Internal force field in selected proteins

Magdalena Ptak-Kaczor1,2, Mateusz Banach1, Leszek Konieczny3 and Irena Roterman1✉

1Department of Bioinformatics and Telemedicine, Jagiellonian University – Medical College, Kraków, Poland; 2Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, Kraków, Poland; 3Chair of Medical Biochemistry, Jagiellonian University – Medical College, Kraków, Poland

The fuzzy oil drop model suggests that the tertiary conformation of a protein – particularly a globular one – can be likened to a spherical micelle. During the folding process, hydrophilic residues are exposed on the surface, while hydrophobic residues are retained inside the protein. The resulting hydrophobicity distribution can be mathematically modeled as a 3D Gaussian. The fuzzy oil drop model is strikingly effective in explaining the properties of type II antifreeze proteins and fast-folding proteins, as well as a vast majority of autonomous protein domains. This work aims to determine whether similar mechanisms apply to other types of nonbonding interactions. Our analysis indicates that electrostatic and van der Waals forces do not conform to the Gaussian pattern. The study involves a reference protein (titin) which shows a high agreement between the observed distribution of hydrophobicity and the theoretical (Gaussian) distribution, a selection of amyloid structures derived from the Protein Data Bank, as well as transthyretin – a protein known for its susceptibility to amyloid transformation.

Key words: hydrophobicity, amyloid, internal force field, external force field

Received: 22 August, 2019; revised: 21 September, 2019; accepted: 15 October, 2019; available on-line: 20 November, 2019

✉E-mail: myroterm@cyf-kr.edu.pl

Acknowledgements of Financial Support: This research was funded by the Jagiellonian University – Medical College system of grants – grant # N41/D/B/00211. This research was also supported in part by the PLGrid Infrastructure at ACC Cyfronet AGH.

Abbreviations: vdW, van der Waals, type of non-bonding interaction; RD, Relative Distance; OH, distribution O (observed hydrophobicity) in comparison with distribution based on intrinsic hydrophobicity (H); OE, distribution O (observed hydrophobicity) in comparison with distribution based on electrostatic interaction (E); OV, distribution O (observed hydrophobicity) in comparison with distribution based on vdW interaction (V); 3D, three-dimensional; PDB, Protein Data Bank; DKL, Kullback-Leibler divergence entropy

INTRODUCTION

The protein force field includes various types of interactions: electrostatic, vdW and torsion potentials (Jo et al., 2017; Vanommeslaeghe et al., 2010; Brooks et al., 2009; Hynninen & Crowly, 2018; Pavlova et al., 2018; Boonstra et al., 2016; Mirijanian et al., 2014; Miller et al., 2008; Yu et al., 2012; Best et al., 2012; Patel & Brooks, 2004; Baker et al., 2010; Hornak et al., 2006; Wang et al., 2017; Shao & Zhu 2018; Hoffmann et al., 2018; Wang et al., 2004; Wroblewska & Skolnick, 2007; Liu & Bevridge, 2002; Vymetál et al., 2019; Skjevik et al., 2015; Chen et al., 2015; Malolepsza et al., 2010; Rubenstein et al., 2018; van der Spoel et al., 2005; Lundborg & Lindahl, 2015; Aragones et al., 2013, van der Spoel et al., 2012; Bjelkmar et al., 2010; Abraham & Gready 2011; Hess et al., 2008). In addition, most force fields recognize the effects of hydrogen bonds, while some take into account structural flexibility, where the bond length and dihedral angles may change as a result of stretching and shearing vibrations, respectively. An overview of the force fields used in simulation toolkits such as Amber, Charmm and ECEPP can be found in (Roterman et al., 1989a; Roterman et al., 1989b; Lundborg & Lindahl, 2015; Guvench & MacKerell, 2008).

In addition to the above, some force fields also include a separate component related to the presence of disulfides (Roterman et al., 1989a).

