geometries of residues are characterized here by graph theory. This characterization is sufficient to enable reliable hinge predictions from a single static structure, and notably, this can be from either the open or the closed form of a structure. This new method to identify hinges within protein structures is called PACKMAN. The predicted hinges are validated by using permutation tests on B-factors. Hinge prediction results are compared against lists of manually-curated hinge residues, and the results suggest that PACKMAN is robust enough to reproduce the known conformational changes and is able to predict hinge regions equally well from either the open or the closed forms of a protein. A group of 167 protein pairs with open and closed structures has been investigated. Examples are shown for several additional proteins, including Zika virus non-structured (NS) proteins where there are 6 hinge regions in the NS5 protein, 5 hinge regions in the NS2B bound in the NS3 protease complex and 5 hinges in the NS3 helicase protein. Results obtained from this method can be important for generating conformational ensembles of protein targets for drug design. PACKMAN is freely accessible at (<https://PACKMAN.bb.iastate.edu/>).

Keywords

protein hinge prediction; alpha shape; zika virus hinges; flexible peptide linkers

Introduction

Proteins are key players in cellular activities, and their dynamics is usually critical for their function. Structure-based drug design, however, often uses only a single structure, despite the fact that most proteins have important function-related dynamics, with a range of different conformations, sometimes with even very large-scale conformational changes. Understanding the dynamics of proteins is important for the comprehension of function and mutations, as well as signaling and transport pathways. It is essential to know the range of conformational states available to a protein, especially the important larger global movements, in order to define a protein's complex functional dynamics. Globular proteins

*Corresponding author's: jernigan@iastate.edu (Robert L. Jernigan).

are themselves characterized by high packing densities, which are non-uniform in nature - with high packing densities in some parts and sometimes lower densities other places. The moving parts can be rigid densely packed structural communities that move to function [1,2]. The overwhelming conclusion from a broad range of dynamics studies is that the motions are, for any given protein, highly restricted and relatively limited types of motions.

Multiple conformational states and motion trajectories have been used to develop mechanistic insights into protein motions and their associated functions. Protein motions at the atomic level can, in principle, be extracted from the various conformational snapshots of the proteins obtained from separate structure determinations by X-ray crystallography, NMR or Cryo-EM, but even with the more than 100,000 structures in the Protein Data Bank (PDB) [3], there are still not sufficient numbers of closely related structures to provide a full set of such forms, and probably never will be, to specify fully the dynamics of all proteins. The available structures are thus usually supplemented by computed dynamics with approaches such as molecular dynamic (MD) simulations or elastic network models (ENM). Nonetheless, the PDB is a rich source of protein conformational states which was used previously by Mark Gerstein and co-workers to develop the Database of Macromolecular Motions (MolMovDB) [4] and to compare the experimental structures to computed dynamics [5–10]. The MolMovDB provides useful classifications of the different types of motions in proteins. In that database, the most common types of motions are hinge motions and shear motions. (Hinges motions are rotations around some axis.) These hinge motions account for 45% of the motions collected in MolMovDB, which is, in turn, comprised of 31% domain hinge motions and 14% fragment hinge motions [11]. Shear motions involve a sliding movement between two parts of a protein. Hinge motion is simpler and involve rotations of one part of a protein against another (typically two domains) and these motions can be characterized by a rotation around a line between two planes. Such motions usually involve multiple residues that undergo significant conformational changes. The overall protein motions can also be more complex, such as combinations of hinge and shear motions. Protein motions can occur on relatively long timescales, which makes it challenging, or expensive, to obtain them from MD simulations. What is needed are reliable predictions of protein motions, and these should lead to a deeper understanding of how proteins function on a fundamental level. Here we tackle the problem of identifying hinges in single static protein structures, based on their geometries.

Hinge motions are characterized by relatively few large changes in torsion angles of the protein backbone, changes in hydrogen bonding and changes in the packing interactions across the hinge. To qualify as a hinge-motion, the residues involved should have lower overall constraints from their specific packing arrangement. Therefore, when a chain exhibits a hinge motion in the region connecting two structural domains, each domain may behave as a rigid body whereas the hinge residues themselves act as a more flexible region that allows these rigid domains to undergo significant motion, relative to one another, even though the local motion around the hinge will be small. Ligand or substrate binding can drive these motions. Other factors such as pH, temperature and salt concentration can stabilize one or the other conformation, either intrinsically or by creating a higher free energy barrier between the two conformers of the. In addition, there is the possibility that

exothermic reactions can exert ballistic driving forces directly causing conformational hinge motions [12].

Hinges from open to closed forms can often be identified as one of the important normal modes in the elastic network models; however, ENMs are usually unable to predict the hinge motions from closed to open forms [13–17]. A similar problem exists for molecular dynamics simulations because the closed forms are usually more stable than open forms. Hinge prediction methods such as Translation Libration Skew Motion Determination (TLSMD) [18], partition protein chains into parts that are modeled as rigid bodies undergoing TLS (Translation/Libration/Screw) vibrational motions in order to predict the flexible regions. StoneHinge [19] uses network analysis of individual protein structures for hinge prediction; FlexOracle [20] uses an energy-based approach where it computes the energy of the fragments compared with the undivided protein, and predicts hinges by minimizing this quantity; and HingeMaster [21] is a meta-tool (integration of tools) that uses various approaches such as optimization algorithms for residue selection, normal modes, graph theory, free energy and sequence information. Moreover, there are alignment and fragment superposition-based methods such as FlexProt [22], HingeFind [23] and DynDom [24], both provide good hinge prediction accuracy. However, all these tools require multiple conformers (such as open and closed) to predict hinges with high precision, with the exception of FlexOracle, which incidentally recommends caution while using metal-bound structures.

The approach here utilizes protein packing, graph theory and statistical approaches to detect the hinge regions in a protein. Figure 1 shows the overall schema for the approach. Delaunay tessellations and alpha shapes are used to estimate the packing profile of an input protein structure, which is then converted into a graph model to determine the hinge regions. Predicted hinge regions are cross validated by using permutation test statistics. The present method allows a user to predict hinges using a single structure; either open or closed, of any given protein. It has several tuning parameters and also generates p-value statistics for each predicted hinge. The software, PACKMAN, is an open source Python-based package, which can be freely accessed at (<https://PACKMAN.bb.iastate.edu/>). Users can also access the source code and the instructions at (<https://github.com/Pranavkhade/PACKMAN>).

Results and Discussion

Protein Packing and B-Factors

Protein flexibility is an inherent global property of a protein structure and usually is the basis for its functional mechanism that almost universally requires internal motions. B-factors which result from local ‘uncertainty’, i.e., either from insufficient data to define an atom position or from the presence of multiple conformations of the protein. And, they can also result from combinations of multiple factors. However, for sufficiently high-resolution crystal structures, the B factors are more likely originate only from the local flexibility of the protein. Absence of strong electron density, which is reflected in the B-factors, often indicates conformational flexibility and dynamics of an atom. Therefore hinges, being relatively more flexible, might be expected to have distinctively larger B-factors when compared to those from more rigid parts of the structure.

Alpha-shape theory and Delaunay tessellations have been widely implemented to understand protein packing, estimate geometric properties, and model their structures. Moreover, they have been widely applied to study the relationship among protein atoms as well as among amino acids. Ban et al., 2006 implemented Delaunay tessellations to predict protein-protein interaction interfaces [25]. Delaunay tessellations are useful for modelling protein packing because the atomic positions have such a high level of apparent geometric irregularity that the calculations are highly unlikely to yield non-unique solutions. Delaunay tessellations can be non-unique if and only if more than four points in the given set form the same circumsphere with no points inside it. Moreover, even if non-unique solutions occur, PACKMAN results are unaffected because the alpha shape created using Delaunay tessellations are converted into a graph where all points in the tessellations are connected to one another. Our results indicate that the use of protein packing and B-factors provide an efficient way to predict hinges. PACKMAN uses Delaunay tessellations and Alpha shapes to predict the hinge residues in a protein, which is further validated by using permutation test statistics using B-Factors.

Hinge prediction parameters

PACKMAN uses three distinct parameters (alpha values, clustering parameter (k) values, and a minimum hinge length parameter) to predict hinge sites in a protein. Hinge predictions are carried out for 167 protein structure pairs for alpha values ranging from 0 to 5 in 0.1 step increments (a total of 51 different alpha values) [Table S1–S7]. The predicted hinges are compared for overlaps and other statistics. Our result shows that an increase in the value of alpha increases the number of predicted hinges (Fig 2A) while the total number of hinge residues remains approximately the same (Fig 2B), generally leading to the prediction of multiple small hinges. This indicates that using large alpha value tends to predict sequentially smaller and fragmented hinges whereas using small alpha values leads to the prediction of larger hinges. Fig. 2C further supports this since the number of overlaps of predicted hinges between the open and closed conformation states increases as the overlaps become discontinuous and match with multiple hinge fragments and are counted as separate overlaps.

