Exploring protein flexibility and allosteric signalling mechanism with rigidity theory

Adnan Sljoka[1 2] and Nobuyuki Tsuchimura[1]*
1 Department of Informatics, School of Science and Technology, Kwansei Gakuin University, Osaka, Japan
2 CREST, Japan Science and Technology Agency (JST)
* Corresponding author, email: adnanslj@kwansei.ac.jp

Abstract—Advancements in the field of mathematical rigidity theory have opened up a number of exciting opportunities for computational predictions of protein flexibility and their dynamics. Starting with a 3D protein structure, several programs such as FIRST model the protein as a constraint multigraph, consisting of vertices (atoms) and edges (covalent bonds, hydrogen bonds, electrostatic interactions, and hydrophobic contacts). FIRST applies the pebble game algorithm on the resulting multigraph which rapidly decompose the protein into rigid clusters and flexible regions. Using an extension of FIRST and the pebble game algorithm we propose a computational approach for studying a biological phenomenon 'allostery'. Allostery refers to an effect of binding at one site to another, often significantly distant functional site on the protein, allowing for regulation of the protein function. Most dynamic proteins are allosteric and allostery has even been coined the 'second secret of life', however the molecular mechanisms that give rise to allostery are currently poorly understood. Extending our earlier seminal work, we have developed a rigidity-transmission allostery (RTA) algorithm which predicts if mechanical perturbation of rigidity (mimicking ligand binding) at one site of the protein can propagate across a protein structure and in turn cause a transmission and change in degrees of freedom and conformation at a second distant site, resulting in allosteric transmission. Since RTA algorithm is computationally fast, we can rapidly scan many unknown sites for rigidity-based allosteric communication, identifying potential new allosteric sites and quantify their allosteric effect. We will review the functional importance of protein flexibility and mathematical and algorithmic background of rigidity theory and method FIRST. In this originative expose we describe rigidity based mechanistic allostery communication model. We will also provide a few illustrations of rigidity-based allostery communication on actual protein structures, including the important signaling G-protein coupled receptors. This method will have important consequence in general understanding of allostery and in aid of design of allosteric drugs.

Keywords— *allostery, protein structure, protein flexibility, rigidity theory, pebble game algorithm, molecular theorem, FIRST, rigidity-transmission allostery algorithm.*

I. INTRODUCTION AND ALLOSTERY

The post genomic era has revived the focus on characterizing the protein function and their regulation [1, 26]. Because proteins are fundamental to most cellular function, there is a significant interest to understand how these macromolecules perform various complex tasks. Allosteric regulation of protein function, the concept which was first introduced by Monod and Jacob in 1960s [2], refers to transmission of signals from ligand/substrate binding site(s) to a topographically distinct and remote orthosteric/active site. The binding of a ligand, drug or other protein partner at the allosteric site(s) triggers a local conformational change that can propagate a substantial distance across the protein structure to cause a rearrangement and change in conformation and dynamics at the distant active site. Allostery is integral to the control of metabolic and signalling pathways and it provides organisms the ability to adapt to constant changes in cellular and environmental conditions [1, 3, 5]. Allostery is one of the most powerful and prevalent means of regulating protein activity and has been referred to as 'the second secret of life' [3] second only to the genetic code. It is now becoming widely accepted that all dynamic proteins are allosteric [5]. Many physiological activities are controlled by allostery and allosteric effect has direct impact on disease states [5]. Allostery can both cause disease and contribute in the development of new therapeutics [15].

Remarkably, even after 50 years since the concept of allostery was first introduced, the mechanism of allosteric communication still remains poorly understood. Despite the lack of any major breakthroughs surrounding allostery, in 2004 the US Food and Drug Administration (FDA) approved the first allosterically designed drug and since then many more have been approved. Over the past few years, there were over 1000 papers published on allostery each year, demonstrating its all-encompassing importance in biology and medicine. Since allostery is a crucial biological phenomenon for understanding biological systems, disease and is critical in therapeutics and drug design, decoding the allosteric mechanism remains one of the key long-standing unsolved problems in biological sciences.

With the advancements in high resolution X-ray crystallography and Nuclear Magnetic Resonance (NMR), the static structures of many proteins have been solved, but ultimately protein function is controlled by its dynamic character [26]. One of the great challenges in understanding any protein function and its regulation, including the elusive allosteric mechanism, involves deep knowledge and modeling of protein flexibility and its dynamics. Although high quality static 3-dimensional structures provide some insights, a cherished desire of any protein scientist is to watch proteins move in real time at atomistic level as they perform their functions. Despite many advances in experimental biochemistry and computational molecular dynamic simulations we are still far from realizing such a dream.

978-1-5090-5753-5/16 $31.00 © 2016 IEEE
DOI 10.1109/APWC.on.CSE.2016.45

Recent advancements in the field of rigidity theory [8,24] have opened up a number of exciting opportunities for computational predictions of protein flexibility and their dynamics. In rigidity theory, proteins are modeled as geometric frameworks consisting of atoms and various connecting intermolecular forces. Programs such as FIRST [10] apply the mathematical results of rigid and flexible structure and decompose a protein framework into flexible and rigid regions, and starting with such a decomposition, fast Monte-Carlo methods such as FRODA [21] were developed for simulating the protein motions.

In section II we provide a brief review of the close link between protein flexibility and protein function and current available techniques for measuring protein flexibility. In section III we give a brief description of mathematical rigidity theory and how combinatorial characterizations of flexible and rigid molecular structures can be used to analyze protein flexibility and motions as is implemented in software FIRST and FRODA. We will demonstrate how biologically relevant questions surrounding the effects of mutations on flexibility, which are extremely challenging to probe with experiments and traditional computational approaches, can be tackled with FIRST.

As a main highlight and contribution of this paper we introduce a novel mechanistic description model of allostery that is based on concepts in rigidity theory (section IV), extending our earlier seminal work in mathematical allostery. We will show how our method rigidity-transmission allostery (RTA) algorithm can be used to predict and quantify allosteric interactions in protein structures. Examples of the output of RTA algorithm will be demonstrated on largest receptor class GPCRs and eukaryotic translation initiation factor protein (section V). With RTA algorithm we can identify the allosteric pathways and detect potential novel allosteric sites, which has direct applications in design and detection of allosteric therapeutics. These are all crucial problems in the area of allostery research.

