Information Entropy in Characterizing Conformational Ensembles and Intrinsic Dynamics of Polypeptides

Ph.D. Thesis Booklet

Gergely Gyimesi

Supervisors:

Dr. Péter Závodszky
Dr. András Szilágyi

Consulent:

Dr. Tamás Hegedűs

Péter Pázmány Catholic University
Faculty of Information Technology

Budapest, 2011.

Contents

1. Introduction

The concepts of entropy as introduced in statistical physics and in information theory share a close mathematical analogy. This relation permits the estimation of physical entropy using methods borrowed from information technology, as well as the interpretation of the information content of physical states. In physics, entropy via the quantity of free energy determines the stability of equilibrium states and the direction of spontaneous processes. Among methods estimating free energy, an exceptional role is held by computational methods that are able to compute the absolute conformational entropy of equilibrium states. In the first part of my work, I will be introducing such a method developed by myself. Furthermore, the theoretical formula for equilibrium entropy can be factorized to the sum of the entropy of independent elementary parts (oscillators) forming the molecular system, and to mutual information terms resulting from their two- or many-body interactions. Calculating mutual information itself thus lends us information on internal interactions present in the system and any resulting correlation effects. Correlated motions play an essential role in establishing the function of proteins as molecular systems. An important example of such effects is the phenomenon of allostery in proteins, where spatially distant regions and binding sites interact with each other through correlation effects and modulate each other's function. Allosteric effects appear in a markedly interesting way in ABC (ATP-Binding Cassette) transporter proteins, where they are responsible for coupling ATP binding and hydrolysis to substrate transport. The second part of my work describes my recent advances in understanding this process. The elucidation of the allosteric effects in ABC transporter proteins would open a new frontier on rational drug design targeting these proteins for combating several associated human diseases.

2. Estimating the Conformational Entropy of Molecular Ensembles

2.1. Background and Aims of Research

The equilibrium properties of systems in molecular biology are determined by the physical quantity of free energy. Significant theoretical efforts have been devoted to the calculation of free energy difference between two equilibrium states of molecular systems. For methods that work by integrating along a path between the two states, such as

thermodynamic integration (TI, [1]), free energy perturbation (FEP, [1, 2]) and histogram analysis methods [3], the appropriate definition of a reaction coordinate and possible significant structural differences between the two states can hinder calculations [4]. Other methods, which calculate the absolute free energy of a single state, can pose an alternative. In this case, however, it is essential to devise an efficient method that is able to estimate the conformational entropy of the system in an equilibrium state.

The individual conformations constituting an equilibrium state of the system are usually generated by a sampling method, e. g. molecular dynamics (MD) simulations or Metropolis Monte Carlo (MC) sampling. Several methods exist for entropy estimation of such molecular ensembles, usually approximating the phase space density of the system and calculating the information entropy of the resulting distribution. One of the earliest methods to calculate conformational entropy is based on the quasiharmonic approximation, which estimates the phase space density of the system as a multivariate Gaussian distribution [5]. What makes this method most appealing is that the parameters of the distribution can be readily derived from the data and the entropy of the resulting distribution can be analytically calculated. The interpretation of the quasiharmonic assumption is also simple, as the axes of the Gaussian distribution determine vibrational modes of the system that are considered to be independent by the method, while all higher order correlation effects are neglected.

The original quasiharmonic method employed the BAT ("bond-angle-torsion") coordinate system, which uses bond lengths, bond angles and torsion angles to describe the position of atoms. The use of such internal coordinates is favorable because the different vibrational modes are easily separable, and coordinates responsible for forming specific conformations are identifiable. In practice, however, defining such a non-redundant internal coordinate system is not a trivial task, and the treatment of periodic coordinates can cause difficulties. As a solution, a version of the quasiharmonic method adapted to Cartesian coordinates was introduced by Schlitter, which also takes into account the entropy contribution of the kinetic degrees of freedom, and adds the quantum mechanical treatment of molecular vibrations [6].

Behind the quasiharmonic approximation lies the assumption that the system undergoes harmonic fluctuations around a single, minimal energy state. Systems in molecular biology, however, typically exhibit a complex, elaborate energy landscape and visit multiple locally minimal energy basins at relevant temperatures, which are separated by low energy barriers. These motions contribute to the anharmonicity of vibrational modes in the system. It is also a serious error to assume the independence of the quasiharmonic modes, essentially

ignoring second and higher order correlations which contribute significantly to conformational entropy [5, 7]. Several solutions and methods have been proposed to tackle these problems.

To account for the anharmonicity of vibrational modes, more accurate methods have been applied for sample based density estimation, e. g. parametric estimation based on Fourier series [8], or non-parametric estimation based on anisotropic kernel functions [9]. The entropy estimation method devised by Numata *et al* estimates the value of the density function at the sample points using the *k*-nearest neighbor density estimation method [10], which has the advantage that it doesn't make an *a priori* assumption on the form of the density function. However, the disadvantage of these methods is that a large number of samples (~10^6) is needed to reach convergence.

A different approach is used by the 2D entropy method developed by Wang and Brüschweiler, which performs density estimation with Gaussian kernels along each quasiharmonic component, and uses numeric integration to calculate their information entropy content [11]. The 2D entropy method also incorporates anharmonic effects, but the bandwidth of the Gaussian kernel is left as a parameter which is not transferable among different systems [11, 12], and which heavily influences the calculated value of entropy.

The contribution of second and higher order correlations to conformational entropy can be separated using mutual information expansion (MIE), which expresses the entropy *S* of the system [13] as

$$S(q_1,\ldots,q_N) = \sum_{i=1}^{N} S_1(q_i) - \sum_{i<j} I_2(q_i,q_j) + \sum_{i<j<k} I_3(q_i,q_j,q_k) - \ldots$$

Here, S_1 are the univariate, marginal entropies of individual coordinates, and the mutual information terms I_k can be shown to quantify the *k*th order correlation between coordinates [14-16]. The *k*th order mutual information terms can be calculated from the *k*th and lower order entropies, e. g. for $k = 2$,

$$I_2(q_1,q_2) = S_1(q_1) + S_1(q_2) - S_2(q_1,q_2)$$

The mutual information expansion is used in several entropy estimation methods [10, 13, 17]. The method of Numata *et al* uses second order mutual information terms ($k = 2$) to enhance entropy estimation.

