Calculating Ensemble Averaged Descriptions of Protein Rigidity without Sampling

Luis C. González1, Hui Wang2, Dennis R. Livesay1*, Donald J. Jacobs2*

1 Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, North Carolina, United States of America, 2 Department of Physics and Optical Science, University of North Carolina at Charlotte, Charlotte, North Carolina, United States of America

Abstract

Previous works have demonstrated that protein rigidity is related to thermodynamic stability, especially under conditions that favor formation of native structure. Mechanical network rigidity properties of a single conformation are efficiently calculated using the integer body-bar Pebble Game (PG) algorithm. However, thermodynamic properties require averaging over many samples from the ensemble of accessible conformations to accurately account for fluctuations in network topology. We have developed a mean field Virtual Pebble Game (VPG) that represents the ensemble of networks by a single effective network. That is, all possible number of distance constraints (or bars) that can form between a pair of rigid bodies is replaced by the average number. The resulting effective network is viewed as having weighted edges, where the weight of an edge quantifies its capacity to absorb degrees of freedom. The VPG is interpreted as a flow problem on this effective network, which eliminates the need to sample. Across a non-redundant dataset of 272 protein structures, we apply the VPG to proteins for the first time. Our results show numerically and visually that the rigidity characterizations of the VPG accurately reflect the ensemble averaged PG properties. This result positions the VPG as an efficient alternative to understand the mechanical role that chemical interactions play in maintaining protein stability.

Introduction

The set of accessible conformations of a protein is critically dependent upon the arrangement and strength of chemical interactions, which greatly influences the conformational entropy. While there are many different computational models available to characterize protein dynamics [1–4], the computational efficiency and relative accuracy have made network rigidity models particularly attractive [5]. Describing protein structure as a network of constraints that fix the distance between atoms (vertices), the salient feature of network rigidity is to carefully characterize the number of degrees of freedom (DOF) within the network. The number of accessible DOF is generally reduced as chemical interactions are added to the network, which is related to the reduction in phase space upon formation of the interaction. In particular, adding a distance constraint to a flexible region reduces the number of available DOF, while adding one to an already rigid region does not.

While there are a number of graph theoretic algorithms to calculate network rigidity [6–9], the efficient Pebble Game (PG) has emerged as the most popular way to account for protein flexibility [5]. Indeed, various methods based on PG have been developed to analyze network rigidity [10–12], where FIRST [5] has served as the starting point for methods that explore the native conformational dynamics, such as ROCK [13,14] and FRODA [15,16]. Using pebbles to refer to DOF, network rigidity properties of the complete network are quickly calculated based on a strict accountancy of pebbles. Once complete, the PG identifies all flexible/rigid regions within the network. Unfortunately, FIRST, ROCK and FRODA are limited by an athermal formulation, meaning fluctuations within the noncovalent interaction network are not modeled.

Within molecular networks [17], covalent bonds are modeled as quenched constraints (meaning they are ever-present), whereas noncovalent bonds fluctuate on and off. The intermittent nature of the noncovalent interactions reflecting protein dynamics further complicates calculation of average network properties. In this direction, we have developed a statistical mechanical Distance Constraint Model (DCM) [18,19] that is based on a Gibbs ensemble of PG networks that uses network rigidity to account for enthalpy-entropy compensation [20,21]. The result of the DCM approach is that the give and take between protein stability and flexibility is accurately quantified [22–27]. In all works to date, solving of the DCM for protein structures has required average network rigidity properties determined from Monte Carlo sampling across a large sample of network topologies. In this approach, there is a binary on/off designation based on the probability of a constraint to be present or not. This randomness leads to an astronomically large ensemble of networks consisting of $2^N$ possibilities for $N$ constraints that are fluctuating on or off throughout the network. Typically $N$ will range between a few hundred to several thousand in applications. Monte Carlo sampling works markedly well because of self averaging properties of constraint networks. It has been found that for statistical error
Because there can be more than one distance constraint placed between a pair of rigid bodies, the body-bar PG [28] represents the network as a multi-graph, where more than one edge can connect between a pair of vertices. That is, each vertex represents a rigid body, having 6 DOF, and each edge represents a distance constraint. Herein, the framework of the body-bar PG algorithm is generalized, and, interestingly, requires only a minor modification in a way that preserves essentially the same implementation. That is, we have developed a Virtual Pebble Game (VPG) that allows for probabilistic descriptions of the network. The network now has only one edge between a pair of vertices with an assigned weight that defines the capacity for it to absorb DOF. This capacity is given by the average number of constraints that can form between a pair of vertices, thus it needs not be an integer value. The VPG extends the counting of constraints and DOF to real numbers, allowing for fractional DOF, which are viewed as representing the probability to find a DOF. Through this generalization of the PG implementation, the VPG algorithm tracks probability flow that governs where the average number of DOF pool within the effective network, rather than track individual DOF that fluctuate about this average. This approach leads to a dramatic computational speed-up because the PG algorithm dictates sampling over many networks to calculate equilibrium properties, whereas the VPG can probabilistically determine them from a single calculation. The approach of the VPG to calculate ensemble properties without sampling is in the same spirit as other algorithms that tackle important computational biology problems with a very large search space that otherwise would require excessive computation time [29–32].

In a recent report [33], we have demonstrated that the VPG closely reproduces the ensemble averaged counting of DOF within a variety of disordered lattices. The ensemble averaged results over many PG runs is designated \( \overline{PG} \). In this report, key average or consensus network rigidity metrics are directly compared across a variety of disordered lattices. The ensemble averaged results over millions of networks are usually necessary to be sampled \([19]\). The VPG is found to be quantitatively high. As we demonstrate below, the VPG is ideally positioned as a viable alternative to ensemble averaging in the characterization of protein rigidity.