II. PROTEIN FLEXIBILITY AND FUNCTION

Proteins (from the Greek ‘protos’ meaning ‘of primary importance’) are the most versatile macromolecules which perform crucial functions in essentially all biological processes [1]. Proteins function as catalysts, they transport and store other molecules such as oxygen, they provide immune protection, and are also responsible for carrying out important transduction signals in and out of the cells, among many other biologically significant functions [1]. Proteins are polypeptide chains composed of sequences drawn from twenty amino acids, sometimes referred as the building blocks of life, which encodes its 3-dimensional structure. High resolution X-ray crystallography and NMR structure determination experimental techniques have revealed their beautiful structural complexity and in many ways have revolutionized our understanding how proteins function at the atomistic resolution. The experimentally solved 3-dimensional protein structures contain the coordinates of individual atoms and are deposited into the protein data bank (PDB), which can be viewed, modeled and aesthetically appreciated with various molecular visualization software.

The pdb file of a protein contains crucial structural information, but it can also provide a misleading view of protein function as proteins are not best viewed by a single folded structure. Instead, a much richer and complete representation of proteins takes into the account both thermodynamic and kinetic nature of proteins [26]. In other words, proteins are highly dynamic sampling various conformations around the complex multidimensional energy landscape which can undergo conformational dynamics even under conditions that strongly favor a well-defined low energy ‘native’ state [26]. To model and deeply understand protein function and its regulation such as allostery, we must be able to accurately and efficiently predict flexibility of proteins at the atomistic level. Connecting the structural information of a protein with flexibility and rigidity predictions can provide a much richer picture and decode the proteins function and mechanism. With the exceptions of intrinsically flexible proteins (see below on TAU protein in Alzheimers), proteins function at the delicate border of rigidity and flexibility. Most proteins need rigidity to maintain their shape and just the right amount of flexibility to perform their biological functions. A number of diseases are often linked to key proteins adopting non-native ‘faulty’ conformations (i.e. incorrect misfolded shapes) which can ultimately lead to creation of overly rigid and indestructible complexes (i.e. `amyloid and prion diseases' such as Alzheimers, Parkinsons, Mad Cow Disease) [1, 25, 28]. Flexibility and rigidity predictions of proteins is an active area of research in both experimental protein science and computational biology.

Determining flexible and rigid regions in a protein and understanding how they move is a complex task [26]. Typical protein structure have several thousands of atoms and contain thousands of conformational degrees of freedom. The situation can become even more complicated in multiprotein complexes that have quaternary structures, which are a group of two or more interacting polypeptide chains called oligomers. In addition to the large size of proteins and complexes, and high conformational degrees of freedom, the conformational fluctuations can be rapid, transient and result in structures that can be spectroscopically indistinguishable from the ground-state (x-ray crystallography snapshot) [18, 26]. Protein flexibility is largely influenced by the interactions with ligands, drugs, other proteins, mutations and changes in the local biochemical environment (i.e. temperature, ph, etc.). Protein motions are also known to occur on wide range of time scales, ranging from fast short-amplitude motions (i.e. bond vibrations) occurring on femtosecond range, side chain motions on the picosecond to nanosecond timescale all the way up to slow timescales with larger-amplitude collective domain motions which are often important for catalysis, occurring on milliseconds to seconds range [21, 26]. All these factors combine to contribute to the difficulty in obtaining knowledge about proteins flexibility and their motions.

A wide range of experimental data (NMR techniques like order parameter measurements, chemical shifts, hydrogen/deuterium exchange (HDX) data, crystallographic *B*-values etc) can provide some insights into the dynamic nature of proteins [18, 26]. However, these techniques are limited as it is nearly impossible to see individual atoms undergoing dynamics. Experiments are also costly and can take a long time to measure.

There has also been a growing interest to apply and incorporate sophisticated biophysical methods together with mathematical algorithms. Computational simulation methods have provided many critical advancements in modeling and predictions of protein flexibility and dynamics [6, 7, 21]. Molecular dynamics (MD) simulations has long been a traditional computational approach that is used to study molecular and protein motions [7]. The trajectories of molecules or individual atoms are determined by repetitively obtaining a numerical solution of the Newton's classical equations of motions F = ma by forward integration in time. Forces between the molecules and potential energies are approximated by the use of all-atom force fields (energy functions). In an ideal situation and a nearly perfect description of force field, MD simulation can be quite powerful as the precise position of each atom at any instant in time for a single protein molecule can be tracked. However, MD simulations are still largely impractical as it takes a prohibitive amount of computational power to investigate most protein motions, in particular the functionally important larger-amplitude collective motions occurring on the micro to milli-second time-scales. The computational time needed to reach motions on the longer relevant timescales, even with massively parallelized MD runs via costly special-purpose commodity computer clusters such as Anton [17], is beyond practical wide-range application. To overcome the computational limitations combined with the rapid growth in both the number of protein structures deposited into the protein data bank and the increasing size of the newly crystalized structures, there is a pressing need to come up with alternate computational tools that simplify the force fields and yet can still provide accurate and efficient protein flexibility prediction.

One such emerging methodology that has provided numerous advances over the past 15 years in the area of computational predictions of protein flexibility stems from the results inspired from work in structural and mathematical rigidity theory [23]. A fast rigidity-theory method and accompanying protein flexibility analysis software that has gained interest and popularity is the program FIRST [10]. Over the last few years, a number of research groups have developed various spinoffs of these methods.

III. PROTEIN FLEXIBILITY CAN BE EFFICIENTLY ANALYZED WITH RIGIDITY THEORY

A. Combinatorial Rigidity Theory of molecular frameworks

Rigidity theory is the study of rigidity and flexibility of structural frameworks which are specified by geometric constraints such as fixed distances and directions on a collection of points and rigid bodies [24]. Because the rigidity properties of either natural structures (molecules, crystals etc) or engineered structures (bridges, robots etc) is essential to its architecture and function, rigidity theory has many applications in engineering, robotics, material science and biology. Rigidity theory has both geometric and combinatorial characteristics relying on techniques in linear algebra, discrete and algebraic geometry, graph theory and combinatorics. We review the basic results in combinatorial rigidity theory which are the most relevant for analysis of protein flexibility and allostery. For thorough review of rigidity theory see [23, 24].

Two crucial components for computational modeling and prediction of macromolecular flexibility and their motions are: (i) a realistic physical representation of molecules and their interactions and (ii) a mathematically sound theory and algorithms for predicting flexibility.

