universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Structural, evolutionary and energy-based analysis of residue interaction networks of small G proteins

verfasst von / submitted by

Raphael Peer, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 834
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Molekulare Biologie
Betreut von / Supervisor:	Dr. Bojan Zagrovic
Mitbetreut von / Co-Supervisor:	

Acknowledgements

I would like to thank my senior supervisor M. Madan Babu for giving me the chance of working in the exciting research environment of the MRC Laboratory of Molecular Biology. Throughout my time in Cambridge, Madan encouraged me to think critically and allowed me to pursue my own research interests. As a result, I acquired valuable experience in project design and project management. Moreover, I could learn a great deal from his clear and structured way of presenting scientific results. I am thankful for the fascinating experience of working in Madan's research group.

I am grateful to Tilman Flock for supervising this master's project and teaching me numerous valuable skills in the process. The project started out by Tilman sharing with me his ideas for future research projects inspired by his previous work on consensus residue interaction networks. He thought me how to design a research project and his creative and analytical way of thinking as well as his can-do attitude proved invaluable. I am thankful for all the time and effort Tilman invested in supervising this project and for his support and friendship throughout my in time Cambridge. On top of that, he kindly let me win in table football most of the time. For all those reasons it has been a great experience to work with him. It goes without saying that Tilman would make an excellent group leader.

I would like to thank Bojan Zagrovic for agreeing to be the University of Vienna's official supervisor for this master's project and for reviewing this thesis. I am honored by his trust in my work and his feedback improved this thesis considerably. Moreover, I would like to add that his fascinating research and his excellent teaching inspired me to specialize in computational structural biology in the first place. The skills and concepts I learned during my internship in Bojan's group and in his courses were invaluable for this project and will continue to be so in my future projects.

I would like to thank Sreenivas Chavali for his mentoring, his support and for sharing his extensive research experience. It was a pleasure to work with him and to share plenty of interesting conversations.

In a side project, I had the pleasure of collaborating with summer student Andrija Sente. Andrija sets the standard for being passionate about science. His curiosity, his ability to learn remarkably fast and his motivation to work around the clock allowed the project to progress faster than anyone thought possible. One can't help but to suspect that the lived in the LMB during that time and I would not be too surprised if he still did. In any case, without his omnipresent cheerful nature, his genuine interest in research and his creative use of inside jokes, my time in the LMB would not have been the same.

Furthermore, I want to thank my colleges and friends Alexey Morgunov, Natasha Latysheva, Charles Ravarani, Melis Kayikci, Hannes Harbrecht, Greet De Baets, Guilhem Chalancon, Balaji Santhanam, Marion Ouédraogo, Rita Pancsa, Norbert Feher, Alex Gunnarsson and Alexandre Erkine for interesting discussions, sharing my sense of humour and epic table football matches. It was great to have so many interesting conversations about biology, history and travel with Alexey. I miss his regular rants about British cuisine which almost made it easier to endure. Moreover, I regret not being available for epic table football duels any more and fear he may have grown overconfident in my absence. I am grateful to Natasha for introducing me to data science, for demonstrating how much can be achieved with a creative approach and a positive attitude and for coming up with an infallible table football strategy. Her cheerful nature makes the lab a brighter place. My thanks go to Charles for saving my day with his R-coding skills on more than one occasion and for showing me where to find good food in Cambridge. I want to thank Greet for showing interest in my research and providing constructive feedback and for the dry ice. I am thankful to Hannes for providing helpful feedback during the course of the project and sharing his experience in molecular modelling. I had the honour to meet and have many interesting discussions with Alexandre Erkine. It was fascinating to see how after decades of research experience his scientific curiosity did not diminish one bit. His attitude towards science and life is a great inspiration for young researchers.

I was lucky to have the support of Jannie Lightfoot and Maud Pilkington who patiently guided me through administrative matters.

Last but not least, I would like to acknowledge the outstanding support from Jake Grimmett and Toby Darling from the scientific computing facility. Their well maintained computational infrastructure and technical support are an invaluable asset for computational biologists at the LMB.

Abstract

Structural, evolutionary and energy-based analysis of residue interaction networks of small G proteins

by Raphael Peer

The growing number of protein structures available in the PDB provides a wealth of information for structural biologists. However, analysing large numbers of structures is computationally challenging. A promising method to address this challenge is to represent proteins as networks of contacting residues. Residue contact networks (RCNs) define residues as nodes and contacts between residues as edges in the network. In this master's project, a bioinformatics pipeline was developed to compare RCNs across proteins of a given protein family. The well studied small G protein family served as a model system. In order to allow a comparison of residue contacts of over 500 small G proteins, a common residue numbering system for this protein family was created. Moreover, contact energy was calculated using the Rosetta protein modelling suite. The result provides two parameters for every residue contact: the fraction of protein structures which have an equivalent residue contact, termed 'contact conservation', and contact energy. Using this information, the relation between contact conservation, contact energy, sequence conservation and co-evolution is explored. Furthermore, comparison with mutational data provides support for the method developed in this project. Taken together, the results demonstrate that the developed bioinformatics pipeline allows to combine detailed information from a large number of protein structures and provides valuable insights.

Contents

Chapter 1

Introduction

1.1 Background

1.1.1 Systems structural biology

Currently, more than 100,000 protein structures have been solved experimentally and deposited in the public available protein data bank (PDB) [1]. Around 90% of these structures have been determined by X-ray crystallography, around 10% were solved using NMR and a small (below 1%) but growing number of protein structures is determined with electron microscopy [2]. The number of protein structures in the PDB is expected to continue to grow in the future.

Several databases, such as SCOP [3] and CATH [4], have been created to classify the growing number of structures. By classifying proteins according to their three dimensional fold, these projects have shown that the protein universe consists of surprisingly few distinct protein folds. The number of distinct protein folds varies strongly from as low as 650 [5] to several thousands [6][7]. Regardless of the precise number, the majority of protein folds are so called 'unifolds' containing only one protein family. Therefore, it can be assumed that the majority of protein families (and hence the majority of proteins) belong to about 1000 common folds [8].

These studies show how large scale analysis of protein structures - also known as systems structural biology - can give insights beyond individual proteins. Given the steady increase in the number of protein structures available, more research opportunities are expected to arise in this field.

1.1.2 Residue contact networks

Classification of proteins into folds and families, as described above, is done by comparing tertiary protein structure. However, insights into function can only be gained from a perspective of atomic interactions. All interactions between residues in a given protein can be represented

in a residue contact network (RCN) [9]. RCNs provide an alternative representation of a protein structure as shown in figure 1.1.

The traditional 3D representation - often simplified as cartoon - shows the 3-dimensional arrangement of secondary structure elements in an intuitive manner. It allows to immediately see similarities and differences between two or a few structures. On the other hand, systematically analysing large numbers of 3D-structures in atomic detail is difficult. In contrast, residue contact networks (RCNs) are computationally more readily accessible. They allow to compute the shortest path between two residues, to identify the most connected residues and to calculate many more network parameters [10]. Arguably, one of the most important benefits of RCNs is that they facilitate comparison of a large number of structures. From multiple RCNs, a network of shared residue contacts can be obtained to explain the common function of multiple structures [11][12][13]. On the downside, RCNs are not intuitive to interpret which can be seen in figure 1.1.

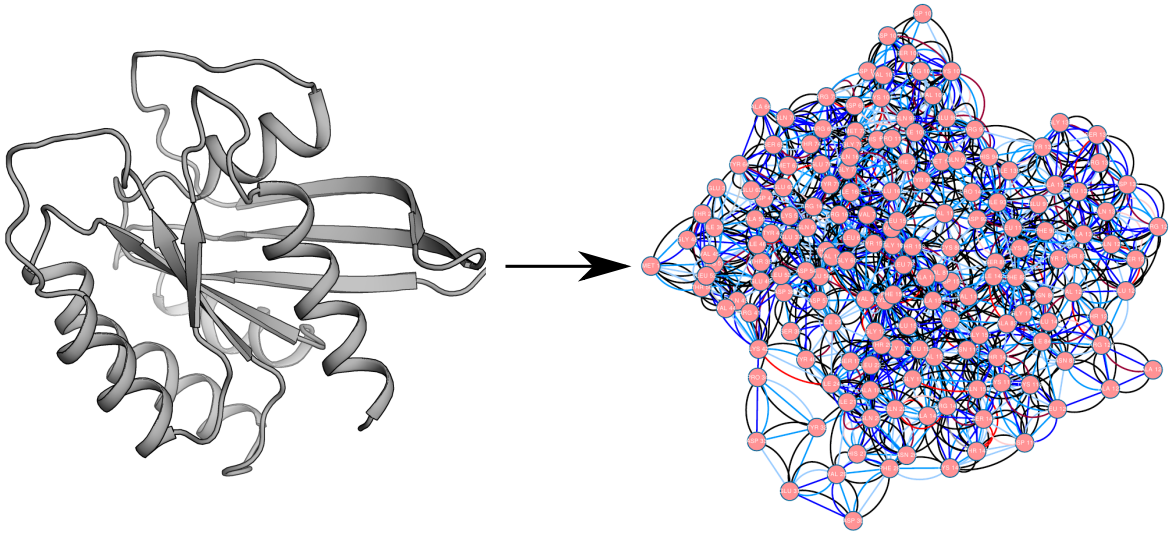

FIGURE 1.1: 3D cartoon and network representation of the small G protein HRas (PDB-ID 3K8Y [14]). The 3D cartoon, as well as all other protein illustrations in this thesis, were created with PyMOL [15]. The network was calculated with the program RINerator [16].

Taken together, while the 3D representation of a protein structure is perfect for visualization and manual inspection, a network of residue contacts has advantages for large scale computational analysis. However, several issues remain which limit the application of RCNs to address biological questions. In the remainder of this section, I will review the literature on RCNs, point the limitations of existing approaches and propose an alternative approach developed in this project.

Residue contacts are formed via various biochemical interactions such as hydrogen bonds, hydrophobic packing, electrostatic interaction of charged groups, π-π or π-cation interactions. For a small number of residues, the presence or absence of contacts can be determined manually. The corresponding distance constraints can be looked up in the literature and molecular viewers, such as PyMOL or UCSF Chimera, can aid by calculating atomic distances. However, when automating residue contact calculation, an algorithm has to be created which does not rely on human judgement.

In the past, many different definitions have been used. A simple way to obtain residue contacts is to calculate the inter-residue distances between C-α atoms [17]. Every two residues whose C-α atoms are within a certain distance (e.g. 8 Ångström) are assumed to be in contact. This approach is easy to implement but has severe flaws. It is possible that two residues with less then 8 Å distance between their C-α atoms do not have any two atoms in close enough proximity to form a biochemical interaction. This may be due to small side chain volume or orientation of the side chains in different directions. Therefore, a network created in this manner can be expected to contain false positives. On the other hand, amino acids with large side chains can form biochemical interactions even if their C-α atoms are not within 8 Å of each other - for instance a salt-bridge between glutamine and lysine. Hence, important residue contacts could be missing from a network created on the basis of Cα distances. Using Cβ distances instead eliminates one of the flaws because side chain orientation is taken into account [18]. However, C-β distances are also a poor indication of a biochemical interaction between two residues as side chain size is not considered. From these considerations, it becomes clear that using only a single, pre-defined atom per amino acid is not a reliable method to calculate residue contacts.

A more precise method is to calculate the distance between the closest two atoms of two residues. Green et al. defined a residue contact as two residues which have at least one pair of heavy atoms (non-hydrogen atoms) within a distance of less than 5 Å of each other [10]. Although this method is more accurate than Cα or Cβ networks, the issue remains that it does not take into account that different atoms have different van der Waals radii. No distance cut-off can be ideal for all types of contacts. Moreover, the approach does not explicitly consider hydrogen atoms.

To overcome this limitation, more sophisticated methods have been developed, most notably RINerator [16][19]. Before the analysis, hydrogen atoms are added to the structure with the program Reduce [20]. Then, the program Probe [21] determines the contacts or overlaps of the van der Waals spheres of atoms. Two van der Waals spheres within 0.25 Å of each other are assumed to represent a favourable interaction. In contrast, an overlap of van der Waals

spheres represents a clash [21]. Clashes are energetically highly unfavourable and usually due to a problem in the structure. An RCN generated with RINerator can be visualized in the popular network graphics software Cytoscape [22] via the plugin RINalyzer [16]. RINerator is the most precise publicly available distance based method of residue contact calculation as it explicitly considers hydrogen atoms as well as the different van der Waals radii of atoms. An RCN created with RINerator and visualized in Cytoscape is shown in figure 1.1.

However, calculating RCNs based solely on inter-residue distance fails to differentiate between energetically highly favourable residue contacts and less favourable contacts. The structural and functional importance of residue contacts is difficult to interpret without contact energy. For this reason, several studies have calculated residue contact energy networks. For instance, interaction strength has been approximated by the number of atom pairs of two residues which are within 4.5 Å of each other (normalized by a value which accounts for different side chain sizes) [11]. However, the number of atom pairs in close proximity is not necessarily indicative of interaction energy. For instance, this number may be higher in hydrophobic interactions of branched-chain amino acids than in salt bridges between oppositely charged amino acids.

Another approach of residue contact energy networks has been made in the context of molecular dynamics (MD) simulations [23]. The interaction energies were calculated with the g_energy module of the MD-simulation software GROMACS [24]. As expected, the authors found that energies of different types of interactions vary considerably. In general, charge mediated interactions were found to be highly favourable (large negative values) and hydrophobic interactions between small residues to be less favourable (small negative values) [23]. While these findings are not surprising, they show that computed interaction energies agree well with biochemical knowledge. Moreover, the wide range of interaction energies reported in this paper highlights the need to consider energies when interpreting residue contact networks.

Another study by the same authors uses the method to compare thermophilic and mesophilic proteins [25]. In this paper, the RCNs were filtered by excluding all contacts above (less negative) a certain threshold. Then, the number of clusters per network was computed for various energy thresholds. A cluster in a network is a set of nodes directly or indirectly connected with each other. The number of clusters per network can be anything from one (each node is at least indirectly connected to all other nodes) to the number of nodes in the network (in which case there are no edges in the network). The number of clusters was shown to depend on the energy threshold. Thermophilic proteins were shown to have a higher number of clusters than mesophilic proteins at any given energy threshold [25]. Moreover, the number of hubs was

compared between thermophilic and mesophilic proteins. Hubs are highly connected nodes in a network. In this case, the authors considered a residue which interacts with at least four other residues as a hub. Again, the number of hub depended on the energy threshold. The number of hubs was shown to be higher in thermophilic proteins regardless of energy threshold. The higher number of both, clusters and hubs, in thermophilic proteins suggest that residues in these proteins are more connected than in mesophilic proteins. These results explain the higher stability of thermophilic proteins. Moreover, the study demonstrates that calculating contact energy can give insights into protein structures which cannot be gained from solely distance based RCNs.

Taken together, RCNs can be improved by calculating residue contact energy.

1.1.3 Comparison of multiple residue contact networks

As figure 1.3 shows, RCNs can be rather large and difficult to interpret. For this reason, gaining functional insight from residue contact analysis is not trivial. In order to interpret an RCN, it can be useful to know whether for a given residue contact, an equivalent contact is present in related proteins. A value from zero to one indicates the fraction of related proteins which have an equivalent residue contact. This value is termed 'contact conservation' and is not to be confused with sequence conservation. It is independent of the amino acid types which form the residue contact - it only indicates that two residues in structurally equivalent positions form a contact in a certain fraction of a given set of related proteins. A so-called consensus RCN can be obtained from multiple RCNs of related proteins by extracting only contacts present in all of the proteins.

Comparative analysis of multiple RCNs has been successfully applied in several projects. For instance, Bhattacharyya et al. have compared the active and inactive form of the beta2-adrenergic receptor (β2-AR) and found greater connectivity of the active form RCN indicating tighter packing [11]. The authors also used the differences between the RCNs of the active and inactive from to shed light on the allosteric communication between the ligand binding site and the G protein interface.

In a more general study, Soundararajan et al. calculated RCNs of 8698 protein domains [12]. Solvent accessibility was computed to determine which residues can be regarded as core residues. Then, contacts between residues in the solvent inaccessible protein core were extracted. The resulting networks were termed protein core atomic interaction network (PCAIN). The authors found that PCAINs but not RCNs are specific for a given protein fold indicating that PCAINs define a protein fold.

In earlier work, RCNs of Gα subunits of trimeric G proteins were compared by Flock et al. [13]. Gα subunits are activated upon binding to G protein coupled receptors (GPCRs) and in turn activate down-stream signalling. The aim of the study was to find out whether Gα proteins have a common activation mechanism and if so to describe the mechanism. The analysis of the Gα-GPCR interface was limited to the only complex structure available. However, for analysis of the inactive state, a consensus RCN could be created from 11 structures. This consensus network included only residue contacts present in all inactive structures investigated. In addition to integrating information from multiple structures, a large scale sequence alignment identified conserved residues. The consensus network was then filtered by sequence conservation to give a consensus RCN of conserved residues only. The mechanism of the common allosteric activation could be explained by analysing the difference between the intra G-α residue contacts of the complex structure and the consensus network of the inactive structures. This study served as a guideline for this master's project as, to my knowledge, it is the first study to create a common numbering system for a protein family and use it to compare RCNs from a large number of protein structures. The RCNs were calculated with the distance based method RINerator [16] (described in the previous section 1.1.2). As this method does not indicate how energetically favourable a given contact is, it would be interesting to include contact energy in a similar study.

In summary, multiple studies have shown that comparing RCNs of related proteins can give interesting insights into structure and function. As the number of available protein structures is expected increase steadily, this approach could become even more useful in the future.

1.1.4 How do residue contacts define the protein fold?

Non-covalent interactions between residues define both secondary and tertiary structure of proteins. In fact, protein structures can be reconstructed fairly well from RCNs [26]. Therefore, the information for specifying a given protein family's fold has to be contained in its residue contacts. However, it was found that intra-family correlation of RCNs is similar to inter-family correlation. In other words, RCNs are not very specific for a protein family [12]. The reason for this apparent paradox could lie in residue contacts with energy close to zero (which means that they are neither favourable or unfavourable). These contact are not expected to be important for stabilizing a given protein fold. More likely, they arise because two residues happen to be in proximity as a consequence of the fold. Therefore, identifying residue contacts which define a given protein fold is not trivial. One approach is to only consider contacts between residues of the protein core as described earlier [12]. A network of such contacts was found to be more

specific for a given protein fold than a conventional RCN. Another approach is to only consider contacts between sequence-conserved residues as described in the previous section [13]. The rationale behind this approach is that contact between sequence-conserved residues ought to be important for all members of a protein family - otherwise the residues would not need to be sequence-conserved. Then, a consensus network (of contacts present in all members of the protein family) was calculated from these contacts [13].

An interesting alternative approach would be calculate a consensus RCN of a protein family using only highly favourable residue contacts. This could be achieved by combining the two approaches discussed above - residue contact energy calculation and comparing multiple RCNs (section 1.1.2 and 1.1.3). Section 1.3.3 will discuss how this alternative approach was tested in this master's project.

1.1.5 Relation between sequence and structure

Protein structure is determined by amino acid sequence [27]. Chothia and Lesk found that structural difference between related proteins increases exponentially with sequence divergence [28]. However, inferring structure from sequence is not trivial. The most successful approaches use related proteins as templates for structure prediction. Predicting protein structure without prior information, called ab initio prediction, is even more difficult [29]. Predicting protein structure from sequences is one of the most important challenges of structural biology and bioinformatics [30].

Protein structure is more conserved than amino acid sequence. This means that proteins with highly dissimilar sequences can still have similar folds [31]. On the other hand, proteins with similar sequence almost certainly fold into similar structures. (There are exceptions to this rule. For instance, proteins with similar sequence but different folds were designed in the lab. However, it is important to note that - possibly with very few exceptions - such cases do not occur naturally [32].) When sequence similarity is too low to readily assume evolutionary relatedness between proteins (20-25% sequence identity), the term 'twilight zone' is used [31]. Interestingly, proteins in the twilight zone usually still have similar folds. While it is common text book knowledge that protein structures are more conserved than sequences, it is in my opinion still surprising how sequences with barely detectable similarity fold into strikingly similar structures.

Among related proteins with low sequence-conservation, the phenomenon of co-evolving residues can be observed. This means that mutation of one residue has to be balanced by a mutation in the other residue in order for the protein to be functional [33]. Although residues can

also co-evolve without being in direct contact (for instance via indirect contact mediated by another residue), most co-evolving residues physically interact each other. For the identification of co-evolving residues, large sequence alignments are screened in search of statistically correlated positions [34]. The information of co-evolving pairs of residues can be used for functional interpretation of proteins and even aid protein structure prediction [33].

As discussed before, protein structure is determined by residue contacts. Mutation of one residue usually disrupts a contact (unless it is a change to a similar amino acid) - unless the second residue mutates in a way which balances the effect of the first mutation. In this manner, related proteins can diverge in sequence but still maintain similar residue contacts and thus similar structure. As this project investigates conserved residue contacts, is important to keep in mind, that related proteins can have equivalent residue contacts but the amino acids forming these contacts may be different (see section 1.3.5).

1.2 The model system: small G proteins

In the previous section, the literature in the field of residue contact networks was reviewed and several challenges and open questions were pointed out. The family of small G proteins is used in this project to address these questions. This section gives an overview of this remarkable family of proteins. The aims of the project will be described in the next section.