Additionally, PACKMAN uses a clustering parameter to generate k clusters from the eccentricity values obtained from the alpha shape graph. This parameter takes an integer value as input (N) and separates the gradient of eccentricity values into N distinct clusters. Amino acids within the cluster having the lowest eccentricity value are reported as a hinge region. Our results indicate that changing the cluster parameter does not change the position of predicted hinge region however larger values may lead to increased numbers of false negatives (Fig 3). Due to the low packing density (low eccentricity) of protein hinge regions and the relatively high packing density (high eccentricity) of the connected domain regions, use of the cluster parameter value 2 results in the prediction of true hinge regions with the highest accuracy (discussed below). Furthermore, PACKMAN uses a minimum hinge length (number of amino acid residues) parameter to set the minimum length for predicted hinge regions. This parameter is a filter that allows a user to add a hinge length constraint to the prediction whenever the approximate hinge length is known. Predicted hinge regions below

the minimum hinge length parameter value are discarded. In this study, we have used a minimum hinge length of 4 residues.

Hinge predictions for structure set

Here, hinge predictions have been carried out first for 167 pairs of two different conformations of the same protein, one open and one closed, collected in one of our previous studies [6]. The present results achieve consistency between the predicted hinges when either the open and closed conformational states are used, for all 167 structure pairs [Table S1], which demonstrates that PACKMAN is robust to conformational change. This is an important finding since for conventional ENMs and MD hinge motions are easily found to move from open to the closed structure but simulations often are unable to show motions from the closed structure towards the open structure. This result mean that that it is now possible to use these computed hinges to reliably predict the ways in which a closed protein opens. Moreover, it indicates that hinge prediction with PACKMAN can be carried out using any conformational state of a protein to generate a set of conformations along the open-closed axis in either direction. Next, we show some specific examples of the results of applying PACKMAN.

Inorganic Pyrophosphatase (PPases) (Family II)

PPases are a family of essential enzymes engaged in the regulation of inorganic pyrophosphate (PPi) cellular concentration and maintaining various biosynthetic reactions such as nucleic acid and protein synthesis in cells [26]. PPases are mainly found in bacterial and archaeal lineages where many of them are human pathogens. The PPase molecular mechanism is regulated by rotation of the C-terminal domain at its hinge by about 90 degrees, exposing the active site for substrate binding [27]. PACKMAN predicts hinge residues to be 180–210 from the closed form and residues 185–195 from the open form (Fig 4A) with p-values of 0.008 and 0 respectively, and clearly, these two predictions have significant overlap and are in close correspondence. These residues are located between the C-terminal and N-terminal domains at locations appropriate for rotation between the open and closed conformations. X-ray crystallography identified amino acids 190 in the closed conformation and 188 in the open conformation as parts of the hinge region [27], validating the prediction of PACKMAN. These results also show the hinge from the closed conformation to be slightly shorter than those from open conformations, but such differences are not found to be consistent for other proteins.

Rat DNA Polymerase β (Pol β)

Pol β undergoes a large conformational change centered around a hinge to regulate DNA base excision repair and a variety of other cellular processes including meiosis [28]. PACKMAN predicts residues 79–146 as a hinge region in the closed form and residues 86–146 in the open form (Fig 4B). In the closed form, two hinges are predicted with p-values of 0 and 0.0002. However, since there was only one residue the gap between the two predicted hinges, these may be considered to be one continuous hinge. In the open form, the p-value of the predicted hinge was 0. Here p-value of 0 indicates that out of 10,000 permutations, the approach found no cases of the randomized hinge and non-hinge sample B-factor mean differences being greater than the differences in the mean B-factor values of the PACKMAN

predicted hinge and non-hinge regions. X-ray crystallography indicates amino acid residues 88–151 to be part of the “finger domain” that regulates the motion of the 8-kD domain towards the Palm domain to form a channel [29], and this overlaps well with the PACKMAN predicted hinge region at residues 79–146 from the closed and 86–146 from the open forms.

Calmodulin (CaM)

CaM is a small protein of 148 amino acid residues that belongs to a class of ubiquitous proteins having similar structures characterized by their distinctive helix-loop-helix Ca^{2+} -binding motifs, the so-called EF hands [30]. CaM mediates regulation of Ca^{2+} -dependent signaling pathways through distinct protein domain movements centered on a hinge region. PACKMAN predicts a hinge at amino acid residues 68–97 for the closed form and amino acid residues 62–86 for the open form (Fig 4C). X-ray crystallography results have reported amino acid residues 74–82 as a “flexible linker” region [31] which lies within the predicted hinge regions obtained from both the open and the closed forms by PACKMAN.

Human Immunodeficiency Virus (HIV) Protease

HIV-1 protease is an effective therapeutic target and has been widely used for effective antiviral drug design against HIV-1 infection. Models built using substrate and inhibitor complexes of HIV-1 protease mutants indicate the hinge regions in the dimerization region (residue 5–10), the active site (residues 25–27), the flap (residues 45–55), and the substrate cleft (residues 80–90), which display the smallest fluctuations in their mean positions and coordinate the essential motions of the protein [32]. PACKMAN predicts the first hinge region for the closed form to be residues 21–33 and from the open form as residues 21–32, with the second hinge region in the closed form at residues 83–88 and for the open form at residues 82–90 (Fig 5A). PACKMAN was able to detect two hinges correctly for both the closed and the open conformations. The third hinge is not predicted.

Uracil-DNA Glycosylase (UDG)

UDG plays an important role in restoring the chemical integrity of DNA by flipping uracil nucleotides out of the DNA base stack using a “pinch-push-pull” mechanism [33]. UDG undergo a small conformational change in the open and the closed forms to regulate its functional dynamics. PACKMAN predicts three different hinges in the closed form (137–144, 151–162 and 190–205) and three hinges in the open form (136–143, 152–162 and 189–205) (Fig 5B). Hinge residue region 136–143 are adjacent to the hinge residues 129–132 which has been identified as a remote “hinge” region implicated in the dynamics of clamping of UDG [34]. UDG amino acid residues 145–148 were named as the recognition and catalytic sites [35] and PACKMAN predicts appropriately adjacent hinges on both sides in open as well as closed forms. Moreover, Phe158 that was identified as an important site for nucleotide flipping [36] is part of the second hinge region predicted by PACKMAN.

Ribose Binding Protein (RBP)

RBP is a bifunctional soluble receptor found in the periplasm of *Escherichia coli*. RBP assists in the transfer of ribose across the cytoplasmic membrane by the interaction of various ligand binding proteins with the ribose high affinity transport complex. RBP

contains four amino acid segments (38–41, 64–73, 90–100, 129–138), which are loops that form flaps over the substrate binding cavity [37]. Segment 109–118 forms a helix on the opposite side of the molecule and is located immediately adjacent to one of the three hinges of RBP [37]. Furthermore, amino acids 223–231 form a helix and a loop segment near the C-terminus of the protein [37]. PACKMAN predicts three hinge regions for both the open and the closed conformations (Fig 5C). The hinge predicted for the closed form is (102–110) and for the open form (100–112) that lies adjacent to the helical segment 109–118, which may help to regulate its flexibility and resultant substrate binding. Furthermore, the hinge predicted in the closed form (129–145) and the open form (135–148) partly overlap the loop at 129–138. A third hinge region is predicted in the closed form (233–240) and the open form (232–244) that lies adjacent to the C-terminus loop segment (223–231). Our result indicates that PACKMAN results are consistent for both the open and the closed form and it is useful for determining the hinge regions that may play a vital role in regulating the loop dynamics associated with substrate binding activity.

Zika virus hinges

Zika virus (ZIKV), a mosquito-borne flavivirus, has emerged since 2013 as a significant public health concern. This was earlier reported to cause mild disease; whereas recent evidence suggests its more serious association with neuropathy, neonatal microcephaly, Guillain-Barré syndrome [38] and cases of eye dysfunction, hearing deficits, and impaired growth. The Zika genome encodes a polyprotein, which undergoes co- and post-translational processing by the viral and host proteases to produce structural (capsid, pre-membrane and envelope) and NS proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5) [39]. ZIKV NS2B-bound NS3 protease and NS5 polymerase have been important targets for drug discovery.

