Conformation-specific statistical coupling analysis of the α7 acetylcholine receptor

Rebecca Dean

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the
Master's degree in Chemistry

Department of Chemistry and Biomolecular Sciences
Faculty of Science
University of Ottawa

Candidate	Supervisors
_____	_____
Rebecca Dean	Corrie daCosta and Maria Musgaard

ABSTRACT

It is well known that information contained in a protein sequence is what allows it to fold into its three-dimensional shape, which performs a specific function. It has been possible for some time to search for proteins with similar sequences, using bioinformatics tools such as BLAST. But it is also known that proteins with similar, or even the same sequence can adopt different structures and vice-versa. With this in mind, we look to use a method called Rosetta-HMMER to perform conformationally specific sequence searches in order to exploit this property of proteins. This method involves the use of Rosetta to redesign protein structures to fit a specified α-carbon backbone, and then uses HMMER to generate a sequence profile. This profile can then be used to query for sequences able to adopt the specified backbone structure. These collected sequences can then be aligned for the purpose of performing statistical coupling analysis. We have used this Rosetta-HMMER method in conjunction with available structures of the α7 acetylcholine receptor to show that distinct sequence profiles generated from different conformations of the same protein are capable of retrieving unique sets of natural sequences when used as a query. We have also shown that when these unique sets of natural sequences are used to perform statistical coupling analysis, different residues are identified as statistically coupled, potentially generating insight into residues that have more potential importance for one backbone conformation over another.

ACKNOWLEDGEMENTS

Firstly, I would like to thank Professor Maria Musgaard for all her support and guidance. Through my work as an honours student and my initial work as a graduate student I learned many valuable skills thanks to her patience and teachings, in a field that was outside of my comfort zone. I would also like to thank her for being a great mentor and friend, as well as for taking a chance on me. Maria, you will always be the first person to see my potential. You are the catalyst for everything I have achieved.

Much of the credit for this work goes to Professor Corrie daCosta, for being the first professor to inspire my passion for proteins (and for being the best professor I had during my bachelor's), for being a model example of what makes a good scientist, for mentorship, support and fearlessness in the face of uncertainty and most importantly for stepping up to help me finish this degree. I would not have been able to do this without him. My gratitude goes beyond what I can put into words.

I would like to thank everyone in the daCosta and Musgaard Labs for their support, in no particular order, Anna, Megan, Mariam, Greg, John, Christian, Louise, Ray, Brittany, Ali, Vasilisa, Yasir, James and Patrick. It was great to get to work close to so many amazing people. Your insights and suggestions were appreciated and helped make me a better scientist.

I would like to thank Dr. Jesus Banda-Vazquez for assisting with this project, for being patient in teaching me the skills I needed to complete it and for being a sounding board for ideas. You were an invaluable collaborator and friend. I enjoyed working with you.

I would like to thank my family for their belief that I could do this work and do it well. I would like to thank my dad and my sister for teaching me the value of hard work and determination. I would like to thank my cousin and my uncle for being there when I was far from home.

I would like to thank all the friends I have made along the way for their support, specifically David, Noah, Joëlie, James, Carley, Alina and Ethan. Je veux remercier aussi Laurie, Marine et Olga. You've made my time in Ottawa both amazing and memorable. I am grateful to have met you all.

Lastly, I would like to thank the funding agencies that made this research possible, NSERC and the University of Ottawa.

DEDICATION

This work is dedicated to my mom, who never got to finish her degree. I miss you and I hope that you're proud of me. I wish you could be here to celebrate with me.

"Still round the corner there may wait
A new road or a secret gate
And though I oft have passed them by
A day will come at last when I
Shall take the hidden paths that run
West of the Moon, East of the Sun"
-J.R.R Tolkien

Table of Contents

PREFACE vi

FIGURES vii

ABBREVIATIONS ix

Chapter 1. Introduction ..... - 1 -

I - Ion channels ..... - 1 -

II – Pentameric Ligand-Gated Ion Channels ..... - 3 -

III – Nicotinic acetylcholine receptors ..... - 5 -

IV – Functional states of α7 ..... - 7 -

V – The Relationship Between Structure and Sequence in α7 and Protein Sequence Redesigns ..... - 9 -

VI – Statistical coupling analysis of protein sequences ..... - 10 -

VII – Motivation for the present work ..... - 11 -

VIII – Hypothesis and Objective ..... - 12 -

CHAPTER 2: Statistical Coupling Analysis (BLAST-Based Sequence Searches) ..... - 13 -

I – Objective ..... - 13 -

II – Approach ..... - 13 -

III – Methods ..... - 14 -

IV - Results ..... - 26 -

V - Discussion ..... - 29 -

CHAPTER 3: Statistical Coupling Analysis (Structure-Based Sequence Searches) ..... - 30 -

I – Objective ..... - 30 -

II – Approach: ..... - 30 -

III – Results: ..... - 31 -

IV – Discussion: ..... - 47 -

V – Methods: ..... - 50 -

CHAPTER 4: Discussion and Conclusion ..... - 53 -

I – Sequence based SCA vs Structure based ..... - 53 -

II – SCA can identify residues unique to a conformation ..... - 54 -

III – Next steps: what comes after SCA? ..... - 56 -

IV – SCA can be applied to other proteins ..... - 60 -

V – Effect of Cα Backbone RMSD on Shared SCA Results ..... - 60 -

VI – Limitations of SCA ..... - 65 -

VII – AlphaFold: from sequence to structure and back again ..... - 66 -

VIII – Conclusion ..... - 67 -

Supplementary Information ..... - 76 -

PREFACE

This MSc began in the laboratory of Dr. Maria Musgaard, where the initial project was to simulate Ryanodine Receptors. Upon Dr. Musgaard's return to Oxford in 2021, I joined the daCosta lab, where the idea for this project unfolded. Supervision for this degree continued under both Dr. Musgaard and Dr. daCosta. Dr. Jesus Banda Vazquez was instrumental in teaching me to use the scripts and bioinformatics tools needed to kickstart this project and to obtain the data I needed. All results depicted in the following work are my own and were obtained and analyzed between fall of 2021 and fall of 2022.

FIGURES

Figure 1.1. Schematized Representation of Ion Channel Gating

Figure 1.2. General Architecture of a Pentameric Ligand Gated Ion Channel

Figure 1.3. Functional Conformational transitions of α7

Figure 2.1. BLAST Sequence Alignment Coloured by Position

Figure 2.2. Pairwise Percent Identity Analysis of BLAST Sequence Alignment

Figure 2.3. Positional Conservation Analysis of full α7 Sequence from BLAST Results

Figure 2.4. SCA matrix of full length α7 sequence multiple sequence alignment generated using BLAST results

Figure 2.5. Eigenvalue decomposition of α7 sequence multiple sequence alignment generated using BLAST results

Figure 2.6. SCA clustering matrix and tree of full length α7 sequence multiple sequence alignment generated using BLAST results

Figure 2.7. SCA sector results of a sequence-based search mapped onto α7 sequence

Figure 2.8. SCA sector results of a sequence-based search mapped onto α7 homology model

Figure 2.9. Key regions of the α7 receptor mapped onto α7 homology model

Figure 3.1. Backbone Cα root mean squared deviation comparisons between all six available PDBs of the α7 acetylcholine receptor

Figure 3.2. Pairwise Percent Identity Analysis of Rosetta Redesigns

Figure 3.3. Pairwise Percent Identity Matrix of Profile Consensus Sequences

Figure 3.4. Total Sequence Space for Individual Structures from Sequence Search Results

Figure 3.5. Core Sequence Space for Structures from Sequence Search Results

Figure 3.6. Sector Results Mapped onto Sequence of 7eki

Figure 3.7. Sector Results Mapped as Surface Representation for PDB 7eki from all datasets

Figure 3.8. Core Sector Results Mapped onto Conformations of α7

Figure 3.9. Total Sector Results Mapped onto Conformations of α7

Figure 3.10. Unique Sector Results Mapped onto Conformations of α7

Figure 3.11. Flowchart Representing Methods Used to Obtain SCA Results

Figure 4.1. Residues identified for mutation by SCA of unique dataset

Figure 4.2. Residues identified for mutation by SCA of Unique Dataset Mapped onto α7 sequence

Figure 4.3. Percentage of Coupled Sites in Common plotted against RMSD of the core dataset

Figure 4.4. Percentage of Coupled Sites in Common plotted against RMSD of the total dataset

Figure 4.5. Percentage of Coupled Sites in Common plotted against RMSD of the unique dataset

ABBREVIATIONS

pLGICs	Pentameric Ligand-Gated Ion Channels
$GABA_AR$s	Gamma-Aminobutyric Acid Receptors
5-HT_3Rs	5-Hydroxytryptamine 3 Receptors
GlyRs	Glycine Receptors
nAChRs	Nicotinic Acetylcholine Receptors
CNS	Central Nervous System
PNS	Peripheral Nervous System
Cryo-EM	Cryogenic Electron Microscopy
ECD	Extracellular Domain
TMD	Transmembrane Domain
ICD	Intracellular Domain
SCA	Statistical Coupling Analysis
MSA	Multiple Sequence Alignment
BLASTp	Basic Local Alignment Search Tool: protein
HMM	Hidden Markov Model
HMMER	Hidden Markov Modeller
CD-HIT	Cluster Database at High Identity Tolerance
RMSD	Root Mean Square Deviation

Chapter 1. Introduction

I - Ion channels

Ion channels are membrane proteins that act as gatekeepers of the cell, controlling the inward and outward diffusion of ions. They contribute to a variety of cellular and metabolic processes, and act as signal amplifiers for cellular responses such as transmission of nerve signals and excitation and contraction of muscles.[1] Since solving the first ion channel structure in 1998[2], several additional ion channel structures have become available, contributing to understanding ion channel structure and function. Ion channels are a broad class of proteins, present within all uni- and multi-cellular organisms. For example, the Erwinia Ligand-Gated ion channel is a prokaryotic ion channel found in *Dickeya Dadantii*[3], responsible for plant rot, while Transient Receptor Potential channels[4], are temperature sensitive ion channels responsible for our cooling sensation upon consumption of compounds such as menthol[5].

While ion channels are a diverse group of proteins, they share several functional hallmarks. All ion channels allow the passive diffusion of ions across cell membranes, down their electrochemical gradients – a process called "conduction". Most ion channels are charge selective, with some conducting only cations (Na^+, K^+, Ca^{2+}, etc.) and others conducting only anions (Cl^-) [6,7]. Furthermore, some channels are selective for a particular ionic species. For example, K^+ channels conduct K^+ ions ~1000x more efficiently than Na^+ ions [8,9]. The opening and closing of most ion channels, a process known as "gating", is highly regulated, with some channels gating in response to changes in transmembrane voltage, or the binding of a small chemical ligand. Some channels gate in response to changes in temperature, pressure, and even pH[1] (Figure 1.1).

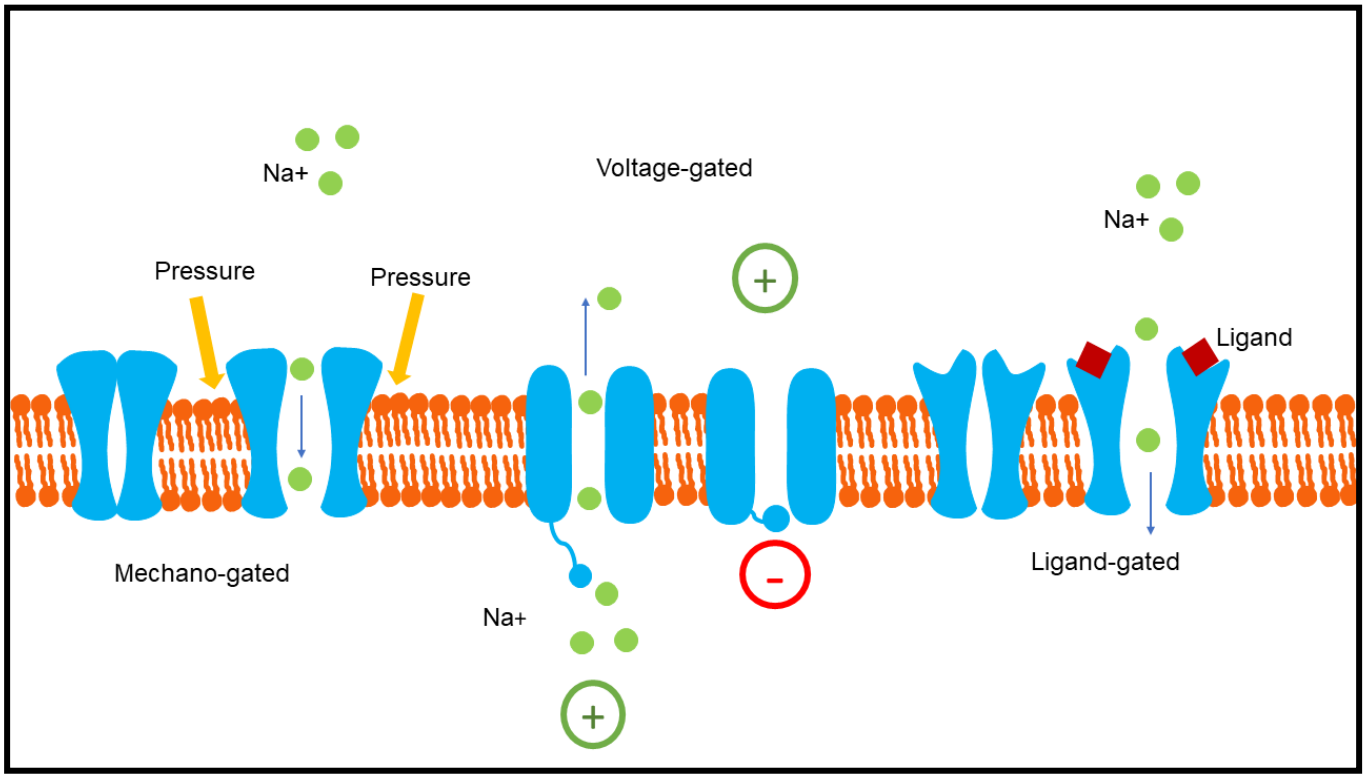

Figure 1.1. Schematized Representation of Ion Channel Gating. Mechano-sensitive (left), voltage-gated (middle), and ligand-gated (right) ion channels. Diagrams depicting schematized ion channel function are also depicted, arrows representing ion channel flow are shown in blue and arrows representing force are shown in yellow. Mechano-gated channels are activated by changes in pressure, voltage-gated channels are activated by changes in current, and ligand-gated are activated by ligand (red diamonds) binding.

II – Pentameric Ligand-Gated Ion Channels

Pentameric ligand-gated ion channels (pLGICs) are a class of ion channel that have two defining characteristics: (1) they are ligand gated channels, meaning that they are opened and closed by the binding of a small molecule, and (2) they are pentamers, in that they are composed of five identical or homologous subunits arranged around a central ion-conducting pore (Figure 1.2). There are prokaryotic and eukaryotic pLGICs. Mammalian pLGICs fall into four subfamilies: Gamma-aminobutyric acid receptors ($GABA_ARs$)[10,11], 5-Hydroxytryptamine 3 receptors ($5\text{-}HT_3Rs$)[12,13], Glycine receptors (GlyRs)[14,15], and nicotinic acetylcholine receptors (nAChRs)[16,17]. In humans, pLGICs play an important role in neurophysiology, have been implicated in numerous diseases, and are thus important drug targets. Understanding the relationship between pLGIC structure and function is therefore not only important for a molecular understanding of the nervous system, but also has implications for human health[18,19].

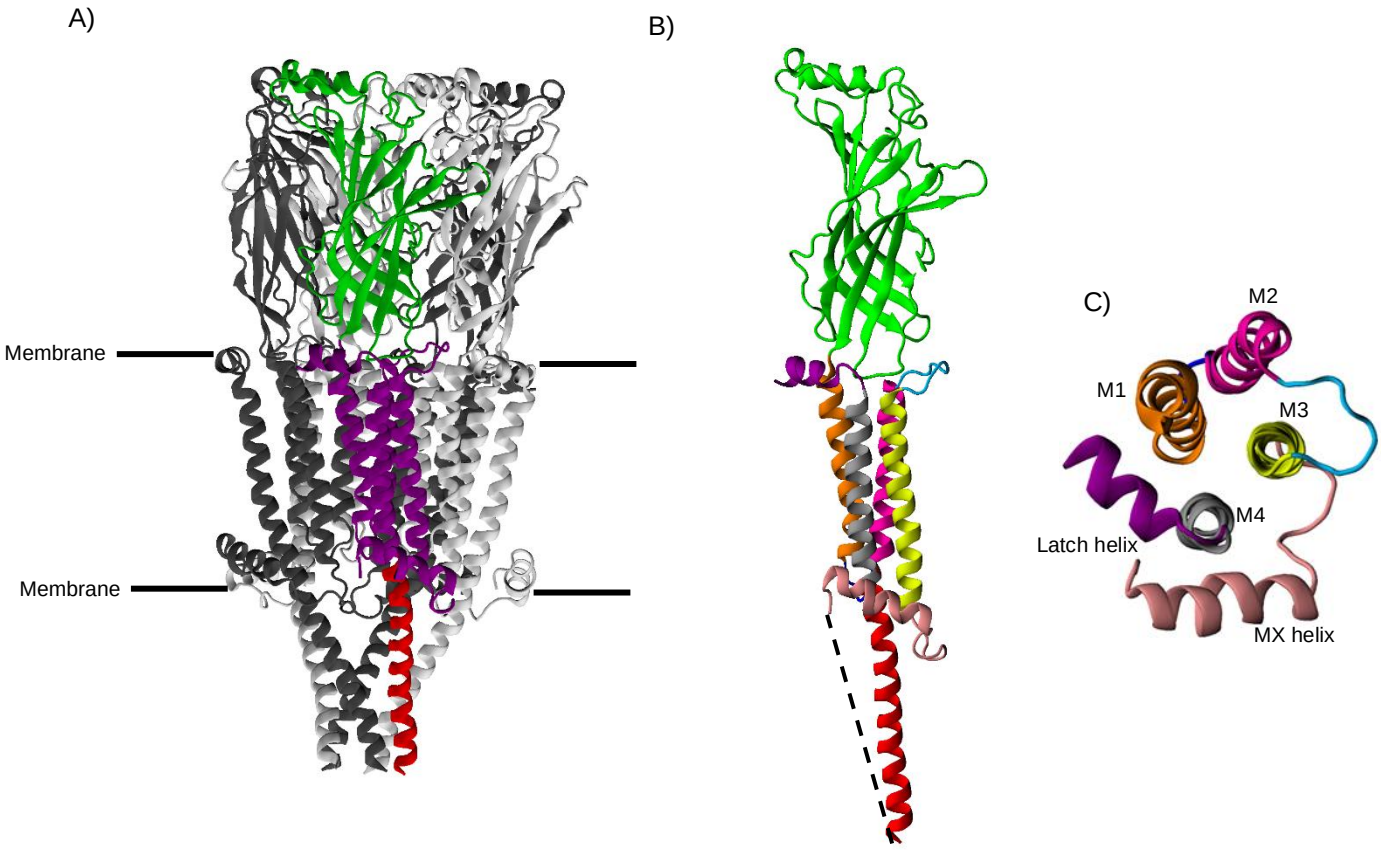

Figure 1.2. General Architecture of a Pentameric Ligand Gated Ion Channel. Structure of the unliganded α7 nicotinic receptor (PDBID: 7eki[32]). (A) Full pentamer with single subunit coloured according to domain: β-sandwich in green, M1-M4 helical transmembrane region in purple, intracellular region in red. (B) Single subunit with individual M1-M4 helices individually coloured, missing loop region of the intracellular domain is represented by the dashed line. (C) Top-down view of the transmembrane domain including the labelled M1 (orange), M2 (magenta), M3 (yellow) and M4 (grey) helices, MX helix (light pink) and latch helix (purple). Images generated in VMD.

III – Nicotinic acetylcholine receptors

In humans, nAChRs and 5-HT_3Rs are cation-selective[13,17], while $GABA_A$Rs and GlyRs are anion-selective[11,15]. All pLGIC subunits share a similar architecture, with a ~200 amino acid extracellular ligand binding domain (ECD) that folds into an immunoglobin-like β-sandwich, and a transmembrane domain (TMD) consisting of four membrane spanning α-helices (M1-M4) (Figure. 1.2). In addition, each subunit contains a cytoplasmic domain of variable length that is the least conserved region of each subunit. In general, structure and function of this cytoplasmic domain is relatively poorly characterized[16,20] and there is no complete structure available. However recent work by Bondarenko et al. has proposed a highly flexible intracellular domain (ICD) that still seems somewhat speculative[21].

The superfamily of pLGICs is structurally and functionally diverse, with each subfamily having different roles within the nervous system. This work focusses specifically on the family of acetylcholine receptors. The first acetylcholine receptor was isolated in 1970 by Changeux, Kasai and Lee, following their work with the *Electrophorus Electricus*[22]. As suggested by their name, nicotinic AChRs are activated by both endogenous acetylcholine, and exogenous nicotine[23]. There are both muscle- and neuronal-type nAChRs and the main difference between them is both their subunit composition, as well as where they are expressed in the body. As one might guess from their names, muscle-type receptors are found at the motor endplate of skeletal muscle and sit at the neuromuscular junction[24], whereas neuronal receptors are mainly expressed in the central and peripheral nervous system (CNS and PNS)[23]. The structural and functional characterization of these receptors remains an important goal due to their involvement in diseases, and technical

advances in electron Cryogenic Electron-Microscopy (Cryo-EM) have led to a renaissance in the structural characterization of both neuronal and muscle-type nAChRs[25].

IV – Functional states of α7

The α7 nicotinic acetylcholine receptor is a homopentameric pLGIC. Activation of α7 leads to pre- and post-synaptic excitation in the CNS, specifically in the hippocampus of the brain[26,27], suggesting that α7 is involved in memory and learning. The α7 receptor is uniquely characterized amongst eukaryotic acetylcholine receptors by its increased permeability to Ca^{2+} ions[28], as well as its unusually rapid desensitization. α7 has been linked to diseases such as schizophrenia and Alzheimer's, as well as in medical events such as stroke and myocardial infarctions[29–31]. These links to disease make the human α7 receptor an important drug target.

The α7 receptor transitions between at least three functionally relevant conformations, an open/active conformation, a closed/resting conformation, and a closed/desensitized conformation. The α7 receptor starts in the resting state, transitions to the active state, and then transitions from the active to the desensitized state, whereupon it returns to its original resting state. The active state is considered ion conducting while the resting and desensitized are non-ion conducting. Structures of the human α7 receptor are available in these three conformations and are the focus of this thesis (Figure. 1.3)[28,32].

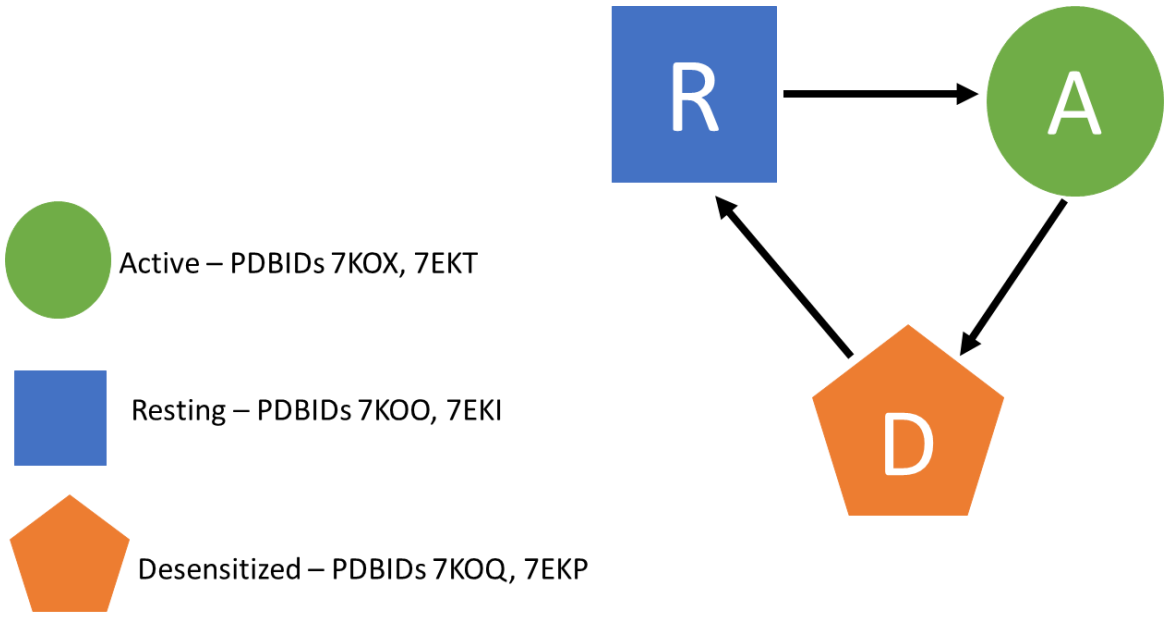

Figure 1.3. Functional Conformational transitions of α7. Conformations of α7 are labelled and coloured according to their assignment proposed by their respective publications (7ekt, 7eki and 7ekp from Zhao *et al.*[32] 7kox, 7koq and 7koo from Noviello *et al.*[28]). Schematized cycle of α7 is shown on the right. The active conformation is open and ion conducting, while the R and D conformations are closed and non-conducting.

V – The Relationship Between Structure and Sequence in α7 and Protein Sequence Redesigns

Anfinsen's dogma tells us that a protein's structure is dictated by its sequence[33]. The diversity of available pLGIC subunits in nature means we have access to a large and diverse set of homologous proteins that share a similar overall fold and structure, and we aim to exploit this diversity and homology to understand structure-function relationships in α7. The α7 receptor is a dynamic protein that exists in multiple functionally relevant conformations, and mutations alter this conformational landscape[34]. The mechanisms by which mutations alter α7 conformational change are unknown. We hypothesize that certain amino acids contribute asymmetrically in terms of their importance towards one conformation over another. In this thesis, we explore an approach for identifying such residues using a combination of structure-based sequence redesign (Rosetta-HMMER) and statistical coupling analysis.

Rosetta is a protein modelling package capable of *ab initio* protein folding[35]. First developed in the late 1990's, Rosetta has become an invaluable tool for redesigning proteins. Rosetta can also stochastically redesign amino acid sequences to fit the energy constraints of a specified Cα backbone. In the present work we exploit this capability of Rosetta to produce sets of *synthetic* sequences that energetically fit the original backbone structure of various functionally relevant α7 conformations (Figure 1.3).

HMMER (Hidden Markov Modeller) is a computational package that can be used to build sequence profiles for a specified sequence alignment[36]. This profile resembles a consensus sequence and corresponds to the probabilities of each amino acid being the output at each position in a protein sequence. The set of Rosetta redesigned sequences is used to build this sequence profile. Since the profile is built upon a unique set of 3D backbone constraints, a tacit assumption

is that it encodes conformationally specific information and can therefore be used to search for natural sequences capable of adopting the specified backbone conformation. The retrieved sequences can then be used to create a multiple sequence alignment that is biased to a specific conformation and thus three-dimensional structure.

VI – Statistical coupling analysis of protein sequences

The functional properties of proteins are shaped by the pattern of physical interactions between atoms of their amino acid residues. In other words, the sum of physical interactions within a protein structure determines the function of the protein. Physical interactions between amino acids in proteins are short range (3-5 Å) and depend on geometry. This means that atoms will only directly affect other atoms close to them. But it also means that there can be a set of interactions between neighbouring amino acid residues that contact each other in a cascading fashion, linking up distant sites that themselves are not in direct contact. It can be difficult to understand these long-range "interactions" because we do not see energy in protein structures[37]. We used statistical coupling analysis (SCA), a method for measuring amino acid covariation in a protein sequence, to look for specific amino acids in the human α7 receptor that could be more important for one specific conformation over another. This method satisfies three characteristics needed for addressing the problem of amino acids "interacting" at a distance within protein structures: the assumption that the most crucial functional features are contained in only a few amino acid residues, that amino acid residues coevolve, and that amino acids linked through coevolution are also linked to a specific function[38,39].

