Protein wild-type and mutant ensemble database

by

Ce Zhang

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Computer Science

Program of Study Committee:

Guang Song, Major Professor

Robert L. Jernigan

Xiaoqiu Huang

Iowa State University

Ames, Iowa

2016

DEDICATION

I would like to dedicate this thesis to my wife Yuanyuan Zhou and to my parents without whose support I would not have been able to complete this work. I would also like to thank my friends and family for their loving guidance and financial assistance during the writing of this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Prof. Guang Song for his guidance, patience and support throughout this research and the writing of this thesis. His insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would also like to thank my committee members for their efforts and contributions to this work: Prof. Robert L. Jernigan and Prof. Xiaoqiu Huang. I would additionally like to thank Dr. Hyuntae Na for his guidance throughout the initial stages of my graduate career and knowledge of structural biology.

ABSTRACT

Protein structures have been determined and deposited into Protein Data Bank at an increasing rate. In this work, we organize all the protein structures in the PDB and form a wild type and mutant structure database. The database groups the wild type and mutant structures of the same protein together. One direct benefit of the database is thus the easy accessibility of the structure ensembles of all the proteins. Such ensembles are known to be highly useful for representing the native states of proteins and for understanding their functions. For each protein, mutants are sorted by the number of mutations and the location(s) of the mutations. What distinguishes our work from other mutation databases is that it is structure-based and includes all the existing structures of the PDB. Synchronization with the PDB database will be maintained. As an application, we carry out an experimental structure-based statistical analysis of the effects of mutations, on both protein structure and protein dynamics. A key question we address in this work is: is it valid to use mutant structures (or variants from different species) to represent a native state sample of a given protein? Our results indicate that mutations can cause significant structure changes and dynamics changes, more than commonly expected. This implies that cautions must be taken when mutation structures are considered to be included as representative samples of the conformation space of a given protein.

CHAPTER 1. INTRODUCTION

PDB [1] has over 120,000 structures. Among these the vast majority are proteins or protein complexes. Only about 2.5% are DNA/RNA. About 90% of these structures are determined by X-ray, 10% by NMR, and 1% by cryo-EM or other means. And more structures are being deposited at an ever increasing rate. Many of these structures are of the same protein that has already one or more structures deposited in PDB.

There exist many structures of the same protein and these structures form an ensemble of the protein which can be used study dynamics that exists in the ensemble. Best et al. [2] shows that some of these ensembles are able to reproduce different NMR measurements and may represent the true native-state ensembles. Others [3, 4, 5, 6] showed that the dynamics within the structure ensembles obtained by Principal Component Analysis(PCA) matches well with the dynamics obtained by normal mode analysis [7, 8, 9].

Most of the existing structure determination methods solve for a single average structure. However, during the last decade or so, there has been a lot of effort in determining protein structure ensembles directly from experimental data. It was realized that a single structure was not sufficient to satisfy all the experimental constraints observed. Attempts have been made to determine an ensemble of two conformations or more [10, 11, 12]. The main challenge for ensemble determination is overfitting. And there is no guarantee that the structures solved represent the true native state ensemble even if they reproduce the observed dynamics well and there is a lack of confidence in the quality of individual structures. Ways to reduce overfitting were proposed [13, 14]. The abundance of structures in PDB provides an excellent alternative for constructing structure ensembles, especially if the ensemble is able to reproduce well NMR measurements.

In this work, we organize all the protein structures in the PDB and form a wild type and mutant structure database. The database groups the wild type and mutant structures of the same protein together. A direct benefit of the database is the easy accessibility of the structure ensembles of these proteins. Such ensembles are known to be highly useful for representing the native states of proteins and for understanding their functions. For each protein, mutants are sorted by the number of mutations and the location(s) of the mutations.

