JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

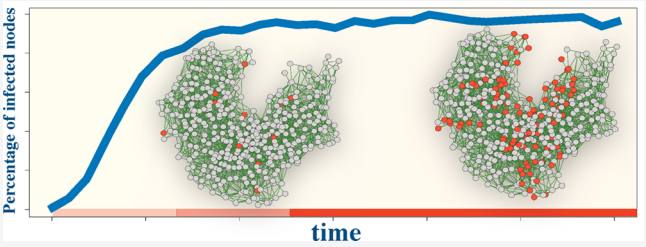
Simulated Epidemics in 3D Protein Structures to Detect Functional Properties

Mattia Miotto,[‖] Lorenzo Di Rienzo,[‖] Pietro Corsi, Giancarlo Ruocco, Domenico Raimondo,* and Edoardo Milanetti*

Cite This: J. Chem. Inf. Model. 2020, 60, 1884−1891 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The outcome of an epidemic is closely related to the network of interactions between individuals. Likewise, protein functions depend on the 3D arrangement of their residues and the underlying energetic interaction network. Borrowing ideas from the theoretical framework that has been developed to address the spreading of real diseases, we study for the first time the diffusion of a fictitious epidemic inside the protein nonbonded interaction network, aiming to study network features and properties. Our approach allows us to probe the overall stability and the capability of propagating information in complex 3D structures, proving to be very efficient in addressing different problems, from the assessment of thermal stability to the identification of functional sites.



INTRODUCTION

Proteins are large biomolecules responsible for the majority of life-sustaining tasks in cells.[1,2] Their great versatility is due to the complex three-dimensional structure they can acquire, which arises as a result of physical and chemical interactions among all of the constituent amino acids. In particular, the global structure is uniquely defined once the sequence of amino acids composing the molecule is specified,[3] with different sequences that can give, up to local rearrangements, the same overall 3D architecture.[4,5]

The peculiar structural conformation each protein assumes is the result of a long evolutionary optimization.[6] Proteins are adapted to carry on specific tasks, usually binding to other molecules while being embedded in a complex dynamical environment in the presence of thermal noise. In this scenario what evolution does is to select sequences that allow proteins to exert their tasks more efficiently in the environment in which they live while maintaining the same overall 3D architecture.[7,8]

Understanding which changes in the amino acid sequence can improve protein efficiency while preserving the biological function has both theoretical and practical implications. Many works investigated the role of different amino acids in the protein structure, folding, stability, and dynamics.[9] In this respect, methods based on graph theory approaches have contributed considerably to the understanding of protein structural flexibility, their hierarchy of structures, and in the identification of key residues.[10−14] All those findings demonstrated that a network-based analysis can be pivotal to shedding light on the complex aspects relative to the organization of protein structures.[15] However, network approaches have often focused on a static description of the system while interesting properties, especially at the level of the single residue, are related to the dynamical behavior of the network.[16]

Theoretical epidemic modeling indeed is a typical approach to study the dynamical behavior of an interaction network, describing the evolution of a contagion process across a population.[17,18]

In the last decades, epidemic models have seen applications in several fields[19,20] thanks to the growth of network sciences. From the spread of real diseases to the diffusion of news in social networks, epidemic models give a measure about the diffusion of information within either the whole network or from a particular node to any other.

Here, we combine for the first time a graph-based schematization of proteins with an epidemic diffusion approach to study the overall stability and the capability to propagate perturbations (or information) in their complex 3D structures.[21,22] In particular, our novel approach proved to be very efficient in characterizing thermal stability and functional sites of proteins.

METHODS

Data Sets. To investigate the capability of a diffusion protocol to grasp the essential feature of the protein structure

Received: November 6, 2019
Published: February 3, 2020

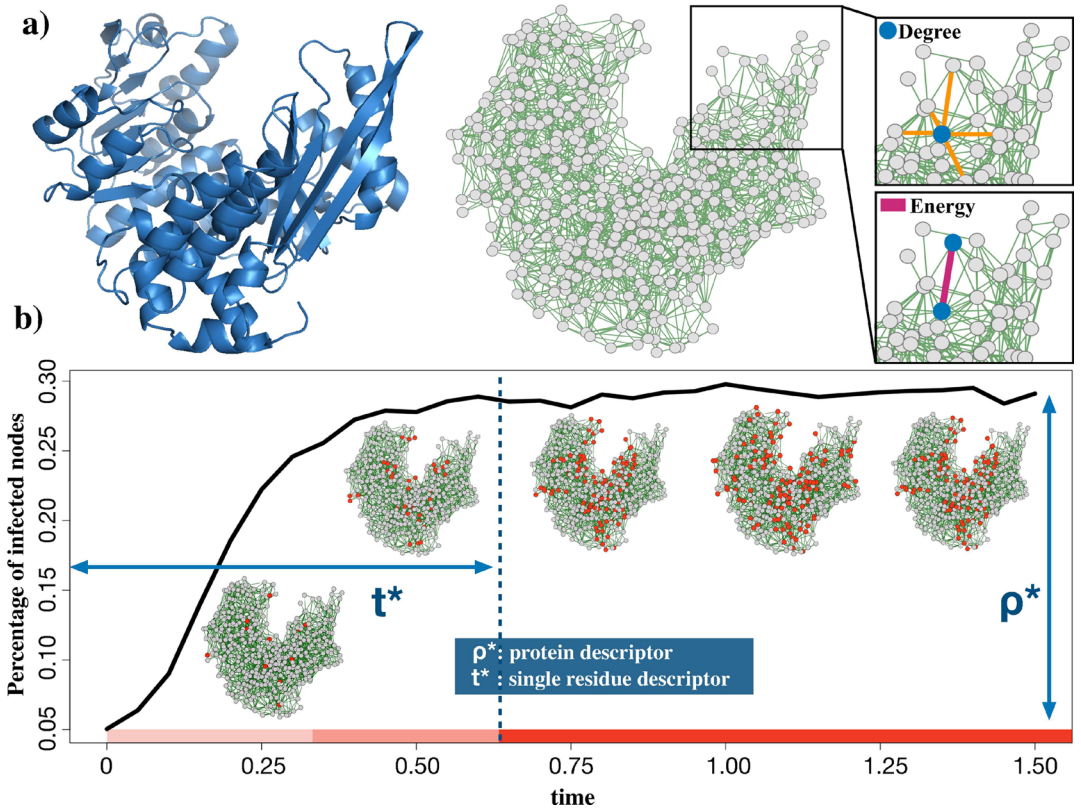


Figure 1. Scheme of the diffusion procedure. (a) Representation of the glutamate dehydrogenase (PDB id: 1HRD) protein structure as ribbons (left) and as a residue interaction network (RIN). Protein residues are considered as nodes and the nonbonded energetic interactions between residues constitute the links between nodes. (b) Outcomes of an epidemic diffusion using interaction energy between residues and node degree as a proxy of infection and recover rates, respectively (as displayed in panel a). Two parameters can be defined: the density of infected nodes at the stationary state, ρ^*, and the time necessary to reach the equilibrium value, t^*. The red nodes represent infected residues at a different time of the epidemic time evolution.