This form of analysis can be used in a conformationally specific manner through the implementation of the aforementioned Rosetta-HMMER approach for retrieving sequences for

alignment, as we first design a set of sequences unique to a specified Cα backbone/conformation and then design a profile based on those sequences. We then use these profiles to retrieve sequences capable of adopting the specified template conformation, which are then aligned. These conformationally-specific alignments form the basis of individual statistical coupling analyses. This allows us to look for variations in co-evolving residue groups (called sectors) across the set of conformations available to us.

VII – Motivation for the present work

Several structures of the human α7 acetylcholine receptor in various conformations are available. The different conformations can be used to generate a set of synthetic sequences, from which are built distinct sequence profiles, where each sequence profile is optimized to a specific backbone conformation. In turn, each sequence profile can be used to retrieve and score a set of natural sequences that are most likely to adopt the backbone conformation specified by each sequence profile. Will the set of sequences retrieved in each case be different? Furthermore, given the dependence of SCA on the input sequence alignment, will SCAs based upon the different sets of retrieved sequences also be different and meaningful?

VIII – Hypothesis and Objective

We hypothesize:

1. That distinct sequence profiles generated from different backbone conformations of the same protein can be used to retrieve unique sets of natural sequences from public databases.

And further,

2. That the unique sets of natural sequences can be used to perform conformationally specific statistical coupling analyses, which will provide insight into residues important for a particular backbone conformation.

Based upon these hypotheses, our overall objective is to determine if SCAs based upon different conformations of the same protein will give rise to different sets of statistically coupled residues? If so, future experiments will determine if these differences are meaningful.

CHAPTER 2: Statistical Coupling Analysis (BLAST-Based Sequence Searches)

I – Objective

The objective of this section was to perform a traditional SCA, using a multiple sequence alignment where all sequences were retrieved from a standard BLASTp search with the human α7 amino acid sequence used as the "query" sequence. This traditional SCA, from a sequence homology-based sequence alignment, represents a control for our SCAs generated from structural homology-based sequence alignments, in order to look at more general information in α7 that is not conformationally biased.

II – Approach

In this section, we have used traditional methods to retrieve homologous sequences as opposed to structurally biased methods that will be discussed in the following chapter. While structurally biased methods are useful for details specific to individual conformations, information can be missing due to incomplete structures used as templates (in our case the unresolved intracellular domain is missing from all our structural templates). The following method not only serves as a control for the downstream structure based SCAs, but also provides results relating to the complete mature α7 sequence, including the structurally unresolved intracellular domain.

III – Methods

Multiple Sequence Alignments

The first step in running any SCA is building a multiple sequence alignment (MSA). A Basic Local Alignment Search Tool: protein (BLASTp)[40] search of the uniref100[41] database was performed using the full-length wild type α7 sequence (supplementary information) and returned 53,212 hits. These sequence hits were initially aligned to a Hidden Markov Model (HMM) using the *hmmalign* function of HMMER 3.2.1 suite. This step was necessary to get an initial alignment because the full set of hits was too large for other alignment algorithms to handle. Sequences in this case are only aligned to the whole human α7, so regions of other sequences were missing (regions of sequence that best match the human α7 sequence are aligned while the rest of the sequence can be discarded in some cases). The HMM was taken from a single chain run of alphafold[42], which constructs MSAs in a manner similar to the method described in chapter three. To reduce repetition, sequences were clustered using Cluster Database at High Identity Tolerance (CD-HIT)[4344]. An identity cut-off was set to 95%, that is, no sequence more than 95% identical to any other sequence was kept, and a minimum sequence length cut-off was set to 400 amino acids. Sequences were then aligned with MAFFT[45] to get a final alignment used to perform the SCA. This alignment contained 7,546 sequences.

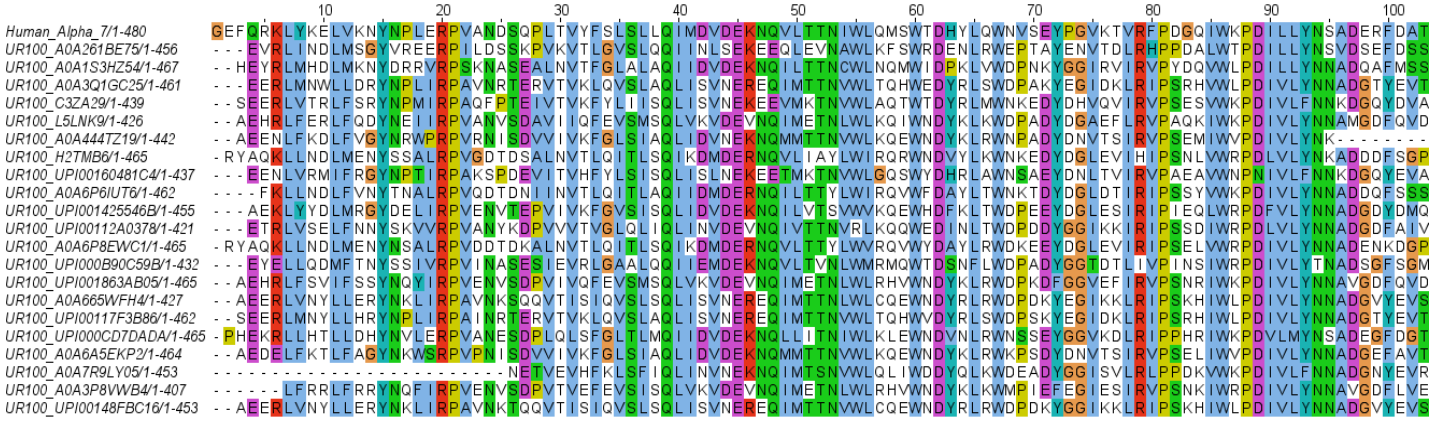

Figure 2.1. BLAST Sequence Alignment Coloured by Position. Portion of the final sequence alignment containing 7,546 natural sequences used for analysis in following sections. Full alignments used for all SCAs can be found in the figshare repository link in supplementary information. Full human α7 sequence is at the top of the alignment. Sequences are in a ranked alignment.

Alignment Analysis

The next step in running a SCA is the pairwise percent identity analysis. The SCA v.5.0 toolbox[46] was used to analyse the final multiple sequence alignment (MSA). This is a matrix comparison plot in conjunction with a histogram plot that can be used to evaluate the quality of the alignment. This matrix compares the identity of every sequence in the alignment against every other sequence in the alignment (pairwise) and computes the identity as the fraction of amino acids common between them. Note that positions with more than 40% gaps were ignored. Only half of the matrix needs to be evaluated due to symmetry and so it is also binned as a histogram plot. Figure 2.2 demonstrates that the alignment is reasonably homogeneous, or "equally unalike" by SCA standards, since most sequences fall between 0.15 and 0.45 pairwise fraction ID. By contrast, an alignment that is non-homogeneous would have multiple, clearly defined large peaks. It is important to have a homogeneous alignment because it avoids pre-biasing of sequences into groups, which allows analysis to be both more accurate and more efficient. As shown by the example in the supplementary material, this became an issue with certain alignment algorithms.

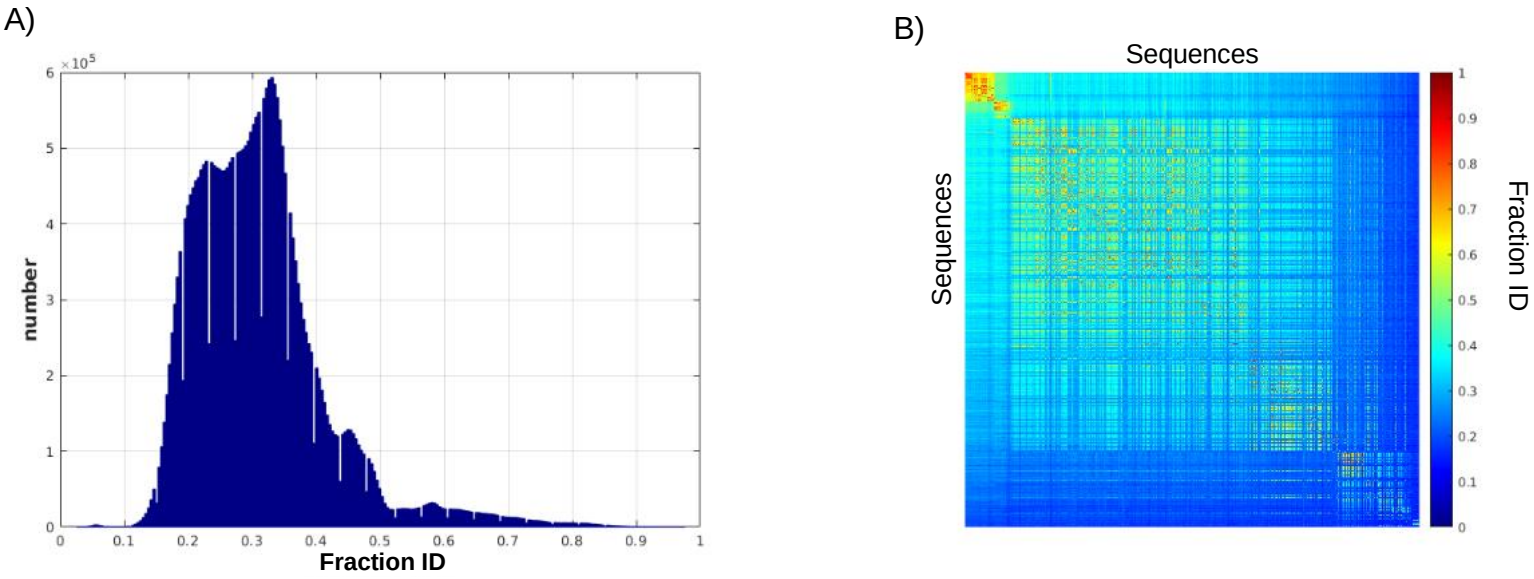

Figure 2.2. Pairwise Percent Identity Analysis of BLAST Sequence Alignment. (A) Histogram of pairwise percent sequence identity (Pairwise %ID) between each sequence against the number of sequences from the alignment that fall into the bin for a given pairwise fraction. Pairwise fraction represents how similar each sequence in the alignment is to each other sequence in the alignment. The histogram is one half of the diagonal (B) matrix which runs each sequence against each other and colours the resulting pixel by pairwise percent identity (heat map on the right).

Conservation Analysis and SCA matrix

The next step in the analysis was to determine the positional conservation of each amino acid in the sequence of interest, in this case our full length α7 sequence, using the generated MSA. Panel A of Figure 2.3 is a histogram where all amino acids with the same degree of conservation have been binned together, while panel B is a plot of the conservation by sequence position over the longest non-gapped alignment, which in this case is 464 amino acids long. Looking at the number of amino acids vs conservation plot (panel A) we can see that there not many highly conserved positions (positions with a conservation value over 2.5) in this alignment. This is further reflected in the conservation vs sequence position plot in panel B.

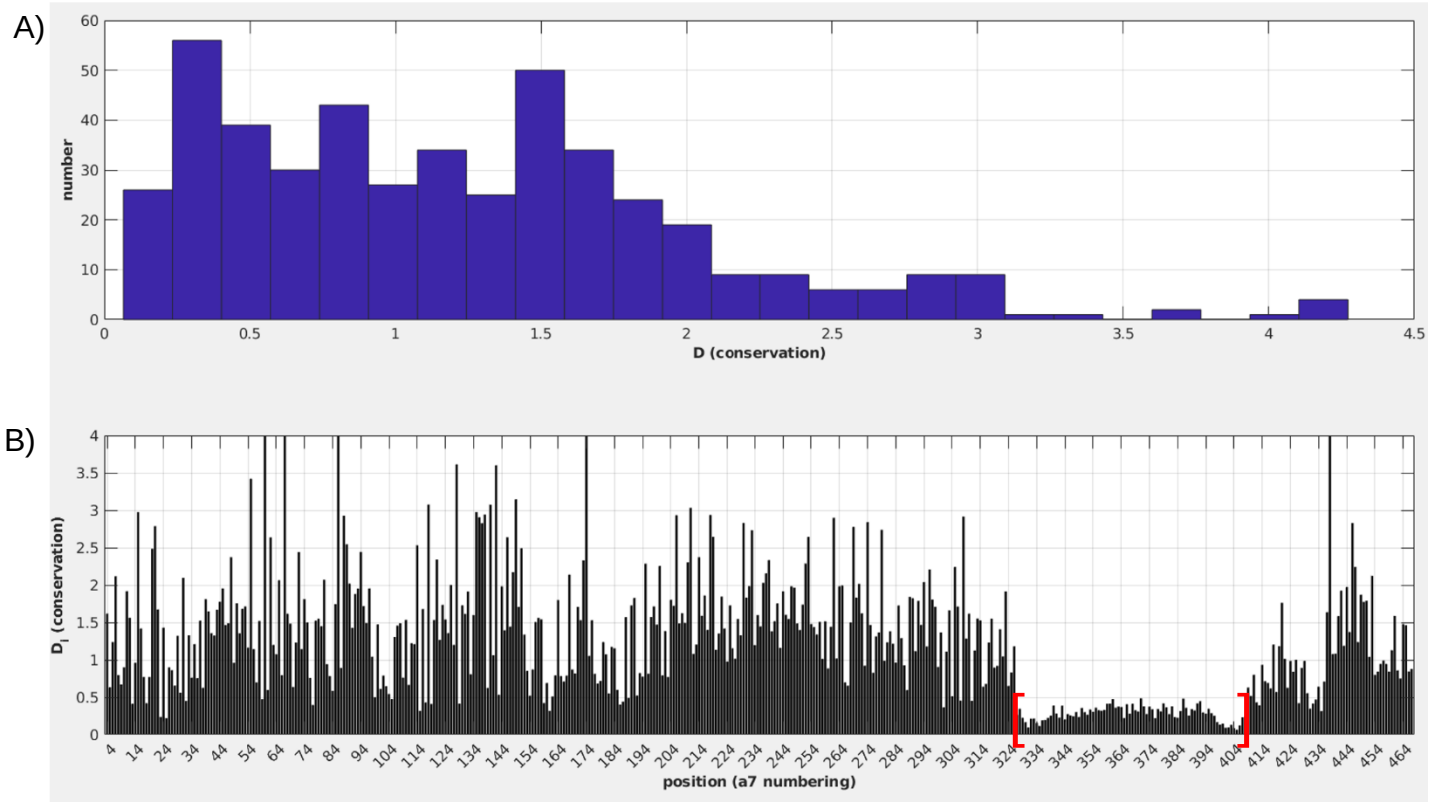

Figure 2.3. Positional Conservation Analysis of full α7 Sequence from BLAST Results. (A) Computed conservation of amino acids vs number of amino acids in each bin of the histogram plot. (B) Amino acid sequence position in wild-type numbering against conservation computed in "A". A low conservation gap delineates the poorly conserved intracellular region is visible in the bottom plot, bracketed in red.

Following the conservation analysis, the SCA matrix is computed. The SCA matrix is a representation of the degree of pairwise coupling between all positions in the amino acid sequence alignment. It is calculated by first determining how much the observed distribution at each position deviates from the random distribution and then examining the degree to which the outcome of one position will influence every other position. In Figure 2.4, we can see an area of low coupling, which corresponds to the poorly conserved intracellular domain of α7. The SCA matrix is useful and can be further broken down to yield additional important information.

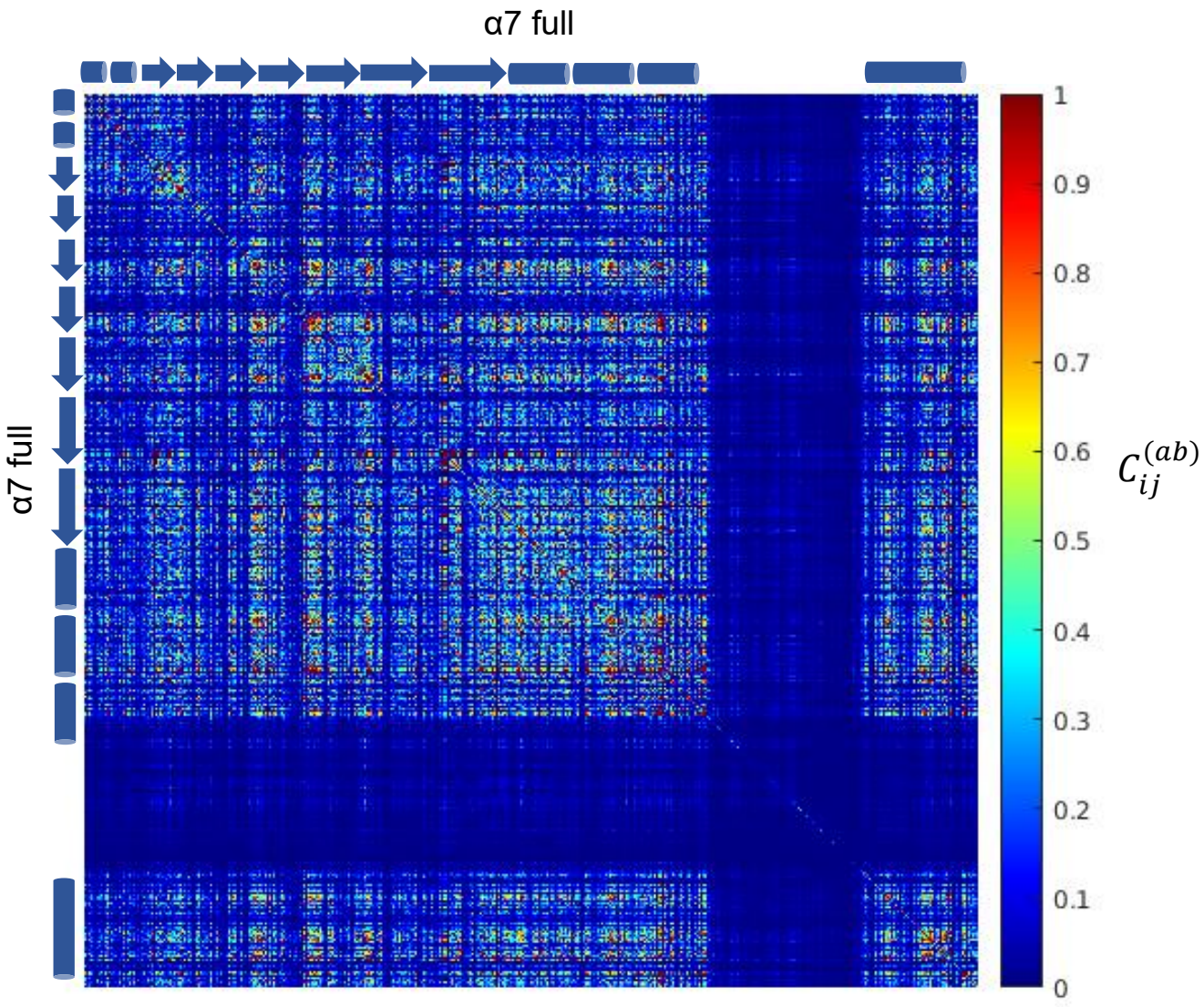

Figure 2.4. SCA matrix of full length α7 sequence multiple sequence alignment generated using BLAST results. SCA matrix shows scattered couplings throughout the sequence, with the exception of the diagonal which is each amino acid coupled against itself, and the missing ICD region which appears as a large blue gap of zero correlations. Details on SCA calculations can be found in supplementary information. C is equal to conservation D of amino acids a and b at positions i and j.

Sector Determination

Eigenvalue decomposition is the first step in determining how many sectors are identified by the SCA process. In this case, the eigenvalue represents N groups of amino acids that are considered statistically linked. From the presence of ungrouped eigenvalues that fall outside the main distribution (Figure 2.4), multiple sectors were suggested. The top 4 eigenvalues were chosen for independent component analysis[39] (kmax 4), which yielded three independent components for sector analysis. Eigenvalue cutoff is subjective and differing numbers of eigenvalues here could mean more than three sectors, which is a limitation addressed in chapter four.

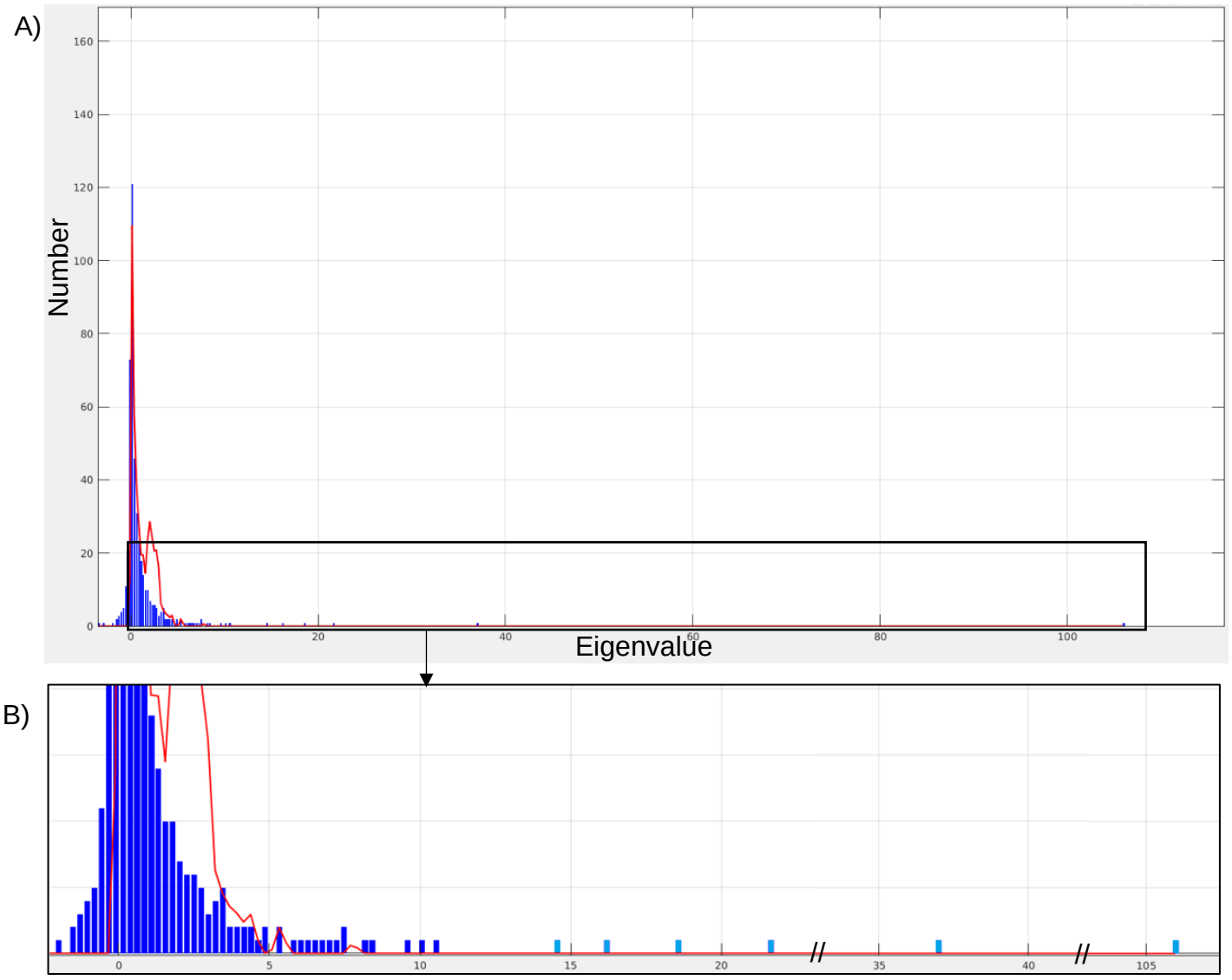

Figure 2.5. Eigenvalue decomposition of α7 sequence multiple sequence alignment generated using BLAST results. (A) The eigenvalue represents the groups of amino acids that are considered statistically linked, while number corresponds to the number of amino acids that remain correlated after analysis. Multiple ungrouped modes are visible in this case. Eigenvalues are represented by blue ticks. (B) Inset of A with axis breaks represented by slashes to better show values considered to fall outside the main distribution. Lighter blue values in this case are beyond the main distribution.

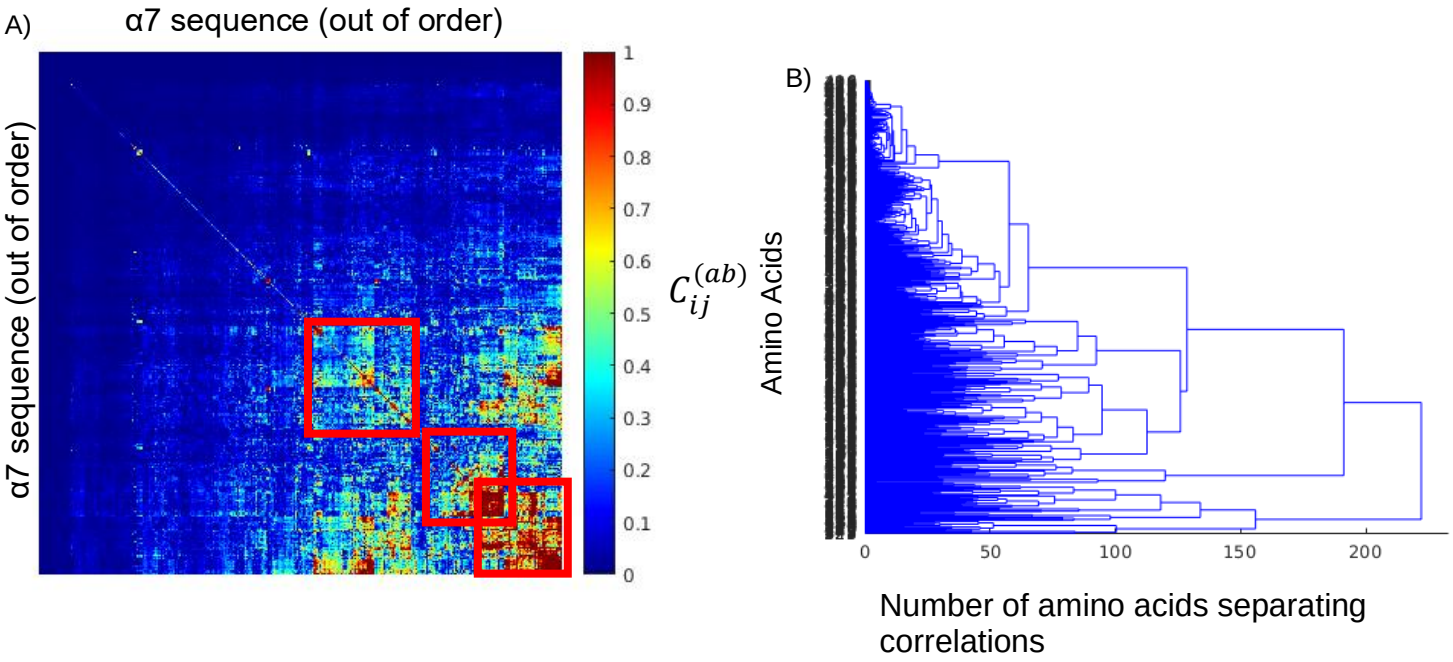

Figure 2.6. SCA clustering matrix and tree of full length α7 sequence multiple sequence alignment generated using BLAST results. (A) Clustering matrix. Sector groups along the diagonal are considered, with a conservation value of 0.5 to 1. Clustering was computed according to calculations described in supplementary information. The α7 sequence (out of order) in this case simply refers to the fact that the amino acids in sequence have been grouped according to their correlations rather than in sequence order as shown in Figure 2.4. (B) Distance tree of half the clustering matrix. Distance is reported as the number of amino acids separating correlations. Three sectors have been chosen to move forward with this analysis, in order to preserve consistency across all performed SCAs.