G proteins (guanine nucleotide-binding proteins) are a protein family with a common fold and the ability to bind GDP or GTP. The conformation is slightly different depending on whether GDP of GTP is bound. This feature enables them to act as molecular switches in cellular signalling. The GTP-bound form - also called the active form - allows a signal to be transmitted to downstream factors. On the other hand, no signal is transmitted in the GDP-bound form - also known as the inactive form [35]. G proteins posses intrinsic GTPase activity. Therefore, any bound GTP is eventually hydrolysed. Furthermore, GDP is eventually released and replaced with GTP as the cytoplasm has considerably higher GTP- than GDP-concentrations. In this manner, a G protein naturally cycles between its active and inactive state. However, the natural rates of GTP-hydrolysis and GDP-release are slow [36]. The fact that both, GTP-hydrolysis and GDP-release, can be regulated by a multitude of factors makes G proteins useful for signalling. Guanine nucleotide exchange factors (GEFs) induce GDP-release, thus activating the G protein. GTPase-activating proteins (GAPs) catalyse GTP-hydrolysis which inactivates G proteins [37].

There are two kinds of G proteins. The first kind is a family of monomeric GTPases of about 160 residues which is known as small G proteins, small GTPases or Ras superfamily.

The second kind is also know as heterotrimeric G proteins and consists of three subunits. The α-subunit, which is related to small GTPases, and the β- and γ-subunit. These G proteins are activated by a family of trans-membrane receptors termed G protein coupled receptors (GPCRs). Hence, GPCRs are GEFs specific for the α-subunit of trimeric G proteins. However, in contrast to GEFs of small G proteins, GPCRs bind the G-α subunit almost 30 Å away from the GDP-binding site, triggering GDP-release via an allosteric mechanism. The common activation mechanism of Gα-subunits has been identified by Tilman Flock [13]. In contrast, small G proteins are bound close to the GDP/GTP binding site by GEFs and GAPs.

The structure of GDP-bound HRas is show as a representative small G protein in figure 1.2. The GDP/GTP-binding site is formed by five loops termed G1 through G5 and coloured in blue. The two regions with different conformations in the GDP- and GTP-bound state are called switch regions and coloured red [35]. Interaction of GEFs and GAPs with the switch regions alters G protein activity [37].

The small G protein family contains more than 150 human proteins with orthologs in even distantly related species such as *C. elegans* and yeast. It is divided into five main subfamilies: Ras, Rho, Ran, Rab and Arf. A large variety of cellular processes is regulated by small G proteins including cell proliferation, actin organization, vesicular transport, nucleocytoplasmic transport, exocytosis and endocytosis [38]. Small G proteins are also medically highly relevant. For instance, mutations in the Ras subfamily are found in 20-30% of all human cancers making them a promising target for cancer treatment [39].

Given their importance for human health, small G proteins are also one of the best studied protein families. At the time of the start of this project, 543 structures were available in the PDB. The availability of such a large data set, the relatively simple fold as well as the great research interest makes this family of proteins an excellent model system for this project.

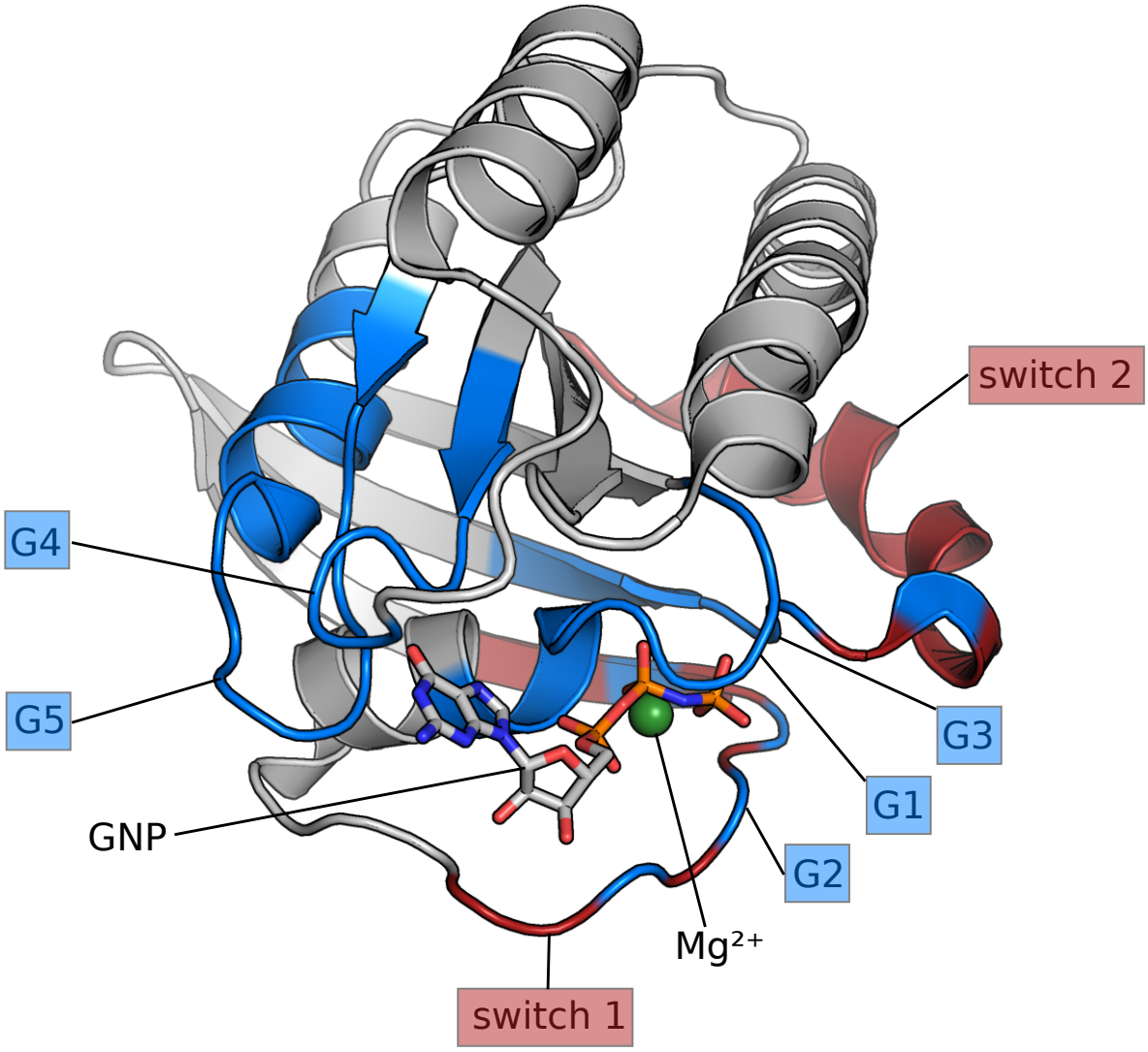

FIGURE 1.2: Structure of human HRas as example for a small G protein. The structure at hand (PDB-ID 3K8Y) was published in reference [14]. In this figure, information from other references [35][36] has been added to describe important structural elements of small G proteins in general. Five regions, critical for the cycle of GDP-GTP exchange and GTP hydrolysis are coloured in blue and labelled as G1 through G5. Two switch regions, which have different conformations in the GDP and GTP-bound states, are coloured in red. When two labels apply, the residues are coloured in alternating blue and red.

1.3 Motivation

1.3.1 Aim 1: Find a way to calculate residue contact energy

As explained in section 1.1.2, RCNs can be improved by considering contact energy. Therefore, the first aim of this project was to find method to calculate residue contact energy, suitable for

large-scale analysis. The Rosetta protein modelling suite was found match this purpose [40]. Section 2.3 in the next chapter describes in detail why this tool was chosen, how it works and how it was applied. Here, a preliminary study is presented to illustrate the importance of including contact energy in RCNs.

Distance-based calculations do not consider the interaction energy of contacting residues. For instance, RINerator [16], one of the most sophisticated tools to calculate distance-based RCNs, defines a contact whenever two atoms of two residues are in proximity (0.25 Å) of each other. In other words, distance-based RCNs provide a binary contact definition (present or not), although the 'biological relevance' of a contact highly depends on its energetics (e.g. different number of contacting atoms, different orientations of aromatic interactions, etc.). Naturally, the interaction energy varies greatly depending on the atoms involved.

In order to examine the range of interaction energies present in a typical RINerator RCN, I used Rosetta to calculate contact energies (the Rosetta energy function will be discussed in detail in section 2.3.1). Then, the RINerator and Rosetta residue contacts were compared. As a result, an energy value obtained from Rosetta could be assigned to each residue contact given by RINerator. As expected, the RINerator network includes not only highly favourable interactions but also residues in proximity with an interaction energy close to zero. Figure 1.3b shows two example contacts from the H-Ras (PDB-ID 3K8Y) RCN, created with RINerator. One example is a contact between an arginine and an aspartate formed by two hydrogen bonds. Due to the electro-negativity of the nitrogen atoms in the arginine, the N-bound hydrogen atoms have a positive partial charge. On the other hand, the aspartate side chain has a negative charge (shared between the two side chain oxygen atoms) at physiological PH. Therefore, it is not surprising and in agreement with biochemical knowledge that those two residues interact favourably. On the other hand, the second example contact is based on the proximity of an arginine and a leucine side chain. Partially positively charged hydrogen atoms from the arginine are in proximity of nonpolar hydrogen atoms from the leucine side chain. No favourable interaction is expected between those two groups. Indeed, Rosetta gives an interaction energy close to zero. This means that the contact is energetically neither favourable nor unfavourable. The proximity is probably a consequence of other, favourable interactions, such as, the backbone-backbone hydrogen bonds (not between those two residues) which hold the two beta-strands together. Favourable, non-covalent residue contacts, such as the ARG-ASP example contact, are important for protein structure. The term 'primary contact' will be used in future sections to refer to highly favourable contacts. In contrast, residue contacts with an interaction energy

close to zero, such as the ARG-LEU example contact, arise only as consequence of the protein structure. These contacts will be referred to as 'secondary contacts'. In a way, secondary residue contacts can be regarded as noise in a RCN. Ideally, an RCN should contain primary contacts but not secondary contacts - at least there should be the possibility to distinguish between them. A residue contact energy network can be filtered using an energy cutoff to obtain a network of only highly favourable residue contacts. Moreover, such a network is smaller and therefore easier to interpret than a purely distance based RCN.

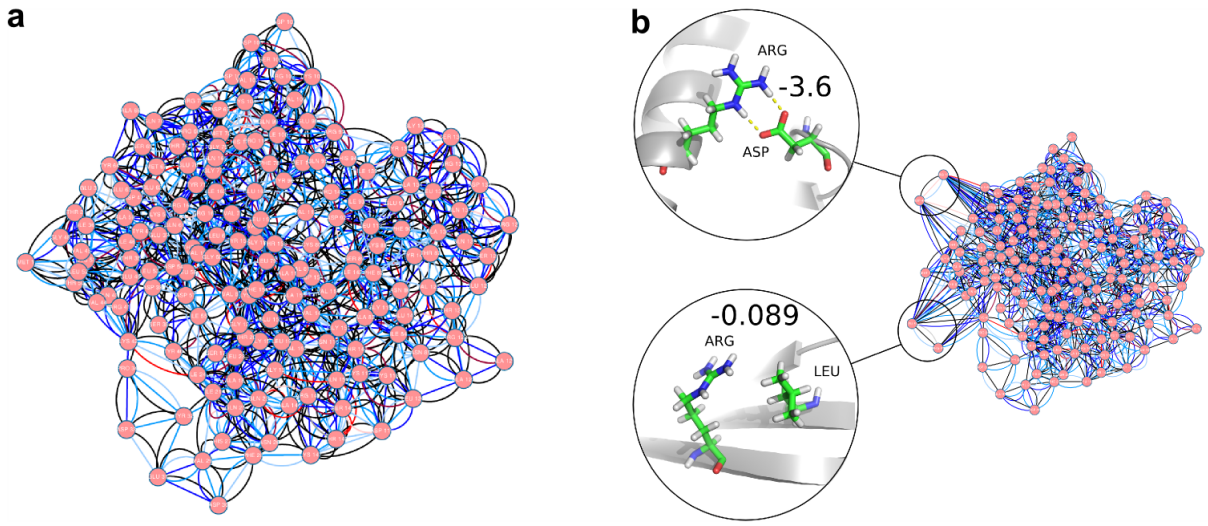

FIGURE 1.3: a: Residue interaction network of HRas created with RINerator (PDB-ID 3K8Y) and visualized in Cytoscape [22]. b: The RINerator network was cross-referenced with Rosetta energy calculation to obtain residue contact energies. Two example contacts are shown. The upper contact is energetically highly favourable while the lower one has an interaction energy close to zero. Interaction energy is given in Rosetta Energy Units (REU) which is explained in section 2.3.1.

In summary, adding interaction energy is a substantial improvement of traditional RCNs. For this reason, this project aims at incorporating interaction energy into large scale calculation of RCNs. How this was achieved by using tools from the Rosetta protein modelling suite [40] will be described in depth in section 2.3 of chapter 2.

1.3.2 Aim 2: Develop a bioinformatics pipeline to compare multiple residue contact networks

Multiple papers have been published which show the usefulness of comparing multiple RCNs as reviewed in section 1.1.3. In this project, residue contact energy networks of 511 small G proteins were calculated and compared. The protein family of small G proteins is described in section 1.2.

Both, residue contact energy networks of individual structures and comparative analysis of multiple distance-based RCNs have been shown in earlier work to give important insights into the structure-function relationship of proteins. However, to my knowledge, no study has yet combined the two methods. This new approach gives two values for each contact: contact conservation (fraction of proteins which have an equivalent contact) and energy. Comparing the two values allows to ask whether conserved contacts (shared across all or most proteins of a protein family) are more likely to be energetically highly favourable. Another advantage of ths approach is the possibility to extract a sub-network of contacts which are both conserved across related proteins and energetically highly favourable. Such residue contacts are expected to be structurally important. To test this hypothesis, it is interesting to compare the results to data from protein stability measurements (see next section: 1.3.3).

In summary, combining multiple residue contact energy networks is a novel approach which allows to investigate the relation between contact conservation and contact energy. Section 2.4 describes how a bioinformatics pipeline was built for this purpose.

1.3.3 Aim 3: Compare residue contact energy networks to mutational data

One aim of this project is to extract the residue contacts which define the common fold of a given protein family. These contacts will be referred to as the structural scaffold of the protein family.

The first step towards identifying the structural scaffold is to find contacts which are conserved across all or most members of a protein family. This step requires a common residue numbering system as explained in section 2.4.1. In order to identify the structural scaffold, it would be intuitive to only consider contacts present in all members of the protein family. However, in this case, a threshold for contact conservation of 0.9 was applied which means that contacts present in at least 90% of all structures are considered. The reason for this lies in the large number (511) of small G protein structures. There are several possibilities why a structural scaffold contact may not be found in every structure. For once, not all side chains are refined in all structures. Plenty of cases were found where only the Cβ was present in the PDB-file. If there is no or only ambiguous electron density for the side chain, atoms or residues can be missing in PDB files. However, this means that it is possible for structural scaffold contacts not to be present in every single small G protein structure. Given the large number of structures, it is impossible to manually asses all individual contacts. Therefore, a contact conservation of 0.9 or higher was deemed appropriate to consider a given contact as potential part of the structural scaffold.

Contact conservation alone is not yet sufficient to define structural scaffold contacts. Even a conserved contact could be a secondary contact (with near zero energy). The reason for this is that the protein fold can force two residues into proximity even with near-zero contact energy. Given the common fold, such a contact could still be conserved. For this reason, contact energy, calculated with Rosetta was used to separate primary from secondary contacts.

The result is a network of highly favourable, conserved residue contacts. This network is hypothesized to represent the structural scaffold of a protein family. Hence, mutation of the corresponding residues is expected to destabilize the structure. To test this hypothesis, the presumed structural scaffold RCN is compared to mutational data. For this purpose, the effect of individual residue mutations on protein stability is required. An alanine scanning study of Gα, performed by Sun et al. and analysed in collaboration with Tilman Flock, was used as experimental validation in this thesis [41]. For each residue, an alanine mutant (or a glycine mutant if the original amino acid is alanine) was created and its thermo-stability was compared to the wild type. A literature search revealed that no comparable study has been done on small G proteins. Nevertheless, to test the principle of identifying the structural scaffold, the Gα family is suited equally well. Exploiting the semi-automated workflow (see section 2.4.2), the entire computational analysis could be repeated for Gα in a relatively short time. Then, the question was asked whether residues from the structural scaffold are also identified as structurally important by the alanine scan. Indeed, the highly stabilizing residues identified by the alanine scan are contained within the computationally proposed structural scaffold. The results of this study are described in section 3.2.

1.3.4 Aim 4: Investigate relation between contact energy and contact conservation

The approach of this project gives RCNs with two parameters for each contact: energy and conservation (explained in section 1.3.1 and 1.3.2, respectively). An interesting question to ask is whether there is a relationship between those two parameters. In other words, are contacts which are present in most or all proteins of a protein family more likely to be energetically highly favourable? Although there is no linear relationship between the two parameters, conserved contacts are on average more favourable than non-conserved contacts. The findings from this analysis shed light on the method or consensus RCNs as discussed in the next chapter (section 3.3).

1.3.5 Aim 5: Explore the relation between sequence conservation and contact conservation

Somehow, the information determining similar protein folds has to be encoded in the sequences, even if the sequences share little similarity. Two explanations come to mind. The first possibility is that only few conserved residues suffice to give rise to similar folds. This would mean that most residues can be varied throughout evolution without changing the overall fold. The second possibility is a conservation of residue contacts rather than residues defines the common fold. As described earlier in the context of co-evolution(section 1.1.5), the effect of mutating one residue can be balanced by a mutation the contacting residue. This means that biochemically similar interactions can be made by different amino acids. In the simplest case, a contact between two residues can maintain its biochemical properties if the two amino acids switch place. This concept is illustrated in figure 1.4. The sequence (at that two positions) would be highly dissimilar but give rise to a biochemically similar interaction.

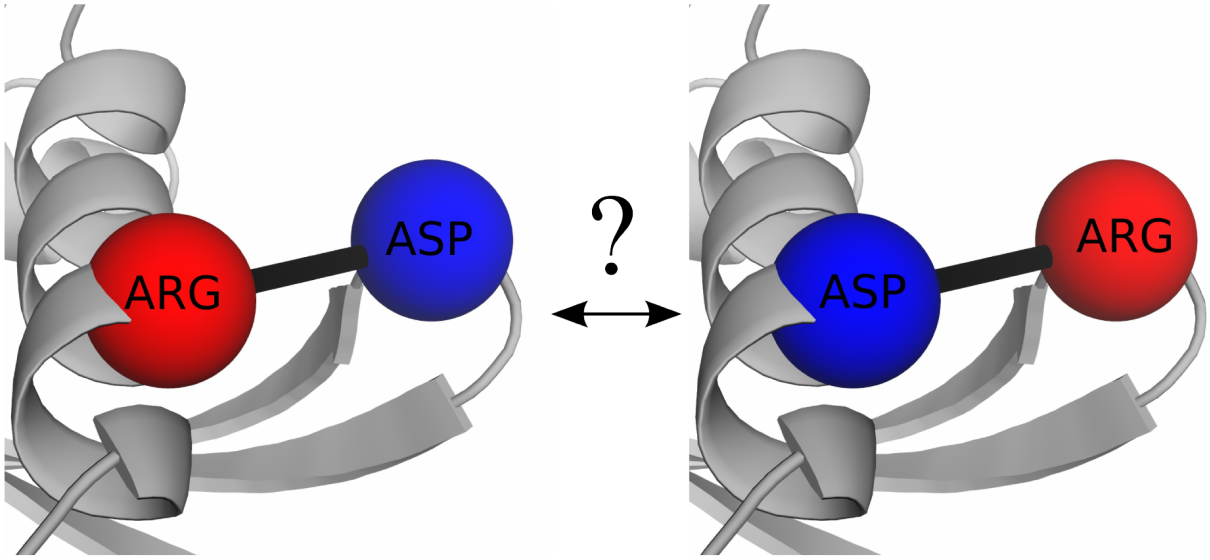

FIGURE 1.4: Concept of comparable residue contacts of different sequences. On the left, a real residue contact from the HRas protein is shown in a simplified representation. On the right, the labels have been exchanged to illustrate the question whether a similar contact could be formed if the positions of the residues were exchanged.

In this project, conserved and highly favourable contacts are identified. Then, the question is asked if they tend to be sequence-conserved (i.e. the same or very similar amino acids in all proteins) or not. Non-sequence conserved residues forming conserved contacts would be a strong indication for the co-evolutionary explanation. This would be an elegant model to explain how diverse sequences can fold into similar structures. On the other hand, if residues

forming conserved contact tend to be conserved in sequence as well, this would indicate that a small number of conserved residues is sufficient to define a common fold. Indeed, the results show that resides forming conserved, highly favourable side-chain contacts tend to be sequence-conserved. Furthermore, the mutual information between contacting resides, obtained from a large alignment of small G protein sequences, is compared to contact conservation and contact energy. No increased level of mutual information was found between residues forming conserved, highly favourable contacts. These results are described in detail in section 3.4.

Chapter 2

Results A: Method development

2.1 Overview

Analysing residue energy contact networks of a whole protein family presents several challenges. First, all suitable structures for the protein fold of interest - in this case small G proteins - have to be selected and downloaded. Then, the residue contact energies have to be calculated for all the structures. In order to compare different residue contact networks, equivalent residues in all the structures have to be identified. This is achieved by creating a common residue numbering system. Finally, all these steps have to be automated in order to ensure reproducibility as well as to allow time efficient recalculation of the results with different data.

2.2 Selecting structures

The aim was to retrieve all structures of small G proteins in the PDB. The family of small G proteins is defined in the Pfam database [42]. The Pfam identifier 'PF00071' was used to retrieve all 543 structures of small G proteins from the PDB. Out of these 543 structures, 32 were excluded from analysis due difficulties with unusual residue numbering (e.g. 1A, 1B, 2A, 2B, 3, 4, etc.). Due to the small number of problematic PDB files, it was decided to exclude them from analysis for simplicity. Section 5.2.3 provides a detailed description of this issue.

