Hyperpolarization-Activated Cyclic Nucleotide-Gated Channels—Structure and Evolution

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Vom Fachbereich Biologie der Technischen Universität Darmstadt zur Erlangung des akademischen Grades eines *Doctor rerum naturalium* genehmigte Dissertation von

Dipl.-Biol. Stephanie Weißgraeber geb. Held aus Wiesbaden

1. Referent: Prof. Dr. Kay Hamacher
2. Referent: Prof. Dr. Gerhard Thiel

Eingereicht am: 12. August 2014
Mündliche Prüfung am: 13. Oktober 2014

Darmstadt 2015 – D17

Summary

Computational models can shed light on protein function and the underlying mechanisms, where experimental approaches reach their limit. We developed an *in silico* mechanical model to analyze the process of cAMP-induced modulation in hyperpolarization-activated cyclic nucleotide-gated (HCN) channels, which conduct cations across the membrane of mammalian heart and brain cells. The structural analysis revealed a quaternary twist of the four subunits of the HCN channel tetramer. This motion has previously been shown to be part of the voltage-gating mechanism of other ion channels.

The insight gained from the mechanical approach was supported by results of analyses of intramolecular coevolution: Covariation of amino acids is induced by compensating mutations that maintain vital functions of a protein. Therefore, these covariations can be used to locate positions relevant for protein function. We found long-range coevolutionary relationships in HCN that suggest the existence of large domain rearrangements like the ones we found for the allosteric conformational change upon cAMP binding.

This thesis can be divided into two approaches: one based on structural data and another which analyzes sequence information. Together these results contribute to a deeper understanding of the gating mechanism of HCN channels.

- Mechanics of the HCN channel
    - A homology model of the transmembrane domain of the HCN4 channel was developed and joined with the crystal structure of the C-terminal domain to create a combined model of HCN4.
    - Release of cAMP from the binding pocket was simulated using an elastic network model and linear response theory to study the resulting conformational change.
    - The displacement from this allosteric change was compared to intrinsic low frequency modes of the protein structure.
    - Contacts were switched off one by one to identify key players of the observed motion.
- Intramolecular coevolution of HCN channels
    - Parameter sets for multiple sequence alignments were analyzed with a visual analytics approach to improve alignment quality prior to coevolutionary analysis.
    - Graph measures of the coevolutionary network of HCN were compared to four other proteins and two null models.

- We identified pairwise relationships that show long-range coevolution between the transmembrane region and the C-terminal domain.
- Three-dimensional mutual information revealed coevolving groups of residues at the interface between neighboring subunits of the tetramer.

Zusammenfassung

Computermodelle können zur Aufklärung von Proteinfunktionen und den ihnen zugrunde liegenden Mechanismen beitragen, wo Experimentalansätze an ihre Grenzen stoßen.

Es wurde ein *in silico* Modell entwickelt, um die Modulation des HCN-Kanals durch seinen Liganden cAMP zu untersuchen. HCN-Kanäle (*hyperpolarization-activated cyclic nucleotide-gated channel*) sind spannungsgesteuerte Kationenkanäle, die in der Membran von Herz- und Gehirnzellen in Säugetieren vorkommen. Die Ergebnisse unserer Strukturanalyse zeigen, dass das Binden von cAMP eine Torsion der vier Untereinheiten des Kanals auslöst. Ein ähnlicher Mechanismus ist bereits von anderen Kanälen bekannt und ist dort Teil des Öffnungsmechanismus der Kanalpore.

Die aus der Analyse des mechanischen Modells gewonnenen Erkenntnisse werden von Ergebnissen der intramolekularen Koevolutionsuntersuchung bestätigt. Korrelierte Mutationen von Aminosäuren treten dort auf, wo wichtige Proteinfunktionen erhalten werden müssen. Darüber kann man Positionen identifizieren, die an solchen Funktionen beteiligt sind. Es konnten mehrere koevolutionäre Verbindungen mit langer Reichweite identifiziert werden, die auf die Existenz einer globalen Konformationsänderung hindeuten. Eine solche konnte mit dem Strukturmodell gezeigt werden.

Die Untersuchungen in dieser Arbeit lassen sich in zwei Kategorien unterteilen: strukturelle und sequenzbasierte Ansätze. Die Kombination der Ergebnisse trägt zu einem besseren Verständnis der Steuerung von HCN-Kanälen bei.

- Mechanik des HCN-Kanals
  - Ein Homologiemodell der Transmembrandomäne des HCN4-Kanals wurde entwickelt und mit der Kristallstruktur der C-terminalen Domäne verbunden.
  - Dissoziation von cAMP aus der Bindetasche wurde anhand eines Kugel-Feder-Modells mit einem Kraftvektor simuliert, um die daraus resultierende Konformationsänderung zu untersuchen.
  - Die Verschiebung aufgrund dieses allosterischen Effekts wurde mit den intrinsischen Schwingungen der Proteinstruktur verglichen.
  - Durch Einzelkontaktabschaltung konnte die Relevanz von Interaktionen für die beobachtete Bewegung analysiert werden.
- Intramolekulare Koevolution des HCN-Kanals
  - Multiple Sequenzalignments mit verschiedenen Parametersätzen wurden analysiert um die Qualität von Alignments für die Koevolutionsanalysen zu verbessern.

- Netzwerkmaße von Koevolutionsgraphen des HCN-Kanals wurden mit denen von vier anderen Proteinen und mit zwei Nullmodellen verglichen.
- Räumlich weit entfernte Aminosäuren im Transmembranbereich und der C-terminalen Domäne koevolvieren.
- Koevolution zwischen Dreiergruppen von Aminosäuren an der Kontaktfläche zwischen benachbarten Untereinheiten deutet auf einen komplexen Bindungsmechanismus hin.

Publications and Contributions

Parts of this thesis are based on or have already been presented in the following publications.

- Weißgraeber S, Hamacher K (**2012**) Generalized correlations in molecular evolution: A critical assessment. *From Computational Biophysics to Systems Biology (CBSB11)–Celebrating Harold Scheraga's 90th Birthday* 8:231 [165]

  I wrote the manuscript.

- Wächter M, Jäger K, Weißgraeber S, Widmer S, Goesele M, Hamacher K (**2012**) Information-theoretic analysis of molecular (co) evolution using graphics processing units. In *Proceedings of the 3rd International Workshop on Emerging Computational Methods for the Life Sciences*, 49–58, ACM [155]

  I supervised the student who performed the analysis, participated in evaluation and biological interpretation of the results, contributed to the manuscript and presented our findings at the ECMLS workshop.

- Wächter M, Jäger K, Thürck D, Weißgraeber S, Widmer S, Goesele M, Hamacher K (**2014**) Using graphics processing units to investigate molecular coevolution. *Concurrency and Computation: Practice and Experience* 26(6):1278–1296 [154]

  This study is an extension of Wächter et al. (2012), see above.

- Heß M, Bremm S, Weißgraeber S, Hamacher K, Goesele M, Wiemeyer J, von Landesberger T (**2014**) Visual exploration of parameter influence on phylogenetic trees. *IEEE Computer Graphics and Applications* 34(2):48–56 [74]

  I prepared and evaluated the protein data set.

- Weißgraeber S, Thiel G, Moroni A, Hamacher K (**2014**) A reduced mechanical model for cAMP-modulated gating in HCN channels. *manuscript in preparation*

  This manuscript will present the mechanical analysis of Chapter 2.

Contents

1 Introduction to Ion Channels

Ion channels are proteins that facilitate the transport of charged atoms across the cell membrane through a hydrophilic pore that connects the extracellular space to the cytosol. In contrast to transporter proteins, ion conduction through channel proteins always occurs down the electrochemical gradient and therefore does not need to consume metabolic energy [26]. Most channels are selective: a certain type of ion is conducted with high efficiency while others are excluded. Their function is vital for the regulation of the membrane potential, which is particularly important in muscles and cells of the nervous system. In these cells, changes in the membrane potential transmit signals which, for example, cause a muscle to contract. To allow regulation of conduction activity, control mechanisms for channel opening are necessary. The regulated process of opening and closing of the channel's pore is called *gating*. The major classification of ion channels depends on two properties: gating stimulus (e.g., voltage, ligand binding, temperature) and ion selectivity (e.g., sodium, potassium, chloride) [75].

Voltage-gated potassium channels have a tetrameric setup (see Figure 1.1). The simplest type in eukaryotes is a homomer with four identical subunits arranged around the membrane-spanning pore. Each subunit consists of six transmembrane helices (named S1–S6) as shown in Figure 1.2. The voltage sensing domain is composed of the helices S1–S4. The helices S5, S6 and the region between them form the core of the channel [172]. Some bacterial and viral potassium channels, such as Kcsa and Kcv, even consist of just this two-helix domain [49, 60]. The pore loop between S5 and S6 holds the selectivity filter—the narrowest part of the pore, which can only be passed by a specific type of ions. Gating of the channel occurs in the C-terminal part of the S6 helix that points toward the cytoplasm: a conserved sequence motif (a proline-containing motif or a single glycine) acts as a hinge which allows a bending motion of the S6 helix. Thereby, the S6 helices of the four subunits cross and close the channel [83].

In voltage-gated channels, opening and closing of the channel depends on changes in the membrane potential. Most channels open when a certain threshold of membrane depolarization is reached. The heart of the voltage-sensing domain is the S4 helix. It holds several conserved, positively charged residues (mostly arginines) that cause the helix to move according to the current charge gradient. In the depolarized state, S4 is located toward the extracellular space, whereas it moves toward the cytosol during repolarization [150, 152]. This movement is conveyed to the pore region via the S4–S5-linker. The transmission is based on amino acid-specific interactions between the linker and the C-terminal end of the S6 helix, which can be disturbed by alanine mutations [81, 149]. Contacts between the S4 and S5 helix of neighboring subunits ensure cooperative gating of the tetramer [148].

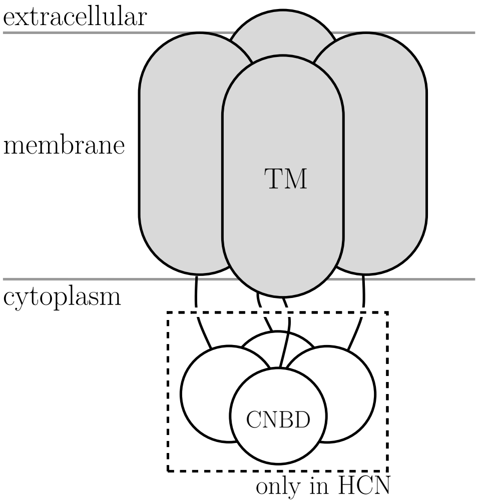


Figure 1.1: Sketch of a voltage-gated potassium channel homotetramer with the transmembrane domain (TM) embedded in the cell membrane. Each of the four subunits contributes to the ion conducting pore that is formed in the middle of the assembly. The cytosolic C-terminal domain (CNBD: cyclic nucleotide-binding domain) is only present in HCN channels.

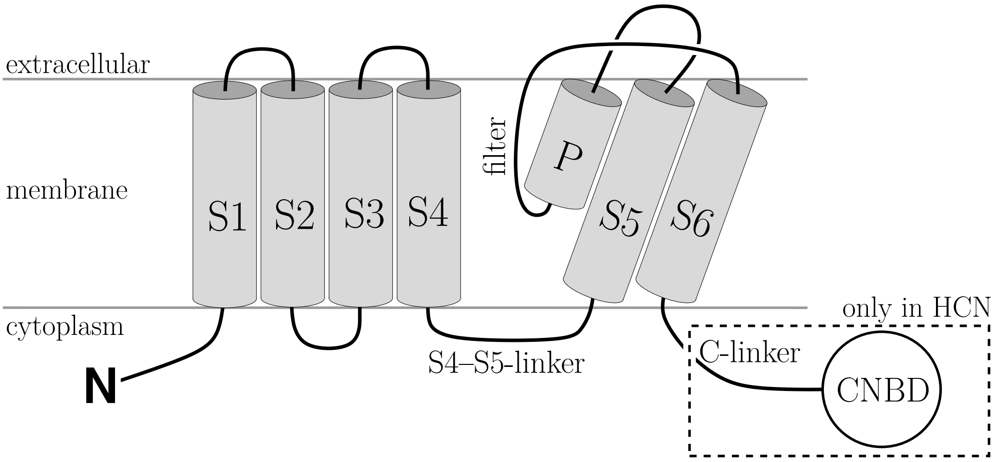


Figure 1.2: Monomer of a voltage-gated potassium channel in the cell membrane. Transmembrane helices are labeled S1–S6, P denotes the pore helix. At the cytosolic N-terminus, the channel can hold a tetramerization domain (e.g., a PAS domain in hERG channels; for a list of abbreviations see Appendix B). The C-terminal domain composed of C-linker and CNBD is only present in HCN channels.

1.1 HCN Channels

Hyperpolarization-activated cyclic nucleotide-gated (HCN) channels—in contrast to most other voltage-gated channels—open upon hyperpolarization of the membrane. They share the setup of a homotetramer and the selectivity filter sequence (GYG) with voltage-gated potassium channels (see Figures 1.1 and 1.2) and are therefore assigned to this protein family. In spite of this classification, they are only weakly selective for potassium: the conductance ratio of K^+ to Na^+ is 4:1 [45]; for other potassium channels this ratio is approximately 1000:1 [106].

As their name implies, HCN channels bind cyclic nucleotides, which influences their gating behavior. When cAMP binds to the C-terminal cyclic nucleotide-binding domain (CNBD), the hyperpolarization threshold that triggers opening of the channel is moved toward the resting potential, i.e., to more positive voltages [46]. The principle of this modulation is *disinhibition*: The C-terminal domain of the HCN channel has an autoinhibitory function. It is hypothesized that binding of intracellular cAMP induces an allosteric conformational change in the CNBD and C-linker that removes their inhibiting effect on the channel [156, 157]. However, the exact structure of the HCN channel and the rearrangements during cAMP binding and voltage gating remain largely unknown because of difficulties in crystallizing membrane proteins.

HCN channels open when the membrane becomes hyperpolarized. As mentioned above, this is quite unusual, since most voltage-gated potassium channels open upon depolarization. Nevertheless, the composition and movement of the voltage-sensing S4 helix in the membrane in HCN are the same as in other channels: positively charged residues cause the S4 helix to move outward and inward upon de- and hyperpolarization, respectively. Therefore, the different behavior seems to originate from a difference in transmission of the voltage signal from S4 to the pore region [107, 108]. Up to now, this process is not fully understood. Furthermore, cAMP binding and voltage both influence channel gating. The S4–S5-linker, that transmits the signal of a change in voltage to the pore, has been shown to interact with the C-linker [43].

HCN channels and their exceptional property of activation upon hyperpolarization of the membrane contribute to several processes in neurons and cardiac muscle cells: they control rhythmic activity in thalamus cells [98], modulate synaptic transmission [17], participate in determination of the resting potential of the cell membrane [47] and influence light response in the retina [88]. Perhaps their most important function is the cardiac pacemaker current, that controls initiation and regulation of the heartbeat. In the sinoatrial node—the region of the heart that initiates cardiac contraction—the channels are activated by the hyperpolarization phase between two action potentials, which causes an inwardly directed cation current. This current induces a depolarization phase and with it the next action potential. Therefore, a shift of the channel's activation threshold toward more positive voltages accelerates the heartbeat. Since this is accomplished by cAMP binding to the HCN channel, the intracellular cAMP level influences the heart rate [96, 127]. In neurons, HCN channels influence long-term potentiation in synapses. By this means, they regulate learning and memory [117].

There are four subtypes of mammalian HCN channels. Although their sequence in the transmembrane and the C-terminal domain is very similar, they exhibit some distinct features and show tissue-specific expression patterns. HCN1 shows the fastest activation kinetics combined with the most positive voltage threshold, which is only slightly shifted even by saturating cAMP levels. It is expressed in certain areas of the brain as well as in the sinoatrial node. HCN2 is expressed throughout the heart and brain and is very sensitive to the intracellular cAMP concentration [73]. The least studied of the four types is HCN3, which is mainly due to technical problems [32]. It seems to be largely insensitive to cAMP. The reason for this remains unknown. In the heart, HCN4 is the most common subtype. It shows the slowest opening kinetics. As well as HCN2, it is strongly modulated by cAMP. All mammalian subtypes of HCN can also be modulated by cGMP [24].

Due to the regulatory tasks that HCN channels perform, dysfunction can cause arrhythmia in the heart [138]. In mice, HCN mutations have been shown to cause learning defects [118] and absence epilepsy [104]; some mutations are lethal in the embryonic phase [137]. Several approved drugs that affect the nervous system modulate HCN channel activation [123]. Recently, HCN channels have been found to play a role in pain signaling and are therefore discussed as drug target for analgesics [53]. As a consequence, the study of HCN channels has become a promising field of research over the last years.

2 Mechanics of the HCN Channel

This chapter describes the modeling of the hyperpolarization-activated cyclic nucleotide-gated channel 4 (HCN4 channel) structure and the process of its modulation by cyclic adenosine monophosphate (cAMP). The structural model is obtained via homology modeling based on a related K_v channel. Based on this model the allosteric reaction of the channel to ligand binding is analyzed. A simulation of the motion of the channel upon ligand dissociation is presented in Section 2.3.3. In Section 2.3.4 this motion is compared to oscillations intrinsic to the protein structure. Finally, groups of residues as well as individual contacts that play a role in the global domain motion are identified in Section 2.3.5 and Section 2.3.6, respectively.

2.1 Introduction

Proteins are not static objects but subject to thermal fluctuations [56]. Additionally, for many proteins, motion is a vital part of their function: they are molecular machines [3]. Therefore, studying their dynamics is important for understanding how they work.

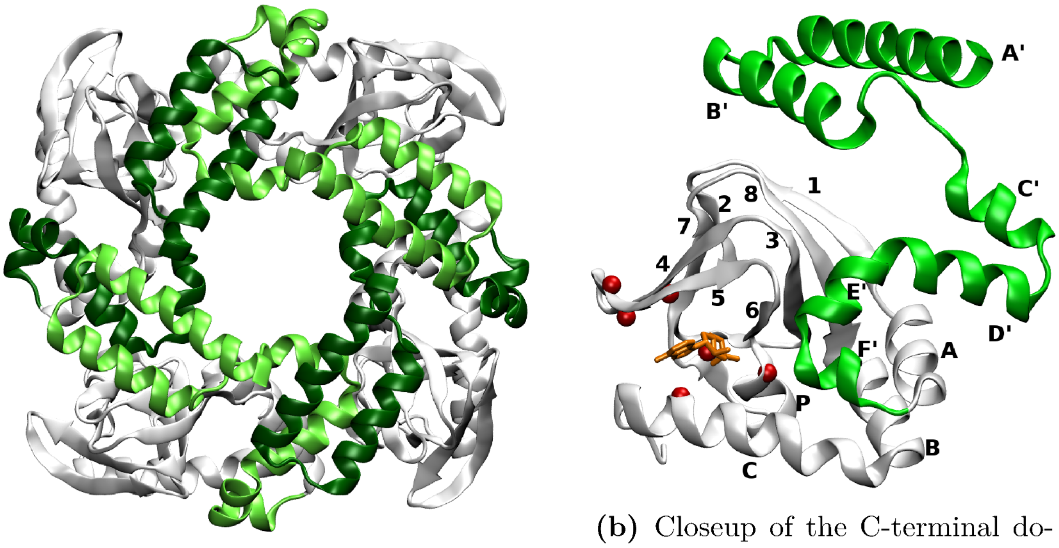
2.1.1 Three-Dimensional Structure of the HCN Channel

In order to study protein mechanics, a structural model is required. Due to the difficulties of crystallizing membrane proteins, the full structure of the HCN channel still remains unknown.

The transmembrane domain of HCN has also not been crystallized yet. It is expected to be closely related to that of other ion channels with six transmembrane helices. In 2005, Giorgetti et al. created homology models of the transmembrane region of murine HCN2 and sea urchin HCN [62]. However, they did not publish any coordinates of their model. As a consequence, no adequate model of the HCN transmembrane domain is available at present.

For the cyclic nucleotide-binding domain (CNBD), on the other hand, structural data is available. Lolicato et al. were able to crystallize the C-terminal domain of HCN1, HCN2 and HCN4 containing the C-linker and the CNBD in the ligand-bound state (PDB IDs 3U0Z, 3U10 and 3U11) [101]. So far, no one has been able to capture the CNBD of HCN in the ligand-free state.

The crystal structure of HCN4 C-terminal domain is shown in Figure 2.1. The resolved C-terminus consists of two parts: the C-linker highlighted in green and the CNBD colored white. Figure 2.1a shows the HCN4 tetramer. The C-linkers of adjoining subunits form



(a) View from the membrane plane onto the tetramer of the C-terminal domain. The C-linkers of adjacent subunits are colored in different shades of green.

(b) Closeup of the C-terminal domain of one subunit. The C-linker is highlighted in green, cAMP (orange stick representation) is located in the binding pocket (red spheres). Helices and β strands are labeled.

Figure 2.1: Crystal structure of the C-terminal domain of HCN4 (PDB 3U11).

an "elbow-on-the-shoulder" motif: the bend between helix A' and B' ("elbow") of one subunit rests on the C' and D' helix ("shoulder") of its neighbor [173].

The interface between the subunits is located almost exclusively in the C-linker region. The CNBD is a domain conserved in many cyclic nucleotide-binding proteins (see Section 2.1.2). It consists of a β roll comprised of eight β strands and four α helices. The cAMP binding pocket is located between β strands 5 and 6, the P helix and the C helix, which closes the binding pocket like a lid. Six amino acids are known to interact with cAMP [4]: V389, T391, K395, E407, R416, R457 (for numeration see Appendix A); they are highlighted in red in Figure 2.1b.

The CNBD is a conserved domain wide-spread among various protein families, such as protein kinases [145], bacterial transcription factors [97] and cyclic nucleotide-gated (CNG) channels [85].

2.1.2 Related Proteins as Modeling Template Candidates

Up to now, there are no cAMP-free experimental structures of HCN channel CNBDs at atomic resolution. The fold of cAMP-regulated channel K1 of the bacterium *Mesorhizobium loti* (MlotiK1) is highly similar to that of HCN. The main difference is found in the C-linker, which is much shorter in MlotiK1. No voltage-sensitivity has been reported for MlotiK1 [37, 116].

Schünke et al. [130] crystallized MlotiK1 in complex with cAMP as well as its ligand-free form. These crystal structures reveal that the cAMP binding pocket is more open in

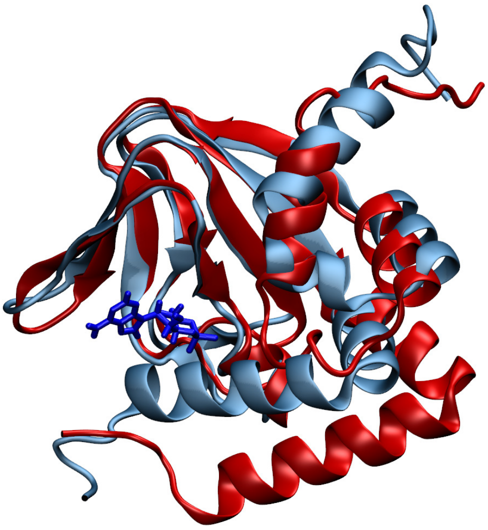


Figure 2.2: Cyclic nucleotide-binding domain of MlotiK1 ligand-free (red, PDB ID: 2KXL) and in complex with cAMP (blue, PDB ID: 2K0G). Ligand is shown as stick model in blue.
Structural fit was performed over β strand 1, 3 and 8. Image was rendered in VMD [78].

the ligand-free conformation. Figure 2.2 shows the B and C helix tilted away from the β roll motif when cAMP is not bound. The binding of cAMP to MlotiK1 seems to cause a contraction of the binding pocket, which largely consists of a hinge bending motion of the C helix toward the β roll.

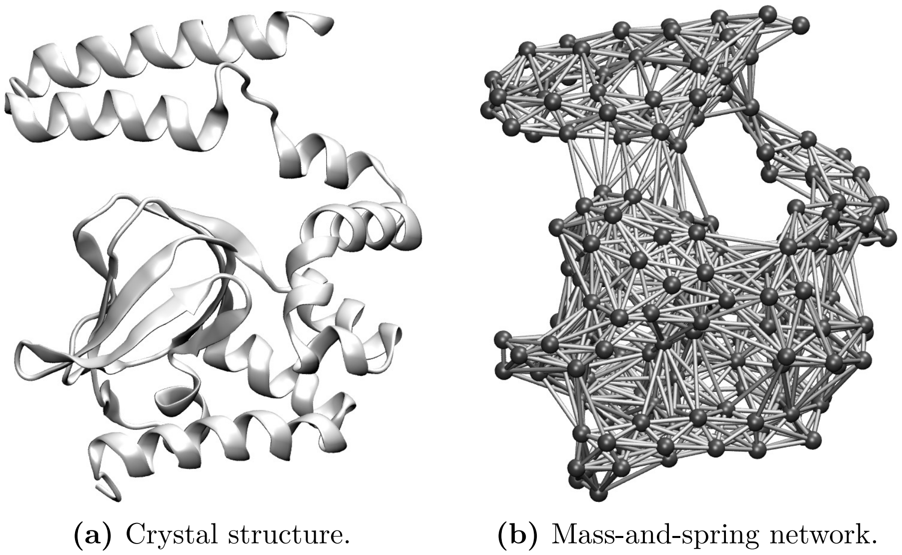
Similarly, the cAMP-binding domain of the catabolite activator protein (CAP) closes upon ligand binding [120, 121]. Although CAP fulfills a cellular function very different from that of ion channels, its CNBD is conserved and has the same basic structure.

Following these observations, the model for the release of cAMP from the binding site of the HCN channel was created: the ligand was removed and forces were applied to widen the binding pocket in accordance to the structural changes in the cAMP-binding domains of MlotiK1 and CAP.

As mentioned in Section 1.1, HCN channels belong to the protein family of voltage-gated potassium channels. These channels are generally named K_V channels (K for potassium, V for voltage-gated). There are 40 different genes coding for K_V channels in the human genome. So far, crystallization of these six-transmembrane channels has only been accomplished in the open pore conformation, i.e., including a depolarized voltage sensor domain. Crystal structures in the closed state have only been resolved for two-transmembrane helix channels [92].

2.1.3 A Coarse-Grained Model for Protein Mechanics: the Anisotropic Network Model

Normal mode analysis uses an approximation of the potential harmonic functions and therefore captures the dynamics as oscillations around the equilibrium state [27, 99]. The motion of a protein is a combination of its normal modes. Thorough energy minimization of the experimentally determined structure is necessary before normal mode analysis can be performed. Furthermore, it takes a lot of computation time to calculate the interaction potentials [164] between all atoms in the system.



(a) Crystal structure. **(b)** Mass-and-spring network.

Figure 2.3: Illustration of an elastic network model of the C-terminal domain of the HCN4 channel (PDB 3U11). Each amino acid is represented by a bead centered at its C^α atom. All nodes within a specified cutoff distance of each other (here: 10 Å) are connected by springs.

Elastic network models (ENMs) are a simplification of normal mode analysis. They approximate the protein to a mass-and-spring network, where each amino acid is a node and the springs are the interactions between them [8, 10, 146]. The coordinates of the nodes in three-dimensional space are the coordinates of their C^α atoms $\boldsymbol{x}_i = (x_i, y_i, z_i)$ with $i = 1, \ldots, N$, N being the number of nodes. An illustration of the elastic network model is given in Figure 2.3.

The anisotropic network model (ANM) is a special type of elastic network model. The potential V is calculated as follows:

$$V = \frac{1}{2} \sum_{i,j}^{N} \gamma_{ij} (R_{ij} - R_{ij}^0)^2, \tag{2.1}$$

where γ_{ij} is the spring constant, which can be interaction type-specific, $R_{ij} = |\boldsymbol{x}_i - \boldsymbol{x}_j|$ is the distance between residues i and j and $R_{ij}^0 = |\boldsymbol{x}_i^0 - \boldsymbol{x}_j^0|$ the distance of these residues in the equilibrium state, i.e., the starting structure [76]. In the case of ENMs no energy minimization is required prior to mode analysis, therefore, a crystal structure or similar models can be used directly.

The Hessian matrix $\mathcal{H}$ of the system is the second derivative of the potential. It contains $N \times N$ super-elements $\mathcal{H}_{ij}$ of the form:

$$\mathcal{H}_{ij} = \begin{pmatrix} \frac{\partial^2 V}{\partial x_i \partial x_j} & \frac{\partial^2 V}{\partial x_i \partial y_j} & \frac{\partial^2 V}{\partial x_i \partial z_j} \\ \frac{\partial^2 V}{\partial y_i \partial x_j} & \frac{\partial^2 V}{\partial y_i \partial y_j} & \frac{\partial^2 V}{\partial y_i \partial z_j} \\ \frac{\partial^2 V}{\partial z_i \partial x_j} & \frac{\partial^2 V}{\partial z_i \partial y_j} & \frac{\partial^2 V}{\partial z_i \partial z_j} \end{pmatrix}, \tag{2.2}$$

where x, y and z are the spatial coordinates of the residues i and j. Since each super-element's dimension is 3×3, the dimension of the complete Hessian matrix is $3N \times 3N$.