In conjunction with ICD, analysis by hierarchical clustering can be performed to visualize correlations more easily. In this case, positions in the sequence that have similar levels of correlation will be grouped close to each other. In keeping with our other analyses, we have chosen to group three larger "boxes" of medium to high conservation (yellow to red in colour) along the diagonal (Figure 2.6). Clustering for sectors in this case is subjective, which is a limitation addressed in chapter four.

The final step is the definition of sectors. A sector in this case simply represents a group of amino acids that are statistically linked to each other. Most datasets from all SCA analyses performed suggested three to four eigenmodes, which were transformed into three maximally independent components and so we have chosen to define three sectors in all cases, including this one. Sectors are defined according to the method developed by Halabi et. al[39]. Note that while in the original work detailing SCA there was clear functional data linked to specific sectors, in our case we consider the sectors arbitrary and instead our focus is on the statistically important residues identified irrespective of sector. This is a limitation which is further addressed in chapter four.

IV - Results

Table 1. SCA Residue Results By Sector from BLAST Search. Sectors are mapped in wild-type mature α7 numbering. Sector colouring is defined in Matlab.

Sector	Residues
Blue	C219, H296
Red	N16, N53, P120, L209, I244, T264, H298, G302, G303, M305, L313, A318, M323, P326, T461
Green	Q39, N47, Q48, Y93, D131, W149, Y188, C190, C191, L215, F230, E259, S285, T289

The single coupling represented by blue is not representative of a true sector in this case, which is a limitation discussed in chapter four. Colouring here is representative of what is predetermined in the provided scripts and is assigned by the method, which is also a limitation discussed in chapter four. These results can be mapped onto structures to visualize regions that may be important and identify any patterns (Figures 2.7 and 2.8). Key areas of the protein that will be mentioned in discussion are displayed in Figure 2.9.

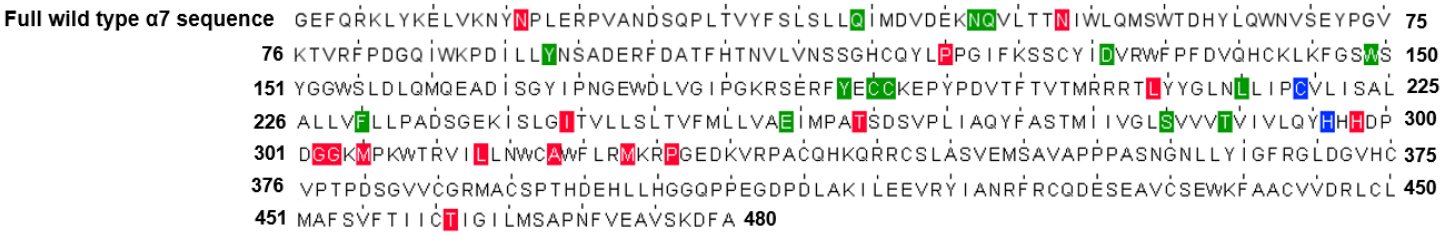

Figure 2.7. SCA sector results of a sequence-based search mapped onto α7 sequence. Mapped blue, green, and red sector residues from Table 1. Colouring is based upon assigned sectors in Matlab. Alignment was generated in Jalview[75].

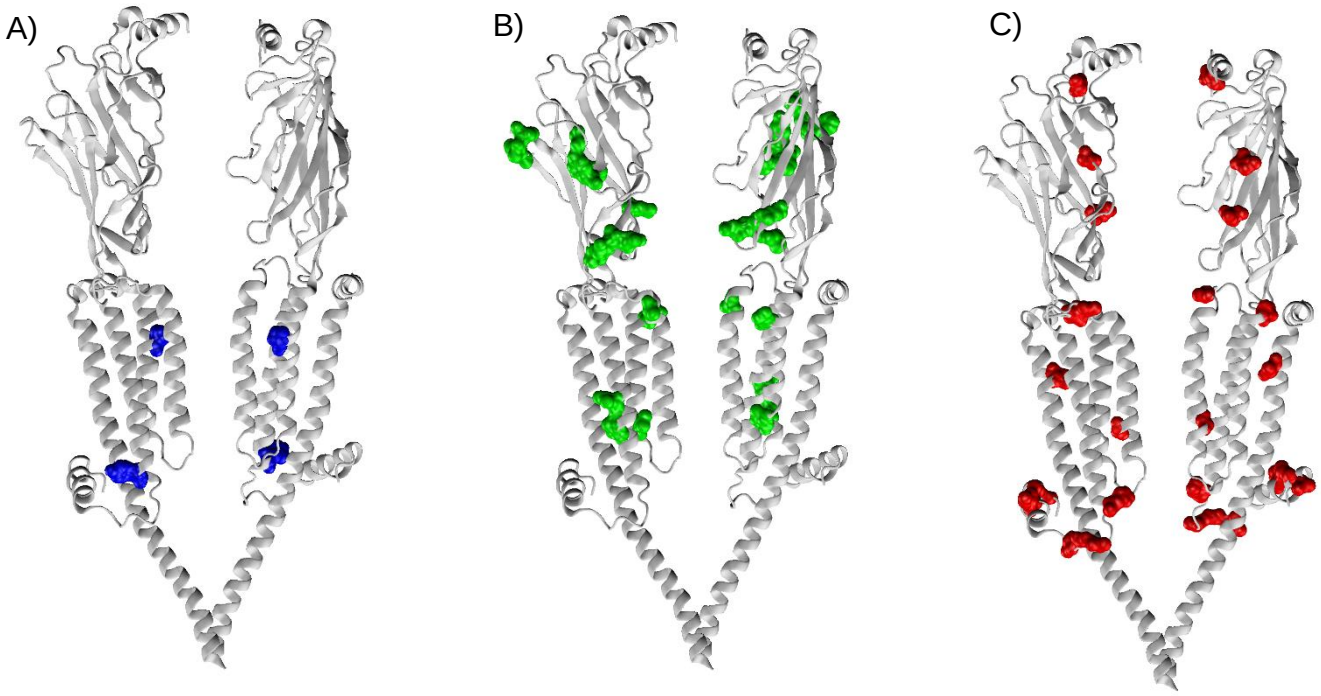

Figure 2.8. SCA sector results of a sequence-based search mapped onto α7 homology model. Mapped residues from Table 1 with (A) as blue, (B) as green and (C) as red. Structure is PDB ID 7eki (unliganded). Sector residues are shown in surface representation on chains A and C. Colouring is based upon assigned sectors in Matlab. Images were generated with VMD.

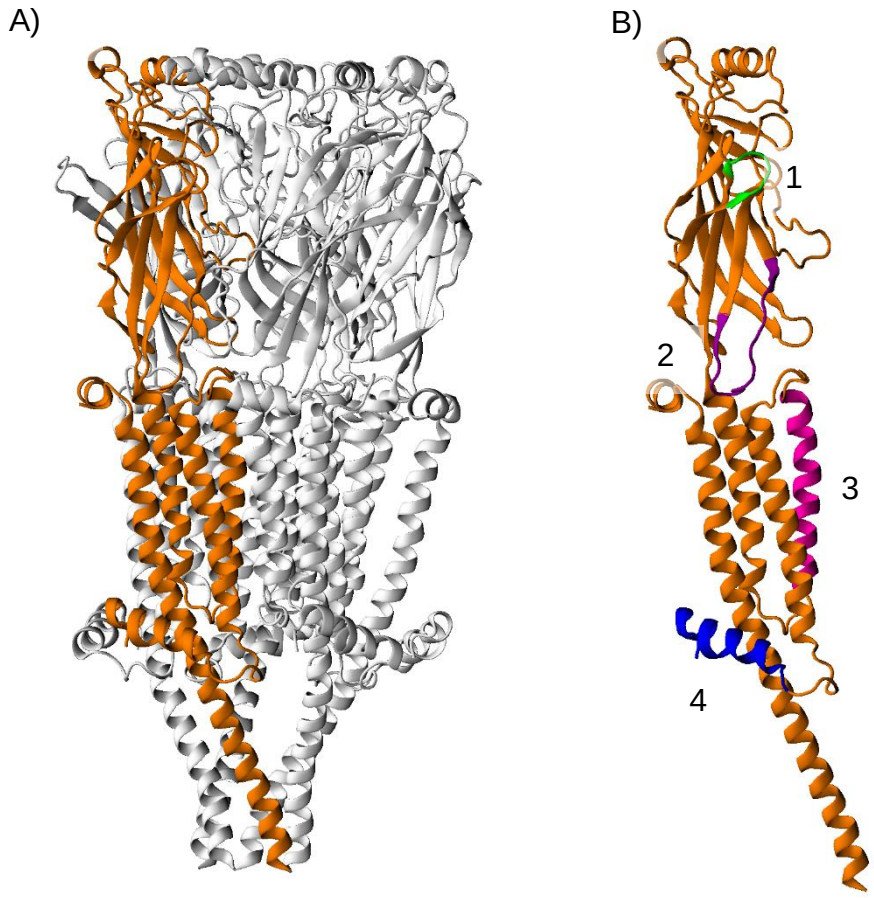

Figure 2.9. Key regions of the α7 receptor mapped onto the unliganded α7 Structure. Structure is PDB ID 7eki (unliganded). (A) Overview of entire structure with chain A in orange. (B) Overview of key regions numbered as follows: 1) Loop C, close to allosteric binding site (green). 2) Cys-loop region close to ECD-TMD interface (purple). 3) M2 pore-lining helix (pink). 4) MX helix which is important for receptor assembly (blue). Images generated in VMD.

V - Discussion

The above set of results contains residues with described experimental importance in literature. In Figures 2.7 and 2.8, the following residues in loop C were identified: Y188, C190 and C191. Loop C is an area important for ligand binding and recognition[47–49]. Y188 is an important residue for high affinity binding [47,50]. C190-C191 is a highly conserved disulfide bond amongst the α-subunits of nicotinic acetylcholine receptors[51]. Other residues that were identified are residues L313 and A318 in the MX helix, which has been linked to receptor assembly[52]. E259 (E20'), was identified as well. It is a pore lining residue located along the M2 helix which is known to be important for conductance and gating[28]. G302-G303 which is not well described in the literature, was also identified in the alignment. It is located in the pre-MX loop, potentially making it important for receptor flexibility and helical recognition during assembly. A mutational experiment that could be used to investigate this hypothesis using traditional receptor cell-surface expression assays or electrophysiological methods would be to swap these two glycines for alanines, to examine whether flexibility is important in this region. These results and the method described herein represent a sample of the information available to us from running a more traditional SCA on the α7 receptor. The structure guided SCAs will be discussed in more detail in chapter three, while comparisons between the two, as well as certain limitations of the method, will be discussed in chapter four.

CHAPTER 3: Statistical Coupling Analysis (Structure-Based Sequence Searches)

I – Objective

The objective of this section is to determine if statistical coupling analyses based upon sequences retrieved using sequence profiles designed from different conformations of the human α7 acetylcholine receptor give rise to differences in statistically coupled residues.

II – Approach:

As discussed in Chapter one, the structure-based approach begins with Rosetta, which is a protein modelling software[35] used for computational protein design. Rosetta uses a specified α-carbon backbone as a constraint to determine an optimal amino acid sequence capable of adopting the specified conformation without energetically disturbing the structural constraints. To produce a multiple sequence alignment biased to a particular backbone conformation, a Hidden Markov Model (HMM), which serves as a sequence profile, is built from a set of Rosetta redesigns specific to a given α7 conformation. The resulting sequence profiles, based upon different α7 conformations, were used to retrieve natural sequences capable of adopting the associated Cα backbone conformations.

III – Results:

RMSD Structural Comparisons

There are six structures of the human α7 acetylcholine receptor available. We examined the root mean square deviations (RMSDs) of the Cα backbones of these six different structures (Figure 3.1) to determine their similarities and differences, as well as how closely they compare to each other. Given the initially observed differences we hypothesized that Rosetta redesigns based upon various structures of α7 should be different, as will be their corresponding sequence profiles. The differences between structures, in particular 7kox which has the highest RMSD, can be attributed to differences in structural quality and characterization protocol, as well as preliminary data that suggests 7kox is the only true active conformation structure. However, as stated in chapter one, we are classifying them according to how they were denoted in their original publication.

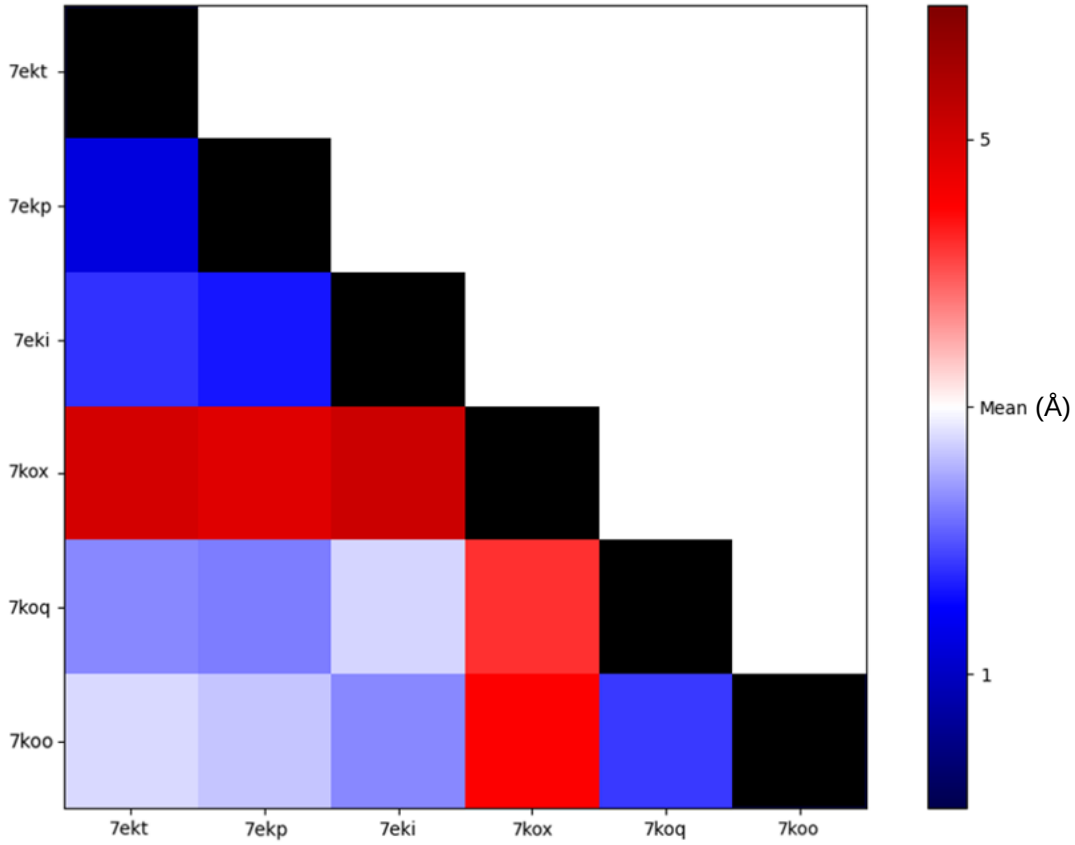

Figure 3.1. Backbone Cα root mean squared deviation comparisons between all six available PDBs of the α7 acetylcholine receptor. Structures were aligned and RMSDs calculated in the VMD software package. RMSD heatmap was plotted in matplotlib using pyplot. RMSD for all structures was calculated using the shortest sequence common to all structures (in this case 7kox, which is 386 amino acids in length). Structures are grouped in the following conformations: Active (7ekt, 7kox), Desensitized (7koq, 7ekp) and resting (7koo, 7eki).

Rosetta Redesigns

We next made direct comparisons of single Rosetta redesigned sequences. Repeat designs of the same chain of a structure, in this case 7kox A, A2 and A3, have more similarity (Figure 3.2, Table S2) than the same chain of different structures (chain A of 7kox, 7koo and 7koq, Figure 3.2, Table S2). Note that while only 7koo, 7koq and 7kox are examined in the following Figure, redesigns were produced for all structures. While all the chain A 7kox repeats had ~60% identity to each other, the chains of the different structures only get as high as 33%, indicating that Rosetta has recognized structural differences and chosen to optimize for a particular conformation (Figure 3.2). Similarity of the redesigns to wild type was found to be 25% or less in all cases. In each case, 150 redesigned sequences were used to build the sequence profiles.

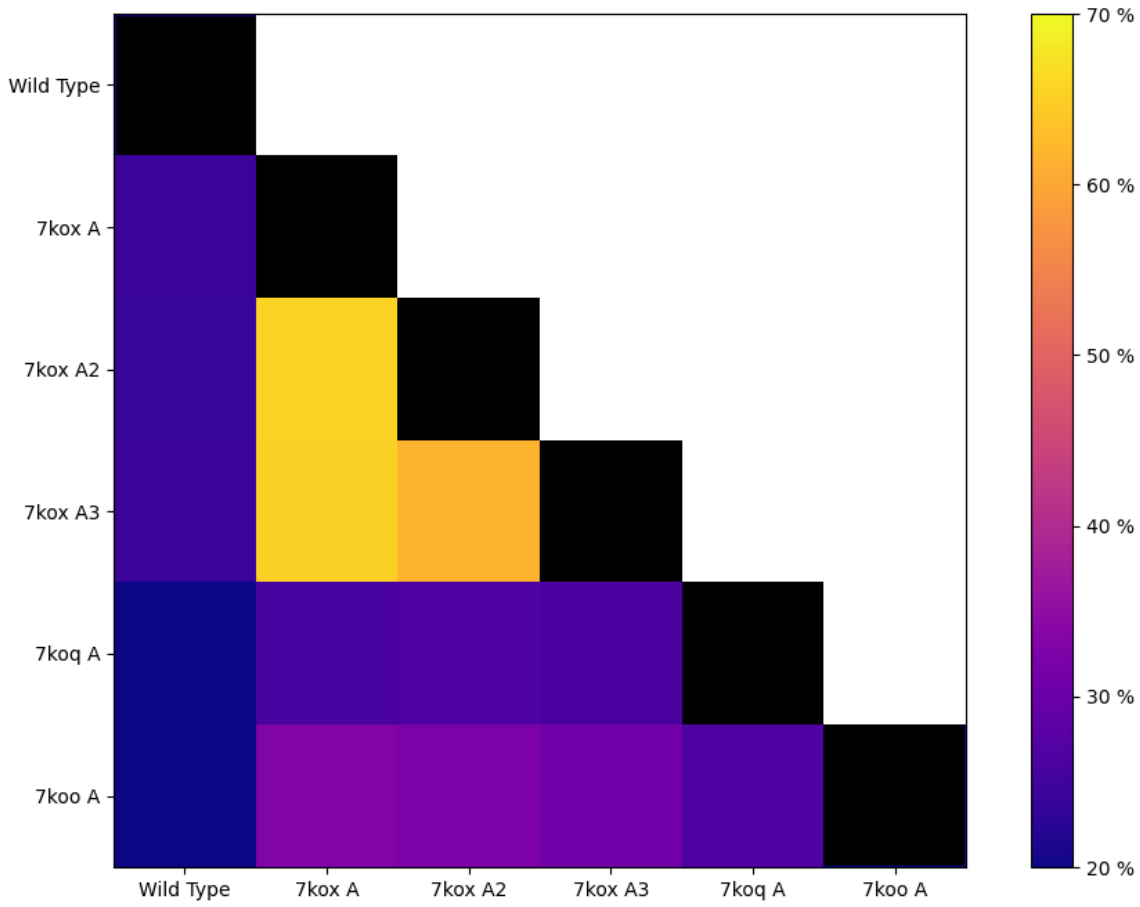

Figure 3.2. Pairwise Percent Identity Analysis of Rosetta Redesigns. Redesigns were generated using Rosetta all atom fixbb function[35]. Pairwise identity was calculated in Jalview[75], sequences were aligned in Jalview[75]. Heatmap was plotted in matplotlib using pyplot. Structures are grouped in the following conformations: Active (7kox), Desensitized (7koq) and resting (7koo).

Sequence Profiles

Once the HMM profile has been generated, it is used to query a database for sequences that best fit the profile. An important control in this experiment is the capacity of the structure to retrieve its own sequence after undergoing this *in silico* process. This has been proposed to be an objective measure of the quality of structural models built upon x-ray or electron microscopic data[53]. Gratifyingly, using our sequence profile in all cases retrieved the human α7 within the top 10 results. We can compare sequence profiles by looking at the pairwise % identity of their consensus sequences (highest probability amino acid at each position) against the wild-type sequence as well as against each other. No sequence was more than 48% identical to any other sequence (Figure 3.3), and the % identity to wild type was even lower, 25% or less in all cases. If we compare these results with our RMSD results from Figure 3.1, there is some correlation between structures with similar RMSDs having higher % identity. For example, 7ekt, 7ekp and 7eki all had RMSDs between one and two angstroms and these have the highest % identity to each other. There also appears to be high % identity between the two resting state conformations (7koo and 7eki) in this case, though their RMSD was closer to the mean. 7kox (putative active state) appears to have similarly lower % identity to all other conformations, which correlates with the calculated RMSD, which was the highest compared to every other structure.

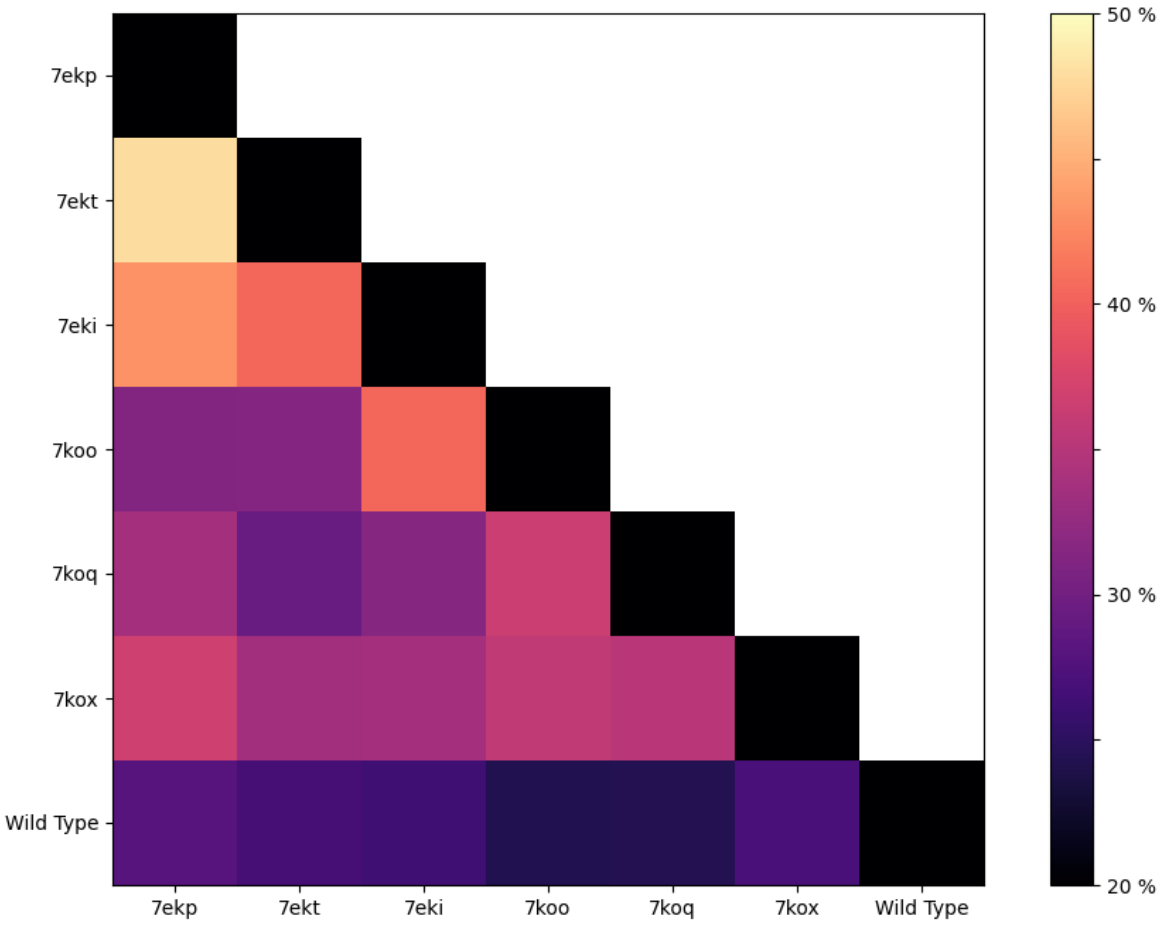

Figure 3.3. Pairwise Percent Identity Matrix of Profile Consensus Sequences. Consensus sequences were generated using the hmmer emit function. Pairwise identity was calculated in Jalview[75], sequences were aligned in Jalview[75] to the shortest sequence present (7kox, 386 amino acids) to preserve gapping and identity at sequence positions. Structures are grouped in the following conformations: Active (7ekt, 7kox), Desensitized (7koq, 7ekp) and resting (7koo, 7eki).

Sequence Searches

Using our sequence profiles as queries to search led to the retrieval of over 35,000 sequence hits in all cases. The search was done for each chain and only the common hits between five chains were retained to control for minor variation between chains of the same structure. Going back to one of our original objectives, we are interested in seeing if differences in the backbone of our six structures leads to differences in SCA results, which lead to the following three types of datasets: (1) a "multiverse" or **core** set of sequences that were common to all searches and thus backbones, (2) a set of sequences **unique** to each conformation/structure and (3) the combination of the previous two sets into what we call the **total** set of sequences for each conformation/structure. The total set of hits is shown in the pentamer Venn diagrams and designated by the red circles (Figure 3.4). The core set is the hexamer Venn diagram shared sequence space, while the unique is the total set minus the core. A full breakdown of how this is defined is shown in Figures 3.4 and 3.5.