2.3 Aim 1: Energy-based residue contact networks

In the beginning of the project, the GROMACS module g_energy was considered for the calculation of interaction energies. This module has been successfully used to calculate residue interaction energies in previous studies as described in section 1.1.2[23][25]. GROMACS uses a state of the art force field and provides as good an estimate of interaction energy as can be obtained computationally. Moreover, it is perfect for analysing RCNs over the course of molecular

dynamics (MD) simulations as it allows to perform all analyses within GROMACS without the need of additional computational tools. However, in order to use this module, the same preparation required for an MD simulation has to be performed (adding explicit solvation and ions, energy minimization of the protein structure, equilibration of the solvent). Moreover, a suitable forcefield has to be chosen. Force fields for MD simulation of proteins contain parameters for all atoms and bonds encountered in the 20 standard amino acids. However, non-protein chemical groups (labeled 'HETATM' in PDB-files), such as nucleotides or metal ions are not parameterized in all force fields. Thus, for every MD simulation, a force field has to be chosen which contains parameters for all chemical groups in the system. In general, using MD simulation as an 'out of the box' method (i.e. using a default setting without consciously and carefully choosing every parameter) is not recommended. Therefore, using the g_energy module would require the same expertise as performing MD simulations. A large scale analysis of 511 structures using GROMACS would be a project in itself and not feasible as only one part of a master's project. For this reason, a simpler method for calculating residue interactions had to be found. Nevertheless, in a larger project, it would be highly interesting to use GROMACS to analyse residue contacts over the course of MD simulations.

As a more practical alternative to GROMACS, the Rosetta modelling suite [40] was chosen. Rosetta energy calculation requires less preparation than GROMACS as it uses continuous solvation rather than explicit solvation. Moreover, it requires less computational resources and all steps are easily automated (using the python interface 'PyRosetta'). The next sections will describe how the Rosetta energy function works and how it was applied in this project.

2.3.1 Rosetta energy function

The Rosetta protein modelling suite was originally developed for the protein structure prediction competition CASP (Critical Assessment of Protein Structure Prediction) [43]. However, many features, such as the energy function, turned out to be useful for other applications as well [44]. Rosetta's all-atom energy function explicitly models hydrogen atoms but uses implicit solvation instead of representing water molecules in atomic detail. It comprises several individual energy terms, such as the Lennard-Jones potential (split into an attractive and a repulsive component), a solvation potential which penalizes burial of polar groups, an electrostatic term and hydrogen bond potentials (classified into short range and long range as well as side chain interactions and backbone-backbone interactions.) Additional terms consider side chain conformations as well as amino acid preferences for certain phi/psi angles [40][45]. From

all these energy terms, the total energy of a state - for instance a certain side chain conformation or an interaction between two groups - can be derived.

Instead of physical energy units (Joule), Rosetta uses generic units on an arbitrary scale called Rosetta Energy Units (REU). The reason for not using SI units is that the energy function is a combination of physics-based and statistical potentials. Using SI units would mislead the user by suggesting a purely physical energy calculation. Moreover, energy values are not necessarily consistent between different protocols [46]. For instance, results from the coarse grained energy function, used in the early stages of structure prediction, are not consistent with the full-atom energy function. Therefore, an arbitrary scale rather than a fixed scale is used by Rosetta. It is interesting to note that Rosetta Energy Units can be converted into Joule using known experimental energies as benchmarks. Comparing Rosetta energy with experimentally measured energy gives a conversion factor which is specific for a given protocol [47]. This project aims at distinguishing between strong and weak or even unfavourable contacts. Relative residue interaction energy values specific for the protocol at hand are sufficient for this purpose.

In summary, the Rosetta energy function is well suited calculating residue contact energy in this project.

2.3.2 Residue contact energy

The energy function is used in a variety of applications of the Rosetta protein modelling suite, such as structure prediction, structure relaxation, docking, interface energy calculation and protein design. The Rosetta program used in this project is called 'residue energy breakdown' and sums up atomic energy terms on a per residue basis. The output is a list of pairwise residue interactions split into various energy terms. The term total energy is given as the sum of all individual energy terms [48]. Residue energy breakdown is a command line application which outputs a comma separated text file where residues are numbered as the occur in the structure, starting at 1 (which does not necessarily correspond to residue numbers in the PDB-file). In order to quickly execute the program for hundreds of structures, a python script was written to automate the execution for multiple PDB files as well as to simplify the output (by retaining only the total energy) and write it to a single text file. In addition, the script uses PyRosetta (a python interface to Rosetta programs) to output a file which links PDB- and Rosetta-numbers for all residues in all structures. Moreover, the script contains checks for inconsistencies in the Rosetta numbering (caused by unconventional residue numbers in the PBB-file, such as 1A, 2A, 1, 2, 3, etc.) and automatically excludes structures from analysis if the Rosetta residue

numbers cannot be unambiguously mapped to PDB-positions. See section 5.2.3 for a more detailed explanation of the challenges of residue numbering.

The semi-automated pipeline allows the same analysis to be applied to a different protein family. This has been done with the Gα family as described in section 3.2. What is more, the analysis can be repeated on the same set of structures pre-processed in a different way. For instance, the structures were relaxed using different protocols (see section 2.3.3) and their effect on residue contact energy was investigated.

2.3.3 Rosetta structure relaxation

Investigating residue contact energy of PDB-structures revealed the presence of a few outliers with highly repulsive energy (large positive values). While even highly favourable interactions, such as salt bridges, never exceed -5 Rosetta Energy Units (REU), repulsions can be up to several hundred REU. One can safely assume that a folded protein cannot have an unfavourable interaction that would require up to 50 salt bridges to compensate for. The van der Waals energy term (derived with the Lennard-Jones potential) was found to be responsible for the repulsive energy. Manual inspection revealed an overlap of the van der Waals radii of at least two atoms from the contacting residues. Overlapping van der Waals radii would indeed be highly unfavourable according to the Lennard-Jones potential. Therefore, Rosetta is right to indicate high repulsion in these cases. The issue is described in literature and structure relaxation is recommended to address it [49].

Protein structure relaxation means moving the structure into its energetic minimum by (usually slight) alterations of its atomic positions. This can be done with the Rosetta-relax application which uses the atomic energy function described in earlier (2.3.1). One might argue that an experimentally determined structure of a folded protein is supposed to be already in an energetic minimum thus rendering structure relaxation obsolete. However, there are several arguments in favour of relaxing a structure before analysis. First, the Lennard-Jones potential is highly sensitive to inter-atomic distances. If two atoms are only a fraction of an Ångström closer than their energetically ideal distance, an otherwise favourable interaction can become highly repulsive. Sparse or missing electron density can make the placement of side chains difficult and the final coordinates may not be unambiguously derived from the electron density in every structure. The second reason applies to complex structures. In some structures, the small G protein is bound to another protein or there are multiple small G proteins in close proximity in the unit cell. However, this project considers only small G proteins. One small G protein chain is extracted from each PDB file regardless of how many other proteins or chains are

present in the PDB file. The energy minimum in the experimentally solved complex structure might not be exactly the same as the minimum of the small G protein in isolation. Relaxing the protein in isolation addresses this issue. Finally, relaxing all structures using the same protocol makes the subsequent energy calculation more comparable across structures. Along these lines, the documentation of the Rosetta relax program states that structure relaxation does not necessarily make structures objectively more correct - it just makes them better suited for analysis with Rosetta [50]. That being said, clashes generally represent a defect in the structure.

When relaxing experimentally solved protein structures, it is advisable to keep the deviation from the original structure very small (which does not necessarily apply to predicted structures). In its default setting, structure relaxation places side chains as well as the backbone into the calculated energetic minimum. Such a protocol can alter the structure considerably. To limit deviation from the original structure, the relaxation program provides options, such as fixing main chain atoms to their start coordinates and limiting the movement of side chains. The protocol used in this project keeps the backbone fixed during relaxation but does not restrict side chain movement. Given the high number of structures investigated (511), it was not possible to assess them individually. Therefore, it was decided not to restrain side chain atom coordinates in order to reliably remove all clashes from all structures. On the other hand, fixing the backbone ensured that the overall secondary and tertiary structure remained unchanged. The relaxation program applies several cycles of side chain repacking. As the energy of a given side chain rotamer depends on neighbouring side chains, iterative repacking is required to find the energetic minimum of all side chains [49]. The protocol applied in this project performs five cycles of rotamer repacking. A more thorough protocol using 15 cycles has also been tested and was found to give roughly the same results. A possible explanation is that experimentally derived structures already have reasonable side chain positions and hence do not require many cycles of rotamer repacking to find the energetic minimum.

Apart from choosing appropriate parameters, structure relaxation represents computational challenges. Relaxing a short protein such as a small G protein (using the quick protocol of five cycles) takes about an hour on one CPU. As each cycle of rotamer repacking depends on the previous cycles, the process of relaxing a single structure cannot be parallelized. Therefore, relaxation of 511 small G protein structures requires about 511 CPU-hours. It is obvious that this task cannot be done on personal computers and requires cluster computing. Furthermore, the process of starting a relaxation process for each structure also has to be automated in order to ensure that the process can be started easily and quickly using different parameters or structures. A python script was written to submit a structure relaxation process for a number

of PDB-files to a cluster of CPUs. One relaxation process is executed on one core of the cluster (i.e. one CPU). In this case, the script submitted 511 structure relaxation process to 511 independent processing units. Naturally, the same parameters are used for all structures. Using this automated approach it was possible to apply the same relaxation protocol to Gα proteins as well in order compare residue contact energy calculation with mutational (see section 3.2).

In summary, structure relaxation is necessary for residue contact energy calculation using Rosetta. The process has been automated for application to multiple structures so that all structures of a given protein family can be simultaneously relaxed using the same protocol.

2.4 Aim 2: Automated comparison of multiple residue contact networks

2.4.1 Common residue numbering

In order to compare residue contacts of multiple structures, it is necessary to define which residues are structurally equivalent in different structures. This is not a trivial task as even for the very same protein, various structures differ in the number of residues due to the presence of purification tags or missing electron density. While this issue could be overcome by using UniProt-positions rather than PDB-positions, comparing protein structures form different species requires a more sophisticated approach. The first intuition could be to create a sequence alignment. However, in a protein family with low sequence conservation, such as small G proteins, multiple reasonable sequence alignments are possible. While sequence similarity between different small G proteins is low, they are structurally highly similar. Therefore, structural alignment is the only way to reliably identify equivalent residues in a diverse protein family, such as small G proteins.

The software MUSTANG was applied for structural alignment. This program superimposes multiple structures by translation and rotation in a way that minimizes the root mean square deviation (RMSD) of the Cα atoms [51]. From the superposition, a sequence alignment is created in which the columns indicate residues that are close in 3D space in the superposition. The approach succeeds at identifying structurally equivalent residues in secondary structure elements (SSEs) such as α-helices and β-strands. However, intrinsically disordered regions, such as loops, are variable between structures and cannot be directly superimposed which leads to the insertion of many gaps. The result, shown in figure 2.1 (top), is a very long alignment in which (SSEs) are well aligned but loop regions contain many columns with mostly gaps.

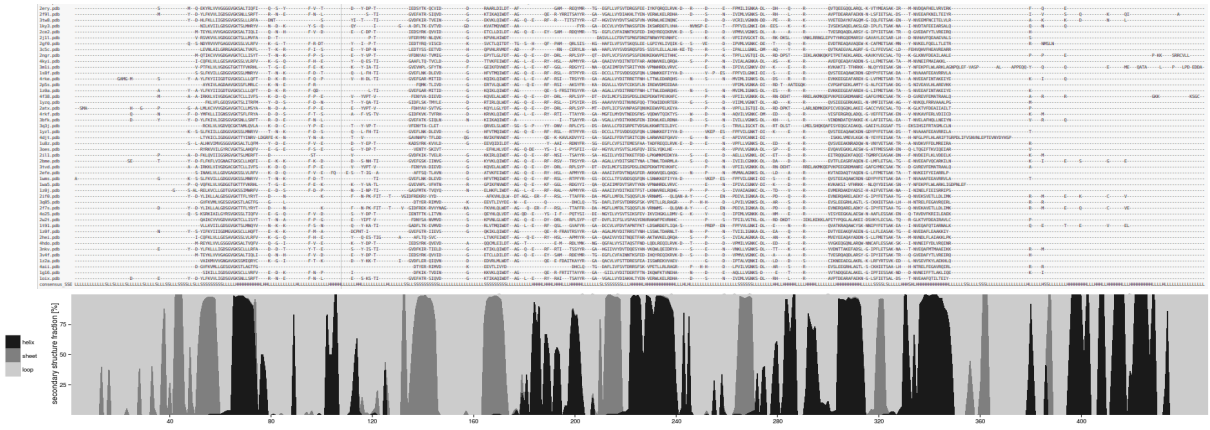

FIGURE 2.1: Raw output from the MUSTANG structural alignment software (top). The fraction of structures with a given secondary structure is provided for each position (bottom). For simplicity, only 47 out of 115 sequences are shown. Many columns with mostly gaps are visible in the alignment illustrating the need for improvement.

To address this issue, the loop regions have been aligned with the MUSCLE sequence alignment tool [52] while the alignment of SSEs was not altered. The bioinformatics toolkit UGENE [53] was used for this purpose as it allows to apply an operation only to selected columns of an alignment. This step requires secondary structure annotation for all 115 proteins in the alignment. While secondary structure information is present in some PDB-files, it is missing in others. Moreover, when comparing many proteins, the secondary structure should be calculated with the same method for all proteins. Therefore, the secondary structure of all proteins in the alignment (115) was computed using the widely used algorithm DSSP[54]. Unfortunately, integrating secondary structure information from 115 (poorly formatted) text files is not trivial. A python script was written to extract the secondary structure annotation of every residue, map it to the alignment and repeat this process for every protein. The output gives, for every position in the alignment, the number of structures with a given SSE such as α-helix or β-sheet. For simplicity, all eight secondary structure categories, given by DSSP, are summarized into three categories: helix, sheet, and loop (disordered). 'Helix' is assigned to the DSSP-types α-helix, 3_{10}-helix and π-helix. DSSP also distinguishes between sheets which consist of at least two consecutive residues forming β-bridges and isolated β-bridges. The category 'sheet' was assigned to consecutive β-bridges only. Finally, all other categories where summarized as loop. This includes the DSSP-types hydrogen bonded turn, bend, isolated β-bridge and 'uncharacterised' (no-output) which stands for irregular or coil. An area-plot showing the fractions the three secondary structure categories - helix, sheet and loop - for every position in the MUSTANG alignment is given in figure 2.1 (bottom).

The final reference alignment derived from refining the MUSTANG alignment is shown in

figure 2.2. Columns represent structurally equivalent residues in different proteins as illustrated in figure 2.3. Taken together, using alignment position as a common residue numbering system allows to compare residues and contacts across different structures.

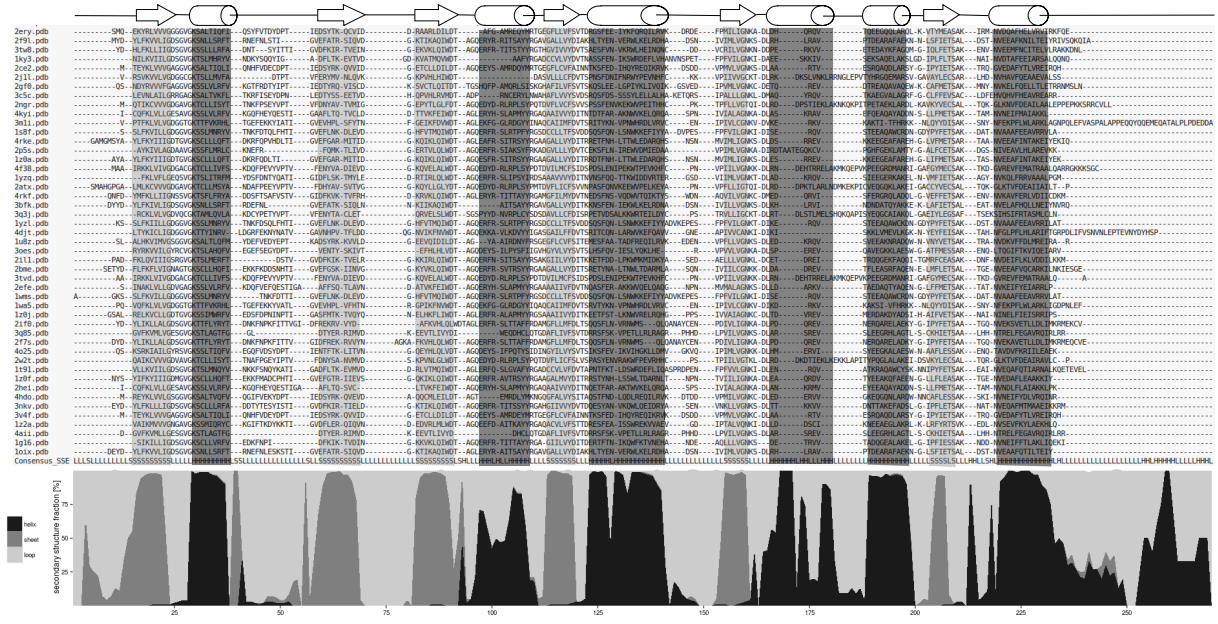

FIGURE 2.2: Refined reference alignment. For simplicity, only 47 out of 115 sequences are shown. Secondary structure elements (SSEs) are indicated on the top (helices as cylinders and sheets as arrows). The fraction of secondary structures (helix, sheet, loop) present at a given position is shown on the bottom. The columns of the alignment are coloured according to secondary structure.

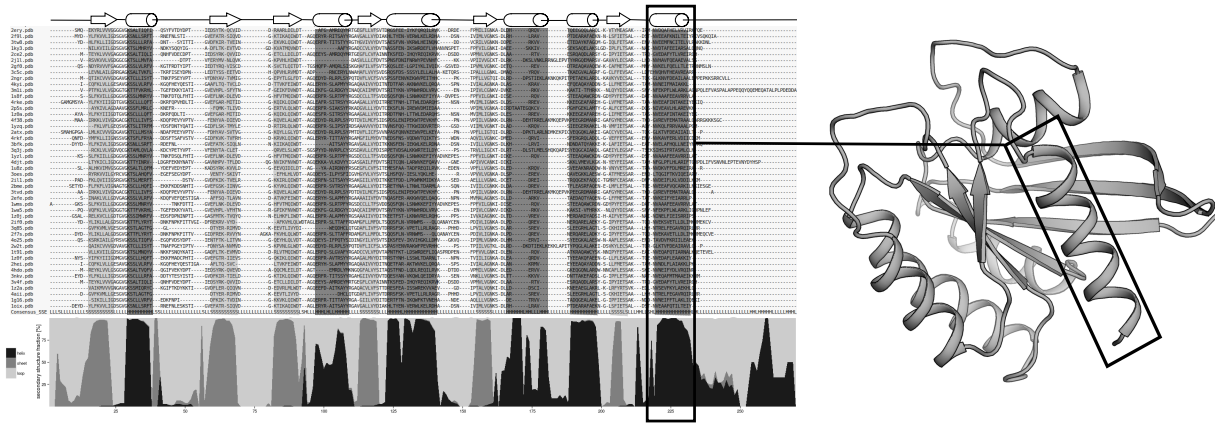

FIGURE 2.3: Reference alignment in comparison to an example structure (HRas, PDB-ID 3K8Y). The alignment is the same as in the previous figure (2.2). Columns represent topologically equivalent positions. As an example, the C-terminal helix is highlighted in both, the alignment and the structure.

2.4.2 Automated bioinformatics pipeline: workflow

In order to allow the analysis to be quickly repeated using different parameters or a different data set, all steps in the bioinformatics pipeline (except refining the alignment) were automated. An overview of the workflow can be seen in figure 2.4. Three python scripts were developed to automate structure preprocessing (selecting a single chain and checking residue numbering), structure relaxation and residue contact network calculation using Rosetta. A this point, the result was a text file giving the Rosetta energy parameters for all residue contacts in the 511 structures of the dataset. In total, this amounts to 619738 residue contacts. Residues are numbered according to Rosetta's generic numbering system which starts with 1 at the N-terminus and continues to enumerate residues as they occur in the structure. This means that residue contacts calculated by Rosetta cannot be related to the structure they were obtained from because PDB-numbers do not necessarily start at 1 and there can be missing residues. Also, there is no way to compare Rosetta's residue numbers across structures as small G proteins vary in length.

Therefore, a common numbering system was devised as explained in section 2.4.1. Structural alignment of 115 structures using MUSTANG takes about an hour to complete on a single CPU (and cannot be parallelized). This step was computed on a cluster although, in theory, it could also be done on a personal computer. The subsequent manual refinement of the alignment, described in section 2.4.1, is the only non-automated step in the pipeline. Therefore, when applying the pipeline to a novel protein family, most of the work goes into manually refining the reference alignment.