The procedure for obtaining a 3D structure for a given input sequence relies on optimization methods that minimize internal energy (Jo et al., 2017; Vanommeslaeghe et al., 2010; Brooks et al., 2009; Hynninen & Crowly, 2018; Pavlova et al., 2018; Boonstra et al., 2016; Mirijanian et al., 2014; Miller et al., 2008; Yu et al., 2012; Best et al., 2012; Patel & Brooks, 2004; Baker et al., 2010; Hornak et al., 2006; Wang et al., 2017; Shao & Zhu 2018; Hoffmann et al., 2018; Wang et al., 2004; Wroblewska & Skolnick, 2007; Liu & Bevridge, 2002; Vymetál et al., 2019; Skjevik et al., 2015; Chen et al., 2015; Malolepsza et al., 2010; Rubenstein et al., 2018; van der Spoel et al., 2005; Lundborg & Lindahl, 2015; Aragones et al., 2013, van der Spoel et al., 2012; Bjelkmar et al., 2010; Abraham & Gready 2011; Hess et al., 2008).

Textbook knowledge says that the tertiary structure stability is ensured by the presence of a hydrophobic core, along with a system of disulfide bonds (Devlin, 2010). The latter is evolutionarily conditioned, while the former owes its properties to interactions between the protein chain and its aqueous environment (Dygyt et al., 2016; Konieczny et al., 2006; Kalinowska et al., 2015; Roterman et al., 2011; Banach et al., 2012; Banach et al., 2018). Optimization of such interactions involves the exposure of hydrophilic residues to the surface and internalization of hydrophobic residues, ensuring entropically beneficial contact with the solvent. We may therefore consider the protein (particularly a globular one) as a quasi-spherical micelle. Various amino acid sequences have different propensities for generating micellar structures. In particular, fast-folding proteins (Roterman et al., 2011), type II antifreeze proteins (Banach et al., 2012; Banach et al., 2018) and a vast majority of protein domains (treated as stand-alone structural units) exhibit strong accordance between the observed distribution of hydrophobicity and the corresponding model distribution, theorized for a perfect spherical micelle (Salapa et al., 2012). The latter “idealized” (or “theoretical”) distribution can be mathematically modeled as a 3D Gaussian (Konieczny et al., 2006). The scope and extent of deviations from this distribution are the characteristic
properties of each protein and are often associated with its biological function. Discordance may take on one of two guises: a local excess of hydrophobicity on the surface and a local deficiency of hydrophobicity. The latter typically indicates the presence of a binding cavity, capable of housing a ligand or substrate (Prymula et al., 2011), while the former suggests a potential complexation site (Brylinski et al., 2007). The targeted nature of discrepancies between the observed and theoretical distributions of hydrophobicity ensures high specificity of molecular interactions.

Unlike internal energy optimization algorithms, which do not involve any preconditions regarding the structure of the polypeptide, the fuzzy oil drop model assumes that the folding process is driven by the presence of the aqueous solvent, which promotes micellization of the residue chain (Gadzała et al., 2019).

The phenomenon of misfolding – particularly in the context of amyloid proteins – has attracted much scientific attention over the years. Notably, amyloid structures exhibit distinct hydrophobicity distribution patterns than globular proteins. In amyloids, this distribution is linear with alternating bands of high and low hydrophobicity stretching along the fibril’s axis (Roterman et al., 2017; Roterman et al., 2016; Dulak et al., 2018). Consequently, the amyloid may be described as a ribbonlike micelle.

This work aims to determine whether the distribution of hydrophobicity in polypeptide chains is mirrored by the distribution of other nonbonding interactions. To answer this question, electrostatic and vdW interactions were subjected to statistical analysis, similar to the one performed for hydrophobicity. The observed distributions were compared with two reference models: the theoretical Gaussian distribution (denoted T) and the uniform distribution, where each residue was ascribed the same value of a given interaction (denoted R). Relative Distance (RD) was then calculated as a quantitative measure of deviations from either of these distributions, with values lower than 0.5 indicating that the observed distribution reflected T, while values higher than 0.5 showing stronger affinity for R.

The process of performing the above computations has been detailed in numerous publications devoted to the applicability of the fuzzy oil drop model to various classes of proteins, including amyloids (Kalinowska et al., 2015).

If our analysis reveals that nonbonding interactions are distributed in a similar way to hydrophobicity, we can conclude that internal energy optimization algorithms should recognize and reflect this property of the internal force field.

MATERIALS AND METHODS

Data. Our analysis involves one protein highly consistent with the 3D Gaussian distribution (and therefore with the fuzzy oil drop model), as well as the amyloids listed in the Protein Data Bank (Berman et al., 2000). Table 1 presents the list of proteins selected for our study.