Singular value decomposition yields the eigenvalues and eigenvectors of the Hessian [64]. Six of the Hessian's eigenvalues vanish because the protein is not fixated and therefore has three rotational and three translational degrees of freedom. From the $3N - 6$ non-zero eigenvalues λ and their corresponding eigenvectors v the Moore-Penrose pseudoinverse [113, 122] of the Hessian $\mathcal{H}^{-1}$ can be computed:

$$\mathcal{H}^{-1} = \sum_{k=1}^{3N-6} \frac{\boldsymbol{v}_k \boldsymbol{v}_k^{\mathrm{T}}}{\lambda_k}, \tag{2.3}$$

resulting in the covariance matrix of the ANM, which contains information about how the residues in the elastic network are coupled.

Similar to normal mode analysis, the eigenvectors of the Hessian matrix of an ENM correspond to the fluctuations a protein is subject to. Each of these eigenvectors contains $3N$ entries since the motion of each residue in all three spatial dimensions are required to describe the complete oscillation of the protein. Their respective eigenvalues are the square of the frequencies of these fluctuations. High frequency fluctuations are local, stabilizing movements, while low frequency modes (also called soft modes) describe global, collective motions, that affect large parts of the protein [12].

Several studies in the past years have shown that low frequency modes obtained by normal mode analysis or via ENMs provide valuable insight into protein mechanics [11]. According to Tobi et al., ligand binding does not induce conformational change but stabilizes conformations that are already accessible to the ligand-free form of the protein [147]. Therefore, the native state can be employed to gather information about allosteric reaction to ligand binding.

Research in the past years has shown that allosteric conformational changes can often be described by one or a subset of low frequency modes if the motion has a high collectivity, i.e., involves a high percentage of residues. Notable studies in the field were conducted by Xu et al., who were able to describe the conformational change between the tense and relaxed forms of hemoglobin using an elastic network model [170], Wang et al., who analyzed the ratchet-like motion of the 70S ribosome [159] and many others [13, 142, 171].

It has been shown that ENM approaches are well-suited to study membrane proteins, regardless of the fact that membrane effects are neglected [12, 128].

Elastic network models are computationally inexpensive and can be used at different levels of coarse-graining [48]. The global shape of the molecule is the dominant feature, whereas the ENM's characteristics are robust to variations in adjustable parameters such as interaction strength [103].

2.1.4 Linear Response Theory—Studying the Reaction of a Protein to Force Application

In 2005, Ikeguchi et al. [79] published an approach to describe conformational changes of proteins upon ligand binding. They used linear response theory (LRT) [71] to simulate the binding of ligands to proteins. LRT states that the response to a perturbation due to ligand binding is related to the equilibrium fluctuations of the receptor in the unperturbed state.

The expected coordinate shift for each residue $\Delta \boldsymbol{r}_i$ can be computed from the covariance matrix $\mathcal{H}^{-1}$ of the ligand-free state and the perturbation upon ligand binding, which is simply a force acting on the binding pocket:

$$\Delta \boldsymbol{r}_i \simeq \beta \sum_{j=1}^{N} \left[\mathcal{H}^{-1}\right]_{ij} \boldsymbol{f}_j, \tag{2.4}$$

with β being the force constant and $\boldsymbol{f}_j$ the force vector. $\Delta \boldsymbol{r}_i$ consists of three components, one for each direction in space (x, y and z).

LRT is a very useful tool if a hypothesis concerning the allosteric reaction of the protein under investigation is at hand.

2.1.5 Assessment of Residue Contacts Using Switch-Off Analysis

In order to identify interactions important for the reaction of the protein to ligand binding, we perform a gedankenexperiment: all contacts are switched off one by one, i.e., the interaction strength for the respective amino acid pair is set to zero in Equation 2.1 before applying the force that mimics the effect of ligand binding. This method is similar to an alanine scan in the laboratory, but there are two major differences: first, an alanine mutation only reduces interaction strength [112], it does not completely remove it. Second, mutation influences all interactions of a residue, while our switch-off model allows investigation of single interactions between two residues [68, 70].

2.2 Methods

Protein images were rendered with `VMD` [78]. Plots were created in `R` [126] using the `ggplot2` library [168].

2.2.1 Homology Modeling

The homology model of the HCN4 channel transmembrane domain was constructed with `SWISS-MODEL` [5, 66]. The HCN4 sequence with GenBank [21] identifier 29840776 was chosen as target sequence. The transmembrane domain of a related $\mathrm{K_v}$ channel (PDB [23] ID: 3LNM, chain B, biological assembly [143]) served as template. Sequence alignment and modeling were performed with `SWISS-MODEL`.

The loops between the transmembrane helices were remodeled *de novo* using the loop modeling function of `MODELLER` [129]. `STRIDE` [72] data were used to identify loop regions.

Since `SWISS-MODEL` is only able to build monomers, a tetramer had to be constructed manually. Coordinate superposition over C^{α} atoms of the modeled monomer and each of the four subunits of the template was performed to create a symmetric tetramer of the

Parameter	Setting
force field	amber
integrator	steep
nsteps	20 000
ns_type	grid
coulombtype	shift
rcoulomb	1.3
rcoulomb_switch	1.0
vdwtype	shift
rvdw	1.3
rvdw_switch	1.0
pbc	xyz
rlist	2.0
optimize_fft	yes
emtol	500

Table 2.1: `GROMACS` parameter settings for energy minimization. All parameters not listed were set to default values.

homology model. To improve assembly of the tetramer and avoid atom clashes, energy minimization was performed *in vacuo* with `GROMACS 4.5` [124] (for parameter settings see Table 2.1; all parameters not listed were left at default values).

To obtain a complete model of the HCN4 channel, the homology model of the tetrameric transmembrane domain was joined to the crystal structure of the tetrameric C-terminal domain (PDB ID: 3U11 [101]). An overlap of three amino acids between the two tetramers was used to merge the chains by superposing the coordinates of the C^{α} atoms of residue 268 (the starting residue of the crystal structure of the C-terminal domain; for residue numbering see Appendix A) and adjusting the PDB file. Afterwards, the joined model was energy minimized *in vacuo* with `GROMACS 4.5` (parameter settings are listed in Table 2.1).

2.2.2 Anisotropic Network Model

An anisotropic elastic network model (ANM) of the HCN4 joined model was built using the `BioPhysConnectoR` package [77] in `R`. The contact cutoff was 10 Å (as in [79]), covalent bonds were set to $82\,RT/\text{Å}^2$ with $R = k_B N_A$ (k_B: Boltzmann constant; N_A: Avogadro constant; T: temperature) [70]. Non-covalent interactions were set to $3.166\,RT/\text{Å}^2$, which is the average value of non-covalent interactions in the Miyazawa-Jernigan matrix [112]. The Hessian matrix of the system was calculated and the pseudoinverse of the Hessian was derived by singular value decomposition resulting in the covariance matrix of the ANM (see Equation 2.3).

2.2.3 cAMP Dissociation Model

As explained in Section 2.1.4, a force was simultaneously applied to six residues which constitute the cAMP binding pocket [4] (V389, T391, K395, E407, R416, R457; for numeration see Appendix A). The direction of the force was chosen to point from the geometric center of the heavy atoms of the bound cAMP toward the C^{α} atom of the forced residue. All force vectors were normalized to the same length so that a force of the same strength was applied to each binding pocket residue.

2.2.4 Low Frequency Mode Analysis

A singular value decomposition of the Hessian matrix of the HCN tetramer was performed in R. Modes which introduce the same change in all four subunits are considered non-degenerate. They were shown to best describe cooperative transitions of multimeric proteins [9]. Therefore, our analysis was restricted to these non-degenerate modes. To perform the symmetry check the magnitude of each three-dimensional $\Delta\boldsymbol{r}_i$ (see Equation 2.4) was computed. For the resulting vector of magnitudes, Pearson's correlation coefficient was computed between the first, second, third and fourth quarter. Modes were considered non-degenerate if the correlation between all quarters was higher than 0.95.

The overlap between each non-degenerate eigenvector and the displacement vector $\Delta\boldsymbol{r}$ from LRT was computed. The overlap I between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is

$$I_{\boldsymbol{ab}} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}|\,|\boldsymbol{b}|} = \cos\phi_{\boldsymbol{ab}}, \tag{2.5}$$

where ϕ_{ab} is the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$ [110].

We also computed linear combinations $\boldsymbol{l}$ of k non-degenerate modes $\boldsymbol{m}$ that best described the displacement upon LRT, i.e., had the highest overlap with $\Delta\boldsymbol{r}$:

$$\boldsymbol{l}_k = \sum_{i=1}^{k} \alpha_i\, \boldsymbol{m}_i, \tag{2.6}$$

with α_i being the scaling factor for the ith mode that is calculated as follows: Since the eigenvectors of the Hessian form an orthonormal basis, we can determine the part of $\Delta\boldsymbol{r}$ that can be described through mode $\boldsymbol{m}_i$ by normalizing $\Delta\boldsymbol{r}$ ($\boldsymbol{s} := \frac{\Delta r}{|\Delta\boldsymbol{r}|}$) and projecting $\boldsymbol{s}$ onto $\boldsymbol{m}_i$ [80]:

$$\boldsymbol{s}_{\boldsymbol{m}_i} = \frac{\boldsymbol{m}_i \cdot \boldsymbol{s}}{\underbrace{|\boldsymbol{m}_i|^2}_{=1}}\, \boldsymbol{m}_i = \underbrace{(\boldsymbol{m}_i \cdot \boldsymbol{s})}_{=|\boldsymbol{s}_{\boldsymbol{m}_i}|=\alpha_i}\, \boldsymbol{m}_i. \tag{2.7}$$

Thus, the scaling factor α_i is the dot product of $\boldsymbol{m}_i$ and $\boldsymbol{s}$, which also corresponds to the overlap between $\boldsymbol{m}_i$ and $\boldsymbol{s}$:

$$I_{\boldsymbol{m}_i\boldsymbol{s}} = \frac{\boldsymbol{m}_i \cdot \boldsymbol{s}}{\underbrace{|\boldsymbol{m}_i|\,|\boldsymbol{s}|}_{=1}} = \boldsymbol{m}_i \cdot \boldsymbol{s} = \alpha_i, \tag{2.8}$$

The overlap of these linear combinations $\boldsymbol{l}_k$ with $\Delta\boldsymbol{r}$ was then computed.

2.2.5 Clustering of Residue Groups

The distance between all C^α atoms in an HCN subunit was measured for the homology model and for the structure after force application, yielding a C^α distance matrix for each protein structure. Based on the difference matrix of these two matrices, residues were hierarchically clustered using the `hclust` function with complete-linkage method in `R`. The emerging tree was cut off at a height of 60 resulting in twelve groups of residues.

2.2.6 Switch-Off Model

Switching off a contact causes a perturbation in the covariance matrix. To this end, the respective elements of the Hessian need to be set to zero. To avoid building the Hessian and its inverse again for every switch-off, the Sherman-Morrison formula [132] was applied [69]:

$$\mathcal{H}^{-1}_{(m,n)} = (\mathcal{H} + \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \mathcal{H}^{-1} - \frac{\mathcal{H}^{-1}\boldsymbol{u}\boldsymbol{v}^T\mathcal{H}^{-1}}{1 + \boldsymbol{v}^T\mathcal{H}^{-1}\boldsymbol{u}}, \tag{2.9}$$

where the subscript (m, n) denotes the row and column of the inverse Hessian matrix entry that will be changed. $\boldsymbol{u}$ and $\boldsymbol{v}$ are $3N$-dimensional vectors quantifying the perturbation. $\boldsymbol{u}$ is the difference of the unit vector in the direction m ($\hat{\boldsymbol{e}}_m$) and the unit vector in the direction n ($\hat{\boldsymbol{e}}_n$); $\boldsymbol{v}$ is the product of the entry of the Hessian that is to be canceled out ($\mathcal{H}_{mn}$) and the vector $\boldsymbol{u}$.

$$\boldsymbol{u} = \hat{\boldsymbol{e}}_m - \hat{\boldsymbol{e}}_n \tag{2.10}$$

$$\boldsymbol{v} = \mathcal{H}_{mn}\boldsymbol{u} \tag{2.11}$$

Only one matrix entry at a time can be altered with this method. Since the Hessian contains entries for x, y and z coordinates of each residue, three perturbations are necessary for each contact switch-off.

To maintain symmetry in the tetramer all equivalent contacts were switched off in all subunits simultaneously.

2.3 Results and Discussion

2.3.1 Evaluation of the HCN4 Model

At present, there is no crystal structure data for the transmembrane domain of an HCN channel. Therefore, a homology model was constructed using the transmembrane domain of a related K_v channel as template.

Several modeling attempts using different homology modeling programs (`SWISS-MODEL`, `I-TASSER`, `MODELLER`) and templates (3LNM, 2R9R, 3BEH, 3LUT) were compared. To this end, the pairwise sequence alignments of HCN4 and related K_v channels were inspected. The best alignment of HCN4 could be achieved with the template 3LNM. The regions of the HCN4 sequence that were annotated as transmembrane helices and pore loop in GenBank were well aligned with the corresponding structural elements of 3LNM. For this reason and because of the fact that the structure of the six transmembrane helices should be fairly conserved, 3LNM was chosen as template sequence, in spite of low (11%) sequence identity to HCN4.

SWISS-MODEL accomplished the best alignment of HCN4 with the template 3LNM as well as the best homology model concerning structural integrity, secondary structure elements matching sequence annotation by GenBank and QMEAN [19] scores. Therefore, this approach was used to create the final model.

The homology model of the transmembrane region was added to the crystal structure of the C-terminal domain (PDB 3U11) to obtain a model including residues 254 to 718 of the functional channel comprising the full transmembrane domain, the C-Linker and the cyclic nucleotide-binding domain (CNBD). Figure 2.4 shows the assembled tetramer: side view with the transmembrane region in the upper part in (a) and top view from the extracellular space in (b). A single subunit is depicted in Figure 2.4c and an annotated illustration of the homology modeled transmembrane domain can be found in Figure 2.4d.

The QMEAN server was used to evaluate the joined model. Figure 2.5 shows the estimated error per residue of the HCN4 joined model. Unfortunately, the QMEAN score is not a completely reliable measure for judging the quality of membrane protein models, since the data set on which its computation is based mostly consists of soluble proteins [18].

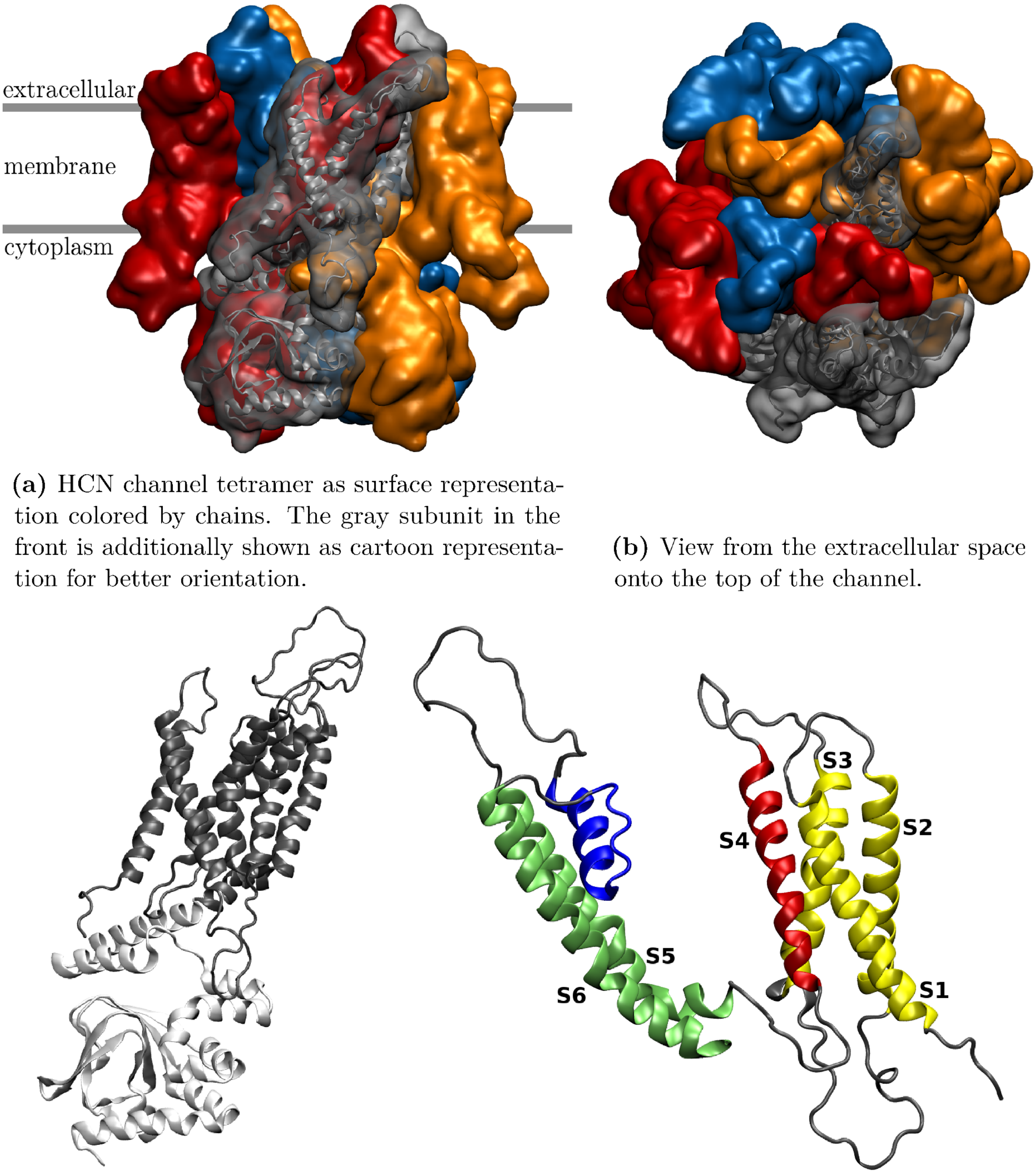
Nevertheless, the model is qualified for our coarse-grained approach as most parts are within an error range of less than 3.5 Å. However, we need to keep in mind that results for the loop regions should be treated with caution.

2.3.2 Modeling the Release of cAMP via Force Application

As mentioned in Section 2.1.2, the model for obtaining the cAMP-free conformation of HCN was based on the MlotiK1 and CAP cAMP-binding domains. Force vectors were chosen in order to widen the binding pocket thereby mimicking the dissociation of the ligand (see Section 2.2.3).

Interaction Cutoff Ikeguchi et al. used a 10 Å cutoff for the interaction network in their LRT experiments [79]. Since other studies using anisotropic network models suggest larger cutoffs [8, 70], the LRT in this thesis was performed under three different cutoff conditions: 10, 13 and 15 Å. Other than requiring a larger force constant to obtain the same amount of distortion, the results for the larger distances (data not shown) did not differ from those of the 10 Å cutoff. Therefore, the subsequent analyses were performed for one cutoff distance only, choosing 10 Å as in the original LRT paper.

Force Application to Single Subunits cAMP binding in HCN channels occurs cooperatively whereby HCN exhibits a dimer of dimers behavior—the probability for one subunit of a dimer to bind a ligand is increased if the other is already occupied [93].



(a) HCN channel tetramer as surface representation colored by chains. The gray subunit in the front is additionally shown as cartoon representation for better orientation.

(b) View from the extracellular space onto the top of the channel.

(c) Single subunit of the HCN tetramer with transmembrane domain (gray) and C-terminal domain (white).

(d) Closeup of the transmembrane domain of one subunit with transmembrane helix S1–S3 (yellow), S4 (red), S5 and S6 (green) and the pore helix and filter region (both blue).

Figure 2.4: Model of the HCN4 channel. Transmembrane domain is a homology model based on PDB 3LNM. It was joined to the crystal structure of the HCN4 C-terminal region (PDB 3U11, see Figure 2.1).

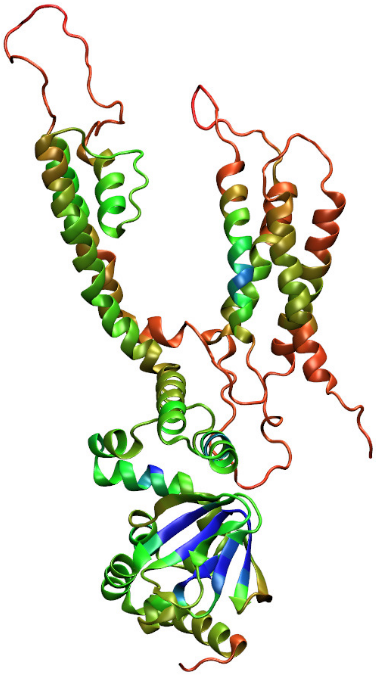


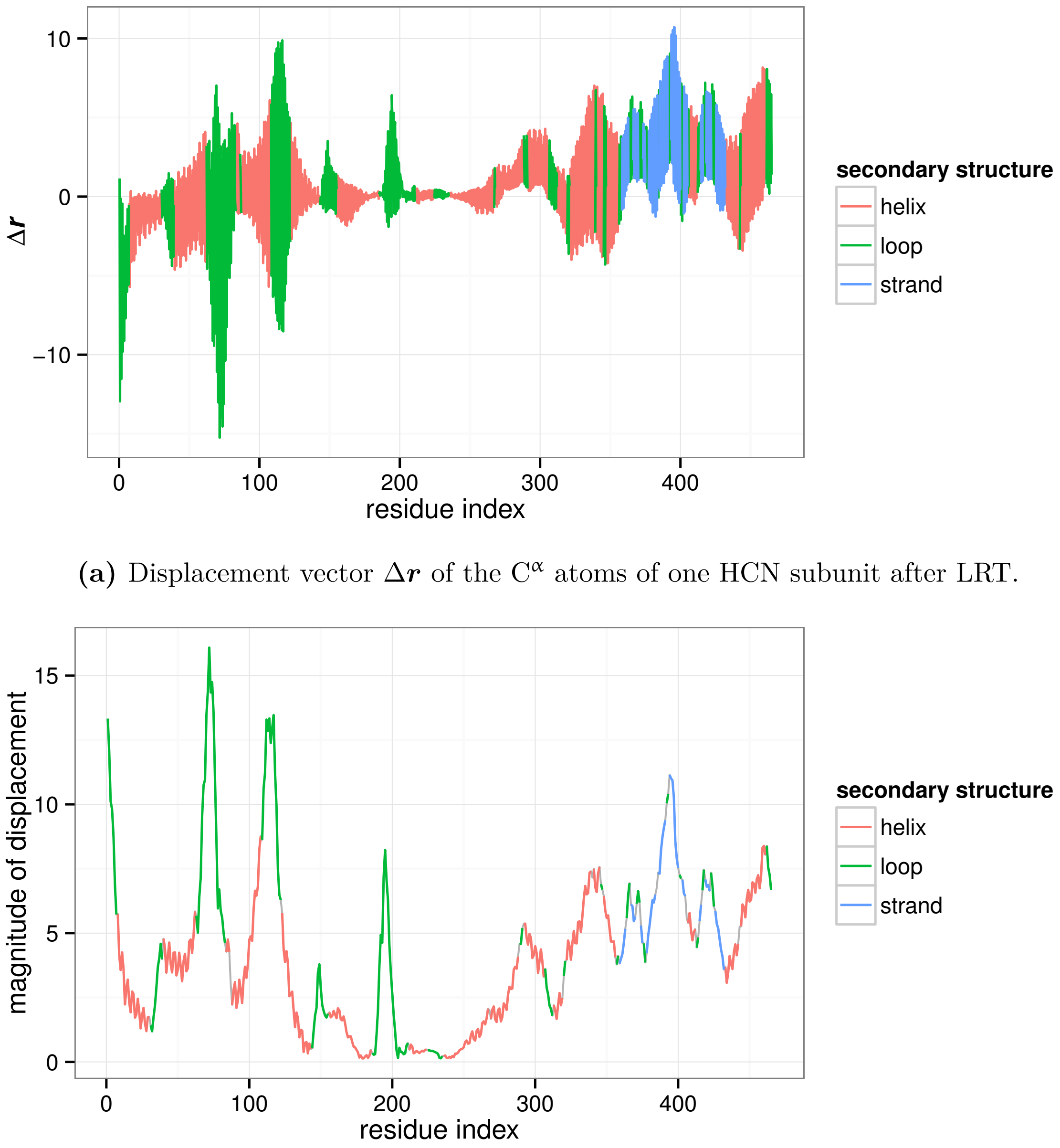
Figure 2.5: Error per residue of the HCN4 channel (joined model) visualized as color scale ranging from blue (error < 1.0 Å) via green to red (error > 3.5 Å).

To test whether this cooperativity influences the reaction of the channel in the ligand dissociation model, LRT was performed on only one as well as on all four binding pockets. The displacement of nodes in the subunits without force application was smaller than that of the forced subunit but the general reaction of the channel remained the same regardless of the number of forced subunits. This observation is in line with a study by Benndorf et al. [20] that showed that the binding of the fourth cAMP molecule further contributes to the conformational change.

2.3.3 Conformational Change Upon Ligand Dissociation

Figure 2.6 shows the displacement of one subunit that occurs upon force application. Due to the fourfold symmetry of the tetramer, the displacement for the other three subunits is analogous to that of the first. The transmembrane domain ranges from residues 1 to 267 followed by the C-terminal domain (residues 268 to 465). Note that LRT is not suited for quantitative assessment of protein motion. Hence, the overall trend of the movement and the displacement of protein parts relative to each other can be interpreted, while the total magnitude cannot be evaluated, i.e., the scale of the y-axis cannot be used.

Large loop regions in the transmembrane domain are the most flexible parts of the protein, which explains their large displacement upon LRT. The strong fluctuation of the plotted line in Figure 2.6a is due to the fact that x, y and z coordinates of every residue are plotted. We can see that the transmembrane domain has a tendency toward negative values, whereas the C-terminal domain is drifting more into the positive sector suggesting movement into opposite directions. The pore and filter region with the conserved CIGYG motif (residues 225 to 229) undergoes almost no displacement. This



(a) Displacement vector $\Delta \boldsymbol{r}$ of the C^α atoms of one HCN subunit after LRT.

(b) Magnitude of displacement of the C^α atoms of one HCN subunit after LRT.

Figure 2.6: Displacement of the C^α atoms of one HCN subunit after LRT. Note that the scale of the vertical axis is given in arbitrary units, since LRT is not suited for quantitative assessment.

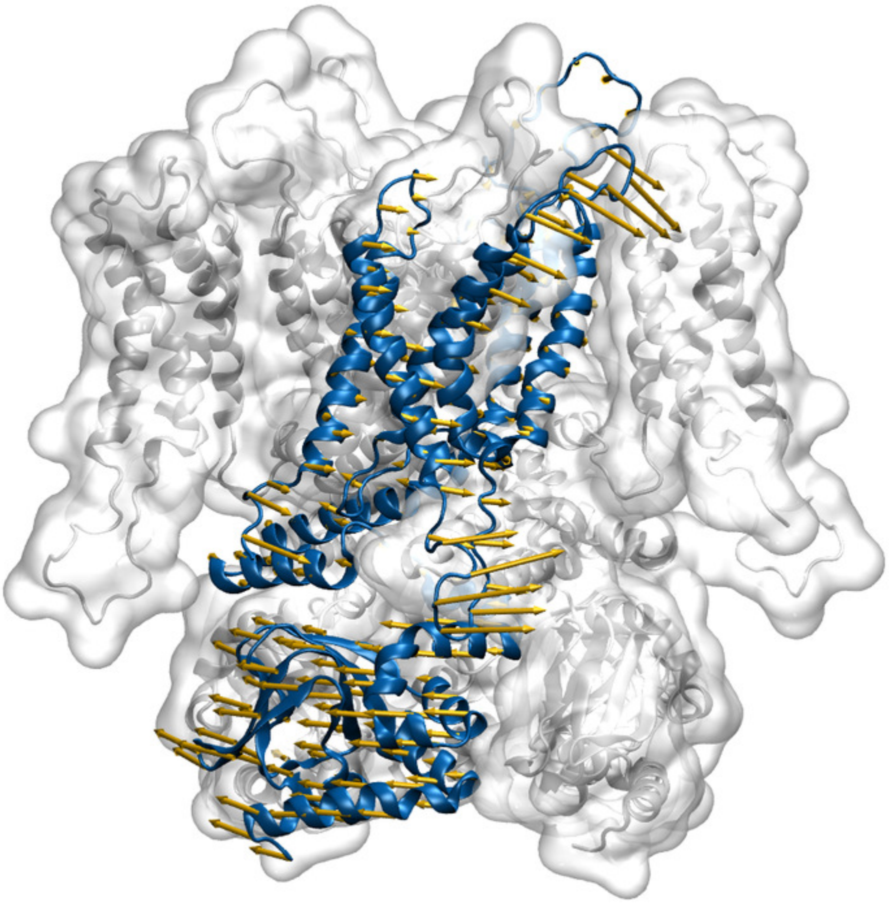


Figure 2.7: HCN4 joined model tetramer, one chain is shown in blue as cartoon representation. Arrows represent the displacement after LRT force application to the cAMP binding pocket residues of all four subunits.

becomes more obvious in Figure 2.6b, where the magnitude of displacement for each residue is visualized. However, directional information is lost here.

The overall motion of the HCN4 protein is shown in Figure 2.7: arrows indicate the direction and their length the magnitude of displacement after LRT. Again, the pictured arrows are only to be interpreted relative to each other, as the LRT method is not quantitatively predictive. Here, we can see that application of force to the six residues of the binding pocket leads to a rotation of the transmembrane domain and the CNBD against each other. The conformation of the outer region of the pore (pore helix and filter region) is maintained during LRT.

Such a torsion of the transmembrane domain against the non-membrane region has been observed for several other channels: the related bacterial MlotiK1 [90], the mechanosensitive MscL channel of *Mycobacterium tuberculosis* [139] and the nicotinic acetylcholine receptor [141].