Proteins are held together with various forces (interactions) of different strengths. A useful modelling assumption is to view such chemical interactions as constraints between atoms. In the rigid geometry assumption of molecules, the angles between the bonds of an atom are fixed and only dihedral angles are allowed to rotate. The locked dihedral (angles) rotations associated with double bonds (i.e. non-rotating bonds) and non-covalent interactions impose additional constraints. With this model in mind, we define a *molecular framework* as a collection of atoms which are treated as fully rigid bodies in 3-space with six trivial degrees of freedom and bonds as hinges, leaving one rotational degree of freedom (dihedral) between the two connecting atoms [24]. In the language of rigidity theory, molecular framework is a special case of general structures known as *body-hinge frameworks*, which consist of rigid bodies connected by revolute hinges. Since each hinge (bond) removes five degrees of freedom between the two bonding atoms, it is useful to model a hinge by a set of five bars connecting the two bonding atoms where each bar removes one degree of freedom. Such structures are called *body-bar frameworks* - a collection or rigid bodies connected by linear bars (Fig. 1 b, c). There is special geometric criteria to be considered when selecting the five bars for a molecular bond as bonds are not generic hinges but this is not important for this paper (see [24] for details). Double or peptide bonds are modelled as a set of 6 bars between the two atoms which locks the rotational degree of freedom. Non-covalent interactions, which are critical for overall protein structure and function are also modeled as a set of bars that further restrict the proteins internal conformational degrees of freedom. Disulphide bonds and hydrogen bonds are typically modelled with 5 bars. Depending on the energy strength and persistence of a hydrogen bond, the number of bars can be adjusted between one and five [18]. Hydrophobic contacts are modelled with 2 bars between any close contacting pairs of carbon-carbon, carbon-sulfer or sulfer-sulfer atoms. Combining covalent and non-covalent interactions, this defines the overall body-bar framework model of a protein (Fig. 1 e).

A body-bar framework is rigid if every motion results in framework that is isometric to the original one (i.e. the framework only has rigid-body motions), otherwise it is

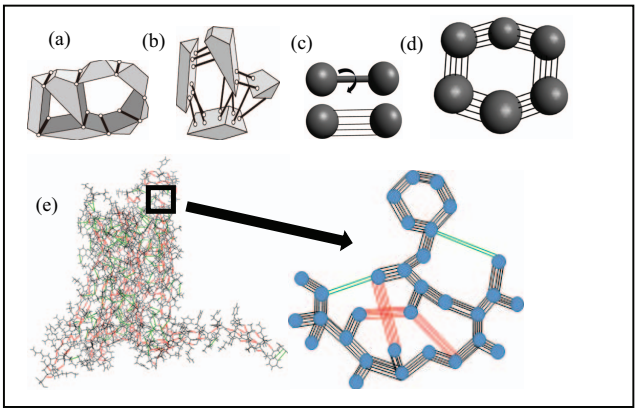

Fig. 1. *(a) Body-hinge structure in 3D composed of rigid bodies connected by hinges (lines). (b) Body-bar framework (c) Molecular framework of a diatomic molecule consisting of two atoms and a single bond can be viewed as a body-hinge structure where atoms are bodies and hinges are bonds. In a diatomic molecule, two atoms rotating about a single bond can be modelled as a generic body-bar structure (two bodies connected with 5 bars). In a diatomic.molecule, there are a total of seven degrees of freedom (6×2 − 5) where each bar has removed a degree of freedom, with remaining ever-present 6 degrees of freedom of trivial rigid body motions and one additional internal degree of freedom corresponding to a rotation around the bond. (d) Generic cyclohexane molecule viewed as a body- bar structure is minimally rigid (|E| = 6(6) −6 = 30 and on all subgraphs the counts prescribed in Molecular Theorem are satisfied.) (e) Protein structure in stick representation (body-hinge) with black, red and green lines corresponding to covalent bonds, hydrogen bonda and hydrophobic contacts, respectively. (e) Body-bar representation of a protein.*

flexible [24]. The combinatorial structure of a general body-bar framework is a multigraph G = (V, E) where V is a set of bodies (atoms) and E is a set of bars. Tay's remarkable theorem [19] shows that the rigidity of any generic body-bar framework (which extends to body-hinge frameworks) is determined only by the underlying multigraph G. Generic here means that the bars avoid special geometries, or more specifically the underlying rigidity matrix has a full rank (see [23, 24] for details). Almost all body-bar frameworks are generic [21]. Tay's theorem also extends to generic body-hinge structures [20]. A powerful result that was conjectured by Tay and Whiteley in 1983 and was recently proved by Katoh and Tanigawa [8] shows the same counting condition stated in Tay's theorem also characterizes the rigidity of generic molecular frameworks. This result is known as t*he Molecular Theorem*. We have combined it with Tay's theorem into one statement. A body-bar framework is *minimally rigid* if removal of any edge (bar) results in a flexible framework.

THEOREM 1: (Tay's Theorem - Molecular Theorem) [8]: A generic body-bar framework (and generic molecular framework where bonds are replaced by 5 bars) on a multigraph G=(V,E) is minimally rigid if and only if $|E| = 6|V| - 6$, and on all subsets of edges $|E'| \leq 6|V'| - 6$.

The $6|V| - 6$ count on a multigraph defines an independent set in a matroid [22, 23], which naturally leads to a greedy algorithm. Using the $O(V^2)$ 'pebble game algorithm' [11] it is possible to track the $6|V| - 6$ count for independence of edges and rigidity in a multigraph. The implementation of the pebble game and Molecular Theorem has led to the development of program FIRST.

B. FIRST and protein flexibility predictions

Starting with a PDB structure (i.e 3-dimensional atomic coordinates), FIRST [10] generates a body-bar framework (a multigraph) of a protein, consisting of atoms (vertices) and edges (covalent bonds, hydrogen bonds, hydrophobic contacts and electrostatic interactions) (Fig. 1e). Hydrogen bonds are critical for protein stability [1, 18]. The strength of each hydrogen bond is calculated using the Mayo energy potential which takes into effect the local donor–hydrogen-acceptor atom geometry [10]. A hydrogen bond cutoff energy value is selected such that all bonds weaker than this cutoff are ignored. Once the final constraint multigraph is obtained, FIRST then applies the pebble game algorithm [13, 19], and decomposes the protein into rigid clusters and flexible regions. A rigid cluster moves as a single rigid body with its trivial 6 degrees of freedom (3 rotations and 3 translations). Every bond in a rigid cluster is non-rotatable. Protein normally consist of several rigid regions connected by flexible linkers (Fig. 2).