There are a number of databases available, as sequence and mutation data are useful in analyzing evolutionary relationship between proteins. The protein mutant database (PMD) [15] includes natural and artificial mutant proteins, which are taken from publications. M.Michael et al. constructed proteins and mutants database for thermodynamic data (ProTherm) [16]. There are also several mutants database focused on structures and/or sequences of specific proteins such as lipase [17], peptaibols [18] and GALT proteins [19]. However, no mutant database is based on sequence and provides coverage of most of the proteins in protein data bank. In this work, We collect all protein sequence similarity information from the protein data bank and develop a wild type and mutant database. Its special features include: (i) most proteins in the protein data bank are included; (ii) it clearly displays sequence difference between the wild type and mutants of each protein; (iii) it provides a convenient accessing point to all the existing structures for any given protein. We build a web page that shows the mutation information on every protein entry. Also an investigation on how mutations affect protein structure and dynamics is carried out. What further distinguishes our work from other mutation databases is that it is structure-based and includes all the existing protein structure of the PDB. Synchronization with the PDB database will be maintained.

CHAPTER 2. MUTANTS DATABASE

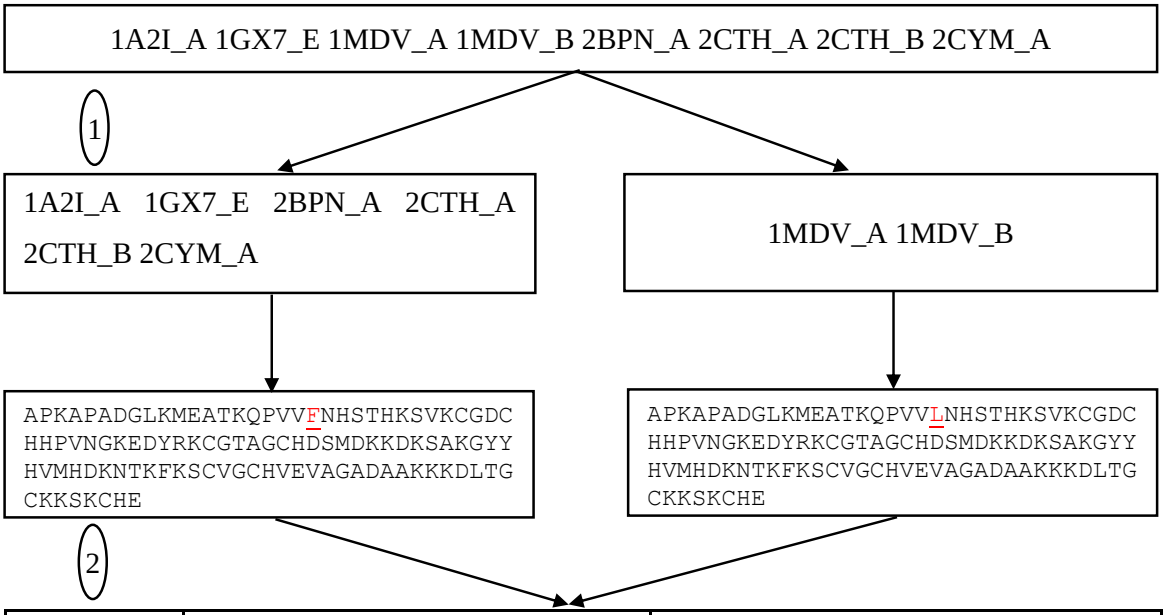
2.1 Construction

There are currently over 120,000 entries in the Protein Data Bank [1] and about 97% are protein structures. Many of these entries are structures of the same protein and thus are highly similar or even identical in sequence. They represent structure mutants, variants from different species, structures in complex with different ligands, or structures determined under different experimental conditions. To group the structures by proteins, we use the clusters produced by blastclust [20]. Blastclust can be run at different levels of sequence similarity. We use two blastclust results: one at 100% sequence similarity and one at 95%. Both results are available online at the PDB website. Each line of bc-100.out.txt lists PDB entries of the same sequence, and each line in bc-95.out.txt lists sequences that are 95% similar. We use these two cluster files to find wild types and mutants for each entry. The detailed procedures are given below.