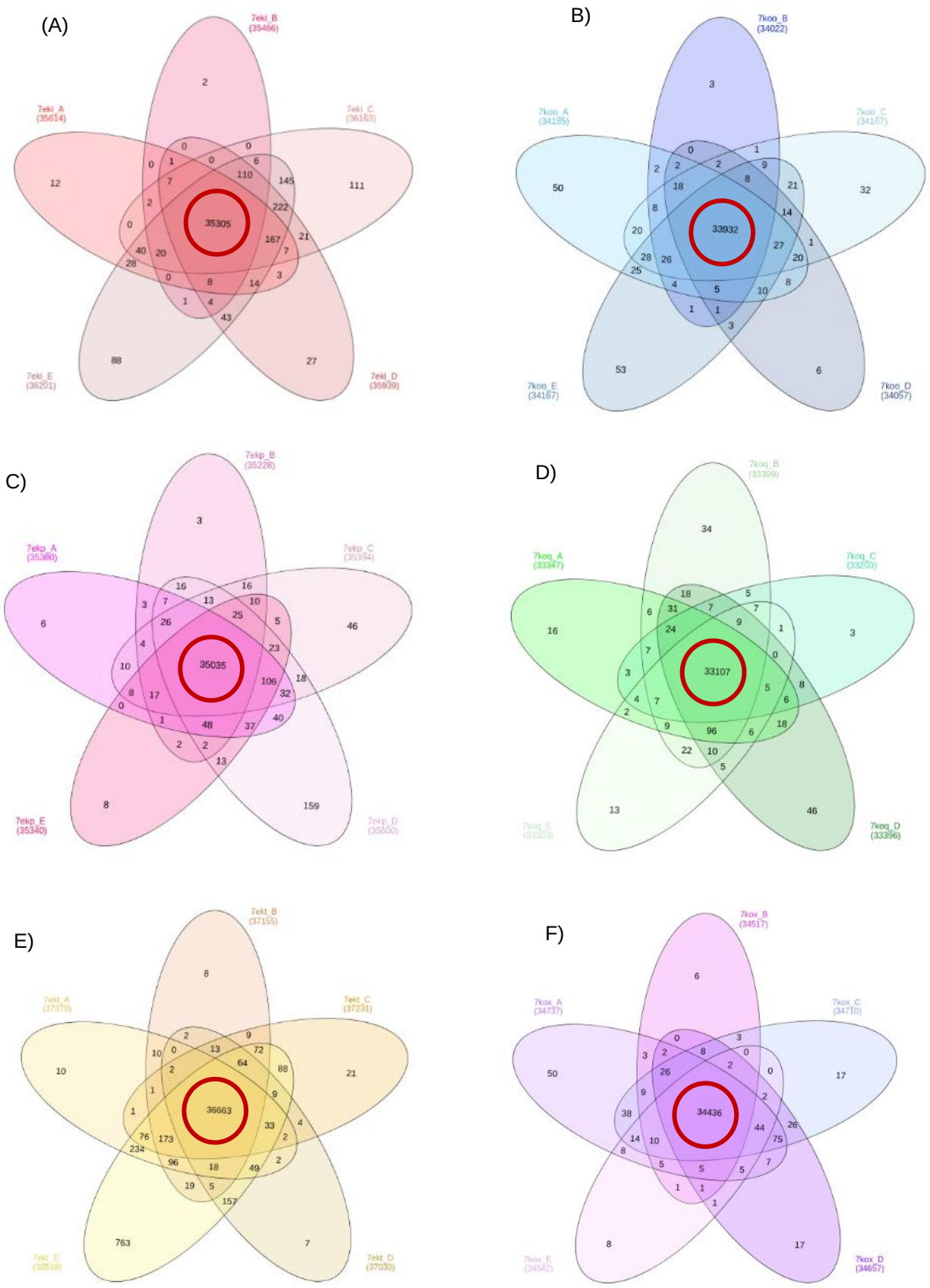

Figure 3.4. Total Sequence Space for Individual Structures from Sequence Search Results. Sequence search was performed using the sequence profiles for each structure and the uniref100 database as downloaded in October 2021. Red circle denotes the set of sequences from all combined chains of each pdb, which represents the total sequence space. (A) 7eki (B) 7koo (C) 7ekp (D) 7koq (E) 7ekt (F) 7kox. Sequence space denoted in red represents sequences common to all five chains, in order to account for minor variation between chain backbones.

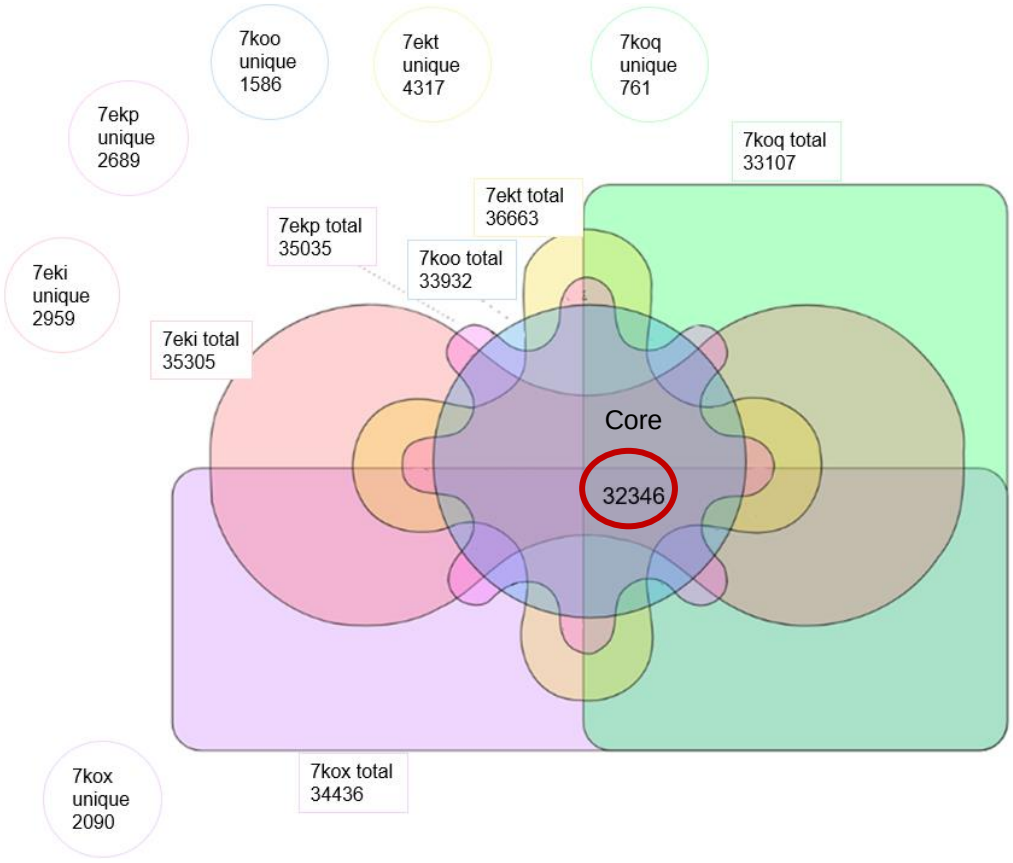

Figure 3.5. Core Sequence Space for Structures from Sequence Search Results. Sequence Search was performed using the sequence profiles for each structure and the uniref100 database as downloaded in October 2021. Red circle denotes the combined sequence space for all structures. Sequence numbers for total datasets are shown in square boxes beneath structure names and are taken from sequence sets in Figure 3.4. Sequence numbers for unique datasets are shown in ovals around the diagram. Note that the total represents all sequences retrieved for each structure, the core is the set of sequences in common between all structures, and the unique is the total minus the core (set of sequences unique to each structure).

SCA

Once all sequence sets were obtained from each backbone conformation, they were aligned to each respective sequence profile. To reduced repetition, sequences were clustered so that no sequence was more than 95% alike using CD-HIT and then aligned. These alignments were then fit (ideally aligned) to the wild type α7 sequence and aligned once more. These ideal fit alignments were subjected to the full procedure for Statistical Coupling Analysis, as detailed in chapter two. The alignment and mapping for 7eki (unliganded, resting state) is shown in Figures 3.6 and 3.7, while the remaining figures are in the supplementary material. The next set of results shown are the sectors themselves, mapped onto the corresponding PDB, as well as mapped onto a sequence alignment of all three datasets. The three different datasets in Figures 3.6 and 3.7 represent what a large, general MSA (core and total datasets) can tell us versus what an MSA unique only to the specified conformation can tell us. There should be more similarity in residues that continuously appear in the total and core datasets than in the unique datasets, because there is more similarity among the sequences used to construct the MSAs. This also represents a control in making sure SCA is capable of recognizing regions of the protein with experimentally determined importance. As discussed in chapter two, sectors are arbitrarily assigned by the method, which may cause some residues to switch sectors between different datasets. There is also no functional data associated with complete sectors. These are limitations that will be discussed in chapter four. Further details on individual residues and regions of interest will be discussed in the section IV discussion as well as in chapter four.

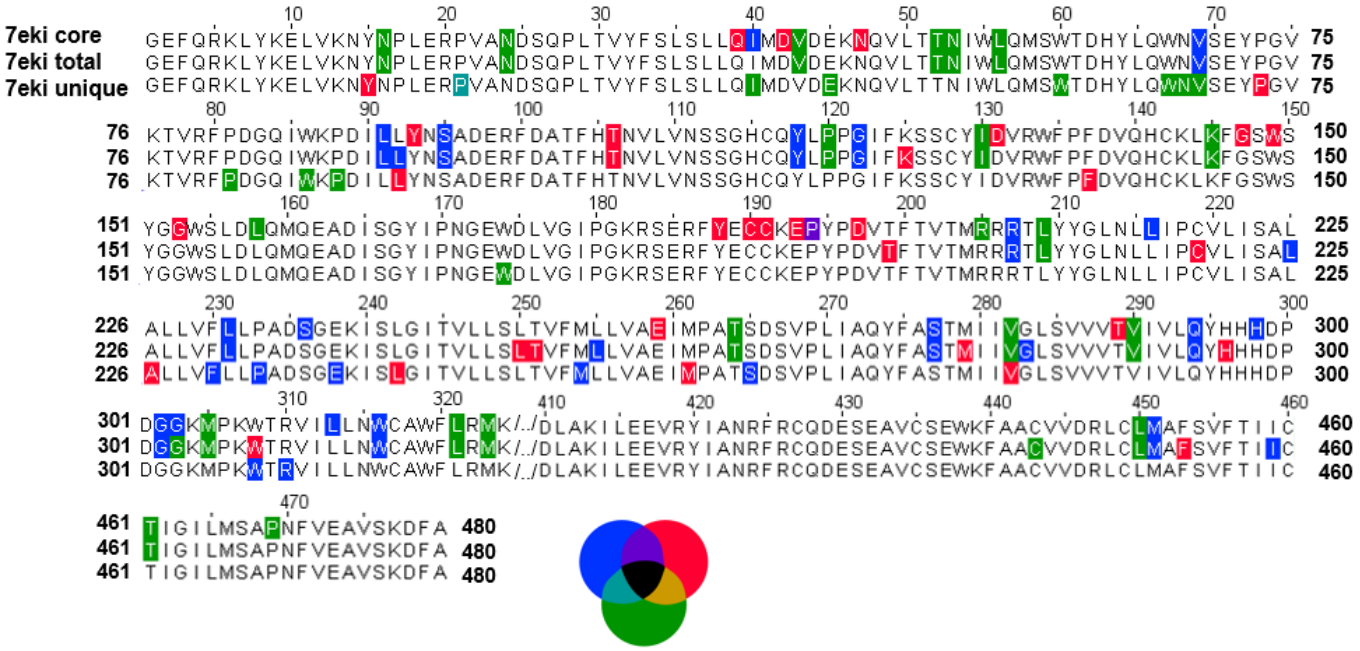

Figure 3.6. Sector Results Mapped onto Sequence of 7eki. 7eki represents a putative resting conformation. Sectors were mapped in wild type mature α7 numbering. Alignments were generated in Jalview[75]. "Unique" represents the set of sequences unique to 7eki, "core" represents the set of sequences common among all conformations and "total" is the sum of the first two sequence sets (Figure 3.5). Complete numbers can be found in Table S1. Colour legend is as follows: red, blue and green represent individual sectors, purple is the combination of red and blue, turquoise is the combination of blue and green, and yellow is the combination of red and green. Black represents residues in all sectors.

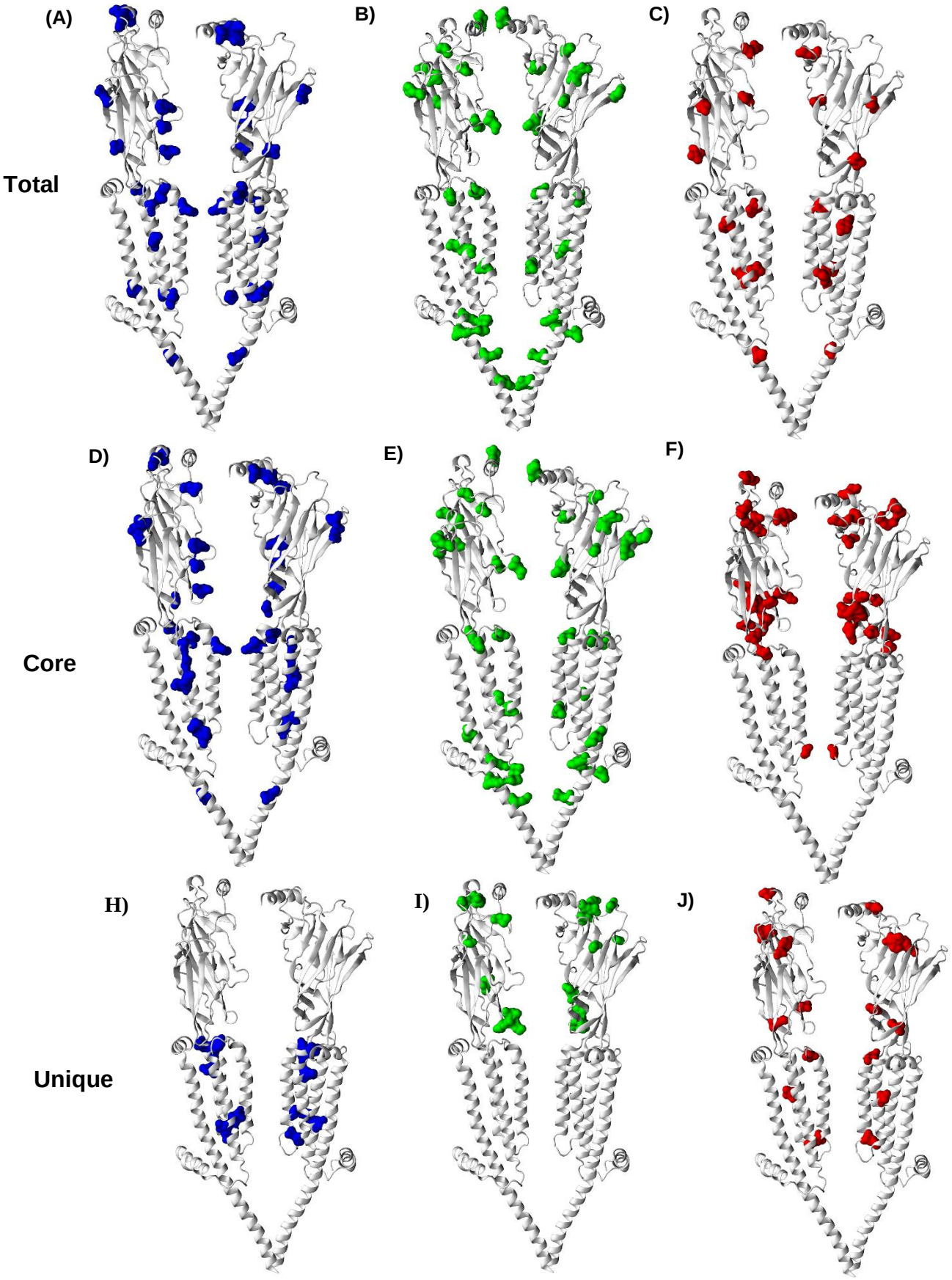

Figure 3.7. Sector Results Mapped as Surface Representation for PDB 7eki from all datasets. Sectors were mapped onto PDB ID 7eki (unliganded resting conformation) in wild type mature α7 numbering. Figures were generated using VMD. **A, B, C)** Blue, Green, and Red sectors for total data set. **D, E, F)** Blue, Green and Red sectors for core data set. **H, I, J)** Blue, Green and Red Sector for unique data set. Complete numbers can be found in Table S1.

The final set of alignments focuses on comparing results from all six conformations against each other, to identify sector residues unique to individual conformations. The goal of this set of alignments is to determine residues that may favour one conformation over another, to generate hypotheses that may be tested experimentally. Each dataset (core, unique and total) has its own alignment containing all six structures. In general, there should be more similarity expected between the residues identified by SCA from conformations with similar RMSDs, which will be discussed in further detail in chapter four. It is also expected that the core and total datasets have more similarity in residues identified by SCA, due to the increased number of shared sequences contained in their alignments, and indeed, when looking at the number of sites specific to one structure vs the complete number of sites identified by SCA, we find that 37% of the sites in the core alignment are specific to one structure, 50% of the sites in the total alignment and 56% in the unique alignment. Note that while there is more similarity in the core dataset, there are still differences because it was aligned to different sequence profiles. Further details on individual residues will be discussed in section IV, while details on candidates for mutation will be discussed in chapter four, with a focus on the unique alignment.

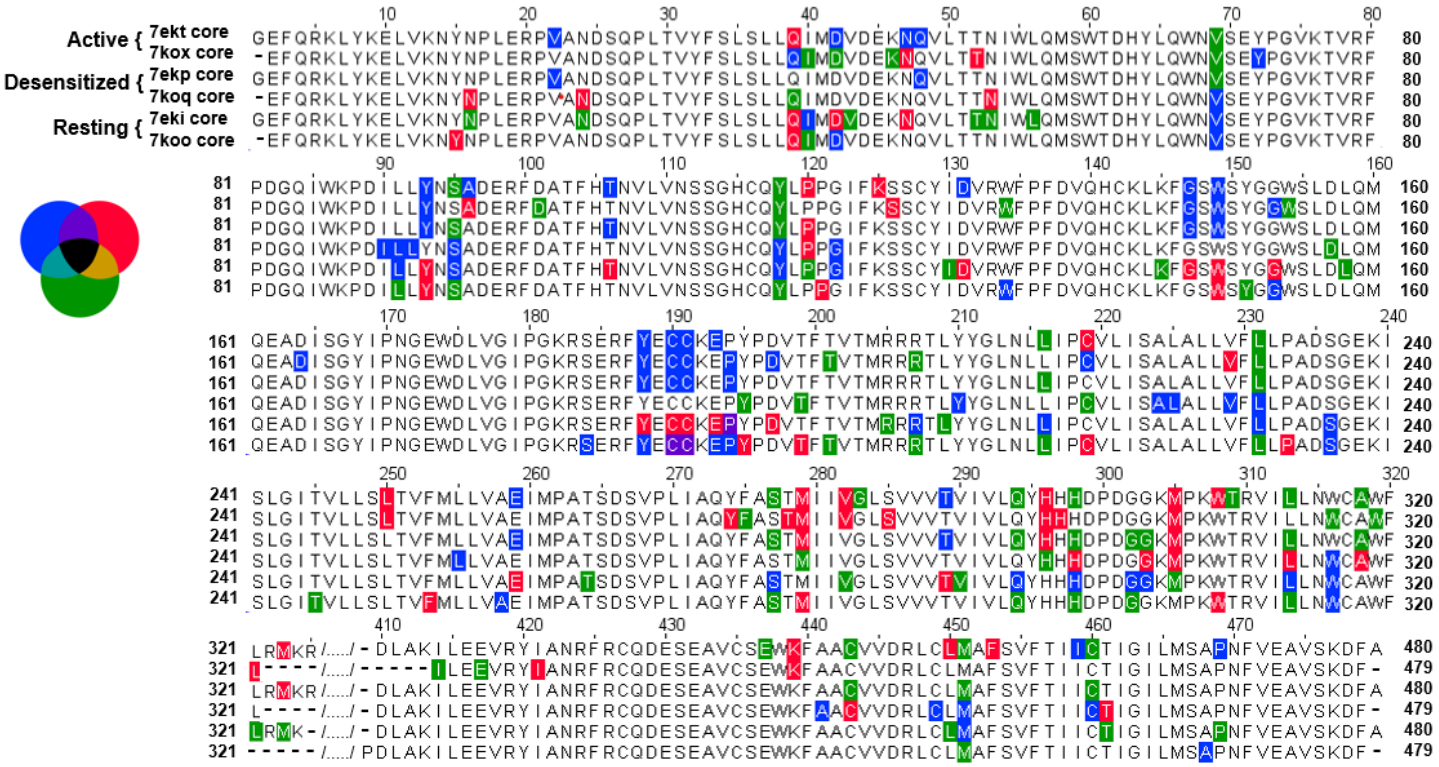

Figure 3.8. Core Sector Results Mapped onto Conformations of α7. Sectors were mapped in wild type mature α7 numbering. Alignments were generated in jalview[75]. This core alignment represents the sectors acquired from the set of sequences common to all structures as detailed in Figure 18. Structures are grouped in the following conformations: Active (7ekt, 7kox), Desensitized (7koq, 7ekp) and resting (7koo, 7eki). Colour legend is as follows: red, blue and green represent individual sectors, purple is the combination of red and blue, turquoise is the combination of blue and green and yellow is the combination of red and green. Black represents residues in all sectors.

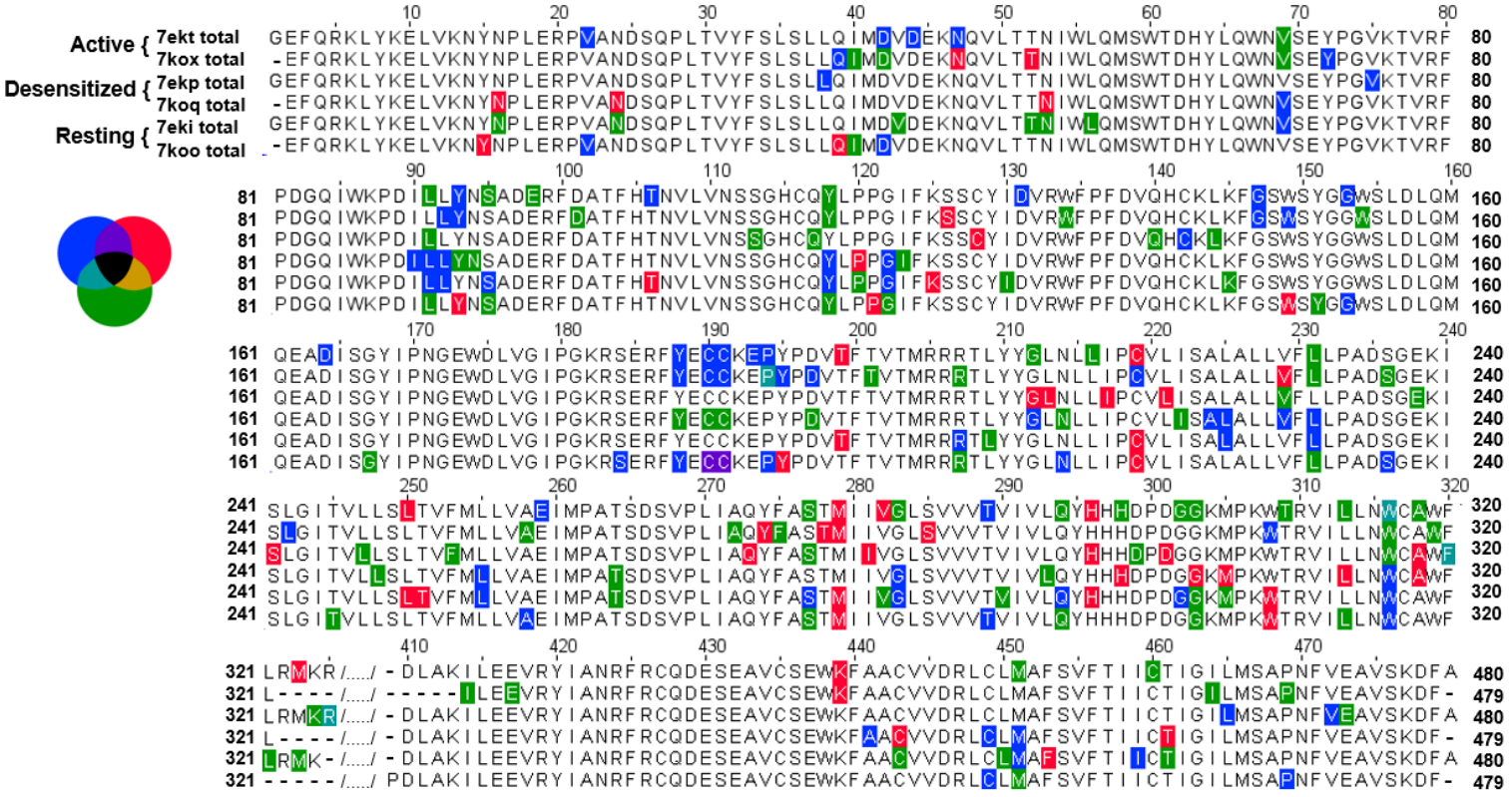

Figure 3.9. Total Sector Results Mapped onto Conformations of α7. Sectors were mapped in wild type mature α7 numbering. Alignments were generated in jalview[75]. This total alignment represents the sectors acquired from the entire set of sequences for each structure as detailed in Figure 17. Structures are grouped in the following conformations: Active (7ekt, 7kox), Desensitized (7koq, 7ekp) and resting (7koo, 7eki). Colour legend is as follows: red, blue and green represent individual sectors, purple is the combination of red and blue, turquoise is the combination of blue and green and yellow is the combination of red and green. Black represents residues in all sectors.

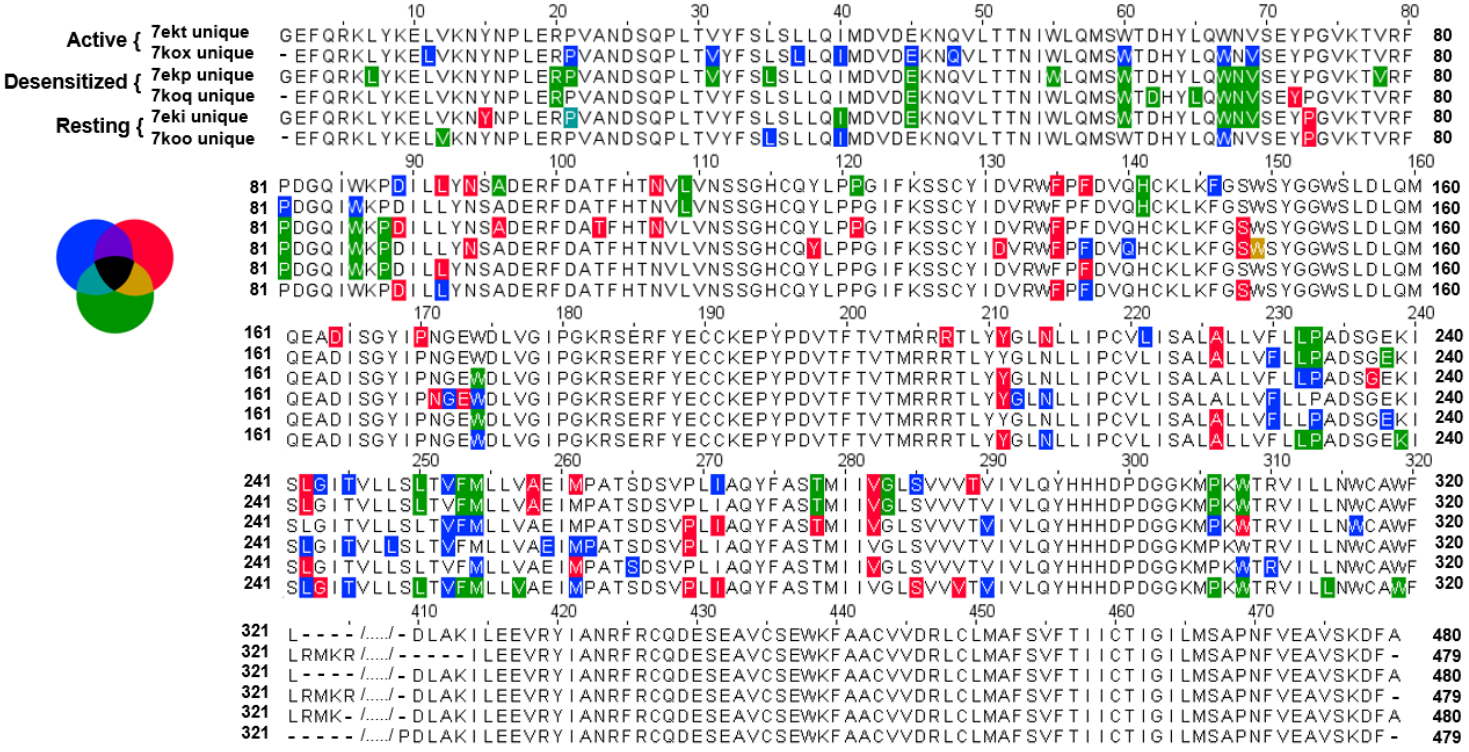

Figure 3.10. Unique Sector Results Mapped onto Conformations of α7. Sectors were mapped in wild type mature α7 numbering. Alignments were generated in jalview[75]. This unique alignment represents the sectors acquired from the set of sequences unique each structure as detailed in Figures 3.4 and 3.5. Structures are grouped in the following conformations: Active (7ekt, 7kox), Desensitized (7koq, 7ekp) and resting (7koo, 7eki). Colour legend is as follows: red, blue and green represent individual sectors, purple is the combination of red and blue, turquoise is the combination of blue and green and yellow is the combination of red and green. Black represents residues in all sectors.