Furthermore, the opening mechanism of several potassium channels has been shown to be a quaternary twist [67, 134]. Alam et al. [2] were able to create a high resolution

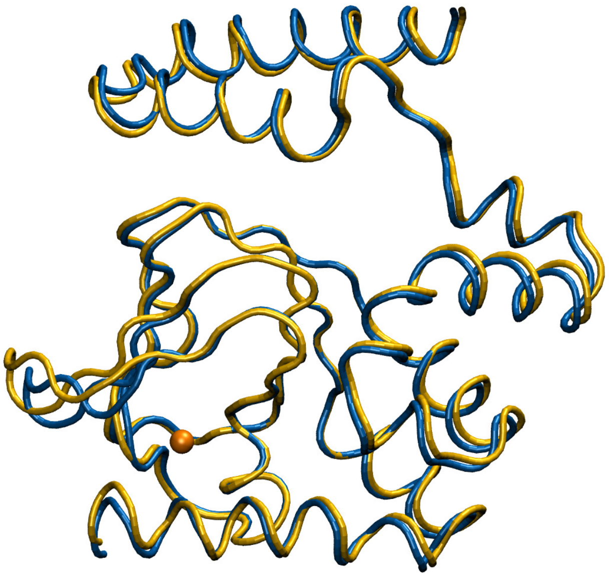


Figure 2.8: HCN4 C-terminal domain before (blue) and after LRT (yellow). The center of the cAMP molecule is represented by an orange sphere. Structural fit was performed over β strands 1, 3 and 8 (see Figure 2.1b)

crystal structure for NaK in the open state and reported a torsion for this sodium and potassium conducting channel compared to the closed form crystallized earlier [133]. They also discovered that the conformation of the filter region of the pore is almost the same in both states, which is in line with observations for the potassium channels KcsA (crystal structure in the closed state [175]) and MthK (crystal structure in the open state [82]). Thus, our observation of a quaternary torsion agrees with previous findings in other ion channels.

Figure 2.8 shows the HCN4 C-terminal domain. A superposition of the two structures before and after LRT was performed using β strands 1, 3 and 8. The largest displacement can be seen in the small loop between β strands 4 and 5. In addition, the helices of the C-linker and the B helix of the CNBD are shifted. The regions where deviations from the crystal structure conformation occur are mostly located toward the interface, which means that inter-subunit contacts in the C-terminal domain are influenced by cAMP dissociation (and binding).

The C helix on the other hand hardly seems affected by the application of force. There have been diverging speculations about the behavior of the C helix during ligand binding. The crystal structure of MlotiK1 (Figure 2.2) clearly shows a hinge-bending motion of the helix away from the rest of the protein when cAMP is released [130]. Since the two proteins and especially their CNBDs are closely related, one would expect a similar mechanism for HCN.

On the other hand, Taraska et al. [144] performed experiments on HCN2 which suggested a disintegration of the C helix upon ligand release without much of a relocation. Their hypothesis is based on a stabilizing effect of cAMP on the C helix [22]. A molecular

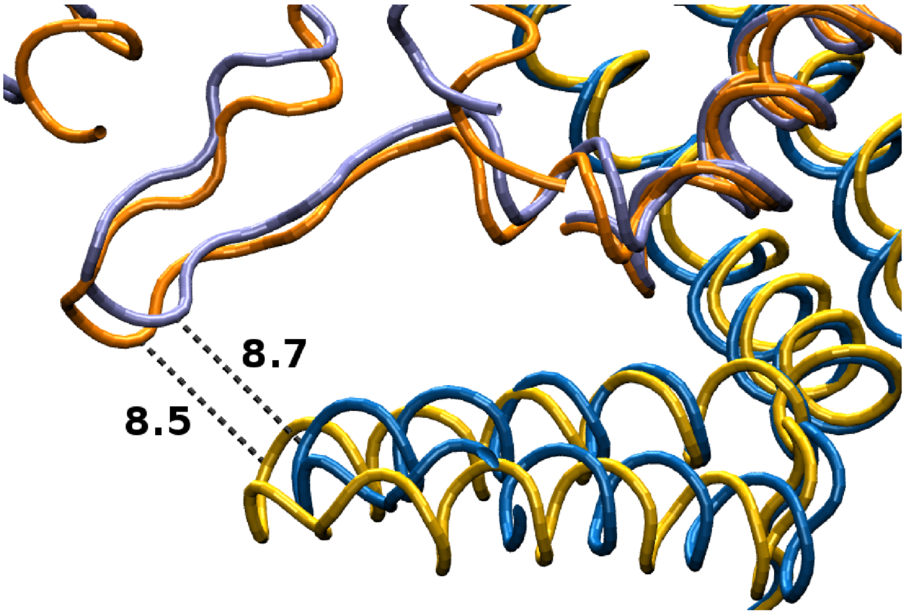


Figure 2.9: Closeup of HCN4 C-linker of one subunit (blue and yellow) and the S4–S5-linker of the neighboring subunit (slate blue and orange). The distance of the subunits is shown before (blue and slate blue: 8.7 Å) and after LRT (yellow and orange: 8.5 Å). Note that the difference is only to be regarded qualitatively since LRT is not a quantitative method, i.e., the distance becomes smaller but we cannot say by how much.

dynamics simulation of the HCN2 C-terminal domain from which cAMP was removed was carried out by Zhou and coworkers [174]. It did not show rigid body displacement of the C helix either.

Nevertheless, it is still possible that the missing hinge-bending motion in our LRT model is an artifact of the elastic network. Interactions between the C helix and the opposite side of the binding pocket, that exist in our model due to the simple contact definition by distance cutoff, are not necessarily realistic. A more sophisticated model might be able to better simulate this behavior. However, as long as it is not clear where the C helix is located in the ligand-free state such a model would be based on too many assumptions and therefore not be reliable.

Figure 2.9 shows a part of the HCN4 protein before and after LRT. The S4–S5-linker of one subunit is in proximity to the C-linker of the neighboring subunit. When the release of cAMP is simulated by force application to the binding pocket, the torsion of the transmembrane and C-terminal domains moves the S4–S5- and the C-linker toward each other and enables closer interaction of these two structural elements. Hence, cAMP removal brings the voltage-sensor coupling element closer to the inhibiting CNBD. Kwan et al. could show that the S4–S5- and the C-linker move relative to each other during gating [94]. Our results suggest that this process is also involved in the channel's reaction to cAMP binding.

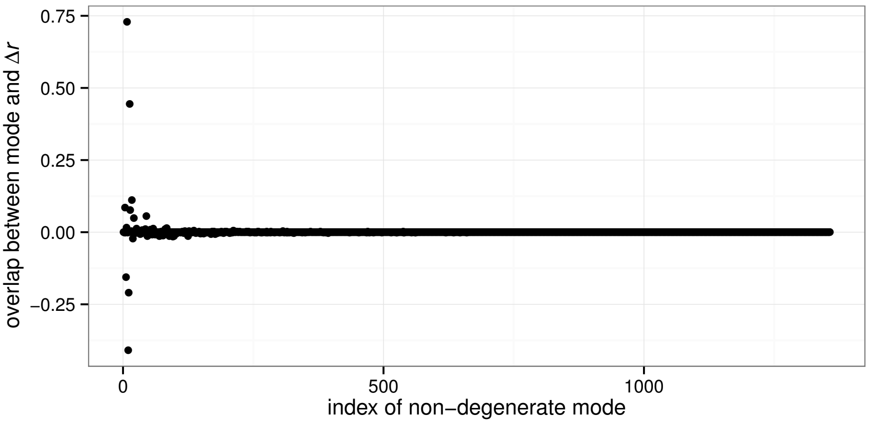


Figure 2.10: Overlap of 1357 non-degenerate modes with $\Delta \boldsymbol{r}$ from LRT. Modes are sorted according to their corresponding eigenvalues starting with the lowest at index 1.

While it has not been revealed yet, exactly how cAMP modulates channel gating, we do know that a complete removal of the CNBD has an effect similar to cAMP binding: the channel already opens at an earlier stage of hyperpolarization (see Section 2.1), i.e., the CNBD in the ligand-free state inhibits channel opening [156].

The S4–S5-linker transmits the reaction of the voltage sensor to the pore-forming parts of the channel. As described in Section 2.1 studies have suggested that an interaction of the C-linker with the S4–S5-linker couples voltage-gating and allosteric modulation of the channel by cAMP.

If cAMP removal brings the C-linkers closer to the S4–S5-linkers of their neighboring subunits, this in turn means that cAMP binding moves them further away thereby preventing interactions. This mechanism could be part of the reason why cAMP binding revokes the inhibitory influence of the CNBD.

2.3.4 Comparison of Δr with Low Frequency Modes

A singular value decomposition was performed on the Hessian matrix to obtain its eigenvalues and -vectors. The overlap (see Equation 2.5) between the non-degenerate eigenvectors and the displacement vector $\Delta \boldsymbol{r}$ from LRT was computed and is shown in Figure 2.10. If the direction of two vectors is identical, their overlap is 1; if they are orthogonal, it is zero.

Since the eigenvectors are sorted from low to high according to their eigenvalues, the low frequency modes—those with low corresponding eigenvalues, also called soft modes—are on the left side of the plot. The highest overlap of a single mode (no. 8 according to the index in Figure 2.10) with $\Delta \boldsymbol{r}$ was 0.73 and is shown in Figure 2.11a. Note that it is irrelevant whether the overlap is positive or negative, since the modes are fluctuations around an equilibrium state [9].

Some movements of proteins are composed of several superimposed modes [41]. To find out whether this is the case for cAMP dissociation in our model, linear combinations of the eigenvectors that featured the highest overlap with $\Delta\boldsymbol{r}$ were computed as described in Section 2.2.4.

Figure 2.11b shows the combination of two eigenvectors (no. 8 and no. 13), which increased the overlap to 0.85. The overlap between $\Delta\boldsymbol{r}$ and the linear combination of three eigenvectors (no. 8, no. 13 and no. 10) was 0.95. It is illustrated in Figure 2.11c. Adding a fourth eigenvector only slightly increases the overlap to 0.97. Thus, three eigenvectors are sufficient to describe the displacement upon force application to the cAMP binding pocket almost perfectly.

Congruence of the eigenmodes (especially the linear combination of three modes) and $\Delta\boldsymbol{r}$ is best in the C-terminal domain and the helical regions of the transmembrane domain. The largest divergence is created by the loops in the transmembrane domain. As mentioned in the evaluation of the homology model, modeled loop regions are most likely to differ from the native state of the protein, which might be the reason for this discrepancy. Nevertheless, the large overlap when using only three eigenvectors shows that the allosteric conformational change we discovered using LRT is an intrinsic property of the HCN4 channel.

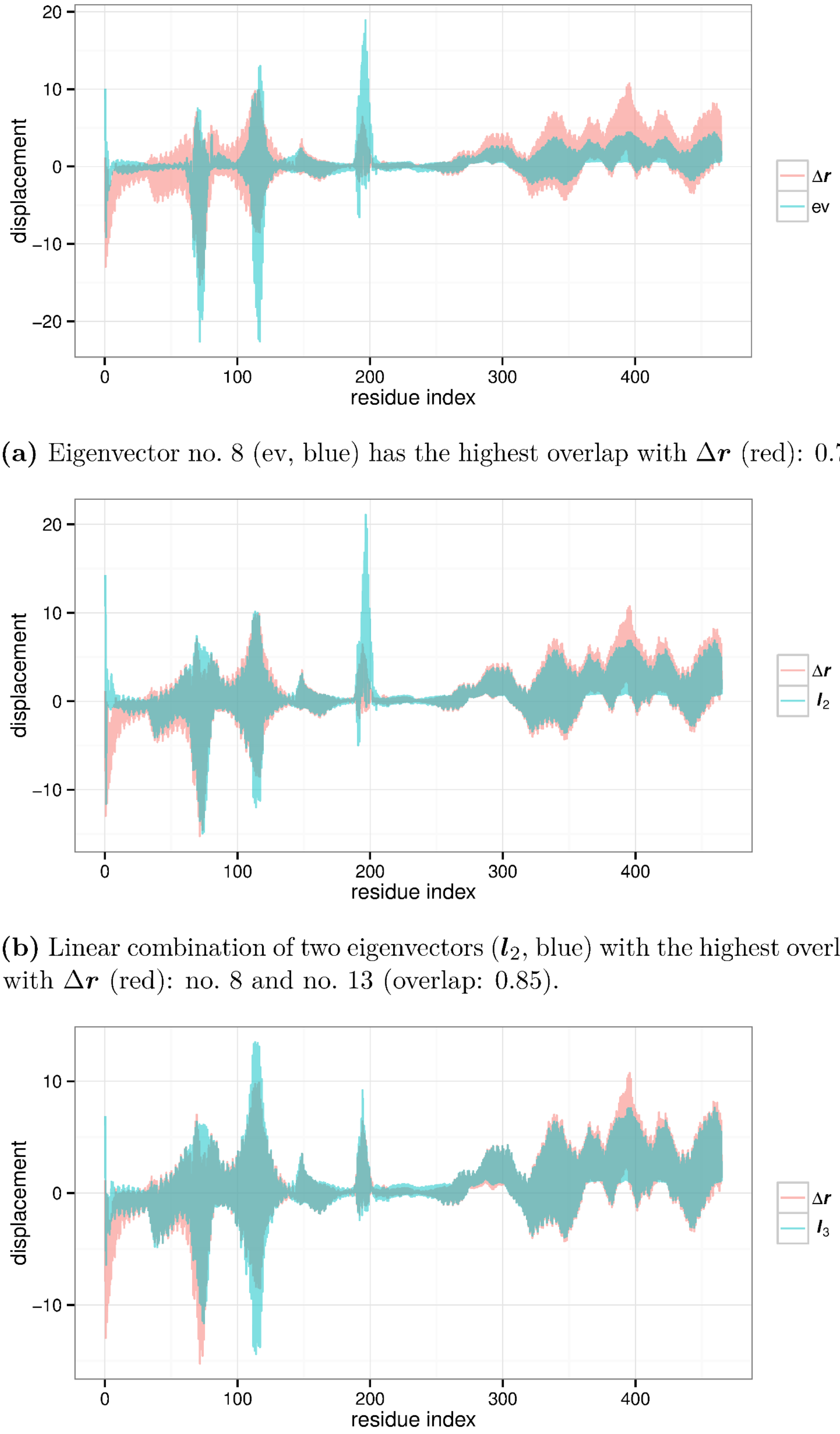
2.3.5 Clustering of Residue Groups Moving in Concert

To identify groups of residues that move together during the conformational change following cAMP release, we used the difference of C^{α} atom distances in an HCN subunit before and after LRT force application (described in Section 2.2.5) as a basis for hierarchic clustering. Thereby, C^{α} atoms whose distance stayed the same during LRT were clustered together. Figure 2.12 shows the HCN channel transmembrane region (a) and C-terminal domain (b) colored by membership in the twelve groups that emerged after cutting the tree at a height of 60.

An intriguing pattern emerges: the largest group (colored in red) comprises the S6 helix and the filter region as well as the C helix, an area in the β sheet and parts of the C-linker. The rest of the CNBD clusters together with the S1 and S3 helix (gray) with the exception of one residue (lime green) in the cAMP binding pocket that forms a group of its own.

The residues of the A′ helix of the C-linker are members in six different groups. This is due to the large increase of displacement between residues 270 and 290 that stretches the A′ helix (cf. Figure 2.6b). The D′ helix (residues 323 to 331, colored in blue) is in a group separate from the rest of the protein but moves as one element. Both events probably result from interactions with the neighboring subunits that cause distortion of the helices.

The S4 voltage sensor helix belongs to two different groups (yellow and orange). Its composition mostly resembles the S5 helix. Since the direct linker loop between the two has the function of passing information about a change in voltage from S4 to S5, a similar pattern of motion is not surprising.

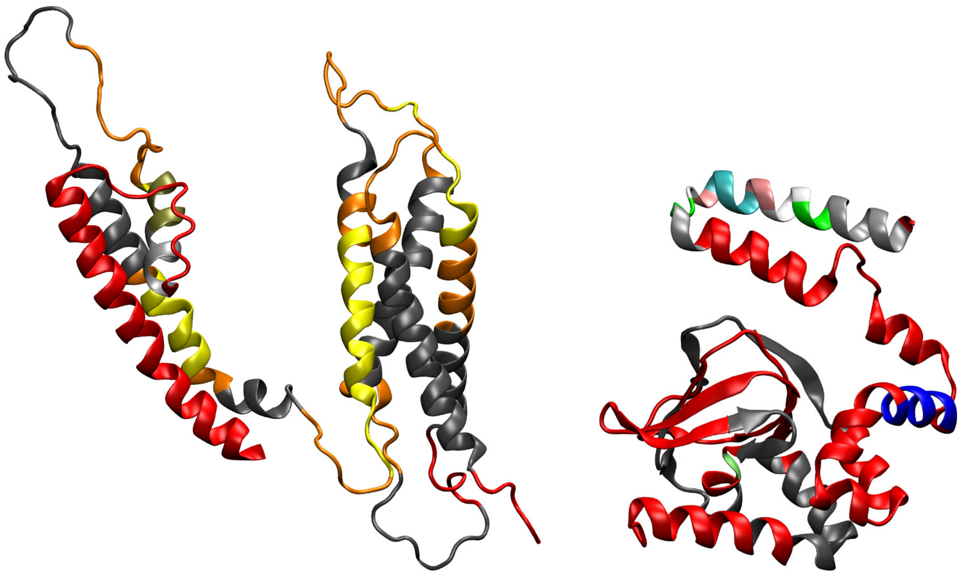


(a) Eigenvector no. 8 (ev, blue) has the highest overlap with $\Delta\boldsymbol{r}$ (red): 0.73.

(b) Linear combination of two eigenvectors ($\boldsymbol{l}_2$, blue) with the highest overlap with $\Delta\boldsymbol{r}$ (red): no. 8 and no. 13 (overlap: 0.85).

(c) Linear combination of three eigenvectors ($\boldsymbol{l}_3$, blue) with the highest overlap with $\Delta\boldsymbol{r}$ (red): no. 8, no. 13 and no. 10 (overlap: 0.95).

Figure 2.11: $\Delta\boldsymbol{r}$ from LRT compared to eigenvectors derived from a singular value decomposition of the Hessian matrix of the system. Plotted vectors were scaled to the same magnitude for better comparability.



(a) Transmembrane region.

(b) C-Linker and CNBD.

Figure 2.12: HCN channel residues colored by groups that move together.

2.3.6 Identifying Key Residues in Channel Modulation

To investigate which contacts in the protein are key players in the conformational change upon cAMP release, a switch-off screening was performed. All non-covalent contacts were switched off individually, i.e., the interaction strength for the respective amino acid pair was set to zero before applying the force that simulates ligand binding. To maintain the symmetry of the tetramer, equivalent contacts were switched off in all subunits at once.

3270 switch-offs were performed and force was applied to open the cAMP binding pocket. The impact was analyzed by comparing the displacement vector of each altered system with one switched off contact to the original $\Delta \boldsymbol{r}$. As one would expect, most switch-offs hardly changed the reaction of the protein to force application. This was either due to other contacts compensating the loss or the contact not being stressed during the conformational change. Equally irrelevant for our analysis were those contacts that caused a change of displacement which was restricted to the close proximity of the broken interaction.

Since we wanted to detect contacts important for global conformational rearrangement, we focused on switch-offs that induced a significant change of displacement (> 0.1) for at least 15 residues of each subunit. This was the case for 21 contacts. The change in magnitude of displacement after LRT they caused is illustrated in Figure 2.13 and their location in the protein is indicated in Figure 2.14.

The upper six panels in Figure 2.13 stabilize the four-helix bundle S1–S4. This is the reason why their interruption induces the largest change in the N-terminal part of the subunit. The contacts 142-154 and 142-155 connect the four-helix bundle with the inner region of the channel via the S4–S5-linker. Switching off one of them (especially the latter) changes the displacement of the whole channel by influencing the torsion of the

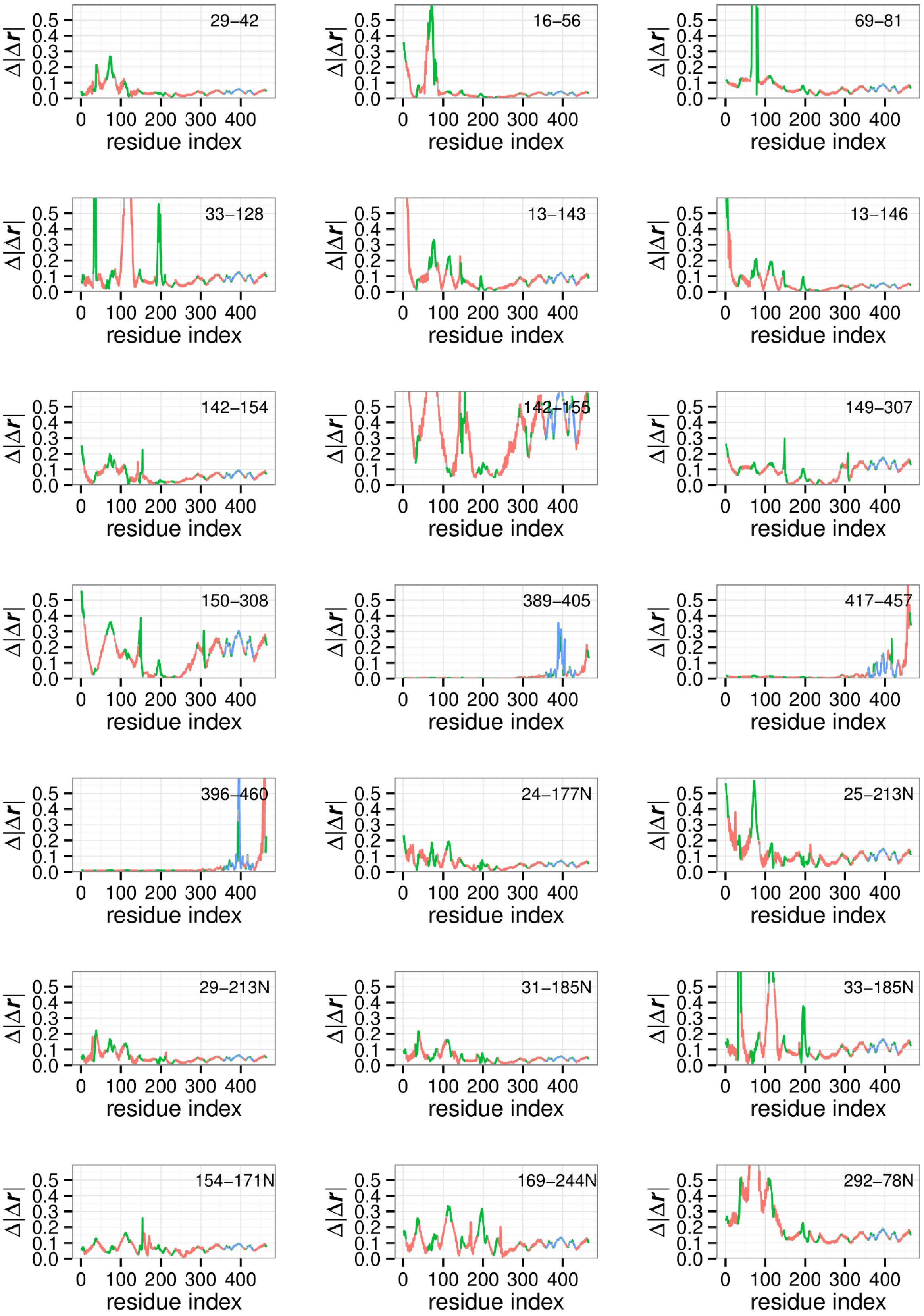
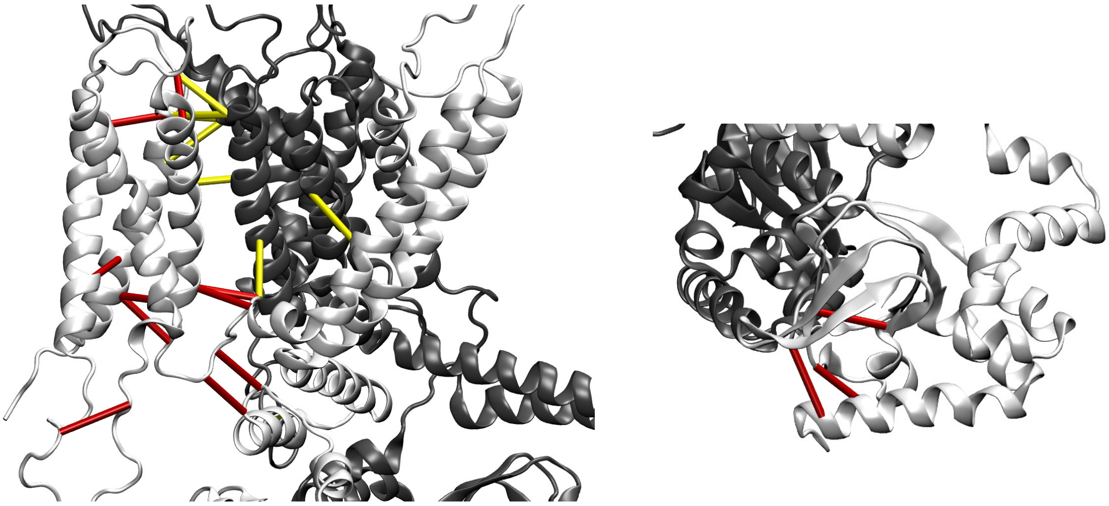


Figure 2.13: Change of the magnitude of displacement upon force application caused by switching off a contact and its counterparts in all four subunits. The switched off contact is given in the upper right corner of each plot. Residue numbers marked with “N” belong to neighboring subunits in an inter-subunit contact.



(a) Transmembrane region and part of the C-linker.

(b) C-terminal domain with part of the C-linker and CNBD.

Figure 2.14: Two neighboring subunits of HCN4 (white and gray). The 21 contacts that cause significant change of displacement when switched off are indicated by cylinders. Intra-subunit contacts are drawn in red, inter-subunit contacts in yellow.

transmembrane and C-terminal domains. The same holds for the contacts 149-307 and 150-308, which connect the C-linker with the transmembrane region within a subunit. The following three panels (389-405, 417-457 and 396-460) are interactions across the cAMP binding pocket. When they are broken, the direct force application in this area causes the C helix or the loop connecting β strands 4 and 5 or both to move farther away from the center of the ligand and thereby the binding pocket.

Eight of the switch-offs plotted in Figure 2.13 involve inter-subunit contacts. Most of them only concern the displacement of the transmembrane domain, which is due to their location: none of the inter-subunit interactions of the C-terminal domain appears in the set of the most influential contacts.

In Figure 2.14 all interactions from Figure 2.13 are pictured as cylinders with red indicating intra-subunit and yellow representing inter-subunit contacts. The noticeable common feature is that all but one of the interactions reach across secondary structure elements. The exception is interaction 69-81: both of these residues lie in the same loop but on opposite sites, thereby influencing the adjoining helices. This shows that interactions between the transmembrane helices play an important role in the allosteric reaction of the channel to cAMP binding and dissociation. Both intra- and inter-subunit contacts are involved in this process.

2.4 Conclusion

We built a homology model of the HCN4 channel transmembrane region and were able to connect it to the crystal structure of the C-terminal domain. The joined model was used to study allosteric conformational change associated with cAMP release. To this end, an elastic network model of the HCN tetramer was constructed. cAMP dissociation was simulated by a force that opened the binding pocket. The resulting conformational change of the HCN tetramer was compared to low frequency modes of the ENM. In addition, we conducted a switch-off screening to identify key residues in the process. Up to now, no such detailed computational analysis of cAMP modulation in HCN has been published.

cAMP binding removes the inhibitory effect of the CNBD and thereby reduces the hyperpolarization threshold that needs to be reached for channel gating. Our results suggest that the quaternary twist, which has been shown to be the opening mechanism for several ion channels, is already part of the allosteric reaction of the channel upon cAMP binding.

We could also show that interaction between the S4–S5-linker in the transmembrane domain and the C-linker is influenced by the allosteric rearrangement. This might be part of the mechanism of how cAMP modulates channel behavior. As the group around Sanguinetti found out, the S4–S5-linker participates in channel gating [36]. They also presented mutational studies which suggested an interaction between the S4–S5-linker and the C-linker [43]. Our results point in a similar direction: The S4–S5-linker seems to be involved not only in channel gating but also in cAMP modulation.

The search for key players in cAMP-induced allosteric conformational change revealed that the most important contacts are those between the helices of the transmembrane domain. Intra- and inter-subunit contacts are relevant in this process.

Since HCN1, HCN2 and HCN4 are very similar in sequence and structure and the methods applied in this study are coarse-grained and insensitive against sequence variation, the insight gained for HCN4 most likely holds for HCN1 and HCN2 as well.

3 Intramolecular Coevolution of HCN Channels

This chapter contains a thorough coevolutionary analysis of the hyperpolarization-activated cyclic nucleotide-gated (HCN) channel by means of several new approaches. In Section 3.2 the influence of parameters during multiple sequence alignment (MSA) construction is analyzed. The section describes the findings of our study “Visual exploration of parameter influence on phylogenetic trees” [74]. Section 3.3 contains a graph theoretical approach to study coevolution in five protein families. Section 3.4 takes a closer look at individual pairwise intramolecular interactions. In Section 3.5 interactions of higher order are examined; the study is based on our publications “Information-theoretic analysis of molecular (co) evolution using graphics processing units” [155] and “Using graphics processing units to investigate molecular coevolution” [154].