1. We divide each entry of 95% cluster into sub clusters based on the 100% similarity clusters(see Figure 2.1). Step one shows an entry of results at 95% sequence similarity corresponds to two entries of results at 100% sequence similarity. Step two shows we compare the sequences of two groups and form the mutation information..

2. We make the assumption that the sub cluster with the largest number of proteins and the least number of mutation tags in each protein file is the wild type. Then we get sequences for each sub cluster and compare them with the wild type to get the types and locations of the point mutations. See Figure 2.1.

3. We retrieve protein names from the compound section (molecule name) inside each pdb file. Protein size is the average number of residues within each sub cluster.



Protein Type	ChainNum#ProteinNum#MuTagNum	PdbID_ChainId_MuTag
WT	6#5#0	1A2I_A_0 1GX7_E_0 2BPN_A_0 2CTH_A_0 2CTH_B_0 2CYM_A_0
F20L	2#1#1	1MDV_A_1 1MDV_B_1

Figure 2.1: An illustration of how results from blustclust are used to identify and group protein wild types and mutants.

4. We display the protein information in table format through the website using html and sort them in descending order by the cluster size (i.e., the total number of wild type and mutant structures). In the table we can see protein name, protein size, the number of wild types and a list of their PDB-ids. mutants are sorted by single mutants, double mutants, or multiple-point mutants. Mutation details for each mutant type are listed and all the PDB ids are given and are hyper-linked to entries in Protein Data Bank.

Fig. 2.2 shows two example pages of the mutant structure database website. The index page is shown in Fig. 2.2(a). There are 45 index pages, each of which has a table that contains 1,000 rows and 9 columns. Each row contains information of a specific protein, including protein name, protein size, the number of wild types, the number of mutants, single mutant number, double mutant number, etc. The last column, *details*, once clicked, leads to a page

about that protein with more details, an example of which is shown in Fig. 2.2(b). It displays a table that contains protein type, chain number, protein number, number of mutation tags and pdbID_chainID_mutation tags. Click any of the pdbID_chainID tags will lead to the corresponding protein entry in the PDB website.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45

ID	Protein Name	# Protein Size	# Wildtypes	# Mutants	# Single point mutants	# Double point mutants	# Other mutants	Files
1	LYSOZYME C	128	543	53	0	34	19	details
2	CARBONIC ANHYDRASE 2	257	358	187	37	89	61	details
3	HIV-1 PROTEASE	98	95	449	40	32	377	details
4	T4 LYSOZYME	162	54	482	20	127	335	details
5	PROTEIN (HUMAN BETA-2 MICROGLOBULIN)	99	483	29	14	5	10	details
6	CATIONIC TRYPSIN	222	364	59	0	1	58	details
7	CELL DIVISION PROTEIN KINASE 2	289	340	15	0	0	15	details
8	THROMBIN HEAVY CHAIN	251	268	83	14	4	65	details
9	BETA-SECRETASE 1	380	240	78	0	0	78	details
10	THROMBIN LIGHT CHAIN	29	312	1	1	0	0	details
11	INSULIN A CHAIN	20	214	40	0	14	26	details
12	INSULIN B CHAIN	28	192	59	0	0	59	details
13	UBIQUITIN	73	202	47	22	4	21	details
14	MYOGLOBIN	152	93	154	24	92	38	details

(A)

