Flexibility analysis of biomacromolecules with application to computer-aided drug design

Simone Fulle, Holger Gohlke[*]

Institute of Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich-Heine-University, Universitätsstr. 1, 40225 Düsseldorf

Running title head: Flexibility analysis of biomacromolecules

*Universitätsstr. 1, 40225 Düsseldorf, Germany. Phone: (+49) 211 81-13662. Fax: (+49) 211 81-13847. E-mail: gohlke@uni-duesseldorf.de.

Summary

Flexibility characteristics of biomacromolecules can be efficiently determined down to the atomic level by a graph-theoretical technique as implemented in the FIRST (*Floppy Inclusion and Rigid Substructure Topology)* and ProFlex software packages. The method has been successfully applied to a series of protein and nucleic acid structures. Here, we describe practical guidelines for setting up and performing a flexibility analysis, discuss current bottlenecks of the approach, and provide sample applications as to how this technique can support computer-aided drug design approaches.

Key words: Flexibility/rigidity analysis, FIRST, ProFlex, statics of biomacromolecules, rigidity theory, constraint counting

1. Introduction

Biomacromolecules are inherently flexible and can undergo functionally relevant conformational changes; these changes occur on a wide range of different amplitudes and time scales. The ability to undergo conformational transitions becomes particularly pronounced in the case of ligand binding to several pharmacologically important protein or RNA structures **(*1*)**, with prominent examples being HIV-1 protease **(*2*)** or HIV-1 TAR RNA **(*3*)**. From an experimental perspective, main sources of information about dynamics of biomacromolecules are crystallographic B-values, atomic fluctuations derived from NMR structural ensembles, NMR relaxation measurements, residual dipolar couplings, and H/D exchange experiments **(*4, 5*)**. From a theoretical or computational perspective, characterizing the dynamics of proteins or nucleic acids is still challenging.

Here, we present concepts from rigidity theory that allow to obtain detailed insights into the intrinsic flexibility characteristics of biomacromolecules in a very efficient manner **(*6*)**. For this, constraint counting is applied to a topological network representation of the biomacromolecule. In the network, vertices represent atoms, and edges represent covalent and non-covalent constraints (Figure 1). Based on the accessibility of rotational degrees of freedom, each bond is identified as either flexible or rigid. Furthermore, the molecule is decomposed into rigid regions and flexible parts in between them. Rigid regions are those parts of a molecule that have a well-defined equilibrium structure and move as a rigid body with six degrees of freedom. Thus, no internal motion is allowed within a rigid region. In turn, flexible regions are hinge regions of the molecule where bond-rotational motions can occur without a high cost of energy.

The approach has been implemented into the FIRST (*Floppy Inclusion and Rigid Substructure Topology*) **(*6*)** and ProFlex **(*6, 7*)** software packages and has been thoroughly validated to identify rigid clusters and collectively moving regions in protein **(*6*)** and RNA structures **(*8*)**. There are ample possibilities of applying flexibility analysis in structure-based drug design, such as for docking or virtual screening approaches; these will be detailed below in the 'Notes section'. Another noteworthy application of flexibility analysis is data-driven protein engineering by identifying structural features that impact protein thermostability **(*9, 10*)** and/or investigating the influence of mutations on protein flexibility and stability **(*9, 11*)**. That way experiments can be guided that aim at optimizing thermostability of proteins and/or improving enzyme activity **(*9, 12*)**. Furthermore, the approach has been successfully used to determine the change in protein flexibility upon complex formation **(*11, 13*)**, to probe the principle of corresponding states on protein structures from mesophilic and thermophilic organisms **(*9, 12*)**, to compare the pattern of flexibility gain during unfolding across different protein families **(*14-16*)**, and to obtain insights into the functional role of the ribosomal exit tunnel structure **(*17*)**. The approach usually takes a few seconds on proteins of hundreds or thousands of residues **(*18*)** so that it can be efficiently applied to large macromolecules, such as a virus capsid **(*19*)** or the ribosomal complex **(*17, 20*)**, too. Recent versions of the program are available for download or interactive use via the FlexWeb site at http://flexweb.asu.edu/ or the ProFlex site at http://www.bch.msu.edu/~kuhn/software/proflex.

--- Figure 1 ---

2. Methods

In the following, we will first outline the concepts of flexibility analysis based on a topological network representation of a biomacromolecule. We will then describe the individual steps for preparing an input structure, performing a flexibility analysis, and visualizing the results.