Discovery of hinges can facilitate the discovery of new drugs that can bind at sites to block the important motions of the viral proteins critical for its function such as replication and immune evasion. Nonstructural protein 5 (NS5) is essential for viral replication and participates in IFN antagonism, allowing the virus to evade the host immune system [40]. It contains a methyl transferase (MT) domain that connects to the fingers subdomain of the RNA-dependent RNA-polymerase (RdRp), assisting the NTP channel of the RdRp to extend outwards to help in RNA synthesis [41]. PACKMAN predicts 5 hinges (307–323, 337–367, 450–483, 568–599 and 601–606) [Table 1, Fig 6A] with p value = 0 and a sixth hinge (741–748) with p value 0.09. NS5 MT interaction with the finger's subdomain of the RdRp is regulated by hydrophobic contacts among MT residues Pro113, Leu115 and Trp121 and Tyr350, Phe466 and RdRp residues Pro584 [41]. These residues are located in close proximity to the 307–323 and 337–367 hinges predicted by PACKMAN (Fig 6A). Zhao et al., 2017 reported that residue segments 312–323 and 742–750, located in the back of the RdRp Motif G, take on two distinct conformations, indicating their role in regulating interactions between MT and the fingers subdomain of RdRp [41]. PACKMAN predicts residue position 307–323 and 741–748 as hinges with p-values of 0 and 0.09, respectively, indicating a strong consistency with the reported results.

NS2B-NS3 protease plays an essential role in viral replication by cleaving the complete viral poly-protein into separate proteins [42]. This behavior of NS2B-NS3 protease makes it an ideal antiviral drug target to inhibit viral replication and growth. The mechanism associated with ZIKV NS2B mediated NS3 protease activation is poorly understood and requires further study. ZIKV NS2B bound NS3 protease adopts a relaxed conformation in the absence of the poly-protein substrate or an inhibitor molecule [43]. The C-terminal loop, comprising residues 69–87, of NS2B adopts a unique conformational state in between the relaxed apo form and the inhibitor-bound state [43], indicating the presence of an adjacent flexible region. In a closely related species Dengue virus (DENV), upon substrate binding, the NS2B residues at positions 63 – 88 move around a hinge at residue 63 [44]. PACKMAN predicts a hinge region (60–69) in NS2B protein (Fig 6B), indicating that the hinge residues 60–69 motion may regulate the movement of NS2B residue segment 63–88. PACKMAN also predicts 4 other hinges (22–25, 34–40, 131–140, 145–152) (Fig 6B) in the NS3 protease. Moreover, additional conformational differences in the ligand-bound form are observed in NS3 protease where, in the absence of ligand, a sharp kink at residue 151 leads residues 152–158 to pack between β -strand 123–126 and β -strand 147–150 [43]. A fourth hinge region (145–152) in NS3 protease is predicted by PACKMAN that includes residue 151 that is adjacent to the 152–167 loop. Since the conformational state of residues 152–167 is key for maintaining enzymatic activity, hinge four might be an important drug target for inhibiting ZIKV replication.

NS3-helicase (NS3-Hel) is present in the C-Terminal region of the NS3 protein and is associated with viral genome replication and RNA synthesis [45]. Inactivation of NS3-Hel in Dengue virus type 2 (DENV 2) virus has been shown to inactivate the replication of virus [46], making it element important for drug design. A recent study reported a β -hairpin (Res 431–444) in Zika virus NS3-Helicase that extends from domain 2 to connect with domain 3 [45], which has been proposed to facilitate the separation of the RNA strands of double-stranded RNA in unwinding [47]. PACKMAN predicts 5 hinge regions (289–297, 425–434, 440–456, 473–494 and 508–517) (Fig 6C) in NS3-Hel. Our results suggest that the β -hairpin (431–444) is adjacent to the second and third hinge regions, suggesting it could also be a drug targetable region. Moreover, these two hinges can explain how the proposed mechanism uses the β -hairpin as a ‘wedge’ to separate the strands of RNA. Moreover, the remaining three hinges lie in close proximity to the β -hairpin and these may assist in regulating details of the motion for this β -hairpin.

The utility of PACKMAN hinge prediction

Hinge regions are central for the regulation of various large-scale and functionally important cooperative motions of a protein. Hinges predicted using PACKMAN can help in speculating possible global motion of a protein. Hinge regions predicted in annexin (Fig 7A) indicate that its global motion can be mediated by the movement of blue and green domains independent of the red domain motion across the hinge axis, which is consistent with the protein bending motion shown by Cregut *et al.* [48]. Moreover, the hinge region predicted in calmodulin (Fig 7B) can help in understanding its global motion. The transition of calmodulin from closed to open form can be regulated by the movement of the blue and red domains across this predicted hinge region. Such motions can depend on the arrangement of

multiple hinges in a structure. In theory, the more hinges the more collectively restricted will be the degrees of freedom for individual hinges. Calmodulin contains only one hinge region between the blue and red domain, allowing several degrees of freedom for the global motion of this particular hinge; whereas in annexin, there are two hinge regions between two domains, leading to fewer degrees of freedom and relatively more restrictions on its motion. Identifying multiple hinge regions can help to understand the degrees of freedom associated with any given protein.

Although there have been various approaches implemented to predict hinge residues, the results were shown to vary substantially when different structures in different conformational states were used. The present approach shows remarkable success in yielding prediction consistency when using different conformational states. PACKMAN can be used for predicting single as well as multiple hinge regions within a protein using any conformational state, yielding only minimal differences in the results. The results are not overly sensitive to the few parameter values. However, from the Fig 2 and 3, we can infer that the best approach to use this tool is to start with low values of both parameters (α and k) and explore if the results are changing as the values are increased. PACKMAN implements a simple and purely geometric method to identify hinge residues, so it can potentially be applied across a wide variety of different molecule types such as RNA or saccharides and is not restricted only to proteins.

Results obtained for three proteins having a single hinge (Inorganic Pyrophosphatase, Rat DNA polymerase β and Calcium-sensing protein calmodulin) and three proteins having multiple hinges (HIV Protease, Uracil-DNA Glycosylase, and Ribose Binding Protein) that were shown above indicate that PACKMAN can be useful for estimating single as well as multiple hinges in a protein. PACKMAN predicts the hinge residues of Zika virus NS5 protein that is likely to regulate its MT and finger subdomain interactions. Furthermore, PACKMAN is able to predict the hinge region associated with the conformational transition of NS2B-NS3 protease from its relaxed apo form to the inhibitor bound form. Moreover, PACKMAN is able to detect the NS3-helicase hinge residues involved in the separation of the RNA strands that mediate viral RNA synthesis. Our results suggest that PACKMAN can be a generally useful tool for understanding molecular mechanisms as well as for providing variant conformations to produce ensembles of drug target conformations. Identifying the mechanism associated with the protein conformational transitions from closed to open has been a challenging problem, which is overcome with PACKMAN. However, with the flexible hotspots (hinges) identified by using PACKMAN, it should be possible to improve simulation methods such as MD and elastic network model to provide the mechanistic insights needed to understand many more protein functions.

Methods and Materials

PACKMAN pipeline

The PACKMAN approach is described in Fig 1. PACKMAN first generates a Delaunay tessellation [49] using backbone atoms of the input protein structure. Alpha shapes [50] are then computed as a subset of Delaunay tessellations in order to obtain a concrete defined shape of the protein. The resultant alpha shape is then converted into a graph model to

determine the network eccentricity of each backbone atom, i.e., its importance. Amino acids having the lowest eccentricity scores are clustered as hinge region by PACKMAN. A permutation test is performed to estimate the significance of a hinge prediction by using B-factor of the corresponding backbone atoms. Details of each step are provided below.

Data collection

167 pairs of matched structures taken from the PDB are considered, each pair having two distinct conformational states [Table S1] [6]. In addition, proteins from the ZIKV NS5 (PDB ID: **5M2Z**), NS2B bounded NS3 protease (PDB ID: **5H4I**) and NS3-Helicase (PDB ID: **6MH3**) and *Homo sapiens* annexin protein (PDB ID: **1AXN**) structures were downloaded from the PDB.

Delaunay Tessellations

Given any set of points, a Delaunay tessellation can be built for any four points given that there is no single point lying inside the circumsphere formed by the four points under consideration. The circumdiameter of the circumspheres formed from the Delaunay tessellations can range from 1 to the maximum length between any two points under consideration. Delaunay tessellations of protein backbone atoms are constructed by computing their convex hull. The input points are lifted to a paraboloid by adding the sum of squares of the atom coordinates to the $n+1$ convex hull.