IV – Discussion:

Key areas of the protein are once again shown to be statistically significant in this set of figures, such as loop C, the orthosteric site and the four M1-M4 TMD helices (depicted in Figure 2.9). In addition to the residues discussed in the previous chapter two being recognized as significant, other residues that are well described in the literature can be recognized from the alignment figures (Figures 3.6, 3.7, 3.8, 3.9 and 3.10). D42 and E45 were both identified in a highly conserved manner. These two residues are part of a group of negatively charged residues thought to be involved in the coordination of a divalent cationic binding site[50,54], along with E173 and D44. D42 was identified in Figures 3.6, 3.7, 3.8 and 3.9 while E45 was identified in Figures 3.6, 3.7 and 3.10. E173 was identified in Figure 3.10 while D44 was identified in Figure 3.9. D42 and E45 fall primarily into the blue and green sectors, but do not seem to be coupled with E173 or E98 (mentioned below). Nonetheless, targets for investigating calcium permeability or specific conductance could be investigated from residues coupled to these sites.

In nearly all the alignments at least one of the R205-R207 motif is recognized in a sector. R205 was identified in Figures 3.6, 3.7 and 3.8 while R207 was identified in Figures 3.6, 3.7, 3.8, 3.9 and 3.10. This triple arginine motif is known to be important for receptor activation in general, but each arginine has its own specific binding interactions[50,55,56]. Since this triad of arginines is known to be important for activation, this seems to suggest more importance in maintaining the R conformation as well as the A conformation.

All residues along the M2 pore lining helix that contribute to the ion permeation pathway[28] were recognized at least once across all alignments. E-1'(238) appeared in Figures 3.6, 3.7, 3.9 and 3.10, S2'(241) appeared in Figure 3.9, T6'(245) appeared in Figures 3.8, 3.9 and 3.10, L9'(248) appeared in Figure 3.9 and 3.10, V13'(252)appeared in Figure 3.10, L16'(255) appeared

in Figures 3.6, 3.7, 3.8 and 3.9, and E20'(259) appeared Figures 3.6, 3.7, 3.8, 3.9, and 3.10. K239 (Figure 3.10) located on the M2 helix, is known to make an important salt bridge interaction that is stabilizing in all conformations[57,58]. A258 (Figures 3.8, 3.9 and 3.10) is thought to play an important role in helical capping as well as drug sensitivity[28,59].

A set of residues that were identified as having an impact on receptor potentiation by the allosteric modulator, PNU-120596, the "TSLMF"[60] mutant, were also identified by SCA. A226 and M254 were identified in Figures 3.6, 3.7 and 3.10, I281 was identified in Figure 3.9 and V288 in Figure 3.10. The final residue did not appear in any SCA alignments. M254 is in the green sector in Figure 3.10, while A226 is in the red. The green and red sector couplings in this case could have more candidates for effects on potentiation by PNU-120596.

A468 and P469, the two residues in the latch turn, were both identified in sectors. P469, which is known to be important for channel expression and trafficking[61] was identified in sectors in Figures 3.6, 3.7, 3.8 and 3.9, while A468, which is involved in interactions with the cys-loop, was identified in Figure 3.8[62,63]. W134, which interacts with P469 and is important for the ECD-TMD coupling[61], was also identified in sectors in Figures 3.8 and 3.9. F137, which is in the middle of the latch and makes important electrostatic contacts with the cys-loop[64,65], was identified in sectors in Figures 3.6, 3.7 and 3.10. P194 (Figures 3.6, 3.7, 3.8 and 3.9) F137 and G153 (3.6, 3.7 3.8 and 3.9) are a triad that regulates toxin selectivity in nicotinic receptors[66].

Residues W149 (Figures 3.6, 3.7, 3.8, 3.19 and 3.10), Y195 (Figures 3.6, 3.7, 3.8 and 3.9) and W55 (Figure 3.10), are known to form the α7 agonist-binding pocket[67,68], in addition to residues Y188, C190 and C191 (figures 3.8 and 3.9). Y93 (Figures 3.6, 3.7, 3.8 and 3.9) assists in packing around the quaternary ammonium of acetylcholine (Ach) in the binding pocket[69]. The side chains of L109 (Figure 3.10) and Q117 (Figure 3.9) were found to assist in the packing of the

acetyl moiety of ACh[69]. Residues E193 (Figure 3.6, 3.7, 3.8, and 3.9), Y195 (Figures 3.8 and 3.9) as well Q117 all seem to be involved structurally in hydrogen bonding of α-bungarotoxin to the receptor[28].

To go into more detail about residues that are conformation specific, we can begin by looking at E98. It appears uniquely in one of the A conformation structures (Figure 3.9, 7ekt) is E98. This glutamate was hypothesized by Noviello et. al.[28] to play a role in calcium permeability. However, it has been suggested in other works[70] that E238 (E-1') is more responsible for this interaction. If we look at residues along the M2 permeation pathway next, the L9' residue is recognized as being the desensitization gate of the α7 channel pore. The only alignment in which this residue appears is the alignment for 7koq. This is a proposed D conformation structure, which is logical when considering the nature of the residue. The L9' is in the green sector in the total alignment and the blue sector in the unique alignment. In the unique alignment this residue is coupled to more residues found along the permeation pathway, which could make couplings in this sector more important for slowing desensitization. Other residues that appear more uniquely in one conformation over another are R205 and R207. The two alignments that the R205 and R207 residues are absent from are the two structures in a proposed D conformation.

The overall proportion of residues identified above that all have known functional roles (34) compared to the average number of sites (64) identified in a SCA is about 53%. This is merely an overview of the types of residues SCA is capable of identifying. More details on candidates for mutation and overall conclusions will be discussed in chapter four.

V – Methods:

3.4.1. Structural Redesign and Sequence Searches. The Rosetta all-atom energy function[35] was used to generate a set of 150 redesigned structures for each chain (5 chains per structure) of the following structures: human α7 AChR in complex with α-bungarotoxin (PDB ID: 7koo[28]), unliganded human α7 AChR (PDB ID: 7eki[32]), human α7 AchR in complex with epibatidine (PDB ID: 7koq[28]), human α7 AChR in complex with EVP-6124 (PDB ID: 7ekp[32]), human α7 AChR in complex with epibatidine and PNU-120596 (PDB ID: 7kox[28]) and human α7 AChR in complex with EVP-6124 and PNU-120596 (PDB ID: 7ekt[32]). These were compiled into a single fasta alignment (150 sequences per chain, 30 alignments total) and used to generate a sequence profile with the *hmmbuild* tool (HMMER 3.2.1 suite[71]). Each sequence profile was used as a query for the uniref100[41] database (October 2021) using *hmmsearch*. The number of sequences returned for each chain was between 32 000 and 37 0000. The common hits were taken from each structure to account for minor variation in the Cα backbone (Figure 3.4). This generated three datasets: a core set of sequences that were common to all searches and thus backbones, (2) a set of sequences unique to each structure and (3) the combination of the previous two sets into what we call the total set of sequences for each structure (Figures 3.4 and 3.5).

3.4.2 Multiple Sequence Alignments. Each dataset was locally aligned against sequence profiles built from the alignment of all redesigned for each structure (150 x 5, 750 sequences for 6 alignments total) using *hmmalign*, to be more representative of the entire structure. Sequences were then clustered to reduce repetition, using Cluster Database at High Identity Tolerance (CD-HIT)[43][44]. An identity cut-off was set to 95%, that is, no sequence more than 95% identical to any other sequence was kept, and a minimum sequence length cut-off was set to 400 amino acids. This

was to ensure a balance between quality of information and amount of information. An alignment was then run using MAFFT[45] with global alignment parameters in the GINSI algorithm and in the input order. These alignments were then pruned to the wild type α7 sequence (aligned maximally to best fit the α7 sequence) using perl scripts (supplementary information). This final set of multiple sequence alignments (MSAs) was then used to perform SCA.

3.4.3 Statistical Coupling Analysis. The SCA v.5.0 toolbox[46] was used to analyse the final MSAs, ignoring positions with more than 40% gaps. Multiple sectors were suggested so the top 4 eigenmodes were chosen for independent component analysis[39], which in turn yielded 3 components for sector consideration[39] A cutoff of 0.98 was considered to define coupled sites. This was chosen to ensure only the top 2% of correlations were considered for sector analysis, in an effort to balance the quality of sector information with the quantity. Other cutoffs were not tested due to time constraints.

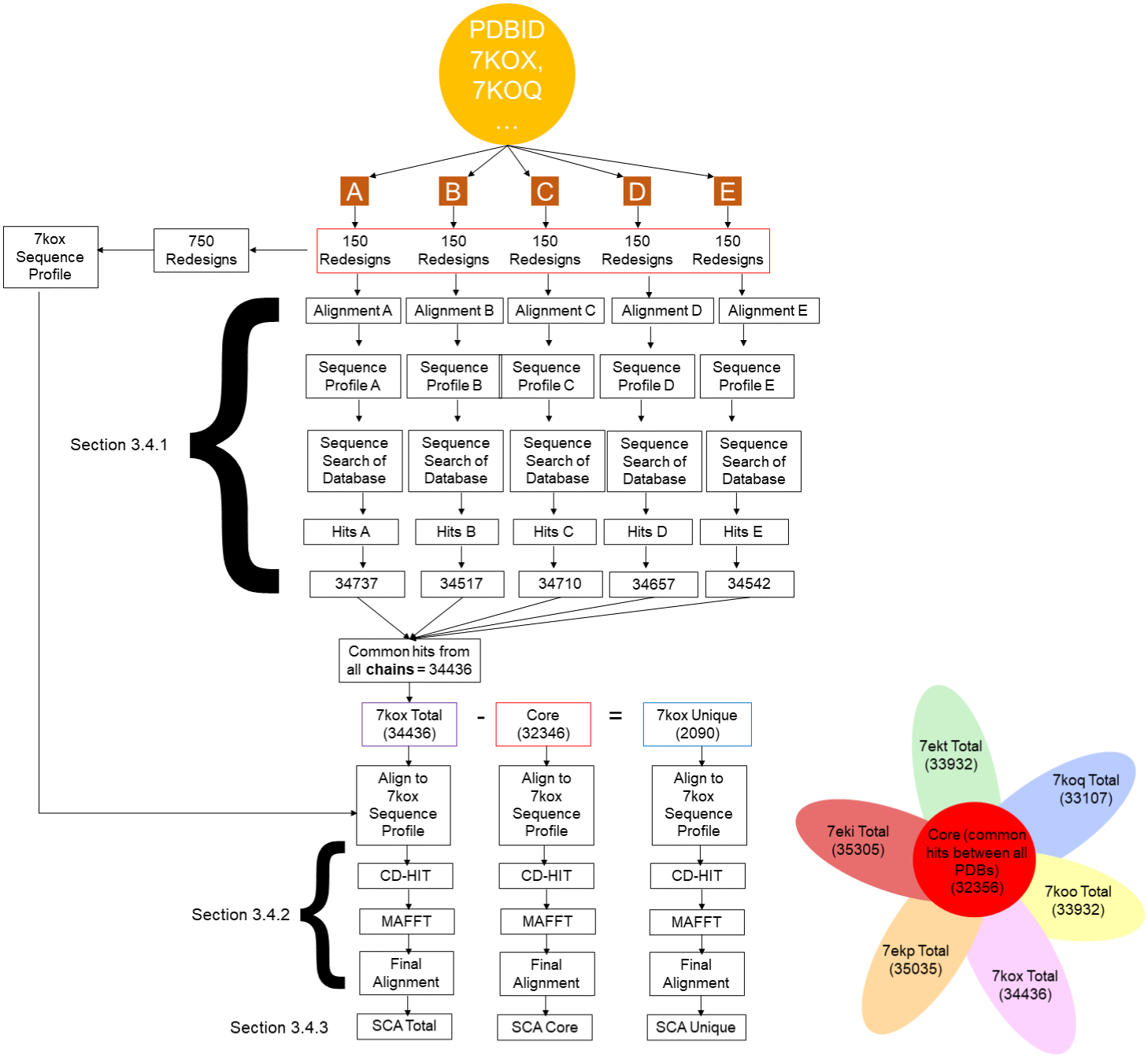

Figure 3.11. Flowchart Representing Methods Used to Obtain SCA Results. Flowchart for a single PDB ID with five chains (7kox, chains A-E; putative active conformation). A similar procedure was ran on all six available structures of α7 (PDB IDs: 7kox (shown), 7koq, 7koo, 7ekt, 7ekp and 7eki yielding: 6 x 5 sets of redesign alignments, 6 x 5 chain sequence profiles, 6 x 5 sequence searches, 6 x 5 sets of hits, six total datasets, six unique datasets, one core dataset, (denoted by the red circle in the venn diagram) 6 x 3 structure profile alignments and 6 x 3 SCAs.

CHAPTER 4: Discussion and Conclusion

I – Sequence based SCA vs Structure based

The objective behind using two different SCA methods is to give a more complete profile of important residues within the α7 protein, as well as to generate information sets for the general sequence versus specific structures. The sequence-based method gave us a set of residues that in theory are important to the overall protein but are not necessarily conformation specific. It also provides a complete sequence profile since the entire sequence is used in the analysis. The structure-based method generates information tailored to each structure and could therefore potentially provide insight into residues important for stabilizing one conformation over another.

Both of these methods are powerful in their own right, but there are limitations to each. The advantage of the sequence-based method is of course a more complete sequence, while the structure-based method only analyzes what is present in the structure. This means that any region with gaps will be excluded. That being said, since the structure-based method generates an alignment that is more "tailored" to a unique structure, the alignments tend to contain fewer blank positions, allowing SCA to analyze proportionally more positions for potential importance.

As we can see from the structural images and tables (Figures 2.7, 2.8 and 3.6 to 3.10), there are more residues present in the sectors for the structure based SCA. This is due to the way that SCA operates as any column with more than 40% gaps will be excluded from analysis. This is likely due to the alignments in the structure based SCAs being much more tailored to the conformations of α7 and therefore not as large, which leads to less ambiguity at more positions. One similarity across both sets of results is the absence of residues of importance in the intracellular domain. This is due to the lack of verified structural information available for that region as well as low conservation and poor sequence alignment amongst pLGICs in this area.

While there are fewer residues of interest in the sequence-based SCA, regions of interest such as loop C, M1-M4 helical region and MX helix still come up in both types of analyses, lending value to both methods.

II – SCA can identify residues unique to a conformation

The final set of results aimed to compare residues of significance across different conformations, with the goal of proposing a set of mutations to test experimentally, as well as identifying residues unique to a single conformation. The core, unique, and total datasets (Figures 3.8-3.10) all yield slightly different SCA results. In the case of the core dataset, only 37% of sites identified by SCA were found to be limited to a single structure. This is in keeping with the expectation that there will be fewer unique sites in the core alignments since they are the same set of sequences, while also having minor variation between alignments due to pairwise alignment to each individual sequence profile. The total and unique datasets yielded 50% and 56% of sites limited to a single structure respectively. The total dataset yields a number of sites closer to the unique since it contains all the sequences from each sequence search, however the unique dataset yields the highest number of sites limited to single structures and so it is to this alignment we will turn our attention (Figure 3.10), with a focus on residues not well described in the literature.

There are three possible ways to develop hypotheses concerning substitutions of residues at these sites for experimental testing. The first is by looking at the physico-chemical properties of the residues occupying the site of interest and choosing mutations with the diverging properties, such as charge swaps in the case of positive and negative residues. The second way is to look at the sets of natural sequence alignments to investigate natural variation of amino acids at these sites. The final way would be to look into the Rosetta redesigned sequences to see what amino acids have been “fitted” in these sites by Rosetta based on 3D backbone restraints. The next paragraphs

detail some examples of logical mutations based on the physico-chemical properties of the amino acids.

If we begin by looking at residues that appear solely in resting conformation structures, V12, Y15, P73, V257, S265, R310, L314 and W319 all appear solely in the two resting conformation structures. P73 falls in the ECD, in a loop at the very start of the β-sandwich. Proline is a helix breaker, meaning that it could play a role in structural organisation of this loop. An experiment to test this would be mutation to glycine, which is flexible due to lack of sterics. S265 falls within the ECD-TMD coupling region, meaning it could be important for coupling ligand binding to channel gating. A way to test this would once again be mutating to a nonpolar residue such as valine. V257 is at the top of the M1-M4 bundle on the M2 helix and points inwards towards the other three helices. There appears to be hydrophobic packing at this region which could be important for the stabilization of this conformation in the resting state. A substitution of a polar residue, such as serine, or a charged residue such as aspartate would be a good test of this. R310, L314 and W319 are all located in the MX helix, which as previously discussed, is important for receptor assembly. A charge swap experiment at R310 for example would be a way to test functionality.

L7, R20, D62, L65, Y72, V78, T103, D131, Q140, N171, G172, G212, P262, and W316 all appear solely in the two desensitized conformation structures. R20 appears to interact with the top of the β8-β9 loop, which could be important for maintaining the architecture of this region. A charge swap would be a suitable experiment to test this. The same could be said of D62, which appears to interact with loops in this area. T103 sits inside the pore and could contribute to loop stabilization or rearrangement. A mutation to a nonpolar residue such as valine would be a good test. D131, N171, and G172 are all located around the proposed divalent cation binding site located

between subunits at the TMD ECD interface. A charge swap of this area would be a useful way to investigate the charged residues. Mutation of G172 to something sterically hindered such as alanine would be a useful experiment. G212 is located in the M3 helix and likely contributes to flexibility but a steric hindrance would be a good way to test this. P262 is located in the M1-M2 loop and likely contributes to maintenance of architecture since proline is a helix breaker. Swapping the proline for an alanine would be a mutation to test this. W319 sticks out from the MX helix, which again could make it important for receptor assembly. A test would be to mutate it for a nonpolar, nonaromatic residue.

Finally, the residues that appear solely in the two active conformations are L11, L37, Q48, H141, F146, D164, P170, L221 and T289. H141, F146, D164 and P170 are all located in the β-sandwich region of the ECD, meaning they could be important for structure, ligand recognition or subunit interactions. Some mutations to test function would be charge swaps for H141 and D164, mutation of F146 to tyrosine, which is a similar sidechain but polar and mutation of P170 to a more flexible residue such as glycine.

III – Next steps: what comes after SCA?

There are many residues from the literature that have already been shown to impact the structure and function of α7, and some of these are in agreement with residues we have identified from our analysis. However, it is our hope that residues we have identified in the previous section can be used to predict mutations that will stabilize different conformations of α7. The set of residues we propose for mutations are V12, Y15, P73, V257, S265, R310, L314 and W319 for the resting conformation, L7, R20, D62, L65, Y72, V78, T103, D131, Q140, N171, G172, G212, P262, and W316 for the desensitized conformation and L11, L37, Q48, H141, F146, D164, P170,

L221 and T289 for the active conformation. Residues are depicted structurally in Figure 4.1 and mapped onto the α7 sequence in Figure 4.2.

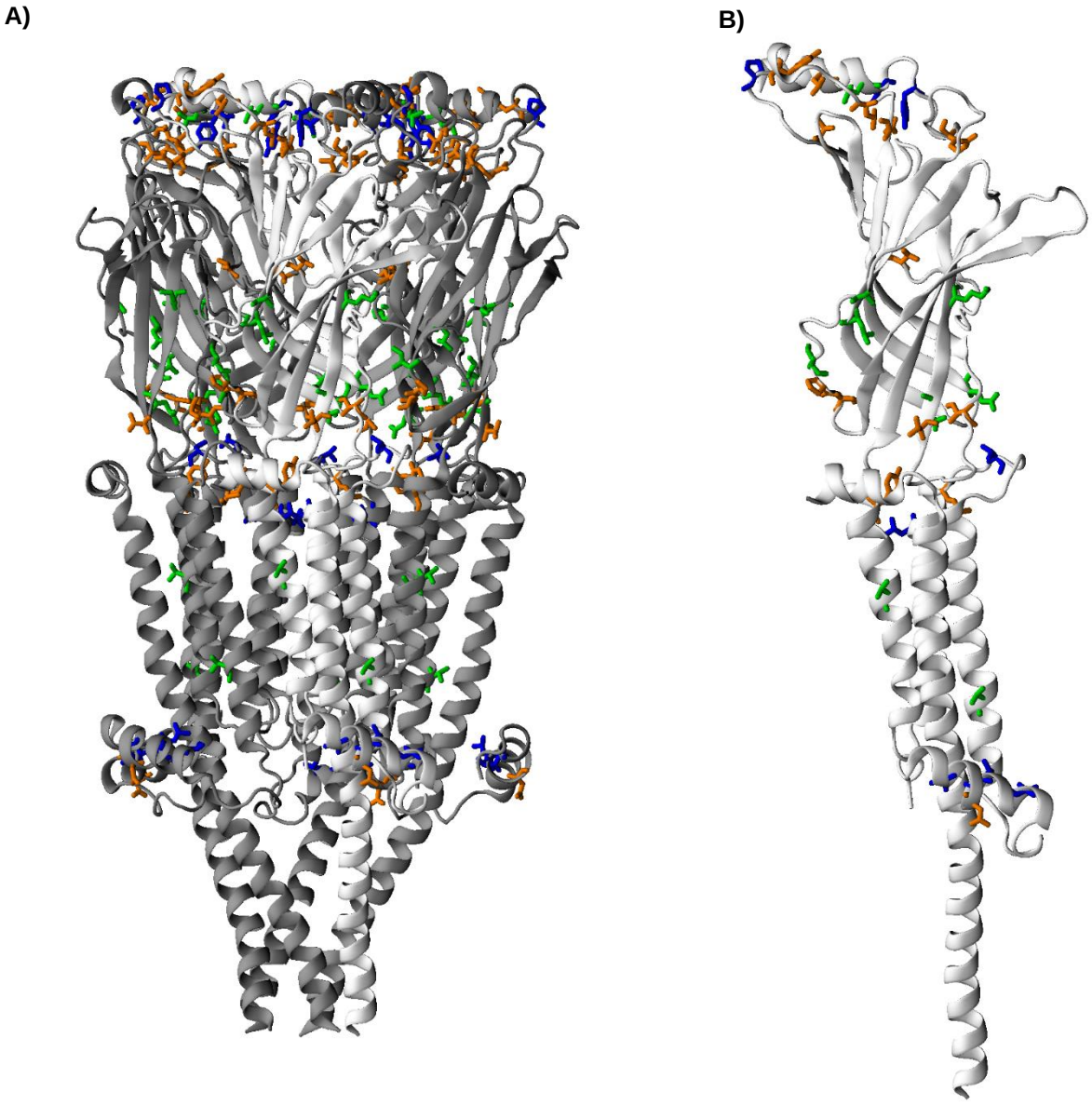

Figure 4.1. Residues identified for mutation by SCA of unique dataset. Residues mapped in VMD onto unliganded resting conformation of α7 (PDBID: 7eki). (A) Overview of sites identified for mutation coloured according to the following: green (active), orange (desensitized) and blue (resting). (B) single chain of structure with same colouring scheme as depicted in (A).

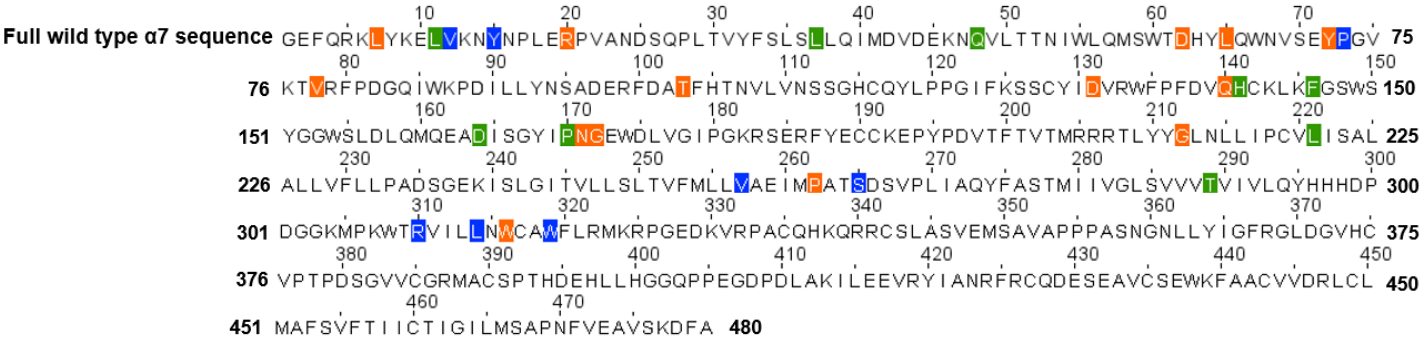

Figure 4.2. Residues identified for mutation by SCA of Unique Dataset Mapped onto α7 sequence. Residues are coloured according to conformations in which they were identified: Active in green, Desensitized in orange and Resting in blue. Alignment was generated in Jalview[75].

IV – SCA can be applied to other proteins

SCA is useful tool across many different protein families. Previous work has shown that SCA can be applied to both small and large proteins so long as there are sufficient sequences to construct homologous alignments. SCA has previously been applied to G-protein coupled receptors[72], Serine proteases[39] and the individual PDZ domain[38]. Sequence based SCA could be applied to any protein with a sequence available, however our structure-based method could be applied to any protein with multiple structural conformations available. The key to good SCA information is the availability of a large, homologous sequence space and so this method would work best for proteins or domains in large, diverse families.

V – Effect of Cα Backbone RMSD on Shared SCA Results

A logical comparison that can be made from the final sets of statistically coupled residues is to see if there is any correlation between the number of residues shared between two structures and their RMSDs. This was plotted as a percentage of residues in common to account for differences in the number of statistically coupled sites vs the Cα RMSD in angstroms (Figures 4.2, 4.3 and 4.4). Note that the core dataset is represented in red, the total in purple and the unique in blue. We expect to see a linear relationship between the % of shared residues and the RMSD. Unfortunately, there does not seem to be a clear correlation between the RMSDs of structures and the number of coupled sites they share. This can be attributed to variations in the alignments used to perform the final sets of SCA. An argument could be made that the core Figure (Figure 4.3)

represents a very slight correlation, which could be due to the high number of shared sequences used in alignments, but it is not clear if this difference is statistically meaningful.