The main workhorse of the program FIRST, the pebble game algorithm, which originates from efficient bipartite matching formulations, checks the counting characterization of generic rigidity of molecular frameworks prescribed by the Molecular Theorem (Theorem 1). Due to the combinatorial nature of the Molecular Theorem, the pebble game does not depend on the atomic coordinates, it is a combinatorial integer algorithm as opposed to numeric, it always gives an exact answer. The pebble game determines if a constraint (edge) is 'independent' (i.e. removes degrees of freedom from the network) or is otherwise 'redundant'. Since it is a greedy algorithm, the order the edges are tested for independence is not important – the final rigidity and degree of freedom prediction is always unique. Pebbles are synonymous with degrees of freedom and a removal of a pebble indicates the inserted constraint (edge) is independent. Redundant constraints do not remove degrees of freedom (pebbles) as their insertion (or deletion) to an already rigid region causes no change in rigidity. A rigid region is 'redundantly rigid' if a removal of any one of its constraints (edges) still keeps the region rigid. If a rigid region in the protein has significant number of redundant edges, then breaking some of its constraints like a hydrogen bond or hydrophobic contact will not alter its rigidity. Redundantly rigid region is what a biochemist would call a 'stable region' as its rigidity is not affected by the flickering nature of the hydrogen bonds [18].

Rigidity prediction in FIRST is performed at a hydrogen bond energy cutoff, with default at -1 kcal/mol (Fig. 2 a, b). Changes in the rigidity can be monitored by a gradual removal of hydrogen bonds one by one (i.e. by lowering of hydrogen bond energy cutoff) in the order of increasing strength, keeping all covalent and hydrophobic interactions intact, and then redoing the rigidity analysis at each step identifying rigid and flexible

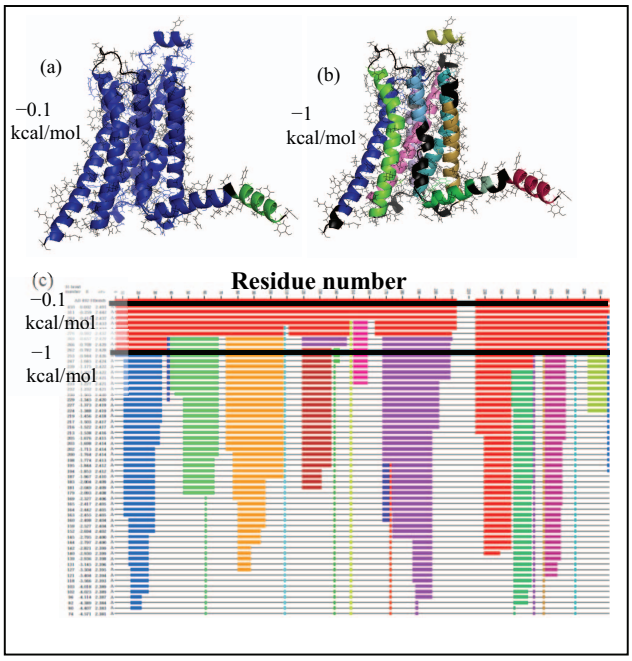

Fig. 2. *R*igidity prediction using program FIRST. A2A adenosine receptor is mainly composed of a single rigid cluster at (a) -0.1 kcal/mol hydrogen bond cutoff as most hydrogen bonds are modeled in the network and (b) several clusters (transmemrane α-helices) at -1.1 kcal/mol. (c) Hydrogen bond dilution plot indicates how protein breaks down as we decrease the cutoff and break more hydrogen bonds. Columns on left are updated and display the hydrogen bond energy levels, corresponding cutoff lines shown at -0.1 and -1.1 kcal/mol are highlighted. Flexible regions are indicated with thin black lines, and rigid regions are indicated with blocks, with separate colours indicating distint rigid clusters. Initially with inclusion of all potential hydrogen bonds, the protein is quite rigid (red block) and as hydrogen bonds are gradually broken with increasing energy, this GPCR decomposes into several rigid clusters.

regions. The change in rigidity can be visualized nicely using the hydrogen bond 'dilution plot' (Fig. 2 c) [8, 18].

While tremendous computational resources are needed to simulate protein flexibility with MD simulations, FIRST can predict the rigid clusters and flexible connections (known as the *rigid cluster decomposition*) very rapidly (less than a second on a standard PC). Previous studies have demonstrated that FIRST gives accurate predictions of flexibility and rigidity in proteins that match well with experimental evidence [8, 18], and it has been applied on large protein assemblies such as viral capsids [6]. It has also been utilized in many practical application such computer drug design, protein engineering, predictions and replications of experimental measures of dynamics such as hydrogen-deuterium exchange [18] and many others. We give one useful illustration of the wide applicability of rigidity-theoretical protein flexibility computations – predicting the effect of mutations on flexibility of proteins.

Faulty mutations are frequently observed in genetic diseases [1]. For instance, in cystic fibrosis disease, a key protein cystic fibrosis transmembrane conductance regulator (CFTR) undergoes a single point mutation ΔF508, a deletion of amino acid phenylalanine [25]. A number of studies have

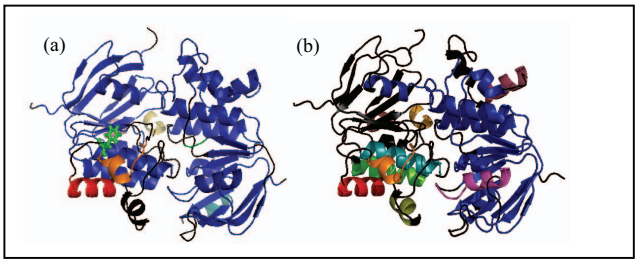

Fig. 3. *R*igid cluster decomposition obtained with FIRST on CFTR protein at −1 kcal/mol hydrogen bond cutoff. Separate colours indicated dinstinct rigid cluster. Blue is the larget rigid cluster. Black represents highly flexible region. (a) Native CFTR prediction with phenylalanine shown in green sticks is dominated by a large rigid cluster (pdb id: 2pze). (b) Mutant ΔF508 with missing phenylalanine is significantly more flexible (pdb id: 2pzf).

demonstrated that the mutant form ΔF508 CFTR has significantly greater conformational flexibility compare to the native (wild-type) structure [25] and this increased flexibility is believed to lead to premature degradation and loss of CFTR function.

To investigate this phenomenon with rigidity theory, we analyzed flexibility of the mutant and non-mutant (wild type) CFTR protein with FIRST (Fig. 3). FIRST prediction, which took only seconds to perform on a standard CPU machine, indicates that the structure without the phenylalanine is significantly more flexible than the non-mutant form. The non-mutant form is dominated by a single large rigid structure while the mutant form consists of many smaller rigid clusters connected by a large flexible region. FIRST analysis is indeed in agreement with the MD simulations and experimental evidence that a deletion of a single amino acid results in a significant overall reduction of rigidity [25]. This also raises a possibility of an allosteric control of CFTR, where an insertion of a single amino acid not only locally rigidifies the nearby residues but propagates and rigidifies other distant parts of the protein.
Allostery of CFTR will not be discussed here, but it certainly paves a prospect for further research avenues to be explored with both experimental studies and computational rigidity allostery detection (see below).