and function, we defined four different data sets: "Thermal data set", "Enzyme data set", "Allosteric data set", and "HIV data set". Details regarding their collection are provided below.

- **Thermal data set.** A set of 32 pairs of homologous proteins with different thermal properties was manually collected from the literature.[23−26] Further information is available in Table S1 of the Supporting Information. Experimentally determined structures were collected from the PDB[27] and filtered according to method (X-ray diffraction), resolution (below 3 Å), and percentage of missing residues (covering more than 95% of Uniprot[28] sequence). Furthermore, in order to focus on the protein's own thermal properties, we excluded all proteins for which experimentally determined structures were only available in a bound state, i.e. in complex with either a ligand or an ion, since the binding is known to alter the thermal stability properties of proteins.[29]
- **Enzyme data set.** It was composed grouping all the enzymes present among the proteins of the Thermal data set. For each enzyme, we retrieved information about the catalytic residues (see Table S1 in the Supporting Information), from the Enzyme Portal of EBI (https://www.ebi.ac.uk/enzymeportal/).[30]
- **Allosteric data set.** We collected from ref 31 proteins whose active and allosteric sites are both known.
- **HIV data set.** It was composed by 18 crystallographic structures of HIV1 and HIV2 proteases (called PR1 and PR2, respectively), both in unbound forms and complexed with 16 different ligands (taken from ref 32). In particular, the HIV data set included: 2 apo dimers (1 of HIV1 and 1 of HIV2) and 16 holo dimers (8 of HIV1 and 8 of HIV2) complexed with 8 different ligands (Table S2 of the SI). For each ligand, we have at least one structure of both PR1 and PR2.

For all data sets, protein structures were minimized using the standard NAMD[33] algorithm and the CHARMM force field[34] in vacuum. A 1 fs time step was used and structures were allowed to thermalize for 10 000 time steps. This procedure aims at removing steric clashes that may be present due to the limitations of the crystallographic refinement.

Network Representation. Protein structures are represented as residue interaction networks[35] (RINs), where each node represents a single residue aa_i. The nearest atomic distance between a given pair of residues aa_i and aa_j is defined as D_{ij}. Two RIN nodes are linked together if $D_{ij} \leq 12$ Å.[33,34] Thus, in our modellization, the location in the 3D space of the node is not defined since in the evaluation of each residue−residue interaction we deal with all the atomic positions of the amino acids.

Furthermore, links are weighted by the sum of two energetic terms: Coulomb (C) and Lennard-Jones (LJ) potentials. The C contribution between the lth atom (a_l) of residue aa_i and the mth atom (a_m) of residue aa_j is calculated as

$$E_{lm}^{C} = \frac{1}{4\pi\epsilon_0}\frac{q_l q_m}{r_{lm}} \quad (1)$$

https://dx.doi.org/10.1021/acs.jcim.9b01027
J. Chem. Inf. Model. 2020, 60, 1884−1891

where q_l and q_m are the partial charges for atoms a_l and a_m, as obtained from the CHARMM force-field; r_{lm} is the distance between the two atoms, and ϵ_0 is the vacuum permittivity. The Lennard-Jones potential is instead given by

$$E_{lm}^{LJ} = \sqrt{\epsilon_l \epsilon_m}\left[\left(\frac{R_{\min}^{l} + R_{\min}^{m}}{r_{lm}}\right)^{12} - 2\left(\frac{R_{\min}^{l} + R_{\min}^{m}}{r_{lm}}\right)^{6}\right] \quad (2)$$

where ϵ_l and ϵ_m are the depths of the potential wells of atom l and m respectively, $R_{\min}^{l}$ and $R_{\min}^{m}$ are the distances at which the potentials reach their minima. Therefore, the weight of the link connecting residues aa_i and aa_j is calculated by summing the contribution of the single atom pairs as

$$E_{ij} = \left[\sum_{l}^{N_i}\sum_{m}^{N_j}(E_{lm}^{C} + E_{lm}^{LJ})\right] \quad (3)$$

where N_i and N_j are the numbers of atoms of the ith and jth residue, respectively.

Diffusion Model on the Protein Network. Epidemic modeling describes the dynamical evolution of the contagion process within a population. An individual (or node) is defined susceptible (S) when it is healthy but could contract the disease, infected (I) when the contagion is transmitted by an adjacent node, and recovered (R) when it manages to recover from the disease. In principle, recovered individuals are immunized and hence they are safe from other infections for a certain time. To study the evolution of the density of infected individuals we have to define the basic processes that rule the transition of individuals states, e.g.

$$\begin{cases} \text{susceptible to infected} & (S \rightarrow I) \\ \text{infected to recovered} & (I \rightarrow R) \\ \text{recovered to susceptible} & (R \rightarrow S) \end{cases} \quad (4)$$

More in details, we must specify (i) the topology of the interaction network, i.e. which nodes directly interact with each other; (ii) the strength of the interaction which is linked to the transmission rate of the infection; and (iii) the recovering rate (i.e. the probability, if present, of returning healthy after having contracted the infection). Depending on the choices one makes for the set of transitions in eq 4, different models and processes can be simulated. A detailed description of the most studied models in classical epidemiology is given in ref 36.