### Materials and Methods

#### Protein Structure Description

We consider a dataset composed of 272 protein structures that are non-redundant at the SCOP [35] family level. Our dataset includes one, two and three domain proteins for PDB codes (see Table 1), that range from 50 to 764 residues. We focus on three types of chemical interactions, which are: intra-residue, linker and hydrogen bond. Note that salt-bridges are considered a special type of H-bond as described previously [5]. The intra-residue interaction models the covalent bonds that exist within a residue. The linker interaction represents the peptide bond that connects the C-N terminal atoms in adjacent residues. The reason we make a distinction between these two types of covalent bonds is due to the number of DOF they consume. While an intra-residue covalent bond (and disulfide bonds if any) consumes five DOF (leaving one for the dihedral angle), the linker consumes six DOF (locking any possible rotation) due to the partial double bond character of the amide group. The last interaction is the hydrogen bond that defines the point where the protein just transitions from being globally flexible to globally rigid, or vice versa. At this rigidity threshold, the greatest fluctuation in network topology occurs, leading to the greatest differences between the \( \overline{PG} \) and VPG quantities. Remarkably, at the rigidity threshold, the similarity in all network rigidity metrics that were calculated using \( \overline{PG} \) and VPG is quantitatively high. As we demonstrate below, the VPG is ideally positioned as a viable alternative to ensemble averaging in the characterization of protein rigidity.

#### Table 1. PDB codes of the proteins in the dataset.

| PDB Code |Description|
|----------|------------|
| 1A1X     | 1A32       |
| 1A76     | 1A8L       |
| 1A92     | 1A9N       |
| 1AEP     | 1AF7       |
| 1AH0     | 1AH5       |
| 1AIH     | 1AK0       |
| 1AKO     | 1AL3       |
| 1ALV     | 1AM6       |
| 1EP3     | 1EQF       |
| 1EF5     | 1E27       |
| 1F02     | 1F0A       |
| 1F0K     | 1F20       |
| 1F5V     | 1F6D       |
| 1FD9     | 1F9N       |
| 1G65     | 1G73       |
| 1G8E     | 1GAK       |
| 1GL4     | 1GPO       |
| 1GQQ     | 1GS5       |
| 1GWU     | 1GWY       |
| 1GXJ     | 1GXY       |
| 1H03     | 1H22       |
| 1H25     | 1H8P       |
| 1H1W     | 1H2N       |
| 1H2A     | 1H2K       |
| 1H2S     | 1H7F       |
| 1H7I     | 1H7J       |
| 1H7L     | 1H7M       |
| 1H7N     | 1H7O       |
| 1H7P     | 1H7Q       |
| 1H7R     | 1H7S       |
| 1H7T     | 1H7U       |
| 1H7V     | 1H7W       |
| 1H7X     | 1H7Y       |
| 1H7Z     | 1H7A       |
| 1H82     | 1H83       |
| 1H84     | 1H85       |
| 1H86     | 1H87       |
| 1H88     | 1H89       |
| 1H90     | 1H91       |
| 1H92     | 1H93       |
| 1H94     | 1H95       |
| 1H96     | 1H97       |
| 1H98     | 1H99       |
| 1H9A     | 1H9B       |
| 1H9C     | 1H9D       |
| 1H9E     | 1H9F       |
| 1H9G     | 1H9H       |
| 1H9I     | 1H9J       |
| 1H9K     | 1H9L       |
| 1H9M     | 1H9N       |
| 1H9O     | 1H9P       |
| 1H9Q     | 1H9R       |
| 1H9S     | 1H9T       |
| 1H9U     | 1H9V       |
| 1H9W     | 1H9X       |
| 1H9Y     | 1H9Z       |
| 1H9A     | 1H9B       |
| 1H9C     | 1H9D       |
| 1H9E     | 1H9F       |
| 1H9G     | 1H9H       |
| 1H9I     | 1H9J       |
| 1H9K     | 1H9L       |
| 1H9M     | 1H9N       |
| 1H9O     | 1H9P       |
| 1H9Q     | 1H9R       |
| 1H9S     | 1H9T       |
| 1H9U     | 1H9V       |
| 1H9W     | 1H9X       |
| 1H9Y     | 1H9Z       |
| 1H9A     | 1H9B       |

doi:10.1371/journal.pone.0029176.t001
Ensemble Averaged Descriptions of Protein Rigidity

The VPG Algorithm

The three main elements of the PG algorithm are pebbles, vertices and distance constraints, which respectively represent DOF, rigid bodies and intramolecular interactions [28]. Note that the justification for the mapping between atoms and rigid bodies, and switching the PG applicable on a bar-joint network to a body-bar network are thoroughly explained in prior works [10,20]. When the body-bar pebble game initiates, all vertices are unconnected and each is ascribed six free pebbles that describe its position and orientation in 3D. When an independent interaction is placed into the network, six trivial DOF are fixed on either one of its incident vertices while the number of distance constraints modeling it are consumed. In the language of the PG algorithm, distance constraints are recursively added to the network, and free pebbles cover the new constraints if they are independent, accounting for DOF removal from the system. Pebbles are not always locally available, but can often be transferred from remote regions of the network. That is, network rigidity is a long-range interaction that can propagate across the network [36,37]. This pebble search function is possible given that pebbles provide directionality in the network, dependent upon which vertex has provided them. A constraint is redundant when a pebble cannot be transferred to cover it.

The search for pebbles in the directed network resembles a network flow problem [38,39], given that the covered capacity of any edge will determine the maximal flow of pebbles through that edge. In a recursive fashion one edge at a time is placed in the network, always following the described process. This way the PG accomplishes its main goal, determining if an edge is independent (fully covered) or redundant (partially or none covered). When a search for pebbles fails and consequently a redundant constraint is found, all the vertices that were involved in the search collapse into a single vertex (with its six trivial DOF), and defines a minimally rigid graph [40], which we loosely refer to as a Laman subgraph [41] because of the analogous concept in two dimensions. This fact allows the PG to run virtually in linear time with the number of vertices in protein-like networks.