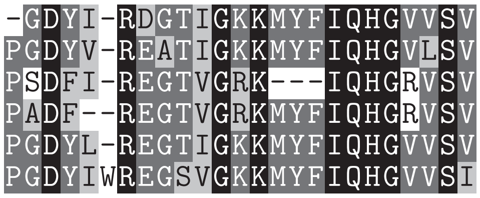
3.1 Introduction

A protein’s structure and function are determined by the chemical properties of the amino acids of which it is composed. Random mutations in the DNA can cause changes in the amino acid composition of a protein [87]. If this change negatively affects the function of the protein in any way, a selective pressure to compensate for this exchange is imposed on all residues interacting with that particular amino acid. Since these compensating mutations provide a selective advantage, they are more likely to survive. This ability to compensate a substitution of their interaction partner induces a covariation of these residues [57, 89]. They are said to be *coevolving*.

Identifying coevolving residues in a protein based on its sequence information is part of promising new approaches including prediction of tertiary structure [84], detection of molecular interactions [42], analysis of binding site specificity [61], search for functional units [63, 166] and catalytic residues [31]. Hence, molecular coevolution of amino acids has been subject to numerous studies in the past years [38, 100].

3.1.1 Multiple Sequence Alignments

When studying intramolecular coevolution in proteins, it is necessary to examine related sequences for changes that occurred during evolution. Usually, these sequences belong to the same or very closely related proteins from different organisms and individuals. First, equivalent regions in the sequences need to be identified and aligned. The result is called a multiple sequence alignment (MSA): Each row in the MSA contains one sequence. Gap



```
-GDYI-RDGTIGKKMYFIQHGVVSV
PGDYV-REATIGKKMYFIQHGVLSV
PSDFI-REGTVGRK---IQHGRVSV
PADF--REGTVGRKMYFIQHGRVSV
PGDYL-REGTIGKKMYFIQHGVVSV
PGDYIWREGSVGKKMYFIQHGVVSI
```

Figure 3.1: Example of a multiple sequence alignment. Structurally corresponding amino acids are written in the same column using gap characters to induce the required shifts.

characters, which represent insertion and deletion mutations (“indels”), are inserted to arrange matching sequence parts underneath each other. A certain level of similarity is required to facilitate an appropriate mapping. A short example MSA is shown in Figure 3.1.

Unfortunately, the computation of a perfect alignment is an NP-complete problem [158]. Since the sequence number in a typical data set is usually too high to compute the exact MSA, heuristics are employed to find an approximate solution. Many computer programs, using a variety of algorithms, have been created to improve automated MSA construction [86]. A common procedure is to use a substitution matrix to evaluate matched amino acid combinations and a penalty to account for the insertion of gaps. Several programs calculate a measure of similarity first and start to combine pairwise alignments beginning with the most similar sequences, e.g., `ClustalW` [95]. Some methods revise the originally created alignment by iterating processing steps (`MUSCLE` [52]), while others take the phylogeny of the sequences into account (`PRANK` [102]). Recently, approaches that utilize hidden Markov models have emerged (`Clustal Omega` [135]).

Most programs offer the possibility to adjust parameter settings such as the type of substitution matrix or the gap penalty. To find out what effects these parameter settings have on the MSA, we used a visual analytics approach, which is described in Section 3.2.

3.1.2 Mutual Information

A well-established approach to detect molecular coevolution is calculating the mutual information (MI) of pairs of columns in a protein multiple sequence alignment [7,68,89]. It is a non-parametric method, which means that—in contrast to parametric methods—the only required input is the MSA; no further assumptions are necessary [39, 165].

MI is an information-theoretical measure based on Shannon’s entropy. In 1948, Shannon defined the entropy H to measure the uncertainty of a random variable X. It is the expectation value of the information content [131]. The information content I of a symbol x_i as a realization of variable X is defined as

$$I(x_i) = \log_2\left(\frac{1}{p(x_i)}\right) = -\log_2\Big(p(x_i)\Big) \tag{3.1}$$

with $p(x_i)$ being the probability of occurrence of symbol x_i. The expectation value of variable X with the possible realizations x_i is the Shannon entropy:

$$H(X) = \sum_i p(x_i)\, I(x_i) = -\sum_i p(x_i) \log_2\Big(p(x_i)\Big). \tag{3.2}$$

It becomes maximal if all realizations occur with the same frequency, i.e., the probabilities of occurrence are uniformly distributed.

The MI M quantifies the amount of information a random variable X provides about another variable Y [105].

$$M(X,Y) = \sum_{i,j} p(x_i, y_j) \log_2 \left(\frac{p(x_i, y_j)}{p(x_i) p(y_j)} \right) \tag{3.3}$$

with $p(x_i)$ and $p(y_j)$ being the marginal probabilities of occurrence of amino acid symbols x_i and y_j in the MSA columns X and Y, respectively, and $p(x_i, y_j)$ their joint probability. In this sense the MI is the Kullback–Leibler divergence of the product of the marginal distributions $P(X)$ and $P(Y)$ from the joint distribution $P(X,Y)$ [91]. Since the joint probability of two independent random variables is simply the product of their marginal probabilities, this divergence—and thereby the MI—becomes zero if X and Y are independent of each other.

MI can also be written as the difference of the sum of the one-point entropies and the joint entropy:

$$M(X,Y) = H(X) + H(Y) - H(X,Y). \tag{3.4}$$

To measure coevolution in a protein, the columns of an MSA are considered to be the random variables. The 21 realizations are the 20 proteinogenic amino acid symbols in addition to the gap character "-". If the two residues evolve independently of each other, their MI equals zero. High MI values indicate coevolution of X and Y.

The highest possible MI value depends on the one-point entropies of the two MSA columns under investigation. The diversity of a combined pair of the columns X and Y is at least as high as that of the more diverse column, therefore

$$H(X,Y) \geq \max\Big(H(X), H(Y)\Big). \tag{3.5}$$

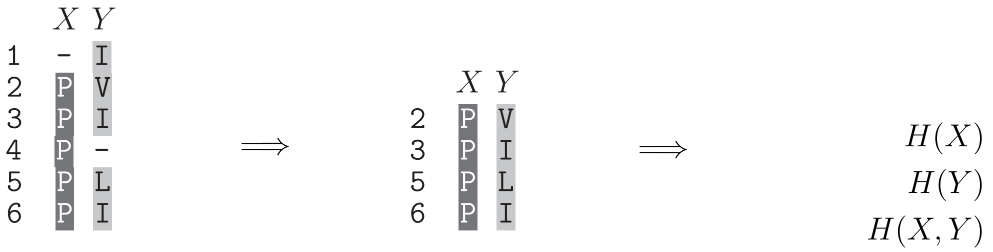
If this relationship is inserted into Equation 3.4, it follows that the maximally possible value of MI is

$$M(X,Y) \leq \min\Big(H(X), H(Y)\Big). \tag{3.6}$$

As mentioned above, the entropy becomes maximal if the occurrence probabilities of all realizations are uniformly distributed. If this is the case for both X and Y and the covariation between the two is absolute, the MI reaches its highest possible value. For a uniform distribution the frequencies for all amino acids become $^1/_{21}$. By inserting this into a combination of Equation 3.2 and Equation 3.6 we get

$$M(X,Y) \leq -21 \times \frac{1}{21} \times \log_2 \frac{1}{21} \approx 4.39. \tag{3.7}$$

Therefore, the MI of positions in a protein MSA lies within the range of $[0, 4.39]$, but strongly depends on the one-point entropies.



(a) MSA columns. **(b)** Gap-free subset. **(c)** MI computation.

Figure 3.2: Illustration of SUMI. A gap-free subset of the column pair is created. Only rows which carry a non-gap character in both columns are chosen to form the subset. $H(X)$, $H(Y)$ and $H(X,Y)$ are computed for the subset and SUMI is derived via Equation 3.4.

Handling Gaps

The most common strategy is to simply regard a gap as an additional symbol in the alphabet. However, there has been much debate about the validity of this approach. Some studies completely exclude MSA columns containing gaps from their analyses [63], while others introduce a cutoff value for the gap content and use only columns with fewer gaps [30]. Unfortunately, procedures like this severely limit coevolutionary analysis. Especially in MSAs of more distantly related sequences many columns contain gaps. It is therefore desirable to find an alternative method to treat gaps.

A promising approach is to sample a gap free subset for each column pairing and compute the MI for this subset only [77,166]. For the computation of this subset mutual information (SUMI), all rows that do not contain a gap in either column of the pair (X,Y) under investigation are chosen. This sampling is shown in Figure 3.2. If the resulting subset contains less than three symbol pairs, the SUMI for the respective column pair is set to zero. SUMI is a compromise between the naive approach of treating gaps the same as any another character and the radical solution of discarding the whole column as soon as one gap is present.

Normalization

Usually, computed MI values are normalized prior to analyzing the results. In the past years, several studies have attempted to account for small sample size, eliminate correlations that are attributed to phylogenetic relationship of the analyzed sequences and bias from overrepresentation of often-sequenced species [28,50,65]. In this thesis, two normalization methods are applied. Z-scores of the MI values are calculated to check statistical significance. The minimal entropy normalization is used to account for limited one-point entropies in MSA columns due to high levels of conservation.

***Z*-score** Statistical significance of MI results can be verified via a null model: The order of amino acid symbols in each MSA column is shuffled to break covariation with other columns but maintain the frequencies of the symbols and thereby the marginal probabilities [163,167]. For each shuffled alignment the MI of all column pairs is computed. After repeating this process many times, the mean $\langle M(X,Y)\rangle$ and standard deviation $\sigma(M(X,Y))$ of the MI over all these shuffled MSAs can be used to calculate a Z-score:

$$Z(X,Y) = \frac{M(X,Y) - \langle M(X,Y)\rangle}{\sigma\Big(M(X,Y)\Big)}. \tag{3.8}$$

Z is a statistical measure that gives the distance of a value from the mean of the distribution in units of one standard deviation. If the MI value of an amino acid pair is considerably higher than the mean value for this same pair derived from the shuffling runs, it is considered statistically significant. Based on other publications $Z > 4$ is chosen as a cutoff value [111].

Minimal Entropy Normalization Many positions in a protein are subject to selective constraints that restrict the space of sequence variation. In these positions only a few amino acid types are non-deleterious replacements. This limits the set of symbols that can appear in the corresponding MSA columns and thereby reduces their entropy. As we know from Equation 3.6, the MI can never become greater than the lesser of the two one-point entropy values of the MSA columns under investigation. Therefore, even if two positions coevolve strongly, their MI value may be reduced by non-maximal one-point entropy of either one or both of the two positions. As a consequence, such an interaction may go unnoticed when screening for the top coevolving positions in a protein of interest.

To better investigate MSA positions with medium to low entropy, the column pair-specific upper bound of the MI can be used for normalization:

$$M_{\mathrm{minH}}(X,Y) = \frac{M(X,Y)}{\min\Big(H(X), H(Y)\Big)}. \tag{3.9}$$

$M_{\mathrm{minH}}(X,Y)$ equals one if the coevolution between the two alignment columns becomes maximal under the given entropy conditions. This approach has already been investigated by Martin et al. [111] and Vinh et al. [153]. We call this normalization method the minimal entropy normalization (minH).

Of course, for SUMI values the pairing-specific one-point entropies need to be used for normalization with the minH method (see Figure 3.2).

3.2 Analyzing the Influence of Parameters on Multiple Sequence Alignments

This section is based on the study “Visual exploration of parameter influence on phylogenetic trees” by Heß et al. [74].

Parameter	ClustalW2	MUSCLE
Clustering method	nj, upgma	nj, upgma
Weight matrix	BLOSUM, Gonnet, PAM	PAM
Distance measure 1	N/A	kbit20-3, kmer20-3, kmer6-6
Distance measure 2	N/A	pctidkimura, pctidlog
Endgaps	yes, no	N/A
Hydrophilic gaps	yes, no	N/A
Rooting method	N/A	midlongspan, minavgleafdist, pseudo
Profile score	N/A	log-expectation, sum-of-pairs
Gap open penalty	1, 2, 5, 10, 25, 50, 100	1, 2, 5, 10, 25, 50, 100
Gap extension penalty	0.05, 0.5, 1, 5, 10	0.05, 0.5, 1, 5, 10
Gap distance penalty	1, 5, 10	N/A

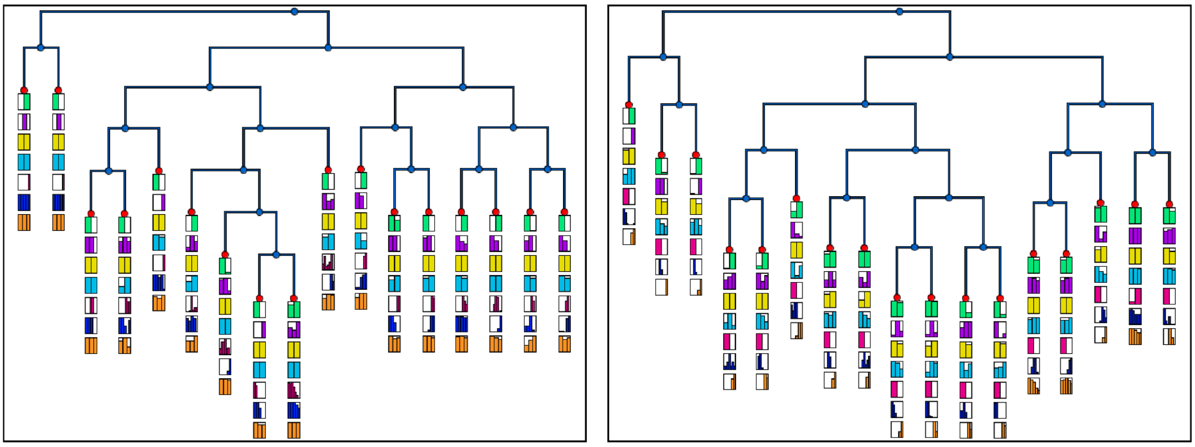
Table 3.1: MSA parameter settings used in this study. Abbreviations: nj (neighbor joining), upgma (unweighted pair group with arithmetic mean)

As mentioned above, MSAs can be constructed using several different algorithms and parameter settings. To find out how these settings influence the composition of the MSA, a large amount of parameter combinations needs to be examined. A visual analytics approach helped to master the challenge of evaluating a large data set of MSAs. The program written by Heß et al. allows for clustering and comparison of trees [74]. Therefore, phylogenetic trees were constructed based on the MSAs and analyzed. The software builds a super-hierarchy by clustering similar trees and gives an overview of their properties. The distributions of parameter settings within clusters is displayed and can be used to find parameters that are characteristic for the respective cluster. Thereby, it is possible to distinguish between parameters that strongly influence the composition of an MSA and those that only have minor impact. This knowledge will help to find a high quality MSA.

3.2.1 Methods

The data set of HCN channel protein sequences was prepared via a BLAST-search of the query sequence of HCN1 *Macaca fascicularis* with the GenBank [21] identifier 355749904 in the non-redundant (nr) database; an E-value cutoff of 0.00001 was applied. To further increase stringency only those sequences that were annotated with the words “hyperpolarization” and “cyclic” were considered. A first MSA was performed (ClustalW2, default parameters [95]) to identify highly dissimilar as well as duplicated sequences and consequently delete them from the final data set, which thereafter contained 211 HCN channel protein sequences.

Based on this data set, MSAs were constructed using two alignment programs: ClustalW2 and MUSCLE [52]. Different settings for various parameters were chosen; they are listed in Table 3.1. By combining all of these settings, 2520 trees were obtained



(a) Clustered superhierarchy of the `ClustalW2` data set. Colors of histograms: “clustering method” (green), “weight matrix” (purple), “endgaps” (yellow), “hydrophilic gaps” (light blue), “gap open penalty” (red), “gap extension penalty” (dark blue), “gap distance penalty” (orange).

(b) Clustered superhierarchy of the `MUSCLE` data set. Colors of histograms: “clustering method” (green), “distance measure 1” (purple), “distance measure 2” (yellow), “rooting method” (light blue), “profile score” (red), “gap open penalty” (dark blue), “gap extension penalty” (orange).

Figure 3.3: Screen shots from the tree visualization software used in the study by Heß et al. [74]. Each of the leafs contains a cluster of similar trees. Histograms show the distribution of MSA parameters in the respective cluster.

for each alignment algorithm. Afterwards, the trees were visually analyzed with the help of a tree comparison software.

3.2.2 Results and Discussion

The 2520 trees for each alignment algorithm were analyzed using the visual tree comparison program. We examined the distribution of parameters and identified characteristic properties of clusters in the super-hierarchy. The two setups showed different cluster-specific properties.

Both `ClustalW2` and `MUSCLE` construct a guide tree according to which the MSA is performed. Screen shots of the visualization software can be found in Figure 3.3. The parameter distributions are displayed in bar plot histograms for each cluster. For the `ClustalW2` super-hierarchy the “clustering method” parameter (green) was most characteristic for all clusters, i.e., the guide trees of almost all MSAs belonging to one cluster were constructed using the same clustering method. This was not the case for the `MUSCLE` setup where this parameter has only slight influence on the structure of the guide tree and thereby the MSA.

The second characteristic property of the `ClustalW2` setup was “weight matrix = Gonnet” (purple in Figure 3.3a) combined with “gap open penalty = 100” (red in Figure 3.3a). None of the other parameters showed any cluster-specific properties in this setup. Since

MUSCLE does not provide a choice of weight matrices, it is not possible to analyze this feature in the MUSCLE setup.

The most characteristic parameter for the MUSCLE super-hierarchy turned out to be "profile score" (red in Figure 3.3b), i.e., the scoring function, which can be set to sum-of-pairs or log-expectation. Sum-of-pairs uses a substitution matrix to compute the score of all pairwise alignments, just as ClustalW2 does [51]. The log-expectation scoring function deploys probabilities computed from VTML 240 [115].

It is very interesting to see that many parameters (e.g., all gap penalty parameters and the distance measure in MUSCLE) seem to have hardly any impact on the MSA structure, while others, such as the choice of clustering method and the scoring function, strongly affect the result.

Conclusion

This analysis helped to improve the construction of a high quality MSA for the HCN channel data set. We can now reduce the search space by dismissing some of the alignment parameters as irrelevant. But the results also show that there is no "optimal" parameter set. Every MSA has to be thoroughly scrutinized. Therefore, we manually curated the MSA we used in the following sections.

3.3 Coevolutionary Networks

Many structures in everyday life are networks: sets of objects that are connected by some kind of relationship [14]. Therefore, it is a natural idea to use graph theoretical approaches to analyze such processes. In the past years, network analysis has been applied in various research areas of natural and social science including interactions of molecules in biological cells [15], neural networks in the brain [136], seismic activity [1] and social networks [33].

Coevolutionary relationships between amino acids in proteins can be expressed as a network as well. However, only a few studies on this topic have been published so far. In 2009, Fatakia et al. used an MI network approach to study coevolution in G protein-coupled receptors [55]. Chakrabarti and Panchenko analyzed coevolution of functional sites in a large data set of proteins [35]. Weil applied graph theoretical analysis to identify binding sites of antibiotics in ribosomes [162].

Up to now, no detailed graph theoretical analysis of intramolecular coevolutionary relationships in a protein has been published to the knowledge of the author. Here, such an analysis is performed for the HCN channel as well as for four control proteins.

Graphs

A graph is made up of a set of objects ("vertices" or "nodes"), some of which are connected by "edges" representing pairwise relations. These edges can point from one vertex to the other (directed graph) or simply connect the vertices without sense of direction (undirected graph). Edges can also be given a weight representing some property of the relationship (weighted graph).

In our study, the vertices of the graph are residues in a protein, the edges represent their coevolutionary interactions. To this end, a measure of coevolution is required, e.g., the mutual information (MI) computed from the columns of an MSA. We construct undirected, unweighted graphs by drawing edges between nodes if the MI value of the two corresponding residues is larger than a certain threshold.

The MI graphs will then be subjected to detailed graph analysis and compared to two null models. Several graph measures will be computed to analyze the structure and properties of coevolutionary networks. In the following, these measures are described:

degree The degree of a vertex v_i $(i = 1, \ldots, N)$ is the number of edges that are connected to it [44].

shortest path A path is a route from one vertex v_i to another v_j along consecutive edges. The shortest path between two vertices in an unweighted, undirected graph is the one leading via the fewest edges [44]. Its length is measured by counting how many edges one passes on the path from v_i to v_j.

betweenness The betweenness B of a vertex v_i is the fraction of all shortest paths that run through it:

$$B(v_i) = \sum_{\substack{j=1 \\ j \neq i}}^{N-1} \sum_{\substack{k=j+1 \\ k \neq i}}^{N} \frac{\sigma_i(v_j, v_k)}{\sigma(v_j, v_k)}, \tag{3.10}$$

where $\sigma(v_j, v_k)$ is the total number of shortest paths from vertex j to vertex k and $\sigma_i(v_j, v_k)$ is the number of those paths that pass through v_i [58].

closeness The farness of a vertex v_i is the sum of the distances d to all other vertices v_j. Closeness C is the reciprocal of farness:

$$C(v_i) = \left\{ \sum_{j=1}^{N} d(v_i, v_j) \right\}^{-1}. \tag{3.11}$$

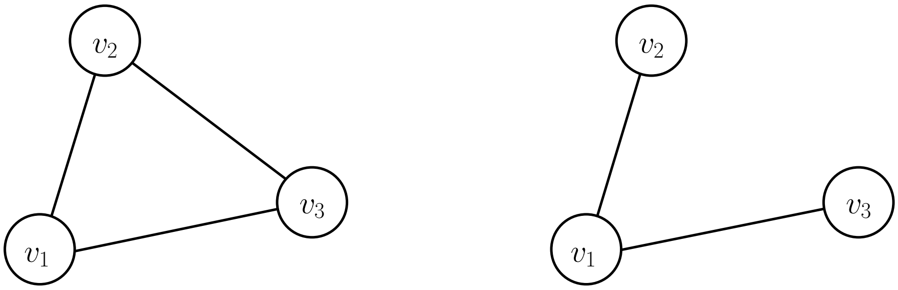
For unconnected vertices a distance of N is assumed, with N being the number of vertices in the network [59].

clustering coefficient The *local* clustering coefficient [160] is defined as

$$\mathcal{C}_{\text{loc}}(v_i) = \frac{\text{Number of triangles linked to node}\, i}{\text{Number of triples centered at node}\, i}. \tag{3.12}$$

A triple consists of three nodes which have at least two edges between them. If the third edge is present as well, the triple is additionally considered a triangle (see Figure 3.4). The *global* clustering coefficient [25] is calculated as follows:

$$\mathcal{C}_{\text{glo}} = \frac{3 \times \text{Total number of triangles}}{\text{Total number of triples}}. \tag{3.13}$$



(a) Triangle: three vertices that are directly connected to each other via all possible edges.

(b) Triple centered at node v_1: a triangle missing the edge between node v_2 and v_3.

Figure 3.4: Sketch to illustrate the definition of triangles and triples in a graph.

The number three in the numerator compensates the fact that each triangle is counted as three triples.

connected component size A graph consists of one or more connected components, in which it is possible to reach every vertex from any other vertex by walking along edges. The connected component encompasses all vertices that can be reached. The size of such a component is the number of vertices it contains [44].

3.3.1 Methods

Datasets

Our protein of interest is the HCN channel. To check whether our findings are unique to this protein, four control proteins were selected. We first chose three Pfam families of ion channel domains. Pfam alignments are manually curated and therefore thought to be of high quality. Similar behavior of our HCN alignment and the Pfam MSAs in the analyses will serve as validation of MSA quality. Calmodulin, a soluble protein, was appointed to be the fourth control to examine if coevolutionary graphs of membrane and soluble proteins share the same properties.

1. Data collection of **HCN channel** sequences is described in Section 3.2.1. Alignment was performed using `MUSCLE` with default parameters and eye-optimized.

2. The full domain alignment of the Pfam family **PF07885** (bacterial two-transmembrane-helix channels) was downloaded from the Pfam database version 26.0 [125]. Since the focus was on potassium channels, only those sequences with UniProt [151] annotations containing the term “potassium” were included in the final data set. To reduce errors we also removed sequences that introduced a gap in all other sequences. Finally, all MSA columns with less than 200 non-gap characters were deleted.

3. The full domain alignment of the Pfam family **PF01007** (inwardly rectifying channels) was downloaded from the Pfam database. Processing of the alignment was as described above for PF07885.

4. The full domain alignment of the Pfam family **PF000520** (eukaryotic six-transmembrane-helix channels) was downloaded from the Pfam database. Processing of the alignment was as described above for PF07885.

5. A BLAST search with human **Calmodulin** (GenBank ID 5542035) in the non-redundant (nr) database was performed; a stringent E-value cutoff of 10^{-50} was applied to exclude related proteins such as centrins and troponin C. An MSA was built using MUSCLE with default parameters. Duplicated entries were removed and MSA columns with less than 200 non-gap characters were deleted.

Graphs

MI (Equation 3.3) of MSA columns was computed using the BioPhysConnectoR package [77] in R [126]. Z-scores were computed according to Equation 3.8 based on 10 000 shuffling runs. MI values of position pairs that had a Z-score greater than four were considered statistically significant [111]. Graphs were constructed from the MI results by connecting those vertices (MSA positions) whose MI was in the top fraction of the significant MI values for the MSA in question. This top fraction was varied to create graphs of different density: 0.5%, 1%, 2%, 5% and 10%. The density of a graph is the portion of possible edges that are realized; a fully connected graph has a density of 100%.

In addition to the MI graph for each MSA, two different null models were created. Null model I held the number of vertices and the degree of each vertex constant while the edges were randomized. Null model II only kept the total number of vertices and edges of the entire graph constant; vertex degrees were free to vary from the underlying MI graph. This null model is a classical random graph, a type of Erdős-Rényi model [54]. We wanted to test whether this random graph behaves differently from an MI graph due to the lack of an underlying coupling of the nodes. 1000 replications for each of the null models were performed.

Graph measures were computed for all graphs using the igraph package [40] in R and the distributions of the MI graphs and the null models were compared.

3.3.2 Results and Discussion

Small-World Networks

The analyzed graphs of all five protein MSAs showed similar properties. They are typical small-world networks with low mean shortest path values and high global clustering coefficient [109,160]. Table 3.2 lists the mean shortest path and global clustering coefficient for all five proteins at a density of 1% in comparison with the average value of 1000 randomized networks of null model II.

	Mean shortest path		Global clustering coefficient	
	MI	NMII	MI	NMII
HCN	2.09	4.00	0.470	0.010
PF00520	3.62	5.88	0.435	0.010
PF01007	3.49	4.83	0.404	0.010
PF07885	2.07	4.84	0.664	0.010
Calmodulin	2.33	7.65	0.524	0.009

Table 3.2: Mean shortest path and global clustering coefficient of the five protein MI networks at a density of 1%. The same measures as mean values over 1000 randomized networks of type II (NMII) are given for comparison. Data for the proteins highlighted in gray are shown in the following analyses.

Comparison of MI Graphs with the Null Models

Several graph measures of the five protein graphs were analyzed. To visualize the results, plots that show the graph measures for a protein network at five different graph densities were created. For each combination of parameters (graph measures, network densities and proteins) histograms of the values were built and normalized so that the number of counts equaled one. The decadic logarithm of these normalized counts was used for the construction of a color scale to enable better discrimination at low counts. This logarithmic color scale is illustrated in Figure 3.5. It is used in the following plots where indicated.

The results for all five proteins are basically the same, the main difference being the number of vertices and the level of conservation, which leads to small variations. Therefore, in the following figures only HCN and two of the control proteins (CaM and PF00520) are pictured, whereas the plots for PF01007 and PF07885, which are highly similar to the others, are omitted.

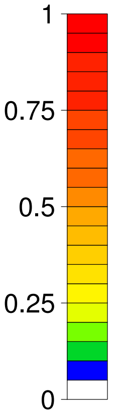
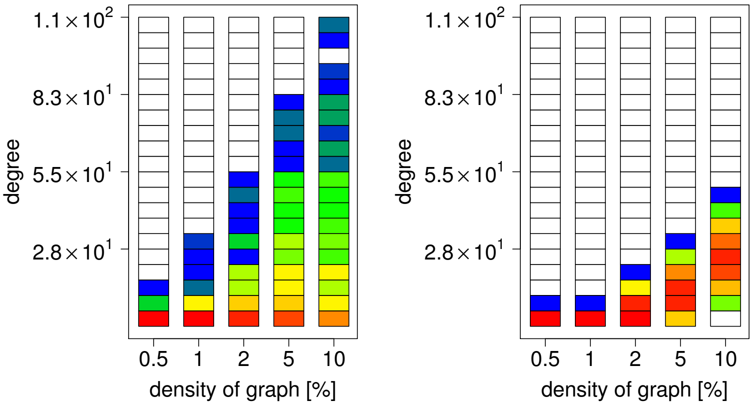


Figure 3.5: Histogram color scale for plots of graph measures. The values for each measure in a graph were counted and normalized so that the number of values equaled one. The decadic logarithm of these normalized counts was taken and the results were translated to this color scale ranging from blue (low counts) to red (high counts), where one corresponds to the number of values in a bar. Counts of zero are colored white.



(a) MI graph

(b) Null model II

Figure 3.6: Degree distribution of PF00520. For definition of color scale, see Figure 3.5.