Protein Type	ChainNum#ProteinNum#MuNum	PdbID_ChainId_MuTag
WT	94#93#0	1UFP_A_0 2EB8_A_0 2EB9_A_0 2JHO_A_0 2W6W_A_0 3U3E_A_0 4NXA_A_0 4NXC_A_0 4PNJ_A_0 104M_A_0 105M_A_0 1AJG_A_0 1AJH_A_0 1BVC_A_0 1BVD_A_0 1BZ6_A_0 1BZP_A_0 1BZR_A_0 1CQ2_A_0 1DUK_A_0 1EBC_A_0 1F6H_A_0 1HJT_A_0 1IOP_A_0 1JP6_A_0 1JP8_A_0 1JP9_A_0 1JPB_A_0 1L2K_A_0 1MBC_A_0 1MBD_A_0 1MBI_A_0 1MBN_A_0 1MBO_A_0 1MYF_A_0 1SPE_A_0 1SWM_A_0 1U7R_A_0 1U7S_A_0 1VXA_A_0 1VXB_A_0 1VXC_A_0 1VXD_A_0 1VXE_A_0 1VXF_A_0 1VXG_A_0 1VXH_A_0 1WVP_A_0 1YOG_A_0 1YOH_A_0 1YOI_A_0 2CMM_A_0 2D6C_A_0 2D6C_B_0 2EKT_A_0 2EKU_A_0 2MB5_A_0 2MYA_A_0 2MYB_A_0 2MYC_A_0 2MYD_A_0 2MYE_A_0 2Z6S_A_0 2Z6T_A_0 2ZSN_A_0 2ZSO_A_0 2ZSP_A_0 2ZSQ_A_0 2ZSR_A_0 2ZSS_A_0 2ZST_A_0 2ZSX_A_0 2ZSY_A_0 2ZSZ_A_0 2ZT0_A_0 2ZT1_A_0 2ZT2_A_0 2ZT3_A_0 2ZT4_A_0 3E4N_A_0 3E55_A_0 3E5I_A_0 3E5O_A_0 3ECL_A_0 3ECX_A_0 3ECZ_A_0 3ED9_A_0 3EDA_A_0 3EDB_A_0 4MBN_A_0 5MBN_A_0 1A6K_A_0 1A6M_A_0 1A6N_A_0
D122N	14#14#8	109M_A_1 110M_A_1 111M_A_1 112M_A_1 1ABS_A_1 1J52_A_1 1JW8_A_0 1TES_A_1 2MBW_A_1 2MGK_A_0 2MGL_A_0 2MGM_A_0 3ASE_A_0 1A6G_A_0
G65T	3#3#3	4H07_A_1 4H0B_A_1 3O89_A_1
K102C	1#1#1	3A2G_A_1
L29H	1#1#1	4IT8_A_1
K42Y	1#1#1	4OOD_A_1
L29E	1#1#1	4PQ6_A_1
F43H	1#1#1	4PQC_A_1
F43Y	1#1#1	4QAU_A_1
K42N	1#1#1	4OF9_A_1
L29F D122N	14#14#10	1JDO_A_1 1MOA_A_0 2G0R_A_1 2G0S_A_1 2G0V_A_1 2G0X_A_1 2G0Z_A_1 2G10_A_1 2G11_A_1 2G12_A_1 2G14_A_1 2SPL_A_0 2SPM_A_0 2SPN_A_0
L29W D122N	10#10#10	1DO1_A_1 1DO3_A_1 1DO4_A_1 1DO7_A_1 1LTW_A_1 2BLH_A_1 2BLI_A_1 2BLJ_M_1 2BW9_M_1 2BWH_A_1
V68F D122N	6#6#3	106M_A_1 107M_A_1 108M_A_1 1MLJ_A_0 1MLK_A_0 1MLL_A_0
H93G DeG153	5#5#5	1IRC_A_1 1DTM_A_1 1DUO_A_1 2EVK_A_1 2EVP_A_1
F46V D122N	4#4#3	101M_A_1 1MTJ_A_1 1MTK_A_1 1MYM_A_0

(B)

Figure 2.2: Our wild type and mutant structure database. (A) the index page and (B) the page of a given protein.