The crucial difference between the VPG and the original PG algorithm is the assignment of pebble capacity to edges, and to handle fractional pebbles. The capacity of an interaction represents the maximum number of DOF that it can consume; for linker and intra-residue interactions the capacity is six and five DOF, respectively. The capacity of a H-bond is defined as the product of its probability, \( P_{\text{nat}} \), and the number of distance constraints used to model it, which is three. Therefore, consider a network consisting of vertices \( \{v_i\}_{i=1,2,\ldots,N} \), with a list of edges \( \{e_m\}_{m=1,2,\ldots,M} \). The capacity for the \( m \)-th edge is denoted by \( c_m \). The VPG follows the following procedures and operations:

1. Initialize the graph with a set of isolated vertices \( \{v_i\} \), with the free DOF of each vertex \( v_i \) being 6.
2. From the list of edges \( \{e_m\} \), insert edge \( e_k \) with capacity \( c_k \) into the graph. Let \( v_i \) and \( v_j \) be the two incident vertices for edge \( e_k \).
3. Collect 6 pebbles for vertex \( v_i \) by doing a breadth first search.
4. Flag vertex \( v_j \) as visited, try to collect \( c_k \) pebbles for vertex \( v_j \) by doing a breadth first search while holding the 6 pebbles on \( v_i \) in place. If not all \( c_k \) pebbles can be found in one trial, continue to collect more pebbles by carrying out the search repetitively until there are enough free pebbles on \( v_j \) to cover edge \( e_k \), or if no new pebbles are found (a failed search).
5. If \( e_k \) or more pebbles are collected on vertex \( v_j \), cover edge \( e_k \) with \( c_k \) pebbles. Otherwise, all the visited vertices within the failed search are condensed into a single vertex. If \( \{e_m\} \) is not empty, go to step 2.
6. End of VPG.

Rigid Cluster Decomposition

After having placed all the constraints, the PG and VPG algorithms determine the number of DOF left in the network. The trivial case is when there are just six DOF remaining, indicating that the network is globally rigid and all vertices are contained in a single rigid cluster. When there are greater than six remaining DOF, pebble location identifies which regions of the protein network are flexible or rigid. Excess pebbles identify flexible

![Figure 1. The respective network descriptions are compared.](image-url)
regions, whereas rigid regions occur when no free pebbles beyond 6 are accessible. From this information, it is possible to apply a Rigid Cluster Decomposition (RCD) to localize groups of vertices that move together as a rigid body. A rigid cluster is a subgraph with all of its vertices completely rigid among themselves.

The process of finding rigid clusters in the VPG proceeds as follows: for any pair of vertices add a hypothetical edge, then try to cover it with $e > 0$ DOF, while six DOF are fixed on one of the incident vertices. If an excess number of DOF is found, then both vertices do not belong to the same rigid cluster, otherwise a failed search is declared and all the vertices involved in the search are part of the rigid cluster. Fig. 2 presents two example RCD cases. Notice that all the edges have been covered and they have different capacities. In the first case (Fig. 2a), there is a total of 7.4 available DOF. Therefore for any pair of vertices in the network, it will always be possible to gather 6+$e$ DOF with $e > 0$ representing excess DOF, which indicates all constraints are independent and the network is globally flexible. For the second case (Fig. 2b), the number of available DOF is exactly six (on vertex four). Therefore, no excess DOF (i.e. $e = 0$) will be found under any circumstance, and this condition indicates that a rigid cluster is present that includes all five vertices.

It is worth emphasizing that the point of these examples is to show that the data structure for the VPG is essentially the same as the PG. On the other hand, the edge pebble capacities are not shown in Fig. 2 for either example. Yet, it can be surmised that the capacities of each edge in example (a) is precisely equal to the numbers assigned to the edges (i.e. 5, 5 and 0.6). If this were not the case, assuming one of these edges had a greater capacity, then some or all of the excess 1.4 DOF that is currently remaining would be used to cover the edges. Conversely, because there are no excess DOF in example (b) it is clear that either all the edges are being covered at their maximum capacity, or, their capacity is larger than the sum of the number of pebbles that cover the edge (on both sides must be added). If the capacity of an edge is larger than the total amount of pebbles covering it, then this would indicate that the edge is redundant. If this were the case, many vertices could be collapsed into a single vertex as explained above in regards to failed pebble searches, and creation of Laman subgraphs.

Network Similarity Metrics

We employ two distinct metrics to compare the networks identified by the VPG and PG algorithms. To quantify rigid cluster similarity, we employ the Rand Measure (RM) [42]. The RM is a very well suited metric to compare clustering within a network. In the case of rigid cluster decomposition, both the body-bar PG and its VPG counterpart assign a unique label to each vertex to indicate the cluster it belongs to within the network. The network will generally consist of many rigid clusters. The RM is a combinatory count of all possible pairs of vertices where it counts all the cluster composition coincidences between the two networks generated by the two approaches. If both networks have the exact same rigid cluster decomposition, then RM is equal to 1. In

**Figure 2. Two different rigid cluster decomposition examples are compared.** In the first example, (a), there are 1.4 free pebbles available (located on vertices $V_1$ and $V_2$), whereas the capacities of edges $(V_1, V_2)$, $(V_1, V_3)$, and $(V_2, V_3)$ is, respectively, 0.6, 5.0, 5.0. If a hypothetical edge is added between any pair of vertices, there is always going to be possible to find DOF, therefore the three vertices result in single bodies (highlighted by color differences). Conversely, in the second example, (b), only the six trivial DOF can be found (on vertex $V_4$). That is, no free pebbles remain in the network (they have all been consumed by the edges). As such, the five vertices belong to a single rigid cluster.

doi:10.1371/journal.pone.0029176.g002
general the RM has a range between 0 to 1. Zero is only possible if one network consist of all vertices within one cluster, while in the other network all vertices are in separate clusters (each vertex has its own unique label).

For a specific pair of vertices, there are two cases in which a match is found between the two networks. In the first case, the two vertices in network 1 belong to the same cluster (they have the same cluster label) and likewise, in network 2 both vertices belong to the same cluster. In the second case, the two vertices have different labels in network 1, indicating they belong to two different clusters, and likewise, in network 2 the two vertices belong to two different clusters. On the other hand, a match is not found if in one network the two vertices belong to the same cluster, while in the other network they belong to two different clusters.

The RM is calculated by the total number of matches divided by the total number of possible pairs. A RM greater than 3/4 is a strong indicator of good agreement. A formal definition as given in [42] is: given \( N \) points, \( X_1, X_2, \ldots, X_N \), and two clusterings of them \( Y = \{ Y_1, \ldots, Y_k \} \) and \( Y' = \{ Y'_1, \ldots, Y'_k \} \) there is defined:

\[
\phi(Y, Y') = \frac{\sum_{i<j} \lambda_{ij}}{\binom{N}{2}}
\]

where \( \lambda_{ij} = 1 \) if \( k \neq k' \) and \( X_i \) and \( X_j \) are in both \( Y_k \) and \( Y'_k \) or if \( X_i \) is in both \( Y_k \) and \( Y'_k \) and \( X_j \) is neither \( Y_k \) nor \( Y'_k \), otherwise \( \lambda_{ij} = 0 \).