The aim of my doctoral work was to develop and test a novel method for the estimation of the conformational entropy of molecular systems, based on Gaussian mixture functions. To test the method, five systems of small peptides whose entropy can be calculated

by an *ab initio* method were generated, and used as reference to evaluate the results yielded by the Gaussian mixture method and other methods described in the literature.

The Gaussian mixture method was implemented in the form of Octave, Python and Perl scripts. The coordinates of the test systems were generated using the SYBYL 7.3 molecular modeling software suite. The energy of individual conformations was estimated using the GROMACS software suite and the GROMOS96 united atom force-field [18]. Other methods taken from the literature were implemented as Octave scripts, except for the method of Numata *et al*, for which the software was kindly provided by the authors.

2.2. Novel Scientific Results

1. Established an entropy estimation method based on Gaussian mixture functions.

The creation of a method to estimate conformational entropy based on Gaussian mixture functions was motivated by the need to have a conceptionally simple, comprehensible method that is able to handle complications arising from the complex energy landscape of larger molecules with many degrees of freedom. Gaussian mixture functions are weighted sums of independent, multivariate Gaussian functions,

$$f(x;\theta) = \sum_{k=1}^{K} p_k g(x; m_k, \sigma_k)$$

where $g(x; m_k, \sigma_k)$ is a Gaussian function centered at m_k with a covariance matrix of σ_k, θ is the set of m_k, σ_k parameters, and K is the number of Gaussian components. Gaussian mixture functions can be used to approximate any smooth function to arbitrary accuracy [19]. The method works by approximating the phase space density of the system by fitting a normalized Gaussian mixture function on the given conformational ensemble using maximum likelihood estimation (MLE), and then estimating the information entropy of the resulting continuous distribution. For likelihood maximization, the expectation maximization (EM) method was used. The entropy of the resulting Gaussian mixture function is then calculated using the method described by Huber *et al* [20]. This method uses second order Taylor series to estimate the contribution of other Gaussian in the neighborhood of each Gaussian component, which can then be calculated analytically.

2. The Gaussian mixture based method is able to accurately estimate the conformational entropy of molecular systems.

The difficulties in estimating the conformational entropy of molecular systems mostly arise due to the complexity of the energy landscape and the anharmonicity of fluctuations around states of local energy minima. The Gaussian mixture method was therefore tested on small peptide systems. The molecules were chosen to demonstrate the applicability of the method to proteins, as well as to select molecules with side-chains having many different rotameric states to model a complex energy landscape.

To evaluate the method, the results were compared first to a reference entropy value calculated by an *ab initio* method, and then to results of other existing methods used to calculate conformational entropy. Four other methods described in the literature were implemented: the original quasiharmonic method by Karplus and Kushick [5], the covariance matrix based method of Schlitter [6], the 2D entropy method of Wang and Brüschweiler [11] and the method of Numata *et al* [10], for which the software was kindly provided by the author.

Five different molecular systems (Ala_3, Ala-Val-Ala, Ace-Ile-Nme, Ace-Val-Nme, Val_2 peptides) were created for the study, having 4, 5, 4, 3 and 4 rotatable torsion angles, respectively (Figure 1).

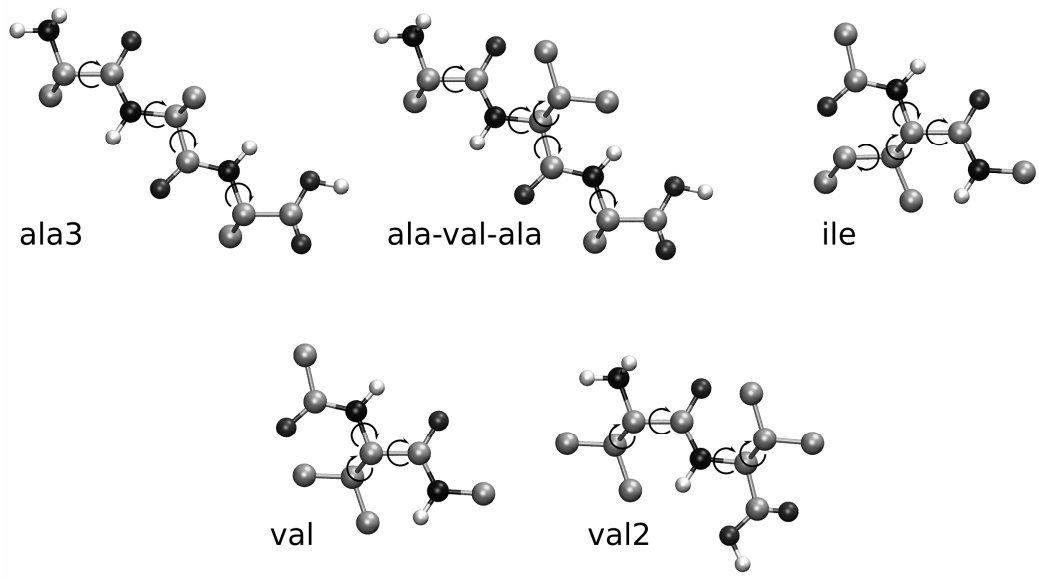

Figure 1. Small peptides and derivatives selected for the calculations. The selected molecules were Ala_3 ("ala3"), Ala-Val-Ala ("ala-val-ala"), Ace-Ile-Nme ("ile"), Ace-Val-Nme ("val") and Val_2 ("val2"). The carboxy and amino termini of all peptides were uncharged. United carbon atoms are shown as grey, oxygen and nitrogen as black, and hydrogen as white spheres. The rotatable torsion angles representing the degrees of freedom of the systems are indicated by arrows.

The first step after generating each test system was to create the full set of conformations, which was acquired by the uniform sampling of the complete internal coordinate space (rotatable torsion angles). The full set of conformations was used for *ab initio* calculation of conformational entropy of the systems based on the statistical physics definition of entropy, by numerical integration over the complete configurational space (referred to as "exact" entropy). The full set of conformations also served as a basis for Monte Carlo (MC) sampling to generate an input data set of a canonical ensemble of conformations. Also, molecular dynamics (MD) simulation with two different settings were performed to generate two additional conformational ensemble datasets for each system. Due to their construction, the entropy values calculated for the MC sampled ensembles are directly comparable to the exact entropy values. The creation of the MD dataset was motivated by the fact that molecular dynamics is a popular sampling method for biomolecules. However, in MD simulations, bond angles are usually treated flexibly, and bond lengths are either allowed to vary or kept fixed. Because of this, two different MD ensembles were created, one having flexible bond angles and fixed bond lengths ("MD/LINCS"), and the other with both bond lengths and angles flexible ("MD/Unc"). The flexible coordinates increase the degrees of freedom of the system, therefore the entropy values calculated for these ensembles are not directly comparable to the calculated exact entropy values.