2.2 Statistics

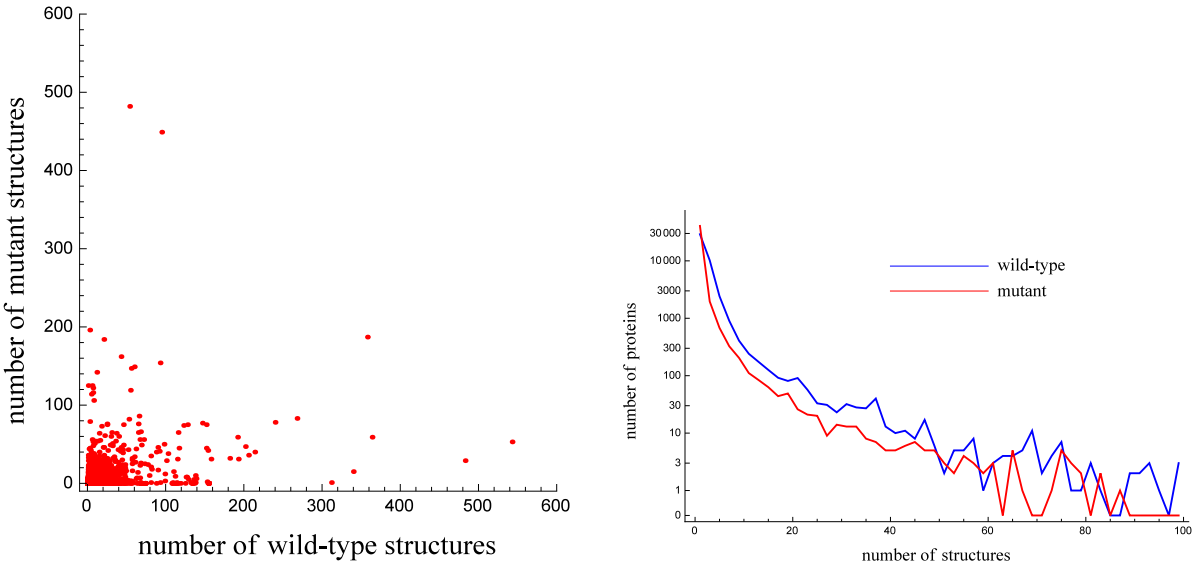
Our mutant database contains 44,035 entries and 139,344 proteins (or structures). Table 2.1 shows the statistics on the number of wild type or mutant structures. The table shows that

among all the proteins, over 1,000 have more than 10 wild type structures, over 500 have more than 20 wild type structures, and so on.

Table 2.1: Some statistics of our mutant database.

Description	statistics
number of proteins	44,305
number of proteins having > 100 WT	68 / 44,305
number of proteins having > 50 WT	146 / 44,305
number of proteins having > 20 WT	573 / 44,305
number of proteins having > 10 WT	1287 / 44,305
number of proteins having > 100 muts	17 / 44,305
number of proteins having > 50 muts	54 / 44,305
number of proteins having > 20 muts	218 / 44,305
number of proteins having > 10 muts	569 / 44,305
number of proteins having > 100 single point mutants	1 / 44,305
number of proteins having > 50 single point mutants	4 / 44,305
number of proteins having > 20 single point mutants	41 / 44,305
number of proteins having > 10 single point mutants	120 / 44,305
number of proteins having > 100 double points mutants	1 / 44,305
number of proteins having > 50 double points mutants	5 / 44,305
number of proteins having > 20 double points mutants	20 / 44,305
number of proteins having > 10 double points mutants	45 / 44,305
number of total wild type structures	109,485
number of total mutant structures	29,859
number of total single point mutant structures	6,391 / 29,859
number of total double point mutant structures	3,716 / 29,859
number of total >2 point mutant structures	19,752 / 29,859

Fig 2.3 shows the distribution of the number of protein structures. In Fig 2.3(a), each point represent a protein, with the abscissa and ordinate values being the total numbers of wild type structures and mutant structures, respectively. Fig 2.3(b) shows that, for the proteins in our database, histogram distributions by the number of wild type (blue) structures or mutant (red) structures. In (a), each point represent a protein whose coordinates are the number of wild type and mutant structures, respectively. (B) Histogram distributions of the proteins by the number of wild type (blue) structures or mutant (red) structures. Both figures show that only a small portion of total proteins have a large number of wild type and mutant structures.