Alpha Shapes

An alpha shape is a subset of the Delaunay tessellations such that each unique tessellation forms a circumsphere with a radius less than the alpha parameter. Each tessellation in the alpha shape is a polyhedron with triangular faces that represents the packing of the associated atoms. The vertices of the alpha shape are the atoms having a strictly positive accessible surface area ($ASA > 0 \text{ \AA}$). For a deeper insight into the alpha shape theory, refer to Alpha shape method articles such as [50]. An introduction to these models is also available from Poupon [51].

Network Eccentricity

Eccentricity is defined here as the largest distance of a given node from all other nodes in the graph. The importance of a node in a graph is estimated by the reciprocal of its eccentricity value. The eccentricity provides a measure of the dominance of a node within a network [52]. Here, the tessellations within the alpha shape are converted into a graph model using the networkx Python module, so that if a tessellation exists, the corresponding protein 3D coordinates of protein backbone atoms are treated as vertices with all to all connections (edges). Vertices of the graph model are clustered on the basis of their eccentricities using k-means clustering. The cluster having the lowest eccentricity value is used to assign the points of a hinge region. If a backbone atom of an amino acid exists in the cluster with lowest eccentricity value, PACKMAN reports it as a hinge region. From these predicted hinge regions, the sequence identities are considered to identify individual hinges – to determine whether clusters include single or multiple hinges.

Permutation Test

Eccentricity based clustering approach allows us to use the local packing densities within a protein to predict hinge residues. Using a B-factor dependent validation of predicted hinge residues adds a dynamic component of a protein molecule to the pipeline. Therefore, a permutation test is conducted to estimate whether the B-Factors of the predicted hinge and non-hinge regions are significantly different from one another. This step analyzes whether the dynamic behaviors of predicted hinge regions are significantly different from the predicted non-hinge regions. The eccentricity derived from the graph and B-factors have no relation or confounding factors in the analysis, therefore, having a significant p-value from the permutation test is a meaningful indicator of a true hinge.

A permutation test is applied for the strong null hypothesis testing that there is no difference between the mean B-factor of hinge backbone atoms and mean B-factor of the non-hinge backbone atoms. Here let $X \sim (x_1, x_2, x_3, \dots, x_n)$ be the B-factors of hinge backbone atoms in a segment of length n and $Y \sim (y_1, y_2, y_3, \dots, y_m)$ be the B-factors of the non-hinge backbone atoms of length m for a protein predicted by PACKMAN. Here $N = n+m$ is the total number of backbone atoms or protein with theoretical t statistics

$$t = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{df}.$$

where, \bar{X} is the mean B-factor value of hinge backbone residues, \bar{Y} is the mean B-factor of non-Hinge backbone residues, S_p is the pooled variance and df is the degrees of freedom. Here pooled variance represents combined variances of both hinge and non-hinge backbone atom B-factor values. The resulting permutation t-statistic is

$$t^{(1)} = \frac{\bar{X}^{(1)} - \bar{Y}^{(1)}}{S_p^{(1)}},$$

where, $\bar{X}^{(1)}$ is the mean of hinge region random permuted B-factors, $\bar{Y}^{(1)}$ is the mean of non-hinge region random permuted B-factors and $S_p^{(1)}$ is the pooled variance of the permuted samples. In this work, we use a critical p-value cutoff $P(t^{(1)} \geq t) = 0.05$ for validating true hinges.

Data Visualization.

Data visualization is carried out with PyMol [53].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

This research has been supported by NSF grant DBI-1661391 and by NIH grants R01-GM127701 and R01-GM127701-01S1. We also thank Research IT@Iowa State University for helping with many aspects of the computing.

Abbreviations:

PACKMAN	PACKing and Motion Analysis
MD	Molecular dynamics
ZIKV	Zika Virus
HIV	Human Immunodeficiency Virus
TLSMD	Translation Libration Skew Motion Determination
MolMovDB	Database of Macromolecular Motions
TLS	Translation/Libration/Screw
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
NS	non-structured
PPases	Inorganic Pyrophosphatase
Polβ	Rat DNA Polymerase β
CaM	Calmodulin
UDG	Uracil-DNA Glycosylase
RBP	Ribose Binding Protein
RdRp	RNA-dependent RNA-polymerase
DENV	Dengue virus
NS3-Hel	NS3-helicase
ASA	Accessible Surface Area

References

- [1]. Chopra N, Wales TE, Joseph RE, Boyken SE, Engen JR, Jernigan RL, Andreotti AH, Dynamic Allostery Mediated by a Conserved Tryptophan in the Tec Family Kinases, *PLOS Comput. Biol* 12 (2016) 1–19. doi:10.1371/journal.pcbi.1004826.
- [2]. Kornev AP, Taylor SS, Dynamics-Driven Allostery in Protein Kinases, *Trends Biochem. Sci* 40 (2015) 628–647. doi:10.1016/j.tibs.2015.09.002. [PubMed: 26481499]
- [3]. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The Protein Data Bank, *Nucleic Acids Res.* 28 (2000) 235–242. doi:10.1093/nar/28.1.235. [PubMed: 10592235]

- [4]. Gerstein M, Krebs W, A database of macromolecular motions, *Nucleic Acids Res.* 26 (1998) 4280–4290. doi:10.1093/nar/26.18.4280. [PubMed: 9722650]
- [5]. Sankar K, Liu J, Wang Y, Jernigan RL, Distributions of experimental protein structures on coarse-grained free energy landscapes, *J. Chem. Phys.* 143 (2015) 243153. doi:10.1063/1.4937940. [PubMed: 26723638]
- [6]. Sankar K, Jia K, Jernigan RL, Knowledge-based entropies improve the identification of native protein structures, *Proc. Natl. Acad. Sci.* 114 (2017) 2928–2933. doi:10.1073/pnas.1613331114. [PubMed: 28265078]
- [7]. Sankar K, Mishra SK, Jernigan RL, Comparisons of Protein Dynamics from Experimental Structure Ensembles, Molecular Dynamics Ensembles, and Coarse-Grained Elastic Network Models, *J. Phys. Chem. B* 122 (2018) 5409–5417. doi:10.1021/acs.jpcc.7b11668. [PubMed: 29376347]
- [8]. Katebi AR, Sankar K, Jia K, Jernigan RL, The Use of Experimental Structures to Model Protein Dynamics, in: Kukol A (Ed.), *Mol. Model. Proteins*, Springer New York, New York, NY, 2015: pp. 213–236. doi:10.1007/978-1-4939-1465-4_10.
- [9]. Isin B, Tirupula KC, Oltvai ZN, Klein-Seetharaman J, Bahar I, Identification of Motions in Membrane Proteins by Elastic Network Models and Their Experimental Validation, in: Vaidehi N, Klein-Seetharaman J (Eds.), *Membr. Protein Struct. Dyn. Methods Protoc*, Humana Press, Totowa, NJ, 2012: pp. 285–317. doi:10.1007/978-1-62703-023-6_17.
- [10]. Eyal E, Chennubhotla C, Yang L-W, Bahar I, Anisotropic fluctuations of amino acids in protein structures: insights from X-ray crystallography and elastic network models, *Bioinformatics.* 23 (2007) i175–i184. doi:10.1093/bioinformatics/btm186. [PubMed: 17646294]
- [11]. Flores SC, Lu LJ, Yang J, Carriero N, Gerstein MB, Hinge Atlas: Relating protein sequence to sites of structural flexibility, *BMC Bioinformatics.* 8 (2007) 1–20. doi:10.1186/1471-2105-8-167. [PubMed: 17199892]
- [12]. Liu J, Sankar K, Wang Y, Jia K, Jernigan RL, Directional Force Originating from ATP Hydrolysis Drives the GroEL Conformational Change, *Biophys. J* 112 (2017) 1561–1570. doi:10.1016/j.bpj.2017.03.004. [PubMed: 28445748]
- [13]. Yang L, Song G, Jernigan RL, How Well Can We Understand Large-Scale Protein Motions Using Normal Modes of Elastic Network Models?, *Biophys. J* 93 (2007) 920–929. doi:10.1529/biophysj.106.095927. [PubMed: 17483178]
- [14]. Yang L, Song G, Jernigan RL, Protein elastic network models and the ranges of cooperativity, *Proc. Natl. Acad. Sci.* 106 (2009) 12347–12352. doi:10.1073/pnas.0902159106. [PubMed: 19617554]
- [15]. Bahar I, On the functional significance of soft modes predicted by coarse-grained models for membrane proteins, *J. Gen. Physiol.* 135 (2010) 563–573. doi:10.1085/jgp.200910368. [PubMed: 20513758]
- [16]. Song G, Jernigan RL, An enhanced elastic network model to represent the motions of domain-swapped proteins, *Proteins Struct. Funct. Bioinforma* 63 (2006) 197–209. doi:10.1002/prot.20836.
- [17]. Bahar I, Cheng MH, Lee JY, Kaya C, Zhang S, Structure-Encoded Global Motions and Their Role in Mediating Protein-Substrate Interactions, *Biophys. J* 109 (2015) 1101–1109. doi:10.1016/j.bpj.2015.06.004. [PubMed: 26143655]
- [18]. Painter J, Merritt EA, {it TLSMD} web server for the generation of multi-group TLS models, *J. Appl. Crystallogr* 39 (2006) 109–111. doi:10.1107/S0021889805038987.
- [19]. Keating KS, Flores SC, Gerstein MB, Kuhn LA, StoneHinge: Hinge prediction by network analysis of individual protein structures, *Protein Sci.* (2009). doi:10.1002/pro.38.
- [20]. Flores SC, Gerstein MB, FlexOracle: predicting flexible hinges by identification of stable domains, *BMC Bioinformatics.* 8 (2007) 215. doi:10.1186/1471-2105-8-215. [PubMed: 17587456]
- [21]. Flores SC, Keating KS, Painter J, Morcos F, Nguyen K, Merritt EA, Kuhn LA, Gerstein MB, HingeMaster: Normal mode hinge prediction approach and integration of complementary predictors, *Proteins Struct. Funct. Genet* (2008). doi:10.1002/prot.22060.