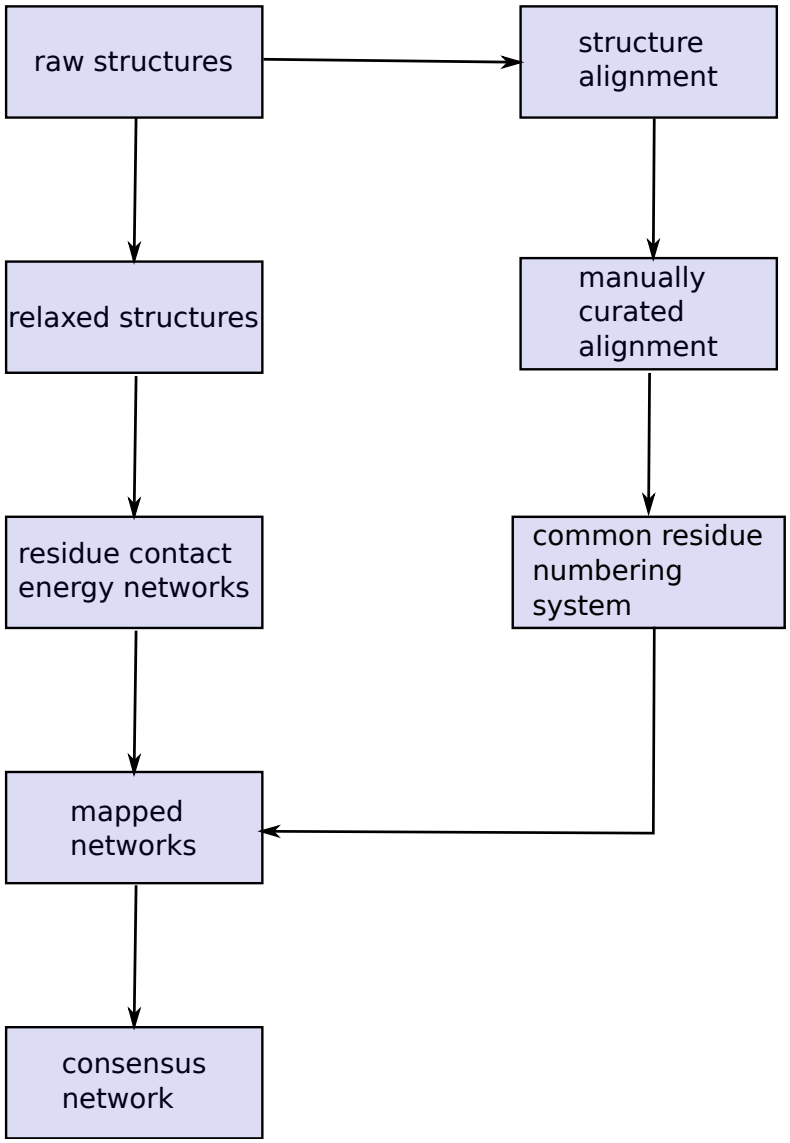

FIGURE 2.4: Overview of the basic workflow of this project. Squares represent the data at a given stage in the pipeline. Arrows show how the steps of the pipeline. Each of this steps is automated by a script, except the manual refinement of the alignment. The pipeline is semi-automated as each step has to executed individually.

Using the reference alignment, the residue contact networks are mapped to alignment position as well as to the PDB-residue numbers. The statistical programming language R was used for this task due to its useful features for large scale data analysis. The result is a text file with all residue contacts of all structures providing Rosetta-residue number, PDB-residue number and alignment position for all residues.

From the mapped residue contacts, it is easy to count how often a given contact between equivalent residues (which have the same alignment positions) occurs. This number is the dived by the number of structures which have both contacting residues (i.e. do not have a gap

at the corresponding alignment positions) to give the so called contact conservation score. This score is a number between 0 and 1, indicating the fraction of structures that have a contact between two given residues.

Finally, an extensive exploratory analysis has been done using the R markdown language knitr [55]. The findings from these analyses are presented in the next chapter.

Chapter 3

Results B: Analysis

3.1 Overview: Energy-weighted consensus contact network

Residue contact energy networks have been calculated for 511 structures of small G proteins using Rosetta. A structural alignment was used to create a common residue numbering of topologically equivalent residues which allows to identify equivalent contacts in different structures. Comparing all the residue contact networks (RCNs), a consensus RCN was calculated. This was achieved by screening all contacts from all structures and extracting only contacts which are present in at least 90% of all structures. Such contacts will henceforth be referred to as 'conserved contacts'. This term is not to be confused with sequence conservation. It means that topologically equivalent residues make the a contact in at least 90% of small G proteins - regardless of the amino acids involved. Taking advantage of the Rosetta contact energy calculation, the mean energy can be assigned to each contact as illustrated in figure 3.1. For more information on the how to derive an energy-weighted consensus network from multiple protein structures, please refer to the chapter 2.4.1.

An energy-weighted consensus network allows to characterize the common features of a protein family rather than features of a single protein structure. The complexity of the network can be reduced by focusing only on highly favourable, conserved contacts. Figure 3.2 shows a matrix representation of all residue contacts in the small G protein family. For simplicity, only long-range contacts (excluding interactions between sequence neighbours up to i+4) are shown. This matrix can be filtered by both contact conservation and contact energy. Only conserved contact are plotted in figure 3.2b. In figure 3.2c, only highly favourable contacts with an energy value below -0.8 REU are shown. On average, contacts with an energy value below -0.8 REU represent the most favourable 16.45% of the contacts in any given structure. (There is some variation between structures but the mean is fairly representative as the standard deviation is only 1.02%). The contact matrix in figure 3.2d is filtered by both contact conservation

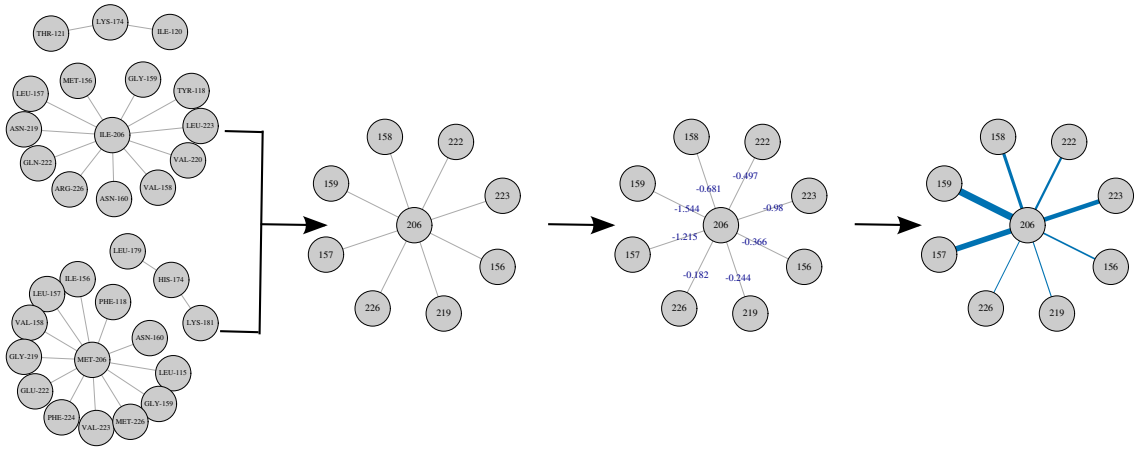

FIGURE 3.1: A consensus network is calculated from multiple individual networks. To illustrate this point, two small sub-networks from two RCNs (PDB-IDs: 3BBP, 1A2B) are shown on the left. Residues are numbered according to the common residue numbering (section 2.4.1). From these two sub-networks, the common contacts are extracted. Then, a label representing the mean contact energy (average of the energies of the contacts from the two initial networks) is added to all common contacts. Finally, contact energy is represented as edge thickness (highly negative energy values are shown as thick edges). The same visual representation of contact energy will be used in further figures.

and energy. While both filtering steps lead to a marked reduction of complexity, filtering by conservation and energy yields the smallest and most interpretable set of contacts. Parallel β-sheets can be seen as sequential contacts parallel to the diagonal. α-helices could be seen very close to the first diagonal had the contacts not been excluded to focus on long range contacts. Anti-parallel β-sheets are orthogonal to the first diagonal. Contacts between the only two anti-parallel β-strands of a typical G protein can be seen between common residue numbers 55-65 on the one hand and the residues 80-90 on the other hand. In summary, a subset of highly favourable, conserved contacts can be obtained from an energy-weighted consensus network.

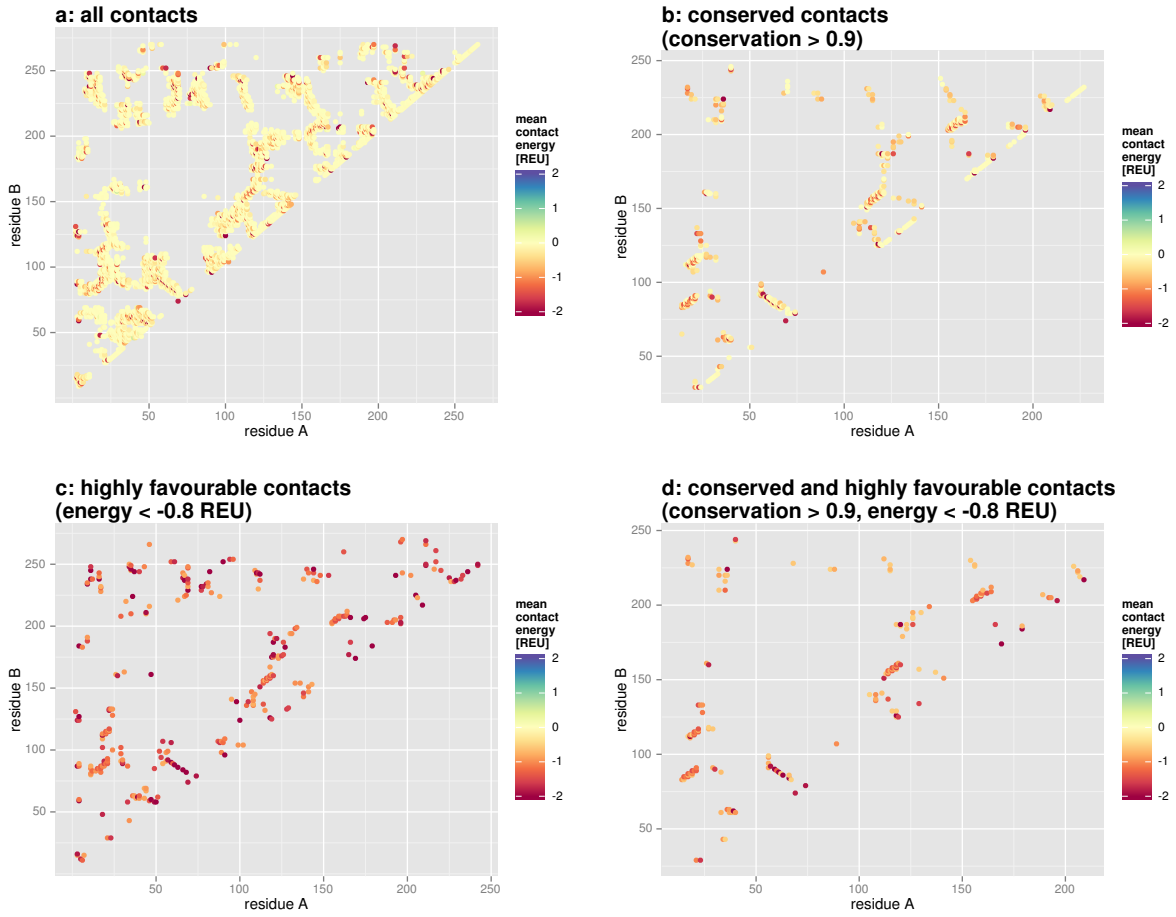

FIGURE 3.2: Matrix of residue contacts. Each dot represents a contact between two residues. The residue numbers on the axes are given in the common numbering system (section 2.4.1). The contacts are coloured according to energy. As the spectrum is limited to the range of -2 to 2, all contact with an energy value of -2 REU or lower are shown in the same colour. However, only a small number of contact energies exceed this range. Limiting the colour spectrum allows to better distinguish different contact energy values for the majority of contacts.

Figure 3.3 shows this network plotted on a representative structure of HRas (PDB-ID 3K8Y). It is reasonable to suspect that these contacts are structurally important. If that is the case, mutations of residues forming these contacts are expected to destabilize the fold. The next section will compare this network (of highly favourable, conserved contacts) to mutational data.

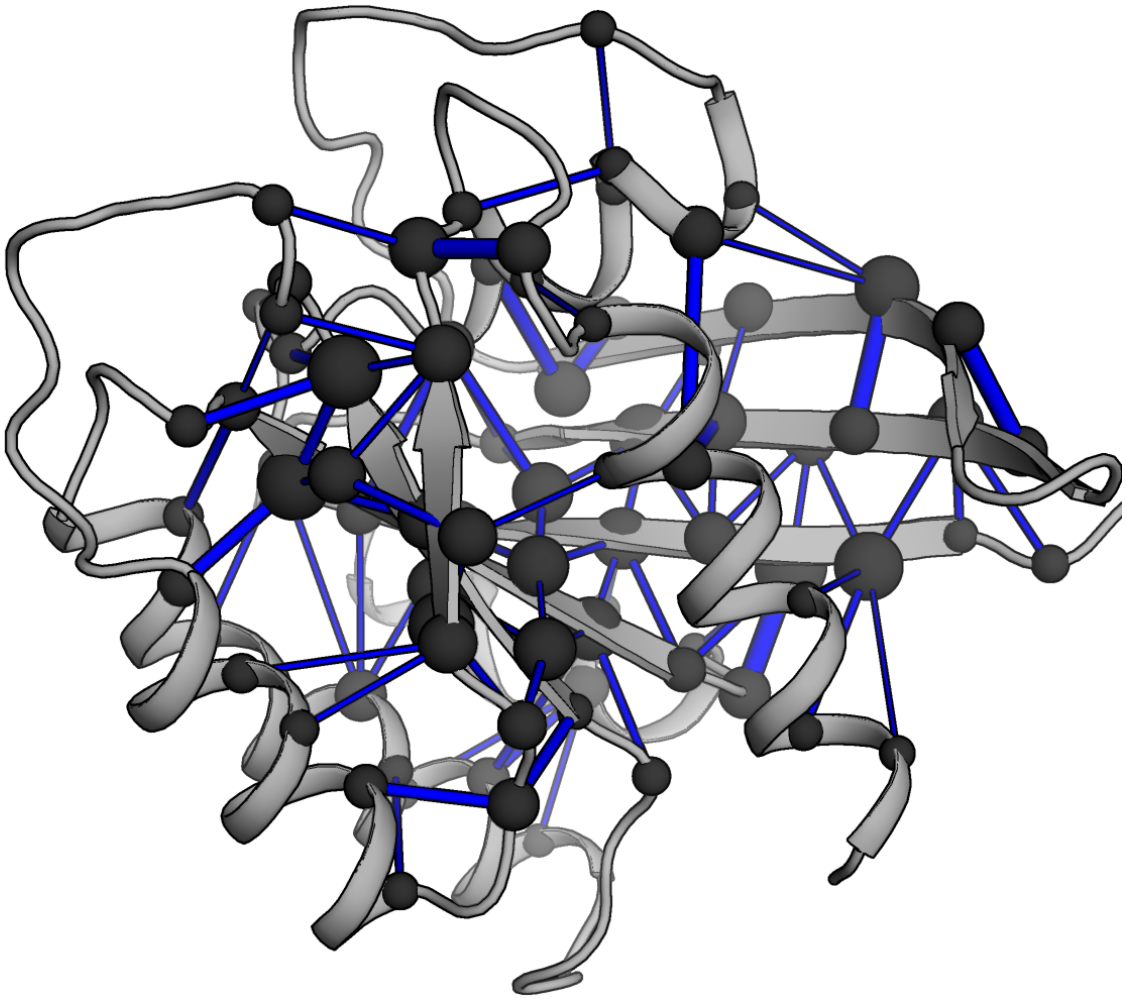

FIGURE 3.3: Conserved, highly favourable contacts plotted onto the structure of HRas (PDB-ID 3K8Y). The cut-off for contact conservation is 0.9 (only contacts present in at least 90% of all structures are shown) while a threshold of -0.8 REU is applied to include only highly favourable contacts. The thickness of the edges is proportional to contact energy. The size of the nodes (residues) is proportional to the sum of the energies of all contacts of a given residue (referred to as cumulative contact energy henceforth). An R script to map a 3D RCN onto a Gα structure (by writing a pml-file for PyMOL) was created by Tilman [13]. This script was modified and applied to a small G protein structure to create the figure at hand.

3.2 Aim 3: Comparison of the consensus network with mutational data

3.2.1 Alanine scanning of Gα

This section investigates whether energy-weighted consensus networks are suitable for predicting the structural scaffold of a given protein fold. Complete alanine scanning data is available for the Gα subunit of heterotrimeric G proteins which is related to small G proteins [41]. This dataset was used to evaluate the method of energy-weighted consensus networks as no comparable dataset exists for small G proteins. Applying the same analysis performed on small G proteins to Gα subunits was facilitated by the automated approach discussed in chapter 2. The only step in the protocol that is not automated is the creation of a reference alignment. However, for Gα-subunits, a residue numbering system, termed CGN (common Gα numbering system) was created by Flock et a. CGN is described in [13] and can be accessed via the web server [56]. Using CGN bypassed the need for creating a reference alignment. For other steps in the bioinformatics pipeline, the same scripts used for small G proteins (workflow shown in figure 2.4) could be applied.

In the Gα alanine scan, for every residue, a mutant protein was constructed with an alanine instead of the original amino acid [41]. Alanines in the original sequence were mutated to glycines. Assuming that alanine does not engage in strong interactions, mutating a residue to alanine breaks all strong interactions. If at least one of these is important for stabilizing the Gα fold, mutation leads to a lower melting temperature compared to the wild type. Contacts important for protein stability are expected to be energetically favourable and conserved across the Gα fold. Hence, residues with such contacts predicted by the computational analysis are expected to lead to a decrease in melting temperature when mutated to alanine.

For each mutant, the paper gives three values: the change in melting temperature (ΔTm) of the GDP-bound state, the ΔTm of the GTP-bound state and the change of complex formation efficiency with a G protein coupled receptor (GPCR) [41]. The authors define clusters of residues according to whether they destabilize the GDP- or GTP-bound state or interfere with complex formations. For comparison with highly favourable, conserved contacts, only residues that stabilize both, the GDP- and the GTP-bound state are of interest. The cluster of residues that stabilizes both states is termed 'stabilization cluster 2' by the authors. The next subsections explore the contact properties of those residues.

3.2.2 Stabilizing residues identified by the alanine scan have more conserved contacts than other residues

For each Gα residue, the number of conserved contacts has been calculated. Then, residues of the stabilization cluster 2 have been compared with all other residues in terms of number of conserved contacts. Given their importance for protein stability, residues from the stabilization cluster are expected to form more conserved contacts. Indeed, figure 3.4a shows that residues from the stabilization cluster 2 have a significantly higher number of conserved contacts. This result agrees with the hypothesis that conserved, highly favourable contacts identified by the computational analysis define the structural scaffold of a protein fold.

3.2.3 Stabilizing residues identified by the alanine scan make stronger contacts than other residues

The next question is how the contact energy obtained from Rosetta compares with the mutational data. The change in melting temperature is provided on a per residue basis, making it a node attribute. However, contact energy is an edge attribute. Therefore, the edge attribute has to be converted into a node attribute. This is achieved by summing up interaction energies of all contacts of a given residue. The result is a single value for each residue which will be referred to as cumulative contact energy. The more contacts a residue makes and the stronger the interactions, the lower (the more negative) this value will be. Then, the cumulative contact energy was compared between residues of the stabilization cluster 2 and all other residues. Figure 3.4b shows that residues from this cluster have a significantly lower cumulative contact energy (lower energy values mean more favourable interactions).

However, the question remains whether residues from the stabilization cluster 2 make stronger contacts or simply have a lower cumulative contact energy due to their higher number of contacts. To address this issue, the average energy per residue contact has been calculated. This value is obtained by normalizing the cumulative contact energy by the number of contacts of a given residue. Indeed, figure 3.4c shows that residues from the stabilization cluster 2 form significantly stronger contacts than other residues. However, the difference between the two distributions is less pronounced. Moreover, the p-value is larger (although still significant at the level of 0.05) than in figure 3.4a and b. Therefore, it seems that the pronounced difference in cumulative contact energy is mostly due to the difference in the number of conserved contacts figure 3.4a.

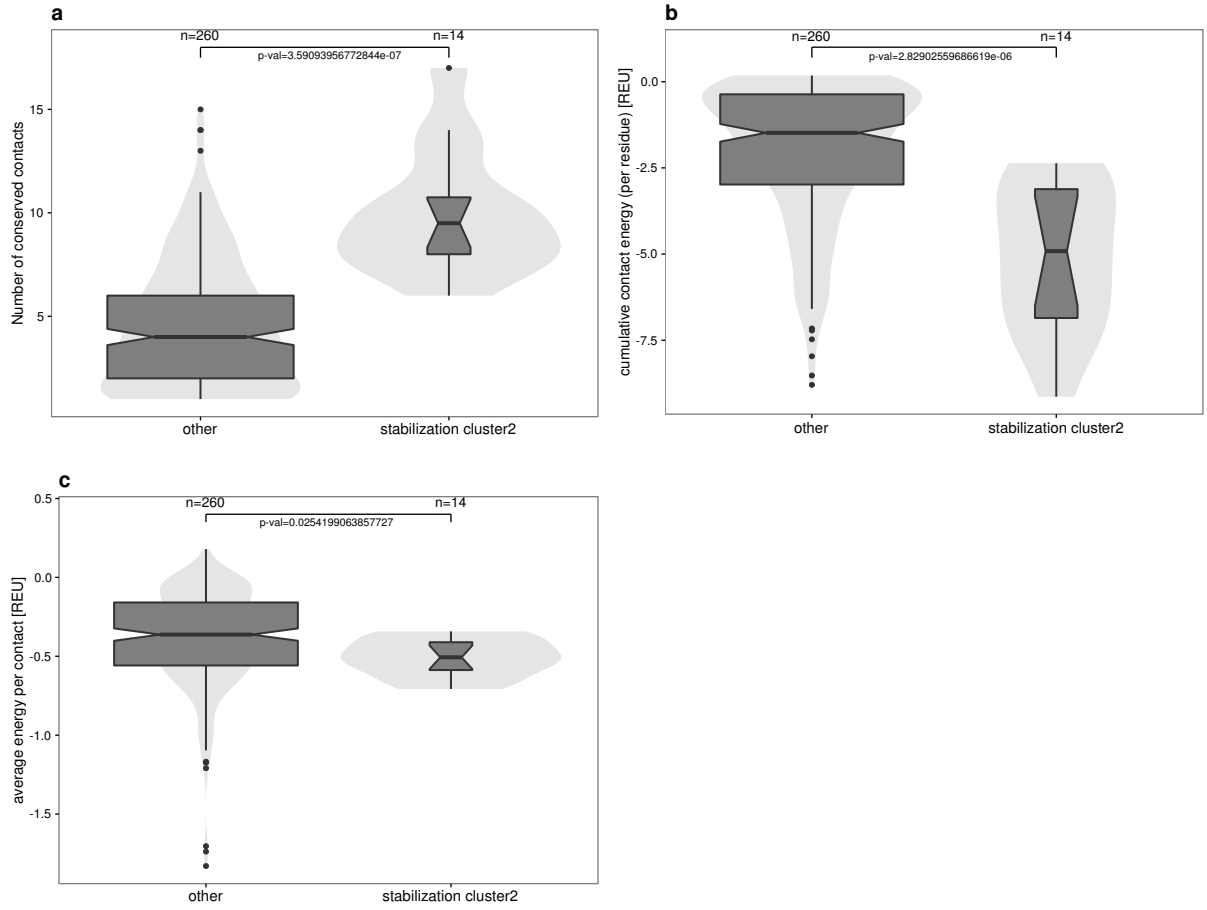

FIGURE 3.4: All contacts formed by residues from the stabilization cluster (identified as cluster of residues stabilizing both, the GDP- and the GTP- bound state in the alanine scanning study) have been extracted and compared to all other residue contacts. The width of the boxes is proportional to the number of observations in each box (indicated on top). Moreover, the underlying distribution is indicated by violins in the background. The p-values obtained from a Mann-Whitney U test are also indicated on top. (a) compares the number of conserved contacts per residue, (b) compares the cumulative contact energy per residue and (c) compares the average energy per contact. The difference between the distributions is significant (at the level of 0.05) in all three plots. The same test as well and the same significance level is used for all box-plots in this thesis. Section 5.1.5 explains the choice of the statistical test and describes in detail which parameters were used to create the box-plots.