FIRST decomposes the protein into rigid clusters and flexible regions, but does not simulate actual protein motions. One powerful extension uses the rigid clusters in geometric simulation algorithms such as FRODA [21] to explore the dynamics. Rigid clusters serve as a natural coarse graining step, where hundreds of degrees of freedom are removed from the overall system. Monte Carlo method FRODA uses the output of FIRST as a preprocessing step to explore the conformational space of the flexible regions. FRODA rapidly (100 000 time speed ups can be obtained when compared to MD simulations) generates conformations that are consistent with bond lengths and angular constraints, while maintaining all rigid clusters. We have developed similar technologies recently and applied it on Tau protein, a key protein in number of pathologies and

dementias such as Alzheimer's disease [28]. Tau protein is part of a class of disordered proteins that are highly flexible and MD simulation are of no practical use due to their high dynamic nature. One of the main challenges in finding successful therapeutics for Alzheimer's is the poor understanding of the atomic structure and dynamics of the TAU protein. We were able to show a first unprecedentedly detailed view of the structural and dynamic nature of both the normal and defective forms of Tau [28]. The outcome of this study provides rich understanding of structural basis of tau pathology. It is clear that rigidity- inspired methods and related emerging technologies are important tools in modern protein science research and drug discovery.

IV. DESCRIPTION OF RIGIDITY TRANSMISSION ALLOSTERIC MECHANISM

The link between protein flexibility and its function is clear and rigidity theory offers powerful computational techniques for understanding this critical relationship. We bring about a further connection and introduce a first rigidity-based description of allosteric control of protein function. The importance of allosteric regulation of protein function is widely established (see section I), but most critical questions surrounding allostery still remain unresolved. One of the key remaining puzzles is to describe the general mechanism of distant coupled conformational change. Specifically, how a structural change in conformation at an allosteric site (in some cases a subtle change) induces a change in conformation at a distant active site. Moreover, what region in the protein is important for this transmission of information (i.e. what are the allosteric pathways?) [3, 5].

We propose transmission of changes in rigidity, more specifically degrees of freedom propagation across the protein network, offers a plausible mechanical model for allosteric coupling between distant sites. This model is founded on our seminal work in the area of rigidity theory [19], which was also considered by Whiteley et al [4] in special geometric frameworks. Our allostery model predicts if mechanical perturbation of rigidity (mimicking ligand or drug binding) at one site of the protein, call it A, can percolate and transmit across a protein structure and in turn cause a transmission and a change in rigidity and conformational degrees of freedom at a second remote site B. Local perturbation of rigidity at site A refers to insertion of edges (constraints) to A up to its rigidification. If there is a reduction in conformational degrees of freedom in site B, due to the perturbation of A, we say that A *transmits* degrees of freedom to B and the two sites are in rigidity-based allosteric communication. The strength of allosteric signal is quantified by the maximum possible amount of transmission (reduction in DOF).

Presence of rigidity-based allostery (transmission of degrees of freedom) is equivalent to a statement that a change in shape (conformation) in site A (i.e. mechanically change the shape as binding might) will lead to rearrangement and change of shape of the second site B. Thus, rigidity-based allostery captures the essence of coupled conformational change inherent in allostery. As we show below, characterizing the allosteric mechanism and quantifying the allosteric response by transmission of degrees of freedom between distant sites can also aid in identifying novel allosteric sites. Our methods and techniques can also be used to map out the allosteric pathway.

We first illustrate the concept of degree of freedom transmission and shape change propagation between remote sites in frameworks. In Fig. 4 (a) we have a 2-dimensional bar and joint framework [23] that is composed of bars (rods) which fix the distances between the connecting flexible ball joints. This framework has a single non-trivial degree of freedom (we can check this with rigidity matrix [23]). If we slightly vary the distance between the two joints u and v in site A (analogous to simulating ligand binding), this initial change in shape propagates across the framework and results in a substantial change in shape and conformation at the distant site B.

Equivalently, if we rigidify site A, fixing the distance between the end joints in A (i.e. insert a bar connecting u and v) will rigidify site B, stopping the motion in B. In this example there is a transmission of one degree of freedom between A and B. In protein analogy, a small ligand that fits in site A can pull on the two vertices, which in turn leads to a change in conformation and a closing motion at site B, allowing site B to more likely dock its binding ligand partner. This example provides a hypothetical analogue to positive allosteric modulation [1] where binding of one ligand enhances the binding to another ligand at a distant site.

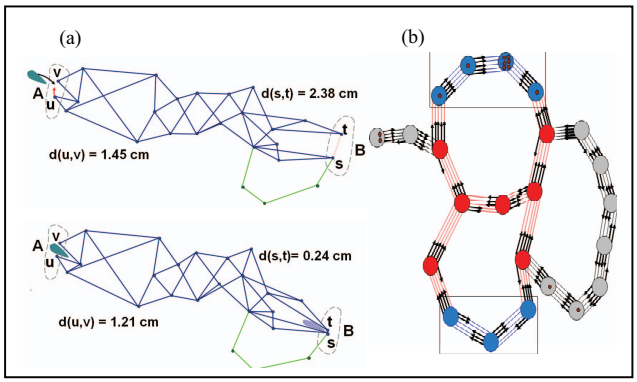

Fig. 4. Allostery in frameworks: transmission of degrees of freedom and shape change. (a) Bar and joint framework has one non-trivial DOF (excluding green edges) and this single DOF can transmit between A and B. If we move u and v closer together (simulating ligand binding) this slight motion in A propagates across the framework and results in a change in conformation in site B. Note that green edges are irrelevant for this transmission; we can remove green edges and the same effect is observed. This example was generated with Geometer's sketchpad software. (b) Applying the RTA algorithm (see below) on this body-bar framework, we see that $DOF^{AB} = 2$ DOF can be transmitted between the sites A and B. Here we have shown the output of the pebble game, which can be used to compute the necessary DOF counts prescribed in RTA algorithm and the relevant region [19] (red vertices and edges) which is the region in the multigraph responsible for transmission of DOF (allosteric pathway).

Similar examples can be constructed that resemble negative allosteric regulation, where a closing motion at one site results in an opening motion (and release of a ligand) at a distant site. We have also illustrated the ability of a body-bar molecular framework to model the transmission of rigidity and changes in degrees of freedom across the structure (see Fig. 4 b).

We now describe the algorithm for detecting transmission of degrees of freedom and allostery in protein structures.