The results calculated for the "val2" system are shown on Figure 2 to illustrate the most essential findings of the study. It was found that for the MC ensembles, entropy values calculated by the Gaussian mixture method show excellent agreement with the exact values (average deviation 2.04%). The quasiharmonic method of Karplus and Kushick (QH) gives a systematic overestimation of the exact value as expected [6], although the average deviation is relatively small (10.4%). The variation of the quasiharmonic method by Schlitter uses Cartesian coordinates instead of internal (torsional) coordinates, this method was implemented in two versions. In the first version, the coordinates of all atoms were input to the method (Schlitter/AA), this gave significantly larger entropy values than the exact value (average deviation 39.97%). In the second version, only atoms defining the internal coordinate torsion angles were used (Schlitter/TA), in this case the exact value was underestimated systematically (average deviation 17.77%). Similarly, the 2D entropy method was also found to underestimate the exact entropy value. With this method, however, the resulting entropy values heavily depended on the σ bandwidth parameter. The method of Numata *et al* yielded results having a relatively good agreement with the exact values (average deviation 17.54%), but underestimated the entropy in three out of five cases.

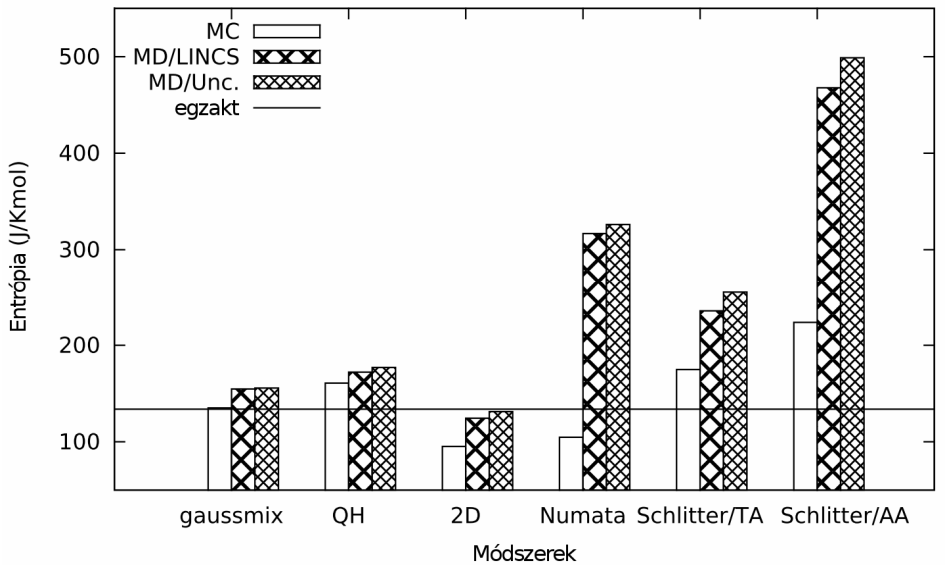

Figure 2. Entropy values calculated for the "val2" system with five different methods for the three ensembles. The exact entropy value is indicated by a horizontal line. In the case of MC ensembles, the Gaussian mixture method gave the best overall agreement with the exact value. The MD ensembles have more degrees of freedom (bond lengths and angles), and their exact entropy is not known. In their case, the calculated entropy depends to a great extent on how the entropy contributions of bond length and angle vibrations are taken into account.

The entropy values of the MD ensembles were found to be generally higher than those of the MC ensembles. The MD/Unc ensembles gave higher entropy values than the MD/LINCS ensembles in almost all cases, the difference, however, was relatively small. This indicates that the contribution of bond length vibrations to the conformational entropy is small. It is also apparent that using Cartesian coordinates (methods of Schlitter and Numata *et al*) yields significantly higher entropy values for both MD ensembles than for MC ensembles. In contrast, for methods employing internal coordinates (quasiharmonic, 2D entropy and Gaussian mixture methods), this difference is much smaller. This shows that the bond angle vibrations contribute significantly to the conformational entropy, but is mostly neglected by methods using internal coordinates.

Based on the results for the five peptide systems, I found that the Gaussian mixture method yields entropy values in excellent agreement with the exact values, while other methods perform significantly poorer (average deviation: 10–40%).

The differences in performance among the methods can have various causes. One could be that they are based on different principles. The Gaussian mixture method, similarly to the quasiharmonic and 2D entropy methods, relies solely on the observed configurational samples, whereas the other two methods are based on physical assumptions behind the observed data. This difference is apparent also from the fact that unlike the methods of

Schlitter and Numata, the Gaussian mixture method does not use atomic masses for the calculations. Methods projecting a physical model behind the data often employ the quantum mechanical treatment of oscillators, which also incorporates the kinetic contribution to the entropy, but these are not taken into account by the Gaussian mixture method.

These two groups of methods differ from each other also in the kind of coordinates used for the calculations. The methods of Schlitter and Numata use Cartesian coordinates and thus take into account contributions arising from the vibration of bond lengths and angles, as opposed to the other three methods. However, bond angle vibrations exhibit a significant extent of coupling to torsion angle fluctuations, which cannot be neglected, and also appear in entropy differences between two states [5]. In the case of methods employing different coordinate systems, the Jacobian of the transformation between the two systems also contributes to the resulting entropy difference, but this is not taken into account by the methods introduced in this work.

3. The Allosteric Mechanism of ABC Transporters

3.1. Background and Aims of Research

Mutual information can directly be related to higher order correlation effects in molecular systems [13], which play an essential role in establishing the function of proteins [21, 22]. Mutual information has been used in various studies to quantify and locate intrinsic correlated motions in proteins [23-25]. Such studies can also aid in identifying functionally important amino acids. In the second part of my work, I am reporting a quantitative study of correlated motions within ABC exporter proteins using mutual information.