2.1. Flexibility analysis based on a topological network representation

1. *Constraint counting*

For understanding the influence of covalent and non-covalent constraints on the flexibility of biomacromolecules consider the following. In 3-space, a structure consisting of n atoms has $3n$ degrees of freedom, six of which describe rotational and translational rigid body motions. The flexibility of the structure is determined by the number of independent internal degrees of freedom *dof*, which is given by subtracting six global degrees of freedom and the number of independent constraints C from the overall number of degrees of freedom (Eq. 1). Thus, with very many (few) constraints present, the biomacromolecule is largely rigid (flexible).

$$dof = 3n - 6 - C \qquad \text{Eq. 1}$$

2. *Treatment of non-covalent constraints*

As the flexibility of biomacromolecules is largely determined by non-covalent interactions, the outcome of a flexibility analysis is mainly governed by the way hydrogen bonds (including salt bridges) and hydrophobic interactions are modeled in the network (Figure 1 (2.)). In general, *hydrogen bonds* are included depending on their geometry and interaction energy. For this, potential hydrogen bonds are ranked according to an energy function that takes into account the hybridization state of donor and acceptor atoms as well as their mutual orientation **(*6*)**. By tuning the energy threshold E_{HB} strong hydrogen bonds can be distinguished from weaker ones. Choosing E_{HB} = -0.6 kcal/mol corresponds to the thermal energy at room temperature and so provides a natural choice **(*6*)**. Choosing E_{HB} = -1.0 kcal/mol has also been reported in the literature **(*21, 22*)** and is currently the default energy cutoff for protein and nucleic acid structures in FIRST. (Note that the default energy cutoff E_{HB} = -0.1 kcal/mol in ProFlex.)

Rather than analyzing a biomacromolecule at a preset E_{HB} value, one can also simulate a thermal unfolding of the underlying topological network representation of the biomacromolecule by successively removing hydrogen bonds in the order of increasing strength. Monitoring the decay of the network by the so-called cluster configuration entropy (Eq. 2) then allows to identify pronounced structural events during the protein unfolding process:

$$H = -\sum_{s} w_s \ln w_s \,, \qquad \text{Eq. 2}$$

where w_s is the probability that an arbitrarily occupied site in the network belongs to a cluster of size s (***23***) or s^2 (***9***). This approach is useful if one aims at investigating changes in the network that are required for a transition to occur between a structurally stable state, where a rigid core is still present within the structure, and a largely flexible state, where this core has ceased to exist (Figure 1 (3.)).

Hydrophobic interactions are considered between pairs of carbon and/or sulfur atoms if the distance between the atoms is smaller than the sum of the van der Waals radii (1.7 Å for carbon, 1.8 Å for sulfur) plus a variable threshold D_{HC}. In most studies, D_{HC} is set to 0.25 (0.15) Å in the case of protein (***9, 12, 18***) (RNA (***17, 24***)) structures.

2.2. Preparing an input structure

2.2.3 Selecting an input structure

A structure in PDB format is required as input for the flexibility analysis, as, e.g., obtained from the Protein Database (PDB), Nucleic Acid Database (NDB), or generated by homology modeling.

1. **X-ray structures** with high resolution allow for the most consistent flexibility characterization. We recommend using X-ray structures resolved to < 2.5 Å. Structures

with resolution > 3.0 Å usually do not allow modeling the underlying constraint network appropriately and should be regarded with care.

2. **NMR structures** are often deposited as ensembles of models that agree with the experimental restraints. In those cases, we recommend to take either the first structure of the ensemble or to cluster all structures of the ensemble and choose the structure closest to the centroid of the largest cluster. With the latter approach, a structure that best represents the ensemble is identified. Many methods are available for clustering, among them the Multiscale Modeling Tools available at http://mmtsb.org/. NMR structures do not provide information about solvation and ion binding properties of the structure and should therefore only be chosen when no X-ray data are available.
3. **Homology Models:** When no experimental structures are available, one is tempted to use molecular modeling techniques to build a structure that can subsequently be used for flexibility analysis. Since the quality of such model-built structures may be low, special care has to be taken in preparing the structure and analyzing the results.
4. In all cases, the quality of the input structure should be checked with the help of the PDBREPORT database **(*25*)**, and no flexibility analysis should be performed on structures labeled “bad”. In the case of statically disordered residues, where two or more conformations are present in the PDB file, only atoms of one conformation should be kept.

See Note 3.1 for comments on the sensitivity of the flexibility analysis to the input structure.

2.2.4 Adding hydrogen atoms and assigning protonation states

In the case of X-ray structures or homology models, missing hydrogen atoms have to be added. This can be done using the WhatIf program **(*26*)**, the REDUCE program **(*27*)**, or the *leap* program from the Amber package **(*28*)**. In addition, for building a proper hydrogen bond network, the orientation of Asn, Gln, and His side chains might have to be corrected; this can be done either with the help of the WhatIf or REDUCE programs or manually. Finally, the protonation states of Asp, Glu, His, Lys, and Arg have to be defined, e.g., either with the help of the H++ webserver **(*29*)** or manually based on an inspection of the molecular environment/hydrogen bond network these sidechains are embedded in.

2.2.5 Treating ions and water molecules

Metal ions should be retained when they are part of the structure. Especially, interactions with divalent ions such as Mg^{2+} are known to affect the conformational flexibility of RNA structures **(*30*)** and should be considered in the flexibility analysis, together with surrounding water molecules when available. Interactions mediated by other structural water molecules, buffer ions, substrates, or cofactor molecules should not be included unless their influence on the flexibility of the biomacromolecule is to be probed; accordingly, these species should be removed from the structure. Unfortunately, water molecules and ions may be wrongly assigned when interpreting the electron density **(*31*)**. Thus, we recommend evaluating this experimental information critically if one wishes to include these species in the flexibility analysis (see Note 3.2). While interactions between water molecules or buffer ions and the biomacromolecule can be modeled as non-covalent bonds in the topological network representation (see below), interactions between metal ions and the biomacromolecule can be modeled as covalent bonds by inserting them manually into the constraint network **(*12*)**.