These results indicates that strong interactions given by the Rosetta energy calculations are structurally relevant. Contacts which are present in only one state (either GDP- or GTP-bound) have to be disrupted in the course of the Gα signalling cycle (the basic cycle of GTP hydrolysis and GDP release is the same as for small G proteins as described in section 1.2). Therefore, it makes sense that such contacts are weaker than contacts present in both states. More generally, the results supports the concept of a stable structural scaffold of conserved, strong contacts on the one hand and weak contacts which can be rewired in different states of a protein on the other hand. A similar concept also applies to different proteins of the same family. While the

structural scaffold defines the common fold, the remaining contacts can vary according to the function of the individual protein. However, further research would be required to expand the concept beyond Gα proteins.

3.2.4 Conclusions from comparing an energy-weighted consensus network with mutational data

The previous section shows that most (12 out of 14) residues important for stabilizing Gα in both states (GDP- and GTP-bound) identified by alanine scanning make highly favourable and conserved contacts. However, many residues which are not found to be important for stability by alanine scanning still form strong and conserved contacts (see figure 3.5). Therefore, the method cannot be used to reliably predict the importance of residues for protein stability. Nevertheless the method succeeds at identifying a subset of residues which - among others - contains all residues indispensable for stability. However, this does not necessarily mean that the network analysis gives a lot of false positives. Melting temperature only indicates stability, not function. Therefore, those highly favourable conserved contacts of residues not identified as stabilizing in the alanine scan could be vital for the common function of Gα subunits rather than their stability.

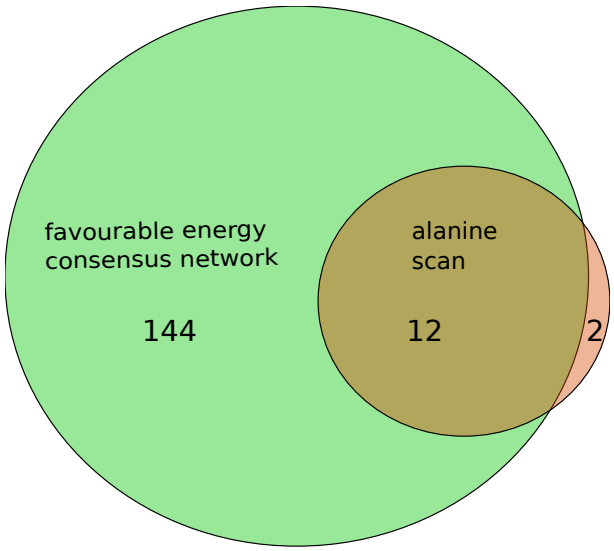

FIGURE 3.5: A network of highly favourable, conserved residue contacts was extracted. In order to be included, a contact has to be present in at least 90% of all small G proteins investigated. Moreover, the contact energy has to be below -0.8 REU (typically, this corresponds to the most favourable 16.45% of residue contacts in a single structure).

Furthermore, contact network analysis might be an interesting addition to mutational data. While alanine scanning only measures the effect of a mutation, investigating contact networks

can identify specific residue contacts whose disruption is likely to be responsible for the decrease in protein stability.

3.3 Aim 4: Relation between contact conservation and contact energy

The next question is whether contacts which are present in all or almost all structures of a protein family are also more likely to be energetically highly favourable. To address this question, all contacts were classified as either conserved (present in at least 90% of all structures), partially conserved (present in 30-90% of all structures) or non-conserved (present in less than 30% of all structures). No linear relation between contact conservation and energy was found as shown in figure 3.6a. The distributions of contact energy of the three categories of contacts are shown in figure 3.6b and c. The significance level α was set at 0.05 which is a commonly used value. All p-values in figure 3.6 are significant due to the large number of observations. However, only conserved contacts tend to be markedly more favourable than the other two groups of contacts. It is important to note that a significant p-value only means that it would be unlikely to observe the results at hand, given the null hypothesis that two distributions are the same. A significant difference does not need necessarily to be large enough to be of interest. This seems to be the case for the difference in contact energy of non-conserved and partially conserved contacts.

Taken together, this result provides interesting insights into the method of consensus network calculation. It shows that conserved contacts tend to be energetically more favourable. Therefore, even without explicitly taking energy into account, a set of conserved contacts is enriched in highly favourable contacts.

Given that, it is interesting to ask, which fraction of the total contact energy in a given protein is provided by conserved contacts. To this end, for each structure, the contact energy of variable and conserved contacts, respectively, was summed up. Figure 3.7a shows that conserved contacts (present in at least 90% of all structures) contribute more towards total contact energy than other (variable) contacts. However, there is some fluctuation between the structures. For this reason, figure 3.7b explores the dependence of total contact energy on protein length. Indeed, a significant correlation (p-value 2.2e-16) could be established. This correlation is not surprising as proteins with more residues have more residue contacts which leads to a lower total contact energy (as favourable energy is given with a negative sign this corresponds to a higher absolute energy value). While small G proteins have very similar lengths, the number of observed residues can be considerably smaller in some structures. Manual inspection

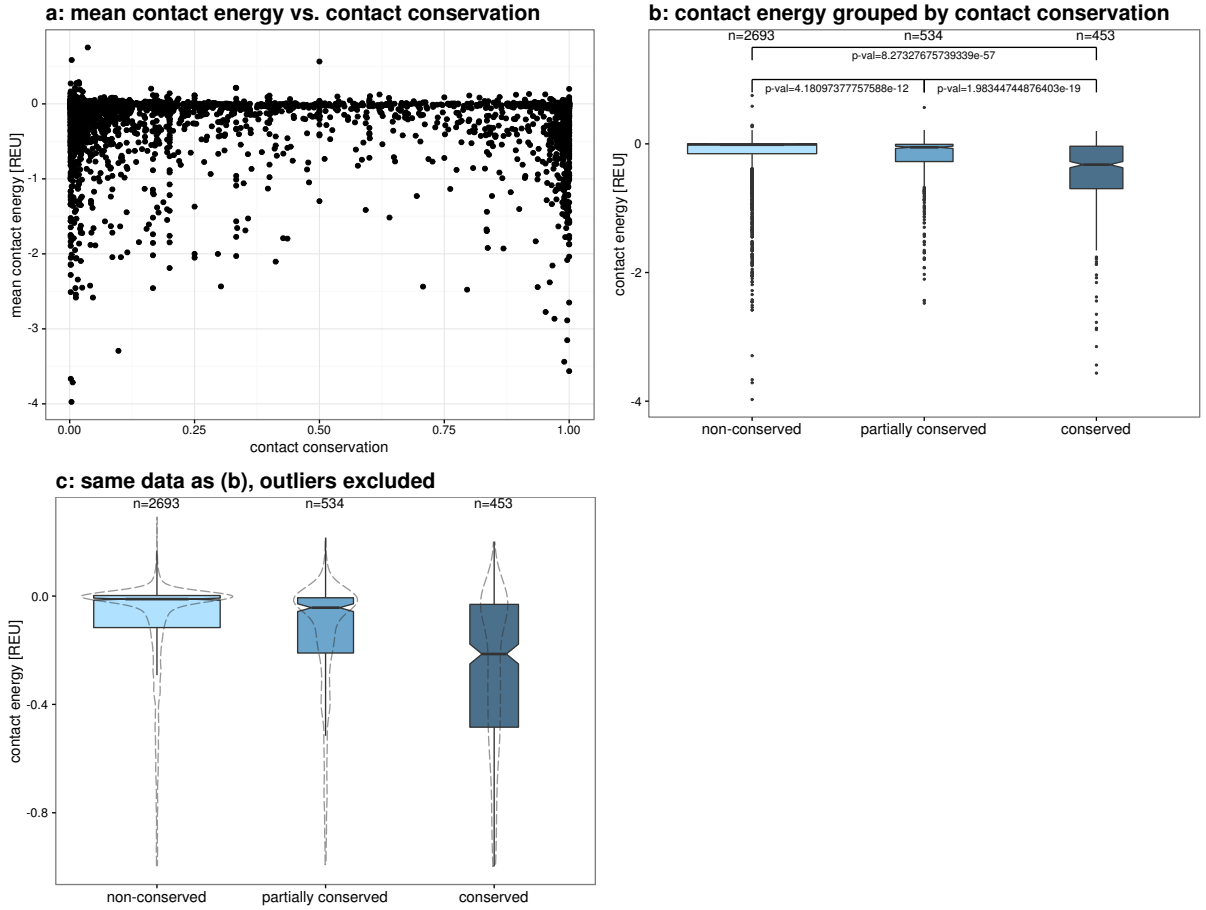

FIGURE 3.6: (a) A scatter plot shows that no linear relation exists between contact energy and contact conservation. (b) A box-plot groups contacts into three categories according to their conservation and shows a significant difference between the groups at the significance level 0.05 (p-values indicated in plot). The box-width is proportional to the number of observations (indicated on top). Due to a small number of outliers, the shape of the boxes is hard to see. Therefore, (c) shows the same data without the outliers and provides a violin plot (dashed lines) to illustrate the underlying distribution. The Mann-Whitney U test used to obtain the p-values as well as paramters chosen for the box plots are described in section 5.1.5.

revealed that many structure have missing residues and therefore fewer residues than a typical small G protein (around 166). These observations suggest that it is reasonable to normalize total contact energy by protein length.

For this reason, figure 3.7c shows the total contact energy normalized by protein length. As expected, there is a marked decrease in variation between structures compared to figure 3.7a. The total contact energy (when normalized by length) is very similar across small G proteins. The next interesting question is to ask whether the ratio of energy of conserved and variable contacts is also similar across structures. figure 3.7d shows the fraction of the total contact energy provided by variable and conserved contacts, respectively. In all structures, conserved

contacts make up about 70-75% of the total contact energy. There is only one major outlier (PDB ID: 1PLJ). While it is normal to observe missing electron density (and therefore missing side chains or residues) in flexible regions on the outside of the protein, this structure has many missing residues in the structured interior of the protein. For this reason, also conserved contacts are affected by missing electron density which explains why the structure has a lower fraction of contact energy of conserved contacts.

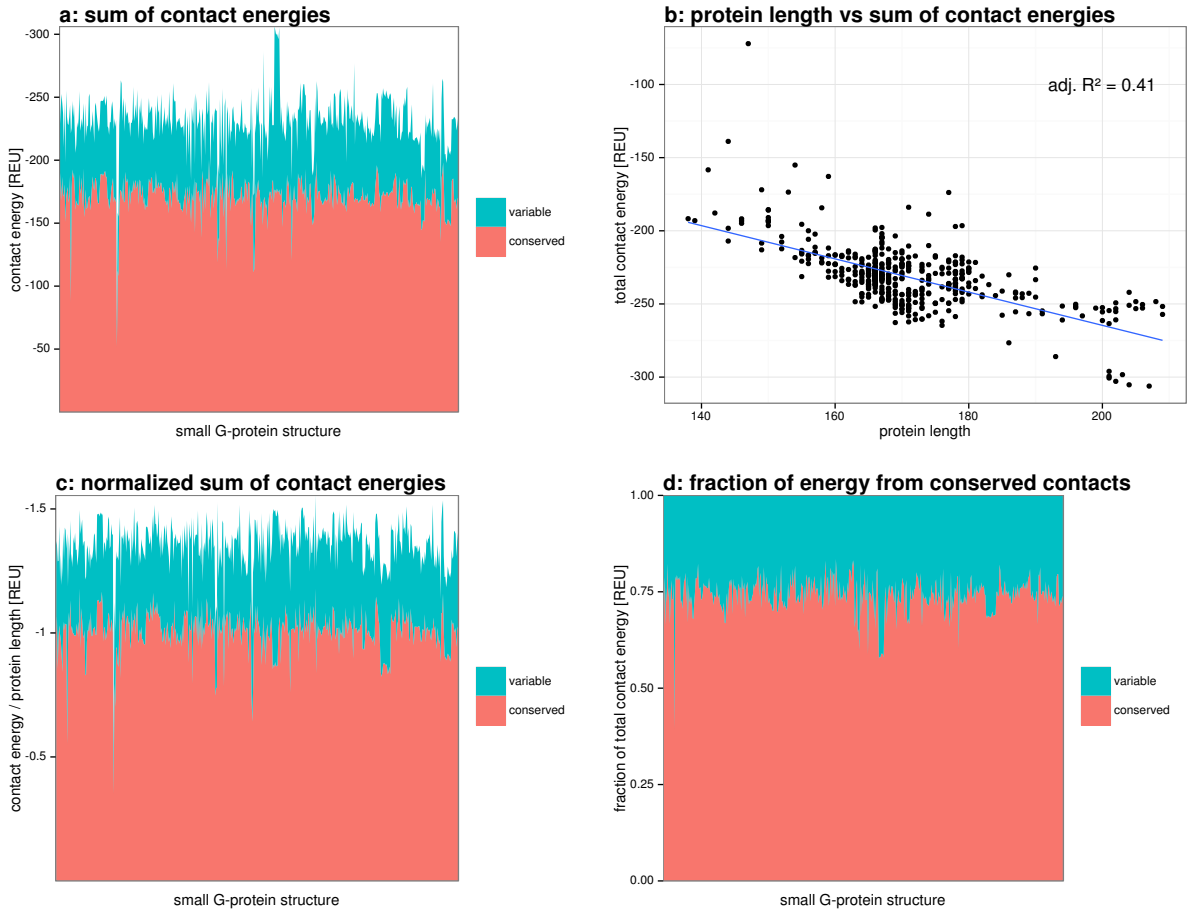

FIGURE 3.7: (a) the fraction of contact energy in conserved (present in at least 90% of all structures) versus variable contacts is shown for each of the 511 small G proteins. Cumulative contact energy is plotted on the y-axis in Rosetta Energy Units (REU). The x-axis represents individual proteins structures; for the sake of clarity, the labels for the 511 structures are omitted. (b) establishes a correlation between total contact energy (energies of all contacts in the protein summed up). (c) shows the same data as *(a)* only normalized by protein length. (d) shows the cumulative energy from conserved contacts as a fraction of total contact energy.

In summary, most of the contact energy in small G proteins is provided by conserved contacts. These contacts can be considered the structural scaffold of this fold as they are shared

across all small G-proteins. The remaining 25-30% of the contact energy is provided by variable contacts present in some small G proteins but not in others. These contacts are likely to fine-tune the protein structure and dynamics to the specific needs of a particular small G protein or subfamily. The findings described here may also apply to other protein families and may even be a universal principle of protein families. However, further research is required to confirm this hypothesis.

3.4 Aim 5: Relation between sequence conservation and contact conservation

An interesting question to ask is whether conserved contacts are only formed by sequence-conserved residues. As discussed in the introduction (section 1.3.5), this does not necessarily need be the case. Despite the highly similar fold of different small G proteins, their sequence conservation can be as low as 30%. Given surprisingly low sequence conservation, the question arises how the information for the common fold is encoded in the sequence. After all, despite the low similarity, all these sequences have to have common features to define the common fold. One explanation could be that the structural scaffold is defined by a small number of conserved residues. This would mean that only a small number of conserved residues can give rise to highly similar structures. Another possibility is that conserved contacts (i.e. contacts present in all structures) define the common fold. These conserved contacts do not necessarily have to be formed by the same amino acid in all proteins as illustrated in figure 1.4 in the introduction. This section explores which model is more suited to describe the structural scaffold of small G proteins.

First insights can be obtained by taking the 3D consensus network described previously and colouring the residues by sequence-conservation (figure 3.8). This figure shows that most residues in the consensus network are sequence-conserved and that non-conserved residues form mostly backbone-backbone contacts. The latter observation makes sense in the light of the fact that the backbone is identical in all amino acids (except proline). Therefore, backbone-backbone contacts are less dependent on amino acid identity.

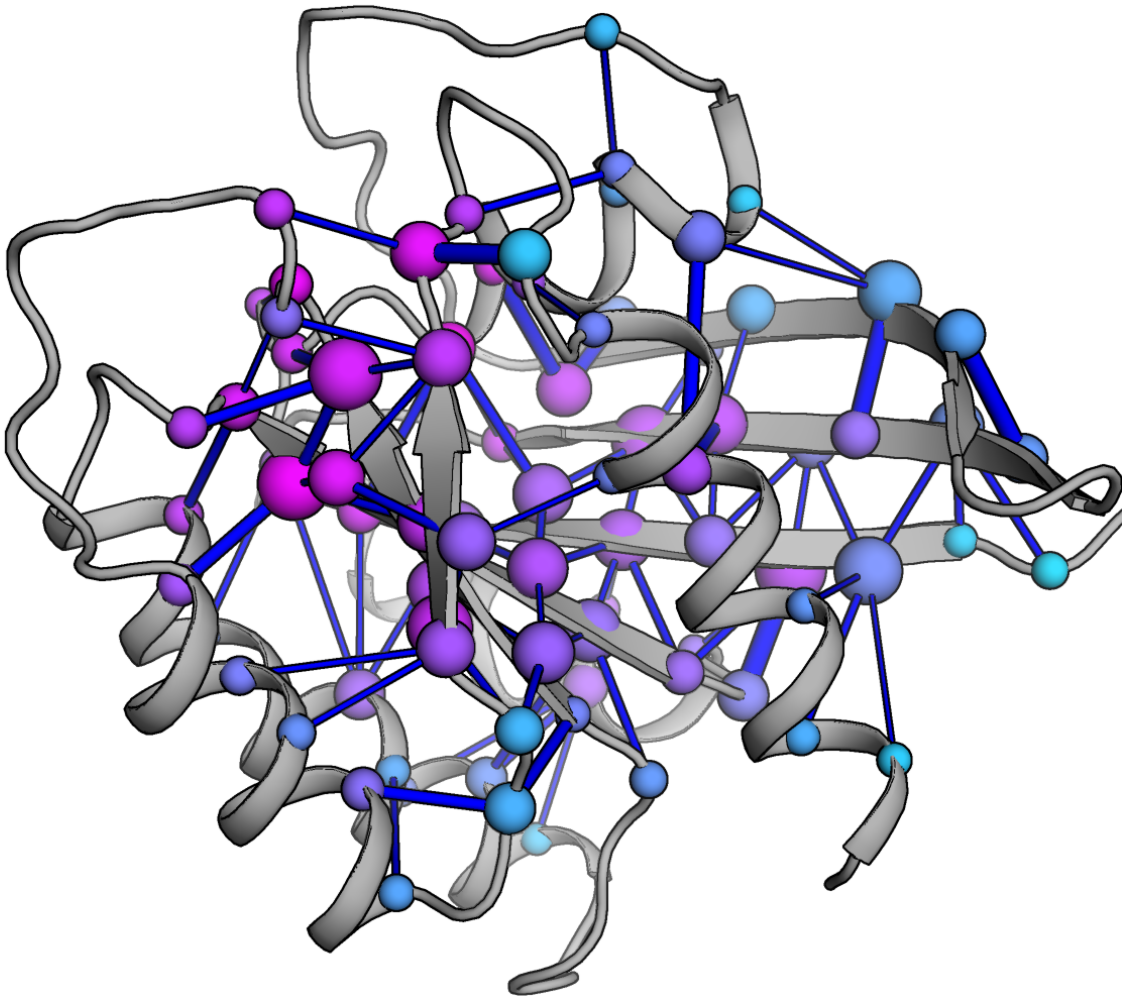

FIGURE 3.8: 3D network with nodes coloured according to sequence conservation. The network is the same as in figure 3.3 with the only difference that residues are coloured according to sequence conservation. Magenta represents conserved residues while non-conserved residues are shown in cyan.

To further investigate the relationship of contact and sequence conservation, all contacts have been grouped into three categories: non-conserved (present in less than 30% of all structures), partially conserved (present in 30-90% of all structures) and conserved contacts (present in at least 90% of all structures). Then, the sequence conservation of these three groups was compared and the significance of the differences was evaluated with the Mann Whitney U-test. Most differences were found to be significant at the level of 0.05. The only two exceptions in figure 3.9b and c are due to the small number of observations in the respective distributions.