We also compare the rigidity assessment between the majority vote from PG to the VPG assignment for each non-linker torsion bond in a protein. This provides a very sensitive metric to assess how well the VPG reflects the consensus results from a large sampling of PG runs. That is, we count the number of times that both approaches agree in their rigid versus flexible assessment, normalized by the total number of comparisons. This calculation leads to an agreement measure \( AM \) that ranges from \(-1 \) to \(1 \).

When the rigidity estimate from VPG matches the majority vote among all PG realization (there is a rigid torsion by consensus in PG corresponds to a flexible torsion in VPG), a match is found between the two networks (VPG deemed as rigid the \( n \)-th torsion).

\[
AM = \text{(measure - 50)} / 49
\]

where \( N_{\text{wrong}} \) is the count of times that VPG disagreed with PG, \( N_{\text{agree}} \) is the count of times that PG matched the VPG, and \( N_{\text{total}} = N_{\text{agree}} + N_{\text{wrong}} \) is the total number of realizations for the PG. For instance, if a particular torsion has a value of \( AM = -1 \), it indicates that the VPG assesses the torsion to be rigid, whereas the PG indicates the opposite (flexible) in all of the realizations. When \( AM = 0 \), this is considered perfect agreement, and when \( 0 < |AM| < 0.1 \), there is disagreement between the consensus PG vote and the VPG prediction, but the minority and majority votes from PG are very close, where the difference is comparable to the intrinsic sampling error bars.

### Rigidity Profiles

To complement the analysis above, we also graphically compare two additional descriptions of network rigidity that resemble contact maps. The Rigid Cluster Map (RCM) is a \( N \times N \) symmetric matrix that identifies co-rigid \( x \)-carbon pairs within protein structure. By definition the main diagonal is rigid (an \( x \)-carbon is rigid with respect to itself). When constructing the RCM matrix, if a pair of \( x \)-carbons belong to the same rigid cluster, a value of 1 is assigned to the intersection of both vertices (specifying the row/column of the matrix), else 0 is given. For one run of the PG, the RCM is a binary plot simply highlighting co-rigid residue pairs. The PG RCM plots are based on a majority rule across the ensemble. That is, if 50% or more of the realizations is rigid a 1 is assigned, otherwise 0 is assigned. For the VPG, there is only one run, and the output will be 1 or 0. Since the RCM is symmetric, the lower triangle shows the VPG, while the upper triangle shows the PG results.

Further, we also employ Mechanical Coupling Maps (MCM) to characterize how flexibility propagates throughout the structure. The MCM quantifies the degree of flexibility of each \( x \)-carbon in a protein relative to a reference \( x \)-carbon, which serves as a rigid body anchor to eliminate the trivial rigid body translations and rotations. To calculate the MCM, the maximum number of excess DOF shared between the \( x \)-carbon of interest and the reference \( x \)-carbon must be determined. Operationally, this is accomplished by first fixing the trivial six DOF on the reference \( x \)-carbon, and then launch a pebble search on the other \( x \)-carbon to gather the maximum number of internal DOF. Note that the result does not depend on which \( x \)-carbon is selected as a reference, as the result depends only on the \( x \)-carbon pairs. For normalization purposes, the number of internal DOF found is divided by six (being the maximum number of DOF that an \( x \)-carbon can have). This information is presented using a color code scheme in the MCM that ranges from 0 to 1. Since the number of DOF that can be found in the VPG can be fractional (not binary like the RCM), the proper comparison to PG requires the MCM values from the PG runs to be averaged across the ensemble. The MCM thus provides a more nuanced view of network rigidity than the RCM. Because the MCM is also symmetric, again the lower triangle shows the VPG, while the upper triangle shows the PG results.

### Results and Discussion

**Characterizing Rigid Cluster Similarity**

Characterizing the rigid clusters offers a unique view in terms of the role that chemical interactions play within proteins. In prior work, we have used rigid cluster decomposition of protein structure to provide a statistically significant description of thermodynamic coupling within double mutant cycles [43]. Moreover, there have been many investigations characterizing the loss of rigidity that occurs upon protein unfolding using a H-bond dilution model [11,34,44–48]. Finally, PG characterizations of rigidity have been used to explain the increased stability of thermophilic proteins [49], RNA function [50], the effects of ligand binding [5] and the identification of critical interactions [51]. In these works, an energy cutoff is used to identify which H-bonds are present. As the energy cutoff is lowered, less H-bonds...
are included in the structure, thus increasing flexibility. Equivalently, as the energy cutoff is raised, more H-bonds are included in the structure, thus increasing rigidity.

In an analogous way, we vary $P_{nat}$ from 0 to 1 in order to control the number of H-bonds present in the network. One technical difference here, is we treat all H-bonds as equivalent, and ignore their energies altogether. In the above mentioned previous works, H-bond energy was used to characterize the strength of a H-bond so that the weakest H-bonds can be removed before the strongest H-bonds to study protein unfolding. The reason for intentionally treating all H-bonds as equivalent in this this work is because we are interested in testing how good the VPG can represent $PG$ over the ensemble. If some H-bonds are almost always present (lowest energy H-bonds) while some H-bonds are almost always broken (high energy H-bonds), the VPG results will be closer to the $PG$ results because less fluctuations will occur in the H-bond network, meaning the comparisons herein correspond to the worst-case scenario. We have tested and verified this dependence on the fluctuations present in the H-bond network, and we will publish a more physically realistic H-bond dilution protocol elsewhere that models protein unfolding. However, the interest in this report is to show that the VPG provides an excellent approach to characterize protein flexibility/rigidity properties even in the extreme case of uniform H-bond strength.