The ABC (ATP-Binding Cassette) family of proteins consists mostly of membrane transporter proteins, which utilize ATP binding and/or hydrolysis to drive the active transport of substrates across biological membranes. There are currently 49 ABC transporters identified in the human genome, whose dysfunction in many cases causes various diseases (e. g. cystic fibrosis [26, 27], Dubin-Johnson syndrome [28], type 2 diabetes mellitus [29], or adrenoleukodystrophy [30]). An often occurring phenomenon during chemotherapy is the multi-drug resistance of cancer cells, the main cause of which is the overexpression of certain ABC transporters that pump the drugs from the cell into the extracellular medium [31-33]. The modulation of various ABC transporters thus bears important clinical significance, for which the detailed knowledge of the structure and mechanism of ABC transporters is essential.

Functional ABC proteins have two cytoplasmic nucleotide binding domains (NBD) and two transmembrane domains (TMD) [34, 35]. The binding of ATP to the nucleotide binding domains induces their symmetric, tight association, with the nucleotides binding on the domain interface [36, 37]. In the ATP-bound (holo) state, the cytoplasmic regions of the transmembrane domains also contact each other, and the transporter forms a “bottom-closed” conformation. In the absence of nucleotides (apo state), X-ray structures show a separation of the NBDs [38-40], and based on the extent of separation, a “bottom-closed” and “bottom-open” apo conformation can be distinguished. The currently known three apo structures show that the extent of NBD separation can vary widely, but it is not consistent with distance measurements based on double electron-electron resonance (DEER) spectroscopy [41] and chemical cross-linking [42]. In light of the artificial conditions used in X-ray crystallography, it remains a question how the resulting protein structures should be interpreted, and what relevance they bear for modeling the *in vivo* state of the transporters.

Several mechanistic models have been proposed for ABC transporter function [43-48]. These differ basically in whether they assume one or two ATP hydrolysis events in a single transport cycle, and whether they assume a constant contact between the NBDs during function. The "processive clamp" or "switch" model [46-48] postulates that the processive hydrolysis of both ATPs is necessary for the dissociation of the NBDs, while the "alternating sites" model [44] proclaims that the binding sites hydrolyse a single ATP molecule per transport cycle, in an alternating fashion. The "constant contact" model [45] also assumes a single hydrolysis event per transport cycle, but in a way that the NBDs do not separate considerably from each other. Thus, there is no universally accepted, consensus model for the mechanism of ABC transporters.

The nucleotide binding and transmembrane domains of ABC transporters communicate with each other via a sophisticated allosteric mechanism. This is indicated by the fact that substrate binding in multidrug exporters, which was shown to happen in the TMDs [49, 50], enhances basal ATPase activity in the NBDs, whereas ATP hydrolysis is necessary for active transport against a concentration gradient. Several loops and residues have been identified between the NBDs and TMDs that could take part in the allosteric mechanism. The most important of these are the coupling helices in the TMDs, which reside at the NBD–TMD interface [51], the Q-loop in the NBD, which is able to differentiate between bound ATP and ADP [51, 52], and the X-loop, which is in contact with the coupling helices [53, 54].

Several studies employing molecular dynamics simulations discuss the stabilizing effect of nucleotides in isolated NBD dimers [45, 55-58]. Although there is no clear indication whether hydrolysis at one site facilitates the hydrolysis at the *trans* site [58, 59], most simulation results and a half-open X-ray diffraction structure of the NBD dimer suggest post-hydrolytic conformational changes where the Q and X-loops move together [45, 55]. Due to the large computational efforts required, few studies report simulations of whole length ABC transporters in a lipid membrane. In the case of the BtuCD bacterial transporter, it has been shown that ATP binding induces conformational changes in both the NBDs and the TMDs [60], and an asymmetric closure of the ATP-bound transporter was observed, which supports the alternating sites model [61]. Simulations performed on the whole length Sav1866 bacterial transporter have highlighted certain residues in the Q and X-loops that can take up several different conformations, and can thus act as molecular switches to transmit conformational changes between the NBD and TMD [52]. Since now, however, there has

been no report of a systematic study of the full pathway of conformational changes originating from the NBD and transmitted toward the TMD.

The aim of my work was to characterize and describe the changes occuring in the pattern of correlated motions in the full transporter due to ATP hydrolysis, furthermore, to compare the collective motions present in the holo and apo states. The motion of the system was simulated by molecular dynamics, and relevant data was extracted from the trajectories. During the simulations, the whole length transporter was embedded into a lipid bilayer and a solvent composed of explicit water molecules to model the physiological environment. All molecules were modeled at an atomic level. To analyse the effect of ATP hydrolysis in the holo state, the X-ray structure of the Sav1866 bacterial transporter, co-crystallized with nucleotides, was used to create an ATP/ATP system (containing ATP at both nucleotide binding sites), and an ATP/ADP system (containing ATP at one site and ADP at the other) to model the state of the transporter before and after hydrolysis. For comparing the dynamics of the holo and apo states, homology models of the human MDR1 protein were used, which were built upon the Sav1866 (holo) and mouse MDR3 (apo) transporter structures as templates (Figure 3).

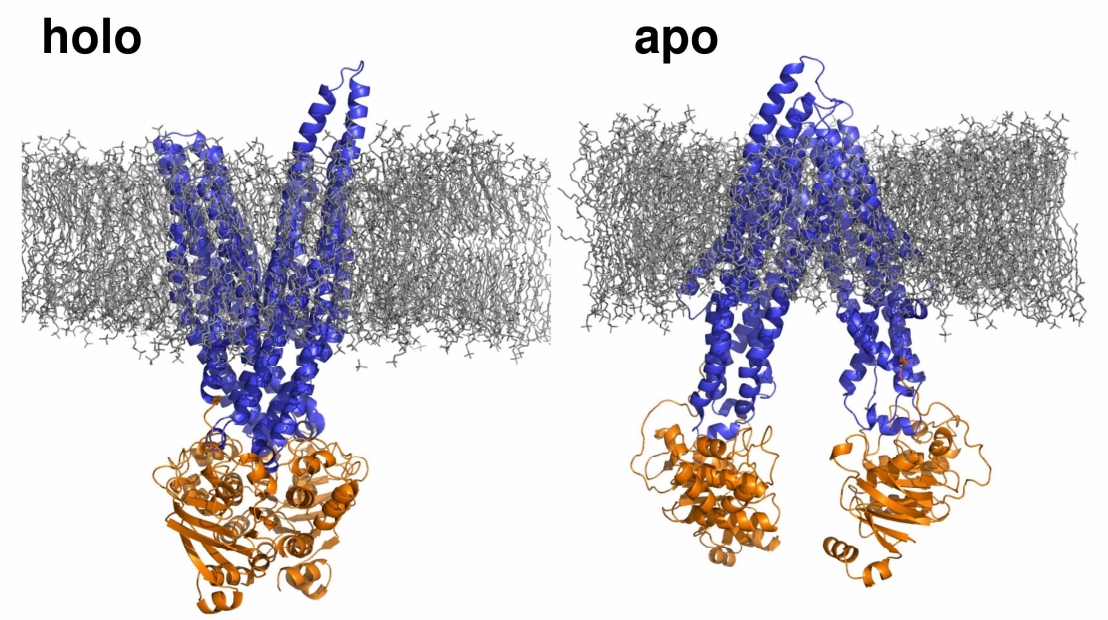

Figure 3. The two conformation of ABC exporters used in molecular dynamics simulations. The figure shows the "bottom-closed" holo (left) and the "bottom-open" apo (right) homology models of the human MDR1 protein.