- [22]. Shatsky M, Nussinov R, Wolfson HJ, FlexProt: Alignment of Flexible Protein Structures Without a Predefinition of Hinge Regions, *J. Comput. Biol* 11 (2004) 83–106. doi:10.1089/106652704773416902. [PubMed: 15072690]
- [23]. Wriggers W, Schulten K, Protein domain movements: Detection of rigid domains and visualization of hinges in comparisons of atomic coordinates, *Proteins Struct. Funct. Genet* 29 (1997) 1–14. doi:10.1002/(SICI)1097-0134(199709)29:1<1::AID-PROT1>3.0.CO;2-J. [PubMed: 9294863]
- [24]. Girdlestone C, Hayward S, The DynDom3D Webserver for the Analysis of Domain Movements in Multimeric Proteins, *J. Comput. Biol* 23 (2016) 21–26. doi:10.1089/cmb.2015.0143. [PubMed: 26540459]
- [25]. Ban Y-EA, Edelsbrunner H, Rudolph J, Interface surfaces for protein-protein complexes, *J. ACM* 53 (2006) 361–378. doi:10.1145/1147954.1147957.
- [26]. Kajander T, Kellosalo J, Goldman A, Inorganic pyrophosphatases: One substrate, three mechanisms, *FEBS Lett.* 587 (2013) 1863–1869. doi:10.1016/j.febslet.2013.05.003. [PubMed: 23684653]
- [27]. Ahn S, Milner AJ, Fütterer K, Konopka M, Ilias M, Young TW, White SA, The “open” and “closed” structures of the type-C inorganic pyrophosphatases from *Bacillus subtilis* and *Streptococcus gordonii* Edited by D. Rees, *J. Mol. Biol* 313 (2001) 797–811. doi:10.1006/jmbi.2001.5070. [PubMed: 11697905]
- [28]. Starcevic D, Dalal S, Jaeger J, Sweasy JB, The hydrophobic hinge region of rat DNA polymerase β is critical for substrate binding pocket geometry, *J. Biol. Chem* 280 (2005) 28388–28393. doi:10.1074/jbc.M502178200. [PubMed: 15901725]
- [29]. Sawaya MR, Pelletier H, Kumar A, Wilson SH, Kraut J, Crystal structure of rat DNA polymerase beta: evidence for a common polymerase mechanism, *Science* (80-.). 264 (1994) 1930–1935. doi:10.1126/science.7516581.
- [30]. Lakowski TM, Lee GM, Okon M, Reid RE, McIntosh LP, Calcium-induced folding of a fragment of calmodulin composed of EF-hands 2 and 3, *Protein Sci.* 16 (2007) 1119–1132. doi:10.1110/ps.072777107. [PubMed: 17473011]
- [31]. Wriggers W, Mehler E, Pitici F, Weinstein H, Schulten K, Structure and dynamics of calmodulin in solution, *Biophys. J* 74 (1998) 1622–1639. doi:10.1016/S0006-3495(98)77876-2. [PubMed: 9545028]
- [32]. Özer N, Özen A, Schiffer CA, Haliloğlu T, Drug-resistant HIV-1 protease regains functional dynamics through cleavage site coevolution, *Evol. Appl* 8 (2015) 185–198. doi:10.1111/eva.12241. [PubMed: 25685193]
- [33]. Wong I, Lundquist AJ, Bernards AS, Mosbaugh DW, Presteady-state analysis of a single catalytic turnover by *Escherichia coli* uracil-DNA glycosylase reveals a “pinch-pull-push” mechanism, *J. Biol. Chem* 277 (2002) 19424–19432. doi:10.1074/jbc.M201198200. [PubMed: 11907039]
- [34]. Sun Y, Friedman JI, Stivers JT, Cosolute paramagnetic relaxation enhancements detect transient conformations of human uracil DNA glycosylase (hUNG), *Biochemistry.* 50 (2011) 10724–10731. doi:10.1021/bi201572g. [PubMed: 22077282]
- [35]. Zharkov DO, V Mechetin G, Nevinsky GA, Uracil-DNA glycosylase: Structural, thermodynamic and kinetic aspects of lesion search and recognition, *Mutat. Res. Mol. Mech. Mutagen* 685 (2010) 11–20. doi:10.1016/j.mrfmmm.2009.10.017.
- [36]. Parikh SS, Putnam CD, Tainer JA, Lessons learned from structural results on uracil-DNA glycosylase, *Mutat. Res. Repair* 460 (2000) 183–199. doi:10.1016/S0921-8777(00)00026-4.
- [37]. Binnie RA, Zhang H, Mowbray S, Hermodson MA, Functional mapping of the surface of *Escherichia coli* ribose-binding protein: mutations that affect chemotaxis and transport, *Protein Sci.* 1 (1992) 1642–1651. doi:10.1002/pro.5560011212. [PubMed: 1304894]
- [38]. Broutet N, Krauer F, Riesen M, Khalakdina A, Almiron M, Aldighieri S, Espinal M, Low N, Dye C, Zika Virus as a Cause of Neurologic Disorders, *N. Engl. J. Med* 374 (2016) 1506–1509. doi:10.1056/NEJMp1602708. [PubMed: 26959308]
- [39]. Bollati M, Alvarez K, Assenberg R, Baronti C, Canard B, Cook S, Coutard B, Decroly E, de Lamballerie X, Gould EA, Grard G, Grimes JM, Hilgenfeld R, Jansson AM, Malet H, Mancini EJ, Mastrangelo E, Mattevi A, Milani M, Moureau G, Neyts J, Owens RJ, Ren J, Selisko B,