Under this H-bond dilution strategy, the capacity defined by $3 \times P_{nat}$ is the number of DOF that will be removed from the VPG when an H-bond is independent (recall each H-bond is described by three distance constraints). For the PG counterpart, a random number between zero and one is assigned to each possible H-bond in each PG realization to determine if it is present or not (H-bond is present if $RAND(0,1) < P_{nat}$). Note that empirically we find that for a given $P_{nat}$, an ensemble of 200 realizations is typically good enough to make robust predictions across our protein dataset. Since we are using $PG$ to define the exact answer to compare against the VPG, we run the PG 1000 times to reduce statistical error bars by $\sqrt{5}$ relative to what is found in actual applications.

As discussed above, the Rand Measure [42] (RM) compares the rigid cluster decompositions from the VPG and $PG$. Fig. 3a presents the RM calculation across the range $P_{nat}$ values for four exemplar protein structures. The four example proteins span a range of sizes (from 64 to 315 residues) and topological architectures. Specifically, they are the chemotaxis receptor methyltransferase CheR structure (pdbid = 1AF7) [52], the FLAP endonuclease from M. jannaschii (pdbid = 1A76) [53], a small scorpion protein toxin (pdbid = 1AHO) [54] and the disulfide oxidoreductase from P. furiosus (pdbid = 1A8L) [55]. In each, there is a region where the RM decreases sharply, which corresponds to the worst-case situation when the fluctuations are

Figure 3. Quantifying $PG$ and VPG similarity. (a) The Rand Measure (RM) is plotted versus $P_{nat}$ for four exemplar proteins that span a range of sizes (from 64 to 315 residues) and topological architectures. All proteins across the full dataset have the same characteristic shape where the minima in RM is related to the protein structure’s rigidity transition. The $P_{nat}$ value corresponding to the worst RM is defined as $PRM$. (b) Histogram detailing $PRM$ values for each protein within our dataset. Encouragingly, an overwhelming majority of cases have RMs greater than 80%. (c) Histogram detailing the agreement measure for each backbone torsion within our dataset at each protein’s respective $PRM$ value. (d) Histogram detailing the Pearson correlation coefficient comparing the $PG$ and VPG mechanical coupling maps across the dataset at each protein’s respective $PRM$ value. doi:10.1371/journal.pone.0029176.g003
maximized. For most networks the point of maximum fluctuation identifies the rigidity threshold, representing the transition from flexible to rigid. A similar pattern was detected across our entire dataset, which appears at values as low as $P_{nat} \approx 0.60$ and ends at values as high as $P_{nat} \approx 0.95$. To calculate RM at each $P_{nat}$, each one of the 1000 PG realizations is compared to the single VPG description. The $\overline{PG}$ RM value is simply the average of the RM 1000 realizations. The $\overline{P_{nat}}$ at the minimum in RM, designated $P_{RM}$, identifies the worst-case scenario.

The high RM values indicate that the rigid cluster decomposition is very similar across the $PG$ and VPG. To emphasize this point, we identify the PG realization that yields the median RM score across the entire ensemble, which is called $PG^{med}$. This rigid cluster decomposition for this point is plotted (using the same four proteins) in the first column of Fig. 4. Color differences indicate different rigid clusters, whereas grey indicates a flexible region. The middle column identifies the rigid clusters identified by the VPG. While the similarity is apparent by just qualitatively comparing the rigid cluster decompositions from each algorithm, the difference plots in the third column are the most compelling. Red coloring identifies regions that disagree in rigid cluster composition, whereas grey indicates agreement.

Expanding to our entire dataset, Fig. 3b plots a histogram of all RM scores at the respective $P_{RM}$ value for each protein. The worst-case RM scores are encouragingly large ($> 80\%$) for an overwhelming majority of the proteins, thus indicating that the rigid clusters identified by the two algorithms are quite similar. Slight shifts to the considered $P_{nat}$ to just above and below $P_{RM}$

| PG | VPG | Disagreement ($|PG-VPG|$) |
|----|-----|--------------------------|
| ![Scorpion protein toxin, $P_{nat} = 0.62$](image) | ![Disulfide oxidoreductase, $P_{nat} = 0.57$](image) | ![Methyltransferase CheR, $P_{nat} = 0.72$](image) |
| ![Flap endonuclease, $P_{nat} = 0.75$](image) | ![Flap endonuclease, $P_{nat} = 0.75$](image) | ![Flap endonuclease, $P_{nat} = 0.75$](image) |

**Figure 4. Rigid Cluster visualizations for four example proteins.** The first column highlights the rigid clusters identified within the PG realization that corresponds to the median RM value, designated $PG^{med}$. The middle column corresponds to the rigid clusters identified by the VPG. Finally, differences between the two algorithms are highlighted in the third column. doi:10.1371/journal.pone.0029176.g004
negligibly affects the histograms. Note that when $P_{nat} \rightarrow 0$ or $P_{nat} \rightarrow 1$ the fluctuations within the network are suppressed, meaning the algorithms become identical. Consequently, the mechanical descriptions converge. Fig. 3a typifies this result, where the two approaches produce identical results ($RM \approx 1$) at small and large values of $P_{nat}$.

**Over versus Under Prediction of Rigidity**

The RM indicates that differences in the rigid cluster decomposition for PG and VPG occur, but the RM does not characterize where and how the differences take place. Therefore, it is important to determine if the VPG tends to systematically over- or under-estimate rigidity in the protein. To determine how often each type of error appears, we quantify similarity within the rigidity of all rotatable backbone ($\phi$ and $\psi$) dihedral angles (torsions). The agreement measure (AM) described above is applied here for three specific proteins, and across the entire dataset. Fig. 5a, b and c present histograms of AM values for three of the proteins from above. In panel (a), it is shown that the VPG slightly overestimates the amount of rigidity within the methyltransferase CheR structure, whereas panel (b) indicates that it slightly underestimates the amount of rigidity within FLAP endonuclease. Panel (c) shows that VPG overestimates the amount of rigidity within the disulfide oxidoreductase. Fig. 3c presents a histogram of the entire dataset. Clearly, the overwhelming majority (>92%) of torsions are in close agreement within statistical uncertainty of $\bar{T}G$. Strong disagreement ($|AM| > 0.1$) between both algorithms is minimal, especially considering the