3.2. Novel Scientific Results

1. The hydrolysis of a single ATP molecule in the Sav1866 transporter changes the conformational correlation patterns of residues, which could initiate the opening of the transporter at the cytoplasmic side.

To assess the differences between the dynamics of the ATP/ATP and ATP/ADP systems, the mutual information between each torsion angle pair was calculated for both systems, using five simulation trajectories, each of at least 50 ns in length, for each system. The sum of mutual information of torsion angles for each amino acid pair was used to select pairs of residues which show significant conformational correlation. The calculation of total mutual information between amino acid pairs was performed using the MutInf method by McClendon et al, which was successfully used to map the allosteric mechanism of the interleukin-2 protein [25].

The network of residue pairs that exhibit significant conformational correlation was built for both systems, which showed tightly correlated clusters of residues as well as several differences between the ATP/ATP and ATP/ADP systems. In the ATP/ATP system, correlated residues form clusters around the Gln422 and Glu473 residues, but these two clusters are practically independent of each other. The conserved Glu residue at position 473 is part of the X-loop, which plays an important role in transmitting conformational changes from the NBD to the TMD [53, 54]. Another important structural element appearing in the network and coupled to the X-loop is the tetra-helix bundle, which is formed by the cytoplasmic regions of the TM3 and TM4 transmembrane helices in the “bottom-closed” conformation.

The correlation network of the ATP/ADP system shows a simpler structure, the residues generally exhibit less coupling. The clusters of residues found in the ATP/ATP system are lost, which indicates the decoupling of the X-loop and the tetra-helix bundle. This decoupling could facilitate the dissociation of the tetra-helix bundle, which has been shown in targeted MD simulations to be an important step in the catalytic cycle [62]. The central role of the Glu473 residue is taken over by the ATP-sensing Gln422 residue in the Q-loop, which appears to be correlated with residues of the coupling helices in the ATP/ADP system. These results show that the hydrolysis of even a single ATP molecule can change the correlated motion patterns of residues in the transporter, that could facilitate the opening at the

cytoplasmic side and the transition to the apo state. Several other residues could be identified based on the correlation network (e. g. Met311, Asn494, Asn495, Asn252, Asp323) which, due to their tight dynamic coupling, could play an important role in establishing the allosteric pathway between the NBD and the TMD required for function.

2. The "bottom-open" apo structure exhibits instability in MD simulations. Based on this instability, and the collective motions observed in the holo conformation, I suggest that the "bottom-closed" apo structure is a physiologically more relevant model of the apo state of the transporters.

The attempt to compare the dynamics of the holo and "bottom-open" apo conformations failed because the apo hMDR1 model exhibited structural instability in molecular dynamics simulations. To verify that this instability was not caused by inaccuracies in the homology model, the simulation was repeated with the mouse MDR3 X-ray structure that was used as a template for model building. However, the experimentally determined structure also proved to be unstable in simulations, undergoing marked distortions similar to the homology model. The RMSD (root mean square displacement) values calculated for the C_α atoms of individual domains and the full transporter show that while the domains show less change in internal structure, their relative position changes significantly. The displacement of the domains is caused by the loss of secondary structure in the TMDs, where only 63.13% of initially helical residues stay dominantly helical, as opposed to 90.04% for the Sav1866 "bottom-closed" conformation. The most affected regions are near the interface between the TMDs and NBDs, causing the detachment and displacement of the NBDs from the cytoplasmic regions of the TMDs.

One of the possible causes of the instability could be the high amount of hydrophobic surface exposed to the solvent in the "bottom-open" structure. A comparison to the "bottom-closed" holo structure shows that the intracellular loops contain several hydrophobic residues whose solvent exposed surface is unfavorably increased in the "bottom-open" structure. The increase can be 20-30 fold for certain residues. The instability could mean that the experimentally derived structure is only stable under the conditions of crystallization, but not under native conditions. The influence of non-native conditions during crystallization is also indicated by the fact that the crystal unit cells of both "bottom-open" structures (mouse MDR3 and MsbA) contain more than one full transporter subunit, which share a significant non-native contact surface near the NBD–TMD interface.

Due to the instability of the "bottom-open" apo structure and the non-native crystal contacts I propose the "bottom-closed" apo conformation, observed in X-ray structures of the MsbA transporter, to be a better model of the apo state. In the "bottom-closed" apo structure, the NBDs display smaller separation and a sideways displacement in opposite directions, relative to the axis connecting their centers of mass. The analysis of collective motions observed in the "bottom-closed" holo conformation indicates that intrinsic motions present in the transporter allow a transition toward the "bottom-closed" apo conformation. Based on these results I suggest that the "bottom-closed" apo structure could serve as a better, physiologically more relevant model of the apo state than the "bottom-open" structure.

4. Possible Utilization and Applicability of Results

During testing of the Gaussian mixture method, systems with a small number of degrees of freedom (3–5) and a fixed number of Gaussian components (10) were used. In these cases, rapid convergence was observed and the resulting entropy values were in good agreement with calculated *ab initio* exact values. The running time of Gaussian mixture fitting and entropy estimation even in the case of the "ala-val-ala" system having the most number of degrees of freedom did not exceed 54 hours on an Intel Xeon E5430 computer. For larger systems, the size of the matrices describing individual Gaussian components increases (d^2 matrix elements for d degrees of freedom), and additionally, a larger number of Gaussian components (k) might be necessary. The scaling of the current implementation of the method is $O(nd^3k)$, where n is the number of conformations in the ensemble. Convergence in larger systems would thus be harder to reach. To reduce the number of degrees of freedom in larger systems (e. g. proteins), several methods could be applied. Using full correlation analysis (FCA, [63]), the system can be partitioned into minimally coupled subspaces [9], and subsequently, mutual information expansion could be used to further reduce the number of dimensions [13]. The entropy of the subsystems could then be easily and rapidly calculated using the Gaussian mixture method. To estimate the number of necessary Gaussian components, clustering methods could be used (e. g. fuzzy clustering) to identify compact, less coupled clusters.