However, in other cases, such as figure 3.9d, the differences between the distributions are very small despite being significant. These small differences are assumed to be irrelevant which is indicated by the grey colour of the p-values. In general, it is quite possible to obtain significant p-values for very small effects if the sample size is large enough. It is important to note that low p-values only mean that, given the null hypothesis, it would be unlikely to observe an effect equal or larger to the observed effect. It does not indicate whether the result is relevant, interesting or worth considering. Therefore, if an effect is too small to be of interest, it is safe to ignore despite statistical significance.

Figure 3.9a shows that conserved contacts tend to be formed by sequence conserved residues while the opposite is true for non-conserved contacts. However, there is substantial overlap between the distributions. The reason for this could be the presence of backbone-backbone contacts which depend less on amino acid identity and therefore sequence conservation. Hence, only side-chain contacts (either side chain-side chain or side chain-main chain) are considered in figure 3.9b. As expected, the difference is more pronounced. Next, only highly favourable side chain contacts (Rosetta energy values below -0.8) were analysed as shown in figure 3.9c. The plot shows that highly favourable, conserved contacts are predominantly formed by sequence-conserved residues. This result makes sense as highly favourable contacts, such as salt-bridges, hydrogen bonds, π-π and π-cation interactions, depend on amino acid identity. On the other hand, weak contacts (with interaction energy close to zero) only require side chain proximity. Finally figure 3.9d shows, as expected, no considerable difference in sequence conservation of residues forming backbone-backbone contacts. Another layer of information in figure 3.9 is the size of each group of contacts as a fraction of all contacts as represented by the width of the box-plot. The fraction of conserved contacts is highest among highly favourable side chain contacts and lowest among backbone-backbone contacts.

While there is a pronounced and significant difference in sequence conservation between conserved side chain contacts and the other two groups of contacts (figure 3.9), there is no appreciable difference between non-conserved and partially conserved contacts. Together, these two groups comprise all contacts with contact conservation less than 0.9. It is interesting to see that residues forming contacts with such a large range of contact conservation show little difference in terms of sequence conservation.

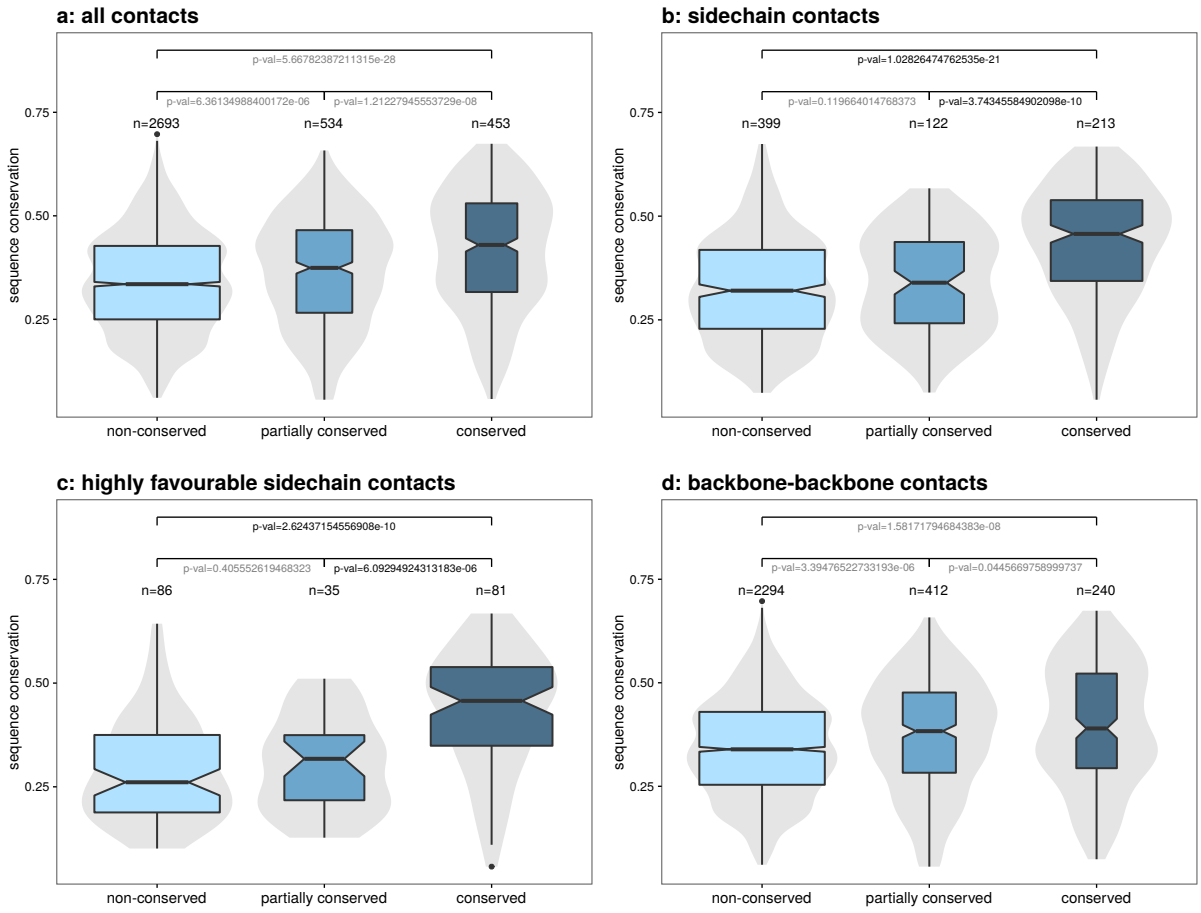

FIGURE 3.9: Relationship between contact conservation and sequence conservation. The number of observations in each group of contacts (non-conserved, partially conserved, conserved) is provided on top of the box-plot. In the background, violin plots show the underlying distribution. The width of the box-plots is proportional to the number of contacts in a given group (note that box-width can only be compared within one plot). p-values indicate the significance of the difference between the distributions. If the difference is either insignificant or too small to be of interest the p-values are coloured in grey. See section 5.1.5 for a description of the Mann-Whitney U test used to obtain the p-values as well as parameters chosen for the box plots. (a) all residue contacts. (b) only contacts involving at least one side chain: either side chain - side chain or side chain-backbone. (c) subset from the contacts in *(b)* which have an energy of -0.8 REU or lower. (d) only backbone-backbone contacts.

Taken together, figure 3.9 indicates that conserved, highly favourable contacts tend to be formed by sequence-conserved residues. For this reason, co-evolution between these residues is not expected to play a major role because different amino acids in related proteins are a pre-requisite for co-evolution - sequence conserved residues do not co-evolve. Nevertheless, the relation between contact conservation and co-evolution was explored. The mutual information between residues was calculated from a sequence alignment (the same alignment used for the common numbering system) with the web-server MISTIC [57]. [insert explanation of mutual information] This server uses a reference structure for residue numbering (in this case 3K8Y).

These residue numbers were than mapped to the common numbering system (see section 2.4.1) in order to compare them to the consensus RCN. Figure 3.10a and b show scatter plots of mutual information (MI) plotted against contact conservation. No considerable correlation was found between the two parameters. In other words, residues forming conserved contacts do not have a greater mutual information than residues forming non-conserved contacts. This result suggests that co-evolution is not the way G proteins achieve a common structural scaffold despite high sequence variability. Instead, it seams that a small number of conserved residues suffices to create a constant structural scaffold despite low overall sequence similarity as indicated by figure 3.8.

However, it is possible that the sequence alignment used for the mutual information calculation was not sufficiently diverse. An alignment of all (37595) UniProt sequences from the Ras superfamily (Pfam-ID PF00071) was downloaded from the Pfam database. (The Pfam database allows to create an alignment of all UniProt sequences from a given protein family. This is achieved by aligning all sequences from that protein family to a pre-defined seed alignment a small number of representative sequences [42].) It is inclusion of more distantly related proteins, such as the Gα family would allow to detect more mutual information between residues. Moreover, it would be highly interesting to do the same analysis on a protein family with lower sequence conservation than small G proteins.

Next, the mutual information data was compared to contact energy. After all, highly favourable interactions are disrupted upon substitution of one of the contacting residues with the 'wrong' type of amino acid. For highly favourable contacts, the 'right' type of amino acid depends on the other contacting residue. Therefore, in a sequence alignment, the two positions of a favourable contact could be connected by mutual information - provided the positions are not conserved. To test this hypothesis, mutual information was compared to the mean contact energy of conserved contacts. However, figure 3.10c and d show no correlation between contact energy and mutual information. In other words, residues forming a favourable contact are not more likely to exhibit mutual information than residues forming a weak contact. The reason for this unexpected result could again lie in the fact that conserved contacts are mostly formed by sequence-conserved residues as shown above. If positions do not vary across sequences of the alignment, no mutual information can be detected.

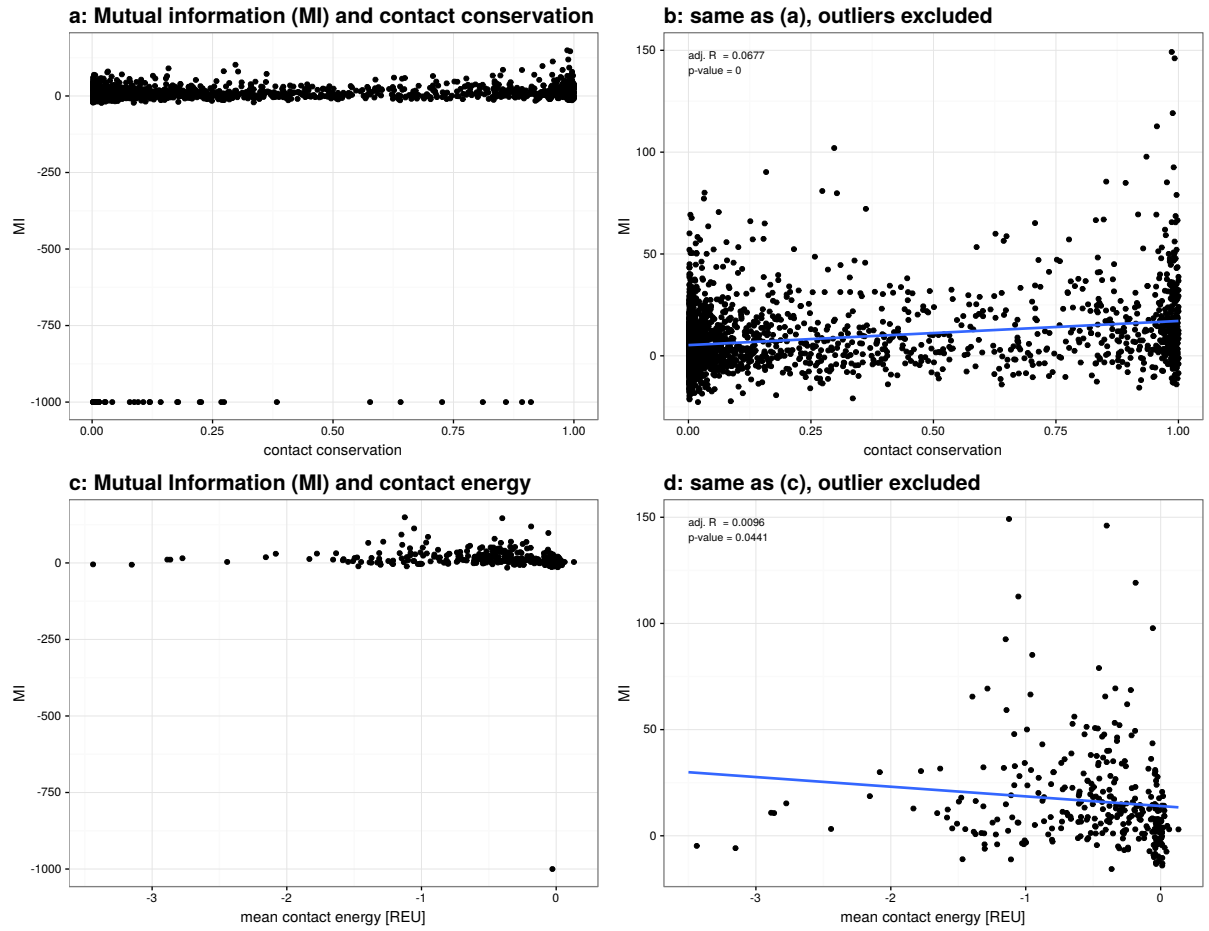

FIGURE 3.10: a: Scatter plot of mutual information against contact conservation. Due to a few outliers, most data points (contacts) lie within a small part of the plot making it difficult to interpret. b: Same plot as (a) only without the outliers in order to focus on the majority of the data. A linear regression shows that the two parameters do not correlate considerably. c: Scatter plot of mutual information against contact energy. Note that only conserved contacts are included (hence the lower number of data points compared to (a)). As conserved contacts are - by definition - present in almost all small G proteins, the contact energy is averaged over all proteins. For this reason, the x-axis is labeled 'mean contact energy'. d: Same plot as (c) only without the outlier which makes the plot easier to interpret. No considerable correlation between mutual information and contact energy was found.

Taken together, comparison of residue contacts with co-evolutionary data did not give conclusive results. However, a more in-depth analysis of a protein family with lower sequence conservation than small G proteins could shed more light on the relation between residue contact conservation and energy on the one hand and co-evolution on the other hand.

The results described in this section provide support for the approach used in the creation of a consensus RCN of Gα proteins [13]. In this study, distance based RCNs were calculated using RINerator [16]. Furthermore, only sequence-conserved residues were included in the consensus RCN. The finding that residues with highly favourable, conserved contacts tend to

be sequence-conserved suggests that this approach is a valid method of extracting presumably important residue contacts. Moreover, distance-based calculation of RCNs is simpler than energy calculation. Therefore, the approach of distance based consensus RCNs of sequence-conserved residues, developed by Flock et al., would be well suited for developing a software to fully automate calculation of consensus RCNs. Section 4.2.3 will explain in more detail why developing such a software would be an interesting project.

In summary, the data suggests that favourable side chain contacts depend on sequence conservation. In other words, highly favourable, conserved side chain contacts tend to be formed by the same amino acids in all proteins. Therefore, a small number of conserved residues seems to suffice to define the G protein fold. This leaves most residues free to mutate and fine-tune the function of specific G proteins without altering the overall fold.

Chapter 4

Discussion

4.1 Limitations

4.1.1 Computational cost

A semi-automatic bioinformatics pipeline was built to allow to quickly reproduce the results or repeat the analysis for a different data set. However, the pipeline requires considerable computational resources. The number of CPU-hours for the analysis is approximately equal to the number of protein structures analyzed (structure relaxation is by far the computationally most demanding step as explained in section 5.1.2). Therefore, applying the pipeline requires access to a computer cluster and basic knowledge of how to use it. Moreover, several bioinformatics tools, such as the Rosetta protein modelling suite, the structural alignment software MUSTANG and several python modules, need to be installed. For these reasons, the analysis can currently only be used by researchers with experience in bioinformatics. An interesting future project would be to create a simplified, fully automated pipeline for more widespread use (see section 4.2.3).

4.1.2 Lack of entropy in energy calculation

The raw data of this project are protein structures from the PDB. These structures, most of which have been solved by X-ray crystallography, represent a conformational average (over space and time) of all proteins conformations present in the crystal [58]. In other words, proteins are represented as rigid structures regardless of protein dynamics. However, in order to understand protein function, dynamics are as important as structure [59][60]. Without information on protein dynamics, the entropic component of any energy calculation has to be neglected.

Gibbs free energy is commonly used to describe interactions in energetic terms. Negative values indicate favourable interactions while positive values indicate unfavourable interactions. Gibbs free energy has an enthalpic and an entropic component. Enthalpic energy calculation relies on parameters such as charge, distance, angles, etc. which can be obtained from a static protein structure. Entropy on the other hand can be regarded as a measure of disorder or flexibility. Any change from a flexible state to a rigid state, for instance two side chains forming an interaction, is entropically unfavourable thus making the interaction less favourable. All physical energy terms of the Rosetta energy function (described in section 2.3), contribute towards the enthalpic component [61]. However, due to the lack of knowledge about dynamics and flexibility, a static structure does not allow the calculation of the entropic component of interaction energy. For this reason, the Rosetta energy function lacks an explicit entropic component. Without entropy, the calculation of residue contact energy is incomplete. Nevertheless, this limitation does not mean that Rosetta energy calculation is completely unreliable. In addition to physical terms, the Rosetta energy function contains knowledge-based potentials such as amino acid interaction preferences [61]. These are obtained from experimentally solved structures. As protein structure is determined by both enthalpy and entropy, it can be argued that knowledge-based potentials implicitly consider entropy.

An interesting way of obtaining an approximation of entropy from static structures would be to use Debye–Waller factors, commonly referred to as B factors. B factors can be regarded as indicators of flexibility of different parts of the structure. Low B factors indicate rigid structure while high B factors indicate disorder and flexibility. For instance, side chains with high B factors can be assumed to be flexible. An interaction between such side chains would restrict their flexibility making it entropically unfavourable. In this manner, B factors could serve as a proxy for interaction energy. Incorporating this information into RCNs would improve energy calculation.

Taken together, the Rosetta energy function gives a viable estimate of residue contact energy. Given that static structures serve as the raw data, it is obvious that the entropic component of Gibbs free energy can only be considered implicitly. But then again, the same flaw is inherent to any analysis based on protein structures.

4.1.3 Results limited to G proteins

In this project, consensus RCNs were calculated for two protein families - small G proteins and Gα subunits. Applying the bioinformatics pipeline to the latter allowed to compare the results to experimental data collected by Sun et al. [41]. For all other analyses, the small G protein

family served as a model system. The results of these analyses may also be true for many other protein families. However, further research would be required to investigate this possibility. For now, the results have to be regarded as limited to G proteins. Nevertheless, the analyses serve as a proof of concept for the bioinformatics pipeline which can also be applied to any other protein family.

4.2 Outlook

4.2.1 Molecular dynamics simulations

As discussed above (section 4.1.2), one limitation of the approach used in this project is that entropy is not considered explicitly. One way of addressing this issue could be to run a molecular dynamics (MD) simulation for every protein structure. An MD simulation would allow to calculate the enthalpic as well as the entropic contribution to the Gibbs free energy of residue interactions [24]. This method would be more effective to identify energetically favourable - and therefore presumably structurally important - residue contacts. On the downside, more computational resources would be required. Moreover, designing a protocol for the MD simulation of up to hundreds of related proteins would require more time and expertise than using the Rosetta modelling suite.

Another application of MD are simulations of both the folded and the unfolded state. The resulting dataset would be highly interesting as it would allow to compare residue contacts in the unfolded and the folded state. The main hypothesis of this project is that a network of highly favourable and conserved contacts stabilizes and defines the common fold of a protein family. This hypothesis predicts the absence of these contacts in the unfolded state while they are expected to be present in the folded state over the course of an MD simulation. In order to test the prediction, a consensus RCN of highly favourable residue contacts from an ensemble of folded simulation snapshots could be compared to RCNs of unfolded state snapshots.

4.2.2 Applications

Consensus RCNs could add another layer of information to various experimental and computational research projects. First, knowledge of residue contacts and the presence of equivalent contacts in related proteins can aid the interpretation of mutational data. Second, consensus RCNs allow to gauge whether information about a given protein applies for related proteins as well. For instance, consider the case of the alanine scanning study which provided the mutational data used in this project (see section 3.2). For every residue of human Gαi1, an alanine

mutant was created and its thermo-stability was measured and compared to the wild type [41]. Such datasets are as valuable as they are rare (no comparable dataset was found for a member of the small G protein family). Therefore, it is important to asses in how far the results apply to related proteins. For residues forming conserved contacts, the effect of mutation can be assumed to be similar for all members of a given protein family. If a residue forms different contacts in different members of the protein family, the effect of mutation may differ as well. Third, consensus RCNs can aid identification of allosteric mechanisms. In allosteric regulation, the activity of an enzyme is altered by binding of an effector molecule to a site other than the active site. Traditional RCNs (of a single protein) have been used to shed light on the mechanism of allosteric regulation in the case of UbcH5b [11]. However, without prior knowledge it is difficult to distinguish between residue contacts involved in the signal transfer from the allosteric to the active site and other contacts.

More information can be obtained from comparison of the active and inactive state RCN, respectively. A universal allosteric mechanism for Gα activation by GPCRs was discovered by comparing the consensus RCN of 11 inactive Gα structures with the RCN of a GPCR-bound Gα structure [13]. This approach allowed to identify residues which change their contacts upon GPCR-binding to alter the active site conformation (leading to GDP release). In the case of the Gα study, only a single structure was available for the GPCR-bound state while multiple structures of the inactive state where available. Instead of simply picking one representative inactive Gα structure, a consensus RCN was calculated from all available inactive α structures. Using a consensus RCN reduced the dependence on a single structure and allowed to find the universal α activation mechanism rather than the activation mechanism of a single Gα protein [13]. A similar approach could be applied to identify allosteric mechanisms in other enzymes. While using consensus RCNs rather than RCNs individual structures is preferable if possible, to my knowledge, no user friendly software is available for this task.

4.2.3 Simplified pipeline

The previous section lists several applications for consensus RCNs. For these applications and possibly many more, it would be very useful to have an easy to use software for calculation of a consensus RCN from a given set of (related) protein structures. This would allow scientists to use consensus RCNs as part of a project. However, to my knowledge, no user friendly software to calculate consensus RCNs from multiple protein structures is available so far. This shortfall could be addressed by simplifying and fully automating the bioinformatics pipeline developed in this project. To this end, the energy calculation would probably have to be discarded due to

its computational cost. In order to avoid having to create (and refine) a structural alignment for the common residue numbering system, an alignment of the protein family of interest could possibly be obtained from the Pfam database [42]. Creating a common residue numbering system is the most difficult step to automate. The remainder of the software could be put together fairly easily from the code written in the course of this project. In summary, developing an easy to use software for quick calculation of consensus RCNs would be an interesting project.