- Speroni S, Steuber H, Stuart DI, Unge T, Bolognesi M, Structure and functionality in flavivirus NS-proteins: Perspectives for drug design, *Antiviral Res.* 87 (2010) 125–148. doi:10.1016/j.antiviral.2009.11.009. [PubMed: 19945487]
- [40]. Grant A, Ponia SS, Tripathi S, Balasubramaniam V, Miorin L, Sourisseau M, Schwarz MC, Sánchez-Seco MP, Evans MJ, Best SM, García-Sastre A, Zika Virus Targets Human STAT2 to Inhibit Type I Interferon Signaling, *Cell Host Microbe.* 19 (2016) 882–890. doi:10.1016/j.chom.2016.05.009. [PubMed: 27212660]
- [41]. Zhao B, Yi G, Du F, Chuang YC, Vaughan RC, Sankaran B, Kao CC, Li P, Structure and function of the Zika virus full-length NS5 protein, *Nat. Commun* 8 (2017). doi:10.1038/ncomms14762.
- [42]. Phoo WW, Zhang Z, Wirawan M, Chew EJC, Chew ABL, Kouretova J, Steinmetzer T, Luo D, Structures of Zika virus NS2B-NS3 protease in complex with peptidomimetic inhibitors, *Antiviral Res.* 160 (2018) 17–24. doi:10.1016/j.antiviral.2018.10.006. [PubMed: 30315877]
- [43]. Chen X, Yang K, Wu C, Chen C, Hu C, Buzovetsky O, Wang Z, Ji X, Xiong Y, Yang H, Mechanisms of activation and inhibition of Zika virus NS2B-NS3 protease, *Cell Res.* 26 (2016) 1260 EP-. 10.1038/cr.2016.116. [PubMed: 27752039]
- [44]. Yildiz M, Ghosh S, Bell JA, Sherman W, Hardy JA, Allosteric Inhibition of the NS2B-NS3 Protease from Dengue Virus, *ACS Chem. Biol* 8 (2013) 2744–2752. doi:10.1021/cb400612h. [PubMed: 24164286]
- [45]. Jain R, Coloma J, García-Sastre A, Aggarwal AK, Structure of the NS3 helicase from Zika virus, *Nat. Struct. & Mol. Biol* 23 (2016) 752 EP-. 10.1038/nsmb.3258.
- [46]. Matusan AE, Pryor MJ, Davidson AD, Wright PJ, Mutagenesis of the Dengue Virus Type 2 NS3 Protein within and outside Helicase Motifs: Effects on Enzyme Activity and Virus Replication, *J. Virol* 75 (2001) 9633–9643. doi:10.1128/JVI.75.20.9633-9643.2001. [PubMed: 11559795]
- [47]. Luo D, Xu T, Watson RP, Scherer-Becker D, Sampath A, Jahnke W, Yeong SS, Wang CH, Lim SP, Strongin A, Vasudevan SG, Lescar J, Insights into RNA unwinding and ATP hydrolysis by the flavivirus NS3 protein, *EMBO J.* 27 (2008) 3209–3219. doi:10.1038/emboj.2008.232. [PubMed: 19008861]
- [48]. Cregut D, Drin G, Liautard JP, Chiche L, Hinge-bending motions in annexins: molecular dynamics and essential dynamics of apo-annexin V and of calcium bound annexin V and I, *Protein Eng. Des. Sel* 11 (2002) 891–900. doi:10.1093/protein/11.10.891.
- [49]. Delaunay B, Sur la sphere vide, *Izv. Akad. Nauk SSSR, Otd. Mat. I Estestv. Nauk* 7 (1934) 793–800.
- [50]. Edelsbrunner H, Mücke EP, Three-dimensional Alpha Shapes, *ACM Trans. Graph* 13 (1994) 43–72. doi:10.1145/174462.156635.
- [51]. Poupon A, Voronoi and Voronoi-related tessellations in studies of protein structure and interaction, *Curr. Opin. Struct. Biol* 14 (2004) 233–241. doi:10.1016/j.sbi.2004.03.010. [PubMed: 15093839]
- [52]. Zhou W, Yan H, Alpha shape and Delaunay triangulation in studies of protein-related interactions, *Brief. Bioinform* 15 (2012) 54–64. doi:10.1093/bib/bbs077. [PubMed: 23193202]
- [53]. DeLano WL, PyMOL: An Open-Source Molecular Graphics Tool, *CCP4 Newsl. Protein Crystallogr* 40 (2002) 82–92.

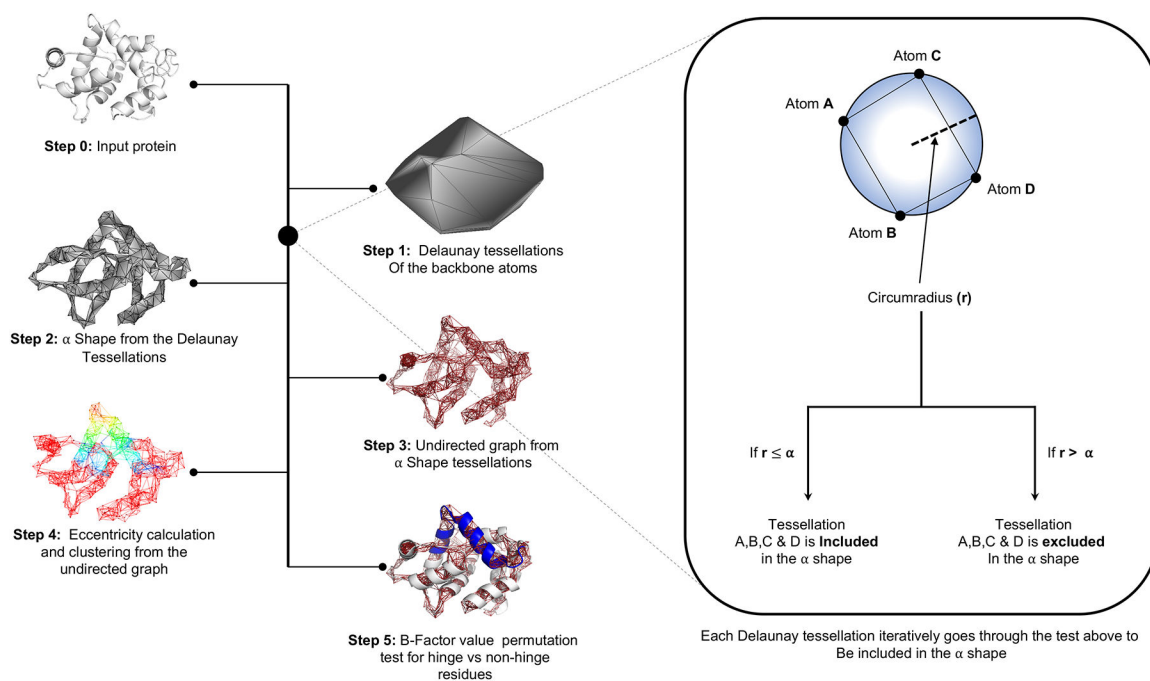


Figure 1.
Method of hinge identification in molecular structures– the PACKMAN pipeline flowchart.

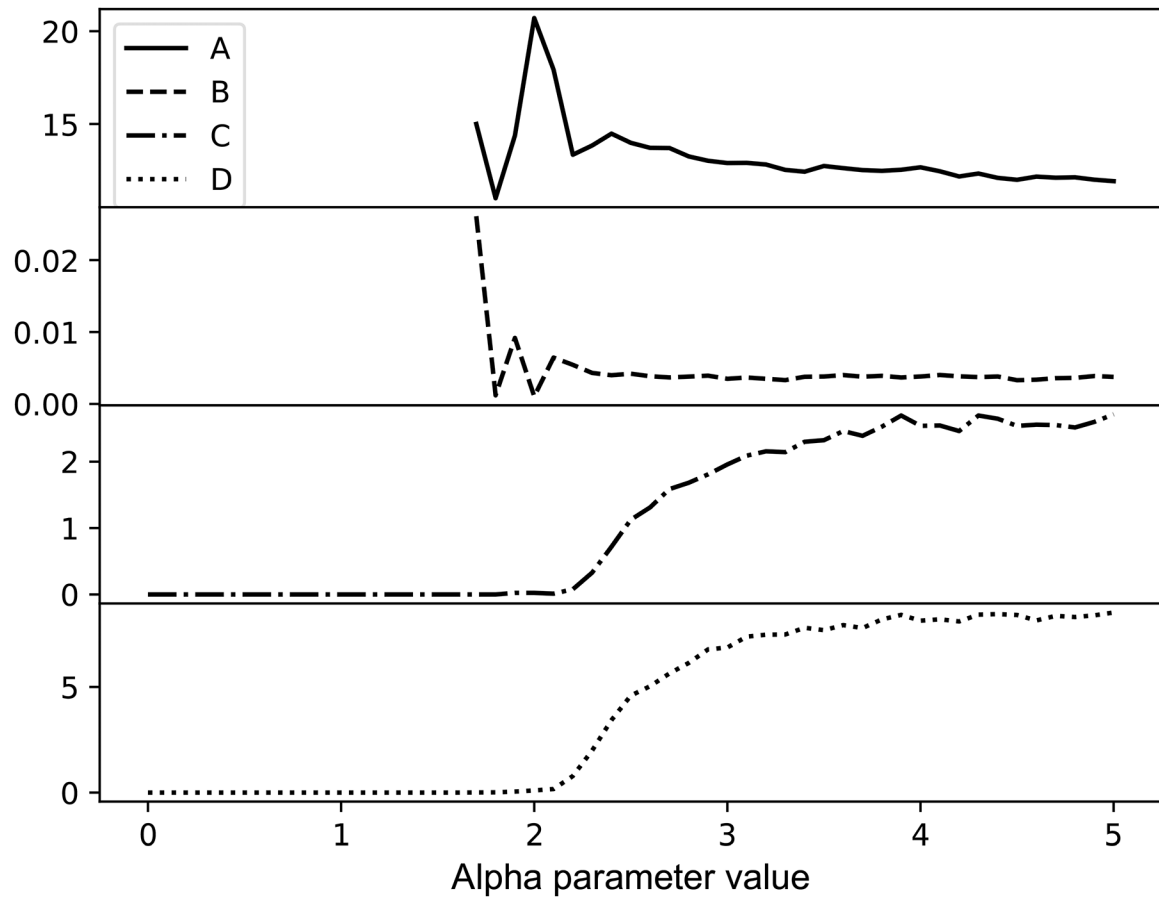


Figure 2. Effect of alpha value on hinge predictions for 167 proteins, each in two conformations.