2.2.6 Treating ligands

Depending on the aim of the flexibility analysis, a ligand molecule can either be included or excluded from the topological network representation. This can be used for computing changes in the receptor flexibility upon ligand binding, which may provide a structural explanation for observed changes in entropy **(*11, 32*)**. If the ligand is included in the flexibility analysis care should be taken to assign appropriate protonation states to the ligand's functional groups.

2.3. Performing a flexibility analysis

1. *FIRST software*

The FIRST software handles protein, RNA, and DNA structures as well as ligands found in PDB entries. As for non-standard nucleosides in tRNA and rRNA, the software can cope with the most commonly occurring modifications of nucleosides such as pseudouridine, where the C5 of uracil is covalently attached to the sugar C1', and methylation of the 2'O position of the ribose sugar. In addition, methylated bases are generally considered if the methyl-carbon atom matches one of the following names: CM1, CM2, CM5, CM7, C5M or C10. See Note 3.3 for further comments on performing a flexibility analysis on RNA and DNA structures.

The FIRST software provides many command-line options for interfering with data input and output, and the program flow. For a detailed discussion, the reader is referred to the program's manual. The three most important options are related to the definition of non-covalent constraints for the topological network representation. For the latest FIRST version (v6.2), these are:

- The energy cutoff for hydrogen bonds E_{HB} can be set via the command line option '-E'. In general, we recommend using the default E_{HB} = -1.0 kcal/mol. As an alternative, a "dilution" of the hydrogen bond network and, hence, a thermal unfolding of the biomacromolecule can be simulated via the option '-dil1'.
- There are three options available for identifying hydrophobic constraints, which can be defined by the command line flag '–H'. We recommend choosing '–H 1', which applies the most commonly used threshold for hydrophobic contacts D_{HC} = 0.25 (0.15) Å for

protein (***9, 12, 18***) (RNA (***17, 24***)) structures, but no additional restrictions. In contrast, the default option for identifying hydrophobic contacts in FIRST is '-H 3', where D_{HC} is set to 0.50 Å (***18***). Furthermore, in this case, a hydrophobic constraint is only included into the network if I) both atoms of the pair are bonded to carbons, sulfurs, or hydrogens (as an indication of a hydrophobic environment) and II) a given atom does not already form a contact with another atom of the residue under consideration.

In summary, a typical FIRST v6.2 run for an input structure myPDB.pdb can be started with

```
:\> FIRST myPDB.pdb –E -1.0 –H 1
```

2. *FlexWeb webserver*

A webserver for flexibility analysis based on the FIRST software is available for public use at http://flexweb.asu.edu. The webserver prompts the user to submit the structure in a PDB format. Hydrogen atoms are added automatically using the REDUCE program (***27***). The user can modify the energy threshold E_{HB}. After the calculation, the results can be investigated on the webpage or downloaded for further analysis.

3. *ProFlex software*

A further implementation of the constraint counting algorithm is provided in the ProFlex software, which is available at http://www.bch.msu.edu/~kuhn/software/proflex. Although small differences in modeling hydrogen bonds and hydrophobic constraints in the topological network representation exist as compared to FIRST and FlexWeb, ProFlex also captures the essential conformational flexibility of proteins. Using a protonated PDB structure `myPDB_wiH.pdb`, a typical ProFlex run is started by

```
:\> PROFLEX –h myPDB_wiH.pdb –e-1.0
```

where

'–e' denotes the energy threshold E_{HB} for hydrogen bonds and

'–h' must be used in the case of a PDB file having hydrogens.

Again, a "dilution" of the hydrogen bond network and, hence, a thermal unfolding of the biomacromolecule can be simulated via the option '-nonh'.

Note that in the current implementation of ProFlex, a hydrophobic constraint between two carbon or sulfur atoms is included into the network I) using a distance threshold D_{HC} = 0.50 Å and II) if both atoms are bonded to carbons, sulfurs, or hydrogens. This corresponds to the flag '-H 2' in the FIRST software.

4. *Generating the topological network representation using Amber*

The topological network representation of a biomacromolecule can also be generated using the ambpdb program of the Amber suite (http://www.ambermd.org) (***28***). This is particularly convenient if snapshots from a molecular dynamics (MD) simulation are available in the "Amber restart file" format, such as to perform flexibility analysis on an MD ensemble of structures. Ambpdb converts a restart file into a FIRSTdataset file, which is essentially a PDB file augmented by information about covalent and non-covalent bonds. The resulting topological network representation is almost identical to the one generated by FIRST if '-H 1' is specified and no energy cutoff for hydrogen bonds is considered. In addition to the restart file, ambpdb requires an "Amber prmtop file" that contains information about the topology of the biomacromolecule. The FIRSTdataset file is generated by

```
:\> ambpdb -first -p myPDB.prmtop < myPDB.restart >
myPDB_FIRSTdataset
```