**Figure 5. Agreement measure (AM) results.** AM histograms for the (a) methyltransferase, (b) FLAP endonuclease and (c) disulfide oxidoreductase at their respective $P_{RM}$ = values. (d) Differences between the $\bar{T}G$ and VPG are mapped to the ribosylglycohydrolase structure from *M. jannaschii*, which is presented as a typical case. Red coloring indicates that the VPG overestimates rigidity relative to $\bar{T}G$, whereas blue indicates an underestimate. Across our dataset, as shown in this example, differences occur most frequently in loop regions.

doi:10.1371/journal.pone.0029176.g005

**Figure 6. Rigid cluster maps (RCM) of chemotaxis receptor methyltransferase CheR structure is plotted at two different $P_{nat}$ values.** Red coloring identifies residue pairs that are co-rigid. $\bar{T}G$ results are presented in the upper triangle, whereas the VPG is presented in the lower. At $P_{nat} = 0.60$, the protein is mostly flexible due to a lack of crosslinking H-bonds. However, the structure becomes increasingly rigid as H-bonds are added to the network. At $P_{nat} = 0.75$, the VPG slightly under-predicts the extent of rigidity. For this protein $P_{RM} = 0.80$.

doi:10.1371/journal.pone.0029176.g006
comparison is at the $P_{RM}$ of each protein that defines the worst-case. Nonetheless, it is interesting to identify when discrepancies are most likely to occur. A survey of the differences reveals that they generally occur in loop regions and edges of secondary structures, as typified in Fig. 5d.

**Rigidity Profile Similarity**

We use Rigid Cluster Maps (RCM) to visually highlight pairwise mechanical couplings within structure, using red marks to highlight $\alpha$-carbon pairs within the same rigid cluster, otherwise no mark is provided. For ease of comparison, the RCM results are presented in the upper triangle, whereas the VPG results are presented in the lower triangle. By construction, the protein backbone corresponding to the RCM diagonal is always rigid in both variants. Using the methyltransferase structure from above as a typical case, the two panels in Fig. 6 correspond to two different values of $P_{nat}$, ranging from a completely flexible (unfolded) structure with few crosslinking H-bonds to a predominantly rigid structure with many crosslinking H-bonds. As one can see, the VPG and RCM algorithms give very similar results.

Going a step further, Fig. 7 presents the RCMs of our four example proteins near $P_{RM}$, thus $P_{nat}$ is corresponding to the critical region. The presented values are slightly shifted from exact $P_{RM}$ values to highlight interesting features. Note that the changes in $P_{nat}$ actually make the RCM plots appear more dissimilar. The large square region along the backbone corresponds to a rigid $\alpha$-helix. A similar pattern is observed in the disulfide oxidore-

---

**Figure 7. Rigid Cluster Maps (RCM) for four different example proteins near their respective $P_{RM}$ values.** RCM results are presented in the upper triangle, whereas the VPG results are presented in the lower. Note that the presented proteins are the same from Fig. 3a. doi:10.1371/journal.pone.0029176.g007

**Figure 8. Mechanical Coupling Maps (MCM) provide a more nuanced description of co-rigidity.** Specifically, the continuous scale provides a normalized description of how many free pebbles (DOF) are shared between each residue pair ($0 = 0$ pebbles, whereas $1 = 6$ pebbles). Again, each MCM is plotted near their respective $P_{RM}$ values for the same four proteins presented Figs. 3a and 7. doi:10.1371/journal.pone.0029176.g008
ductase, which also has few crosslinking interactions at this value of $P_{nat}$. Conversely, the off-diagonal features are mostly conserved in the methyltransferase CheR structure, but the VPG slightly overestimates the extent of rigidity within the core region (residues $\sim 75-150$). The FLAP endonuclease example provides the most interesting visual differences between the two approaches, where the VPG underestimates the $\overline{PG}$ predictions. That is, the VPG fails to identify rigid clusters present within the $\overline{PG}$. However, the differences are found to be much less severe on closer inspection regarding the number of available DOF. While there are no free pebbles within the $\overline{PG}$ in these regions, the probabilistic VPG identifies a tiny nonzero fraction ($3 \times 10^{-3}$). Clearly, this difference is negligible, but the binary RCM makes the difference appear much larger than it actually is.

The Mechanical Coupling Maps (MCM) provide a more nuanced view of rigid cluster decomposition. Unlike the binary RCMs, MCMs are based on a continuous scale that identifies the fractional number of pebbles shared between each $\alpha$-carbon pair. In this sense, the MCMs are similar to the cooperativity correlation plots calculated by our statistical mechanical DCM [18,19,22–25]. Fig. 8 compares the MCMs for the same four proteins in Fig. 7, using the same $P_{nat}$ values. The rigidity over-prediction by the VPG in the methyltransferase example is again clear. However, there is appreciable co-rigidity within the residue pairs contained within the range of residues $\sim 60-80$ and residues $\sim 80-150$, which was identified as flexible by the RCM. Additionally, the MCMs reveal a more interesting set of similarities throughout the plots. In the same way, the similarity within FLAP endonuclease is also more pronounced, although the VPG again somewhat overestimates the extent of rigidity. Conversely, the MCMs actually show more dissimilarity within the two examples without any off-diagonal RCM components. In both, there is marginal co-rigidity identified by the $\overline{PG}$ (the reddish shadowing) due to some rigidity fluctuations throughout the ensemble that is suppressed by the VPG.

Expanding across the entire dataset, Fig. 3d provides a histogram of the Pearson correlation coefficient between the MCM matrices calculated by the $\overline{PG}$ and the VPG. Clearly, the VPG is consistently a good estimator of the $\overline{PG}$ behavior. This point is strengthened by Fig. 9, which compares the Pearson correlation coefficients of the MCM of each unique PG realization to the $\overline{PG}$ plot for the same four proteins considered.
above. That is, the boxplots describe the intrinsic variability across the PG ensemble. Within each boxplot, the horizontal grey line indicates the median RM value across the PG distribution, whereas the top and bottom of the box indicates the upper and lower quartiles. The whiskers describe the rest of the distribution, and the dots identify outliers (corresponding to the default settings of R). The red line corresponds to the similarity between the PG and VPG MCMs, which is encouragingly strong. In fact, the VPG similarity to the PG behavior is better than the third quartile in all cases. This result clearly indicates that VPG approximates PG behavior better than the vast majority of the single PG realizations. These comparisons are calculated at the same value of $P_{RM}$ as above, meaning they again correspond to the critical region.