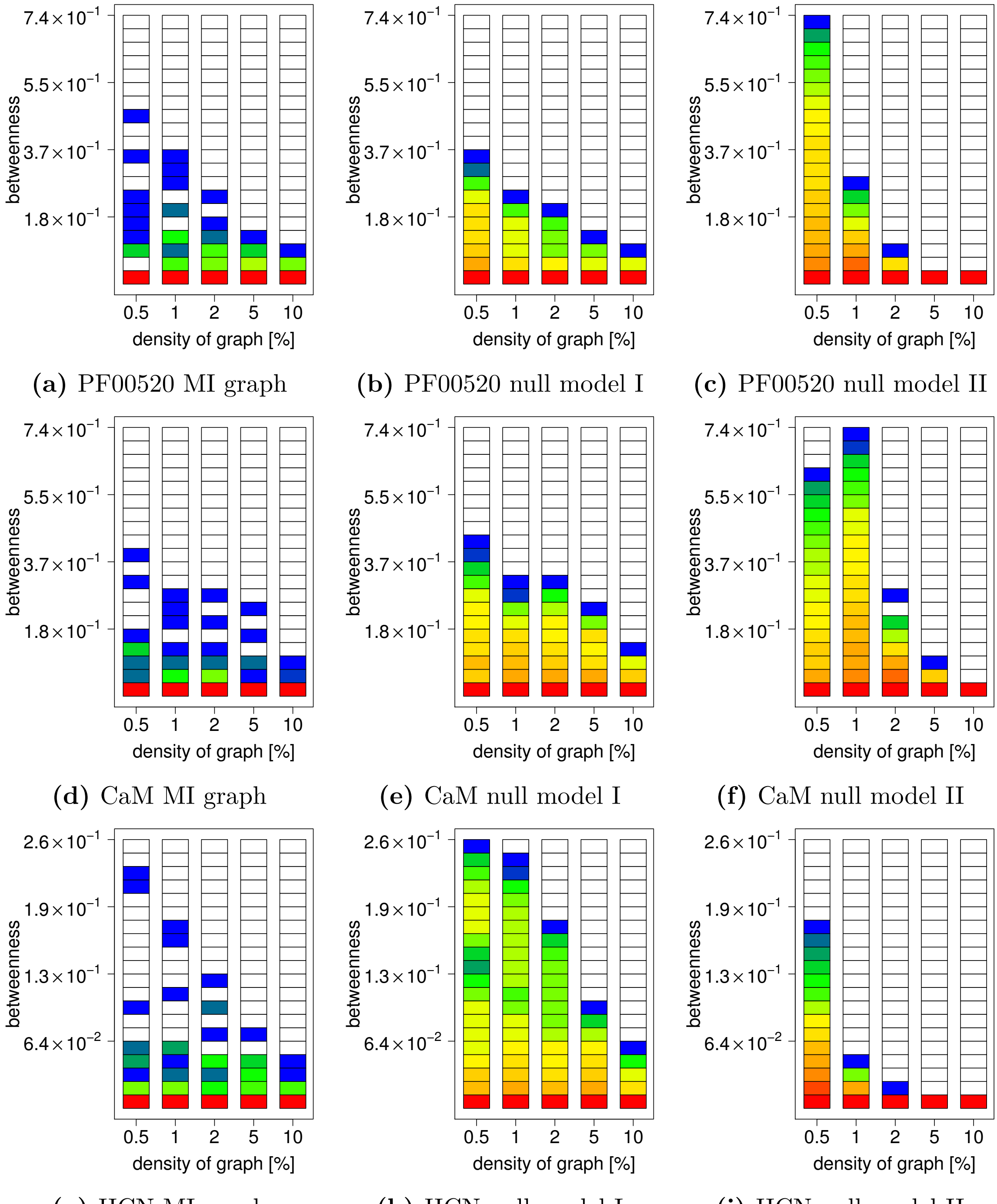
Degree Figure 3.6 shows the degree distribution of the PF00520 graph (a) in comparison to the null model of type II (b), where the number of edges and vertices was maintained during randomization. The comparison to the degree distribution of null model I is unnecessary because the degrees of each vertex were kept fixed during edge randomization. It is therefore the same as that of the MI graph.

As already mentioned above, coevolutionary networks feature the typical properties of small-world networks. Here, we can see again the small-world character of the MI graph with a few high degree hubs and many vertices with medium to low degree, while the graph with random edge placement features mainly low degree vertices. Due to conserved MSA columns with low entropy, there are a lot of vertices with degree zero (cf. Table 3.3). Also, we can see that a density of 10% is set too high: vertices with a degree of 110 appear, which means some vertices are connected to almost half of the overall 239 vertices in the graph.

Betweenness The betweenness is defined as the fraction of shortest paths that run through a vertex. Figure 3.7 reveals a decrease of betweenness with increasing density for all proteins as well as null model I and II.

Most vertices in the MI graphs have very low betweenness and again a few hubs with high betweenness can be found. From this we can recognize that the high degree nodes are indeed those through which the shortest paths lead. Null model I shows a similar pattern with a smoother distribution. This is most likely accounted for by the higher number of values originating from the 1000 randomized graphs. Due to the division by the total number of paths, the betweenness of the hubs drops as more and more paths emerge at higher densities.

Null model II has higher maximum values at first, which rapidly drop with increasing density. Since hubs are missing in null model II, shortest paths are much longer (cf. Figure 3.11), thereby leading through more vertices and increasing their betweenness.



(a) PF00520 MI graph **(b)** PF00520 null model I **(c)** PF00520 null model II

(d) CaM MI graph **(e)** CaM null model I **(f)** CaM null model II

(g) HCN MI graph **(h)** HCN null model I **(i)** HCN null model II

Figure 3.7: Betweenness of PF00520, CaM and HCN. For definition of color scale, see Figure 3.5.

Therefore, at low densities there are only few paths between many vertices. With increasing number of edges more possible routes between null model II vertices emerge, which causes betweenness to decrease. At a density of 2% the betweenness of null model II drops below that of null model I; for HCN this is already the case at 0.5%.

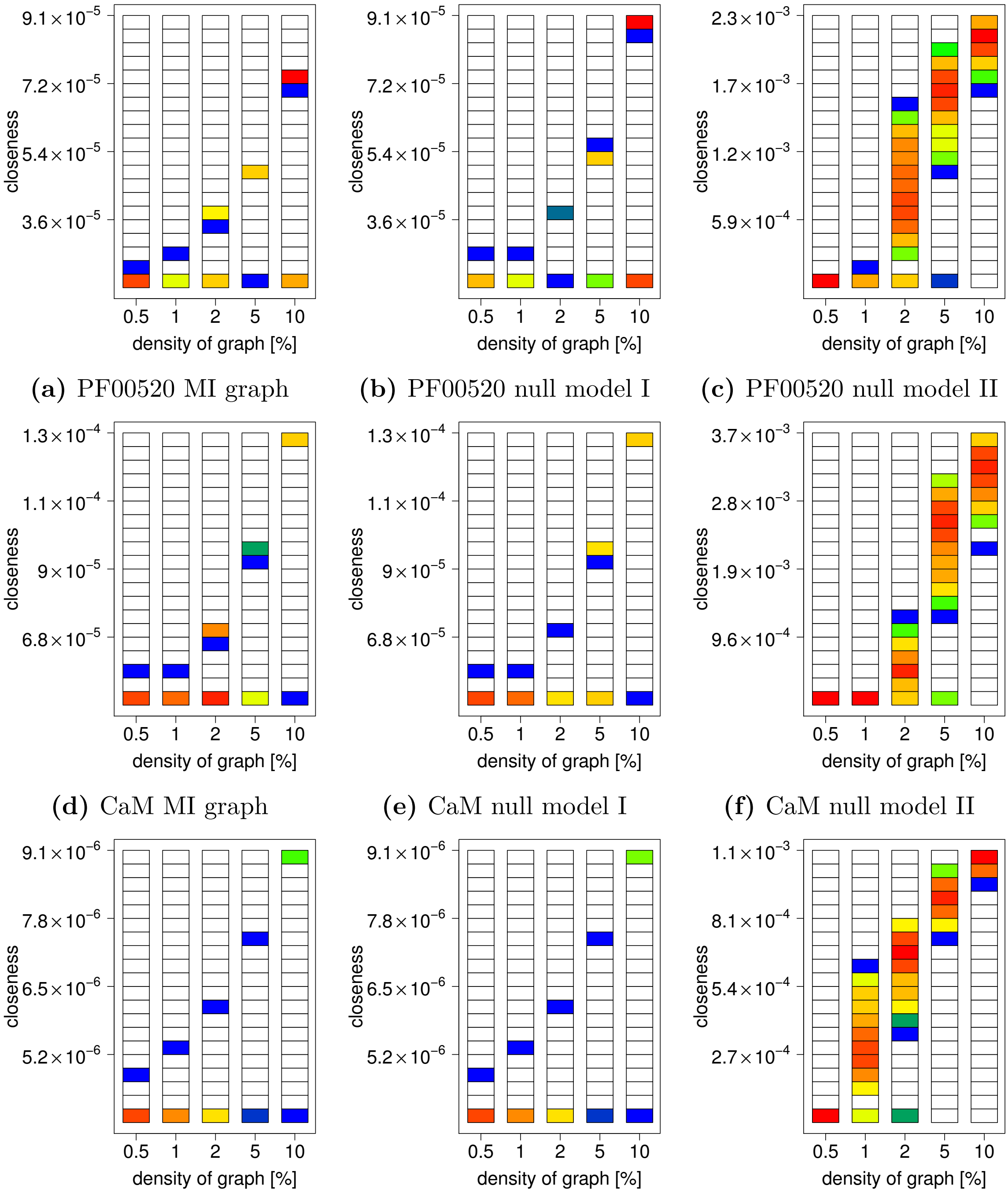
Closeness As closeness is the reciprocal of the sum of the distances of a vertex to all other vertices, the distances of each vertex to all other vertices need to be measured for its computation. By definition, the distance between unconnected vertices is considered to be the total number of nodes in the network (Equation 3.11). This leads to the effect that vertices with a degree of zero strongly reduce the closeness of all other nodes in the network. Since there are fewer zero-degree vertices in null model II, the closeness of all nodes is much higher.

Note that the vertical scale for null model II in Figure 3.8 is two orders of magnitude larger than that for the MI graph and null model I. It is not surprising to see that the closeness becomes larger with increasing graph density. The remainder of low closeness values at high densities of the MI graph and null model I is again due to conserved MSA positions which create zero-degree nodes. Null model II has no such constraint, therefore low closeness values completely disappear at densities of 5% and higher for PF00520 and HCN or 10% for Calmodulin.

Local Clustering Coefficient The local clustering coefficient measures the ratio of the number of triangles connected to a vertex to the number of triples centered on it. It is one of the few graph measures for which the behavior of null model I deviates from that of the original MI network. Figure 3.9 shows that the clustering coefficients for null model I spread over the whole range with a tendency toward the lower boundary while the MI graph features considerably fewer low values. This indicates that the neighbors of a node in the MI graph have a higher probability to be connected with each other than in a graph with random distribution of edges, albeit fixed vertex degrees. This is due to the MI graph being based on a measure of correlation: if a correlates with b and b correlates with c, there is an indirect connection between a and c. Burger and van Nimwegen showed this for coevolutionary relationships in MSAs [29]. Since these indirect correlations are missing in null model I, there are more vertices with lower local clustering coefficients. In line with expectations, null model II (Figures 3.9c, 3.9f and 3.9i) shows lower local clustering coefficients than the MI graph and null model I because the edges are distributed in a random manner across the graph.

Connected Components A connected component of a graph is a subgraph in which all vertices can be reached from all other vertices in the connected component by walking along edges. All reachable vertices belong to this same connected component.

Figure 3.10 illustrates the connected components of the MI graphs and the null models sorted decreasingly by size. For null models the average connected component sizes over 1000 random graphs are plotted. Vertices with a degree of zero are not taken into account.

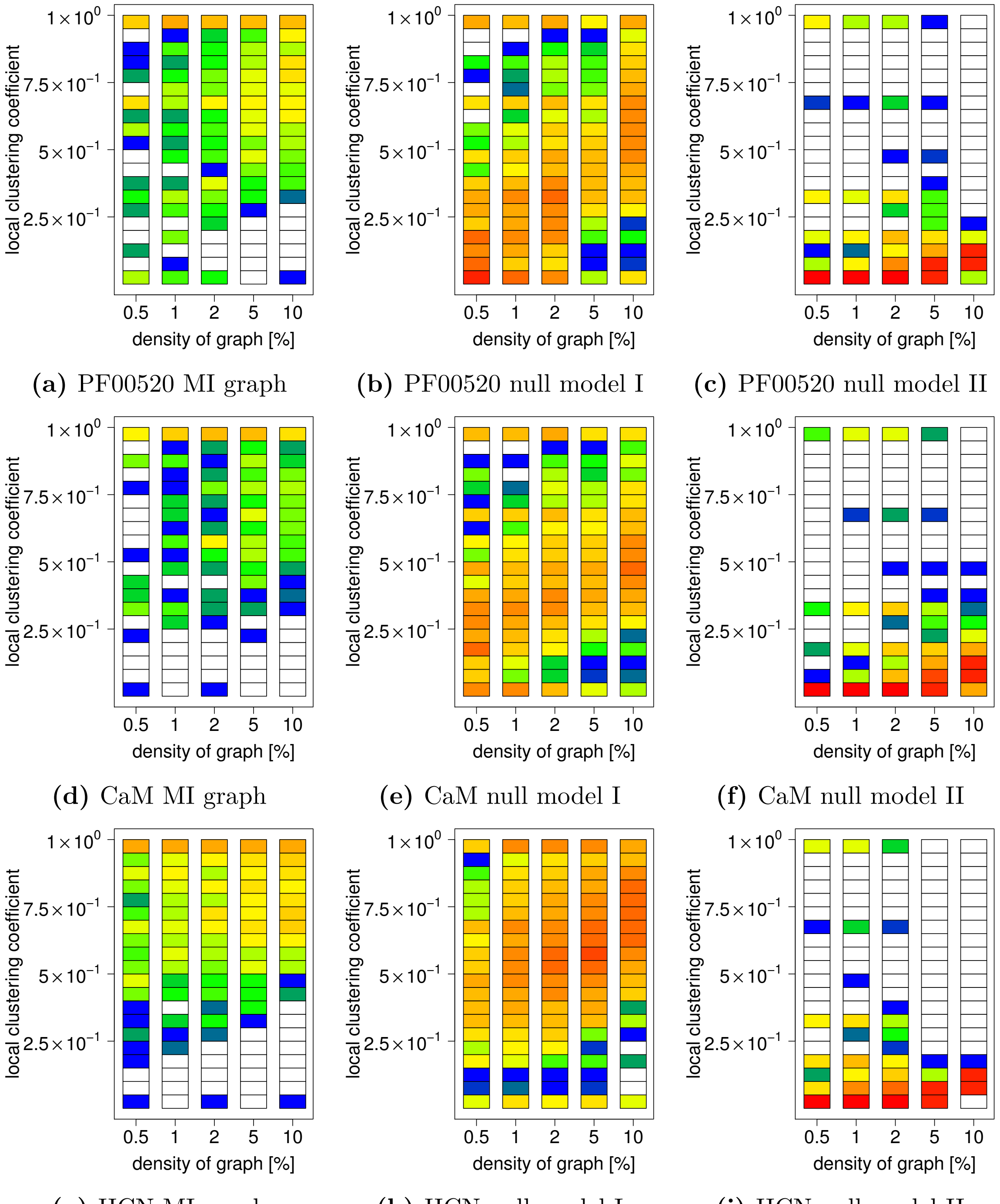


(a) PF00520 MI graph **(b)** PF00520 null model I **(c)** PF00520 null model II

(d) CaM MI graph **(e)** CaM null model I **(f)** CaM null model II

(g) HCN MI graph **(h)** HCN null model I **(i)** HCN null model II

Figure 3.8: Closeness of PF00520, CaM and HCN. For definition of color scale, see Figure 3.5. Note the differences in scale for null model II.



(a) PF00520 MI graph **(b)** PF00520 null model I **(c)** PF00520 null model II

(d) CaM MI graph **(e)** CaM null model I **(f)** CaM null model II

(g) HCN MI graph **(h)** HCN null model I **(i)** HCN null model II

Figure 3.9: Local clustering coefficient of PF00520, CaM and HCN. For definition of color scale, see Figure 3.5.

The most distinct feature is that MI graphs often have a few small components in addition to the main component, while this does not occur in null model I. Most of these small components consist of only two or three vertices. They are pairs of residues that coevolve with each other but lack coevolution with other partners [63].

Null model II contains many small connected components at low densities, which merge into one large component as soon as the density is sufficiently large. In this behavior it differs from null model I and the MI graph because it lacks the large number of vertices with a degree of zero and the hubs that attract many edges. Therefore, the edges can distribute over the whole graph, which leads to several subgraphs without connections to each other.

Shortest Paths The shortest path between two vertices is the one leading along the fewest possible edges.

In Figure 3.11, the distribution of all shortest paths in each graph is plotted. As all lengths of shortest paths were sufficiently small integer numbers, no re-binning was performed to facilitate better resolution. The number of bins in Figure 3.11 therefore corresponds to the longest shortest path measured and differs from the number of bins used in previous figures. We can see that the distribution hardly changes with increasing density both for the MI graphs and null model I. This shows that the topology of the network with a few highly frequented hubs is already present at low densities. Adding more edges only connects more nodes (cf. Table 3.3) to the network but does not significantly change its architecture. Null model II requires higher densities to reduce path lengths because there are fewer zero degree vertices. The short path lengths are a typical property of small-world networks. In contrast, regular lattices display very large shortest path lengths [160].

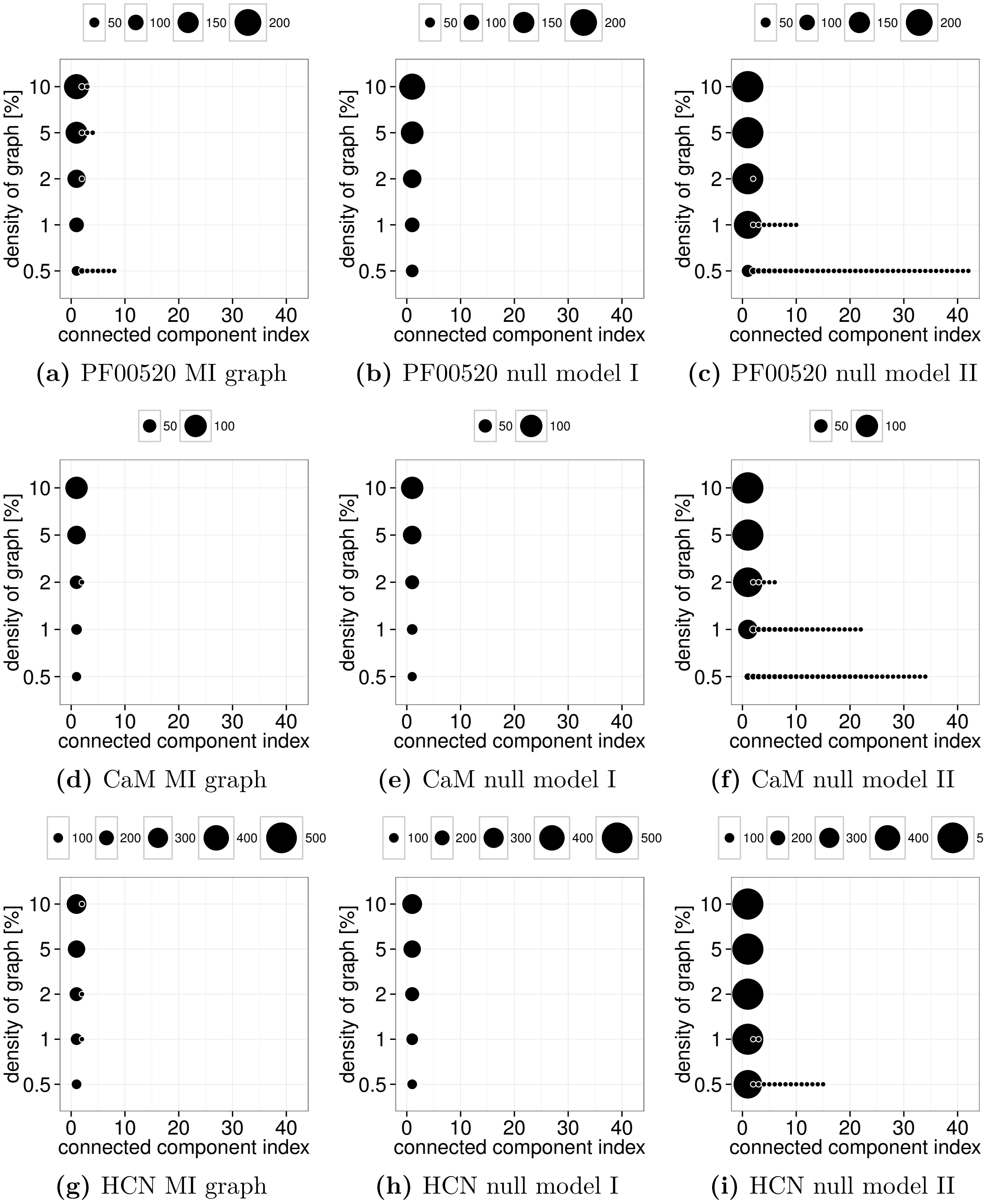
Differences Between the Proteins

The largest difference between the graphs of the five proteins can be attributed to their size and thereby the size of the respective MSA. The number of MSA columns is the number of vertices in the resulting graph. The number of possible edges depends on the number of vertices, since the largest possible number of edges E_{theomax} in an undirected, simple graph is reached if every node is connected to all others:

$$E_{\text{theomax}} = \tfrac{1}{2}V(V-1), \tag{3.14}$$

with V being the number of vertices in the graph. The density of a graph is the fraction of possible edges that are realized. Therefore, at a density of 1% the mean degrees of the protein networks are as follows: HCN (5.08), PF00520 (2.38), PF01007 (3.35), PF07885 (1.21) and CaM (1.49).

These values explain a phenomenon which can best be detected when comparing the plots of null model II, e.g., Figures 3.11c, 3.11f and 3.11i. The plotted distributions for a large protein network at low density resemble distributions of smaller protein graphs at higher densities. Therefore, when comparing graph measure distributions of

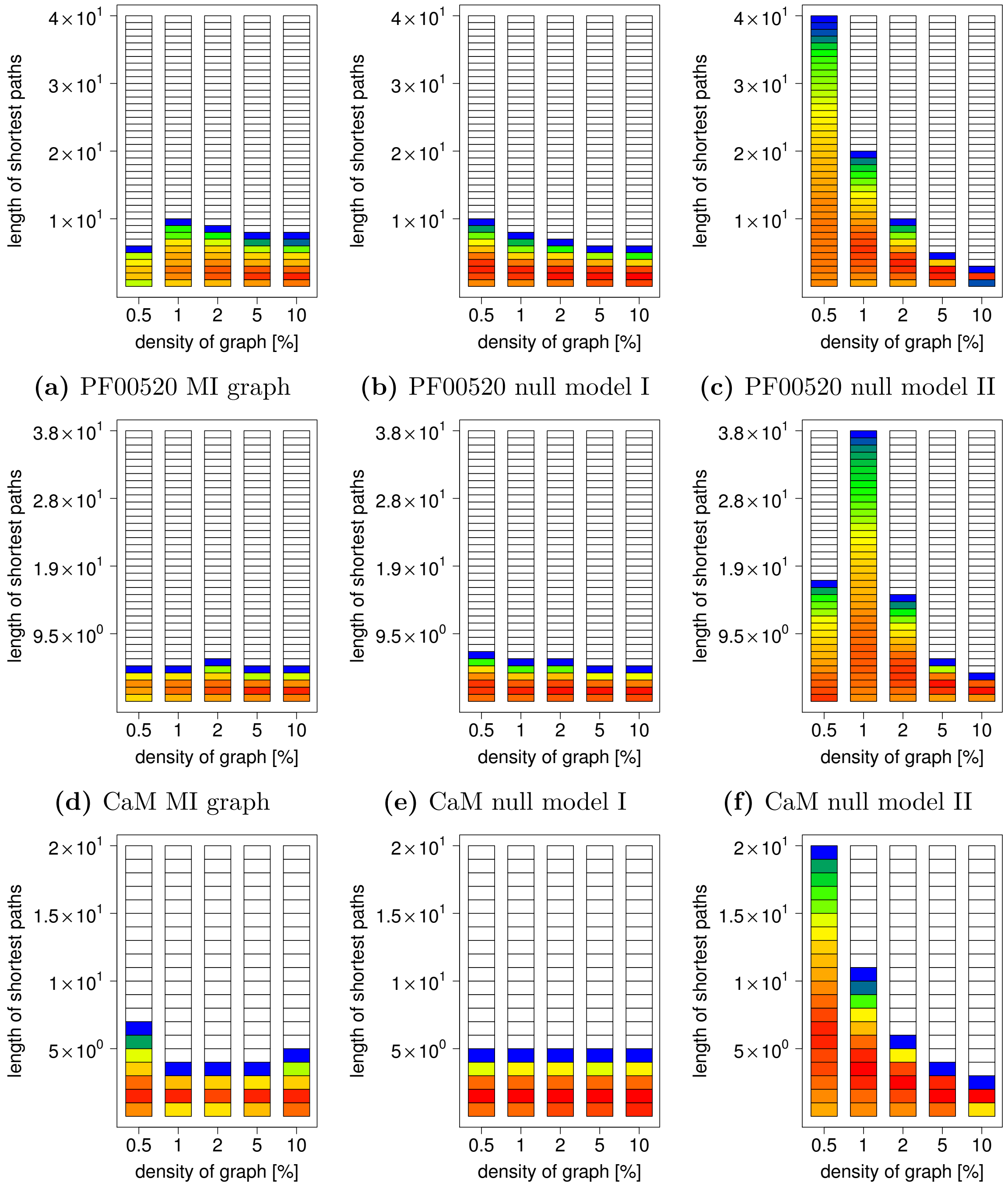


(a) PF00520 MI graph **(b)** PF00520 null model I **(c)** PF00520 null model II

(d) CaM MI graph **(e)** CaM null model I **(f)** CaM null model II

(g) HCN MI graph **(h)** HCN null model I **(i)** HCN null model II

Figure 3.10: Connected components of PF00520, CaM and HCN. Size of the dots represents size of the components.



(a) PF00520 MI graph **(b)** PF00520 null model I **(c)** PF00520 null model II

(d) CaM MI graph **(e)** CaM null model I **(f)** CaM null model II

(g) HCN MI graph **(h)** HCN null model I **(i)** HCN null model II

Figure 3.11: Distribution of length of shortest paths for PF00520. Differences in the number of bins arise from not re-binning the measured values for better resolution. For definition of color scale, see Figure 3.5.

	Entropy	0.5%	1%	2%	5%	10%
HCN (509)	0.98	0.198	0.271	0.360	0.485	0.576
PF00520 (239)	2.49	0.310	0.385	0.527	0.682	0.812
PF01007 (336)	2.29	0.345	0.467	0.604	0.765	0.854
PF07885 (121)	2.36	0.264	0.355	0.496	0.686	0.785
Calmodulin (149)	1.05	0.181	0.235	0.356	0.530	0.671

Table 3.3: Mean entropy of MSA columns and the fraction of vertices with non-zero degree in the five protein MI networks at all analyzed densities. The total number of vertices, which is also the number of MSA columns, is given in parentheses behind each protein name. The proteins highlighted in gray were shown in the analyses of graph measures.

different proteins, one should keep in mind that equivalence of the network systems cannot necessarily be derived from equal density.

Apart from MSA size, i.e., the number of columns in the MSA, another difference between the five protein families is the level of sequence conservation and entropy in the MSAs. The conservation of an alignment position can be measured by computing its Shannon entropy (see Equation 3.2). Completely conserved MSA columns feature a Shannon entropy of zero. The highest possible diversity within an alignment column is the occurrence of each of the 21 symbols (20 amino acids plus a gap character) with the same frequency, resulting in a uniform distribution. In this case, the maximum of the Shannon entropy would be reached, which is $\log_2 21 = 4.39$. Table 3.3 shows that the MSAs of HCN and Calmodulin have lower mean MSA position entropy, i.e., the sequences in the MSA are more strongly related and therefore less diverging than those in the Pfam MSAs. This is due to the stringent selection of sequences we applied to the HCN and CaM data sets.

Highly conserved residues are not able to coevolve because there can be no evolution without variability. Hence, conserved sequence parts increase the number of zero degree nodes in the MI graph. If the edges added to the graph when the density is increased cannot distribute among all nodes, the degree of all non-zero degree nodes rapidly grows. This restriction induces very high degrees in some nodes, while others maintain a degree of zero as can be seen in Figure 3.6. The fraction of vertices with non-zero degree is always lower for MSAs with more conserved sequences, which can be recognized in Table 3.3.

Conclusion

Our analyses have shown that MI graphs have properties that are characteristic for small-world networks. They differ in all graph measures analyzed here from networks with random edge distribution (null model II). Furthermore, they can even be distinguished from randomized networks with maintained vertex degrees (null model I) by examining the local clustering coefficients and the fact that they often show several small connected components besides the main component.

These basic features were observed for all proteins examined in this study. Membrane proteins—here represented by various families of ion channels—do not behave differently from soluble proteins (in our case Calmodulin). This is in line with observations from other groups, who could not find any essential differences in the coevolutionary networks of different proteins.