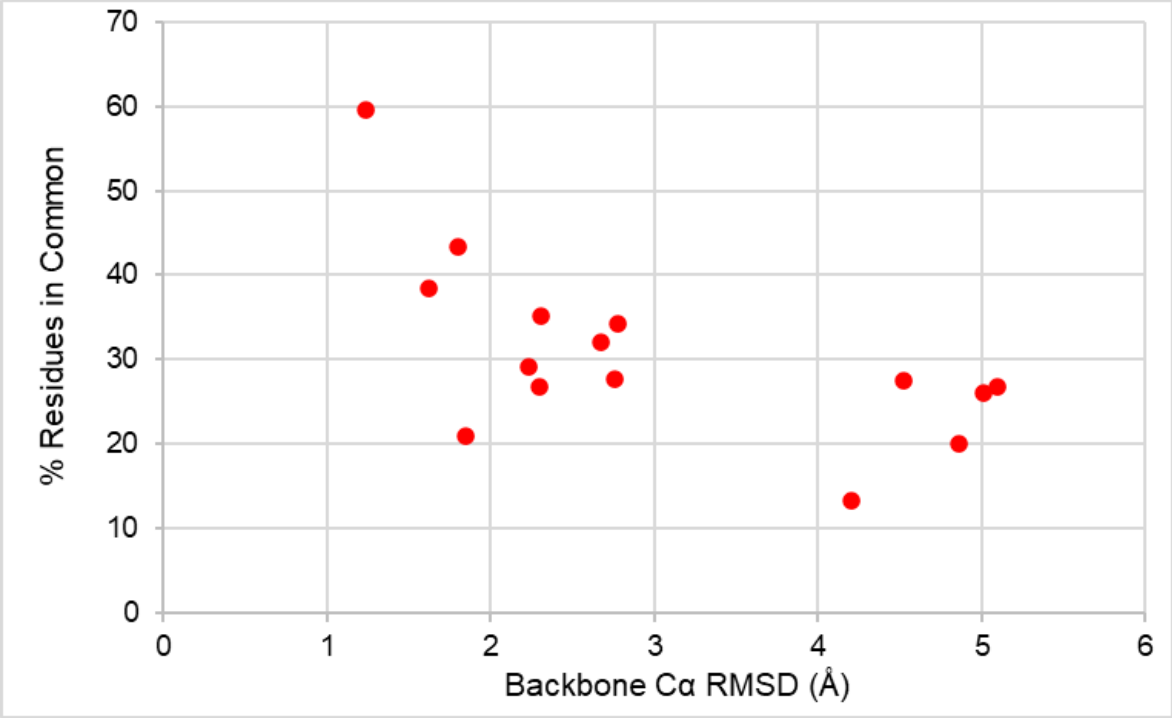

Figure 4.3. Percentage of Coupled Sites in Common plotted against RMSD of the core dataset. Structures were aligned and RMSD calculated in VMD. % residues in common was standardized against the complete number of coupled sites in order to better compare individual data points.

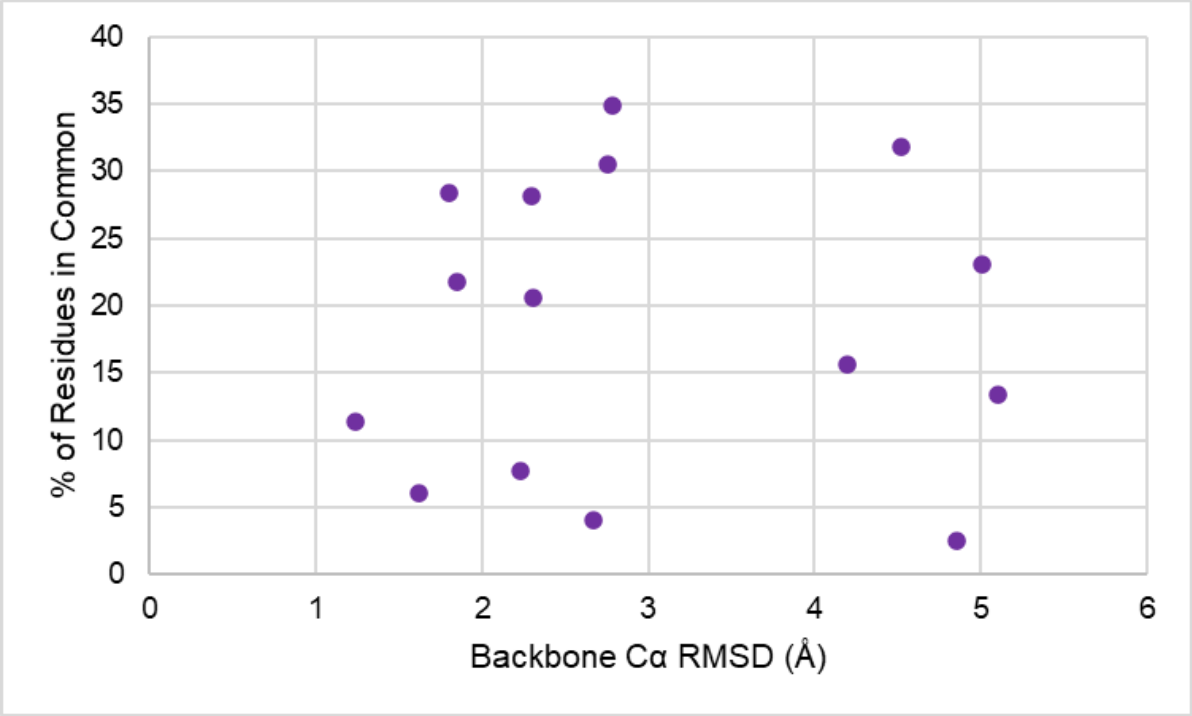

Figure 4.4. Percentage of Coupled Sites in Common plotted against RMSD of the total dataset. Structures were aligned and RMSD calculated in VMD. % residues in common was standardized against the complete number of coupled sites in order to better compare individual data points.

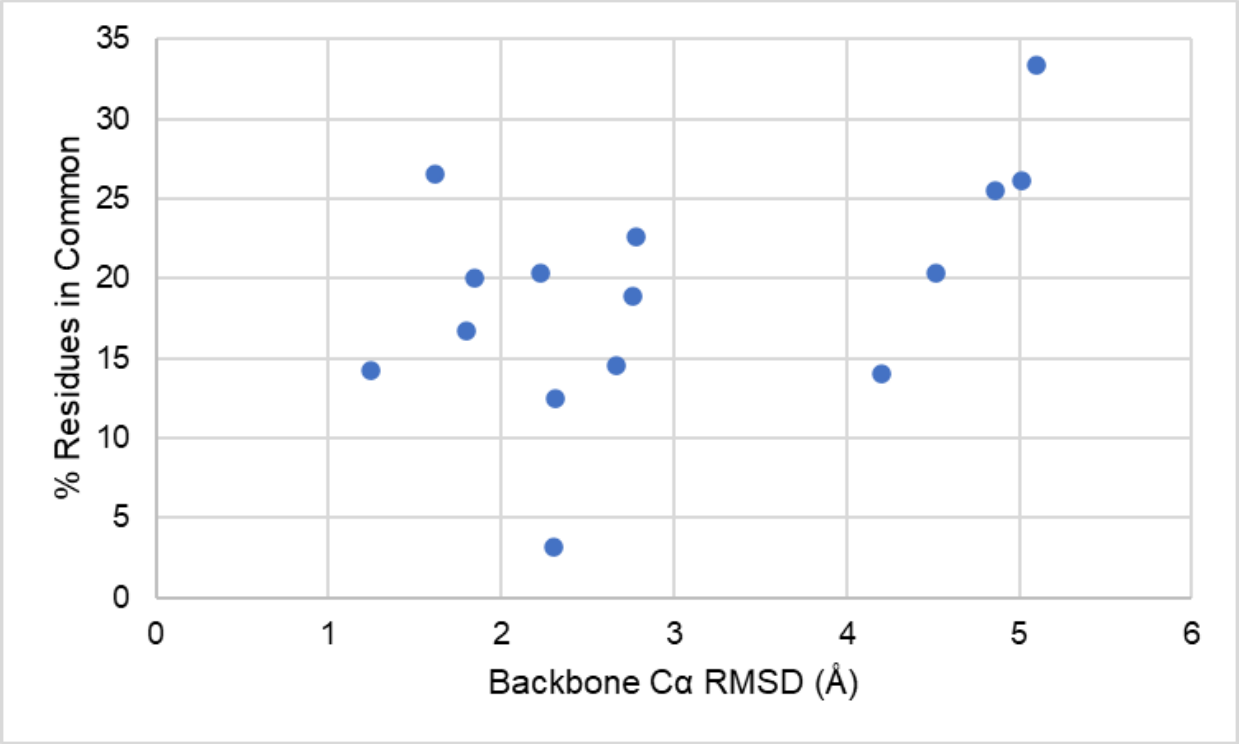

Figure 4.5. Percentage of Coupled Sites in Common plotted against RMSD of the unique dataset. Structures were aligned and RMSD calculated in VMD. % residues in common was standardized against the complete number of coupled sites in order to better compare individual data points.

VI – Limitations of SCA

SCA is a powerful method, but it is limited by its input. In other words, the SCA is robust to the MSA upon which it is run, meaning that the quality of the results is dependent upon the quality of the input data. Due to the variation in alignment size and quality for the inputs of the core, unique and total datasets, there is a lack of uniformity across the sector colours and the residues that fall within those sectors. For this reason, trying to ascribe one function to one sector was not useful in our analysis of structural data, despite the identification of residues of importance by SCA, as discussed in chapters two and three. We also have a lack of experimental data in relation to individual sectors in this work and so future experiments would be needed to mutate individual sectors to try to ascribe particular functions in α7 to a given sector. Colouring is also arbitrarily assigned in this case and so we cannot ascribe meaning to residues falling into one sector over another or switching sectors between different datasets at this stage of analysis. This is more likely a consequence of the method rather than residues switching functional roles, but more data is needed to determine this.

Another limitation of this method is the subjective nature of some points of the analysis. As seen in Figure 2.5, there can be multiple eigenmodes of significance which can represent more than one group of amino acids, and the cutoff imposed for this is subjective[37,73]. In our case, for consistency across all datasets, we chose to analyze the top four eigenmodes for independent component analysis which yielded three components to consider for sector formation. This was based on the majority of the datasets having 3 to 4 modes of significance. However, were we to do this analysis again, the number of sectors could change depending on the value set for independent component analysis, which could yield more specified functions in each sector. The subjective nature of some parameters choices is also a limitation. In chapter two there is a single coupling

present for blue rather than a true sector (7-10% of residues in sequence). This is a result of the 98% cutoff we imposed, but had we varied this number more and with more time to test parameters it is likely that other residues of importance would be identified. Similarly, with our kmax of 4, we are only considering positions that have fewer than 40% gaps in the alignment. Were we to alter this number, it is possible we may end up with different positions for SCA consideration.

The main limitation of our structure based SCA is the lack of information available for areas not characterized in the structures, or areas that have low conservation amongst the pLGIC superfamily. The analysis for regions that are structurally well-defined and homologous was largely in agreement with what is currently known and accepted about the α7 nAChR, but there could be information missing due to the absence of the ICD in the structures and the low conservation of this region.

VII – AlphaFold: from sequence to structure and back again

The coupling of computational techniques to experimentally determined data continues to reveal more about the natural world, and as the field of bioinformatics continues to expand so too does the range of methods available. One such method that turned the world of protein science on its head in 2020 was the rise of AlphaFold, a machine learning algorithm capable of folding a protein sequence into a structure with remarkable accuracy in some cases[42]. While this method does not replace the importance of experimental structure determination, it can be used to provide valuable hypotheses for experimental testing. The inner workings of AlphaFold remain complex due to the nature of the machine learning algorithm, however we do know that the first steps are similar to our own sequence search methods. AlphaFold begins by searching for sequences homologous to a specified input, which it then uses to construct a sequence profile similar to our own. This sequence profile is then used for a template search using the exact same method we used

to retrieve our initial sequence sets. Preliminary comparisons of an AlphaFold structure of the α7 monomer to any of the experimentally determined conformation specific structures suggests that AlphaFold does not favour one conformation in its final output. We therefore pose the following question: if we bypass the initial sequence search steps of AlphaFold, does a structurally biased input alignment, such as the ones we have constructed for the above work, push AlphaFold into folding a structure with a specific conformation? This is an important future direction and if it is indeed possible, it would make AlphaFold an even more powerful tool in the world of protein science.

VIII – Conclusion

To conclude, we have validated that SCA is a method capable of identifying residues with pre-determined experimental importance in the human α7 nAChR. We have also generated a number of hypotheses concerning residues that may functionally contribute to one conformation over another, which can be experimentally tested to determine if these differences are in fact meaningful. Our hypothesis concerning the retrieval of unique sequences by specific conformations was also determined to be true. The results of this work may provide an important step towards our understanding of the structure-function relationships in the human α7 nAChR.

1. Alexander, S., Mathie, A. & Peters, J. Ion Channels. *Br J Pharmacol* **164**, S137 (2011).

2. Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science* **280**, 69–77 (1998).

3. Hilf, R. J. C. & Dutzler, R. X-ray structure of a prokaryotic pentameric ligand-gated ion channel. *Nature 2008 452:7185* **452**, 375–379 (2008).

4. Cosens, D. J. & Manning, A. Abnormal Electroretinogram from a Drosophila Mutant. *Nature 1969 224:5216* **224**, 285–287 (1969).

5. Vandewauw, I. *et al.* A TRP channel trio mediates acute noxious heat sensing. *Nature 2018 555:7698* **555**, 662–666 (2018).

6. Hartzell, H. C. Chloride Channels: An Historical Perspective. *Physiology and Pathology of chloride transporters and channels in the nervous system* 1–15 (2010) doi:10.1016/B978-0-12-374373-2.00001-7.

7. Accardi, A. Structure and Function of CLC Chloride Channels and Transporters. *Advances in Molecular and Cell Biology* **38**, 59–82 (2006).

8. Bernèche, S. & Roux, B. A Gate in the Selectivity Filter of Potassium Channels. *Structure* **13**, 591–600 (2005).

9. Åqvist, J. & Luzhkov, V. Ion permeation mechanism of the potassium channel. *Nature 2000 404:6780* **404**, 881–884 (2000).

10. Olsen, R. W. & Sieghart, W. International Union of Pharmacology. LXX. Subtypes of gamma-aminobutyric acid(A) receptors: classification on the basis of subunit composition, pharmacology, and function. Update. *Pharmacol Rev* **60**, 243–260 (2008).

11. Belelli, D. *et al.* Extrasynaptic GABAA receptors: form, pharmacology, and function. *J Neurosci* **29**, 12757–12763 (2009).

12. Walstab, J., Rappold, G. & Niesler, B. 5-HT(3) receptors: role in disease and target of drugs. *Pharmacol Ther* **128**, 146–169 (2010).

13. Barnes, N. M., Hales, T. G., Lummis, S. C. R. & Peters, J. A. The 5-HT3 receptor--the relationship between structure and function. *Neuropharmacology* **56**, 273–284 (2009).

14. Lynch, J. W. Native glycine receptor subtypes and their physiological roles. *Neuropharmacology* **56**, 303–309 (2009).

15. Yevenes, G. E. & Zeilhofer, H. U. Allosteric modulation of glycine receptors. *Br J Pharmacol* **164**, 224–236 (2011).

16. Millar, N. S. & Gotti, C. Diversity of vertebrate nicotinic acetylcholine receptors. *Neuropharmacology* **56**, 237–246 (2009).

17. Changeux, J. P. Allosteric receptors: from electric organ to cognition. *Annu Rev Pharmacol Toxicol* **50**, 1–38 (2010).

18. Alexander, S. P. H. *et al.* THE CONCISE GUIDE TO PHARMACOLOGY 2017/18: Ligand-gated ion channels. *Br J Pharmacol* **174**, S130–S159 (2017).

19. Lara, C. O., Burgos, C. F., Moraga-Cid, G., Carrasco, M. A. & Yévenes, G. E. Pentameric Ligand-Gated Ion Channels as Pharmacological Targets Against Chronic Pain. *Front Pharmacol* **11**, (2020).

20. Changeux, J. P. & Paas, Y. Nicotinic Acetylcholine Receptors. *Encyclopedia of Neuroscience* 1129–1133 (2009) doi:10.1016/B978-008045046-9.01127-X.

21. Bondarenko, V. *et al.* Structures of highly flexible intracellular domain of human α7 nicotinic acetylcholine receptor. *Nature Communications 2022 13:1* **13**, 1–9 (2022).

22. Changeux, J. P., Kasai, M. & Lee, C. Y. Use of a Snake Venom Toxin to Characterize the Cholinergic Receptor Protein. *Proc Natl Acad Sci U S A* **67**, 1241 (1970).

23. Albuquerque, E. X., Pereira, E. F. R., Alkondon, M. & Rogers, S. W. Mammalian nicotinic acetylcholine receptors: From structure to function. *Physiol Rev* **89**, 73–120 (2009).

24. Albuquerque, E. X. *et al.* Acetylcholine Receptor and Ion Conductance Modulator Sites at the Murine Neuromuscular Junction: Evidence from Specific Toxin Reactions. *Proceedings of the National Academy of Sciences* **70**, 949–953 (1973).

25. Rahman, M. M. *et al.* Structural mechanism of muscle nicotinic receptor desensitization and block by curare. *Nature Structural & Molecular Biology 2022 29:4* **29**, 386–394 (2022).

26. Broide, R. S. & Leslie, F. M. The α7 nicotinic acetylcholine receptor in neuronal plasticity. *Molecular Neurobiology 1999 20:1* **20**, 1–16 (1999).

27. Fabian-Fine, R. *et al.* Ultrastructural Distribution of the α7 Nicotinic Acetylcholine Receptor Subunit in Rat Hippocampus. *Journal of Neuroscience* **21**, 7993–8003 (2001).

28. Noviello, C. M. *et al.* Structure and gating mechanism of the α7 nicotinic acetylcholine receptor. *Cell* **184**, 2121-2134.e13 (2021).

29. de Jaco, A., Bernardini, L., Rosati, J. & Tata, A. M. Alpha-7 Nicotinic Receptors in Nervous System Disorders: From Function to Therapeutic Perspectives. *Cent Nerv Syst Agents Med Chem* **17**, (2017).

30. Wallace, T. L. & Porter, R. H. P. Targeting the nicotinic alpha7 acetylcholine receptor to enhance cognition in disease. *Biochem Pharmacol* **82**, 891–903 (2011).

31. Ma, K. G. & Qian, Y. H. Alpha 7 nicotinic acetylcholine receptor and its effects on Alzheimer's disease. *Neuropeptides* **73**, 96–106 (2019).

32. Zhao, Y. *et al.* Structural basis of human α7 nicotinic acetylcholine receptor activation. *Cell Research 2021 31:6* **31**, 713–716 (2021).

33. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science (1979)* **181**, 223–230 (1973).

34. daCosta, C. J. P., Free, C. R., Corradi, J., Bouzat, C. & Sine, S. M. Single-Channel and Structural Foundations of Neuronal α7 Acetylcholine Receptor Potentiation. *Journal of Neuroscience* **31**, 13870–13879 (2011).

35. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* **13**, 3031–3048 (2017).

36. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).

37. Süel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology 2002 10:1* **10**, 59–69 (2002).

38. Lockless, S. W. & Ranganathan, R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science (1979)* **286**, 295–299 (1999).

39. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774 (2009).

40. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).

41. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).

42. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873* **596**, 583–589 (2021).

43. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).

44. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

45. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066 (2002).

46. McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142 (2012).

47. Huang, S. *et al.* Complex between α-bungarotoxin and an α7 nicotinic receptor ligand-binding domain chimaera. *Biochem J* **454**, 303–310 (2013).

48. Karlin, A. Structure of nicotinic acetylcholine receptors.

49. Stokes, C., Treinin, M. & Papke, R. L. Looking below the surface of nicotinic acetylcholine receptors. *Trends Pharmacol Sci* **36**, 514–523 (2015).

50. Cheng, X., Wang, H., Grant, B., Sine, S. M. & McCammon, J. A. Targeted Molecular Dynamics Study of C-Loop Closure and Channel Gating in Nicotinic Receptors. *PLoS Comput Biol* **2**, e134 (2006).

51. Corringer, P.-J., le Novè, N. & Changeux, J.-P. *NICOTINIC RECEPTORS AT THE AMINO ACID LEVEL*. *Annu. Rev. Pharmacol. Toxicol* vol. 40 (2000).

52. Rudell, J. C., Borges, L. S., Yarov-Yarovoy, V. & Ferns, M. The MX-Helix of Muscle nAChR Subunits Regulates Receptor Assembly and Surface Trafficking. *Front Mol Neurosci* **13**, 48 (2020).

53. Martínez-Castilla, L. P. & Rodríguez-Sotres, R. A Score of the Ability of a Three-Dimensional Protein Model to Retrieve Its Own Sequence as a Quantitative Measure of Its Quality and Appropriateness. *PLoS One* **5**, e12483 (2010).

54. Galzi, J. L., Bertrand, S., Corringer, P. J., Changeux, J. P. & Bertrand, D. Identification of calcium binding sites that regulate potentiation of a neuronal nicotinic acetylcholine receptor. *EMBO J* **15**, 5824 (1996).

55. Lee, W. Y., Free, C. R. & Sine, S. M. Binding to gating transduction in nicotinic receptors: Cys-loop energetically couples to pre-M1 and M2-M3 regions. *J Neurosci* **29**, 3189–3199 (2009).

56. Alves, D. S., Castello-Banyuls, J., Faura, C. C. & Ballesta, J. J. An extracellular RRR motif flanking the M1 transmembrane domain governs the biogenesis of homomeric neuronal nicotinic acetylcholine receptors. *FEBS Lett* **585**, 1169–1174 (2011).

57. Mesoy, S., Jeffreys, J. & Lummis, S. C. R. Characterization of Residues in the 5-HT3 Receptor M4 Region That Contribute to Function. *ACS Chem Neurosci* **10**, 3167–3172 (2019).

58. da Costa Couto, A. R. G. M., Price, K. L., Mesoy, S., Capes, E. & Lummis, S. C. R. The M4 Helix Is Involved in α7 nACh Receptor Function. *ACS Chem Neurosci* **11**, 1406–1412 (2020).

59. Ho, K. K. & Flood, P. Single Amino Acid Residue in the Extracellular Portion of Transmembrane Segment 2 in the Nicotinic 7 Acetylcholine Receptor Modulates Sensitivity to Ketamine. *Anesthesiology* **100**, 657–62 (2004).

60. daCosta, C. J. P., Free, C. R., Corradi, J., Bouzat, C. & Sine, S. M. Single-Channel and Structural Foundations of Neuronal α7 Acetylcholine Receptor Potentiation. *Journal of Neuroscience* **31**, 13870–13879 (2011).

61. Deba, F. *et al.* LY2087101 and dFBr share transmembrane binding sites in the (α4)3(β2)2 Nicotinic Acetylcholine Receptor. *Sci Rep* **8**, (2018).

62. Grutter, T. *et al.* Molecular tuning of fast gating in pentameric ligand-gated ion channels. *Proc Natl Acad Sci U S A* **102**, 18207–18212 (2005).

63. Aldea, M., Mulet, J., Sala, S., Sala, F. & Criado, M. Non-charged amino acids from three different domains contribute to link agonist binding to channel gating in alpha7 nicotinic acetylcholine receptors. *J Neurochem* **103**, 725–735 (2007).

64. Lee, W. Y., Free, C. R. & Sine, S. M. Binding to gating transduction in nicotinic receptors: Cys-loop energetically couples to pre-M1 and M2-M3 regions. *J Neurosci* **29**, 3189–3199 (2009).

65. Jha, A., Cadugan, D. J., Purohit, P. & Auerbach, A. Acetylcholine receptor gating at extracellular transmembrane domain interface: the cys-loop and M2-M3 linker. *J Gen Physiol* **130**, 547–558 (2007).

66. Sine, S. M., Strikwerda, J. R. & Mazzaferro, S. Structural basis for α-bungarotoxin insensitivity of neuronal nicotinic acetylcholine receptors. *Neuropharmacology* **160**, (2019).

67. Li, L. *et al.* The tethered agonist approach to mapping ion channel proteins – toward a structural model for the agonist binding site of the nicotinic acetylcholine receptor. *Chem Biol* **8**, 47–58 (2001).

68. Arias, H. R. Topology of ligand binding sites on the nicotinic acetylcholine receptor. *Brain Res Rev* **25**, 133–191 (1997).

69. Zhang, D., Gullingsrud, J. & McCammon, J. A. Potentials of mean force for acetylcholine unbinding from the Alpha7 nicotinic acetylcholine receptor ligand-binding domain. *J Am Chem Soc* **128**, 3019–3026 (2006).

70. Fucile, S. The distribution of charged amino acid residues and the Ca2+ permeability of nicotinic acetylcholine receptors: A predictive model. *Front Mol Neurosci* **10**, 155 (2017).

71. Löytynoja, A. & Milinkovitch, M. C. A hidden Markov model for progressive multiple alignment. *Bioinformatics* **19**, 1505–1513 (2003).

72. Seo, M. J., Heo, J., Kim, K., Chung, K. Y. & Yu, W. Coevolution underlies GPCR-G protein selectivity and functionality. *Scientific Reports 2021 11:1* **11**, 1–11 (2021).

73. Teşileanu, T., Colwell, L. J. & Leibler, S. Protein Sectors: Statistical Coupling Analysis versus Conservation. *PLoS Comput Biol* **11**, e1004091 (2015).

74. Kullback, S. Information theory and statistics; 1. Auflage. 399 (1968).

75. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

76. Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A* **102**, 10557–10562 (2005).

77. Löytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science (1979)* **320**, 1632–1635 (2008).

Supplementary Information

Data containing alignments used for this project available at:
https://doi.org/10.6084/m9.figshare.21529416

Data containing redesigns used for this project available at:
https://doi.org/10.6084/m9.figshare.21529416

Data containing matlab and perl scripts used for this project available at:
https://doi.org/10.6084/m9.figshare.21558642

Full Length Wild Type α7 Sequence:

```
GEFQRKLYKELVKNYNPLERPVANDSQPLTVYFSLSLL
QIMDVDEKNQVLTTNIWLQMSWTDHYLQWNVSEYPGVKTVRFPDGQIWKPDILLYNSADE
RFDATFHTNVLVNSSGHCQYLPPGIFKSSCYIDVRWFPFDVQHCKLKFGSWSYGGWSLDL
QMQEADISGYIPNGEWDLVGIPGKRSERFYECCKEPYPDVTFTVTMRRRTLYYGLNLLIP
CVLISALALLVFLLPADSGEKISLGITVLLSLTVFMLLVAEIMPATSDSVPLIAQYFAST
MIIVGLSVVVTVIVLQYHHHDPDGGKMPKWTRVILLNWCAWFLRMKRPGEDKVRPACQHK
QRRCSLASVEMSAVAPPPASNGNLLYIGFRGLDGVHCVPTPDSGVVCGRMACSPTHDEHL
LHGGQPPEGDPDLAKILEEVRYIANRFRCQDESEAVCSEWKFAACVVDRLCLMAFSVFTI
ICTIGILMSAPNFVEAVSKDFA
```

PRANK and MUSCLE alignment algorithms:

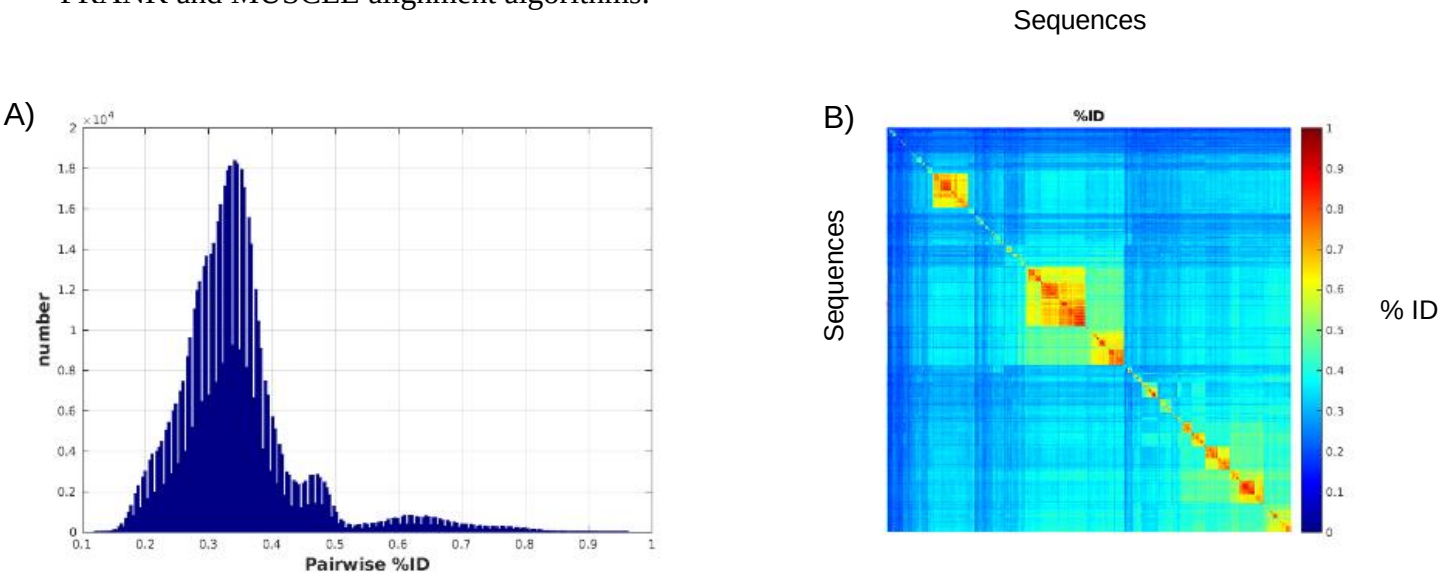

Figure S1. Pairwise Percent Identity Analysis of 7eki core Sequence Alignment using PRANK. (A) histogram of pairwise percent sequence identity (Pairwise %ID) between each sequence against the number of sequences from the alignment that fall into the bin for a given pairwise %ID. Pairwise %ID represents how similar each sequence in the alignment is to each other sequence in the alignment. Histogram is one half of the diagonal (B) matrix which runs each sequence against each other and colours the resulting pixel by pairwise percent identity (heat map on the right). Alignment was generated with PRANK[76] alignment algorithm.