The set includes titin as a representative of a class of strongly accordant proteins (Banach et al., 2014a; Banach et al., 2014b), along with several amyloids whose structures can be found in PDB. The latter facilitate the analysis of the distribution of hydrophobic, electrostatic and vdW interactions in amyloids (note that the distribution of hydrophobicity in amyloids is extensively discussed in several other publications: (Dulak et al., 2018a; Dulak et al., 2018b)). Transthyretin was selected for this study due to its strong propensity for amyloid transformation (Lim et al., 2016).

Fuzzy oil drop model. A detailed presentation of the fuzzy oil drop model can be found in numerous publications (Konieczny et al., 2006; Kalinowska et al., 2015; Roterman et al., 2011). Therefore, we will limit ourselves to a brief recapitulation of its core concepts relevant to the presented results.

The observed distribution of hydrophobicity in a protein molecule is determined by the interactions between its constituent residues. These interactions depend both on the intrinsic hydrophobicity of each residue and on the separation between neighboring residues. Our calculation is based on the function proposed in (Levitt, 1976). Having obtained the observed distribution (O), we may compare it with the corresponding theoretical distribution (T), where each residue is assigned a hydrophobicity value (T) equal to the value of the 3D Gaussian at its location (which is assumed to correspond to the so-called effective atom, i.e. averaged-out values of the positions of all atoms comprising the given residue).

The 3D Gaussian is obtained by encapsulating the protein body in a suitably shaped ellipsoid capsule, whose dimensions are expressed by a trio of sigma coefficients (denoted σ, σ and σ).

Normalization of T and O allows direct comparison of both distributions. This is done by employing the divergence entropy parameter proposed by Kullback and Leibler (Kullback & Leibler, 1951). This parameter, denoted DKL, expresses the “distance” between two distributions: observed (O) and idealized (T), of which the latter is treated as a reference distribution. As a result, T describes a “perfect” globular protein, in which hydrophobicity peaks at the center and remains low on the surface. The value of DKL cannot, however, be interpreted on its own, since it is a measure of entropy. This is why another reference distribution is necessary to facilitate meaningful comparisons. This second reference distribution – called R – is defined as a “polar opposite” of T. Under R, each residue is ascribed a value of 1/N (N being the total number of residues in the chain). Thus, R describes a situation where no concentration of hydrophobicity exists at any point within the protein body. Calculating DKL for both O-T and O-R tells us whether O approximates T or R more closely, with lower values indicating greater similarity. To avoid having to deal with two separate values, a Relative Distance (RD) parameter is derived as the relation of DKL (O-T) to the sum of DKL (O-T) and DKL (O-R). RD < 0.5 indicates the presence of a hydrophobic core.

RD is calculated as follows:

$$RD = \frac{D_{KL}(O-T)}{D_{KL}(O-T) + D_{KL}(O-R)}$$

Given that numerous proteins exhibit strong accordance between the observed distribution of hydropho-

Table 1. List of analyzed proteins.

<table>
<thead>
<tr>
<th>Protein</th>
<th>PDB ID</th>
<th>Chain length</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titin</td>
<td>1TIF</td>
<td>89 aa</td>
<td>(Improta et al., 1996)</td>
</tr>
<tr>
<td>Amyloid Aβ(1-40)</td>
<td>2MVX</td>
<td>40 aa</td>
<td>(Schütze et al., 2015)</td>
</tr>
<tr>
<td>Aβ(1-40)</td>
<td>2MPZ</td>
<td>26 aa</td>
<td>(Spourakis et al., 2015)</td>
</tr>
<tr>
<td>Aβ(11-22)</td>
<td>2MXU</td>
<td>32 aa</td>
<td>(Xiao et al., 2015)</td>
</tr>
<tr>
<td>Tau</td>
<td>5O3L</td>
<td>73 aa</td>
<td>(Fitzpatrick et al., 2017)</td>
</tr>
<tr>
<td>α-synuclein</td>
<td>2N0A</td>
<td>140 aa</td>
<td>(Tuttle et al., 2016)</td>
</tr>
<tr>
<td>Transthyretin</td>
<td>1DVQ</td>
<td>115 aa</td>
<td>(Klabunde et al., 2000)</td>
</tr>
</tbody>
</table>
bicity and the theoretical Gaussian, it is interesting to speculate whether other nonbonding interactions (particularly electrostatic and vdW potentials as computed by Gromacs (van der Spoel et al., 2005)) follow a similar pattern. Consequently, we have computed $E$ and $V$ values for each residue, expressing (respectively) their electrostatic and van der Waals potentials. These observed values were subsequently compared with $T$ to determine whether electrostatic and vdW interactions follow a centralized pattern, similar to hydrophobic interactions.