However, while the MI graph approach reveals network properties in general, it fails to provide detailed information about the specific protein. This can already be seen in the study by Chakrabarti and Panchenko [35], who tested a coevolutionary network approach. They analyzed whether active sites, ligand, metal or protein binding sites exhibited network properties different from those of non-functional sites. The most potent graph measure for this purpose turned out to be the degree of a node, which is simply the number of significant coevolutionary interactions of a residue. Building a coevolution-based graph is not necessary to obtain this information.

Fatakia et al. claimed to have used graph theory to detect coevolved sites in G protein-coupled receptors. However, they merely analyzed degree distributions [55]. Weil tried to identify binding sites of antibiotics in ribosomes by computing graph measures of networks derived from coevolutionary relationships, but none of the measures tested in the study was qualified to accomplish the desired task [162]. To the author's knowledge, no other studies about graph theoretical analysis of MI-based coevolutionary networks in proteins have been published.

To gain more detailed information about intramolecular coevolution, the hubs in the MI network and the coevolutionary bonds featuring the highest MI values need to be analyzed. This task is addressed in Section 3.4. The protein of interest in the following analyses is the HCN channel.

3.4 Coevolving Residues in HCN Channels

Most of the time, the search for coevolving position pairs is the quest for finding key players in protein function [34]. If the change of an amino acid needs to be compensated by coevolving partners, this amino acid is most likely important for the protein in some way. Otherwise there would be no selective pressure and therefore no detectable coevolution. Hence, coevolving residues are, besides conserved amino acids, the major objects of interest. Therefore, in coevolutionary analysis, both residues that coevolve with many others and single interactions between small groups or pairs of residues are examined [42, 63].

For the HCN channel, the protein of interest, several aspects will be checked for coevolutionary relationships. Local links within and between subunits will be analyzed as well as long-range interactions. We also search for hubs, i.e., residues that coevolve with many others.

3.4.1 Methods

Data Set

The HCN protein sequence data set was prepared as described in Section 3.3.1. To facilitate interpretation of the results, the MSA was mapped to the sequence parts that are present in the structure of the homology model described in Section 2.3.1. Analysis was performed only for those MSA areas that have a counterpart in the model structure.

MI Computation

MI values were computed according to two approaches: the first which regards the gap character as another symbol in the alphabet as well as the second which computes MI from a gap-free subset of each MSA column pairing (see Section 3.1.2). To distinguish between the two methods, the former is called "ORMI" (original MI) while the latter is denoted as "SUMI" (subset MI) throughout this section.

Two normalization methods were applied as described in Section 3.1.2. For Z-score normalization MI values with a corresponding $Z > 4$ were considered statistically significant. Values with $Z \leq 4$ were set to zero before result evaluation. The second method was minH normalization, for which values were treated according to Equation 3.9. Both normalization methods were applied to both ORMI and SUMI.

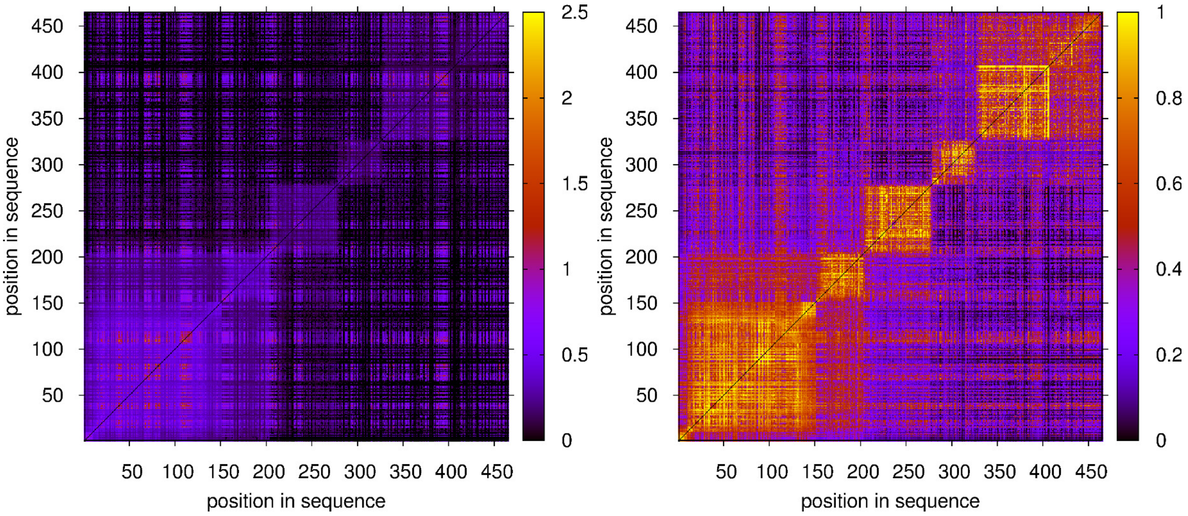
MI was computed with the `BioPhysConnectoR` package in `R`. Matrix plots were created using `gnuplot` [169].

3.4.2 Results and Discussion

Domain Patterns Detected in ORMI Values

The ORMI of the prepared HCN channel alignment was computed to identify coevolving pairs of residues and thereby to shed light on the importance of individual residues and areas in the protein. To this end, we screened for high ORMI values, as these reveal variable positions coupled to each other by coevolutionary constraints. As explained in Section 3.1.2 the maximally possible MI value depends on the one-point entropies of the MSA columns. To detect pairs of residues that coevolve but whose MI value is limited by low one-point entropy, the minH normalization can be used.

Figure 3.12 shows the ORMI matrix with Z-score cutoff (a) and minH normalization (b). Characteristic stripes of low MI values are visible especially in the Z-score cutoff matrix (a). They originate from residues with very low one-point entropy, that show low MI with all other residues due to the dependency of the maximally possible MI value on the one-point entropies. However, another, much more intriguing pattern emerges in both illustrations: areas of high coevolutionary signals appear along the diagonal suggesting the existence of domains in the protein that exhibit coevolution within themselves but little or no exchange between each other. The fact that the pattern is more distinct in the minH normalized matrix (Figure 3.12b) indicates that this phenomenon mostly affects low entropy (i.e., conserved) MSA columns.



(a) ORMI with Z-score cutoff. All ORMI values with a corresponding $Z \leq 4$ were set to zero (see Equation 3.8).

(b) minH normalization. ORMI values are normalized to the minimum of the respective one-point entropies (see Equation 3.9).

Figure 3.12: ORMI matrix of HCN channel alignment. Fields in the image are colored according to the ORMI value of the respective matrix element. The diagonal containing the one-point entropies is set to zero.

This very unusual pattern formation is visualized in the HCN structure in Figure 3.13. All residues pairs featuring a minH normalized ORMI value of one, which is the maximally possible value for this normalization method, are connected by links in the structure of the homology model; the construction of this model is described in Section 2.3.1. Clearly, connections are only present within the areas that appeared in the matrix plots, while inter-domain links are completely absent. The links divide the protein into five regions: the S1–S4 helix bundle, the S5 helix, the S6 helix together with pore helix and filter loop, the C-linker and finally the cyclic nucleotide-binding domain (CNBD); for location of structural elements within the HCN protein see Figures 2.1b and 2.4d.

Normally, both long- and short-range coevolutionary interactions are possible, there is no limitation to certain areas within the protein. To find the root of this highly atypical property, the original MSA was reinspected. Here, we discovered that several rows contained only partial sequences whose locations in the MSA coincided with domain boundaries. Generally, missing parts in a small number of sequences do not cause any problems—they are outweighed by the remaining data. But in a small data set with rather conserved MSA columns, as is the case with our HCN alignment, they induce coevolutionary signals: Within an affected sequence region, a gap in column X always occurs if there is also a gap in column Y. In conserved columns these gaps are (almost) the only variation, therefore the impact on the MI value is greatest in low entropy columns. This is the reason why the resulting signal is more distinct in the minH normalized data set, where coevolution of alignment columns with low entropy is revealed.

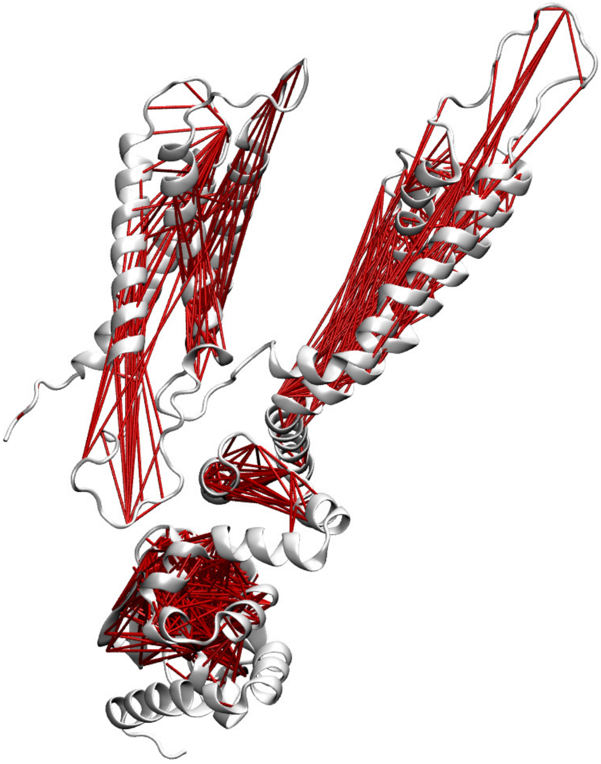


Figure 3.13: One subunit of the HCN tetramer shown in white with red links representing coevolutionary interactions with a minH normalized ORMI value of one.

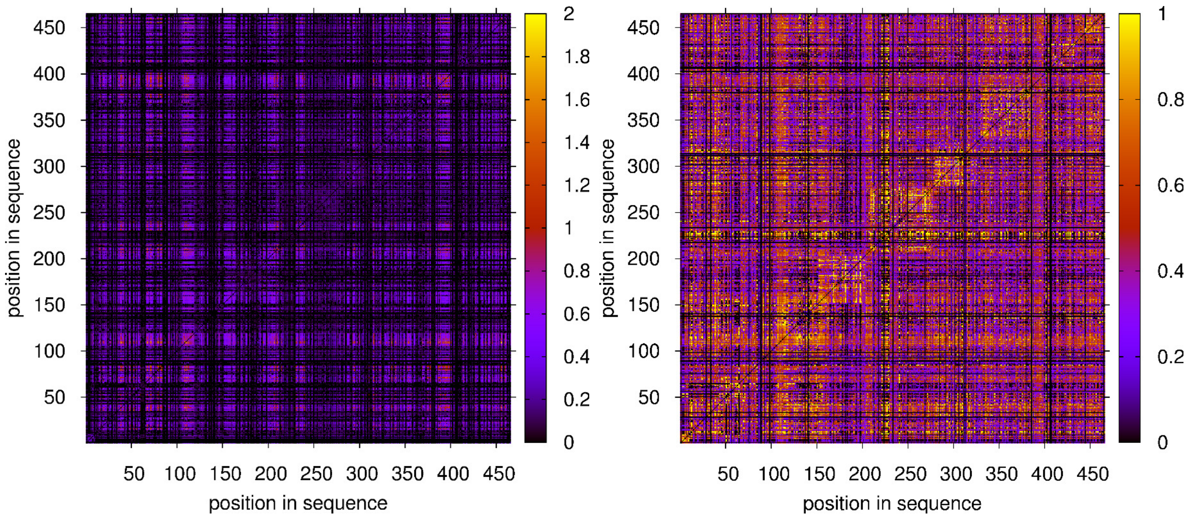
The incomplete sequences were examined. Some of them were annotated as “partial”, while others were simply missing either the N- or C-terminal part without being specified as incomplete. We therefore assume that these missing sequence parts arise from inaccuracies in the sequencing or documentation process rather than genuine deletion mutations in the corresponding protein-coding gene. Thus, the coevolutionary signals they induce are considered false positives. A way to clear the MI signal from these perturbations was looked for and found in the computation method of SUMI.

SUMI Eliminates False Positives

To avoid further reduction of the data set by removing all partial sequences, we decided to employ SUMI. As described in Section 3.1.2, this approach extracts a gap-free subset of sequences for each column pairing, which contains only rows that carry non-gap characters in both of the columns under investigation. The MI is then computed only for this subset. This procedure allows for examination of coevolutionary relationships in the HCN channel without the falsely positive signal that was introduced by stretches of gaps in the alignment.

SUMI data are shown in Figure 3.14. As expected, the pattern of residues coevolving only within domains, which we saw in Figure 3.12, has disappeared in both data sets, the Z-score cutoff (a) as well as the minH normalization (b). The following analyses are performed with these SUMI data sets.

Figure 3.15 shows the top 20 interactions with the highest SUMI values normalized by Z-score cutoff. There are some long-range interactions from the S1–S4 helix bundle to



(a) SUMI with Z-score cutoff. All SUMI values with a corresponding $Z \leq 4$ were set to zero (see Equation 3.8).

(b) minH normalization. SUMI values are normalized to the minimum of the respective one-point entropies (see Equation 3.9).

Figure 3.14: SUMI matrix of HCN channel alignment. Fields in the image are colored according to the SUMI value of the respective matrix element. The diagonal containing the one-point entropies is set to zero.

the CNBD and several short-range links within the S1–S4 helix bundle (for location of secondary structural elements refer to Figures 2.1 and 2.4). None of these top coevolving pairs include residues of the inner part of the channel (S5, S6, pore helix) or the C-linker. Furthermore, many of the links are concentrated at residues not only connected to one but several others. This fact is also illustrated in Figure 3.16. Here, an HCN subunit is colored by the one-point entropy of its residues with low values highlighted in blue and high values in red. Additionally, beads were drawn into the structure indicating hubs in the network of coevolutionary relationships: residues which participated in at least five of the top 100 coevolutionary interactions determined by the highest SUMI values with a Z-score cutoff of $Z > 4$. Expressed in graph theoretical terms, these are the high degree nodes of the MI graph.

We can see that all hubs are colored in shades of red indicating high entropy values. This is due to the MI depending on the MSA column entropy. Residue pairs can never reach MI values higher than the minimum of their one-point entropies. This leaves low entropy positions in the range of lower MI values even if they coevolve as strongly as possible within their constraints, i.e., they cannot be detected by analyzing the top hits of Z-score cutoff normalized SUMI. Thus, while the Z-score cutoff is a well-suited method to detect coevolution between hubs and other high entropy positions, it does not allow for identification of coevolving low entropy sites. This issue is addressed in the next part of this section.

In Figure 3.16 it also becomes apparent that what seems to be one large hub at the top of Figure 3.15 is actually two hubs situated in the S3 helix and the adjacent extracellular

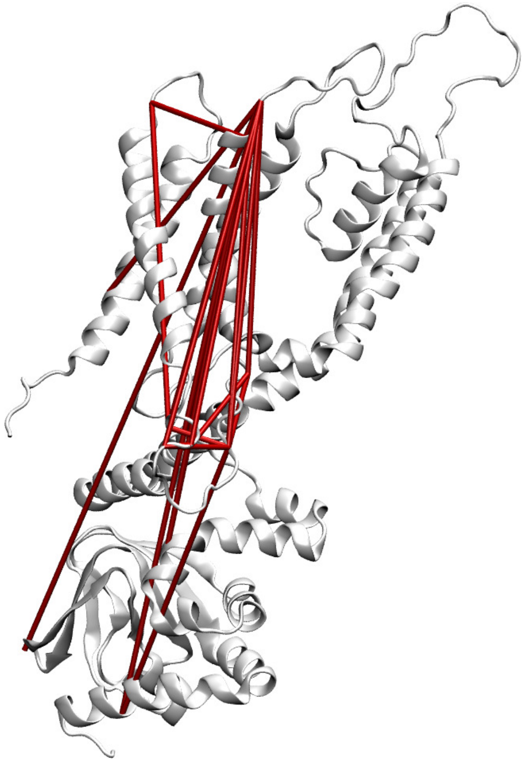


Figure 3.15: Subunit of the HCN tetramer shown in white. The 20 interactions with the highest SUMI values (after Z-score cutoff) are drawn as red links.

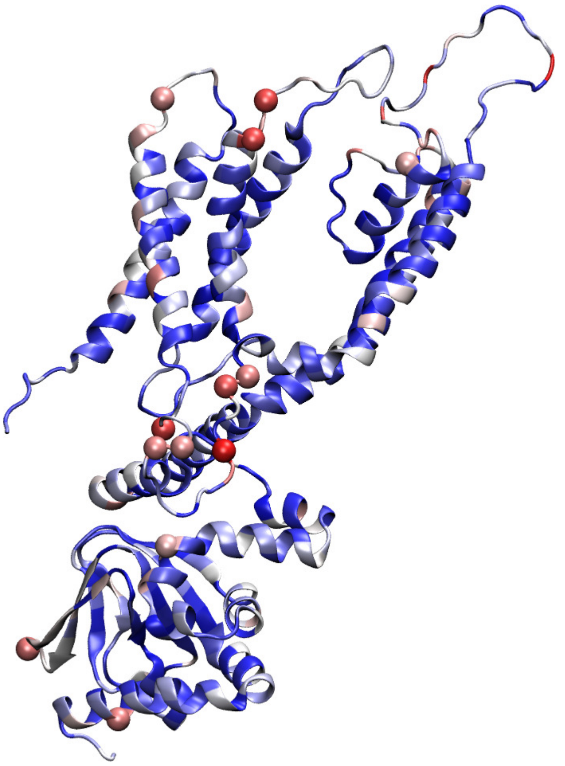


Figure 3.16: Subunit of the HCN tetramer with residues colored by their one-point entropy ranging from low (blue) to high values (red). Beads represent hub residues, that participate in at least five of the top 100 interactions. They are also colored according to their one-point entropy.

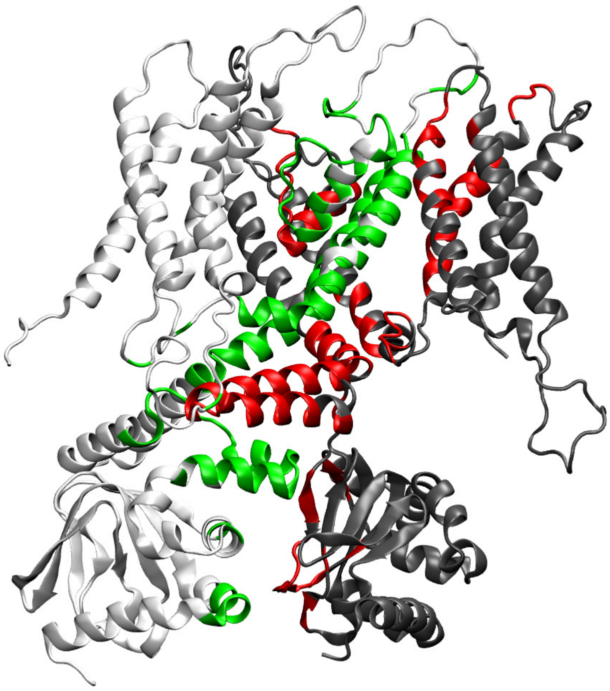


Figure 3.17: Interface between two subunits of the HCN channel tetramer. One subunit is colored in white with the residues constituting the interface highlighted in green. For guidance, the white subunit's orientation is the same as that in Figure 3.15. The neighboring subunit is colored in gray with red interface residues. Since the tetramer is symmetric and all subunits are identical, only one of the four interfaces is depicted. The remaining two subunits are omitted for clarity.

loop: 108 and 110 (for reference of residue numbers see Appendix A). The three hubs located in the CNBD were also already among the top 20 interactions (335, 394 and 455). A residue in the pore helix (214) which was not involved in the top 20 SUMI links is now revealed as a hub taking part in six out of the top 100 interactions. Its coevolutionary connections link it to four other hubs (79, 82, 110, 455) as well as two neighboring residues in the extracellular loop adjacent to the pore helix (209, 210).

A hot spot of coevolution is found in the intracellular loop between S2 and S3, which again is owed to several high entropy values here. However, high entropy is a necessary but not sufficient condition for the existence of a hub. As we can see in the loop region between S5 and the pore helix, there are three residues colored in a brighter shade of red than several of the hubs indicating higher entropy. Nevertheless, none of these positions can be found in the set of hubs.

Since the HCN channel is a homotetramer, it is not possible to distinguish whether coevolutionary signals arise from intra- or inter-subunit interactions. Figure 3.17 shows the interface between two individual subunits. All amino acids of which at least one atom is within a distance of 7 Å from an atom of the adjacent subunit are considered interfacial residues. These residues are colored green in the white subunit and red in the gray subunit. As the tetramer is symmetric and the subunits are identical, only one out of four interfaces is shown. Especially in the transmembrane domain, where the helices are intertwined, a lot of residues are in contact. Note that both subunits also have an interaction partner next to them on their other side, so the area that is in contact with either the subunit to the right or to the left comprises 268 out of 465 residues in one subunit in total.

In the transmembrane region this interface spans across the complete inner channel (S5, S6, pore helix and filter region), a large part of S4 and the end of S1 that points toward the extracellular space leaving only S2 and S3 without inter-subunit contact. The interaction of the monomers in the C-terminal domain is mainly formed by the C-linker: the "elbow-on-the-shoulder" motif described in Section 2.1.1 consist of the A′ and B′ helix ("elbow") resting on the C′ and D′ helix ("shoulder"). Almost all residues of these four helices are part of the interface. Only few residues in the CNBD are within reach of neighboring subunits: isolated amino acids of the A and B helix as well as parts of the β roll on the opposite side.

None of the top 20 coevolutionary links drawn in Figure 3.15 seems to originate from direct inter-subunit interaction since all links are either connecting nearby residues or they run along the (in this illustration) vertical axis of the subunit. As mentioned above, SUMI with a Z-score cutoff can only find coevolutionary links between residues with a fairly high entropy. If we want to identify coevolving pairs in lower entropy regions—which is particularly interesting in the case of MSAs with closely related sequences like ours—we need to use other normalization methods.

Detecting Coevolution in Low Entropy Positions

The minH normalization relates the MI to the minimum of the one-point entropies of the column pair under consideration and thereby the maximally possible MI value of the respective pairing (see Equation 3.9). The normalized MI values range from 0 to 1, which facilitates the extraction of coevolutionary signals from low entropy MSA columns, that would otherwise be overlooked.

As can already be seen in Figure 3.14b, many position pairs show high SUMI if minH normalization is applied. In fact, there are so many position pairs (1691) with a minH normalized SUMI value of 1 that it is not feasible to visualize all of them as links in the protein structure. 365 residues out of 465 in the protein in total participate in these coevolutionary relationships with maximal values of minH normalized SUMI. When inspecting the MSA, we found that many of these maximal values are due to only one or two amino acid changes in highly conserved columns. In general, this is not an issue, quite the contrary: these are the low entropy positions we wanted to analyze. Nevertheless, we face a problem here: similar to the ORMI being biased by stretches of gaps in only a few rows of conserved columns, SUMI is strongly influenced by stretches of random amino acids that are caused by low quality sequences in the alignment. This is most likely to be the case here. Some of the sequences in our data set perfectly match the rest of the MSA in most parts, but exhibit regions of several consecutive positions that mismatch most severely.

This effect is strongest in small data sets because here errors in only a couple of sequences make up a larger percentage of the alignment. Since a certain minimum level of variation is necessary for the existence of any coevolution, we decided to circumvent this problem by applying an entropy filter to the minH normalized SUMI values. Only positions with a one-point entropy of 0.3 or greater were considered in the following analysis.

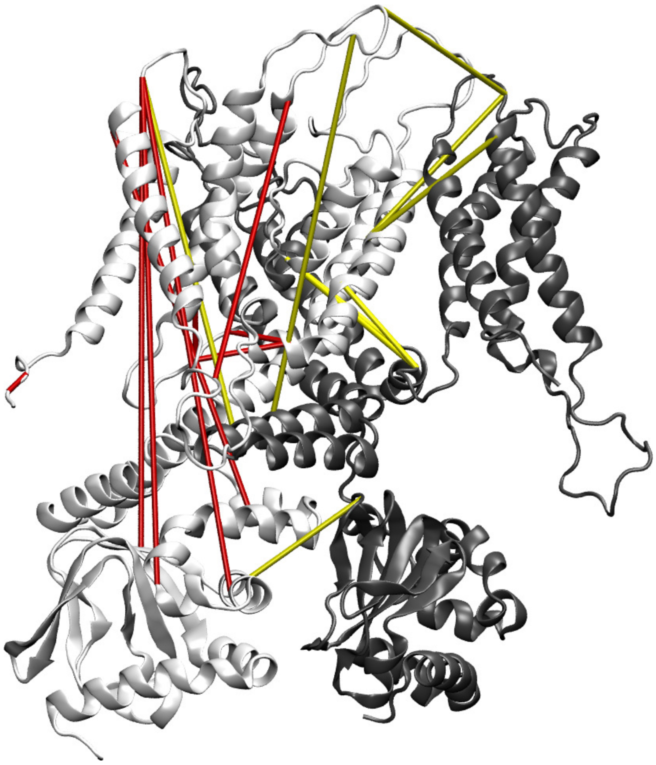


Figure 3.18: Subunit of the HCN tetramer with links of minH normalized SUMI filtered by the one-point entropy of the corresponding MSA columns. Intra-subunit interactions are shown in red, inter-subunit interactions in yellow. For guidance, the subunits are oriented in the same way as in Figure 3.17.

Only 18 interactions remain after application of the entropy filter. Figure 3.18 shows all of them as links in the structure of the HCN channel. Since all four subunits of a homotetramer are identical copies, it is not possible to distinguish between intra- and inter-subunit coevolutionary relationships because the protein is produced by the expression of only one gene. This leads to mixing of the coevolutionary constraints that originate within and between subunits. To draw the links in Figure 3.18, we therefore decided to always connect the closest representatives of each pair. All pairs which had the smaller distance within a subunit are connected by a red link, all those whose distance was smaller when connecting the corresponding amino acids in neighboring subunits are highlighted in yellow.

There are still some long-range interactions along the vertical axis. Additionally, we can see that new links appear, especially between the adjacent subunits. An interesting observation is that two of the hubs at the extracellular end of S3 (108 and 110) that were identified with the Z-score cutoff normalized SUMI are not present any more Figures 3.15 and 3.16. This means that although these particular positions have a high entropy and coevolutionary relationships with many other residues, they do not show maximal coevolution with any of them. For the hub position in S2 (38) the opposite is true: This residue located at the transition between S2 and the extracellular loop was detected by both the Z-score cutoff and the minH method (compare Figures 3.15 and 3.18).

While most short-range coevolutionary relationships simply originate from spatial proximity, the reasons for long-range interactions are more diverse. Ion channels undergo large domain movements during gating. The CNBD of the HCN channel binds cAMP which influences the gating threshold of the transmembrane domain. To this end, an allosteric conformational change is necessary. In fact, we could identify this motion to be a quaternary twist of the transmembrane and the C-terminal domains against each other, see Section 2.3.3. Residues that are located far away from each other but are under shared functional constraints have been described by others [16] and are known to be part of allosteric reactions [34]. Thus, the long-range coevolutionary relationship between the transmembrane domain and the CNBD observed in Figures 3.15 and 3.18 could arise from the allosteric function required for modulation of channel gating. Large parts of the protein participate in this process, which imposes a shared selective pressure to maintain the gating motion.

Conclusion

Pairwise intramolecular coevolution of the HCN channel was analyzed in detail. We were able to eliminate false positive signals resulting from consecutive gap regions by using SUMI, which computes the MI of a gap-free subset for each MSA column pairing.

SUMI is a promising approach to remove bias from gaps in MSAs with partial sequences. However, ignoring gaps in the analysis of coevolution is not always correct. It is possible that the deletion of a position is a way of compensating mutations in other residues and thereby a kind of coevolution. These cases are neglected when using SUMI. An accurate solution of the gap question remains to be found.

Two normalization methods allow for detection of different types of coevolving residues. The Z-score cutoff helps to find statistically significant MI signals in high entropy positions and can be used to identify hubs that exhibit coevolutionary relationships with many proteins. The minH method, on the other hand, can trace coevolution in MSA columns with low one-point entropy. However, the minH method has a major drawback: it is prone to sequencing and alignment errors that introduce the only deviation in an otherwise conserved MSA column. This effect can be compensated by applying an entropy filter to the minH normalized MI values prior to interpretation of the results.

Since both normalization methods detect a different type of coevolving pairs, the choice depends on the problem one wants to tackle. For a thorough analysis, we suggest to apply both approaches and combine the results. Nevertheless, it is always important to carefully validate the results and the quality of the underlying MSA data.

We were able to identify several long-range, coevolutionary interactions between the transmembrane region and the C-terminal domain. Additionally, links between the transmembrane helices of neighboring subunits were detected. These coevolutionary relationships most likely stabilize the allosteric reaction of the channel upon cAMP binding, which is a collective domain movement: a quaternary twist of the transmembrane region against the CNBD.

Due to the high collectivity of this twist, the relationships between pairs of residues analyzed here might not suffice to reveal all coevolutionary correlations. Section 3.5 addresses the question whether interactions between groups of residues can be detected better by higher-dimensional MI analysis.

3.5 MI of Three-Cliques in HCN Channels

This section is based on our publications "Information-theoretic analysis of molecular (co) evolution using graphics processing units" [155] and "Using graphics processing units to investigate molecular coevolution" [154] both by Wächter et al.

Coevolutionary relationships are not necessarily always pairwise. As the existence of coevolving groups suggests, there are interactions that require the coordination of several residues [63]. As early as 1993, Korber et al. found networks of residues that share high MI and suggested these originate from higher-order interactions [89]. The question is, whether these networks can be fully described by pairwise relationships between their members or whether we need a higher-order MI to reveal them. While Morcos et al. claim that MI is composable and that group interactions manifest themselves in the pairwise signal [114], this hypothesis has not been proven so far.