*A. **Rigidity-transmission allostery (RTA) algorithm.***

Input: A molecular body-bar multigraph G_h corresponding to hydrogen bond energy cutoff h. Two disjoint vertex induced subgraphs A and B.

Goal: Predict if rigidifying A transmits a reduction in conformational degrees of freedom (DOF) at B.

1. Calculate the available DOF in site B, call it DOF^B: (DOF^B is the number of independent edges that would need to be added to B that result in its rigidification. Equivalently, we can run the pebble game algorithm on G_h and count the maximum number of free pebbles less six we can gather on vertices in B [19].)
2. Perturb rigidity of A: rigidify A by adding maximum number of independent edges within A.
3. Re-calculate the available DOF in B, call it $DOF^B_{Aperturbed}$.

Output: Transmission of DOF from A to B is:
$\mathbf{DOF^{AB} = DOF^B - DOF^B_{Aperturbed}}$
(i.e. maximum reduction in DOF at B given perturbation of A). When $\mathbf{DOF^{AB}} > 0$, A and B are in allosteric communication.

Rigidity-based allosteric communication is completely symmetric. In other words, the effect of perturbing rigidity of site A on site B and the maximum amount of degree of freedom transmission is identical if we perturb B and observe the effect on site A. For transmission of degrees of freedom to be feasible, both sites A and B have to have some internal flexibility [19].

Example of RTA algorithm: Refer to framework in Fig. 4 (b). We obtain the following counts: $DOF^B = 2$, $DOF^B_{Aperturbed} = 0$, $DOF^{AB} = 2 - 0 = 2$. Thus, in this example there is a transmission of 2 degrees of freedom from A to B. In other words, rigidifying A results in a reduction of 2 DOF in B and subsequent complete rigidification of site B. Transmission of DOF occurs only over the red region in the graph (relevant region), which can be extracted with a relevant region detection algorithm we have previously developed [19]. In contrast, the gray region in the multigraph is irrelevant for transmission of DOF from A to B, that is removal of gray region does not alter this transmission.

Remark1: As the count of transmission of DOF is symmetric, B can also transmit 2 DOF to A. However, as site A has 3 DOF instead of 2, note that rigidification of site B would result in a reduction in A from 3 DOF to 1 DOF, leaving some flexibility.

Remark2: The proofs of correctness of the degree of freedom counts extracted from RTA algorithm, the pebble game extensions that allow fast computations of counts in step 1 and 3, and the relevant region detection algorithm for detection of allosteric pathway are not discussed here, for this we refer to initial work on rigidity-based allostery transmission [19].

V. APPLICATIONS OF RIGIDITY-TRANSMISSION ALLOSTERY IN PROTEINS

To show the applicability of rigidity-based modeling of allostery on actual protein structures, we apply the RTA algorithm on two protein classes, a G-protein coupled receptor (GPCR) and eukaryotic translation initiation factor eiFA4E (or 4E for short). We only provide a brief biological discussion here. Full detailed analysis and discussion of biological consequence of the finding presented will appear in forthcoming papers.

The largest class of receptors in the human genome are G-protein coupled receptors (GPCRs), which are essential components of signal transduction throughout the body [1, 12, 27]. GPCRs mediate most transmembrane signal transduction across large distances over the cellular membrane by responding to an enormous variety of extracellular stimuli (drugs, hormones, neurotransmitters and other proteins). Since GPCRs are responsible for the control of most information that passes into the cell, they play a critical role in disease and are the most commonly targeted receptor class in drug design. About 50% of all modern medicinal drugs bind to GPCRs [27].

In humans there are over 800 GPCRs [1], whose 3-dimensional structure consists of 7-transmembrane alpha-helices. Signal transduction and activation is regulated by extracellular ligands. Activating ligands of GPCRs are called agonists and inactivating ligands are called antagonists or inverse agonists and they roughly bind at the same location at the extracellular region of receptor known as the 'orthorsteric site' [12]. Activation can also occur through binding events at other allosteric sites. Upon binding of the agonist ligand, it is believed that this information propagates across the receptor and results in relative movement of α-helices and a change in conformation at the intracellular side of the receptor, so it can activate and engage binding of its G-protein partner [12, 27]. GPCRs are naturally allosteric but how they transmit the allosteric signals across the membrane is still not well understood and is a major area of research in both academia and pharmaceutical industry.

To shed some light on the possible allosteric mechanism in GPCRs, we have performed the analysis with RTA algorithm on human adenosine A2A receptor (Fig. 5). This GPCR is a drug target for various disorders such as inflammation, cancer, diabetes, infectious diseases and neuronal defect disorders [12]. We defined site A as the orthosteric site (all atoms and induced edges on the receptor that are interacting with the bound ligand) and residues 230 and 291 at the intracellular side where G-protein binds as site B (Fig. 5 a). We tested for rigidity-based communication in four different available x-ray crystal structure of A2A receptor, three which are in active configuration. RTA algorithm was performed for all hydrogen bond energy cutoffs

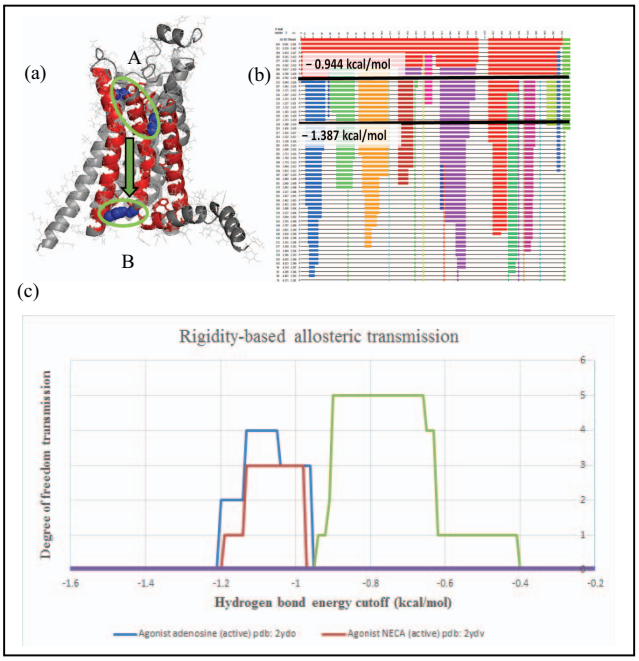

Fig. 5. RTA algorithm applied on human adenosine A2A receptor. (a) Transmission of DOF was tested between site A (orthorsteric site) and site B (G-protein binding region). Here we have shown the relevant region between the two sites A and B in blue pointing out the allosteic pathway (in red) connecting A and B in the adenosine-bound structure (pdb 2ydo) at -1 kcal/mol cutoff. Gray parts are irreleant for allosteric transmission. (b) The range of cutoffs where transmission occurs in adenosine-bound structure (2ydo) is superimposed on its dilution plot, which starts at -0.944 kcal/mol and stops at -1.387 kcal/mol. (c) Plot of tranmission of DOF as a functin of the energy cutoff in four different A2A receptor crystal structures. In all three active-like structures bound to agonists, transnission of DOF occurs. In the inactive state, tranmission of DOF is not seen. When the cutoff is close to 0 kcal/mol, no transmission is possible as the whole protein is rigid including the sites A and B. As we lower the cutoff and break more hydrogen bonds the protein becomes less rigid and eventually allosteric transmission starts. Transmission of DOF continues for some range of cutoffs and stops once the significant portion of hydrogen bond network has been diluted.