The Rigidity Transition

Following earlier works [18,19,22,25,34], we define $P_t$ (for transition) as the peak in the rigid cluster susceptibility (RCS) curve, which is defined as the reduced second moment of rigid cluster size. That is, the peak in RCS identifies the point in which the rigid cluster sizes are maximally fluctuating, indicating a transition from a globally flexible to globally rigid network. Twelve examples (including the four proteins discussed above) of RCS curves using the PG and VPG approaches are shown in Fig. 10, all of which are qualitatively similar. As shown by the scatter plot in Fig. 11a, the rigidity transitions identified by the PG and VPG algorithms are highly correlated. In addition, the Average Cluster Size (ACS) at $P_t$ is also highly correlated across the two algorithms (Fig. 11b). Since $P_t$ identifies the $P_{nat}$ value with maximal

Figure 10. Rigid Cluster Susceptibility (RCS) is plotted versus $P_{nat}$ for 12 typical protein examples (PG = solid line and VPG = dashed line). Note that the proteins presented in the first column are the same from Fig. 3a.
doi:10.1371/journal.pone.0029176.g010
variability within the rigid cluster sizes, it is expected to be related to the $P_{RM}$ because the single VPG mean-field calculations suppresses fluctuations. This is indeed the case as indicated by the strong correlation between $P_t$ and $P_{RM}$ for both the PG and VPG algorithms (Fig. 11c–d).

Conclusions

In this report, we demonstrate that ensemble averaged PG properties, which requires sampling, is approximated well by a single mean field calculation. That is, the probabilistic VPG accurately reproduces a number of ensemble-averaged network rigidity properties. The high values of the RM clearly indicate that the rigid cluster decompositions are very similar, especially for $|P_{nat} - P_{RM}| > 0.2$. The AM and structural comparisons of the rigid clusters respectively provide quantitative support for this point. Comparisons of the RCM and MCM rigidity profiles between the PG and VPG variants also indicate that the calculated rigidity properties are highly similar. In fact, the PG to VPG MCMs are much more similar than the intrinsic variability across the ensemble of PG snapshots. Finally, the mechanical transitions identified by the peak in the rigid cluster susceptibility curves are highly correlated across the two variants. Taken together, these results collectively demonstrate the utility and power of the virtual pebble game that deals with the probability of finding pebbles rather than the pebbles themselves.

Author Contributions

Conceived and designed the experiments: DRL DJJ. Performed the experiments: LCG. Analyzed the data: LCG. Contributed reagents/materials/analysis tools: LCG HW. Wrote the paper: LCG DRL DJJ.

Figure 11. Rigidity transition effects. (a) The rigidity transition ($P_t$) is compared across the PG and VPG algorithms. (b) Similarly, the average cluster size (ACS) at their respective $P_t$ values are compared across the two algorithms. The value of $P_{nat}$ with the worst RM (called $P_{RM}$) is compared to $P_t$ calculated using the (c) VPG and (d) PG. The linear relationships occur because the mean field approximation is maximally inaccurate in this range. Note, a few proteins do not have completed peaks in their rigid cluster susceptibility curves because the protein never crosses the rigidity transition, which have been excluded from panels (c) and (d).

doi:10.1371/journal.pone.0029176.g011
References

1. Sabhijy J (2010) Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. Corr Open Pharmacol 10(6): 738–744.
2. Swier MC, Chong LT (2010) Reaching biological timescales with all-atom molecular dynamics simulations. Corr Open Pharmacol 10(6): 745–752.
3. Jacobs DJ (2010) Ensemble-based methods for describing protein dynamics. Corr Open Pharmacol 10(6): 760–769.
4. Rader AJ (2010) Coarse-grained models: getting more with less. Corr Open Pharmacol 10(6): 753–759.
5. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF (2001) Protein extensibility prediction using graph theory. Protein Struct Func Genet 44(2): 150–165.
6. Sughara K (1980) On redundant bracing in plane skeletal structures. Bull Electrotech Lab 44: 376–386.
7. Imai H (1985) On combinatorial structures of line drawings of polyhedra. Discrete Appl Math 10(1): 79–92.
8. Gabow HH, Westermann HH (1988) Forests, frames and games: Algorithms for matroid sums and applications. In: 20th Annual ACM Symposium on the theory of Computing, pp 407–421.
9. Hendrickson B (1992) Conditions for unique graph realizations. SIAM J Comput 21(1): 65–84.
10. Jacobs DJ (1998) Generic rigidity in three-dimensional bond-bending networks. J Phys A: Math Gen 31(31): 6633–6668.
11. Jacobs DJ, Kuhn LA, Thorpe MF (1999) Flexible and rigid regions in proteins. In: Thorpe MF, Dudby PM, eds. Rigidity theory and applications, Plenum Publishing, NY, pp 357–384.
12. Thorpe MF, Chubynsky Y, Hesperheide BM, Menor S, Jacobs DJ, et al. (2005) Current topics in physics, Imperial College Press, London, chapter 6, pp 97–112.
13. Lei M, Zavodsky MI, Kuhn LA, Thorpe MF (2004) Sampling protein conformations and pathways. Journal of Computational Chemistry 25(9): 1135–1148.
14. Zavodsky MI, Lei M, Thorpe MF, Day AR, Kuhn LA (2004) Modeling correlated main-chain motions in proteins for exible molecular recognition. Proteins: Structure, Function and Bioinformatics 57(2): 243–261.
15. Wells SA, Menor S, Hesperheide BM, Thorpe MF (2005) Constrained geometric simulation of diffusive motion in proteins. Phys Biol 2: S127S136.
16. Farrell DW, Kirill S, Thorpe MF (2010) Generating stereochemically acceptable protein pathways. Proteins 78: 2908–2921.
17. Greene LH, Higman VA (2003) Uncovering network systems within proteins. J Mol Biol 334: 781–791.
18. Livesay DR, Dallakyan S, Wood GG, Jacobs DJ (2004) A exible approach for understanding protein stability. FEBS Lett 576(3): 468–476.
19. Jacobs DJ, Dallakyan S (2005) Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. Biophys J 88(2): 903–915.
20. Vorov OK, Livesay DR, Jacobs DJ (2009) Helix/coil nucleation: A local response to global demands. Biophys J 97(11): 5000–5009.
21. Vorov OK, Livesay DR, Jacobs DJ (2011) Nonadditivity in a conformational entropy upon molecular rigidification reveals a universal mechanism affecting folding cooperativity. Biophys J 100(4): 1129–38.
22. Livesay DR, Jacobs DJ (2006) Conserved quantitative stability/exibility relationships (qse) in an orthogonal muscle 1 pair. Protein 62(1): 130–143.
23. Jacobs DJ, Livesay DR, Hules J, Tasayco ML, Jacobs DJ (2006) Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family. Chem Cent J 2(17).
24. Mortensen JM, Xu M, Jacobs DJ, Livesay DR (2009) Unifying mechanical and thermodynamic descriptions across the thioroaxin protein family. Protein 73(3): 610–627.
25. Verma D, Jacobs DJ, Livesay DR (2010) Predicting the melting point of human c-type lysozyme mutants. Curr Prot Pept 11(7): 562–572.
26. Mortensen JM, Jacobs DJ, Livesay DR (2010) Allometric response is both conserved and variable across three eel orthologs. Biophys J 99(7): 2242–2254.
27. Hesperheide BM, Jacobs DJ, Thorpe MF (2004) Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. J Phys: Condens Matter 16: S5055–64.
28. Roberts KE, Cushing PR, Boisguerin P, Madden DR, Donald BR (2011) Design of protein-protein interactions with a novel ensemble-based scoring algorithm.