A) Here the Y-axis shows the number of hinges per pair of protein conformational states for a given values of alpha, B) Here the Y-axis shows the average number of amino acid residues within the predicted hinge regions for a given value of alpha, C) Here the Y-axis shows the average number of overlapping hinge regions predicted using closed and the open conformational state of 167 proteins, and D) Here the Y-axis shows permutation p-values of the predicted hinges for the given values of alpha. Figures A) and C) indicate that as the value of the alpha increases, the number of predicted hinges and the overlaps increase. This also points towards the possibility that increases in the value of alpha may lead to the prediction of fragmented hinge regions. This is further supported by B) which indicates that the average length of the hinges decreases as alpha increases. Figure D) shows that the p-values remain similar for most values of alpha. The spike in p-value on the left occurs because of a small number of cases.

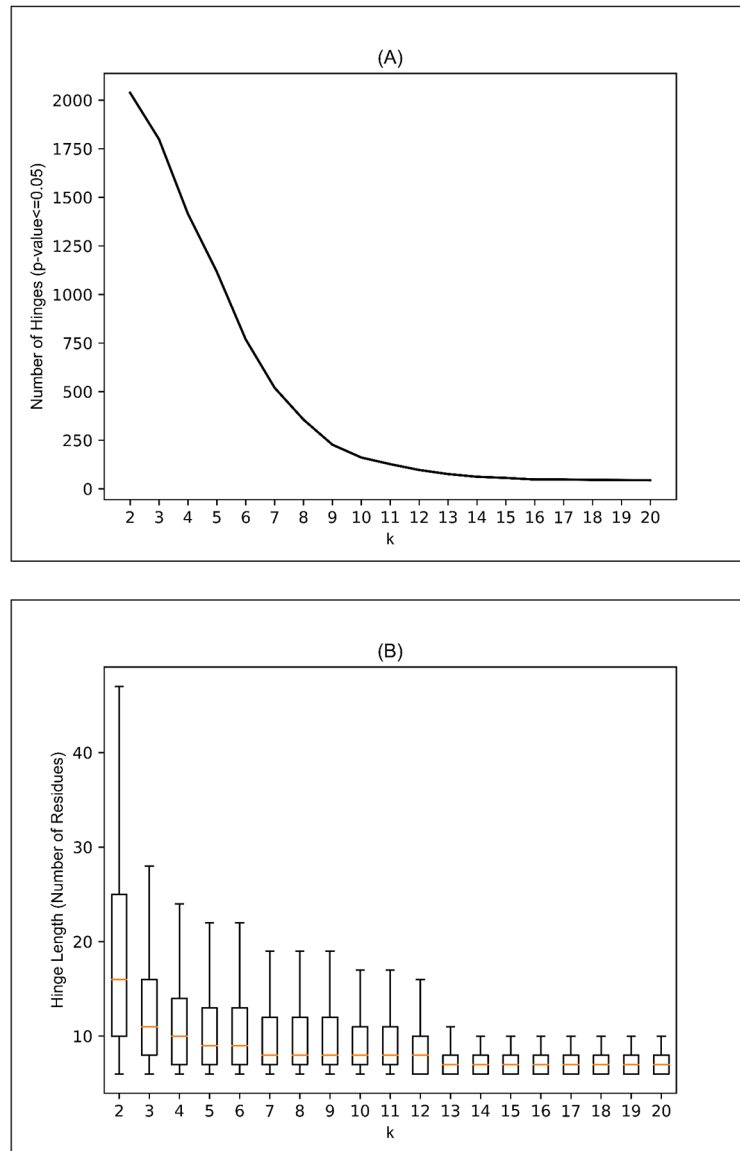


Figure 3. Impact of changing the clustering parameter (k).

(A) Effect of k on the number of hinges (with $p < 0.05$). This indicates that for larger values, hinges may be overlooked, and also that higher values lead to smaller hinges. (B) The effect of k on the hinge length. We observe that the length of the hinge decreases as the k is increased which further indicates that smaller hinges will be observed.

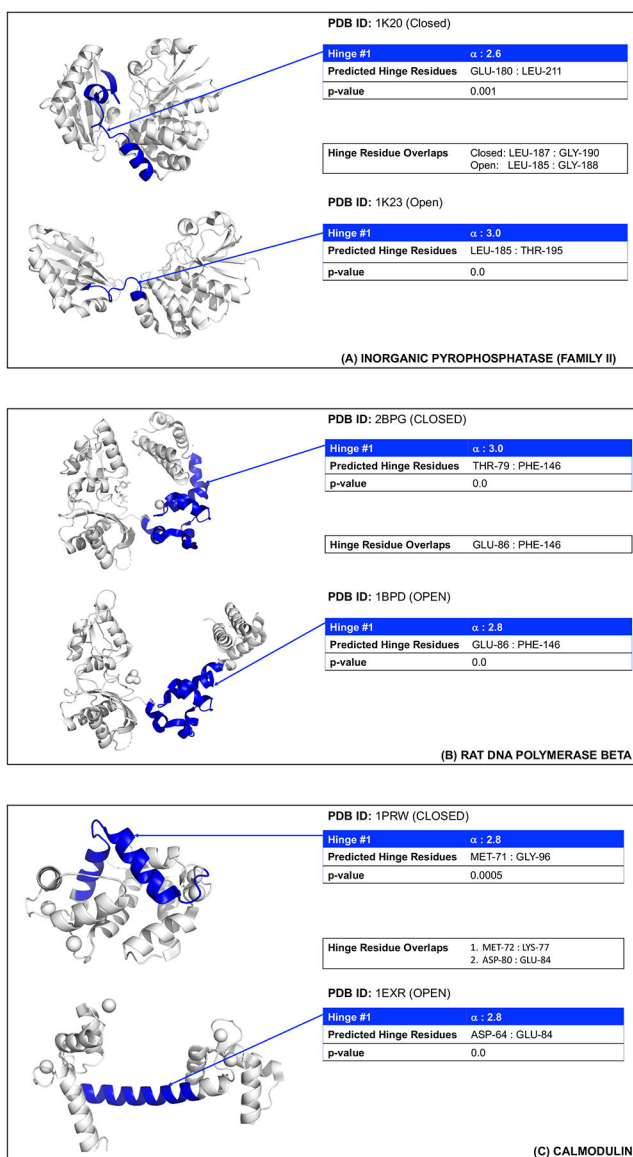


Figure 4.
The single hinges (blue) identified in: a) Inorganic pyrophosphatase (Family II), b) RAT DNA Polymerase Beta and c) Calmodulin.

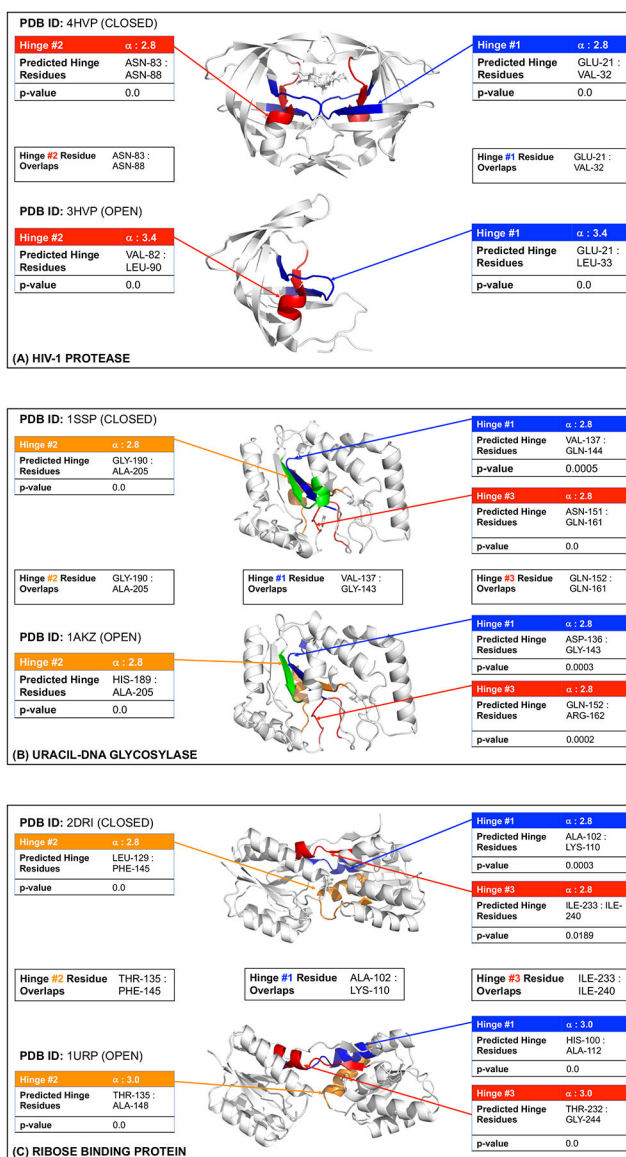


Figure 5.
Multiple hinges identified in: a) HIV protease, b) Uracil-DNA glycosylase and c) Ribose binding protein.