(a) (b)

Figure 2.3: Distributions of number of wild type and mutant structures.

Table 2.2 shows 23 proteins that have more than 100 wild type or mutant structures in our database. If one desires to obtain all the structures and their PDB-id's for any of these proteins, they can simply run a search with the given protein name, the database will bring all the PDB entries of that protein. This feature can be helpful to those who are interested in obtaining structure ensembles of PDB structures.

Table 2.2: Protein samples having more than 100 wild type or mutant structures.

PROTEIN	# WT	# mut
LYSOZYME C	543	53
DNA POLYMERASE BETA	206	36
CARBONIC ANHYDRASE 2	358	187
MITOGEN-ACTIVATED PROTEIN KINASE 14	152	75
HIV-1 PROTEASE	95	449
HEMOGLOBIN (DEOXY) (BETA CHAIN)	147	77
T4 LYSOZYME	54	482
HEMOGLOBIN SUBUNIT ALPHA	193	31
BETA-2 MICROGLOBULIN (β)	483	29
RIBONUCLEASE A	182	32
CATIONIC TRYPSIN	364	59
REVERSE TRANSCRIPTASE/RNASEH	60	149
CELL DIVISION PROTEIN KINASE 2	340	15
THROMBIN HEAVY CHAIN	268	83
LYSOZYME	43	162
BETA-SECRETASE 1	240	78
INSULIN A CHAIN	214	40
CYTOCHROME C PEROXIDASE, MITOCHONDRIAL	21	184
INSULIN B CHAIN	192	59
P51 REVERSE TRANSCRIPTASE	56	147
UBIQUITIN	202	47
TRANSTHYRETIN	128	75
MYOGLOBIN	93	154

CHAPTER 3. APPLICATION

Since there are a large number of wild type or mutant structures available for many of the proteins in our database, we can carry out many statistical analysis of these structures. Here we focus on the effects of mutations on protein structure and protein dynamics.

To this end, we analyze the differences between wild type ensembles and mutant ensembles of a selected subset of proteins and use the differences to infer how mutations alter protein structure and dynamics.

3.1 The Dataset

To have a statistically meaningful ensemble analysis, a subset of 559 protein are selected from our mutant database. These proteins are selected since for each of them there exist at least 15 wild type structures.

3.2 Methods

3.2.1 Structural alignment

Proteins are not static and fluctuate around their native states. The wildtype/mutant ensemble of a protein represents the structure variations due to such fluctuations. Mutations may cause a significant structure deviation to the protein. To see if this is true and to distinguish structure changes caused by mutations from those naturally exist even among wild type structures due to protein fluctuations, we compare the structural variations in the wild type ensemble and those in the mutant ensemble of the same protein. If mutation indeed causes a significant structural change to a given protein, we should expect to see that structural variations of its mutant ensemble are distinctly different from those of its wild type ensemble.

To that end, we first carry out a structural alignment of all the wild type structures of all the proteins in the data set. This is done as follows:

- Select one of the wild type structures. Align the rest of wildtype structure to it by applying the optimal rotation and translation that minimize the root mean square distance (RMSD) [21].
- Compute the geometric average of all the aligned structures.
- Find the wild type structure that has the smallest RMSD distance to the geometric average and label it as the reference structure.
- Align all wild type and mutant structures of the same protein to the reference structure and get RMSD for each structure
- Compute the average RMSDs of wild type structures and mutant structures respectively.

3.2.2 Principal components analysis

Principal component analysis (PCA) is applied to wild type and mutant ensemble structures [3]. Before applying PCA to an ensemble, we first determine the reference structure and align all the structures in the ensemble (see the last section).