In the present work, we simulated an epidemic diffusion over the protein RINs (see Figure 1). While we preserved the full topological information, we restrained to susceptible−infected−susceptible (SIS) epidemic model, where each node (i.e. residue), once infected, can transmit the infection to near neighbor nodes in the network. A residue can recover from the infection, returning to the susceptible state (meaning that the transition $R \rightarrow S$ is instantaneous). From a biological point of view, the diffusion of the infection inside the network could mimic the effects of a perturbation of some sort taking place on some node/residue. The origins of such a perturbation can be as different as the binding of a ligand, a point mutation, or some post-translational modifications. All those events are expected to produce a local alteration of the state of the residue (the start of the infection) whose effects should propagate along the protein/network. The perturbation propagates differently, depending on the interactions the nearby residues (the susceptible nodes) have with the perturbed one.

In this scenario, the probability of finding a node i in the infected state is given by

$$\frac{dp^i(t)}{dt} = -\delta_i p^i(t) + [1 - p^i(t)]\sum_{j=1}^{N}\beta_{ij}p^j(t) \quad (5)$$

where δ^i is the rate with which node i recovers from infection, while β_{ij} represents the infection rate of node i given that node j is infected at time t.[37]

Equation 5 can not be solved analytically for complex topologies like the RIN ones, but a numerical treatment is required. In the Supporting Information, we provide both a short treatment of the mean-field approximation (where instead it is possible to analytically solve eq 5 for different choices of the transitions in eq 4) and an analytical in-depth on the long-time limit of eq 5. In our case, for each RIN node we identify the recovering rate, δ_i, with the node degree (i.e., the number of connections that a node has with other nodes). While the infection rate between node i and j, β_{ij}, is given by the weight of the link ($\beta_{ij} = E_{ij}$ as given by eq 3). Once defined the infection and recovering rates, we simulated the diffusion process into the protein 3D structure, starting from a specific set of residues or by picking an initially random set, and looking at the mean density of infected residues over time:

$$\rho(t) = \left\langle \frac{N_I(t)}{N_{\text{tot}}} \right\rangle \quad (6)$$

where $N_I(t)$ is the number of infected residue at time t, N_{tot} is the total number of protein residues, and the symbols $\langle\cdot\rangle$ indicate the mean over the M realizations of the diffusion process (all results presented here are obtained setting M = 1000 in order to avoid large fluctuations that are unavoidable in a single realization). It has been found that, depending on the connectivity matrix architecture and the sets of $\{\delta_i\}$ and $\{\beta_{ij}\}$ parameters, the system can exhibit different behaviors. As $t \rightarrow \infty$, the infection, starting from some nodes, propagates in the whole network and reaches a stationary regime where a certain density ρ^* of nodes is constantly infected at each time, independent from the size and the identity of the initial set of infected nodes. Intuitively, $\rho^* = 0$ if the number of nodes that recover from the infection overcomes those that become infected. On the other hand, $\rho^* = 1$ when the infection is too aggressive. The nontrivial scenario ($0 < \rho^* < 1$) is achieved when the network architecture and the parameters allow having a balance between the number of nodes that become infected and the ones that recover. We defined the transient time t^* as the time after which $\rho(t^*) = \rho^* - \delta$, with $\delta \rightarrow 0$ (see Figure 1b); in other words, t^* is the time needed by the epidemic to reach its stationary state. In particular, one can define a node specific descriptor, t_i^*, which is the time required for the infection to reach the stationary state, when the infection starts from the i-esime node. Moveover, since the epidemic originating from a single node usually is characterized by a fast extinction, for each residue, we selected also its two closest neighbors in sequence as a contagion starting point.

Statistical analysis was performed by using R package stats.[38] In particular, clustering analysis performed on the HIV data set was made using the HeatMap function, applying the Euclidean distance matrix (given by the "dist" function) and the "hclust" method for the clustering algorithm.

https://dx.doi.org/10.1021/acs.jcim.9b01027
J. Chem. Inf. Model. 2020, 60, 1884−1891

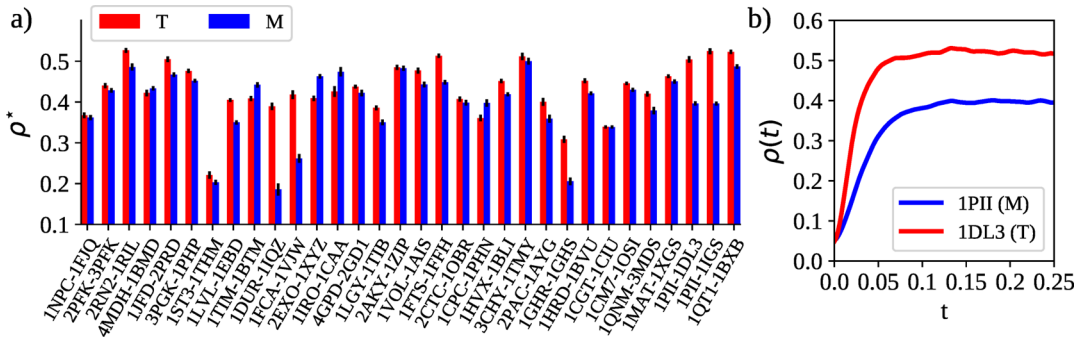


Figure 2. Stationary density of infected nodes which gives information on the thermal stability of the protein. (a) Bar-plot representation of the density of an infected node at the stationary state (ρ^*) for the 32 mesostable (blue) and thermostable (red) proteins of the Thermal data set. (b) Mean density of infected nodes (ρ) as a function of time (t) for an explicative homologous couple, the phosphoribosylanthranilate isomerase (PRAI) protein (PDB id: 1PII-1DL3).