h in increments of 0.01 kcal/mol and we obtained the DOF transmission amounts for each cutoff, which can be visualized in a plot showing DOF transmission as a function of an energy cutoff (see Fig. 5 c).

The RTA prediction shows that in all three agonist-bound (active) structures, perturbation of rigidity at the orthorsteric site propagates across the receptor to the remote G-protein binding region and transmits a change in degrees of freedom. Remarkably, in the inactive structure there is no degree of freedom transmission (Fig. 5 c). These findings are consistent with the general roles of agonists vs inverse agonists on activation and inactivation of the GPCRs [1, 12, 27]. Moreover, this analysis suggests that transmissions of rigidity upon binding of agonist is important for structural and conformational changes required for activation of a GPCR and unlike agonists the antagonist binding prevented this change in conformation to propagate. This points to the role of rigidity-based communication of functionally important in allosteric control of A2A receptor.

Remark: The two agonist ligands adenosine and NECA are structurally very similar. The authors of the study that solved the crystal structures of the A2A receptor with these two ligands point out that the two configurations of the receptor are only partially active [12]. The UKA agonists is somewhat different and bigger than the other two agonists and the authors of the pertaining study have suggested that UKA-bound receptor conformation they obtained is in a fully active conformation [27]. This may explain why adenosine and NECA bound structures transmit DOF at an almost identical hydrogen bond energy range, whereas UKA-bound receptor transmits more DOF and also at an earlier and wider range of hydrogen bond energy cutoffs.

We now illustrate the RTA algorithm on protein eiF4E (or 4E for short). 4E directly binds to messenger RNA and plays a crucial role in eukaryotic protein synthesis [1, 14]. 4E becomes overexpressed in cancer cells and is an important drug target for various cancers [14]. 4E strongly interacts with initiation factor protein eiF4G (4G) and when the 4E-4G complex is formed, 4E is in the active state. Disrupting the 4E-4G interactions is crucial for development of anticancer agents [14]. Finding allosteric sites on 4E is important as binding events at the allosteric sites could propagate and disrupt 4G-binding.

As our method is computationally fast, we can rapidly scan many unknown sites for rigidity-based allosteric communication, identifying potential new allosteric sites. Starting with the crystal structure of 4E bound to 4G, we performed the RTA analysis by testing if remote sites on 4E are in rigidity-based allosteric communication with 4G protein. In order to not drastically disrupt the rigidity of 4E, we applied the initial rigidity perturbation on the flexible loop regions which are far removed from 4G. In terms of RTA algorithm, 4G was taken to be site B. We then sequentially perturbed the rigidity of every of 3 consecutive residues (site A) on the flexible loops of 4E in a window sliding approach and then monitored if there is a transmission of DOF to 4G. This was again performed for various cutoff ranges h. If for any window of 3 consecutive residues (site A in the RTA algorithm) we found transmission of degrees of freedom to site B, we calculated the area under the transmission curve (refer back to Fig. 5 c) to quantify the intensity of degree of freedom propagation.

In Fig. 6 a we have shown the output of the allostery predictions. We find that most of the loops on 4E (coloured in blue) are not in allosteric communication with 4G, with exception to one region around residues 77-82 (coloured in red). Wagner and colleagues [14] discovered when eIF4E/eIF4G inhibitor compound 4EGI-1 binds to 4E, which prevents 4G binding, it bound at a remote location different from 4G binding site suggesting that the mechanism was allosteric (Fig. 6 b). Remarkably, this inhibitor bound on the same region where we found the most degree of freedom transmission to 4G. In fact, we see that the inhibitor is in direct contact with residues 77-82 (Fig. 6 c). This finding clearly points to a propagation of rigidity and degree of freedom as being functionally important for allosteric transmission in 4E.

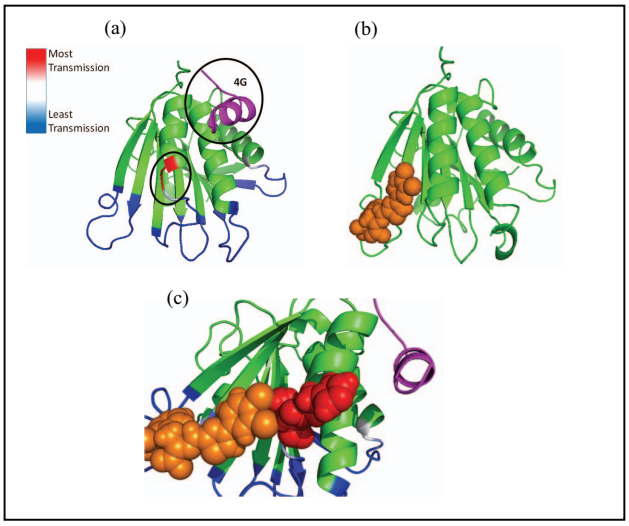

Fig. 6. Detection of allosteric hotspots using RTA algorithm applied on eiFA4E (4E). (a) 4E structure bound to 4G (purple) (pdb id: 1ejh). Distant loops 4E are coloured based on the intesity of DOF transmission. Most of the loops (blue) are poor tranmitters of DOF with the exception of residues 77-83 (red and circled) which we predict to be an allosteric hotspot. (b) 4E structucture with the bound inhibitor drug 4EGI-1 (orange spheres) (pdb id: 4tpw) which allosterically disrupts 4G binding. (c) Close up view showing the spatial proximity and direct contact of the drug 4EGI-1 (orange) and regions on 4E we predict with RTA algorithm to be the most allosteric. 4E structure with inhibitor (4E was removed) was superimposed with Pymol software onto 4E structure bound to 4G.