To answer the question we computed three-dimensional MI for MSAs of protein sequence data sets. Our study is the first analysis of higher-order relationships in intramolecular coevolution. In 2009, Weigt et al. mentioned that their method to detect coevolution could be extended to examine correlations of three or more positions, but estimated that the amount of sequence data was not sufficient for computing these correlations reliably [161]. We resolve this problem by using Z-scores to determine statistical significance. As for the two-point MI, many shuffling runs for each combination of MSA columns are necessary to compute Z-scores. While this is still feasible by conventional means when considering all combinations of *pairs*, it becomes a task with high computational costs when MI for all possible *triplets* needs to be calculated. Therefore, we employed graphics processing units (GPUs), which contain thousands of cores, to tackle this problem [119].

Computation of MI is well-suited for parallelization because each column pairing can be treated independently of the others. In addition, the shuffling runs for the computation of Z-scores can be parallelized. Exploiting the large number of cores on a GPU allows for massive parallelization and thereby facilitates the computation of three-dimensional MI. The program `CoMIC` (Coevolution via MI on CUDA) used in this study is described in detail in the aforementioned publications.

In our publications mentioned above, three-dimensional MI analysis was performed for Calmodulin, a calcium-binding messenger protein, and variable surface glycoprotein of *Trypanosoma brucei*, the human pathogen that causes sleeping sickness. In this thesis, the HCN channel will be analyzed.

Three-Dimensional Mutual Information

Similar to the two-dimensional MI (Equation 3.4) its three-dimensional counterpart can be computed as

$$M(X,Y,Z) = H(X) + H(Y) + H(Z) - H(X,Y,Z). \tag{3.15}$$

Again, just as for the two-dimensional MI, the highest possible three-dimensional MI value depends on the one-point entropies of the three MSA columns under investigation. Analogous to Equation 3.5 the lower bound of $H(X, Y, Z)$ is

$$H(X, Y, Z) \geq \max\Big(H(X), H(Y), H(Z)\Big). \tag{3.16}$$

From this it follows that the three-dimensional MI can never be greater than the sum of the two smallest elements of the set $A = \{H(X), H(Y), H(Z)\}$.

$$M(X, Y, Z) \leq \min(A) + \min\Big(A \setminus \min(A)). \tag{3.17}$$

Therefore, in contrast to two-dimensional MI, the theoretical maximum of three-dimensional MI is 8.78 when considering an MSA with an amino acid alphabet comprising 21 characters (cf. Equation 3.7).

There is another important difference between $M(X, Y)$ and $M(X, Y, Z)$: Two-dimensional MI computation yields a matrix (or two-dimensional array) that carries the one-point entropies of the respective MSA columns on its diagonal (cf. Equation 3.6). For the diagonal values of the three-dimensional MI array $M(X, X, X)$ we derive the following when inserting into Equation 3.15:

$$\begin{aligned} M(X, X, X) &= H(X) + H(X) + H(X) - \underbrace{H(X, X, X)}_{=H(X)} && (3.18) \\ &= 2\, H(X). && (3.19) \end{aligned}$$

Z-scores for three-dimensional MI are calculated in the same way as for the two-dimensional case. As mentioned above, here, the shuffling null model is computationally more expensive because all combinations of triplets need to be shuffled. From the resulting mean and standard deviation Z-scores are computed according to Equation 3.8.

To facilitate comparability of two- and three-dimensional MI despite different scales, we compute percentiles $c(X, Y)$ and $c(X, Y, Z)$ of MI. $c(X, Y)$ is the fraction of shuffled MI values $\widetilde{M}(X, Y)$ that is smaller than the MI for the original, non-shuffled MSA columns $M(X, Y)$. $c(X, Y, Z)$ is defined accordingly.

3.5.1 Methods

Plots were created using `R`; protein images were rendered in `VMD` [78].

Data Set

The sequence data set and the MSA of the HCN channel were prepared as described in Section 3.3.1. To enable comparability of the results to those from Section 3.4 and facilitate interpretation, the MSA was mapped to the sequence parts that are present in the structure of the homology model described in Section 2.3.1. Analysis was performed only for those MSA areas that have a counterpart in the model structure.

MI Computation

Three- and two-dimensional MI of the MSA were computed along with their corresponding Z-scores. 10 000 shuffling runs were performed for computation of Z-scores and percentiles c. MI values with a corresponding $Z > 4$ were considered statistically significant. Values with $Z \leq 4$ were set to zero before result evaluation. As mentioned above, computation of MI, Z-scores and percentiles was performed with the program `CoMIC` on GPUs. The algorithm is described in detail in our publications [154, 155].

3.5.2 Results and Discussion

Our aim was to find out whether all coevolutionary interactions can be described by pairwise relationships detected through two-dimensional MI analysis or whether three-dimensional MI is necessary to reveal correlations of higher order. To this end, we needed to compare the obtained MI data. Since computation of three-dimensional MI yields three-dimensional arrays, a simple plot of the entire data (as in Figure 3.14 for two-dimensional MI) is not possible.

In Figure 3.19 three-dimensional MI of residue triplets is compared to the maximum of the three pairwise MI values of the same triplet or *three-clique*. Values that featured a $Z \leq 4$ were set to zero; they cause the stretches of data points parallel to the axes. A series of data runs along the diagonal of the plot. It reflects the relation of the diagonal entries of the MI arrays $M(X, X)$ and $M(X, X, X)$: the diagonal of the two-dimensional MI array holds the one-point entropies $H(X)$, while the three-dimensional MI array diagonal elements contain $2 \times H(X)$ (see Equation 3.19). We can use this line with a slope of 0.5 as a reference for the relationship between the two MI types. For the diagonal entries of the MI arrays, the third dimension cannot provide any additional information about the MSA column. All information about column X is already given by the column itself, it is irrelevant whether we observe a third, identical column X. Therefore, the line represents the case in which three-dimensional MI yields the same amount of information as the two-dimensional MI.

As we can see in Figure 3.19, there are three-cliques of residues on both sides of this boundary. The data in the lower right triangle are three-cliques that show greater three-dimensional MI values than can be explained by the two-dimensional MI between all pairs of the three MSA columns. Thus, higher-order correlations cannot be completely revealed by examining pairwise, two-dimensional MI.

To facilitate better comparability of two- and three-dimensional MI, we use MI percentiles. $c(X, Y)$ or $c(X, Y, Z)$ is defined as the fraction of shuffled MI values from the null model that is smaller than the MI for the original column pairing (X, Y) or (X, Y, Z), respectively. Therefore, naturally, MI percentiles lie within a range of $[0, 1]$. Figure 3.20 shows the maximum of the pairwise percentiles in three-cliques plotted against their corresponding three-dimensional MI percentile. Based on this picture, we classify three-cliques into two groups: One group contains all triplets with $c(X, Y, Z)$

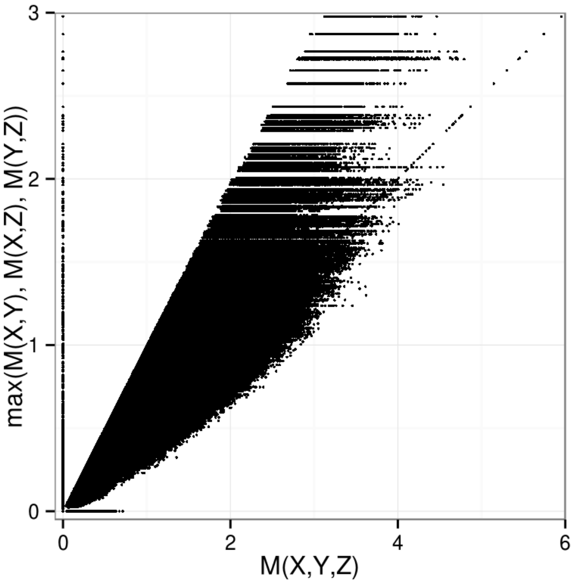


Figure 3.19: Maximum of two-dimensional MI values of a three-clique plotted against the respective three-dimensional MI. All MI values with a corresponding $Z \leq 4$ were set to zero before plotting, which causes the stretches of zeros parallel to the axes.

greater than the maximum of their two-dimensional MI percentiles. From this group, we define the set $\mathcal{C}$:

$$\mathcal{C} = \{(X, Y, Z)|\max(c(X, Y), c(X, Z), c(Y, Z)) \leq c(X, Y, Z)\}. \tag{3.20}$$

The second group comprises all three-cliques for which the opposite is true.

Three-cliques that belong to the set $\mathcal{C}$ show more three-dimensional MI than can be explained from the pairwise relationships. This signal originates from higher-order correlations and suggests that coordination of functional units is supported by coevolution which cannot be detected by two-dimensional MI. This finding contradicts the assumption that MI is composable as Morcos et al. hypothesized [114].

To further analyze these three-cliques with higher-order coevolution, Figure 3.21 was created. Each residue in the protein structure is colored by the frequency with which it contributes to the three-cliques of set $\mathcal{C}$. In addition, the ten residues with the highest frequency are marked as beads in the 3D structure. Here, we can see that higher-order correlations are much more abundant in the C-terminal domain than in the transmembrane region.

At first, this seems rather surprising since we know that strong coevolutionary signals are present in all regions of the protein (cf. Figures 3.15 and 3.18). However, we have to keep in mind that for Figure 3.21 only instances with two-dimensional MI percentile values less than the three-dimensional MI percentile were counted. Therefore, blue color of a residue does not necessarily imply low three-dimensional MI. It just shows that the MI of the residue's three-clique interactions is not greater than that of the corresponding

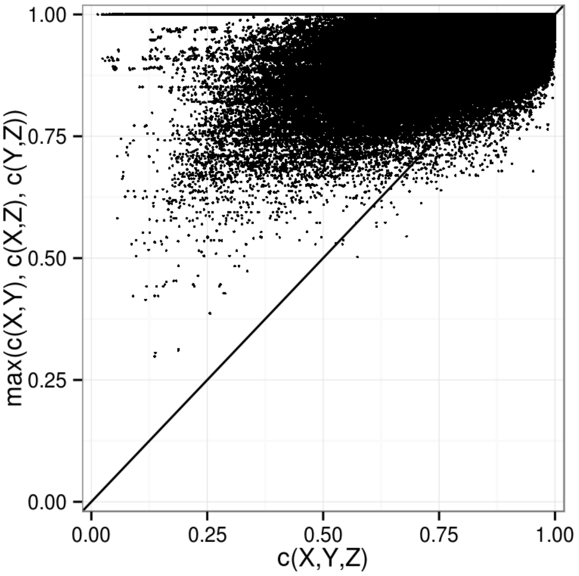
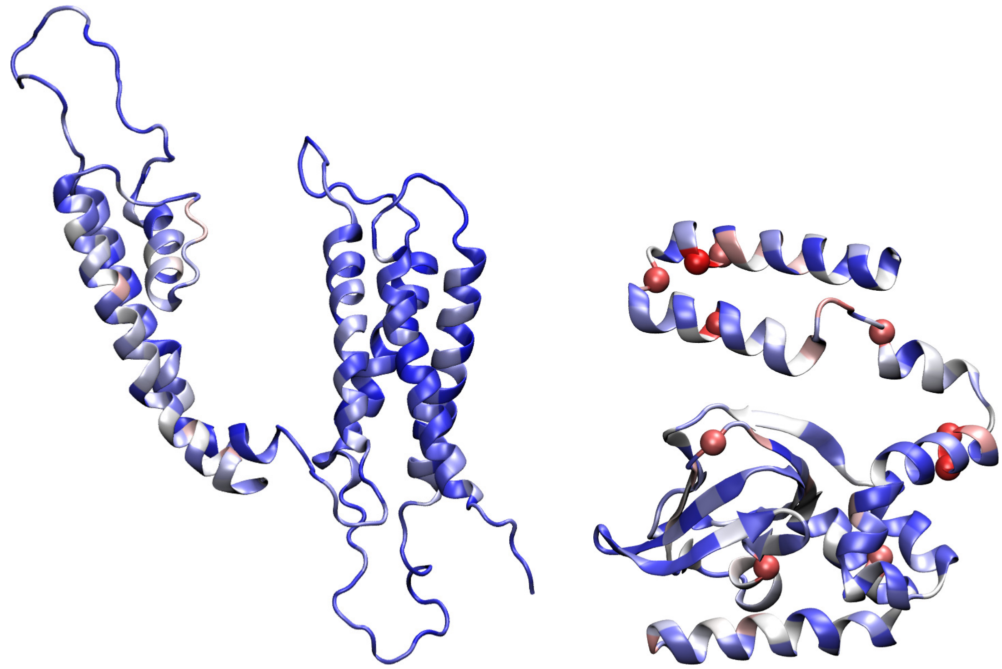


Figure 3.20: Maximum of percentile-normalized two-dimensional MI of a three-clique plotted against the respective percentile-normalized three-dimensional MI. Data below the diagonal line belong to set $\mathcal{C}$ (see Equation 3.20).

pairwise interactions. We chose this way of illustration because we want to focus on coevolutionary relationships which cannot be detected by two-dimensional MI.

The residues depicted as beads in Figure 3.21 are also listed in Table 3.4 (see Appendix A for numbering of residues). We can see a hot spot of three-clique coevolution in the area of the turn between the A′ and B′ helices in the C-linker (residues 284, 285, 290, 298). Effectively, residue 312 belongs to this area as well since the HCN channel is a homotetramer. As described in Section 2.1.1, the turn between the A′ and B′ helices rests on the C′ helix. Thereby, residue 312 of one subunit is in contact with the neighboring subunit's A′–B′ helix turn mentioned above. Besides being part of the inter-subunit interface, the C-linker plays an important role in the allosteric reaction upon cAMP binding. It is the connection of the C-terminal domain, which contains the cAMP binding pocket, to the transmembrane region, where voltage sensing, channel gating and ion conduction occur. Many residues participate in this connecting function. This requires higher-order interactions and thereby imposes a common selective pressure, which causes group coevolution. The three-dimensional MI signal for the involved residues is therefore greater than pairwise, two-dimensional MI.

Two of the residues that contribute the most to three-clique coevolution are located in the D′ helix, another can be found in the B helix. All of these three residues interact with neighboring subunits. Thus, eight out of the top ten three-clique coevolving residues can be assigned to the inter-subunit interface region. The remaining two lie in the turn between β7 and β8 and in the P helix, which constitutes part of the cAMP binding pocket.



(a) Transmembrane domain. **(b)** C-terminal domain.

Figure 3.21: HCN4 channel domains colored by the frequency with which they contribute to the set $\mathcal{C}$ of three-cliques (blue: low values; red: high values). Additionally, the ten residues with the highest values are shown as spheres; all of them are located in the C-terminal domain. For better visibility, the two domains are shown separately.

Residue number	Residue type	Location
284	M	A′ (C-linker)
285	E	A′ (C-linker)
290	R	Turn between A′ and B′ (C-linker)
298	I	B′ (C-linker)
312	L	C′ (C-linker)
325	E	D′ (C-linker)
326	D	D′ (C-linker)
407	T	P (CNBD)
424	Y	Turn between β7 and β8 (CNBD)
439	V	B (CNBD)

Table 3.4: Residues appearing most frequently in three-cliques belonging to set $\mathcal{C}$ (for definition see Equation 3.20). These amino acids are marked by spheres in Figure 3.21.

There is an interesting contrast between the pairwise coevolutionary relationships identified in Section 3.4 and the three-clique coevolution found here: The dominant pairwise interactions were found between the transmembrane and the C-terminal domain. Although some short-range links were detected, long-range interactions prevailed (see Figure 3.18). Furthermore, important hub residues, which share high coevolution with many other positions, were present in both domains. Strong coevolution of three-cliques, on the other hand, is restricted to the C-terminal domain and can be found mainly at the interface between subunits of the homotetramer.

While conduction of the allosteric change from the CNBD to the transmembrane region certainly also requires group interaction, we were able to detect the resulting coevolutionary network by simply analyzing two-dimensional MI. This is probably due to strong pairwise coupling being present in addition to higher-order communication. In contrast to this, coevolution at the interface between subunits was not detectable with two-dimensional MI. Only three-dimensional MI could reveal these inter-subunit coevolutionary relationships.

This observation matches our findings regarding *Trypanosoma brucei*'s variable surface glycoprotein (VSG) [154]—a homodimer. Its two- and three-dimensional MI were computed; the analysis was performed in a similar fashion as for HCN in this study. We found significant three-clique coevolution for VSG at the interface between the two monomers. Just like the HCN channel, VSG is only functional in its oligomeric state. Therefore, maintaining the interactions that link the subunits is of vital importance. The fact that these interactions can be detected better by three-dimensional MI suggests a binding mechanism that is more sophisticated than simple pairwise matching.

Conclusion

We studied coevolutionary relationships of three-cliques of residues in the HCN channel. In this thesis, as well as in the underlying studies on Calmodulin and VSG [154, 155], we could answer the question that was asked at the beginning of this section: is it possible to fully describe group coevolution with pairwise relationships between their members or do we need a higher-order MI to reveal them?

The answer is that, indeed, some events of group correlation cannot be detected by analyzing pairwise MI. To identify these functional groups, correlations of higher order have to be examined. Here, we found a coevolutionary interaction network located at the interface between neighboring subunits in the homotetrameric HCN channel applying three-dimensional MI analysis. This is in agreement with a finding from our previous study regarding VSG: the binding of the subunits of an oligomer to each other seems to be a collective effort of the interface residues, which is maintained by group coevolution—and thus detectable by higher-order MI analysis.

Our studies show the need for higher-order correlation measures, such as three-dimensional MI, to identify coevolution within functional groups. However, similar to the question whether three-dimensional MI helps to reveal new insight, one might even go one step further: what would happen if we analyzed coevolution of four-cliques, five-cliques and beyond? At the moment, limited computational power prohibits such experiments, but an increasing degree of parallelism might pave the way for interesting new analyses in the future [6].

3.6 Conclusion

Intramolecular coevolution is a mechanism that compensates detrimental mutations in a protein. Analyzing correlations between positions in a multiple sequence alignment can detect these coevolutionary relationships and thereby reveal interactions between pairs and groups of residues in the protein structure.

In this chapter, we examined coevolution of the HCN channel as well as other proteins by means of measuring mutual information between MSA columns. Prior to that, careful manual curation of the data sets was necessary because quality of some sequences in the databases is low and alignment algorithms sometimes produce inadequate results. Furthermore, MI needs to be normalized and tested for statistical significance before results can be analyzed. To this end, we employed the methods of Z-score cutoff and minimal entropy normalization. An alternative handling of gaps is offered by the computation method of subset MI (SUMI). This approach helped to clear artifacts that originated from low quality sequences in the MSA.

Coevolution was analyzed on three different levels: First, we analyzed the properties of coevolutionary graphs built from MI data and found that MI graphs of all proteins under investigation resembled typical small-world networks. The second analysis was a more detailed view of pairwise coevolutionary relationships in the HCN channel. It revealed important long-range interactions between the two major domains: the transmembrane region and the C-terminus. These interactions could be attributed to preservation of the quaternary twist, a motion inherent in many channel proteins as part of the gating mechanism. Third and finally, we investigated group coevolution in HCN. We were able to demonstrate that some higher-order correlations can only be detected by three-dimensional MI. By this means, we could show that residues at the interface between tetramer subunits participate in group coevolution, which suggests a complex binding mechanism.

4 Final Conclusion

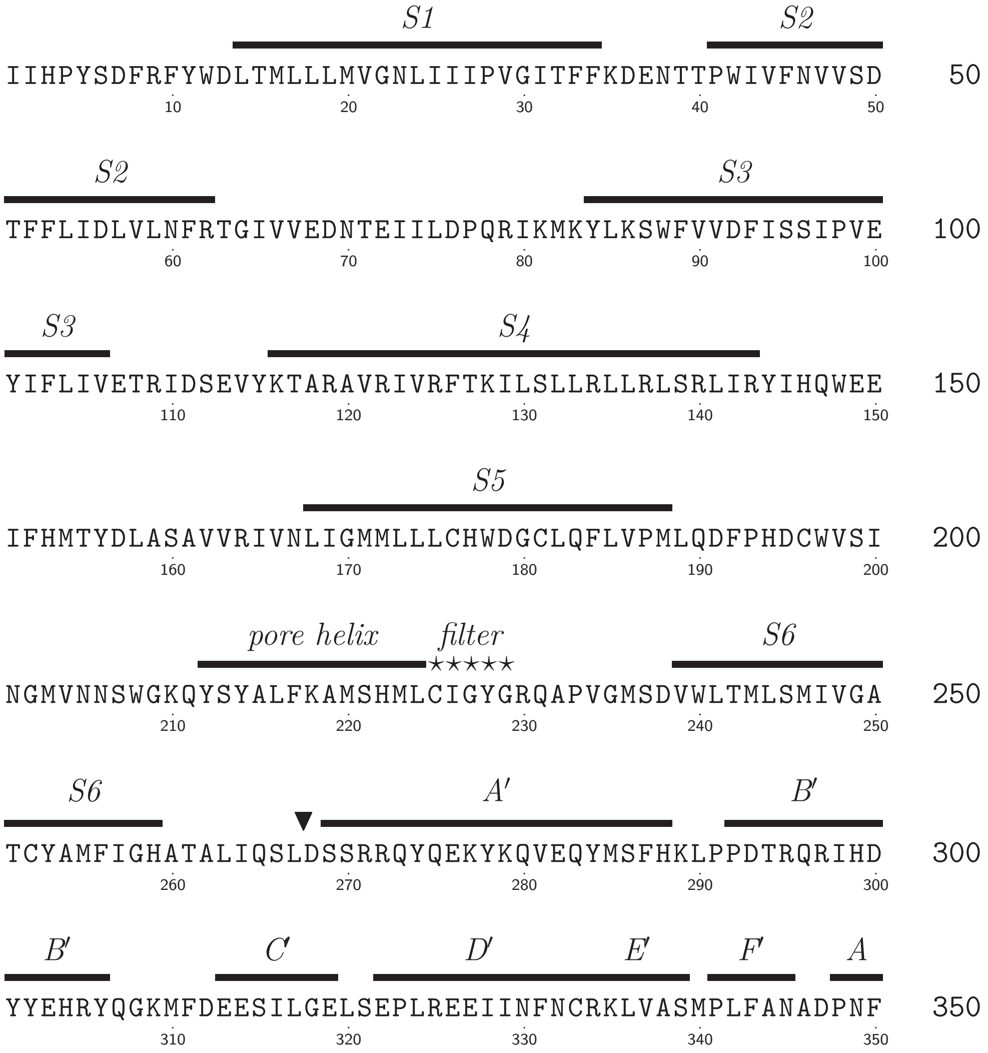
Since no crystal structure of the cyclic nucleotide-binding domain of HCN in its ligand-free state has been published up to now, we developed an *in silico* model of this domain. It was joined to a homology model of the HCN channel's transmembrane region. Using an elastic network approach combined with linear response theory, we could simulate the release of cAMP from the binding pocket. A quaternary twist motion of the four subunits, that was already identified as part of the voltage-gating mechanism of potassium channels, turned out to be also involved in HCN modulation by cAMP. We detected several contacts that play an important role in this process through a switch-off screening. Most of the relevant interactions were found in the transmembrane region, especially between helices of different subunits. Intra-subunit contacts between the S4–S5-linker and the C-linker were also found among these key players. These two structural elements have previously been shown to interact. Our results confirm this interaction and suggest that it is involved in the coupling of cAMP modulation and voltage-dependent gating.

We also found hints indicating allosteric domain rearrangements in the intramolecular coevolution of HCN channels: A detailed analysis of pairwise mutual information revealed long-range coevolutionary links between the transmembrane region and the C-terminal domain. Interactions of shorter range were detected between adjacent transmembrane helices of neighboring subunits. Since pairwise relationships cannot fully capture higher-order correlations, three-dimensional MI was computed as well. Hot spots of group coevolution were found at the interface between subunits of the HCN channel tetramer.

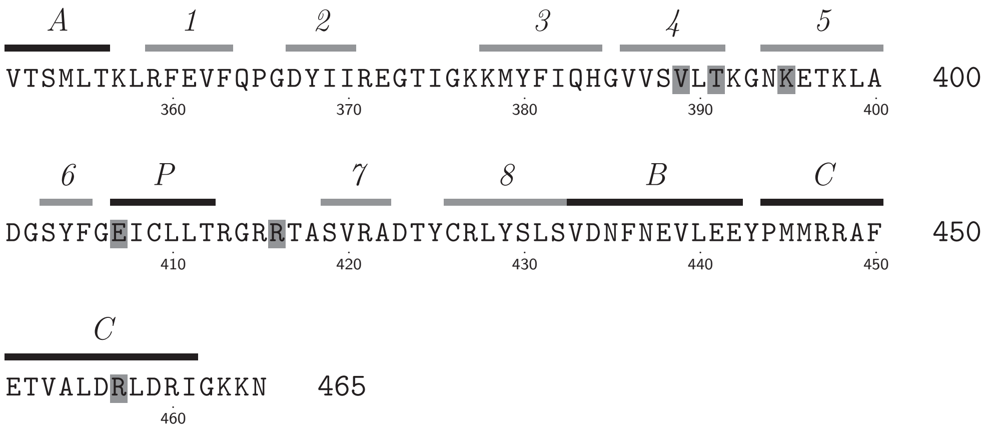
Our results contribute to the understanding of HCN channel activation: cAMP binding starts a quaternary twist of the four subunits. This domain rearrangement has been shown to be the opening mechanism of related channels, which is most likely also the case for HCN. The global allosteric reaction upon cAMP binding could be shown on the structural level and leaves its footprint in the coevolutionary patterns of the protein. The combination of structure- and sequence-based approaches revealed not only the nature of the conformational change but also identified participating residues. This insight paves the way for future studies on structure and mechanics of HCN channels including their role as a drug target.