In: Proceedings of the 15th Annual international conference on Research in computational molecular biology, Berlin, Heidelberg: Springer-Verlag, RECOMB’11. pp 361–376. URL: http://dl.acm.org/citation.cfm?id=1985747.1985762.
30. Frey KM, Georgiev I, Donald BR, Anderson AC (2010) Predicting resistance mutations using protein design algorithms. Proceedings of the National Academy of Sciences 107: 13707–13712.
31. Chen CY, Georgiev I, Anderson AC, Donald BR (2009) Computational structure-based redesign of enzyme activity. Proceedings of the National Academy of Sciences.
32. Georgiev I, Lilien RH, Donald BR (2008) The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. Journal of Computational Chemistry 29: 1527–1542.
33. González LG, Wang H, Livesay DR, Jacobs DJ (in review) A virtual pebble ensemble average for rigidity.
34. Rader AJ, Hesperheide BM, Kuhn LA, Thorpe MF (2002) Protein unfolding: Rigidity lost. PNAS 99(6): 35803545.
35. Murzin AG, Brenner SE, Hubbard T, Chotia C (1995) Scop: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4): 536–549.
36. Guyon E, Roux S, Hansen A, Bêteau D, Troadec JP, et al. (1990) Non-local and non-linear problems in the mechanics of disordered systems: application to granular media and rigid bodies. Phys Rev E 46: 373–419.
37. Jacobs DJ, Thorpe MF (1995) Generic rigidity percolation: The pebble game. Phys Rev Lett 75(22): 4051–4054.
38. Ford LR, Fulkeron DR (1962) Flow in Networks Princeton University Press.
39. Cornen TH, Leiserson CE, Rivest RL (1991) Introduction to Algorithms. MIT Press and McGraw-Hill.
40. Tay TS (1984) Rigidity of multi-graphs. i. linking rigid bodies in n-space. Journal of Combinatorial Theory, Series B 36(1): 95–112.
41. Laman G (1978) On graphs and rigidity of plane skeletal structures. J Eng Math 4(4): 331–340.
42. Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66(336): 846–856.
43. Istomin AV, Gromiha MM, Vorov OK, Jacobs DJ, Livesay DR (2000) New Insight into long-range nonadditivity within protein double-mutant cycles. Proteins 70(3): 913–924.
44. Hesperheide BM, Rader AJ, Thorpe MF, Kuhn LA (2002) Identifying protein folding cores from the evolution of exible regions during unfolding. J Mol Graphics Modeling 21(3): 193–207.
45. Rader AJ, Anderson G, Ising B, Khosran HA, Bahar I, et al. (2004) Identification of core amino acids stabilizing rhodopsin. PNAS 101(19): 7246–7251.
46. Rader AJ, Bahar I (2004) Folding core predictions from network models of proteins. Polymer 45(2): 659–668.
47. Tan H, Rader AJ (2008) Identification of putative, stable binding regions through exibility analysis of hm-1 ggs120. Proteins 74(4): 801–894.
48. Wells RA, Jimenez-Roldan IC, J. C. Romero RA (2009) Comparative analysis of rigidity across protein families. Phys Biol 6(4).
49. Radestock S, Gohilke H (2011) Protein rigidity and thermophilic adaptation. Proteins 79(4): 1089–1108.
50. Folle S, Gohilke H (2009) Constraint counting on rna structures: Linking exibility and function. Methods 49(2): 181–188.
51. Fox N, Streinu I (2011) Redundant interactions in protein rigid cluster analysis. In: 1st IEEE International Conference on Computational Advances in Bio and medical Sciences (ECCABS).
52. Djordjevic S, Stock AM (1997) Crystal structure of the chemotaxis receptor methyltransferase cher suggests a conserved structural motif for binding s-adenosylmethionine. Structure 5(4): 434–438.
53. Wang K, Back K, Kim H, Cho Y (1998) The crystal structure of ap endonuclease-1 from methanococcus jannaschii. Nat Struct Biol 5: 707–713.
54. Smith GD, Blessing RH, Ealick SE, Fontecilla-Camps JC, Hauptman HA, et al. (1990) Non-local and non-linear problems in the mechanics of disordered systems: application to granular media and rigid bodies. Phys Rev E 46: 373–419.
55. Ren B, Tibbelin G, de Pascale D, Rossi M, Bartolucci S, et al. (1998) A protein diulcidic oxido-duce from the archean pyrococcus furiosus contains two thioredoxin fold units. Nat Struct Biol 5(7): 602–611.

Ensemble Averaged Descriptions of Protein Rrigidity