Detection of allosteric hotspots is very powerful as it can provide novel guidance in controlling the function of proteins. Drugs which interact with allosteric surfaces on the proteins are therapeutically preferable since they can provide greater specificity and potency than the traditional drugs that interact directly with the active site [3, 5]. Unlike allosteric sites, active sites are largely conserved and structurally homogeneous across other functionally diverse proteins.

VI. CONCLUSION

Rigidity theory provides powerful tools for studying protein flexibility and its function. A straightforward method that describes how allosteric signals are transmitted across protein structures has been previously difficult to design [3, 5]. Our allostery detection method provides a new mechanistic view of allostery through transmissions of rigidity and conformational degrees of freedom and it can describe functionally important features in protein signalling. Using the RTA prediction we can quantify the strength of allosteric signals, detect previously unknown allosteric sites and depict the allosteric pathways that are crucial for the allosteric communication across the network. Moreover, due to the speed of the pebble game algorithm our techniques are suitable for high throughput allostery analysis. Current research aims to incorporate knowledge about the hydrogen bonding network from multiple protein structures (ensembles from available MD simulations or NMR structures). This should improve the robustness of allosteric predictions and remove any sensitivities related to a construction of a network using a single structure as we have previously demonstrated on flexibility predictions [18]. Forthcoming papers will further demonstrate our predictions are in agreement with experimental data and provide strong case for rigidity-based allostery as a mechanistic description of allosteric regulation. This should eventually allow us to obtain a better understanding of allostery and tackle the more difficult signalling events in the cell.

REFERENCES

[1] B. Alberts, A. Johnson, J. Lewis, M. Ra, K. Roberts and P. Walter, Molecular Biology of the Cell, Garland Science, New York, Fourth Edition, 2002.

[2] J.P. Changeux, F. Jacob and J. Monod, Allosteric proteins and cellular control systems, J Mol Biol, 6:306-329, 1963.

[3] A. W. Fenton, Allostery: An illustrated definition for the “second secret of life,” Trends Biochem. Sci., 33, pp. 420–425, 2008.

[4] W. Finbow-Singh, W. Whiteley, Isostatic Block and Hole Frameworks. Siam J. Discr. Math. 27, 991–1020 (2013).

[5] K. Gunasekaran, B. Ma and R. Nussinov, Is Allostery an Intrinsic Property of all Dynamic Proteins?, Proteins: Struct., Funct., Bioinf., 57, 433443, 2004.

[6] B.M. Hespenheide, D.J. Jacobs and M.F. Thorpe, Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus, J. Phys.: Condens. Matter 16 S5055-64, 2004.

[7] L.E. Kavraki, G.N. Phillips JR. and M.L. Teodoro, Journal of Computational Biology, Volume 10, Numbers 34, p. 617-634, 2003.

[8] N. Katoh and S. Tanigawa, A Proof of the Molecular Conjecture, Discrete & Computational Geometry 45: 647-700, 2011.

[9] D.M. Krüger et al, CAN web server, H. Nucleic Acids Res. 2013.

[10] L.A. Kuhn, D.J. Rader and M.F. Thorpe, Protein exibility predictions using graph theory, Proteins, 44:150-65, 2001.

[11] K. Leach, P.M. Sexton and A. Christopoulos, Allosteric GPCR modulators: taking advantage of permissive receptor pharmacology, Trends Pharmacol Sci., 28(8):382-9, 2007.

[12] G. Lebon, T. Warne, P.C. Edwards, K. Bennett, C.J. Langmead, A.G.W. Leslie and C.G. Tate, Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation, Nature, Vol 474, 521, 2011.

[13] A. Lee and I. Streinu. Pebble game algorithms and sparse graphs, Discrete Mathematics, 308, 1425 1437, 2008.

[14] E. Papadopoulos, S. Jenni, E. Kabha, K.J. Takrouri, T. Yi, N. Salvi, R.E. Luna, E. Gavathiotis, P. Mahalingam, H. Arthanari, et al. Structure of the eukaryotic translation initiation factor eIF4E in complex with 4EGI-1 reveals an allosteric mechanism for dissociating eIF4G Proc. Natl. Acad. Sci. U S A, 111, pp. E3187–E3195, 2014.

[15] Nussinov, R. & Tsai, C. J. Allostery in disease and in drug discovery. Cell 153, 293–305(2013).

[16] B. Schulze, A. Sljoka A. and Whiteley W., How does symmetry impact the flexibility of proteins?, Philosophical Transactions of Royal Society A 372, 20120041, 2014.

[17] D.E. Shaw et al., Anton, A Special-Purpose Machine for Molecular Dynamics, Communications of the ACM (ACM) 51 (7): 9197, 2008.

[18] Sljoka A. and Wilson D., Probing Protein Ensemble Rigidity and predictions of Hydrogen-Deuterium exchange, Physical Biology, 10 056013, 2013.

[19] Sljoka A., Algorithms in rigidity theory with applications to protein flexibility and mechanical linkages, PhD dissertation, York University, Toronto, 2012.

[20] T.S. Tay, Rigidity of Multigraphs I: Linking Rigid Bodies in n-space, Journal of Combinatorial Theory Series B, Vol. 26, pp. 95-112, 1984.

[21] S.A. Wells, S. Menor, B.M. Hespenheide and M.F. Thorpe, Constrained geometric simulation of difusive motion in proteins, Phys. Biol., 2 S12736, 2005.

[22] N. White, Encyclopedia of Mathematics and its Applications: Theory of Matroids, Cambridge University Press, Cambridge, 1986.

[23] W. Whiteley, Some Matroids from discrete applied geometry. In J. Bonin, J. Oxley, and B. Servatius, editors, Matroid Theory, volume 197 of Contemp. Math., pages 171-311. Amer. Math. Soc., Providence, 1996.

[24] W. Whiteley, Counting out to the fexibility of molecules, Phys. Biol. 2, S116-S126, 2005.

[25] G. Wieczorek and P. Zielenkiewicz, DeltaF508 mutation increases conformational flexibility of CFTR protein. J Cyst Fibros 7: 295-300, 2008.

[26] H.Wildman K and D. Kern. Dynamic personalities of proteins Nature. 450: 964-972, 2007.

[27] F. Xu, H. Wu, V. Katritch, G.W. Han, K.A. Jacobson, Z.G. Gao, V. Cherezov and R.C. Stevens, Structure of an Agonist-Bound Human A2A Adenosine Receptor, Science 332, 322, 2011.

[28] Zhu S., Shala A., Bezginov A., Sljoka A., Audette G. and Wilson D., Hyperphosphorylation of Intrinsically Disordered Tau Protein Induces an Amyloidogenic Shift in Its Conformational Ensemble, PLoS ONE, 10(3), 2015.