A Sequence of the HCN4 Homology Model

Illustration of the sequence part of HCN4 that is present in the homology model. Secondary structure elements are marked with bars (black for α helices, gray for β strands) and labeled. The selectivity filter region is highlighted with stars and the cAMP binding residues are shaded in gray. The black triangle at residue number 268 marks the start of the crystal structure 3U11.



```
IIHPYSDFRFYWDLTMLLLMVGNLIIIPVGITFFKDENTTPWIVFNVVSD    50
TFFLIDLVLNFRTGIVVEDNTEIILDPQRIKMKYLKSWFVVDFISSIPVE   100
YIFLIVETRIDSEVYKTARAVRIVRFTKILSLLRLLRLSRLIRYIHQWEE   150
IFHMTYDLASAVVRIVNLIGMMLLLCHWDGCLQFLVPMLQDFPHDCWVSI   200
NGMVNNSWGKQYSYALFKAMSHMLCIGYGRQAPVGMSDVWLTMLSMIVGA   250
TCYAMFIGHATALIQSLDSSRRQYQEKYKQVEQYMSFHKLPPDTRQRIHD   300
YYEHRYQGKMFDEESILGELSEPLREEIINFNCRKLVASMPLFANADPNF   350
```


A
1
2
3
4
5
VTSMLTKLRFEVFQPGDYIIREGTIGKKMYFIQHGVVSVLTKGNKETKLA
360
370
380
390
400
400
6
P
7
8
B
C
DGSYFGEICLLTRGRRTASVRADTYCRLYSLSVDNFNEVLEEYPMMRRAF
410
420
430
440
450
450
C
ETVALDRLDRIGKKN
460
465

B List of Abbreviations

arnt aryl hydrocarbon receptor nuclear translocator protein

CaM Calmodulin

cAMP cyclic adenosine monophosphate

c-di-GMP cyclic di-guonosine monophosphate

CNBD cyclic nucleotide-binding domain

CNG cyclic nucleotide-gated (channel)

CoMIC coevolution via MI on CUDA

HCN hyperpolarization-activated cyclic nucleotide-gated (channel)

hERG human ether-a-go-go related gene

LRT linear response theory

MI mutual information

minH minimal entropy (normalization)

MSA multiple sequence alignment

nj neighbor-joining

ORMI original mutual information

PAS per-arnt-sim (domain)

per periodic circadian protein

sim single minded protein

SUMI subset mutual information

TM transmembrane

upgma unweighted pair group method with arithmetic mean

Bibliography

[1] Abe S, Suzuki N (**2004**) Small-world structure of earthquake network. *Physica A: Statistical Mechanics and its Applications* 337(1):357–362

[2] Alam A, Jiang Y (**2009**) High-resolution structure of the open NaK channel. *Nature Structural & Molecular Biology* 16(1):30–34

[3] Alberts B (**1998**) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* 92(3):291–294

[4] Altieri SL, Clayton GM, Silverman WR, Olivares AO, De la Cruz EM, Thomas LR, Morais-Cabral JH (**2008**) Structural and energetic analysis of activation by a cyclic nucleotide binding domain. *Journal of Molecular Biology* 381(3):655–669

[5] Arnold K, Bordoli L, Kopp J, Schwede T (**2006**) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22(2):195–201

[6] Asanovic K, Bodik R, Catanzaro BC, Gebis JJ, Husbands P, Keutzer K, Patterson DA, Plishker WL, Shalf J, Williams SW, et al. (**2006**) The landscape of parallel computing research: A view from Berkeley. *Tech. rep.*, Technical Report UCB/EECS-2006-183, EECS Department, University of California, Berkeley

[7] Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (**2000**) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Molecular Biology and Evolution* 17(1):164–178

[8] Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I (**2001**) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal* 80(1):505–515

[9] Bahar I (**2010**) On the functional significance of soft modes predicted by coarse-grained models for membrane proteins. *The Journal of General Physiology* 135(6):563–573

[10] Bahar I, Atilgan AR, Erman B (**1997**) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design* 2(3):173–181

[11] Bahar I, Chennubhotla C, Tobi D (**2007**) Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Current Opinion in Structural Biology* 17(6):633–640

[12] Bahar I, Lezon TR, Bakan A, Shrivastava IH (**2010**) Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chemical Reviews* 110(3):1463–1497

[13] Bahar I, Lezon TR, Yang LW, Eyal E (**2010**) Global dynamics of proteins: bridging between structure and function. *Annual Review of Biophysics* 39:23–42

[14] Barabási AL (**2002**) *Linked: How everything is connected to everything else and what it means.* Plume Editors

[15] Barabási AL, Oltvai ZN (**2004**) Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* 5(2):101–113

[16] Baussand J, Carbone A (**2009**) A combinatorial approach to detect coevolved amino acid networks in protein families of variable divergence. *PLoS Computational Biology* 5(9):e1000488

[17] Beaumont V, Zucker RS (**2000**) Enhancement of synaptic transmission by cyclic AMP modulation of presynaptic I_{h} channels. *Nature Neuroscience* 3(2):133–141

[18] Benkert P, Biasini M, Schwede T (**2011**) Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 27(3):343–350

[19] Benkert P, Künzli M, Schwede T (**2009**) QMEAN server for protein model quality estimation. *Nucleic Acids Research* 37(suppl 2):W510–W514

[20] Benndorf K, Kusch J, Schulz E (**2012**) Probability fluxes and transition paths in a Markovian model describing complex subunit cooperativity in HCN2 channels. *PLoS Computational Biology* 8(10):e1002721

[21] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (**2010**) GenBank. *Nucleic Acids Research* 38:46–51

[22] Berman HM, Ten Eyck LF, Goodsell DS, Haste NM, Kornev A, Taylor SS (**2005**) The cAMP binding domain: an ancient signaling module. *Proceedings of the National Academy of Sciences of the United States of America* 102(1):45–50

[23] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (**2000**) The Protein Data Bank. *Nucleic Acids Research* 28:235–242

[24] Biel M, Wahl-Schott C, Michalakis S, Zong X (**2009**) Hyperpolarization-activated cation channels: from genes to function. *Physiological Reviews* 89(3):847–885

[25] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (**2006**) Complex networks: Structure and dynamics. *Physics Reports* 424(4):175–308

[26] Bolsover SR, Hyams JS, Shephard EA, White HA, Wiedemann CG (**2004**) *Cell biology: A short course.* John Wiley & Sons, Inc., Hoboken, New Jersey

[27] Brooks B, Karplus M (**1983**) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences of the United States of America* 80(21):6571–6575

[28] Brown CA, Brown KS (**2010**) Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my! *PLoS One* 5(6):e10779

[29] Burger L, van Nimwegen E (**2010**) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Computational Biology* 6(1):e1000633

[30] Buslje CM, Santos J, Delfino JM, Nielsen M (**2009**) Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics* 25(9):1125–1131

[31] Buslje CM, Teppa E, Di Doménico T, Delfino JM, Nielsen M (**2010**) Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. *PLoS Computational Biology* 6(11):e1000978

[32] Cao-Ehlker X, Zong X, Hammelmann V, Gruner C, Fenske S, Michalakis S, Wahl-Schott C, Biel M (**2013**) Up-regulation of hyperpolarization-activated cyclic nucleotide-gated channel 3 (HCN3) by specific interaction with K+ channel tetramerization domain-containing protein 3 (KCTD3). *Journal of Biological Chemistry* 288(11):7580–7589

[33] Carrington PJ, Scott J, Wasserman S (editors) (**2005**) *Models and methods in social network analysis.* Cambridge University Press

[34] Chakrabarti S, Panchenko AR (**2009**) Coevolution in defining the functional specificity. *Proteins: Structure, Function, and Bioinformatics* 75(1):231–240

[35] Chakrabarti S, Panchenko AR (**2010**) Structural and functional roles of coevolved sites in proteins. *PLoS One* 5(1):e8591

[36] Chen J, Mitcheson JS, Tristani-Firouzi M, Lin M, Sanguinetti MC (**2001**) The S4–S5 linker couples voltage sensing and activation of pacemaker channels. *Proceedings of the National Academy of Sciences of the United States of America* 98(20):11277–11282

[37] Clayton GM, Silverman WR, Heginbotham L, Morais-Cabral JH (**2004**) Structural basis of ligand activation in a cyclic nucleotide regulated potassium channel. *Cell* 119(5):615–627

[38] Codoñer FM, Fares MA (**2008**) Why should we care about molecular coevolution? *Evolutionary Bioinformatics Online* 4:29–38

[39] Codoñer FM, O'Dea S, Fares MA (**2008**) Reducing the false positive rate in the non-parametric analysis of molecular coevolution. *BMC Evolutionary Biology* 8:106

[40] Csardi G, Nepusz T (**2006**) The igraph software package for complex network research. *InterJournal* Complex Systems:1695

[41] Cui Q, Bahar I (**2005**) *Normal mode analysis: theory and applications to biological and chemical systems.* CRC press

[42] de Juan D, Pazos F, Valencia A (**2013**) Emerging methods in protein co-evolution. *Nature Reviews Genetics* 14(4):249–261

[43] Decher N, Chen J, Sanguinetti MC (**2004**) Voltage-dependent gating of hyperpolarization-activated, cyclic nucleotide-gated pacemaker channels molecular coupling between the S4–S5 and C-linkers. *Journal of Biological Chemistry* 279(14):13859–13865

[44] Diestel R (**2010**) *Graph Theory (Graduate Texts in Mathematics 173).* 4th edn., Springer, Heidelberg

[45] DiFrancesco D (**1981**) A study of the ionic nature of the pace-maker current in calf Purkinje fibres. *The Journal of Physiology* 314(1):377–393

[46] DiFrancesco D, Tortora P (**1991**) Direct activation of cardiac pacemaker channels by intracellular cyclic AMP. *Nature* 351(6322):145–147

[47] Doan T, Kunze D (**1999**) Contribution of the hyperpolarization-activated current to the resting membrane potential of rat nodose sensory neurons. *The Journal of Physiology* 514(1):125–138

[48] Doruker P, Jernigan RL, Bahar I (**2002**) Dynamics of large proteins through hierarchical levels of coarse-grained structures. *Journal of Computational Chemistry* 23(1):119–127

[49] Doyle DA, Morais Cabral J, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R (**1998**) The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science* 280(5360):69–77

[50] Dunn SD, Wahl LM, Gloor GB (**2008**) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24(3):333

[51] Edgar RC (**2004**) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1):113

[52] Edgar RC (**2004**) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5):1792–1797

[53] Emery EC, Young GT, McNaughton PA (**2012**) HCN2 ion channels: an emerging role as the pacemakers of pain. *Trends in Pharmacological Sciences* 33(8):456–463

[54] Erdős P, Rényi A (**1960**) On the evolution of random graphs. *Magyar Tudományos Akadémia Matematikai Kutatóintézetének Közleményei* 5:17–61

[55] Fatakia SN, Costanzi S, Chow CC (**2009**) Computing highly correlated positions using mutual information and graph theory for G protein-coupled receptors. *PLoS One* 4(3):e4681

[56] Feynman RP, Leighton RB, Sands ML (**1963**) *Mainly mechanics, radiation, and heat*, vol. 1. Addison Wesley Publishing Company

[57] Fitch W (**1971**) Rate of change of concomitantly variable codons. *Journal of Molecular Evolution* 1(1):84–96

[58] Freeman LC (**1977**) A set of measures of centrality based on betweenness. *Sociometry* 35–41

[59] Freeman LC (**1979**) Centrality in social networks conceptual clarification. *Social Networks* 1(3):215–239

[60] Gazzarrini S, Severino M, Lombardi M, Morandi M, DiFrancesco D, Van Etten JL, Thiel G, Moroni A (**2003**) The viral potassium channel Kcv: structural and functional features. *FEBS Letters* 552(1):12–16

[61] Gianni S, Haq SR, Montemiglio LC, Jürgens MC, Engström Å, Chi CN, Brunori M, Jemth P (**2011**) Sequence-specific long range networks in PSD-95/discs large/ZO-1 (PDZ) domains tune their binding selectivity. *Journal of Biological Chemistry* 286(31):27167–27175

[62] Giorgetti A, Carloni P, Mistrik P, Torre V (**2005**) A homology model of the pore region of HCN channels. *Biophysical Journal* 89(2):932–944

[63] Gloor GB, Martin LC, Wahl LM, Dunn SD (**2005**) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* 44(19):7156–7165

[64] Golub G, Kahan W (**1965**) Calculating the singular values and pseudo-inverse of a matrix. *SIAM Journal on Numerical Analysis* 2(2):205–224

[65] Gouveia-Oliveira R, Pedersen AG (**2007**) Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms for Molecular Biology* 2(1):12

[66] Guex N, Peitsch MC, Schwede T (**2009**) Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* 30(S1):S162–S173

[67] Haliloglu T, Ben-Tal N (**2008**) Cooperative transition between open and closed conformations in potassium channels. *PLoS Computational Biology* 4(8):e1000164

[68] Hamacher K (**2008**) Relating sequence evolution of HIV1-protease to its underlying molecular mechanics. *Gene* 422(1-2):30–36

[69] Hamacher K (**2011**) Free energy of contact formation in proteins: efficient computation in the elastic network approximation. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics* 84(1 Pt 2):016703

[70] Hamacher K, McCammon JA (**2006**) Computing the amino acid specificity of fluctuations in biomolecular systems. *Journal of Chemical Theory and Computation* 2(3):873–878

[71] Hansen J, McDonald I (**1986**) *Theory of Simple Liquids, 2nd.* Academic, New York

[72] Heinig M, Frishman D (**2004**) STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research* 32(suppl 2):W500–W502

[73] Herrmann S, Stieber J, Ludwig A (**2007**) Pathophysiology of HCN channels. *Pflügers Archiv – European Journal of Physiology* 454(4):517–522

[74] Heß M, Bremm S, Weißgraeber S, Hamacher K, Goesele M, Wiemeyer J, von Landesberger T (**2014**) Visual exploration of parameter influence on phylogenetic trees. *IEEE Computer Graphics and Applications* 34(2):48–56

[75] Hille B (**2001**) *Ion channels of excitable membranes.* Sinauer Sunderland, MA

[76] Hinsen K (**1998**) Analysis of domain motions by approximate normal mode calculations. *Proteins: Structure, Function, and Genetics* 33(3):417–429

[77] Hoffgaard F, Weil P, Hamacher K (**2010**) BioPhysConnectoR: Connecting sequence information and biophysical models. *BMC Bioinformatics* 11(1):199

[78] Humphrey W, Dalke A, Schulten K (**1996**) VMD: visual molecular dynamics. *Journal of Molecular Graphics* 14(1):33–38

[79] Ikeguchi M, Ueno J, Sato M, Kidera A (**2005**) Protein structural change upon ligand binding: linear response theory. *Physical Review Letters* 94(7):078102

[80] Jänich K (**2008**) *Lineare Algebra.* Springer

[81] Jensen MØ, Jogini V, Borhani DW, Leffler AE, Dror RO, Shaw DE (**2012**) Mechanism of voltage gating in potassium channels. *Science* 336(6078):229–233

[82] Jiang Y, Lee A, Chen J, Cadene M, Chait BT, MacKinnon R (**2002**) Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* 417(6888):515–522

[83] Jiang Y, Lee A, Chen J, Cadene M, Chait BT, MacKinnon R (**2002**) The open pore conformation of potassium channels. *Nature* 417(6888):523–526

[84] Kamisetty H, Ovchinnikov S, Baker D (**2013**) Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America* 110(39):15674–15679

[85] Kaupp UB, Seifert R (**2002**) Cyclic nucleotide-gated ion channels. *Physiological Reviews* 82(3):769–824

[86] Kemena C, Notredame C (**2009**) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* 25(19):2455–2465

[87] Kimura M (**1984**) *The neutral theory of molecular evolution*. Cambridge University Press

[88] Knop GC, Seeliger MW, Thiel F, Mataruga A, Kaupp UB, Friedburg C, Tanimoto N, Müller F (**2008**) Light responses in the mouse retina are prolonged upon targeted deletion of the HCN1 channel gene. *European Journal of Neuroscience* 28(11):2221–2230

[89] Korber BT, Farber RM, Wolpert DH, Lapedes AS (**1993**) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences of the United States of America* 90(15):7176–7180

[90] Kowal J, Chami M, Baumgartner P, Arheit M, Chiu PL, Rangl M, Scheuring S, Schröder GF, Nimigean CM, Stahlberg H (**2014**) Ligand-induced structural changes in the cyclic nucleotide-modulated potassium channel MloK1. *Nature Communications* 5

[91] Kullback S, Leibler RA (**1951**) On information and sufficiency. *The Annals of Mathematical Statistics* 79–86

[92] Kuo A, Gulbis JM, Antcliff JF, Rahman T, Lowe ED, Zimmer J, Cuthbertson J, Ashcroft FM, Ezaki T, Doyle DA (**2003**) Crystal structure of the potassium channel KirBac1.1 in the closed state. *Science* 300(5627):1922–1926

[93] Kusch J, Thon S, Schulz E, Biskup C, Nache V, Zimmer T, Seifert R, Schwede F, Benndorf K (**2012**) How subunits cooperate in cAMP-induced activation of homotetrameric HCN2 channels. *Nature Chemical Biology* 8(2):162–169

[94] Kwan DC, Prole DL, Yellen G (**2012**) Structural changes during HCN channel gating defined by high affinity metal bridges. *The Journal of General Physiology* 140(3):279–291

[95] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R (**2007**) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947–2948

[96] Larsson HP (**2010**) How is the heart rate regulated in the sinoatrial node? Another piece to the puzzle. *The Journal of General Physiology* 136(3):237–241

[97] Lawson CL, Swigon D, Murakami KS, Darst SA, Berman HM, Ebright RH (**2004**) Catabolite activator protein: DNA binding and transcription activation. *Current Opinion in Structural Biology* 14(1):10–20

[98] Leresche N, Jassik-Gerschenfeld D, Haby M, Soltesz I, Crunelli V (**1990**) Pacemaker-like and other types of spontaneous membrane potential oscillations of thalamocortical cells. *Neuroscience Letters* 113(1):72–77

[99] Levitt M (**1983**) Molecular dynamics of native protein. I: Computer simulation of trajectories. *Journal of Molecular Biology* 168(3):595–620

[100] Little DY, Chen L (**2009**) Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution. *PLoS One* 4(3):e4762

[101] Lolicato M, Nardini M, Gazzarrini S, Möller S, Bertinetti D, Herberg FW, Bolognesi M, Martin H, Fasolini M, Bertrand JA, Arrigoni C, Thiel G, Moroni A (**2011**) Tetramerization dynamics of C-terminal domain underlies isoform-specific cAMP gating in hyperpolarization-activated cyclic nucleotide-gated channels. *Journal of Biological Chemistry* 286(52):44811–44820

[102] Löytynoja A, Goldman N (**2008**) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320(5883):1632–1635

[103] Lu M, Ma J (**2005**) The role of shape in determining molecular motions. *Biophysical Journal* 89(4):2395–2401

[104] Ludwig A, Budde T, Stieber J, Moosmang S, Wahl C, Holthoff K, Langebartels A, Wotjak C, Munsch T, Zong X, et al. (**2003**) Absence epilepsy and sinus dysrhythmia in mice lacking the pacemaker channel HCN2. *The EMBO Journal* 22(2):216–224

[105] MacKay DJC (**2003**) *Information theory, inference, and learning algorithms.* Cambridge University Press

[106] MacKinnon R (**2003**) Potassium channels. *FEBS letters* 555(1):62–65

[107] Macri V, Nazzari H, McDonald E, Accili EA (**2009**) Alanine scanning of the S6 segment reveals a unique and cAMP-sensitive association between the pore and voltage-dependent opening in HCN channels. *Journal of Biological Chemistry* 284(23):15659–15667

[108] Männikkö R, Elinder F, Larsson HP (**2002**) Voltage-sensing mechanism is conserved among ion channels gated by opposite voltages. *Nature* 419(6909):837–841

[109] Marchiori M, Latora V (**2000**) Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications* 285(3):539–546

[110] Marques O, Sanejouand YH (**1995**) Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins: Structure, Function, and Bioinformatics* 23(4):557–560

[111] Martin LC, Gloor GB, Dunn SD, Wahl LM (**2005**) Using information theory to search for co-evolving residues in proteins. *Bioinformatics* 21(22):4116–4124

[112] Miyazawa S, Jernigan RL (**1985**) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 18(3):534–552

[113] Moore E (**1920**) On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society* 26:394–395

[114] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (**2011**) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America* 108(49):E1293–E1301

[115] Müller T, Spang R, Vingron M (**2002**) Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Molecular Biology and Evolution* 19(1):8–13

[116] Nimigean CM, Shane T, Miller C (**2004**) A cyclic nucleotide modulated prokaryotic K+ channel. *The Journal of General Physiology* 124(3):203–210

[117] Nolan MF, Malleret G, Dudman JT, Buhl DL, Santoro B, Gibbs E, Vronskaya S, Buzsáki G, Siegelbaum SA, Kandel ER, et al. (**2004**) A behavioral role for dendritic integration: HCN1 channels constrain spatial memory and plasticity at inputs to distal dendrites of CA1 pyramidal neurons. *Cell* 119(5):719–732

[118] Nolan MF, Malleret G, Lee KH, Gibbs E, Dudman JT, Santoro B, Yin D, Thompson RF, Siegelbaum SA, Kandel ER, et al. (**2003**) The hyperpolarization-activated HCN1 channel is important for motor learning and neuronal integration by cerebellar Purkinje cells. *Cell* 115(5):551–564

[119] Owens JD, Houston M, Luebke D, Green S, Stone JE, Phillips JC (**2008**) GPU computing. *Proceedings of the IEEE* 96(5):879–899

[120] Passner JM, Schultz SC, Steitz TA (**2000**) Modeling the cAMP-induced allosteric transition using the crystal structure of CAP-cAMP at 2.1 Å resolution. *Journal of Molecular Biology* 304(5):847–859

[121] Passner JM, Steitz TA (**1997**) The structure of a CAP-DNA complex having two cAMP molecules bound to each monomer. *Proceedings of the National Academy of Sciences of the United States of America* 94(7):2843–2847

[122] Penrose R (**1955**) A generalized inverse for matrices. In *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, 406–413, Cambridge University Press

[123] Postea O, Biel M (**2011**) Exploring HCN channels as novel drug targets. *Nature Reviews Drug Discovery* 10(12):903–914

[124] Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, van der Spoel D, et al. (**2013**) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29(7):845–854

[125] Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. (**2012**) The Pfam protein families database. *Nucleic Acids Research* 40(D1):D290–D301

[126] R Development Core Team (**2009**) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0

[127] Robinson RB, Siegelbaum SA (**2003**) Hyperpolarization-activated cation currents: from molecules to physiological function. *Annual Review of Physiology* 65(1):453–480

[128] Roux B, Karplus M (**1988**) The normal modes of the gramicidin-A dimer channel. *Biophysical Journal* 53(3):297–309

[129] Šali A, Blundell TL (**1993**) Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* 234(3):779–815

[130] Schünke S, Stoldt M, Lecher J, Kaupp UB, Willbold D (**2011**) Structural insights into conformational changes of a cyclic nucleotide-binding domain in solution from *Mesorhizobium loti* K1 channel. *Proceedings of the National Academy of Sciences of the United States of America* 108(15):6121–6126

[131] Shannon CE (**1948**) A mathematical theory of communication. *Bell System Technical Journal* 27(3):379–423

[132] Sherman J, Morrison WJ (**1950**) Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics* 21(1):124–127

[133] Shi N, Ye S, Alam A, Chen L, Jiang Y (**2006**) Atomic structure of a Na^+-and K^+-conducting channel. *Nature* 440(7083):570–574

[134] Shrivastava IH, Bahar I (**2006**) Common mechanism of pore opening shared by five different potassium channels. *Biophysical Journal* 90(11):3929–3940

[135] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. (**2011**) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7(1)

[136] Sporns O, Chialvo DR, Kaiser M, Hilgetag CC (**2004**) Organization, development and function of complex brain networks. *Trends in Cognitive Sciences* 8(9):418–425

[137] Stieber J, Herrmann S, Feil S, Löster J, Feil R, Biel M, Hofmann F, Ludwig A (**2003**) The hyperpolarization-activated channel HCN4 is required for the generation of pacemaker action potentials in the embryonic heart. *Proceedings of the National Academy of Sciences* 100(25):15235–15240

[138] Stieber J, Hofmann F, Ludwig A (**2004**) Pacemaker channels and sinus node arrhythmia. *Trends in Cardiovascular Medicine* 14(1):23–28

[139] Sukharev S, Durell SR, Guy HR (**2001**) Structural models of the MscL gating mechanism. *Biophysical Journal* 81(2):917–936

[140] Sutanthavibul S, Smith BV, Sato T, many others (**2013**), Xfig 3.2.5c. `http://www.xfig.org/`

[141] Szarecka A, Xu Y, Tang P (**2007**) Dynamics of heteropentameric nicotinic acetylcholine receptor: implications of the gating mechanism. *Proteins: Structure, Function, and Bioinformatics* 68(4):948–960

[142] Tama F, Sanejouand YH (**2001**) Conformational change of proteins arising from normal mode calculations. *Protein Engineering* 14(1):1–6

[143] Tao X, Lee A, Limapichat W, Dougherty DA, MacKinnon R (**2010**) A gating charge transfer center in voltage sensors. *Science* 328(5974):67–73

[144] Taraska JW, Puljung MC, Olivier NB, Flynn GE, Zagotta WN (**2009**) Mapping the structure and conformational movements of proteins with transition metal ion FRET. *Nature Methods* 6(7):532–537

[145] Taylor SS, Kim C, Vigil D, Haste NM, Yang J, Wu J, Anand GS (**2005**) Dynamics of signaling by PKA. *Biochimica et Biophysica Acta (BBA) – Proteins and Proteomics* 1754(1):25–37

[146] Tirion MM (**1996**) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical Review Letters* 77(9):1905–1908

[147] Tobi D, Bahar I (**2005**) Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proceedings of the National Academy of Sciences of the United States of America* 102(52):18908–18913

[148] Tombola F, Pathak MM, Isacoff EY (**2006**) How does voltage open an ion channel? *Annual Review of Cell and Developmental Biology* 22:23–52

[149] Tristani-Firouzi M, Chen J, Sanguinetti MC (**2002**) Interactions between S4-S5 linker and S6 transmembrane domain modulate gating of HERG K+ channels. *Journal of Biological Chemistry* 277(21):18994–19000

[150] Tronin AY, Nordgren CE, Strzalka JW, Kuzmenko I, Worcester DL, Lauter V, Freites JA, Tobias DJ, Blasie JK (**2014**) Direct evidence of conformational changes associated with voltage-gating in a voltage sensor protein by time-resolved x-ray/neutron interferometry. *Langmuir*

[151] UniProt Consortium, et al. (**2014**) Activities at the universal protein resource (UniProt). *Nucleic Acids Research* 42(D1):D191–D198

[152] Vargas E, Yarov-Yarovoy V, Khalili-Araghi F, Catterall WA, Klein ML, Tarek M, Lindahl E, Schulten K, Perozo E, Bezanilla F, et al. (**2012**) An emerging consensus on voltage-dependent gating from computational modeling and molecular dynamics simulations. *The Journal of General Physiology* 140(6):587–594

[153] Vinh NX, Epps J, Bailey J (**2010**) Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research* 9999:2837–2854

[154] Wächter M, Jäger K, Thürck D, Weißgraeber S, Widmer S, Goesele M, Hamacher K (**2014**) Using graphics processing units to investigate molecular coevolution. *Concurrency and Computation: Practice and Experience* 26(6):1278–1296

[155] Wächter M, Jäger K, Weißgraeber S, Widmer S, Goesele M, Hamacher K (**2012**) Information-theoretic analysis of molecular (co) evolution using graphics processing units. In *Proceedings of the 3rd International Workshop on Emerging Computational Methods for the Life Sciences*, 49–58, ACM

[156] Wainger BJ, DeGennaro M, Santoro B, Siegelbaum SA, Tibbs GR (**2001**) Molecular mechanism of camp modulation of HCN pacemaker channels. *Nature* 411(6839):805–810

[157] Wang J, Chen S, Siegelbaum SA (**2001**) Regulation of hyperpolarization-activated HCN channel gating and cAMP modulation due to interactions of COOH terminus and core transmembrane regions. *The Journal of General Physiology* 118(3):237–250

[158] Wang L, Jiang T (**1994**) On the complexity of multiple sequence alignment. *Journal of Computational Biology* 1(4):337–348

[159] Wang Y, Rader A, Bahar I, Jernigan RL (**2004**) Global ribosome motions revealed with elastic network model. *Journal of Structural Biology* 147(3):302–314

[160] Watts DJ, Strogatz SH (**1998**) Collective dynamics of ‘small-world’ networks. *Nature* 393(6684):440–442

[161] Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (**2009**) Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America* 106(1):67–72

[162] Weil P (**2012**) *Koevolution in molekularen Komplexen*. Ph.D. thesis, TU Darmstadt

[163] Weil P, Hoffgaard F, Hamacher K (**2009**) Estimating sufficient statistics in co-evolutionary analysis by mutual information. *Computational Biology and Chemistry* 33(6):440–444

[164] Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta SJ, Weiner P (**1984**) A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society* 106(3):765–784

[165] Weißgraeber S, Hamacher K (**2012**) Generalized correlations in molecular evolution: A critical assessment. *From Computational Biophysics to Systems Biology (CBSB11) – Celebrating Harold Scheraga’s 90th Birthday* 8:231

[166] Weißgraeber S, Hoffgaard F, Hamacher K (**2011**) Structure-based, biophysical annotation of molecular coevolution of acetylcholinesterase. *Proteins: Structure, Function, and Bioinformatics* 79(11):3144–3154

[167] White RA, Szurmant H, Hoch JA, Hwa T (**2007**) Features of protein-protein interactions in two-component signaling deduced from genomic libraries. *Methods in Enzymology* 422:75–101

[168] Wickham H (**2009**) *ggplot2: elegant graphics for data analysis*. Springer New York

[169] Williams T, Kelley C, many others (**2013**), Gnuplot 4.6: an interactive plotting program. `http://www.gnuplot.info/`

[170] Xu C, Tobi D, Bahar I (**2003**) Allosteric changes in protein structure computed by a simple mechanical model: hemoglobin T <–> R2 transition. *Journal of Molecular Biology* 333(1):153–168

[171] Yang L, Song G, Jernigan RL (**2007**) How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophysical Journal* 93(3):920–929

[172] Yellen G (**2002**) The voltage-gated potassium channels and their relatives. *Nature* 419(6902):35–42

[173] Zagotta WN, Olivier NB, Black KD, Young EC, Olson R, Gouaux E (**2003**) Structural basis for modulation and agonist specificity of HCN pacemaker channels. *Nature* 425(6954):200–205

[174] Zhou L, Siegelbaum SA (**2007**) Gating of HCN channels by cyclic nucleotides: residue contacts that underlie ligand binding, selectivity, and efficacy. *Structure* 15(6):655–670

[175] Zhou Y, Morais-Cabral JH, Kaufman A, MacKinnon R (**2001**) Chemistry of ion coordination and hydration revealed by a K^+; channel–Fab complex at 2.0 Å resolution. *Nature* 414(6859):43–48

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit entsprechend den Regeln guter wissenschaftlicher Praxis selbstständig und ohne unzulässige Hilfe Dritter angefertigt habe.

Sämtliche aus fremden Quellen direkt oder indirekt übernommenen Gedanken sowie sämtliche von Anderen direkt oder indirekt übernommenen Daten, Techniken und Materialien sind als solche kenntlich gemacht. Die Arbeit wurde bisher bei keiner anderen Hochschule zu Prüfungszwecken eingereicht.

Darmstadt, den 12. August 2014

Stephanie Weißgraeber

Danksagung

Zum Schluss möchte ich all jenen danken, die mich während meiner Doktorandenzeit und beim Schreiben dieser Dissertation unterstützt haben.

Als erstes danke ich natürlich Kay Hamacher für die Möglichkeit in seiner Arbeitsgruppe zu promovieren, die Betreuung und dafür, dass ich stets die Freiheit hatte eigene Ideen in meine Forschung einzubringen. Ein besonderer Dank geht an meinen Zweitgutachter Gerd Thiel sowie an Anna Moroni für die konstruktiven Gespräche über Ionenkanäle und Diskussionen über meine Ergebnisse.

Den aktuellen und ehemaligen Mitgliedern der AG *Computational Biology and Simulation* und Gisela Schaffert danke ich für hilfreiche und für sinnlose Gespräche, all den Spaß, den wir hatten, und die gute Arbeitsatmosphäre, in der sich alle gegenseitig helfen. Mit euch war es großartig!

Jorge und Randall danke ich dafür, dass sie mir die Welt und die Wissenschaft immer wieder aufs Neue veranschaulicht haben.

An meine Freunde, insbesondere an Filiz, geht ein großes Dankeschön dafür, dass sie sich stets geduldig meine Probleme angehört haben; vielen Dank auch an meine fleißigen Korrekturleser Patrick, Sandro, Julia, James und Alex.

Meinen Eltern und Geschwistern bin ich sehr dankbar für ihre Unterstützung und die Gewissheit, dass ich mich immer auf sie verlassen kann. Und schließlich: Philipp, danke, dass du mich verstehst, für mich da bist und immer an mich glaubst!