4.3 Summary and Conclusion

In this master's project, multiple residue contact energy networks were calculated from protein structures and combined in a consensus RCN using a common residue numbering system. The two approaches of energy-based RCNs and consensus RCNs have not been combined so far. Using the example of the well studied small G protein family, the relation of contact conservation (the fraction of proteins which have an equivalent residue contact), contact energy, sequence conservation and co-evolution was explored. The results from these analyses suggest that residue contacts present in all or almost all members of the protein family tend to be energetically more favourable (see section 3.3). Moreover, residues forming conserved, highly favourable contacts tend to be sequence-conserved (see section 3.4). Comparing the consensus RCN of G α with mutational data from this protein family provided support for the relevance of the energy-based RCNs but also highlighted the difficulty of predicting structural importance from static protein structures(see section 3.2). Taken together, these results demonstrate how to identify residues and residue contacts which stabilize the common fold of a protein family - and distinguish them from residues and contacts specific for an individual protein.

Furthermore, the scale of the project which made use of a dataset of 511 small G protein structures exceeds the scale of previous studies of consensus RCNs. A semi-automatic bioinformatics pipeline was developed to process this dataset in a reproducible manner (see chapter 2). In essence, information (RCNs) from all these structures was extracted to obtain information about the protein family (consensus RCN) rather than individual proteins. For many structural biology projects, it is interesting to know which features of a given protein are inherent to the whole protein family and which are specific for the protein of interest. Making use of the growing wealth of information in the PDB will become an even more important part of structural biology in the future. The results demonstrate that consensus RCNs are a promising method for combining detailed information from a large number of protein structures for a

single project. Moreover, several interesting paths for future research were discovered in the course of the project.

Chapter 5

Methods

5.1 Protocol

Chapter 2 describes the method which was developed for this project and explains the rationale behind the approach. It provides all the information required to understand, review and criticise the approach. This section assumes knowledge of the approach described in chapter 2 and provides all parameters required to reproduce the protocol.

5.1.1 Structure selection

As described in section 2.2, 543 structures of small G proteins were downloaded from the PDB using the Pfam identifier PF00071.
Out of these, 32 structures were excluded from analysis:
1e96, 1foe, 1g4u, 1he1, 1hh4, 1i4d, 1i4l, 1i4t, 1mh1, 1ryf, 1ryh, 2fju, 2h7v, 2nz8, 2p2l, 2vrw, 2wkp, 2wkq, 2wkr, 2yin, 3b13, 3bji, 3c5h, 3ryt, 3sbd, 3sbe, 3su8, 3sua, 3th5, 4gzl, 4gzm, 4iru
The remaining 511 structures of small G proteins were analysed in this project using the bioinformatics pipeline describe in chapter 2:
121p, 1a2b, 1a2k, 1a4r, 1agp, 1am4, 1an0, 1bkd, 1byu, 1c1y, 1cc0, 1clu, 1ctq, 1cxz, 1d5c, 1doa, 1dpf, 1ds6, 1ek0, 1ftn, 1g16, 1g17, 1gnp, 1gnq, 1gnr, 1grn, 1gua, 1gwn, 1gzs, 1he8, 1huq, 1i2m, 1iaq, 1ibr, 1ioz, 1jah, 1jai, 1k5d, 1k5g, 1k8r, 1kao, 1ki1, 1kmq, 1ky2, 1ky3, 1kz7, 1kzg, 1lb1, 1lf0, 1lf5, 1lfd, 1m7b, 1n6h, 1n6i, 1n6k, 1n6l, 1n6n, 1n6o, 1n6p, 1n6r, 1nf3, 1nvu, 1nvv, 1nvw, 1nvx, 1oiv, 1oiw, 1oix, 1ow3, 1p2s, 1p2t, 1p2u, 1p2v, 1plj, 1plk, 1pll, 1q21, 1qbk, 1qg2, 1qg4, 1qra, 1r2q, 1rrp, 1rvd, 1s1c, 1s8f, 1t91, 1tu3, 1tu4, 1tx4, 1u8y, 1u8z, 1u90, 1uad, 1ukv, 1vg0, 1vg1, 1vg8, 1vg9, 1wa5, 1wms, 1wq1, 1x1r, 1x1s, 1x3s, 1x86, 1xcg, 1xcm, 1xd2, 1xj0, 1xtq, 1xtr, 1xts, 1yhn, 1yu9, 1yvd, 1yzk, 1yzl, 1yzn, 1yzq, 1yzt, 1yzu, 1z06, 1z07, 1z08, 1z0a, 1z0d, 1z0f, 1z0i, 1z0j, 1z0k, 1z22, 1z2a, 1z2c, 1zbd, 1zc3, 1zc4, 1zvq, 1zw6, 221p, 2a5j, 2a78, 2a9k, 2atv, 2atx, 2bcg, 2bku, 2bmd, 2bme, 2bov, 2c2h, 2c5l, 2ce2, 2cjw, 2cl0, 2cl6, 2cl7, 2clc, 2cld, 2cls, 2d7c, 2dfk, 2dpx,

2e9s, 2efc, 2efd, 2efe, 2efh, 2eqb, 2erx, 2ery, 2evw, 2ew1, 2f7s, 2f9l, 2f9m, 2fe4, 2ffq, 2fg5, 2fn4, 2fol, 2fu5, 2fv8, 2g0n, 2g3y, 2g6b, 2g77, 2gcn, 2gco, 2gcp, 2gf0, 2gf9, 2gil, 2gjs, 2gzd, 2gzh, 2hei, 2ht6, 2hup, 2hv8, 2hxs, 2ic5, 2iey, 2iez, 2if0, 2il1, 2j0v, 2j1l, 2ngr, 2nty, 2nzj, 2o52, 2ocb, 2ocy, 2odb, 2oil, 2ot3, 2ov2, 2p5s, 2q21, 2q3h, 2qme, 2qrz, 2quz, 2rap, 2rex, 2rga, 2rgb, 2rgc, 2rgd, 2rge, 2rgg, 2rgn, 2rhd, 2uzi, 2v55, 2vh5, 2w2t, 2w2v, 2w2x, 2wbl, 2wm9, 2wmn, 2wmo, 2wwx, 2x19, 2x1v, 2y8e, 2yc2, 2yc4, 2zet, 3a58, 3a6p, 3bbp, 3bc1, 3bfk, 3brw, 3bwd, 3c5c, 3cbq, 3cf6, 3ch5, 3clv, 3con, 3cph, 3cpj, 3cue, 3cwz, 3ddc, 3dz8, 3e5h, 3ea5, 3eg5, 3gcg, 3gft, 3gj0, 3gj3, 3gj4, 3gj5, 3gj6, 3gj7, 3gj8, 3gjx, 3i3s, 3icq, 3jza, 3k8y, 3k9l, 3k9n, 3kkm, 3kkn, 3kko, 3kkp, 3kkq, 3kuc, 3kud, 3kz1, 3l0i, 3l8y, 3l8z, 3law, 3lbh, 3lbi, 3lbn, 3lo5, 3lw8, 3lwn, 3lxr, 3m1i, 3mjh, 3msx, 3nby, 3nbz, 3nc0, 3nc1, 3nkv, 3oes, 3oiu, 3oiv, 3oiw, 3pir, 3pit, 3q3j, 3q72, 3q7p, 3q7q, 3q85, 3qbt, 3qbv, 3rab, 3ran, 3rap, 3ref, 3reg, 3rry, 3rrz, 3rs0, 3rs2, 3rs3, 3rs4, 3rs5, 3rs7, 3rsl, 3rso, 3rwm, 3rwo, 3sea, 3sfv, 3t06, 3t5g, 3tgp, 3tkl, 3tnf, 3tso, 3tvd, 3tw8, 3v4f, 3vhl, 3w3z, 3wyf, 3wyg, 3x1w, 3x1x, 3x1y, 3x1z, 3zjy, 421p, 4aii, 4c0q, 4c4p, 4cym, 4cz2, 4d0g, 4d0l, 4d0m, 4d0n, 4did, 4djt, 4dkx, 4dlr, 4dls, 4dlt, 4dlu, 4dlv, 4dlw, 4dlx, 4dly, 4dlz, 4drz, 4dsn, 4dso, 4dst, 4dsu, 4dvg, 4dxa, 4efl, 4efm, 4efn, 4epr, 4ept, 4epv, 4epw, 4epx, 4epy, 4f38, 4fmb, 4fmc, 4fmd, 4fme, 4g01, 4g0n, 4g3x, 4gmx, 4gpt, 4hat, 4hau, 4hav, 4haw, 4hax, 4hay, 4haz, 4hb0, 4hb2, 4hb3, 4hb4, 4hdo, 4hdq, 4hlq, 4i1o, 4itr, 4js0, 4jvs, 4k81, 4klz, 4kvg, 4kyi, 4l8g, 4l9s, 4l9w, 4ldj, 4lhv, 4lhw, 4lhx, 4lhy, 4lhz, 4li0, 4lpk, 4lrw, 4luc, 4lv6, 4lwz, 4lx0, 4lyf, 4lyh, 4lyj, 4m1o, 4m1s, 4m1t, 4m1w, 4m1y, 4m21, 4m22, 4m8n, 4mgi, 4mgk, 4mgy, 4mgz, 4mh0, 4mit, 4nmm, 4nyi, 4nyj, 4nym, 4o25, 4o2l, 4o2r, 4obe, 4ojk, 4ol0, 4phf, 4phg, 4phh, 4pzy, 4pzz, 4q01, 4q02, 4q03, 4q21, 4q9u, 4qxa, 4rke, 4rkf, 4u5x, 4uru, 4urv, 4urw, 4urx, 4ury, 4urz, 4us0, 4us1, 4us2, 4xh9, 4yc7, 4ydh, 521p, 5p21, 621p, 6q21, 721p, 821p

5.1.2 Structure relaxation

The reason for using structure relaxation as well as the choice of parameters were described in section 2.3.3. The following command was used to execute structure relaxation:

```
./relax.linuxgccdebug  # path to Rosetta relax executable
-database rosetta_database  # path to Rosetta database
-in:file:s 3ky8.pdb  # input PDB-file
-out:file:fullatom  # outfile format: atomic coordinates (like in PDB-file)
-relax:constrain_relax_to_start_coords  # keep backbone atoms fixed
-relax:ramp_constraints false  # keep constraints in place throughout protocol
-relax:quick  # only use 5 cycles of rotamer repacking
```

5.1.3 Rosetta residue energy breakdown

The Rosetta program 'residue energy breakdown' was executed using this command:

```
./residue_energy_breakdown.linuxgccrelease
-database rosetta_database  # path to Rosetta database
-in:file:s 3ky8.pdb  # input PDB-file
-out:file:silent res_en_breakdown_3k8y.txt  # output file name
```

5.1.4 Structural alignment with MUSTANG

MUSTANG [51] was used to create a structural alignment. The program outputs the structural alignment in PDB-file containing all structures aligned in the same coordinate system. Moreover, a sequence alignment derived from the structure alignment is also provided. A matrix of RMSD values of all structures against all structures can also be obtained by passing an extra argument. The three output files share the same name and have the extensions .pdb, .afasta and .rms_rot, respectively.

```
MUSTANG_v3.2.2/bin/mustang-3.2.1  # path to executable
-f structure_list.txt  # file with a list of structures to be aligned
-r ON  # output RMSD matrix
-F fasta  # output sequence alignment in fasta format
-o mustang_alignment_name  # name of the output files
```

5.1.5 Statistics

Mann-Whitney U test

Two figures, figure 3.6 and figure 3.9, required testing, whether the difference between two distributions is significant. Visual inspection revealed that the distributions are neither normal nor fit any other parameterized distribution. Therefore, a non-parametric test had to be applied. Moreover, residue contact are independent from each other (as opposed to matched samples). For these condition, a Mann-Whitney U test (also called Mann-Whitney-Wilcoxon test or Wilcoxon rank-sum test) is suited well as it requires neither a particular distribution nor matched samples [62]. In R, this test can be calculated with the function 'wilcox.test()'. (The name of this function can be somewhat confusing because it indicates the Wilcoxon signed-rank test (a non-parametric test for matched samples). The reason for this is that the function 'wilcox.test()' can calculate both, Wilcoxon signed-rank test as well as the Mann-Whitney U

test. While the U test is applied by default, the Wilcoxon signed-rank test can be performed using the argument 'paired=TRUE'.)

Box-plots

The box-plots (figure 3.4, 3.6 and 3.9), as well as all other plots presented in this thesis, were created using the R-package ggplot2 [63].

In this package, the lower and upper limits of the boxes are defined as the first and third quartiles (the 25th and 75th percentiles), respectively. This definition is a widely accepted standard for box-plots [64]. The distance between the fist and the third quartile is termed interquartile range (IQR). The median is shown as a horizontal line in the box.

There is no universal convention of drawing whiskers in box-plots. In this thesis, the whiskers extend to the most extreme values which are still inside $1.5 * IQR$, as defined in the package ggplot2 [65]. Observations outside this range (outliers) are shown as dots. No outliers are excluded in the box-plots in this thesis unless mentioned explicitly (3.6c). If no dots are present outside the range of the whiskers, as in figure 3.9b, this means that there is no value higher than the third quartile plus $1.5 * IQR$ and no value lower than the fist quartile minus $1.5 * IQR$.

Notches are added to the box-plots to enable a quick estimate whether the difference between two median is significant. The notch extends $1.58 * IQR/\sqrt{n}$ from the median in both directions [65]. This corresponds roughly to the 95% confidence interval [66]. Therefore, notches allow to quickly estimate the significance of differences between boxes just by looking at the plot. However, a more thorough analysis was done using the Mann-Whitney U test as described above (5.1.5).

All box-plots in this thesis are combined with violin-plots in order to illustrate the underlying distributions. In order to allow comparison of distributions with different numbers of observations, all violins have the same area. The number of observations is provided on top of each box and violin. All distributions were trimmed to the range of the data; violins only reach from the highest to the lowest value. For small samples, this may give the impression that the tails of the distributions were cut (for instance in figure 3.4a). This setting was chosen because the violin plots were added to show the distribution of the raw data - extrapolating the distributions would violate this aim. The violins in figure 3.6 are visualized differently than in the other plots for optical reasons only - the shape of the violins is derived in the same way in all figures.

Finally, the widths of all boxes are proportional to $\sqrt{n}$ within each figure. Note that the box-widths are not comparable across figures. For this reason the number of observations is provided on top of each box.

5.1.6 Programming

Python

The general purpose programming (scripting) language Python (version 2.7) was used for processing and checking PDB-files. Furthermore, it was used for scripts which execute other software (called wrapper scripts) and process the output. Python wrapper scripts were written for the Rosetta programs 'relax' and 'residue_energy_breakdown' as well as MUSTANG [51] (for structural alignment) and DSSP [54]. Analysis of biological data, such as PDB-files or sequence alignments make extensive use of Biopython modules, a collection python functions and classes for bioinformatics [67]. In particular, the Biopython module 'PDB' proved invaluable for reading, checking and altering PDB-files [68]. Abstracted data, such as residue contact networks, were stored and processed using modules from the SciPy [69] library (most importantly pandas [70] and NumPy [71]). All scripts were developed in the Spyder integrated development environment (IDE) in an IPython [72] environment.

R

The programming language R was developed primarily for statistics, data analysis and visualization. While both, Python and R, are well suited for this task, R (version 3.2) [73] was chosen due to personal preference. Apart from R's basic functionality, the packages plyr (mainly for the function 'ddply') [74] and reshape2 [75] proved particularity useful. All plots presented in this thesis were created using the data visualization package ggplot2 [63]. All scripts were written and exploratory was performed in the RStudio IDE. R-code was combined with plain text for automatic generation of HTML- or PDF-reports using the package knitr [55]. The output contains R-code, plots and explanatory plain text and is structured into sections. This approach, called literate programming, enhances readability of analytic code and facilitates sharing results with collaborators.

5.1.7 Molecular graphics

All illustrations of protein structures in this thesis were created with the molecular graphics system PyMOL [15]. For exploratory analysis, another molecular viewer, UCSF Chimera [76],

was used.

5.2 Computational challenges

Chapter 2 presents the project's approach in a straightforward way. For the sake of clarity and simplicity, some computational challenges encountered along the way were not explained in detail. This section will outline the main problems that had to be solved in order arrive bioinformatics pipeline developed during this master's project. The information provided here is not vital to understand and critically review the results of the project. However, when reproducing the results, applying the method in another project or creating a similar pipeline, reading this section is highly recommended and could save a lot of time.

5.2.1 Preparing structures for Rosetta applications

Small G protein downloaded from the PDB are very diverse. Some structures are complexes of small G proteins with other proteins. The size of the unit cell is different from PDB file to PDB-file which means that the number of chains per structure is variable. However, in order to compare residue contact networks of different structures it is important to have only one small G protein per structure. Initially, all chains featuring a small G protein domain (identified by the Pfam-ID 'PF00071') were extracted and written to new PDB files (one file per chain) while chains from other proteins were discarded. However, this approach would create a bias as structures with many chains would contribute more residue contacts to the dataset (provided there are similar residue contacts in all chains) than structures with only one chain. Therefore, only one chain per structure was selected for analysis. A python script was written to parse all downloaded PDB files (using the Biopython module 'PDB'). The first chain with a small G-protein domain is isolated and written to a new PDB file.

Moreover, Rosetta does not parse all lines in a PDB file. It is therefore recommended to prepare PDB files before analysis with Rosetta by removing all lines except those starting with 'ATOM' or 'TER'. The header information as well as any HETATM records (except selenomethionine) are lost in this process. The python script for extracting chains described above also cleans the PDB-files in this manner.

5.2.2 Limitation of the number of structures in a structural alignment

From a structural alignment, a common residue numbering system is obtained for all structures. In an all against all structural alignment, the number of pairwise alignments increases

as the square of the number of structures. This is computationally not feasible for 511 structures. The state of the art software MUSTANG was found to fail frequently when the number of structures exceeds 100. Therefore, only a subset of all structures could be used for structural alignment. The subset was selected by using only one protein per orthologous group. Many proteins have been crystallized from several different species. By selecting the best resolution structure for each protein, a subset of 115 structures from 115 different species is created. These structure were aligned using MUSTANG and the resulting alignment was refined as described in section 2.4.1.

This approach created the challenge of mapping residue contact networks of structures not present in the alignment to alignment positions. While orthologous proteins in different species do not necessarily have the same PDB residue numbering or the exact same number of residues, they map to the same UniProt sequence. PDB and UniProt information is cross-referenced by SIFTS (Structure Integration with Function, Taxonomy and Sequences) [77]. A script to automate retrieval of residue to residue correspondence between multiple PDB structures and UniProt sequences was developed by Tilman Flock for his work on G-α subunits [13] and was kindly provided for this project. Using this information, each residue of a protein not present in the alignment could be related to a residue of a homologous protein in the alignment. In this manner, alignment positions could be mapped to all residue contact networks and serve as a common residue numbering system. Hence, the workflow for the small G-protein family is slightly more complicated than illustrated in figure 2.4. A more detailed overview of the workflow is provided in figure 5.1. However, the simple workflow was applied to the Gα subunit. The simple workflow is recommended unless the number of structures is too high for a structural alignment.

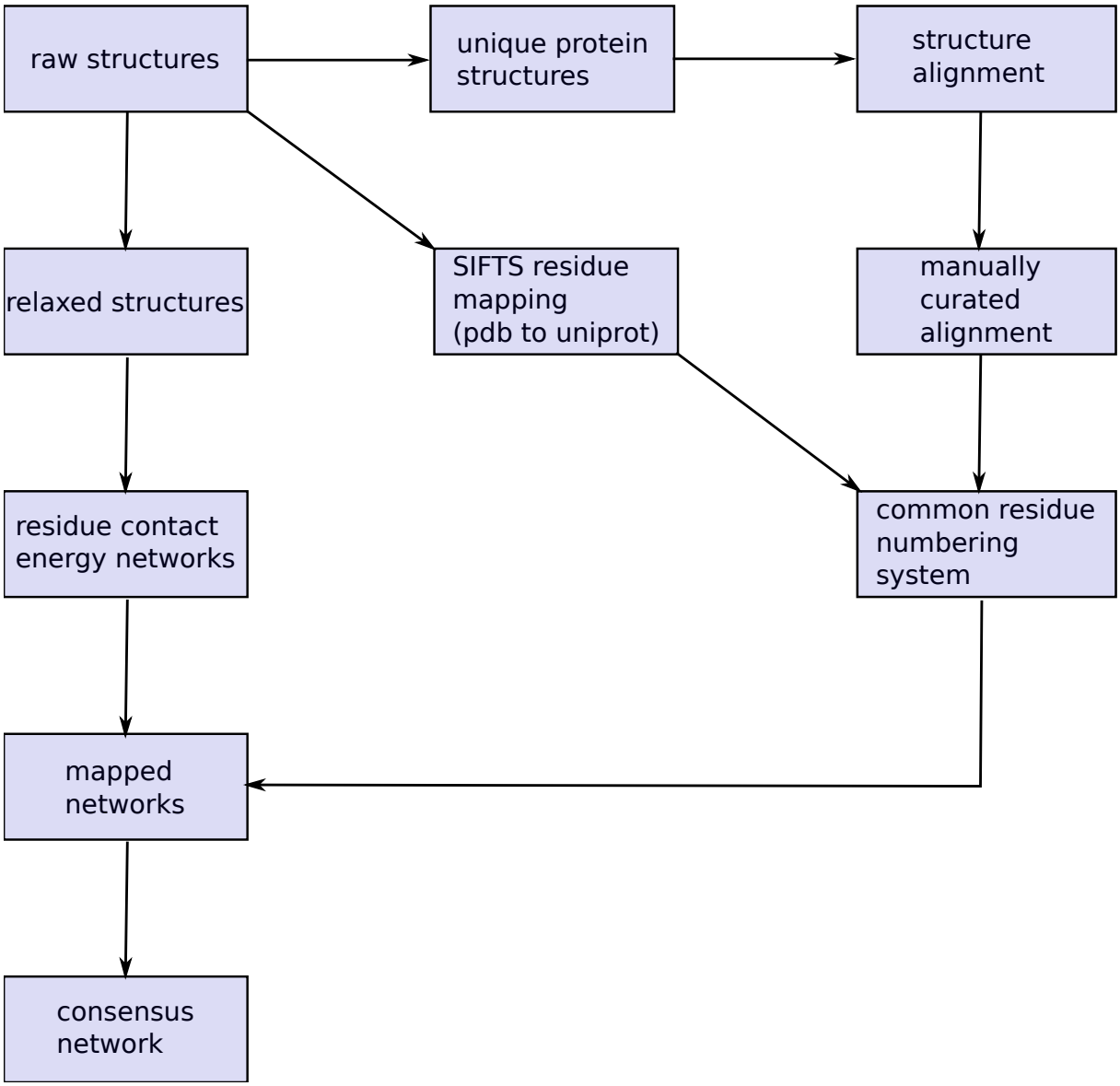

FIGURE 5.1: SIFTS workflow.