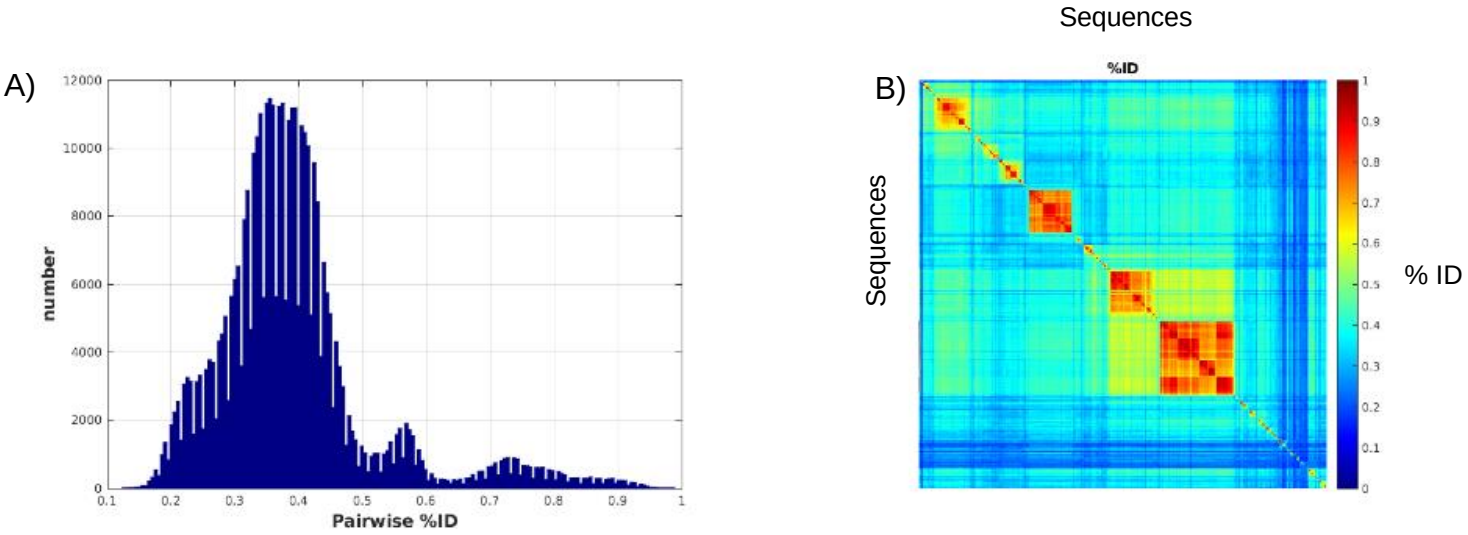

Figure S2. Pairwise Percent Identity Analysis of 7eki core Sequence Alignment using MUSCLE. (A) histogram of pairwise percent sequence identity (Pairwise %ID) between each sequence against the number of sequences from the alignment that fall into the bin for a given pairwise %ID. Pairwise %ID represents how similar each sequence in the alignment is to each other sequence in the alignment. Histogram is one half of the diagonal (B) matrix which runs each sequence against each other and colours the resulting pixel by pairwise percent identity (heat map on the right). Alignment was generated with MUSCLE[77] alignment algorithm.

The % ID matrices along with the histograms show that alignments have been pre-grouped into subfamilies (bright red boxes along the diagonal, multiple peaks) rather than having a uniformly "unalike" alignment. This adds a layer of complexity to SCA because these subfamilies would have an impact on the results of couplings during SCA calculations and for this reason these alignments were not analyzed further.

SCA calculations:

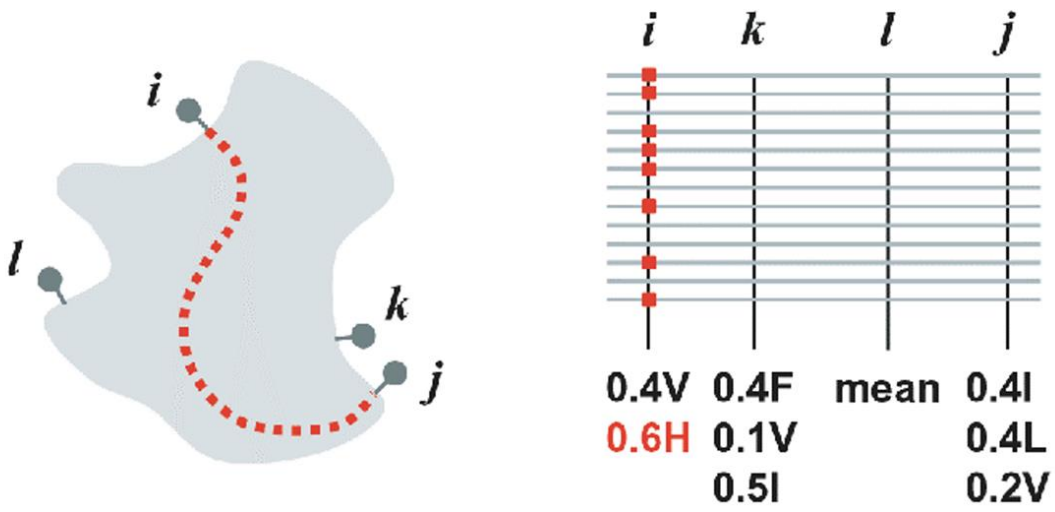

Figure S3. A statistical perturbation method for measuring interactions between residues in proteins. a, A hypothetical protein, showing four sites i, j, k and l with the following energetic properties: (i) l makes no contribution to structure or function of the protein, but i, j and k contribute in some way, and (ii) l and k act independently of one another, and i and j act cooperatively. Thus, l is energetically valueless, i and k are energetically additive, and i and j are energetically coupled. b, Schematic representation of a large and diverse MSA of the protein family, where horizontal lines represent individual protein sequences. Site l (which is unconstrained) should show the mean distribution of amino acids found randomly in all natural proteins, whereas i, j and k should show some level of conservation (deviance from the mean distribution). Figure and caption reproduced with permission from springer nature biology, license no. 5361510644431.

The first step in working out a SCA is to collect a series of sequences in the protein family and generate a multiple sequence alignment. For example, say we have a hypothetical protein with amino acid residues at sites *i*, *l*, *k* and *j*, as shown in Figure S3.A. A multiple sequence alignment for this protein is shown in Figure S3.B. If we take the frequency of each amino acid type at each position in the alignment, we get the observed frequencies at each position as shown along the bottom. Position *l* in this case is deemed to be unimportant because it represents the mean or the observed frequency of amino acids in all proteins and exhibits very little conservation. Positions *k* and *j* can be deemed to be moderately conserved, while position *i* is very conserved. In this

case, 0.6H means we see histidine at this position 60% of the time. The probability of seeing a 60% chance of a histidine at this position when the mean frequency of histidine across all proteins is 5%, is very low. The improbability of this outcome at this position is a measure of evolutionary constraint. We can represent this with the following equation:

Equation I.

$$C_i^a = D(f_i^a \,||\, q^a)$$

D = probability

f= frequency

i = position

a = amino acid

q = mean background frequency

C=conservation of a at position i

Wherein the probability *D* of seeing the frequency *f* of amino acid *a* at position *i* is given by the background of *q*. This equation can be reduced into the Kullback-Leibler entropy[74]:

Equation II.

$$C_i^a = f_i^a \ln\frac{f_i^a}{q^a} + (1-f_i^a)\ln\frac{1-f_i^a}{1-q^a}$$

D = probability

f= frequency

i = position

a = amino acid

q = mean background frequency

C=conservation of a at position i

For the Kullback-Leibler entropy, the degree to which the observed distribution deviates from the random distribution is represented by first order statistics (in other words, simple positional conservation). This is an excellent way to determine the evolutionary importance of a specific amino acid at a given position. However, referring to our earlier protein example, we do not expect positions *i* and *j* to evolve independently. The outcome at position *j* will influence position *i*. To examine this link, we introduce a second order statistical term:

Equation III.

$$C_i^{(a)} = C_i^{(a)} + \sum C_{ij}^{(ab)}$$

f= frequency

i,j = position

a,b = amino acids

C=conservation of a,b at position i,j

The addition of the pairwise term to the initial term from eq II allows us to consider coevolution. We do not need to go to a 3rd order term because we cannot currently constrain higher order statistics in the current genome database. This pairwise term allows us to generate a statistical coupling matrix, in which the degree that every position is correlated to every other position can be illustrated[37–39].

Complete sector lists:

Table S1. Complete list of all residues described in sector by structure.
All residues are in wild type α7 numbering.

structure name	Alignment		structure name	Alignment	
	core			**core**	
7kox	sector 1 (red)	47 52 96 126 229 250 274 278 279 282 285 296 297 305 321 421 439	7ekt	sector 1 (red)	39 120 125 219 250 279 282 296 305 308 323 439 450 453
	sector 2 (green)	40 42 46 69 101 118 134 154 201 207 231 275 316 319 414 417		sector 2 (green)	69 91 95 98 118 216 231 277 283 294 298 302 303 309 313 318 437 443 451 460
	sector 3 (blue)	39 72 93 147 149 153 164 188 190 191 194 197 219		sector 3 (blue)	22 42 47 48 93 96 106 131 147 149 188 190 191 193 259 289 469
	unique			**unique**	
	sector 1 (red)	226 242 258 282		sector 1 (red)	92 94 107 135 137 164 170 207 211 214 226 242 258 261 282 289
	sector 2 (green)	109 141 232 233 238 250 253 254 278 283 306 308		sector 2 (green)	89 146 221 243 245 252 270 285
	sector 3 (blue)	11 21 31 37 40 45 48 60 67 69 81 86 230		sector 3 (blue)	199 219 250 279 282 296 323 439
	total			**total**	
	sector 1 (red)	47 52 126 229 274 278 279 285 439		sector 1 (red)	69 91 95 98 118 122 212 216 231 277 283 294 298 302 303 309 313 316 318 451 460
	sector 2 (green)	40 42 69 101 118 134 154 194 201 207 231 236 258 272 275 316 319 414 417 464 469		sector 2 (green)	22 42 44 47 93 106 131 147 153 164 188 190 191 193 194 259 289 316

	sector 3 (blue)	39 72 92 93 147 149 188 190 191 194 195 197 219 242 308		sector 3 (blue)	22 42 44 47 93 106 131 147 153 164 188 190 191 193 194 259 289 316
	core			**core**	
7koo	sector 1 (red)	15 39 93 121 149 190 191 195 199 219 233 253 279 308	7eki	sector 1 (red)	39 42 47 93 106 131 147 149 153 188 190 191 193 194 197 259 289
	sector 2 (green)	40 91 95 118 151 201 207 216 231 245 277 294 298 302 313 451		sector 2 (green)	16 24 43 52 53 56 120 130 145 158 205 209 264 282 290 305 321 323 325 326 450 461 469
	sector 3 (blue)	42 69 134 153 184 188 190 191 193 194 236 258 316 468		sector 3 (blue)	40 69 91 95 118 122 194 207 216 231 236 277 294 298 302 303 313 316 451
	unique			**unique**	
	sector 1 (red)	73 89 135 148 211 226 243 269 271 285 288 331		sector 1 (red)	15 73 92 137 173 226 242 261 282
	sector 2 (green)	12 232 233 239 250 253 254 257 306 308 314 319		sector 2 (green)	21 40 45 60 67 68 69 81 86 88 174
	sector 3 (blue)	35 40 67 92 137 174 214 242 245 252 261 290		sector 3 (blue)	21 230 233 238 254 265 308 310
	total			**total**	
	sector 1 (red)	15 39 93 121 149 190 191 195 219 279 308		sector 1 (red)	106 125 199 219 250 251 279 296 308 453
	sector 2 (green)	40 91 95 118 122 151 167 207 231 245 277 294 303 313 451		sector 2 (green)	16 24 43 52 53 56 120 130 145 209 264 282 290 303 305 321 323 325 326 443 450 461

	sector 3 (blue)	22 42 153 184 188 190 191 194 214 236 258 289 316 449 469		sector 3 (blue)	69 91 92 95 118 122 207 225 231 255 277 283 294 302 316 451 469
	core			**core**	
7koq	sector 1 (red)	16 24 53 120 298 303 305 313 318 323 325 326 443 461	7ekp	sector 1 (red)	120 279 296 305 323
	sector 2 (green)	39 157 195 199 219 279 296		sector 2 (green)	69 95 118 216 231 277 294 298 302 303 313 318 443 451 460
	sector 3 (blue)	69 90 91 92 95 118 122 210 224 225 229 231 255 316 441 449 451 460		sector 3 (blue)	22 48 93 106 147 149 188 190 191 194 259 289
	unique			**unique**	
	sector 1 (red)	72 94 118 131 135 148 149 171 173 211 269		sector 1 (red)	89 96 103 107 121 135 148 211 237 269 271 278 282 308
	sector 2 (green)	20 45 60 62 65 67 68 69 81 86 88 149		sector 2 (green)	7 20 21 31 35 45 55 60 67 68 69 78 81 86 88 174
	sector 3 (blue)	137 140 172 174 212 214 230 242 245 248 252 259 261 262		sector 3 (blue)	232 233 252 253 254 289 306 316
	total			**total**	
	sector 1 (red)	16 24 53 120 298 303 305 313 318 323 325 326 443 461		sector 1 (red)	128 212 213 217 221 241 273 281 296 301 318
	sector 2 (green)	93 94 123 188 190 191 193 194 197 214 222 248 264 293		sector 2 (green)	91 113 117 140 144 229 238 247 253 277 299 316 320 324 325 335 338 473
	sector 3 (blue)	69 90 91 92 118 122 212 224 225 229 231 255 283 316 441 449 451		sector 3 (blue)	38 75 142 320 325 327 335 340 345 347 348 465 472

Heatmap Data:

Table S2. Complete data used for chapter 3 heatmaps. RMSDs were calculated in VMD while % identity was calculated in Jalview[75]. Redesigns refers to the Rosetta redesigns while HMM Consensus sequences are the consensus sequences generated from each structural HMM.

RMSD							
7ekt	0						
7ekp	1.24	0					
7eki	1.8	1.62	0				
7kox	5.01	4.86	5.1	0			
7koq	2.3	2.23	2.76	4.2	0		
7koo	2.78	2.67	2.31	4.52	1.85	0	
	7ekt	7ekp	7eki	7kox	7koq	7koo	
Redesigns % Identity							
Wild type	0						
7kox A	24.17	0					
7kox A2	23.98	65.54	0				
7kox A3	24.43	65.28	61.66	0			
7koq A	19.85	25.83	26.6	26.02	0		
7koo A	19.8	33.08	32.48	31.12	26.46	0	
	Wild type	7kox A	7kox A2	7kox A3	7koq A	7koo A	
HMM Consensus % Identity							
7ekp	0						
7ekt	47.98	0					
7eki	43.18	40.4	0				
7koo	31.22	31.47	40.46	0			
7koq	33.76	29.37	31.55	36.48	0		
7kox	36.87	33.5	33.76	35.81	35.29	0	
Wild-Type	28.03	26.77	26.33	24.23	24.43	26.94	0
	7ekp	7ekt	7eki	7koo	7koq	7kox	Wild-Type

Individual Structural SCA Results in alignment and mapping form

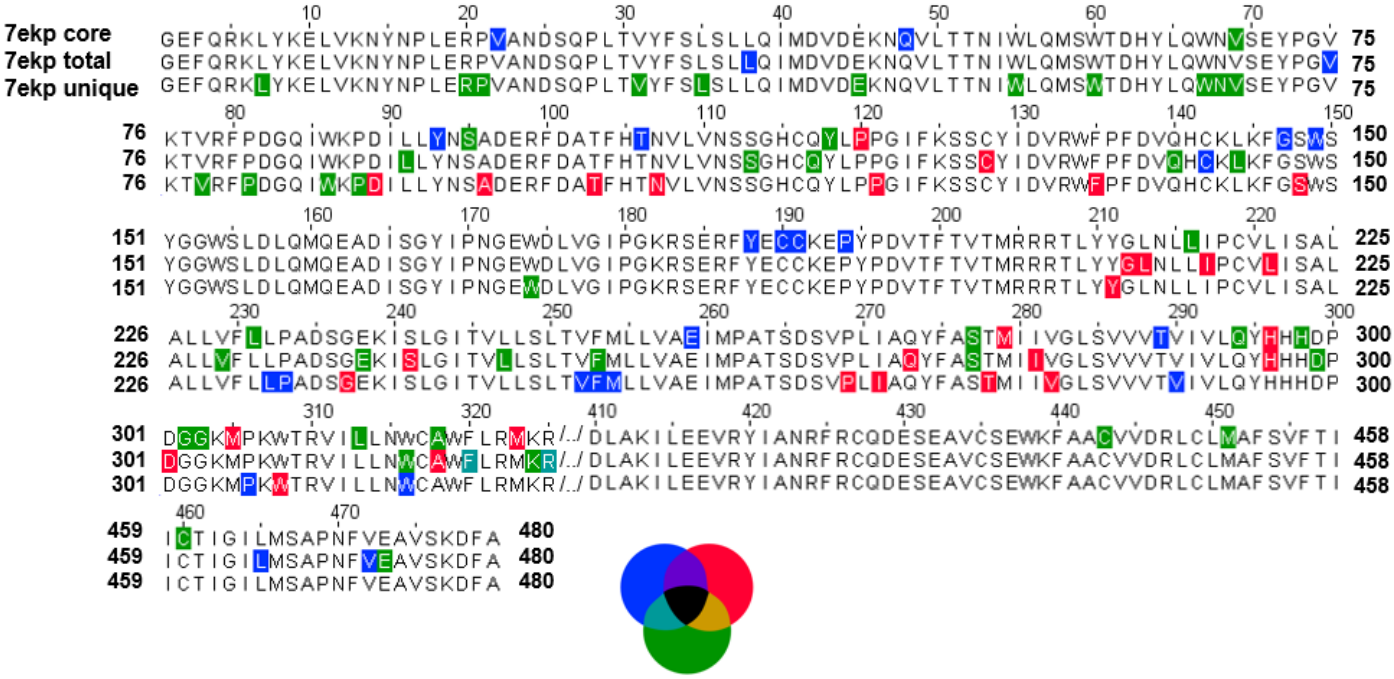

Figure S4. Sector Results Mapped onto Sequence of 7ekp. 7ekp represents a putative desensitized conformation. Sectors were mapped in wild type mature α7 numbering. Alignments were generated in Jalview[75]. Unique represents the set of sequences unique to 7ekp, core represents the set of sequences common among all conformations and total is the sum of the first two sequence sets. Complete numbers can be found in Table S1. Colour legend is as follows: red, blue and green represent individual sectors, purple is the combination of red and blue, turquoise is the combination of blue and green and yellow is the combination of red and green. Black represents residues in all sectors.

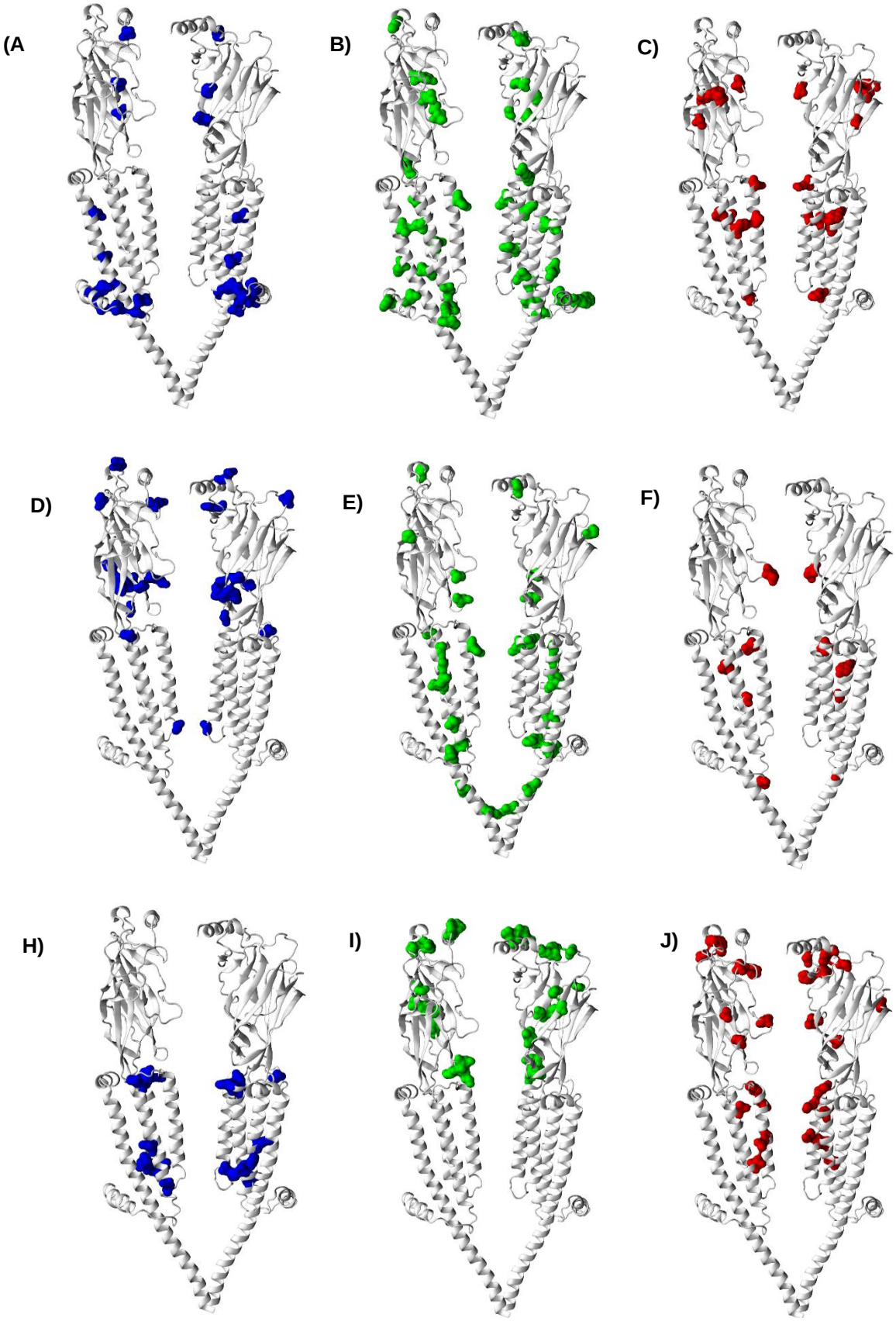

Figure S5. Sector Results Mapped as Surface Representation for PDB 7ekp from all datasets. Sectors were mapped onto PDB ID 7ekp (putative desensitized conformation) in wild type mature α7 numbering. Figures were generated using VMD. **A, B, C)** Blue, Green, and Red sectors for total data set. **D, E, F)** Blue, Green and Red sectors for core data set. **H, I, J)** Blue, Green and Red Sector for unique data set. Complete numbers can be found in Table S1.

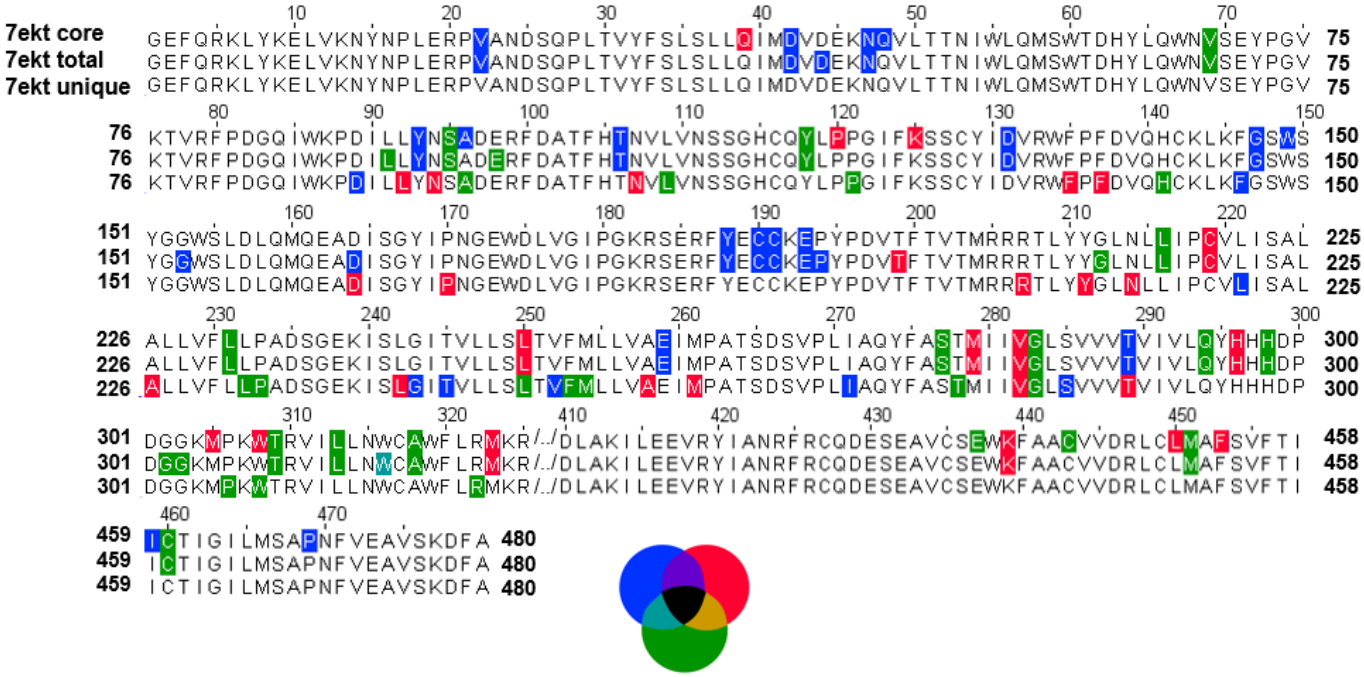

Figure S6. Sector Results Mapped onto Sequence of 7ekt. 7ekt represents a putative active conformation. Sectors were mapped in wild type mature α7 numbering. Alignments were generated in Jalview[75] (reference). Unique represents the set of sequences unique to 7ekt, core represents the set of sequences common among all conformations and total is the sum of the first two sequence sets. Complete numbers can be found in Table S1. Colour legend is as follows: red, blue and green represent individual sectors, purple is the combination of red and blue, turquoise is the combination of blue and green and yellow is the combination of red and green. Black represents residues in all sectors.