The Gromacs force field – electrostatic and vdW potentials. The distribution of observed electrostatic and vdW potentials was performed using Gromacs (van der Spoel et al., 2005) running on computing resources supplied by Cyfronet AGH in the PL-Grid infrastructure (http://plgrid.pl/ (August 2019)). The selected molecule, whose structure had been derived

Table 2. RD parameters values for each interaction.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Electrostatic</th>
<th>vdW</th>
<th>Hydrophobic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1TIT</td>
<td>0.839</td>
<td>0.807</td>
<td>0.382</td>
</tr>
<tr>
<td>2MPZ</td>
<td>0.749</td>
<td>0.809</td>
<td>0.541</td>
</tr>
<tr>
<td>2MXU</td>
<td>0.806</td>
<td>0.757</td>
<td>0.513</td>
</tr>
<tr>
<td>2MVX</td>
<td>0.858</td>
<td>0.826</td>
<td>0.608</td>
</tr>
<tr>
<td>2NOA*</td>
<td>0.870</td>
<td>0.789</td>
<td>0.506</td>
</tr>
<tr>
<td>503L</td>
<td>0.815</td>
<td>0.767</td>
<td>0.664</td>
</tr>
<tr>
<td>1DVQ**</td>
<td>0.850</td>
<td>0.864</td>
<td>0.652</td>
</tr>
</tbody>
</table>

*analysis restricted to the 30–100 fragment, **transthyretin. The values describe the status of the chains treated as part of the complex.

Figure 1. Distribution profiles for 1TIT (reference protein characterized by a low RD value for hydrophobic interactions): A – electrostatic (OE), B – hydrophobicity (OH), C – vdW (OV).

Figure 2. Distribution profiles for amyloid proteins: Aβ(15-40) with D23N mutation (2MPZ). A – electrostatic (OE), B – hydrophobicity (OH), C – vdW (OV).
from PDB (Berman et al., 2000), was subjected to energy minimization in order to eliminate steric clashes introduced by incorporation of H atoms based on X-ray diffraction studies.

Following optimization using Gromacs software (specifically, Gromacs 5.0.7 with the OPLS-AA/L all-atom force field), we computed the strength of electrostatic and vdW interactions between each residue and its respective neighborhood. This yielded a distribution of both types of interactions throughout the protein body, mathematically expressed as a set of $E_i$ and $V_i$ values, which could be directly compared to our reference distributions (T and R) given by the fuzzy oil drop model (Kalinowska et al., 2015).

In our previous publications, RD was used to express the relative distance between the observed distribution of hydrophobicity (O) and two reference distributions (T and R – as described above). Here, instead of hydrophobicity, $O$ denotes the distribution of two other types of interactions (E and V – electrostatic and vdW respectively). If the computed value of RD is greater than 0.5, we conclude that the given interaction follows the uniform pattern rather than the monocentric core pattern.

**Computing the status of individual residues.** In order to compare the status of individual amino acids with the corresponding reference distributions, we need to be aware of the interactions between each residue and its neighbors. Having measured these interactions, we may calculate divergence entropy (Kullback & Leibler, 1951) to determine the scope of alignment between the observed and reference distributions. The corresponding procedure (utilizing the

**Figure 3. Distribution profiles for amyloid proteins: Aβ(1-40) (2MVX).**
A – electrostatic (OE), B – hydrophobicity (OH), C – vdW (OV).

**Figure 4. Distribution profiles for amyloid proteins: Aβ(11-42) with (E22Δ mutation) (2MXU).**
A – electrostatic (OE), B – hydrophobicity (OH), C – vdW (OV).
RESULTS

Table 2 presents the Relative Distance (RD) values for T-O-R models for electrostatic interactions (OE), OV – vdW interactions, and OH expressing hydrophobic interactions. In each case, T (Gaussian) and R distributions were used as references.