My current work serves as a demonstration that the approach using Gaussian mixtures is a viable and accurate method to estimate the configurational entropy of non-diffusive molecular systems. The Gaussian mixture method could be used as a basis for the development of further methods to calculate the configurational entropy of larger systems (e. g. proteins).

The modulation and regulation of the function of ABC transporter proteins could pose a solution for the treatment of several severe human diseases. A possible method for modulation is the design of small molecule drugs that either enhance or inhibit the function of certain ABC transporters. A further fine-tuning of modulation could be achieved by identifying allosteric binding sites that affect the communication between the NBDs and TMDs. In the case of the CFTR (Cystic Fibrosis Transmembrane conductance Regulator) protein, the ΔF508 deletion resulting in cystic fibrosis causes a damaged NBD/TMD interface [54]. Small molecule chemical chaperones have been successfully used to recover the functional structure of the mutant protein, and compounds enhancing the function of the protein have also been identified [64-66]. Understanding the structure and dynamics of transporters could also help in elucidating their mechanism of action. By utilizing the allosteric effect, the function of transporters could even be manipulated from the extracellular side. The viability of this approach is demonstrated by the 5D3 antibody that is able to bind to the ABCG2 transporter from the extracellular side and sense function-related conformational changes occuring on the cytoplasmic side (e. g. ATP binding) [67, 68].

Rational drug design requires the knowledge of the correct membrane-bound structure of the transporters, and in order to interpret the effect of certain small molecule modulators, the detailed knowledge of the mechanism and dynamic properties of the target protein might also be necessary. My work is a contribution using computational methods to unravel the details of the transport cycle after ATP hydrolysis, and to aid the interpretation of atomic resolution structures derived by experimental methods.

5. List of Publications

1. Gyimesi G., Závodszky P., Hegedűs T., and Szilágyi A., “Calculation of configurational entropy from molecular dynamics trajectories using Gaussian mixtures.”, under submission

2. Gyimesi G., Ramachandran S., Kota P., Dokholyan N.V., Sarkadi B., Hegedűs T., “ATP hydrolysis at one of the two sites in ABC transporters initiates transport related conformational transitions.” *Biochimica et Biophysica Acta – Biomembranes*, **2011**. 1808(12): p. 2954-2964.

3. Papp A., Szommer T., Barna L., Gyimesi G., Ferdinándy P., Spadoni C., Darvas F., Fujita T., Ürge L., Dormán G., “Enhanced hit-to-lead process using bioanalogous lead evaluation and chemogenomics: application in designing selective matrix metalloprotease inhibitors.” *Expert Opinion on Drug Delivery*, **2007**. 2(5): p. 707-723.

4. Flachner B., Varga A., Szabó J., Barna L., Hajdú I., Gyimesi G., Závodszky P., Vas M., “Substrate-assisted movement of the catalytic Lys 215 during domain closure: site-directed mutagenesis studies of human 3-phosphoglycerate kinase.” *Biochemistry*. **2005**. 44(51): p. 16853-65.

6. Bibliography

[1] Beveridge D.L., DiCapua F.M., "Free Energy via Molecular Simulation: Applications to Chemical and Biomolecular Systems". *Annu Rev Biophys Biophys Chem*, **1989**. 18: p. 431-92.

[2] Zwanzig R.W., "High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases". *Journal of Chemical Physics*, **1954**. 22(8): p. 1420-26.

[3] Kumar S., Bouzida D., Swendsen R.H., Kollman P.A., Rosenberg J.M., "The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method". *Journal of Computational Chemistry*, **1992**. 13(8): p. 1011-1021.

[4] Meirovitch H., Cheluvaraja S., White R.P., "Methods for calculating the entropy and free energy and their application to problems involving protein flexibility and ligand binding". *Curr Protein Pept Sci*, **2009**. 10(3): p. 229-43.

[5] Karplus M., Kushick J.N., "Method for estimating the configurational entropy of macromolecules". *Macromolecules*, **1981**. 14(2): p. 325-332.

[6] Schlitter J., "Estimation of absolute and relative entropies of macromolecules using the covariance matrix". *Chem Phys Lett*, **1993**. 215(6): p. 617-621.

[7] Baron R., van Gunsteren W.F., Hünenberger P.H., "Estimating the configurational entropy from molecular dynamics simulations: anharmonicity and correlation corrections to the quasi-harmonic approximation". *Trends Phys Chem*, **2006**. 11: p. 87-122.

[8] Hnizdo V., Fedorowicz A., Singh H., Demchuk E., "Statistical thermodynamics of internal rotation in a hindering potential of mean force obtained from computer simulations". *J Comput Chem*, **2003**. 24(10): p. 1172-83.

[9] Hensen U., Grubmuller H., Lange O.F., "Adaptive anisotropic kernels for nonparametric estimation of absolute configurational entropies in high-dimensional configuration spaces". *Phys Rev E Stat Nonlin Soft Matter Phys*, **2009**. 80(1 Pt 1): p. 011913.

[10] Numata J., Wan M., Knapp E.W., "Conformational entropy of biomolecules: beyond the quasi-harmonic approximation". *Genome Inform*, **2007**. 18: p. 192-205.

[11] Wang J., Brüschweiler R., "2D Entropy of Discrete Molecular Ensembles". *J Chem Theory Comput*, **2006**. 2(1): p. 18-24.

[12] Li D.-W., Khanlarzadeh M., Wang J., Huo S., Brüschweiler R., "Evaluation of Configurational Entropy Methods from Peptide Folding−Unfolding Simulation". *J Phys Chem B*, **2007**. 111(49): p. 13807-13813.

[13] Killian B.J., Yudenfreund Kravitz J., Gilson M.K., "Extraction of configurational entropy from molecular simulations via an expansion approximation". *J Chem Phys*, **2007**. 127(2): p. 024107.

[14] Matsuda H., "Physical nature of higher-order mutual information: Intrinsic correlations and frustration". *Phys Rev E*, **2000**. 62(3): p. 3096-102.

[15] Attard P., "Information content of signals using correlation function expansions of the entropy". *Phys Rev E*, **1997**. 56(4): p. 4052-67.