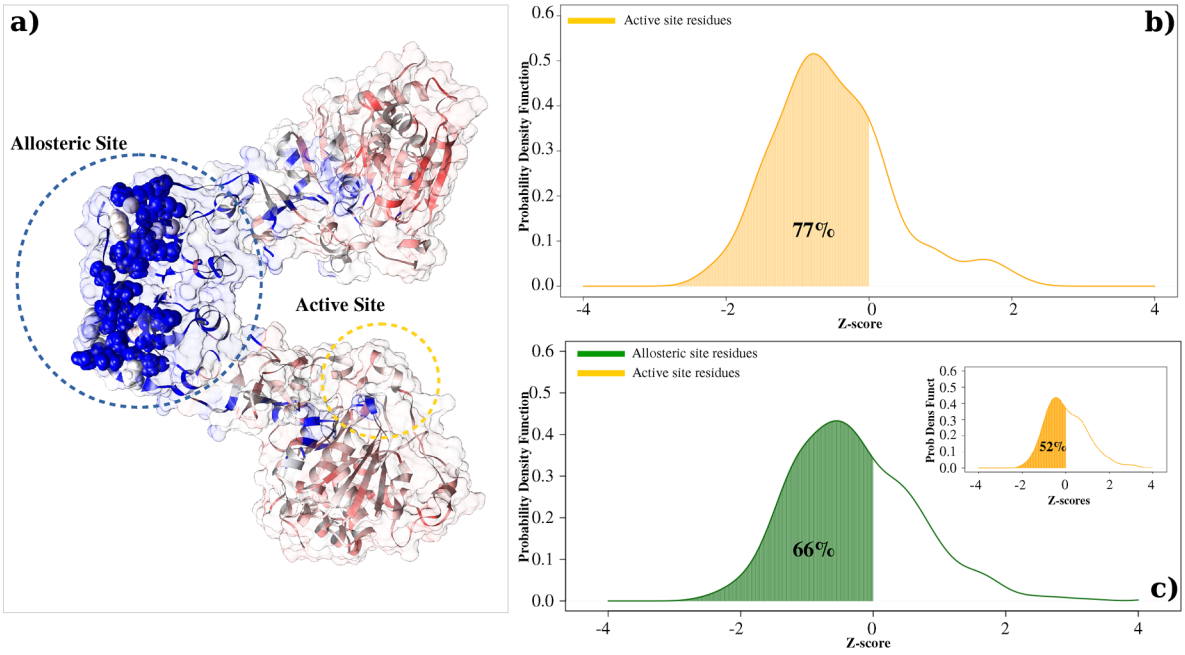


Figure 3. Epidemics starting from protein functional sites spread faster. (a) Ribbon representation of aspartate transcarbamoylase (ATCase) protein (PDB id: 1D09) and colored according to residue t_i^* values, from red (highest values) to blue (lowest values). Allosteric site residues (reported as spheres) present the lowest values of t_i^* parameters. (b) Distribution of the Z-score values regarding the subpopulation of amino acids belonging to the active sites of the 24 enzymes in the Enzyme data set. (c) Distribution of the Z-score values regarding the subpopulations of active sites (orange) and allosteric sites (green) amino acids of the 20 proteins in the Allosteric data set.

RESULTS

Stationary Epidemic Behavior as a Global Measure of Protein Thermal Stability. Different thermal behaviors in homologous proteins have long been studied and several features have been identified as responsible for those differences (such as salt bridges, charged amino acids disposition, etc.[39−45]).

These features are very well-defined in network representation, both in terms of network topology (structure) and link weights (energy). Here we exploit our epidemic-diffusion algorithm to assess the capability of the network to reflect the protein thermostability.

In particular, we compared the stationary state density of infected nodes between all the couples of the Thermal data set, which is composed of 32 pairs of thermophilic-mesophilic homologous proteins. For each protein, the diffusion was simulated, starting each time from a randomly selected set of infected residues. In particular, 5% of the nodes were infected at $t = 0$.

In 84% of cases (27 out of 32 comparisons), thermophilic proteins acquired a higher density of infected nodes with respect to their mesophilic counterparts, when epidemic diffusion reaches the equilibrium (Figure 2a). According to us, this result reflects both the overall higher connectivity and the higher energy of the links in the thermophilic proteins compared to the mesophilic ones. In Figure 2b, we reported an example of diffusion process results where the different steady states are very well visible (PDB id: 1PII-1DL3).

Epidemic Transient Phase Permitting Local Characterization of Protein Structures. After demonstrating that we can properly apply the epidemic diffusion approach exploring global features of a three-dimensional structure of a protein, we investigated our diffusion approach at a single residue level. i.e. we tested if residues that functionally need to

https://dx.doi.org/10.1021/acs.jcim.9b01027
J. Chem. Inf. Model. 2020, 60, 1884−1891