The prmtop file can be generated using the program xleap of the Amber suite and a PDB file as input. As an advantage over applying FIRST or FlexWeb directly to a PDB file, the xleap/ambpdb route allows to also consider ligands that have not yet been deposited in the PDB database. The resulting network representation can serve as input to the FIRST software. For this, use the file ending with "_FIRSTdataset" and run FIRST via:

```
:\> FIRST myPDB_FIRSTdataset –E -1.0
```

2.4. Analyzing and visualizing the results

The outcome of a flexibility analysis of a biomacromolecule can be analyzed at different levels of detail. First, rigid cluster decompositions provide hints about movements of structural parts as rigid bodies; second, flexibility characteristics at the bond level are instructive for analyzing, e.g., binding site regions; finally, flexibility characteristics of larger regions can be related to potential global movements. That way, static properties of a biomacromolecule can be linked to biological function and/or be used to support computer aided drug-design. See Note 3.4 for comments on comparing results from a flexibility analysis to data from experiments.

1. *Rigid cluster decomposition*

   A decomposition of the topological network into rigid clusters (and flexible regions in between) is calculated by both, the FIRST and ProFlex software. With the help of a Pymol script generated by the programs, each rigid cluster can be visualized as a uniformly colored body (Figure 1 (3.)). That way, regions of the biomacromolecule that are expected to have a well-defined equilibrium structure (rigid clusters) can be distinguished from flexible regions where bond-rotational motions can occur without a high cost of energy.

2. *Flexibility index*

While the decomposition into rigid clusters and flexible regions only provides a qualitative picture, a continuous quantitative measure is also available in terms of a flexibility index f_i, which is defined for each covalent bond i. In ProFlex and initial versions of FIRST, f_i is defined as (Eq. 3) **(*6*)**

$$f_i = \begin{cases} \dfrac{F_j}{H_j} & \textit{in an underconstrained region} \\ 0 & \textit{in an isostatically rigid cluster} \\ -\dfrac{R_k}{C_k} & \textit{in an overconstrained region} \end{cases} \qquad \text{Eq. 3}$$

In underconstrained regions j, f_i relates the number of independently rotatable bonds (F_j) to the number of potentially rotatable bonds (H_j). Conversely, in overconstrained regions k the number of redundant constraints (R_k) is related to the overall number of constraints (C_k). Thus, f_i ranges from -1 to 1, with negative values in rigid regions and positive values in flexible ones; the index allows quantifying *how much more flexible (stable)* an underconstrained (overconstrained) region is compared to a minimally rigid region **(*13*)**. For visualizing the results, atom-based flexibility indices can be calculated as average over f_i values of covalent bonds the atom is involved in **(*8, 13*)**. E.g., a flexibility index for C_α atoms has been calculated by averaging over the two backbone bonds (N-C_α and C_α-C'), while a flexibility index for phosphorus atoms has been calculated by averaging over the O5'-P and P-O3' bonds **(*8, 13*)**. The atom based flexibility indices can be visualized by a color-coded mapping onto the biomacromolecule's atoms (Figure 1 (5.)) **(*13, 17*)**. It is common to use bluish colors for indicating overconstrained regions, reddish colors for flexible regions, and green or white for minimally rigid regions **(*6, 8, 17*)**.

In recent versions of FIRST, a flexibility index g_i is now calculated according to (Eq. 4):

$$g_i = \begin{cases} \dfrac{F_j}{6E_j - B_j} & \textit{in an underconstrained region} \\ 0 & \textit{in an isostatically rigid cluster} \\ \dfrac{C_k - (6V_k - 6)}{\frac{6V_k(V_k - 1)}{2} - (6V_k - 6)} & \textit{in an overconstrained region} \end{cases} \qquad \text{Eq. 4}$$

In underconstrained regions *j*, F_j indicates the number of independently rotatable bonds, E_j is the number of edges representing rotatable bonds, and B_j is the total number of constraints from rotatable bonds. In overconstrained regions *k*, C_k is the total number of constraints and V_k indicates the number of atoms in that region. Note that $f_i = g_i$ for bonds in underconstrained regions but $f_i \neq g_i$ for bonds in overconstrained regions: f_i relates the number of redundant constraints to the *actual* number of *all* constraints, while g_i relates the number of redundant constraints to the *maximal* number of *redundant* constraints. This must be considered when comparing flexibility analyses from different programs or program versions.

3. *Hydrogen bond dilution*

By gradually removing non-covalent bonds from the constraint network, the thermal unfolding of biomacromolecule structures can be simulated (***12, 15***). So far, hydrogen bonds and salt bridges have been removed successively from the network in the order of increasing strength. In contrast, the number of hydrophobic contacts has been kept constant because the strength of hydrophobic interactions remains constant or even increases with increasing temperature. A hydrogen bond dilution can be computed by FIRST using the '–dil 1' option and by ProFlex using the '–nonh' option. The dilution simulates a melting of the network and results in a hierarchy of regions of varying stability (***18***). That way, information is gained that complements the above flexibility indices.