5.2.3 Residue numbering issues

As described before, Rosetta has its own residue numbering scheme starting at one at the N-terminus regardless of the residue numbers in the PDB file. The Rosetta's python interface (PyRosetta) was used to obtain a list of Rosetta and PDB residue number correspondences. However, if PDB residue numbers are not numeric but contain letter as well, more than one residue can be assigned the same Rosetta residue number. For instance, for the PDB numbers 1A, 2A, 1, 2, 3, etc., the corresponding Rosetta numbers are 1, 2, 1, 2, 3, etc. Double residue

numbers fail to uniquely identify residues. As only very few structures display non-numeric residue numbers, they were excluded from analysis.

Another issue was identified when for some structures the total number of residues according to Rosetta was lower than the total number of residues in the PDB file. Seemingly arbitrarily and very rarely, Rosetta seemed to ignore a residue without giving a warning message. The result is that the after the skipping, residue number i is instead given number i-1 leading to problems when mapping residues to alignment positions. The reason was found to be occupancy values of zero for atoms of the missed residues. Some PDB files provide alternative side chain conformations (ANISOU records). In these cases, two lines provide two different coordinates for each atom. The column 'occupancy' - a number between zero and one - indicates the fraction of unit cells with the atom at the given coordinates. The vast majority of PDB files provides only one set of coordinates per atom which means that the occupancy is always one. Preparing PDB structures for Rosetta applications requires to remove alternative coordinates. In very few cases, this leads to residues which have atoms with zero occupancy. These residues are then ignored by Rosetta. It is unfortunate that Rosetta does not give an error- or warning-message because of the considerable time and effort that goes into finding the cause of such erroneous output. After the problem was identified, the affected structures were excluded from further analysis.

In total, out of 543 structures, 32 have been excluded from analysis due to numbering issues. While the problems could have been resolved on an individual basis, a disproportionally long time would have been required for only a small part of the dataset.

After these experiences, a function was integrated into the pipeline to check for non-numeric PDB residue numbers, output a warning message and automatically exclude the structures from analysis. Likewise, another function checks if the number of residues recognized by Rosetta is equal to the number of residues in the PDB file (obtained form the Biopython module 'PDB'). Again, a warning message is provided if this is not the case and the affected structures are excluded from analysis. In summary, these experiences were used to create automatic error-checking and warning messages to remove the need for tedious error-tracking in future analyses.

References

[1] H M Berman et al. "The Protein Data Bank." *Nucleic acids research* 28.1 (2000), pp. 235–242. DOI: 10.1093/nar/28.1.235.

[2] "PDB statistics" (). URL: http://www.rcsb.org/pdb/statistics/holdings.do.

[3] A G Murzin et al. "SCOP: a structural classification of proteins database for the investigation of sequences and structures." *Journal of molecular biology* 247.4 (1995), pp. 536–40. DOI: 10.1006/jmbi.1995.0159. URL: http://www.ncbi.nlm.nih.gov/pubmed/7723011.

[4] Ca Orengo et al. "CATH - a hierarchic classification of protein domain structures". *Structure* March (1997), pp. 1093–1109. DOI: 10.1016/S0969-2126(97)00260-8. URL: http://discovery.ucl.ac.uk/170914/.

[5] Z X Wang. "A re-estimation for the total numbers of protein folds and superfamilies." *Protein engineering* 11.8 (1998), pp. 621–626. DOI: 061/7.

[6] C A Orengo, D T Jones, and J M Thornton. *Protein superfamilies and domain superfolds.* 1994. DOI: 10.1038/372631a0.

[7] S Govindarajan, R Recabarren, and R A Goldstein. "Estimating the total number of protein folds". *Proteins: Structure, Function, and Bioinformatics* 35.4 (1999), pp. 408–14. DOI: 10.1002/(SICI)1097-0134(19990601)35:4<408::AID-PROT4>3.0.CO;2-A. URL: /citations?view{_}op=view{_}citation{\&}continue=/scholar?hl=en{\&}as{_}sdt=1,22{\&}scilib=1024{\&}scioq=i-tasser{\&}citilm=1{\&}citation{_}for{_}view=stFuORQAAAAJ:d1gkVwhDpl0C{\&}hl=en{\&}oi=p{\%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/10382668.

[8] Eugene V Koonin, Yuri I Wolf, and Georgy P Karev. "The structure of the protein universe and genome evolution." *Nature* 420.6912 (2002), pp. 218–223. DOI: 10.1038/nature01256.

[9] Lesley H. Greene. "Protein structure networks". *Briefings in Functional Genomics* 11.6 (2012), pp. 469–478. DOI: 10.1093/bfgp/els039.

[10] Lesley H. Greene and Victoria a. Higman. "Uncovering network systems within protein structures". *Journal of Molecular Biology* 334.4 (2003), pp. 781–791. DOI: 10.1016/j.jmb.2003.08.061.

[11] Moitrayee Bhattacharyya, Chanda R. Bhat, and Saraswathi Vishveshwara. "An automated approach to network features of protein structure ensembles". *Protein Science* 22.10 (2013), n/a–n/a. DOI: 10.1002/pro.2333. URL: http://doi.wiley.com/10.1002/pro.2333.

[12] Venkataramanan Soundararajan et al. "Atomic interaction networks in the core of protein domains and their native folds." *PloS one* 5.2 (2010), e9391. DOI: 10.1371/journal.pone.0009391. URL: http://dx.plos.org/10.1371/journal.pone.0009391http://www.ncbi.nlm.nih.gov/pubmed/20186337http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2826414.

[13] Tilman Flock et al. "Universal allosteric mechanism for Gα activation by GPCRs." *Nature* 524.7564 (2015), pp. 173–179. DOI: 10.1038/nature14663. URL: http://www.ncbi.nlm.nih.gov/pubmed/26147082.

[14] Greg Buhrman et al. "Allosteric modulation of Ras positions Q61 for a direct role in catalysis." *Proceedings of the National Academy of Sciences of the United States of America* 107.11 (2010), pp. 4931–6. DOI: 10.1073/pnas.0912226107. URL: http://www.ncbi.nlm.nih.gov/pubmed/20194776http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2841912.

[15] LLC Schrödinger. "The PyMOL molecular graphics system, version 1.6" ().

[16] Nadezhda T. Doncheva et al. *Analyzing and visualizing residue networks of protein structures*. 2011. DOI: 10.1016/j.tibs.2011.01.002. URL: http://dx.doi.org/10.1016/j.tibs.2011.01.002.

[17] Ganesh Bagler and Somdatta Sinha. "Network properties of protein structures". *Physica A: Statistical Mechanics and its Applications* 346.1-2 SPEC. ISS. (2005), pp. 27–33. DOI: 10.1016/j.physa.2004.08.046. arXiv: 0408009 [q-bio].

[18] N Kannan and S Vishveshwara. "Identification of side-chain clusters in protein structures by a graph spectral method." *Journal of molecular biology* 292.2 (1999), pp. 441–464. DOI: 10.1006/jmbi.1999.3058.

[19] Nadezhda T Doncheva et al. "Topological analysis and interactive visualization of biological networks and protein structures." *Nature Protocols* 7.4 (2012), pp. 670–85. DOI: 10.1038/nprot.2012.004. URL: http://www.ncbi.nlm.nih.gov/pubmed/22422314.

[20] J M Word et al. "Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation." *Journal of molecular biology* 285.4 (1999), pp. 1735–1747. DOI: 10.1006/jmbi.1998.2401.

[21] J M Word et al. "Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms." *Journal of molecular biology* 285.4 (1999), pp. 1711–33. DOI: 10.1006/jmbi.1998.2400. URL: http://www.ncbi.nlm.nih.gov/pubmed/9917407.

[22] Paul Shannon et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome research* 13.11 (2003), pp. 2498–504. DOI: 10.1101/gr.1239303. URL: http://www.ncbi.nlm.nih.gov/pubmed/14597658http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC403769.

[23] M.S. Vijayabaskar and Saraswathi Vishveshwara. "Interaction Energy Based Protein Structure Networks". *Biophysical Journal* 99.11 (2010), pp. 3704–3715. DOI: 10.1016/j.bpj.2010.08.079. URL: http://linkinghub.elsevier.com/retrieve/pii/S0006349510011872.

[24] Sander Pronk et al. "GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit". *Bioinformatics* 29.7 (2013), pp. 845–854. DOI: 10.1093/bioinformatics/btt055.

[25] M S Vijayabaskar and Saraswathi Vishveshwara. "Comparative analysis of thermophilic and mesophilic proteins using Protein Energy Networks." *BMC bioinformatics* 11 Suppl 1 (2010), S49. DOI: 10.1186/1471-2105-11-S1-S49.

[26] M Vendruscolo, E Kussell, and E Domany. "Recovery of protein structure from contact maps." *Folding & design* 2.5 (1997), pp. 295–306. DOI: 10.1016/S1359-0278(97)00041-2. arXiv: 9705211 [cond-mat].

[27] C B Anfinsen. "Principles that govern the folding of protein chains." *Science (New York, N.Y.)* 181.4096 (1973), pp. 223–30. DOI: 10.1126/science.181.4096.223. URL: http://www.ncbi.nlm.nih.gov/pubmed/4124164.

[28] C Chothia and a M Lesk. "The relation between the divergence of sequence and structure in proteins." *The EMBO journal* 5.4 (1986), pp. 823–826. DOI: 060fehlt.

[29] R Bonneau and D Baker. "Ab initio protein structure prediction: progress and prospects." *Annual review of biophysics and biomolecular structure* 30.6 (2001), pp. 173–89. DOI: 10.1146/annurev.biophys.30.1.173. URL: http://www.ncbi.nlm.nih.gov/pubmed/11340057.

[30] Yang Zhang. "Progress and challenges in protein structure prediction." *Current opinion in structural biology* 18.3 (2008), pp. 342–8. DOI: 10.1016/j.sbi.2008.02.004. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2680823{\&}tool=pmcentrez{\&}rendertype=abstract.

[31] Su Yun Chung and S Subbiah. "A structural explanation for the twilight zone of protein sequence homology". *Structure* 4.10 (1996), pp. 1123–1127. DOI: 10.1016/S0969-2126(96)00119-0. URL: http://www.cell.com/article/S0969212696001190/fulltext.

[32] M I Sadowski and D T Jones. "The sequence-structure relationship and protein function prediction." *Current opinion in structural biology* 19.3 (2009), pp. 357–62. DOI: 10.1016/j.sbi.2009.03.008. URL: http://www.sciencedirect.com/science/article/pii/S0959440X09000438.

[33] David de Juan, Florencio Pazos, and Alfonso Valencia. "Emerging methods in protein co-evolution." *Nature reviews. Genetics* 14.4 (2013), pp. 249–61. DOI: 10.1038/nrg3414. URL: http://www.ncbi.nlm.nih.gov/pubmed/23458856.

[34] S W Lockless and R Ranganathan. "Evolutionarily conserved pathways of energetic connectivity in protein families." *Science (New York, N.Y.)* 286.5438 (1999), pp. 295–299. DOI: 10.1126/science.286.5438.295.

[35] S R Sprang. "G protein mechanisms: insights from structural analysis." *Annual review of biochemistry* 66 (1997), pp. 639–678. DOI: 10.1146/annurev.biochem.66.1.639.

[36] H R Bourne, D a Sanders, and F McCormick. "The GTPase superfamily: conserved structure and molecular mechanism." *Nature* 349.6305 (1991), pp. 117–127. DOI: 10.1038/349117a0.

[37] Johannes L Bos, Holger Rehmann, and Alfred Wittinghofer. "GEFs and GAPs: critical elements in the control of small G proteins." *Cell* 129.5 (2007), pp. 865–77. DOI: 10.1016/j.cell.2007.05.018. URL: http://www.sciencedirect.com/science/

article/pii/S0092867407006551http://www.ncbi.nlm.nih.gov/pubmed/17540168.

[38] Krister Wennerberg, Kent L Rossman, and Channing J Der. "The Ras superfamily at a glance." *Journal of cell science* 118.Pt 5 (2005), pp. 843–846. DOI: 10.1242/jcs.094300.

[39] Johannes L Bos. "ras oncogenes in human cancer: a review." *Cancer research* 49.17 (1989), pp. 4682–9. URL: http://www.ncbi.nlm.nih.gov/pubmed/2547513.

[40] Rhiju Das and David Baker. "Macromolecular modeling with rosetta." *Annual review of biochemistry* 77 (2008), pp. 363–382. DOI: 10.1146/annurev.biochem.77.062906.171838.

[41] Dawei Sun et al. "Probing Gαi1 protein activation at single-amino acid resolution." *Nature structural & molecular biology* 22.9 (2015), pp. 686–94. DOI: 10.1038/nsmb.3070. URL: http://www.nature.com/doifinder/10.1038/nsmb.3070http://www.ncbi.nlm.nih.gov/pubmed/26258638.

[42] Robert D Finn et al. "The Pfam protein families database." *Nucleic acids research* 38.November 2007 (2010), pp. D211–D222. DOI: 10.1093/nar/gkm960.

[43] John Moult et al. "A Large-Scale Experiment to Assess Protein Structure Prediction Methods". *Proteins: Structure, Function, and Bioinformatics* 23.3 (1995), pp. ii–iv. DOI: 10.1002/prot.340230303.

[44] Kristian W. Kaufmann et al. *Practically useful: What the Rosetta protein modeling suite can do for you*. 2010. DOI: 10.1021/bi902153g.

[45] Steven a Combs et al. "Small-molecule ligand docking into comparative models with Rosetta." *Nature protocols* 8.7 (2013), pp. 1277–98. DOI: 10.1038/nprot.2013.074. URL: http://www.ncbi.nlm.nih.gov/pubmed/23744289.

[46] "Units in Rosetta" (). URL: https://www.rosettacommons.org/docs/latest/rosetta{_}basics/Units-in-Rosetta.

[47] Elizabeth H. Kellogg, Andrew Leaver-Fay, and David Baker. "Role of conformational sampling in computing mutation-induced changes in protein structure and stability." *Proteins* 79.3 (2011), pp. 830–8. DOI: 10.1002/prot.22921. URL: http://doi.wiley.com/10.1002/prot.22921http://www.ncbi.nlm.nih.gov/pubmed/21287615http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3760476.

[48] "Rosetta residue energy breakdown" (). URL: https://www.rosettacommons.org/docs/latest/application{_}documentation/analysis/residue-energy-breakdown.

[49] Lucas Gregorio Nivón, Rocco Moretti, and David Baker. "A Pareto-Optimal Refinement Method for Protein Design Scaffolds". *PLoS ONE* 8.4 (2013), pp. 1435–1439. DOI: 10.1371/journal.pone.0059004.

[50] "Preparing structures for Rosetta" (). URL: https://www.rosettacommons.org/manuals/archive/rosetta3.5{_}user{_}guide/dd/da1/preparing{_}structures.html.

[51] Arun S. Konagurthu et al. "MUSTANG: A multiple structural alignment algorithm". *Proteins: Structure, Function and Genetics* 64.3 (2006), pp. 559–574. DOI: 10.1002/prot.20921.

[52] Robert C Edgar. "MUSCLE: multiple sequence alignment with high accuracy and high throughput". *Nucleic Acid Research* 32.5 (2004), pp. 1792–1797. DOI: 10.1093/nar/gkh340.

[53] Konstantin Okonechnikov et al. "Unipro UGENE: a unified bioinformatics toolkit." *Bioinformatics (Oxford, England)* 28.8 (2012), pp. 1166–7. DOI: 10.1093/bioinformatics/bts091. URL: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts091http://www.ncbi.nlm.nih.gov/pubmed/22368248.

[54] W Kabsch and C Sander. "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." *Biopolymers* 22.12 (1983), pp. 2577–637. DOI: 10.1002/bip.360221211. URL: http://www.ncbi.nlm.nih.gov/pubmed/6667333.

[55] Yihui Xi. "knitr: A General-Purpose Package for Dynamic Report Generation in R." (2016). URL: https://github.com/yihui/knitr/releases/download/doc/knitr-manual.pdf.

[56] "Common Gα numbering" (). URL: http://www.mrc-lmb.cam.ac.uk/CGN/.

[57] Franco L. Simonetti et al. "MISTIC: Mutual information server to infer coevolution." *Nucleic acids research* 41.Web Server issue (2013), pp. 8–14. DOI: 10.1093/nar/gkt427.

[58] Daniela Kruschel and Bojan Zagrovic. "Conformational averaging in structural biology: issues, challenges and computational solutions." *Molecular bioSystems* 5.12 (2009), pp. 1606–1616. DOI: 10.1039/b917186j.

[59] H Frauenfelder, S G Sligar, and P G Wolynes. "The energy landscapes and motions of proteins." *Science (New York, N.Y.)* 254.5038 (1991), pp. 1598–603. URL: http://www.ncbi.nlm.nih.gov/pubmed/1749933.

[60] Natasha S. Latysheva et al. "How do disordered regions achieve comparable functions to structured domains?" *Protein Science* 24 (2015), n/a–n/a. DOI: 10.1002/pro.2674. URL: http://www.ncbi.nlm.nih.gov/pubmed/25752799.

[61] Carol a. Rohl et al. "Protein Structure Prediction Using Rosetta". *Methods in Enzymology* 383.2003 (2004), pp. 66–93. DOI: 10.1016/S0076-6879(04)83004-0.

[62] Michael P Fay and Michael a Proschan. "Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules." *Statistics surveys* 4 (2010), pp. 1–39. DOI: 10.1214/09-SS051. URL: http://www.ncbi.nlm.nih.gov/pubmed/20414472http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2857732.

[63] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Ggplot". 33.1 (2010), pp. 1–3.

[64] Martin Krzywinski and Naomi Altman. "Points of Significance: Visualizing samples with box plots". *Nature Methods* 11.2 (2014), pp. 119–120. DOI: 10.1038/nmeth.2813. URL: http://www.nature.com/doifinder/10.1038/nmeth.2813.

[65] "geom_boxplot" (). URL: http://docs.ggplot2.org/0.9.3.1/geom{_}boxplot.html.

[66] Robert McGill, John W. Tukey, and Wayne a. Larsen. "Variations of Box Plots". *The American Statistician* 32.1 (1978), pp. 12–16. DOI: 10.2307/2683468.

[67] Peter J A Cock et al. "Biopython: Freely available Python tools for computational molecular biology and bioinformatics". *Bioinformatics* 25.11 (2009), pp. 1422–1423. DOI: 10.1093/bioinformatics/btp163.

[68] Thomas Hamelryck and Bernard Manderick. "PDB file parser and structure class implemented in Python". *Bioinformatics* 19.17 (2003), pp. 2308–2310. DOI: 10.1093/bioinformatics/btg299.

[69] "SciPy (scientific python)" (). URL: https://en.wikipedia.org/wiki/SciPy.

[70] Wes McKinney. "Data Structures for Statistical Computing in Python". *Proceedings of the 9th Python in Science Conference* 1697900.Scipy (2010), pp. 51–56. URL: http://conference.scipy.org/proceedings/scipy2010/mckinney.html.

[71] "NumPy (numerical python)" (). URL: https://en.wikipedia.org/wiki/NumPy.

[72] Granger Brian E. Perez Fernando. "IPython: A System for Interactive Scientific Computing, Computing in Science and Engineering". *Computing in Science and Engineering* 9.3 (2007), pp. 21–29. DOI: 10.1109/MCSE.2007.53. URL: http://scitation.aip.org/content/aip/journal/cise/9/3/10.1109/MCSE.2007.53.

[73] R Core Team. "R: A Language and Environment for Statistical Computing" (2015). URL: https://www.r-project.org/.

[74] Hadley Wickham. "The Split-Apply-Combine Strategy for Data". *Journal of Statistical Software* 40.1 (2011), pp. 1–29. DOI: 10.1.1.182.5667. URL: http://www.jstatsoft.org/v40/i01/.

[75] Hadley Wickham. "Reshaping Data with the reshape Package". *Journal of Statistical Software* 21.12 (2007), pp. 1–20. DOI: 10.1016/S0142-1123(99)00007-9. URL: http://www.jstatsoft.org/v21/i12.

[76] E F Pettersen et al. "UCSF chimera - A visualization system for exploratory research and analysis". *Journal of Computational Chemistry* 25.13 (2004), pp. 1605–1612. DOI: 10.1002/jcc.20084. URL: http://onlinelibrary.wiley.com/doi/10.1002/jcc.20084/abstract.

[77] Sameer Velankar et al. "SIFTS: Structure Integration with Function, Taxonomy and Sequences resource". *Nucleic Acids Research* 41.D1 (2013), pp. 483–489. DOI: 10.1093/nar/gks1258.