[16] Reiss H., "Superposition Approximations from a Variation Principle". *Journal of Statistical Physics*, **1972**. 6(1): p. 39-47.

[17] Hnizdo V., Tan J., Killian B.J., Gilson M.K., "Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods". *J Comput Chem*, **2008**. 29(10): p. 1605-14.

[18] Hess B., Kutzner C., van der Spoel D., Lindahl E., "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation". *J Chem Theory Comput*, **2008**. 4(3): p. 435-447.

[19] Loftsgaarden D.O., Quesenberry C.P., "A Nonparametric Estimate of a Multivariate Density Function". *Annals of Mathematical Statistics*, **1965**. 36(3): p. 1049-51.

[20] Huber M.F., Bailey T., Durrant-Whyte H., Hanebeck U.D., "On entropy approximation for Gaussian mixture random vectors", in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*. **2008**. p. 181-188.

[21] Agarwal P.K., Billeter S.R., Rajagopalan P.T., Benkovic S.J., Hammes-Schiffer S., "Network of coupled promoting motions in enzyme catalysis". *Proc Natl Acad Sci U S A*, **2002**. 99(5): p. 2794-9.

[22] Benkovic S.J., Hammes-Schiffer S., "A perspective on enzyme catalysis". *Science*, **2003**. 301(5637): p. 1196-202.

[23] Lange O.F., Grubmuller H., de Groot B.L., "Molecular dynamics simulations of protein G challenge NMR-derived correlated backbone motions". *Angew Chem Int Ed Engl*, **2005**. 44(22): p. 3394-9.

[24] Lange O.F., Grubmuller H., "Generalized correlation for biomolecular dynamics". *Proteins*, **2006**. 62(4): p. 1053-61.

[25] McClendon C.L., Friedland G., Mobley D.L., Amirkhani H., Jacobson M.P., "Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles". *J Chem Theory Comput*, **2009**. 5(9): p. 2486-2502.

[26] Riordan J.R., Rommens J.M., Kerem B., Alon N., Rozmahel R., Grzelczak Z., Zielenski J., Lok S., Plavsic N., Chou J.L., et al., "Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA". *Science*, **1989**. 245(4922): p. 1066-73.

[27] Senior A.E., Gadsby D.C., "ATP hydrolysis cycles and mechanism in P-glycoprotein and CFTR". *Semin Cancer Biol*, **1997**. 8(3): p. 143-50.

[28] Wada M., Toh S., Taniguchi K., Nakamura T., Uchiumi T., Kohno K., Yoshida I., Kimura A., Sakisaka S., Adachi Y., Kuwano M., "Mutations in the canilicular multispecific organic anion transporter (cMOAT) gene, a novel ABC transporter, in patients with hyperbilirubinemia II/Dubin-Johnson syndrome". *Hum Mol Genet*, **1998**. 7(2): p. 203-7.

[29] Reis A.F., Ye W.Z., Dubois-Laforgue D., Bellanne-Chantelot C., Timsit J., Velho G., "Association of a variant in exon 31 of the sulfonylurea receptor 1 (SUR1) gene with type 2 diabetes mellitus in French Caucasians". *Hum Genet*, **2000**. 107(2): p. 138-44.

[30] Mosser J., Douar A.M., Sarde C.O., Kioschis P., Feil R., Moser H., Poustka A.M., Mandel J.L., Aubourg P., "Putative X-linked adrenoleukodystrophy gene shares unexpected homology with ABC transporters". *Nature*, **1993**. 361(6414): p. 726-30.

[31] Eckford P.D., Sharom F.J., "ABC efflux pump-based resistance to chemotherapy drugs". *Chem Rev*, **2009**. 109(7): p. 2989-3011.

[32] Gottesman M.M., Ling V., "The molecular basis of multidrug resistance in cancer: the early years of P-glycoprotein research". *FEBS Lett*, **2006**. 580(4): p. 998-1009.

[33] Szakacs G., Varadi A., Ozvegy-Laczka C., Sarkadi B., "The role of ABC transporters in drug absorption, distribution, metabolism, excretion and toxicity (ADME-Tox)". *Drug Discov Today*, **2008**. 13(9-10): p. 379-93.

[34] Hyde S.C., Emsley P., Hartshorn M.J., Mimmack M.M., Gileadi U., Pearce S.R., Gallagher M.P., Gill D.R., Hubbard R.E., Higgins C.F., "Structural model of ATP-binding proteins associated with cystic fibrosis, multidrug resistance and bacterial transport". *Nature*, **1990**. 346(6282): p. 362-5.

[35] Higgins C.F., "ABC transporters: from microorganisms to man". *Annu Rev Cell Biol*, **1992**. 8: p. 67-113.

[36] Smith P.C., Karpowich N., Millen L., Moody J.E., Rosen J., Thomas P.J., Hunt J.F., "ATP binding to the motor domain from an ABC transporter drives formation of a nucleotide sandwich dimer". *Mol Cell*, **2002**. 10(1): p. 139-49.

[37] Chen J., Lu G., Lin J., Davidson A.L., Quiocho F.A., "A tweezers-like motion of the ATP-binding cassette dimer in an ABC transport cycle". *Mol Cell*, **2003**. 12(3): p. 651-61.

[38] Kerr I.D., Jones P.M., George A.M., "Multidrug efflux pumps: the structures of prokaryotic ATP-binding cassette transporter efflux pumps and implications for our understanding of eukaryotic P-glycoproteins and homologues". *Febs J*, **2010**. 277(3): p. 550-63.

[39] Ward A., Reyes C.L., Yu J., Roth C.B., Chang G., "Flexibility in the ABC transporter MsbA: Alternating access with a twist". *Proc Natl Acad Sci U S A*, **2007**.

[40] Aller S.G., Yu J., Ward A., Weng Y., Chittaboina S., Zhuo R., Harrell P.M., Trinh Y.T., Zhang Q., Urbatsch I.L., Chang G., "Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding". *Science*, **2009**. 323(5922): p. 1718-22.

[41] Zou P., Bortolus M., McHaourab H.S., "Conformational cycle of the ABC transporter MsbA in liposomes: detailed analysis using double electron-electron resonance spectroscopy". *J Mol Biol*, **2009**. 393(3): p. 586-97.

[42] Loo T.W., Bartlett M.C., Clarke D.M., "Human P-glycoprotein is active when the two halves are clamped together in the closed conformation". *Biochem Biophys Res Commun*, **2010**.