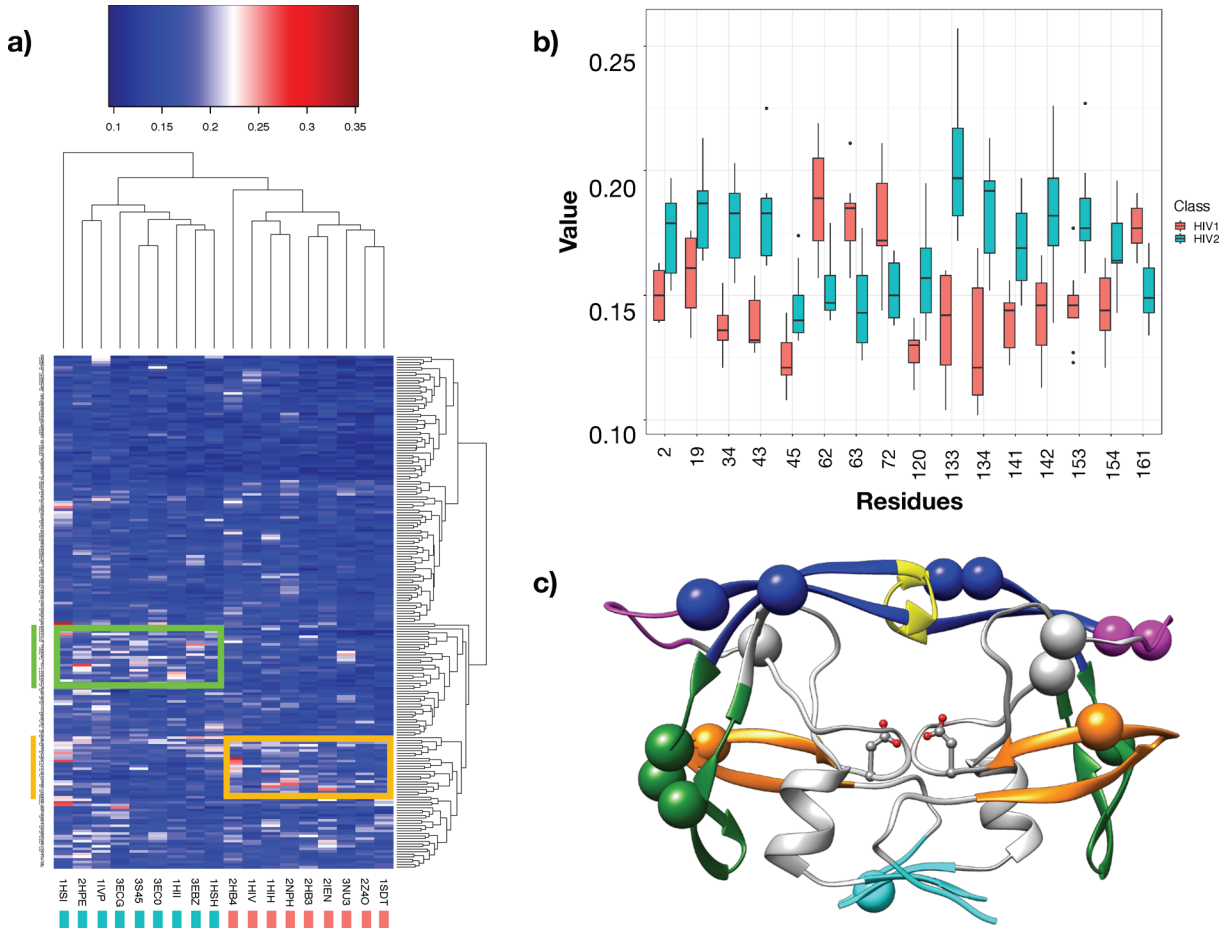


Figure 4. Epidemic diffusion profile discriminates different HIV proteases. (a) Heatmap representation of the clustering analysis performed on the Z_i scores of the proteases of the HIV data set. (b) Boxplot of the distributions of the Z_i scores of the seven most different HIV residues. (c) Dimeric HIV protease (PDB id: 3EBZ), displayed as cartoons and colored according to the functional and structural regions defined by Harte et al.,[46] illustrating the 16 positions at which the HIV-1 protease set differs from the HIV-2 protease one. All amino acid residues most responsible for the differences are shown as spheres, and the color indicates distinct regions. Flaps (blue) residues 43−58 and 143−158; flap tips (yellow) residues 49−52 and 149−152; flap elbow (magenta) residues 37−42 and 137−142; cantilever (green) residues 59−75 and 159−175; fulcrum (orange) residues 10−23 and 110−123; and dimer interface (cyan) residues 1−4, 96−99, 101−104, and 196−199. The catalytic Asp dyads (D25/D125) are displayed using ball-and-stick representation.

have strong communication with the rest of the protein are characterized by peculiar diffusive properties. In this framework, one of the most important challenges in computational biology is the characterization of the active and allosteric sites in proteins. Since the substrate-binding has to be detected also far from the binding region through a cascade of residue−residue interactions,[31] we hypothesize that the diffusive approach could be a perfect approach to capture this aspect.

So we simulated an epidemic originating from every single residue of all proteins in the Enzyme data set and calculated the t_i^* descriptor for each node (see Methods for details). We found that the time t_i^* varies according to some features of the infected initial nodes. In particular, if the epidemic starts from energetically interconnected residues, it is very likely that the stationary state will be obtained in a shorter time when compared with the other residues. In order to compare proteins having sequence of different lengths, we normalized results over each protein size by using the Z-score. The Z-score of the ith residue was defined as

$$Z_i = \frac{t_i^* - \overline{t^*}}{\sqrt{\overline{(t^*)^2} - (\overline{t^*})^2}} \tag{7}$$

where the overbar represents the mean over all the amino acids in the analyzed protein.

By construction, the Z-scores of a population will be distributed with a Gaussian distribution of mean 0 and standard deviation 1, therefore with 50% of the values lower than 0 and 50% higher than 0. If a certain class of residues, i.e. a subpopulation, is characterized by lower (higher) values than 0 we can conclude that the subpopulation has lower (higher) t_i^* with respect to the other residues.

Charged residues exposed on the protein surface and core residues are obviously very fast in propagating the infection, because of the high energy interactions the charged residues are involved in and because of the high number of contacts the core residues have. We preliminarily confirmed this (see the Supporting Information) as shown in Figure S1. We also correlated protein secondary structure location (as calculated by STRIDE[47]) of each residue with its t_i^* because we can suppose that, on average, residues belonging to secondary structures should be assembled in a dense part of the

https://dx.doi.org/10.1021/acs.jcim.9b01027
J. Chem. Inf. Model. 2020, 60, 1884−1891

interaction network. Indeed, α helix components are characterized by lower t_i^* values (see the Supporting Information).

We then proceeded to apply the diffusion protocol in order to analyze t_i^* values for the 22 enzymes in the Enzyme data set for which we know residues forming the active site. The comparison between the t_i^* values of the active site with that of the other residues clearly shows that the former is characterized by a statistically significant lower t_i^* values. The "functional information" the protein receives after ligand binding into the active site needs to be quickly communicated to the whole protein, and our diffusive method is able to well characterize this important biological aspect. In Figure 3b, where we report the distribution of active sites Z-score values, we can notice that the 77% of residues belonging to active sites present a Z-score lower than 0, represented by the orange area under the curve. This indicates stronger connectivity of active site residues with the whole protein than the average value of all other residues.

Then, we have considered the 20 allosteric proteins with known active and allosteric site residues (see Allosteric data set in Methods). Even in this case, we applied our epidemic protocol in order to evaluate the time necessary to reach equilibrium. As shown in Figure 3c, 66% of allosteric residues shows a Z-score value lower than zero, demonstrating that allosteric residues are faster than average residues in propagating information inside the protein network due to their biological functional role in the 3D structure. Interestingly, the active sites of these proteins are not characterized by peculiar diffusive characteristics because just 52% active sites reach Z-score values lower than 0. The reason for this behavior, different from what observed before, could be due to their different binding "state": in the Enzyme data set, proteins are in the apo form, while in the Allosteric data set the proteins are in the holo form, with the ligand occupying the active site. The diffusive approach could be sensitive to these two states.