Furthermore, by applying indices from network theory (***33***), the *microstructure* of a network, i.e., properties of the set of rigid clusters generated by the bond dilution process, and *macroscopic properties* of a network associated with the rigid cluster size distribution, such as a transition from a folded to an unfolded state, have been be analyzed in the context of protein (thermo-)stability (***9, 12***). Calculating these indices is possible within the Constraint Network Analysis (CNA) package (***9, 12***), which is a front-end to FIRST. Such analyses may also become valuable for structure-based drug design when it comes to estimating the effect of ligand binding on the structural stability of a receptor.

3. Notes

Constraint counting on a topological network representation of biomacromolecules provides a deeper understanding of the flexibility characteristics of protein, RNA, and DNA structures down to the atomic level in a computational time on the order of seconds. Compared to MD simulations, the computational time requirement for a flexibility analysis is several orders of magnitude smaller. By now, there is ample evidence that a flexibility analysis provides a picture of biomacromolecular flexibility that agrees with MD results or data from experiments (***6, 8, 9, 13***). Still, several methodological pitfalls exist, and improvements of the topological network representation can be anticipated.

3.1. Sensitivity of flexibility analysis to the input structure

While atomic motions along a MD trajectory are governed by the continuous spectrum of forces exerted by surrounding atoms, the constraints in the topological network are "all-or-nothing" – a bond is either present or absent. Especially in the case of non-covalent interactions, one needs to distinguish forces sufficiently strong, which are included into the network, from weaker ones, which are excluded. In the case of marginally stable biomacromolecules, this can lead to different experimental input structures showing significant differences in flexibility predictions (C. Pfleger, E. Schmitt, H. Gohlke, unpublished results): a region in such structures may switch from flexible to rigid depending on the inclusion of a few (in the extreme, a single) constraint. We thus recommend testing the sensitivity of flexibility analysis by varying the energy cutoff for hydrogen bonds E_{HB} and/or the criteria for inclusion of hydrophobic interactions, and repeating the flexibility analysis. Likewise, conformations extracted along a MD trajectory can also result in different flexibility predictions (***13*, *34***). When available, we thus recommend performing the flexibility analysis on an ensemble of input structures and then average the results (***13***). This is also advantageous because it allows deriving a measure of significance for flexibility predictions on the atomic level in terms of the standard error of the mean. Ensemble-based flexibility analysis can be performed using the CNA package.

3.2. Treatment of water molecules

Interactions mediated by structural water molecules are known to affect the flexibility and stability of biomacromolecules. In most flexibility analysis studies so far, water molecules have not been included in the topological network, mainly due to the problem to distinguish tightly bound water molecules from fast-exchanging ones based on information from experiment. Results from MD simulations can complement experiments in this respect (***35***). However, by incorporating data from computationally expensive MD simulations, the advantage of the highly efficient flexibility analysis with computing times on the order of seconds even for the large ribosomal subunit will be lost. Encouragingly, previous findings showed only a negligible difference in the flexibility characteristics of a protein-protein complex when structural waters were considered (***13***). In addition, the influence of solvent on structural stability is already implicitly considered by including hydrophobic interactions as constraints into the network (***9***).

3.3. Treatment of RNA and DNA structures

Recently, we adapted the approach to RNA structures by developing a new topological network representation for these macromolecules **(*8*)**. The adaptation was necessary because the structural stability of proteins, dominated by hydrophobic interactions, and RNA structures, dominated by hydrogen bonds and base stacking interactions, is determined by different non-covalent forces. Although the new network parameterization already provides crucial insights into the flexibility characteristics of RNA structures **(*8, 17, 36, 37*)**, several improvements of the network representation can be anticipated:

1. Base stacking interactions are known to be both dependent on the type of the bases and the sequential context: I) stacking interactions in general increase in the order pyrimidine-pyrimidine < purine-pyrimidine < purine-purine bases **(*38*)**; II) stacking interactions are larger for sequences rich in G-C rather than A-U base pairs **(*39, 40*)**. Thus, differences in base stacking interactions could be modeled by using varying numbers of constraints for the hydrophobic tethers. This approach has not been pursued so far.
2. Another area of improvement in modelling nucleic acids relates to the question how repulsive forces between negatively charged phosphate groups can be included into the topological network representation. Modeling repulsive forces is difficult within the combinatorial approach followed in the pebble game algorithm because this leads to one-way inequalities, where the constraint length cannot become shorter but longer, compared to two-way equalities, where the constraint length is fixed, used so far **(*41*)**.

In regard to using the RNA parameterization for analyzing DNA structures, one should notice that both types of molecules express different flexibility characteristics in response to the presence or absence of the 2'OH group **(*42*)**. A recent MD study revealed that the differences between flexibility and rigidity in both types of nucleic acids are much more complex than usually believed (***43***): RNA is very deformable along a small set of essential deformations, whereas DNA has a more degenerate pattern of deformability. To date, no validation study for using FIRST on DNA structures has been reported.