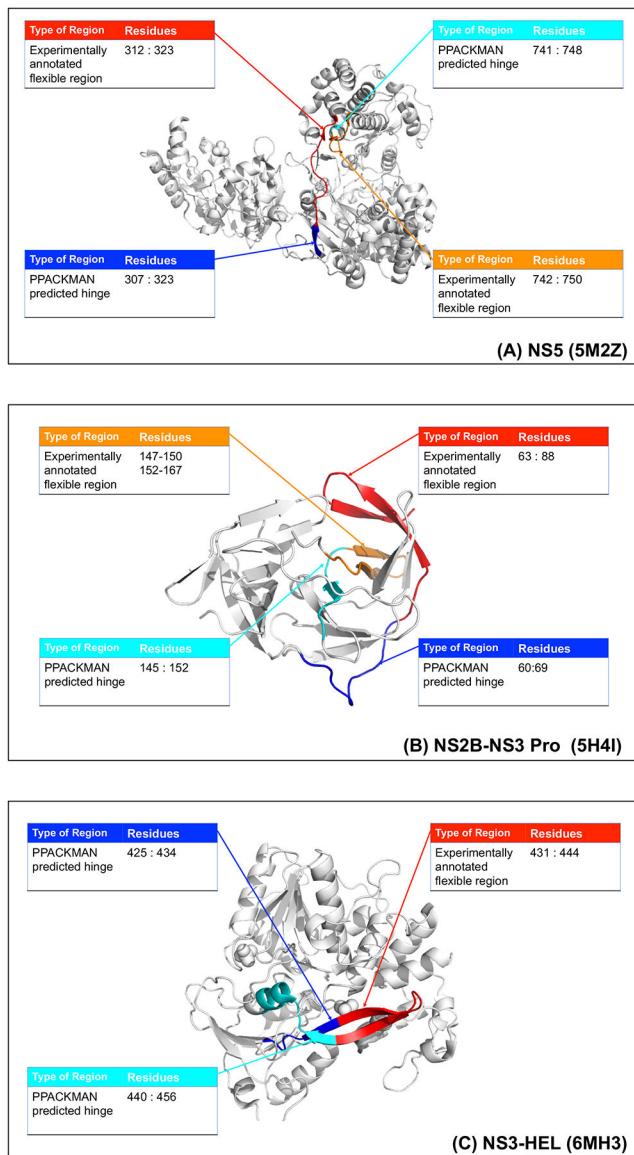


Figure 6.
Multiple hinges in zika virus proteins shown as colored backbone elements a) NS5, b) NS2B bound NS3 protease and c) NS3 helicase.

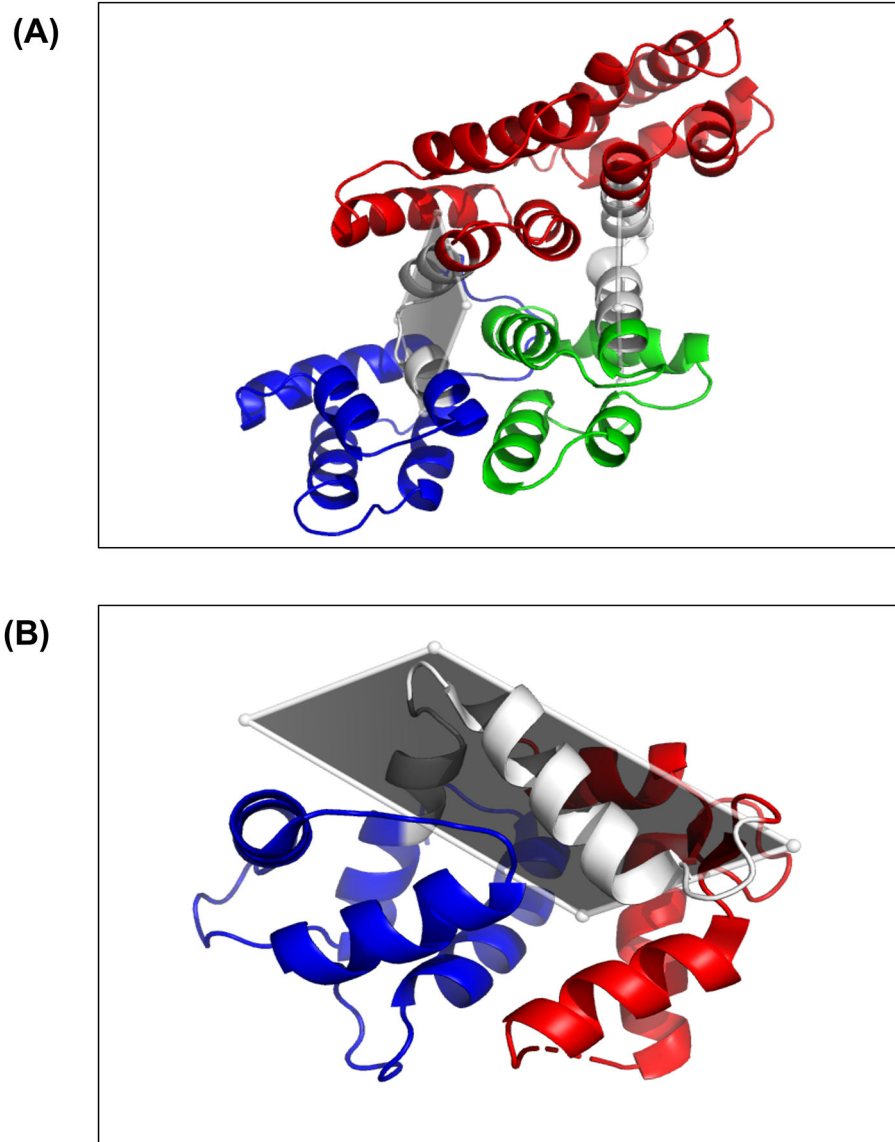


Figure 7: The hinge regions (shown in gray) are important for the movement of protein domains. Here the domains are defined as the area before, after and in between the hinges predicted by PACKMAN in the closed and open conformational state. Predicted hinges in: (A) Annexin (PDB ID: [1AXN](#)) where there are two hinges, one between the red domain (Residues 100–242) and the blue domain (Residues 1:84) and another between the red domain and the green domain (Residues 260–324). (B) Calmodulin (PDB ID: [1PRW](#)) there is only one hinge connecting the red (Residues 1:70) and the blue (97:148) domains.

Table 1.
Hinges predicted for PACKMAN with $\alpha = 2.8$.

Hinge 6 in NS5 has a p-value > 0.05 however, this hinge has been noted as an important flexible region in the literature.

Protein Name	PDB ID	Hinge	Position	p-value
NS5	5M2Z	1	H-307 - N-323	0
		2	T-337 - P-367	0
		3	S-450 - R-483	0
		4	A568 - D-599	0
		5	R-601 - V-606	0
		6	R-741 - W-748	0.097
NS2B-NS3 Pro	5H4I	1	G-21 - V-25	0.0
		2	Q-96 - V-100	0.046
		3	G-133 - I-139	0.001
		4	V-146 - G-151	0.001
NS3-HEL	6MH3	1	F-289 - A-297	0
		2	D-425 - I-434	0
		3	V-440 - R-456	0
		4	Y-473 - L-494	0
		5	L-507 - A-517	0.008