Discrimination of HIV-1 and HIV-2 Proteases by Their Epidemic Diffusion Profiles. Finally, we used epidemic diffusion time analysis in order to discriminate between HIV-1 and HIV-2 proteases. Despite the structural similarities, HIV-1 and HIV-2 proteases show dramatic disparities in susceptibility to HIV-1 protease inhibitors.[32] We have represented 18 HIV proteases (HIV data set) by a diffusion time profile (e.g., the concatenation of single residue t_i^*), and all profiles have been easily compared since all protein sequences have the same length.

The heatmap reported in Figure 4a shows clustering analysis results performed on residues and proteins of the HIV data set. All of the proteins are correctly identified as HIV-1 or HIV-2 proteases demonstrating the possibility of using an epidemic diffusion approach in order to evaluate functional differences related to three-dimensional protein structures.

We then explored key residues responsible for discrimination of the two groups. For each residue of HIV-1 and HIV-2 proteases, we compared their Z_i scores distributions with a t test. The 44 residues showing the most significant difference, when we set a p-value threshold of 0.05 are 2, 13, 16, 17, 19, 25, 28, 29, 34, 35, 36, 40, 42, 43, 45, 60, 62, 63, 68, 71, 72, 88, 91, 102, 118, 120, 124, 126, 132, 133, 134, 137, 139, 140, 141, 142, 144, 153, 154, 161, 167, 172, 182, 187.

Notably, lowering the p-value threshold to 0.005, we identified a subgroup of 16 residues, i.e. 2, 19, 34, 43, 45, 62, 63, 72, 120, 133, 134, 141, 142, 153, 154, 161, shown in Figure 4b. As we said before, although HIV-1 and HIV-2 proteases share a great deal of structural similarity, the reasons for intrinsic protease inhibitor resistance in HIV-2 are not known. Very interestingly, the subgroup of 16 residues we identified occur at sites distant from the active site but mainly (13 out of 16) in functionally and structurally very relevant regions (see Figure 4c).

This leads us to present a hypothesis that perhaps our epidemic approach could have been captured subtle structural changes, imparted by a limited number of residues, causing dramatic functional differences between homologous proteins (HIV-1 and HIV-2 proteases). That is the 16 amino acids outside the HIV-2 protease active site may cause subtle changes in conformation and in long-range effects compared to HIV-1, which might impact protease inhibitor binding affinity.

DISCUSSION

Proteins are complex systems. So evolution must be very proficient in tuning parameters (e.g., selecting mutations) to obtain more fitted proteins with respect to some features while maintaining the protein as functional. For instance, optimizing enzymes to be more efficient at high temperatures (i.e., increasing their thermal stability) must not reduce enzyme flexibility and the ability to change configurations.

Graph theory-based methods represent a powerful approach to investigate protein topological and energetic properties. However, we could consider it a static view of the protein structures that does not allow us to describe their complexity in a complete way. To overcome this limitation several aspects of proteins were investigated through dynamical approaches, like molecular dynamics or perturbation-response approaches, which take into account the dynamical properties. A problem connected with these approaches is that they are typically characterized by a high computational cost.

In this work, we explored for the first time the possibility to adopt an epidemic diffusion-based method as an efficient way to study functional aspects (both local and global) of proteins strictly connected with their three-dimensional structural organization. The new idea we introduced with this approach was to investigate the dynamic properties of the interactions network of a protein structure by using epidemic diffusion-based algorithms, preserving in this way both the topology and the energy properties of the interactions. The most striking advantage of this method is that it is not very computationally expensive allowing for a fast exploration of complex problems related to protein function (diffusion on an average protein requires few minutes on a standard personal computer). Starting from the RIN formalism,[48,49] we studied the diffusion of a fictitious epidemic inside the protein structure represented as a network using energies and node degrees as proxies of infection and recovery rates.

A large number of mathematical models have been formulated to study the spread of infectious diseases, but most of these are just variants of the Kermack and McKendrick epidemic model.[17,18] Reproducing different aspects of the spread of real diseases, all models ultimately provide a measure of the information diffusion throughout the entire network.

Simulations of diffusion processes were performed considering typical network parameters for calculating the probability of transmission of infection (proportional to the link energy) and the probability of each node of returning susceptible (proportional to node degree).

https://dx.doi.org/10.1021/acs.jcim.9b01027
J. Chem. Inf. Model. 2020, 60, 1884−1891

From diffusion simulations, two descriptors were defined, one (ρ^*) providing global information and the other (t^*) local one. In particular, a residue-specific descriptor is of fundamental importance because the identification of functional key residues in a protein structure is a useful aspect for protein design in many open biological questions.

Considering the stationary phase, the mean of the percentage of infected nodes is constant over the steps balancing the rate of infection and recovery. The value of the stationary percentage of infected nodes is a very compact way to quantify the global properties of the entire protein related to residue−residue energetic interactions. A protein characterized by strong interconnectivity will have a very strong energetic coupling between its residues showing, at the equilibrium, a higher number of infected nodes.

Given an overall fold, the arrangement of side chains organizes the intermolecular interaction to better resist the thermal noise. Therefore, we test the sensibility of this formalism applying it on a well-defined set of homologous protein pairs, one protein from a mesostable organism, the other one from a thermostable one.

We found that thermophilic proteins have a significantly higher percentage of infected residues than homologous mesophilic counterparts, meaning that thermophilic proteins organize their network of interactions in order to promote infection. We could, therefore, conclude that thermophilic proteins have, on average, a higher level of interconnectivity than mesophilic proteins.

Another important aspect we explored in this work was the local properties of proteins that often are generated by long-range effects. In this case, the problem was studied by taking into account the transient phase of the diffusion simulation, which is composed of steps between the initial infection and the stationary state.

The time necessary in order to reach the stationary phase, t^*, is depending on the choice of the starting infected nodes, expressing their centrality in the energy network. This local characterization can be utilized in order to identify which kind of residues (or domains) are more central in a protein, in terms of their connection with the rest of the protein. We clearly demonstrated that both residues belonging to enzyme allosteric and active sites typically reach the state of equilibrium with a smaller time than any other residue.