[43] Jones P.M., O'Mara M.L., George A.M., "ABC transporters: a riddle wrapped in a mystery inside an enigma". *Trends Biochem Sci*, **2009**. 34(10): p. 520-31.

[44] Senior A.E., al-Shawi M.K., Urbatsch I.L., "The catalytic cycle of P-glycoprotein". *FEBS Lett*, **1995**. 377(3): p. 285-9.

[45] Jones P.M., George A.M., "Opening of the ADP-bound active site in the ABC transporter ATPase dimer: evidence for a constant contact, alternating sites model for the catalytic cycle". *Proteins*, **2009**. 75(2): p. 387-96.

[46] Janas E., Hofacker M., Chen M., Gompf S., van der Does C., Tampe R., "The ATP hydrolysis cycle of the nucleotide-binding domain of the mitochondrial ATP-binding cassette transporter Mdl1p". *J Biol Chem*, **2003**. 278(29): p. 26862-9.

[47] van der Does C., Tampe R., "How do ABC transporters drive transport?" *Biol Chem*, **2004**. 385(10): p. 927-33.

[48] Higgins C.F., Linton K.J., "The ATP switch model for ABC transporters". *Nat Struct Mol Biol*, **2004**. 11(10): p. 918-26.

[49] Loo T.W., Clarke D.M., "Drug-stimulated ATPase activity of human P-glycoprotein requires movement between transmembrane segments 6 and 12". *J Biol Chem*, **1997**. 272(34): p. 20986-9.

[50] Loo T.W., Bartlett M.C., Clarke D.M., "Transmembrane segment 7 of human P-glycoprotein forms part of the drug-binding pocket". *Biochem J*, **2006**. 399(2): p. 351-9.

[51] Dawson R.J., Locher K.P., "Structure of a bacterial multidrug ABC transporter". *Nature*, **2006**. 443(7108): p. 180-5.

[52] Aittoniemi J., de Wet H., Ashcroft F.M., Sansom M.S., "Asymmetric switching in a homodimeric ABC transporter: a simulation study". *PLoS Comput Biol*, **2010**. 6(4): p. e1000762.

[53] Oancea G., O'Mara M.L., Bennett W.F., Tieleman D.P., Abele R., Tampe R., "Structural arrangement of the transmission interface in the antigen ABC transport complex TAP". *Proc Natl Acad Sci U S A*, **2009**. 106(14): p. 5551-6.

[54] He L., Aleksandrov A.A., Serohijos A.W., Hegedus T., Aleksandrov L.A., Cui L., Dokholyan N.V., Riordan J.R., "Multiple membrane-cytoplasmic domain contacts in cftr mediate regulation of channel gating". *J Biol Chem*, **2008**.

[55] Jones P.M., George A.M., "Nucleotide-dependent allostery within the ABC transporter ATP-binding cassette: A computational study of the MJ0796 dimer". *J Biol Chem*, **2007**. 282(31): p. 22793-803.

[56] Jones P.M., George A.M., "Mechanism of ABC transporters: a molecular dynamics simulation of a well characterized nucleotide-binding subunit". *Proc Natl Acad Sci U S A*, **2002**. 99(20): p. 12639-44.

[57] Newstead S., Fowler P.W., Bilton P., Carpenter E.P., Sadler P.J., Campopiano D.J., Sansom M.S., Iwata S., "Insights into how nucleotide-binding domains power ABC transport". *Structure*, **2009**. 17(9): p. 1213-22.

[58] Wen P.C., Tajkhorshid E., "Dimer opening of the nucleotide binding domains of ABC transporters after ATP hydrolysis". *Biophys J*, **2008**. 95(11): p. 5100-10.

[59] Campbell J.D., Sansom M.S., "Nucleotide binding to the homodimeric MJ0796 protein: a computational study of a prokaryotic ABC transporter NBD dimer". *FEBS Lett*, **2005**. 579(19): p. 4193-9.

[60] Oloo E.O., Tieleman D.P., "Conformational transitions induced by the binding of MgATP to the vitamin B12 ATP-binding cassette (ABC) transporter BtuCD". *J Biol Chem*, **2004**. 279(43): p. 45013-9.

[61] Ivetac A., Campbell J.D., Sansom M.S., "Dynamics and Function in a Bacterial ABC Transporter: Simulation Studies of the BtuCDF System and Its Components". *Biochemistry*, **2007**. 46(10): p. 2767-78.

[62] Weng J.W., Fan K.N., Wang W.N., "The conformational transition pathway of ATP binding cassette transporter MsbA revealed by atomistic simulations". *J Biol Chem*, **2010**. 285(5): p. 3053-63.

[63] Lange O.F., Grubmuller H., "Full correlation analysis of conformational protein dynamics". *Proteins*, **2008**. 70(4): p. 1294-312.

[64] Loo T.W., Bartlett M.C., Clarke D.M., "Rescue of folding defects in ABC transporters using pharmacological chaperones". *J Bioenerg Biomembr*, **2005**. 37(6): p. 501-7.

[65] Wellhauser L., Kim Chiaw P., Pasyk S., Li C., Ramjeesingh M., Bear C.E., "A small-molecule modulator interacts directly with deltaPhe508-CFTR to modify its ATPase activity and conformational stability". *Mol Pharmacol*, **2009**. 75(6): p. 1430-8.

[66] Schultz B.D., Frizzell R.A., Bridges R.J., "Rescue of dysfunctional deltaF508-CFTR chloride channel activity by IBMX". *J Membr Biol*, **1999**. 170(1): p. 51-66.

[67] Ozvegy-Laczka C., Varady G., Koblos G., Ujhelly O., Cervenak J., Schuetz J.D., Sorrentino B.P., Koomen G.J., Varadi A., Nemet K., Sarkadi B., "Function-dependent conformational changes of the ABCG2 multidrug transporter modify its interaction with a monoclonal antibody on the cell surface". *J Biol Chem*, **2005**. 280(6): p. 4219-27.

[68] Ozvegy-Laczka C., Laczko R., Hegedus C., Litman T., Varady G., Goda K., Hegedus T., Dokholyan N.V., Sorrentino B.P., Varadi A., Sarkadi B., "Interaction with the 5D3 monoclonal antibody is regulated by intramolecular rearrangements but not by covalent dimer formation of the human ABCG2 multidrug transporter". *J Biol Chem*, **2008**. 283(38): p. 26059-70.