Figure 1 provides a summary of the results of the T, OH, OE and OV distribution profiles for titin. The green horizontal line represents the R distribution. The blue profile represents the idealized (T) distribution given by the 3D Gaussian. The red line represents the O distribution for the analyzed interaction. The observed profile of hydrophobic interactions is clearly similar to the theoretical model. The RD value for titin is very low at 0.382 (Table 1). Yet, major differences between OH and the other two observed distributions can be seen. The nonbonding interactions appear to be more closely aligned with the uniform distribution (R), and therefore evenly distributed over the entire chain length, with no noticeable peaks. The red line is aligned with the green line in both Fig. 1A and Fig. 1C, and the corresponding values of RD are very high: 0.839 for OE and 0.807 for OV. These high values reveal strong alignment of both OE and OV with R. Based on these observations, we may conclude that a hydrophobic core is clearly present in titin, while neither OE nor OV reveal any “core-like” concentration. In particular, the lack of any electrostatic and vdW concentration is observed in the central part of the molecule, despite close packing of atoms and residues at that location.

Figure 5. Distribution profiles for amyloid proteins: α-synuclein (2N0A). A – electrostatic (OE), B – hydrophobicity (OH), C – vdW (OV).

Figure 6. Distribution profiles for amyloid proteins: tau (5O3L). A – electrostatic (OE), B – hydrophobicity (OH), C – vdW (OV).
A similar situation occurs in amyloids, where both electrostatic and vDW distributions closely match R. While in terms of hydrophobic interactions, amyloids are very different from globular proteins (such as titin), similar differences in the scope of electrostatic and vDW distributions cannot be observed (Figs 2–6).

Regarding transthyretin – its distribution of hydrophobicity may be regarded as an intermediate case. The protein contains a centralized hydrophobic core but also exhibits major deviations from the Gaussian pattern. In contrast, electrostatic and vDW distributions are similar to those observed in other analyzed proteins and largely consistent with the uniform pattern (Fig. 7).

Our analysis underscores the distinction between globular proteins and amyloids. Previous publications highlighted differences related to distributions of hydrophobicity, identifying areas where the observed distribution opposes the theoretical model (Roterman et al., 2016; Lim et al., 2016; Dułak et al., 2019). These fragments are regarded as “amyloid seeds” – however, in the context of electrostatic and vDW interactions, they have no special properties compared to the rest of the amyloid chain.

**DISCUSSION AND CONCLUSIONS**

The fuzzy oil drop model treats amyloids as ribbon-like micelles (in contrast to spherical micelles, which represent globular proteins) (Kalinowska et al., 2017). The presence of short fragments, whose local distribution of hydrophobicity opposes the theoretical model, results from the systemic inability of the amyloid peptide to attain a spherical conformation (Dułak et al., 2018a; Dułak et al., 2018b). This locally suboptimal distribution is compensated by the isolation of discordant fragments in the structure of a ribbon-like micelle. Rapid growth of amyloid fibrils is therefore linked to the overriding need to shield hydrophobic residues by attracting and attaching additional peptides. This process results in the formation of a peculiar type of hydrophobic core in the form of a centrally located hydrophobic band.

No similar mechanism can be identified in the scope of electrostatic and vDW interactions (Marchewka et al., 2011). Therefore, protein structure prediction algorithms which rely solely on the composition of the input sequence are correct in applying the isotropic model to these interactions. Anisotropic distribution appears to occur only in relation to hydrophobic forces, resulting in isolation of hydrophobic residues in the interior of a globular micelle (or in the central part of an amyloid ribbon). Despite a relatively low number of globular proteins subjected to analysis (see: Marchewka et al., 2011; Roterman, 2019) for additional examples), we may conclude that the properties of electrostatic and vDW interactions in amyloids are not significantly different from those observed in globular proteins.

The practical conclusion for protein folding simulations is that non-directed optimization can be applied for energy minimization expressing electrostatic and vDW interactions, while directed optimization is better suited for hydrophobic interactions which tend to produce a hydrophobic core.

**Acknowledgements**

The authors wish to thank Piotr Nowakowski for editorial assistance.

**REFERENCES**


Figure 7. Distribution profiles for transthyretin (1DVQ), a protein known for its propensity for amyloid transformation. A – electrostatic (OH), B – hydrophobicity (OE), C – vDW (OV).