We also investigated the local property of each residue of two HIV-1 and HIV-2 proteases. The method showed its perfect ability to separate the two classes of proteases, in terms of transient phase, elucidating nontrivial differences (HIV-1 and HIV-2 protease share 50% of sequence identity) by analyzing the dynamic properties of residues represented as a network. The epidemic approach was also able to select 16 residues responsible for the discrimination of the two groups, which might impact protease inhibitor binding affinity helping to understand the key differences between HIV-1 and HIV-2 infections. Furtermore, 13 out of the 16 residues identified by our approach can be mapped on very important functional and structural regions of the HIV protease. We believe that the study of the dynamical aspect of the protein structure network is, in general, a promising direction for the future. We intend to investigate how far our results can be generalized to other types of protein functional elements, other types of proteins like membrane proteins, and hopefully to other kinds of macromolecules like nucleic acids.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b01027.

More detailed explanation of epidemics models together with more analyses and information about the collected data sets (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Domenico Raimondo − *Department of Molecular Medicine, Sapienza University, Rome 00161, Italy*;
Email: domenico.raimondo@uniroma1.it

Edoardo Milanetti − *Department of Physics, Sapienza University, Rome 00185, Italy; Center for Life Nanoscience, Istituto Italiano di Tecnologia, Rome 00161, Italy*;
orcid.org/0000-0002-3046-5170;
Email: edoardo.milanetti@uniroma1.it

Authors

Mattia Miotto − *Department of Physics, Sapienza University, Rome 00185, Italy; Center for Life Nanoscience, Istituto Italiano di Tecnologia, Rome 00161, Italy*

Lorenzo Di Rienzo − *Department of Physics, Sapienza University, Rome 00185, Italy*

Pietro Corsi − *Department of Science, Roma Tre University, Rome 00154, Italy*

Giancarlo Ruocco − *Department of Physics, Sapienza University, Rome 00185, Italy; Center for Life Nanoscience, Istituto Italiano di Tecnologia, Rome 00161, Italy*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.9b01027

Author Contributions

‖These authors contributed equally to the present work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Vittorio Padovano for his help in the early stages of this project.

■ REFERENCES

(1) Mannige, R. Dynamic New World: Refining Our View of Protein Structure, Function and Evolution. *Proteomes* **2014**, *2*, 128−153.

(2) Chothia, C.; Hubbard, T.; Brenner, S.; Barns, H.; Murzin, A. Protein Folds In The All-β And All-α Classes. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 597−627.

(3) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. The Protein Folding Problem. *Annu. Rev. Biophys.* **2008**, *37*, 289−316.

(4) Lesk, A. M.; Chothia, C. How Different Amino Acid Sequences Determine Similar Protein Structures: The Structure and Evolutionary Dynamics of the Globins. *J. Mol. Biol.* **1980**, *136*, 225−270.

(5) Ofran, Y.; Margalit, H. Proteins of the Same Fold and Unrelated Sequences Have Similar Amino Acid Composition. *Proteins: Struct., Funct., Genet.* **2006**, *64*, 275−279.

(6) Debès, C.; Wang, M.; Caetano-Anollés, G.; Gräter, F. Evolutionary Optimization of Protein Folding. *PLoS Comput. Biol.* **2013**, *9*, e1002861.

(7) Domingues, F. S.; Koppensteiner, W. A.; Sippl, M. J. The Role of Protein Structure in Genomics. *FEBS Lett.* **2000**, *476*, 98−102.