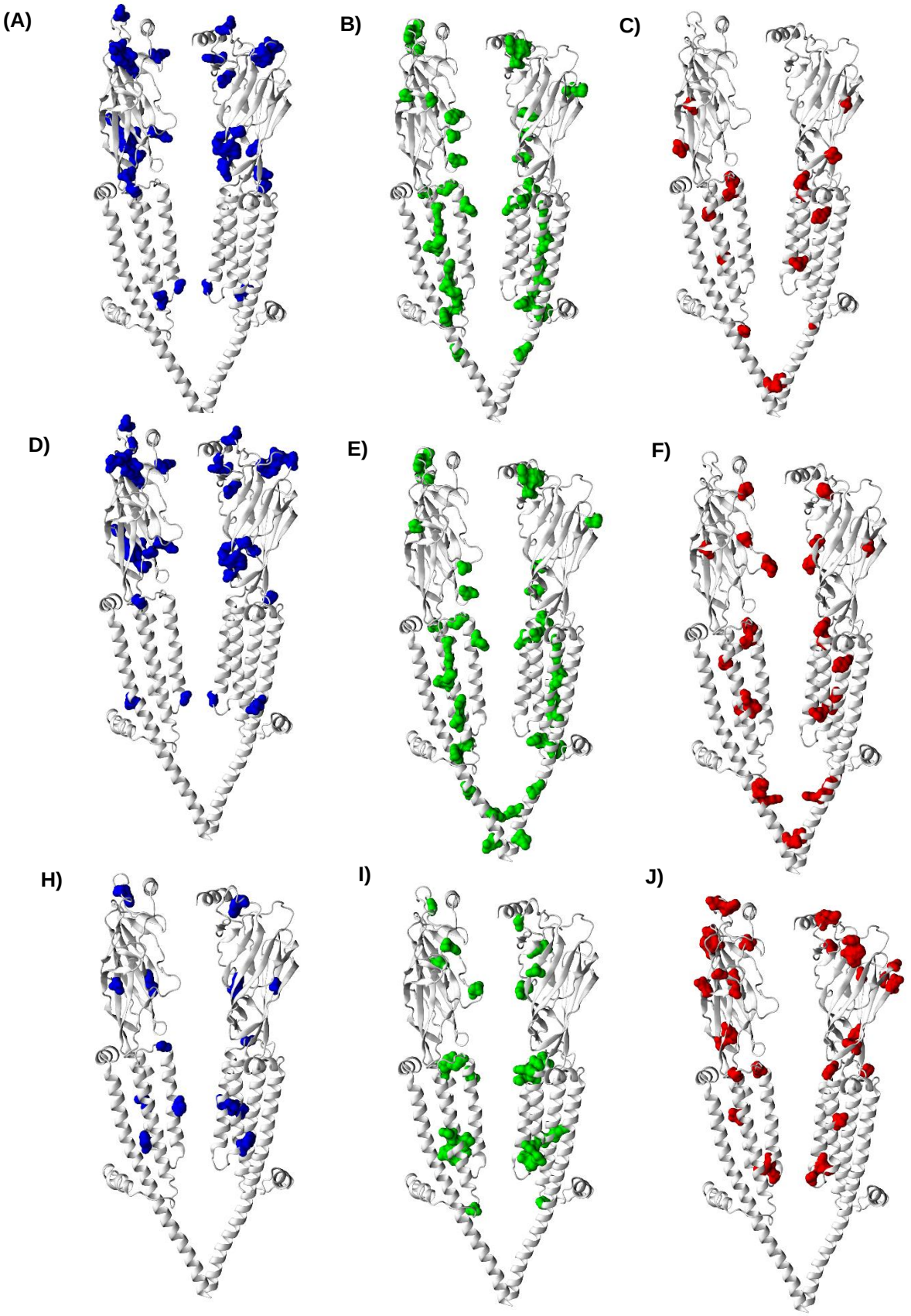

Figure S7. Sector Results Mapped as Surface Representation for PDB 7ekt from all datasets. Sectors were mapped onto PDB ID 7ekt (putative active conformation) in wild type mature α7 numbering. Figures were generated using VMD. **A, B, C)** Blue, Green, and Red sectors for total data set. **D, E, F)** Blue, Green and Red sectors for core data set. **H, I, J)** Blue, Green and Red Sector for unique data set. Complete numbers can be found in Table S1.

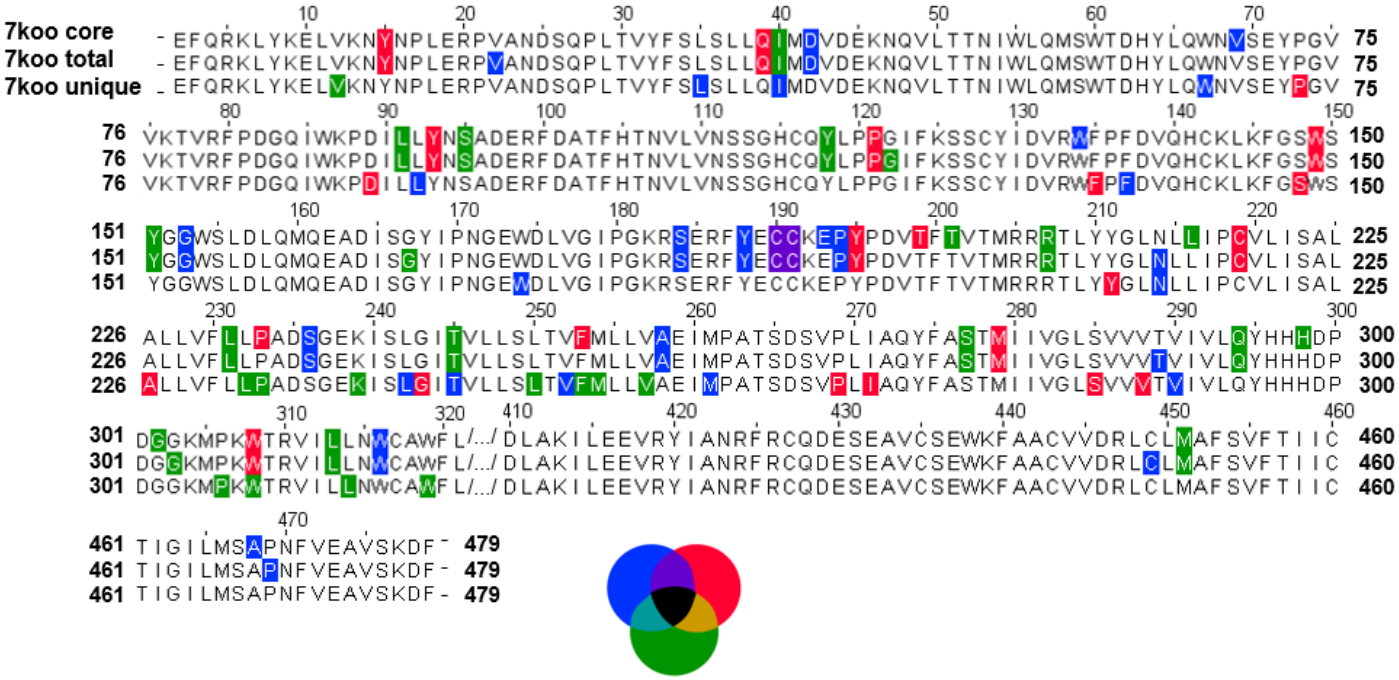

Figure S8. Sector Results Mapped onto Sequence of 7koo. 7koo represents a putative resting conformation. Sectors were mapped in wild type mature α7 numbering. Alignments were generated in Jalview[75]. Unique represents the set of sequences unique to 7koo, core represents the set of sequences common among all conformations and total is the sum of the first two sequence sets. Complete numbers can be found in Table S1. Colour legend is as follows: red, blue and green represent individual sectors, purple is the combination of red and blue, turquoise is the combination of blue and green and yellow is the combination of red and green. Black represents residues in all sectors.

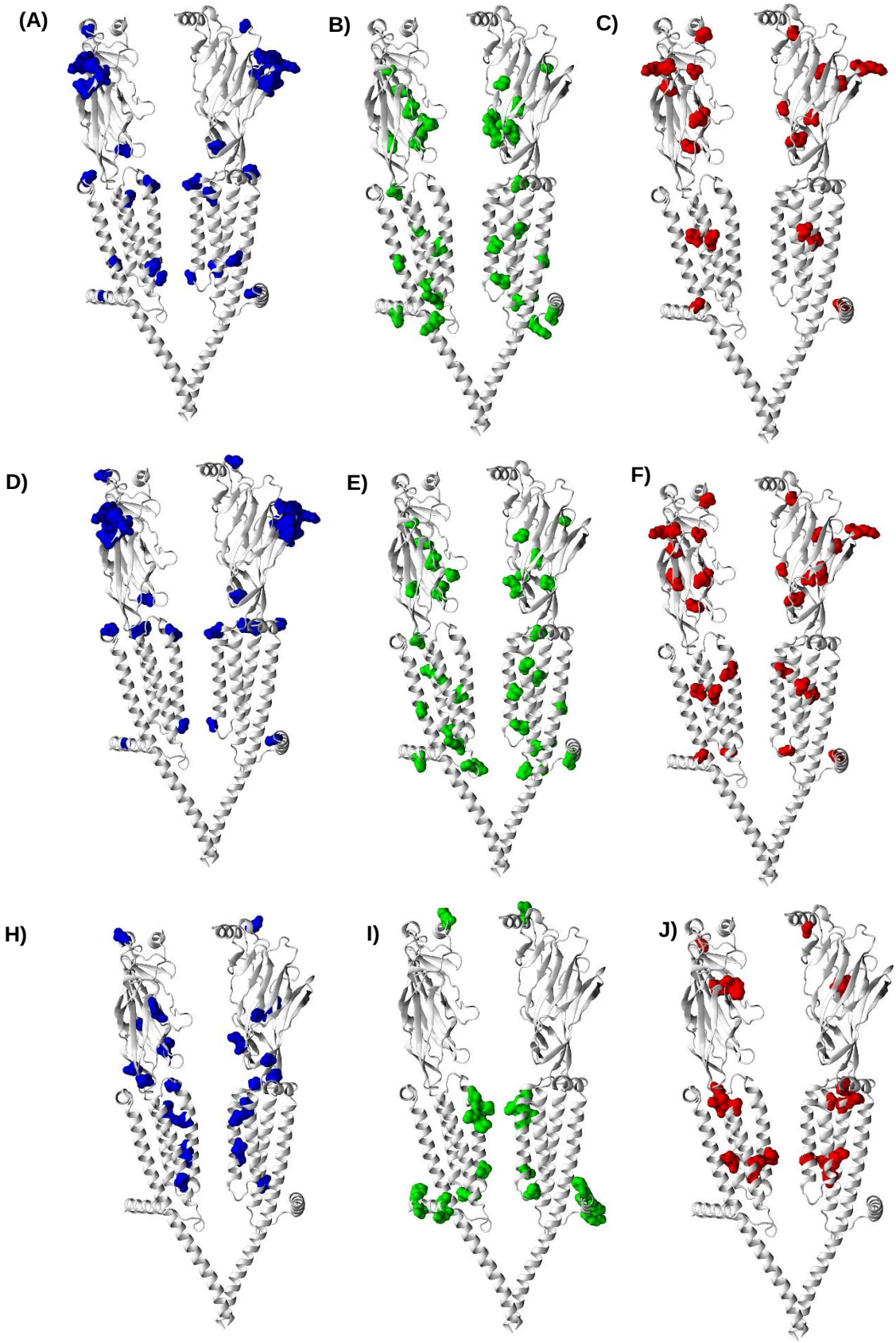

Figure S9. Sector Results Mapped as Surface Representation for PDB 7koo from all datasets. Sectors were mapped onto PDB ID 7koo (putative resting conformation) in wild type mature α7 numbering. Figures were generated using VMD. **A, B, C)** Blue, Green, and Red sectors for total data set. **D, E, F)** Blue, Green and Red sectors for core data set. **H, I, J)** Blue, Green and Red Sector for unique data set. Complete numbers can be found in Table S1.

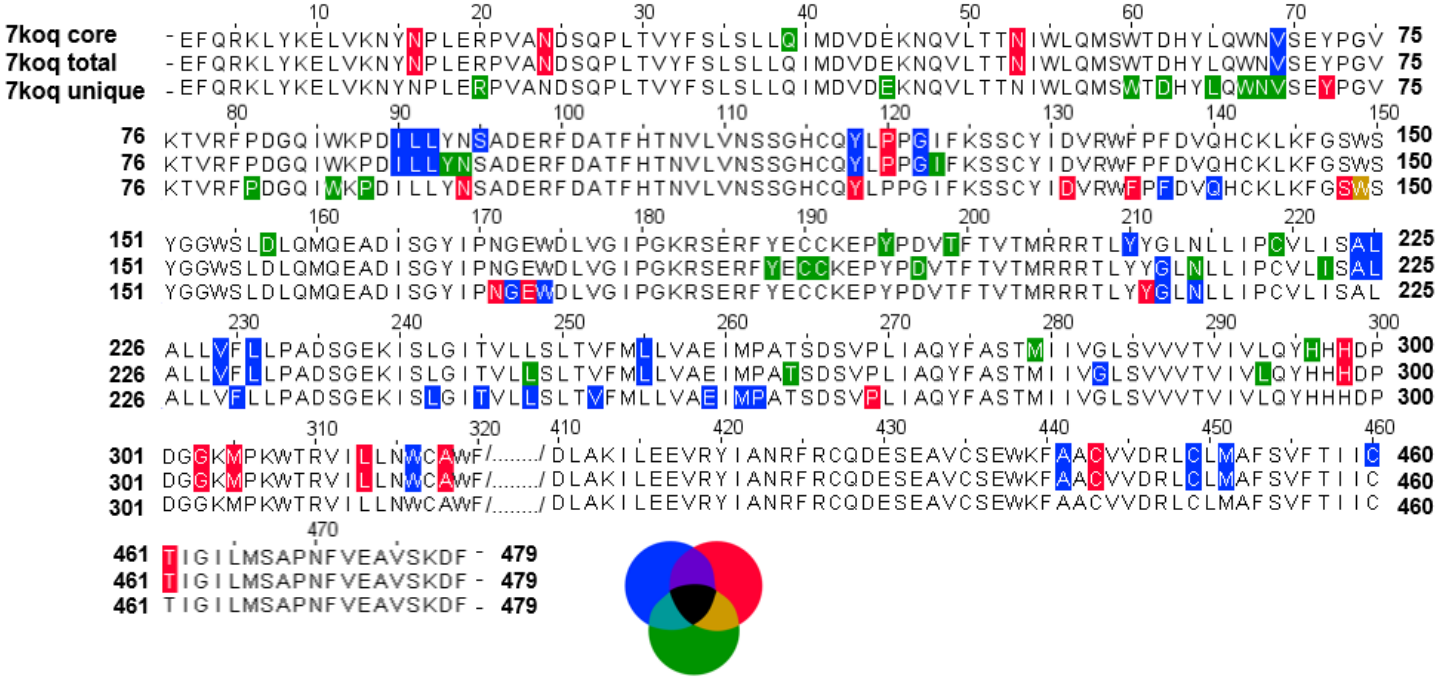

Figure S10. Sector Results Mapped onto Sequence of 7koq. 7koq represents a putative desensitized conformation. Sectors were mapped in wild type mature α7 numbering. Alignments were generated in Jalview[75]. Unique represents the set of sequences unique to 7koq, core represents the set of sequences common among all conformations and total is the sum of the first two sequence sets. Complete numbers can be found in Table S1. Colour legend is as follows: red, blue and green represent individual sectors, purple is the combination of red and blue, turquoise is the combination of blue and green and yellow is the combination of red and green. Black represents residues in all sectors.

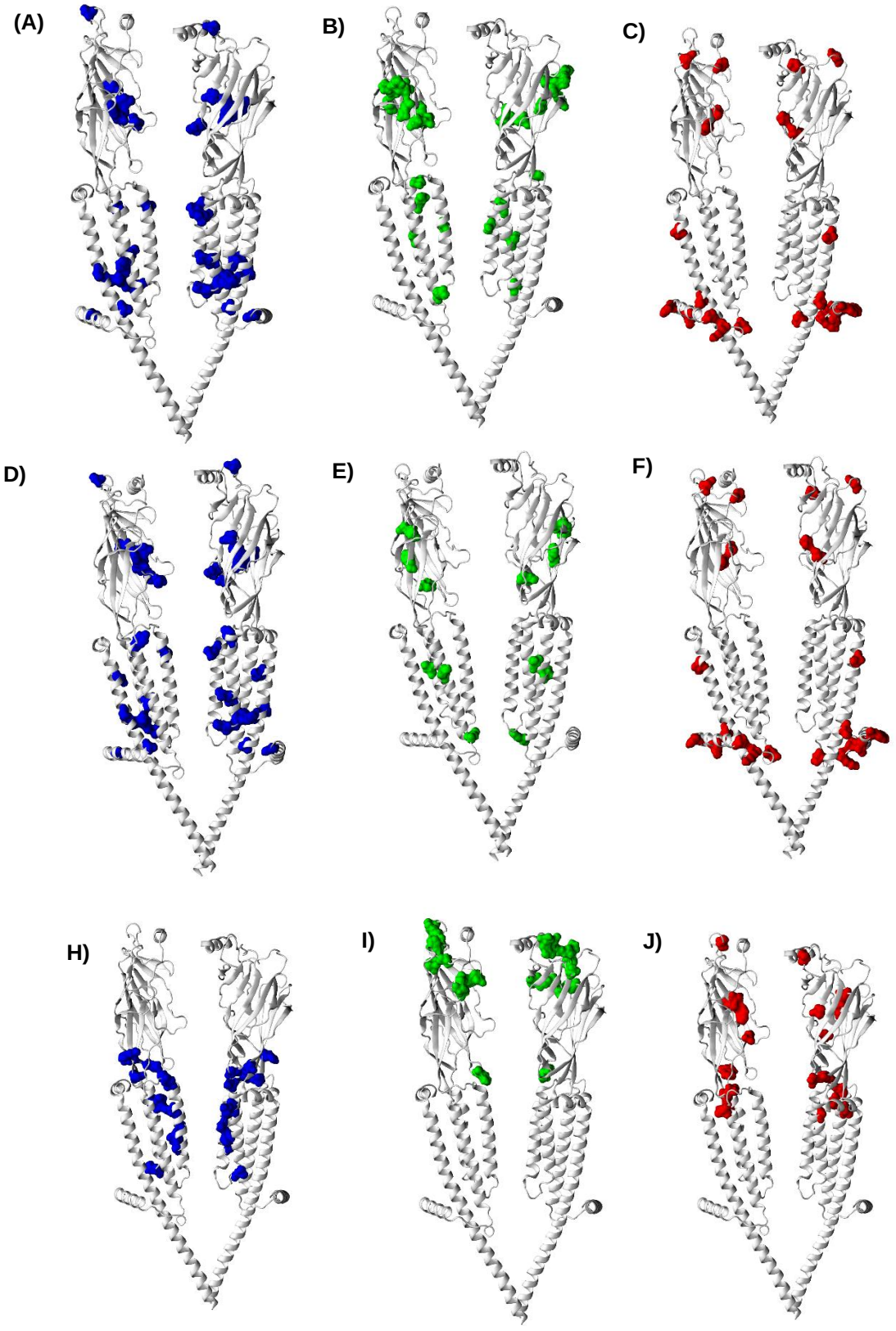

Figure S11. Sector Results Mapped as Surface Representation for PDB 7koq from all datasets. Sectors were mapped onto PDB ID 7koq (putative desensitized conformation) in wild type mature α7 numbering. Figures were generated using VMD. **A, B , C)** Blue, Green, and Red sectors for total data set. **D, E, F)** Blue, Green and Red sectors for core data set. **H, I, J)** Blue, Green and Red Sector for unique data set. Complete numbers can be found in Table S1.

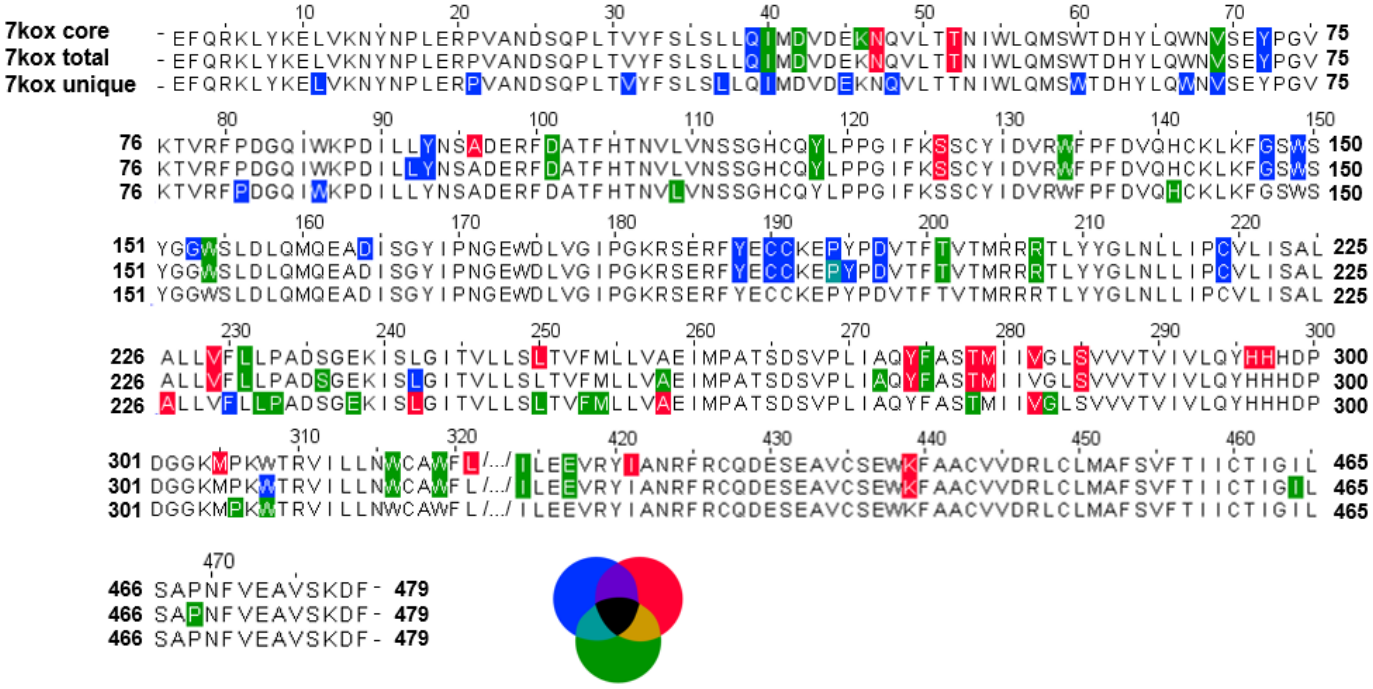

Figure S12. Sector Results Mapped onto Sequence of 7kox. 7kox represents a putative active conformation. Sectors were mapped in wild type mature α7 numbering. Alignments were generated in Jalview[75]. Unique represents the set of sequences unique to 7kox, core represents the set of sequences common among all conformations and total is the sum of the first two sequence sets. Complete numbers can be found in Table S1. Colour legend is as follows: red, blue and green represent individual sectors, purple is the combination of red and blue, turquoise is the combination of blue and green and yellow is the combination of red and green. Black represents residues in all sectors.

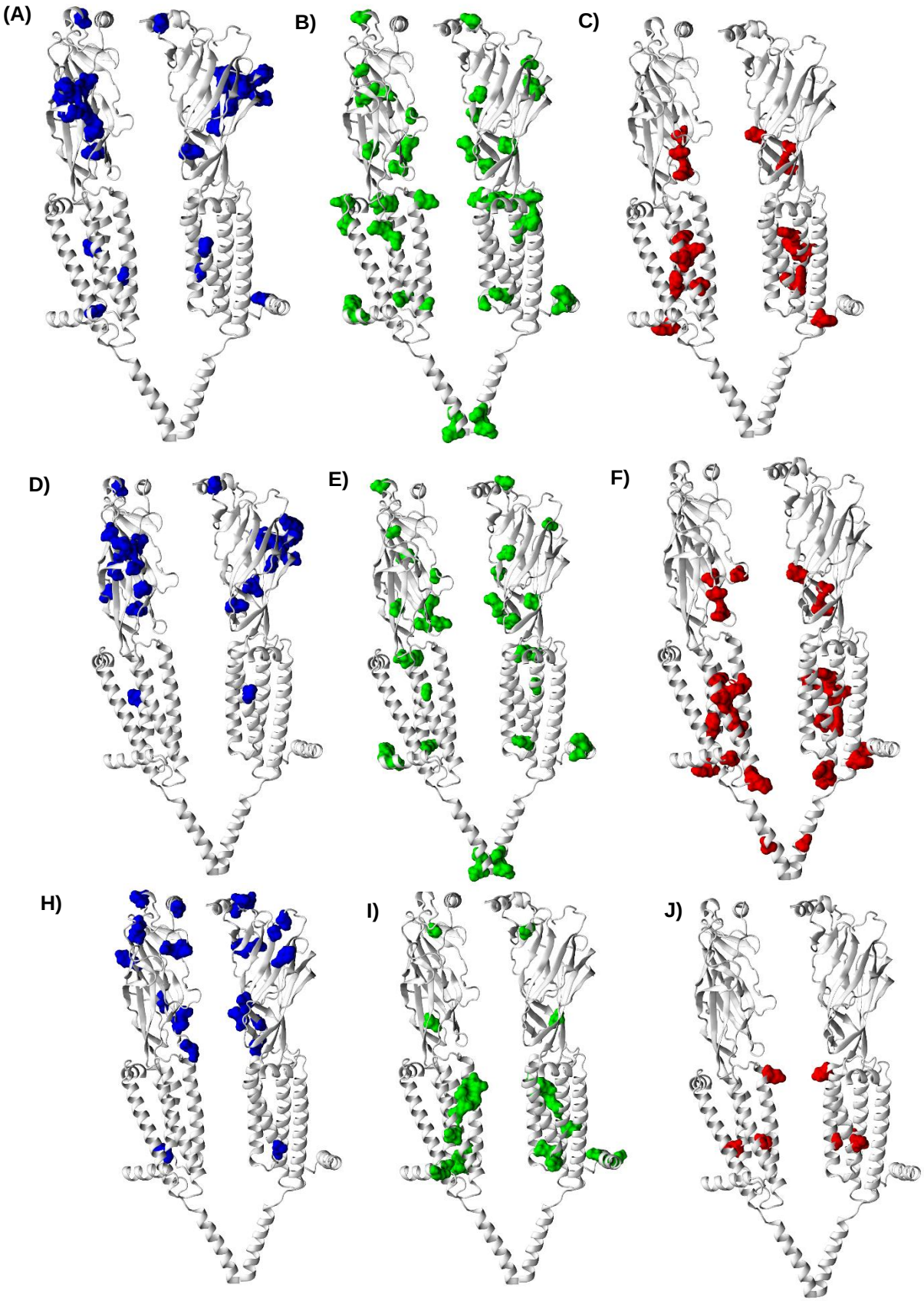

Figure S13. Sector Results Mapped as Surface Representation for PDB 7kox from all datasets. Sectors were mapped onto PDB ID 7kox (putative active conformation) in wild type mature α7 numbering. Figures were generated using VMD. **A, B , C)** Blue, Green, and Red sectors for total data set. **D, E, F)** Blue, Green and Red sectors for core data set. **H, I, J)** Blue, Green and Red Sector for unique data set. Complete numbers can be found in Table S1.