3.4. Comparison of flexibility analysis results with data from experiments

When comparing results from a flexibility analysis with data from experiments, one needs to keep in mind that *flexibility* is a static property, which describes the possibility of motion. Phrased differently, flexibility denotes the ability of a region to be deformed. From the study of flexibility alone, however, no information is available about the direction and magnitude of the possible motions (***44***). In contrast, data from experiments, e.g., crystallographic B-values, or MD simulations, e.g., atomic fluctuations, often report on the *mobility* of atoms. Unsurprisingly, results from flexibility analysis and mobility information from experiment or MD simulation must disagree in the case of a rigid, yet mobile, body (such as a moving helix or domain).

Along these lines, one must take into account that flexibility analysis is better suited to characterize biomacromolecular flexibility that underlies longer timescale motions (***45***). While hydrogen/deuterium exchange experiments are frequently interpreted in the context of such longer timescale motions, NMR S^2 order parameters are generally associated with fast fluctuations in the ns regime. Thus, results of a flexibility analysis and S^2 order parameters must be compared with caution.

3.5. Applications

There are many potential applications for flexibility analysis. Predicted flexibility characteristics of biomacromolecules can either be linked to biological function, which is not in the focus of the present review, or be used to support structure-based drug design. The present challenge in structure-based drug design is that it is not known in advance which conformation a target will adopt in response to binding of a ligand or how to design a ligand for such an unknown conformation **(*1*)**. In this context, it is advantageous that flexibility analysis provides rigidity and flexibility information at various structural levels:

1. Flexibility characteristics at the bond level are instructive for analyzing binding site regions. As such, flexibility analysis can be used to guide the sampling of protein main-chain flexibility during ligand docking as proposed by Keating and coworkers **(*7*)**. In such a case, the identified hinge regions can be used as input for the docking program FlexDock, which handles hinge-bending motions of the receptor molecule during the docking process **(*46*)**. Similarly, a flexibility analysis will also be helpful for identifying potentially flexible sidechains in a binding site. This can be used for docking with AutoDock4 **(*47*)**, which allows to model as flexible only a few sidechains of the binding site during the docking.
2. By investigating ribosomal structures from different organism, we found characteristic flexibility patterns in the highly conserved antibiotics binding pocket at the PTC for different kingdoms. These flexibility patterns have been related to antibiotics selectivity **(*17*)**. These findings point to the importance of considering differences in the degrees of freedom of binding regions upon complex formation, as such differences may entropically influence binding processes. Furthermore, it shows that subtle differences in binding site flexibility might need to be considered for a proper assessment of the drugability of new putative binding sites.
3. Flexibility characteristics of larger regions can be related to potential global conformational changes and provide hints about movements of structural parts as rigid bodies. By determining a hierarchy of regions of varying stabilities of the large ribosomal subunit, we were able to propose a pathway of allosteric signal transmission

from the ribosomal tunnel region to the peptidyl transferase center (PTC) **(*17*)**. Remarkably, this prediction was later confirmed by cryo-EM data of a stalled ribosome structure **(*48*)** and mutation studies **(*49*)**. This shows that the approach can be used to detect coupling between two structural sites, which makes it most interesting for identifying new allosteric binding sites.

4. Finally, the rigid cluster decomposition can serve as input for coarse-grained simulation methods **(*21, 22, 50-52*)**, which sample the conformational space of a biomacromolecule by means of constrained geometric simulation (Figure 1 (4.)). Ligands can then be docked into the ensemble of receptor conformations, as was successfully demonstrated for the cyclic peptide cyclosporine with its receptor cyclophilin **(*53*)** and multiple ligands binding to HIV-1 TAR RNA **(*37*)**. In both cases, docking into an ensemble of simulation-generated structures proofed to be a valuable tool to cope with large *apo*-to-*holo* conformational transitions of the receptor structure, thereby implicitly taking into account conformational changes upon binding.**(*54*)**

Acknowledgement

We are grateful to L.A. Kuhn and M.F. Thorpe for insightful discussions.

Figure captions

Figure 1: Workflow of a flexibility analysis of a biomacromolecule based on constraint counting. A thrombin structure (PDB code 1ETS) was taken as an example. (1.) A PDB structure including polar hydrogen atoms is used as input. (2.) The biomacromolecule is modeled as a topological network. In this network, vertices represent atoms and edges represent covalent and non-covalent bond constraints (strong hydrogen bonds (red lines), salt bridges (red lines), and hydrophobic interactions (green lines)) **(*44*)**. Then, each bond is identified as either part of a rigid region or a flexible link in between. The resulting rigid cluster decomposition of the thrombin structure is shown in (3.), where each rigid cluster is depicted as a uniformly colored body. The left (right) picture shows the rigid cluster decomposition before (after) a phase transition as determined using the cluster configuration entropy (Eq. 2) **(*9, 12*)**. The computed decomposition of the biomacromolecular structure into rigid and flexible regions can be used in a subsequent step as input for coarse-grained simulations **(*21, 22, 44*)**, which explore the molecule's mobility. Panel (4.) shows an ensemble of thrombin conformers generated by such a method, NMsim **(*21, 55*)**, within a few hours of computational time. Finally, a flexibility index (Eq. 3) can be obtained, which is mapped in a color-coded fashion onto the thrombin structure (5.). Overconstrained regions are indicated by blue colors ($f_i < 0$), rigid regions are represented in white ($f_i = 0$), and flexible regions are shown in red colors ($f_i > 0$). The blow up in (5.) shows the active site of thrombin together with a bound ligand and the S1, S2, and S3 subpockets. The flexibility index provides crucial insight into the binding site flexibility at the bond level. For example, the 60-insertion loop (Tyr60A-Trp60D) assumes different orientations in complexes with different inhibitors **(*56*)**. In agreement with this, residues Leu60 and Asp60E-Thr60I are identified to be flexible, which allows the movement of the 60-insertion loop. Finally, potential applications of the approach to computer-aided drug design are listed in (6.).