(8) Karshikoff, A.; Nilsson, L.; Ladenstein, R. Rigidity versus Flexibility: the Dilemma of Understanding Protein Thermal Stability. *FEBS J.* **2015**, *282*, 3899−3917.
(9) Chakrabarty, B.; Parekh, N. NAPS: Network Analysis of Protein Structures. *Nucleic Acids Res.* **2016**, *44*, W375−W382.
(10) Dokholyan, N. V.; Li, L.; Ding, F.; Shakhnovich, E. I. Topological Determinants of Protein Folding. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 8637−8641.
(11) del Sol, A.; Fujihashi, H.; Amoros, D.; Nussinov, R. Residues Crucial for Maintaining Short Paths in Network Communication Mediate Signaling in Proteins. *Mol. Syst. Biol.* **2006**, DOI: 10.1038/msb4100063.
(12) Amitai, G.; Shemesh, A.; Sitbon, E.; Shklar, M.; Netanely, D.; Venger, I.; Pietrokovski, S. Network Analysis of Protein Structures Identifies Functional Residues. *J. Mol. Biol.* **2004**, *344*, 1135−1146.
(13) Vendruscolo, M.; Dokholyan, N. V.; Paci, E.; Karplus, M. Small-world View of the Amino Acids That Play a Key Role in Protein Folding. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2002**, *65*, 061910.
(14) Aftabuddin, M.; Kundu, S. Hydrophobic, Hydrophilic, and Charged Amino Acid Networks within Protein. *Biophys. J.* **2007**, *93*, 225−231.
(15) Miotto, M.; Olimpieri, P. P.; Di Rienzo, L.; Ambrosetti, F.; Corsi, P.; Lepore, R.; Tartaglia, G. G.; Milanetti, E. Insights on Protein Thermal Stability: a Graph Representation of Molecular Interactions. *Bioinformatics* **2019**, *35*, 2569−2577.
(16) Yang, L.-Q.; Sang, P.; Tao, Y.; Fu, Y.-X.; Zhang, K.-Q.; Xie, Y.-H.; Liu, S.-Q. Protein Dynamics and Motions in Relation to Their Functions: Several Case Studies and the Underlying Mechanisms. *J. Biomol. Struct. Dyn.* **2014**, *32*, 372−393.
(17) Kermack, W. O.; McKendrick, A. G. A Contribution to the Mathematical Theory of Epidemics. *Proc. R. Soc. A* **1927**, *115*, 700−721.
(18) Kermack, W. O.; McKendrick, A. G. Contributions to the Mathematical Theory of Epidemics. II. The Problem of Endemicity. *Proc. R. Soc. A* **1932**, *138*, 55−83.
(19) Vespignani, A. Modelling Dynamical Processes in Complex Socio-technical Systems. *Nat. Phys.* **2012**, *8*, 32−39.
(20) Yang, H.-X.; Wang, B.-H. Immunization of Traffic-driven Epidemic Spreading. *Phys. A (Amsterdam, Neth.)* **2016**, *443*, 86−90.
(21) Castellano, C.; Fortunato, S.; Loreto, V. Statistical Physics of Social Dynamics. *Rev. Mod. Phys.* **2009**, *81*, 591−646.
(22) Albert, R.; Barabási, A.-L. Statistical Mechanics of Complex Networks. *Rev. Mod. Phys.* **2002**, *74*, 47−97.
(23) Brinda, K. V.; Vishveshwara, S. A network Representation of Protein Structures: Implications for Protein Stability. *Biophys. J.* **2005**, *89*, 4159−4170.
(24) Kannan, N.; Vishveshwara, S. Aromatic Clusters: a Determinant of Thermal Stability of Thermophilic Proteins. *Protein Eng., Des. Sel.* **2000**, *13*, 753−761.
(25) Mozo-Villarias, A.; Cedano, J.; Querol, E. A Simple Electrostatic Criterion for Predicting the Thermal Stability of Proteins. *Protein Eng., Des. Sel.* **2003**, *16*, 279−286.
(26) Sterner, R.; Liebl, W. Thermophilic Adaptation of Proteins. *Crit. Rev. Biochem. Mol. Biol.* **2001**, *36*, 39−106.
(27) Touw, W. G.; Baakman, C.; Black, J.; te Beek, T. A.; Krieger, E.; Joosten, R. P.; Vriend, G. A series of PDB-related Databanks for Everyday Needs. *Nucleic Acids Res.* **2015**, *43*, D364−368.
(28) Pundir, S.; Martin, M. J.; O'Donovan, C. UniProt Protein Knowledgebase. *Methods Mol. Biol.* **2017**, *1558*, 41−55.
(29) Celej, M. S.; Montich, G. G.; Fidelio, G. D. Protein Stability Induced by Ligand Binding Correlates with Changes in Protein Flexibility. *Protein Sci.* **2003**, *12*, 1496−1506.
(30) Alcantara, R.; Onwubiko, J.; Cao, H.; Matos, P. d.; Cham, J. A.; Jacobsen, J.; Holliday, G. L.; Fischer, J. D.; Rahman, S. A.; Jassal, B.; Goujon, M.; Rowland, F.; Velankar, S.; Lopez, R.; Overington, J. P.; Kleywegt, G. J.; Hermjakob, H.; O'Donovan, C.; Martin, M. J.; Thornton, J. M.; Steinbeck, C.; et al. The EBI Enzyme Portal. *Nucleic Acids Res.* **2013**, *41*, D773−D780.
(31) Amor, B. R. C.; Schaub, M. T.; Yaliraki, S. N.; Barahona, M. Prediction of Allosteric Sites and Mediating Interactions Through Bond-to-Bond Propensities. *Nat. Commun.* **2016**, DOI: 10.1038/ncomms12477.
(32) Triki, D.; Billot, T.; Visseaux, B.; Descamps, D.; Flatters, D.; Camproux, A.-C.; Regad, L. Exploration of the Effect of Sequence Variations Located Inside the Binding Pocket of HIV-1 and HIV-2 Proteases. *Sci. Rep.* **2018**, *8*, 5789.
(33) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(34) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. *J. Chem. Inf. Model.* **2012**, *52*, 3144−3154.
(35) Grewal, R. K.; Roy, S. Modeling Proteins as Residue Interaction Networks. *Protein Pept. Lett.* **2015**, *22*, 923−933.
(36) Pastor-Satorras, R.; Castellano, C.; Van Mieghem, P.; Vespignani, A. Epidemic Processes in Complex Networks. *Rev. Mod. Phys.* **2015**, *87*, 925−979.
(37) Allen, L. J. Some Discrete-time SI, SIR, and SIS Epidemic Models. *Math. Biosci.* **1994**, *124*, 83−105.
(38) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013; ISBN 3-900051-07-0.
(39) Amadei, A.; Del Galdo, S.; D'Abramo, M. Density Discriminates Between Thermophilic and Mesophilic Proteins. *J. Biomol. Struct. Dyn.* **2018**, *36*, 3265.
(40) Tartaglia, G. G.; Cavalli, A.; Vendruscolo, M. Prediction of Local Structural Stabilities of Proteins from Their Amino Acid Sequences. *Structure* **2007**, *15*, 139−143.
(41) Vishveshwara, S.; Brinda, K. V.; Kannan, N. Protein Structure: Insights from Graph Theory. *J. Theor. Comput. Chem.* **2002**, *01*, 187−211.
(42) Vijayabaskar, M.; Vishveshwara, S. Interaction Energy Based Protein Structure Networks. *Biophys. J.* **2010**, *99*, 3704−3715.
(43) Lee, C. W.; Wang, H. J.; Hwang, J. K.; Tseng, C. P. Protein Thermal Stability Enhancement by Designing Salt Bridges: a Combined Computational and Experimental Study. *PLoS One* **2014**, *9*, e112751.
(44) Folch, B.; Dehouck, Y.; Rooman, M. Thermo- and Mesostabilizing Protein Interactions Identified by Temperature-dependent Statistical Potentials. *Biophys. J.* **2010**, *98*, 667−677.
(45) Folch, B.; Rooman, M.; Dehouck, Y. Thermostability of Salt Bridges versus Hydrophobic Interactions in Proteins Probed by Statistical Potentials. *J. Chem. Inf. Model.* **2008**, *48*, 119−127.
(46) Harte, W. E.; Swaminathan, S.; Mansuri, M. M.; Martin, J. C.; Rosenberg, I. E.; Beveridge, D. L. Domain Communication in the Dynamical Structure of Human Immunodeficiency Virus 1 Protease. *Proc. Natl. Acad. Sci. U. S. A.* **1990**, *87*, 8864−8868.
(47) Frishman, D.; Argos, P. Knowledge-Based Protein Secondary Structure Assignment. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 566−579.
(48) Sengupta, D.; Kundu, S. Role of Long- and Short-Range Hydrophobic, Hydrophilic and Charged Residues Contact Network in Protein's Structural Organization. *BMC Bioinf.* **2012**, DOI: 10.1186/1471-2105-13-142.
(49) Böde, C.; Kovács, I. A.; Szalay, M. S.; Palotai, R.; Korcsmáros, T.; Csermely, P. Network Analysis of Protein Dynamics. *FEBS Lett.* **2007**, *581*, 2776−2782.