References

1. Cozzini P, Kellogg GE, Spyrakis F, Abraham DJ, Costantino G, Emerson A, Fanelli F, Gohlke H, Kuhn LA, Morris GM, Orozco M, Pertinhez TA, Rizzi M, and Sotriffer CA. (2008) Target flexibility: An emerging consideration in drug discovery and design. J Med Chem 51:6237-6255.
2. Wlodawer A and Vondrasek J. (1998) Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. Annu Rev Biophys Biomol Struct 27:249-284.
3. Zhang Q, Sun X, Watt ED, and Al-Hashimi HM. (2006) Resolving the motional modes that code for RNA adaptation. Science 311:653-656.
4. Perez A, Noy A, Lankas F, Luque FJ, and Orozco M. (2004) The relative flexibility of B-DNA and A-RNA duplexes: Database analysis. Nucleic Acids Res 32:6144-6151.
5. Getz M, Sun X, Casiano-Negroni A, Zhang Q, and Al-Hashimi HM. (2007) NMR studies of RNA dynamics and structural plasticity using NMR residual dipolar couplings. Biopolymers 86:384-402.
6. Jacobs DJ, Rader AJ, Kuhn LA, and Thorpe MF. (2001) Protein flexibility predictions using graph theory. Proteins 44:150-165.
7. Keating KS, Flores SC, Gerstein MB, and Kuhn LA. (2009) StoneHinge: Hinge prediction by network analysis of individual protein structures. Protein Sci 18:359-371.
8. Fulle S and Gohlke H. (2008) Analysing the flexibility of RNA structures by constraint counting. Biophys J 94:4202-4219.
9. Radestock S and Gohlke H. (2008) Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. Eng Life Sci 8:507-522.
10. Livesay DR and Jacobs DJ. (2006) Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. Proteins 62:130-143.
11. Tan HP and Rader AJ. (2009) Identification of putative, stable binding regions through flexibility analysis of HIV-1 gp120. Proteins 74:881-894.
12. Radestock S and Gohlke H. (2010) Protein rigidity and thermophilic adaptation. Proteins DOI 10.1002/prot.22946.
13. Gohlke H, Kuhn LA, and Case DA. (2004) Change in protein flexibility upon complex formation: Analysis of Ras-Raf using molecular dynamics and a molecular framework approach. Proteins 56:322-337.
14. Wells SA, Jimenez-Roldan JE, and Romer RA. (2009) Comparative analysis of rigidity across protein families. Phys Biol 6:46005.
15. Hespenheide BM, Rader AJ, Thorpe MF, and Kuhn LA. (2002) Identifying protein folding cores from the evolution of flexible regions during unfolding. J Mol Graphics Modell 21:195-207.
16. Rader AJ and Bahar I. (2004) Folding core predictions from network models of proteins. Polymer 45:659-668.
17. Fulle S and Gohlke H. (2009) Statics of the ribosomal exit tunnel: Implications for co-translational peptide folding, elongation regulation, and antibiotics binding. J Mol Biol 387:502-517.
18. Rader AJ, Hespenheide BM, Kuhn LA, and Thorpe MF. (2002) Protein unfolding: Rigidity lost. Proc Natl Acad Sci U S A 99:3540-3545.
19. Hespenheide BM, Jacobs DJ, and Thorpe MF. (2004) Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. J Phys: Condens Matter 16:5055-5064.
20. Wang Y, Rader AJ, Bahar I, and Jernigan R. (2004) Global ribosome motions revealed with elastic network model. J Struct Biol 147:302-314.

21. Ahmed A and Gohlke H. (2006) Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. Proteins 63:1038-1051.

22. Wells S, Menor S, Hespenheide B, and Thorpe MF. (2005) Constrained geometric simulation of diffusive motion in proteins. Phys Biol 2:S127-136.

23. Rader AJ. (2010) Thermostability in rubredoxin and its relationship to mechanical rigidity. Phys Biol 7:16002.

24. Fulle S and Gohlke H. (2009) Constraint counting on RNA structures: Linking flexibility and function. Methods 49:181-188.

25. Hooft RWW, Vriend G, Sander C, and Abola EE. (1996) Errors in protein structures. Nature 381:272-272.

26. Vriend G. (1990) WHAT IF: A molecular modeling and drug design program. J Mol Graph 8:52-56.

27. Word JM, Lovell SC, Richardson JS, and Richardson DC. (1999) Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 285:1747.

28. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, and Woods RJ. (2005) The Amber biomolecular simulation programs. J Comput Chem 26:1668-1688.

29. Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, and Onufriev A. (2005) H++: a server for estimating pK(a)s and adding missing hydrogens to macromolecules. Nucleic Acids Res 33:W368-W371.

30. Draper DE. (2004) A guide to ions and RNA structure. RNA 3:335-343.

31. Hashem Y and Auffinger P. (2009) A short guide for molecular dynamics simulations of RNA systems. Methods 47:187-197.

32. Ahmed A, Kazemi S, and Gohlke H. (2007) Protein flexibility and mobility in structure-based drug design. Front Drug Des Discov 3:455-476.

33. Stauffer D and Aharony A (1994) *Introduction to Percolation Theory.*, Taylor and Francis, London.

34. Mamonova T, Hespenheide B, Straub R, Thorpe MF, and Kurnikova M. (2005) Protein flexibility using constraints from molecular dynamics simulations. Phys Biol 2:S137-147.

35. Vaiana AC, Westhof E, and Auffinger P. (2006) A molecular dynamics simulation study of an aminoglycoside/A-site RNA complex: Conformational and hydration patterns. Biochimie 88:1061-1073.

36. Stoddard CD, Montange RK, Hennelly SP, Rambo RP, Sanbonmatsu KY, and Batey RT. (2010) Free State Conformational Sampling of the SAM-I Riboswitch Aptamer Domain. Structure 18:787-797.

37. Fulle S, Christ NA, Kestner E, and Gohlke H. (2010) HIV-1 TAR RNA spontaneously undergoes relevant apo-to-holo conformational transitions in molecular dynamics and constrained geometrical simulations. J Chem Inf Model 50:1489-1501.

38. Saenger W (1984) *Principles of Nucleic Acid Structure*, Springer-Verlag, New York.

39. Ornstein RL, Rein R, Breen DL, and Macelroy RD. (1978) An optimized potential function for the calculation of nucleic acid interaction energies. I. Base stacking. Biopolymers 17:2341-2360.

40. Gralla J and Crothers DM. (1973) Free energy of imperfect nucleic acid helices: III. Small internal loops resulting from mismatches. J Mol Biol 78:301-319.

41. Whiteley W. (2005) Counting out to the flexibility of molecules. Phys Biol 2:S116-126.

42. Pan Y and MacKerell AD. (2003) Altered structural fluctuations in duplex RNA versus DNA: A conformational switch involving base pair opening. Nucleic Acids Res 31:7131-7140.

43. Noy A, Pérez A, Lankas F, Luque FJ, and Orozco M. (2004) Relative flexibility of DNA and RNA: A molecular dynamics study. J Mol Biol:627-638.

44. Gohlke H and Thorpe MF. (2006) A natural coarse graining for simulating large biomolecular motion. Biophys J 91:2115-2120.

45. Livesay DR, Dallakyan S, Wood GG, and Jacobs DJ. (2004) A flexible approach for understanding protein stability. FEBS Letters 576:468-476.

46. Schneidman-Duhovny D, Inbar Y, Nussinov R, and Wolfson HJ. (2005) Geometry-based flexible and symmetric protein docking. Proteins 60:224-231.

47. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, and Olson AJ. (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem 30:2785-2791.

48. Seidelt B, Innis CA, Wilson DN, Gartmann M, Armache JP, Villa E, Trabuco LG, Becker T, Mielke T, Schulten K, Steitz TA, and Beckmann R. (2009) Structural insight into nascent polypeptide chain-mediated translational stalling. Science 326:1412-1415.

49. Vázquez-Laslop N, Ramu H, Klepacki D, Kannan K, and Mankin A. (2010) The key function of a conserved and modified rRNA residue in the ribosomal response to the nascent peptide. EMBO J 29:3108-3117.

50. Lei M, Zavodszky MI, Kuhn LA, and Thorpe MF. (2004) Sampling protein conformations and pathways. J Comput Chem 25:1133-1148.

51. Ahmed A and Gohlke H. (2009) Multiscale modeling of macromolecular conformational changes, in *1st International Conference on Computational & Mathematical Biomedical Engineering - CMBE09* (Nithiarasu P., and Löhner R., Eds.), pp 219-222, Swansea, UK.

52. Farrell DW, Speranskiy K, and Thorpe MF. (2010) Generating stereochemically acceptable protein pathways. Proteins 78:2908-2921.

53. Zavodszky MI, Ming L, Thorpe MF, Day AR, and Kuhn LA. (2004) Modeling correlated main-chain motions in proteins for flexible molecular recognition. Proteins 57:243-261.

54. Totrov M and Abagyan R. (2008) Flexible ligand docking to multiple receptor conformations: A practical alternative. Curr Opin Struct Biol 18:178-184.

55. Ahmed A, Rippmann F, Barnickel G, and Gohlke H. (2010) A normal mode-based geometric simulation approach for exploring biologically relevant conformational transitions in proteins. submitted.

56. de Amorim HLN, Netz PA, and Guimaraes JA. (2010) Thrombin allosteric modulation revisited: a molecular dynamics study. J Mol Model 16:725-735.