For a protein with r residues, we can represent its r C^{α} atoms using a vector of length $3r$. Assume for this protein, there are n structures. We can write down all the coordinate information of these n structures together in a n by $3r$ coordinate matrix M. We then compute the co-variance matrix C_{ij} in the following manner:

$$C_{ij} =< (M_i - < M >)(M_j - < M >) > \tag{3.1}$$

Brackets <> represent the average of the n structures. We can decompose matrix C as:

$$C = E\Delta E^T, \tag{3.2}$$

where E are the eigenvectors, or the principle components (PCs). The diagonal matrix Δ contains all the variances that correspond to the PCs.

3.2.3 Overlaps between principal components

The first few principal components of an ensemble of protein structures represent the major directions of variations or motions. When including mutants in a structure ensemble that is initially composed of only wild type structures, the inclusion of new structures may alter the dynamics represented by the ensemble. To characterize the effect of mutants on the dynamics, we compute the principal components of motions of the ensemble before and after mutant structures are included. We then compute the overlaps between the corresponding PCs to see to what extent they have been altered.

The overlaps are defined simply as the dot product of the two PCs being compared. Let $\mathbf{p}_i$ and $\mathbf{q}_i$ be the i^{th} PCs of the ensemble before and after mutants are included in the ensemble.

$$overlap_i = \mathbf{p}_i \cdot \mathbf{q}_i. \tag{3.3}$$

A perfect match between two principal components gives an overlap value of 1. The closer to 1 is the overlap, the better is the match.

3.3 Results

To determine if there is significant structural difference as a result of mutations, we compare the structure ensemble of the wild types and the structure ensemble of the mutants for every protein in the data set.

The wild type and mutant database created in this work provides a convenient access to all the available experimental structures for any given protein. These structures of a given protein form an ensemble of conformations that can better describe the native state of the protein than any single structure itself. The ensembles can be used to better understand the native states of proteins and protein functions. Since it includes all the protein structures in the PDB, some systematic studies of all the ensembles may provide new insights. In this section, as one example application, we will use the database to study the effect of mutations on protein structure. the abundance of structure in the database allows us to carry out a statistical analysis of the effect of mutations on structure, and to draw some conclusions about the effects of mutations based solely on experimental structures.

3.3.1 Effects of mutations on protein structure

Mutation in a protein may change the folded structure of the protein. Some mutant structures are significantly different from their corresponding wild type structures, while for the other cases, the changes are insignificant. Fig. 3.1 shows the distribution of structure changes by mutations. Fig. 3.1(a) contains 559 black and red points, each of which represents a protein in our data set (see section 3.1). The average RMSDs of wild types and mutants of 559 proteins are plotted as red and black points. For each point, the range of RMSD fluctuations of the middle 80 percentiles of wild type (and mutant) structures is drawn as horizontal (and vertical) gray line. Red points are the special cases in which the ranges of horizontal and vertical lines do not overlap. For each dot (or protein), the abscissa is the average RMSD between the wild type structures and their reference structure, representing the extent of structure fluctuations within the wild type ensemble. The ordinate is the average RMSD between the mutant structures and their reference structure, representing the extent of structure fluctuations within the mutant ensemble. Red points have a special meaning that will be discussed later. The horizontal/vertical line (or error bar) crossing at each point represents the range of RMSD distribution (again the RMSD is to the reference structure) of the wild type/mutant structures. The range include the middle 80 percentiles only, excluding the extremes (the first and last 10 percentiles). In the figure, most of points (433 over 559) are located above the diagonal line, implying that mutations in general cause a larger structure deviation from the reference structure than what exists in wild type structures.

The 126 points below the diagonal line indicate that the RMSD fluctuations of those proteins are somewhat reduced by mutations. Possibly mutations strengthen some local interactions and make the protein more stable. This implies that, in those cases, mutation suppresses protein flexibility and mutant structures are less deviated from the reference structures than wild type structures.

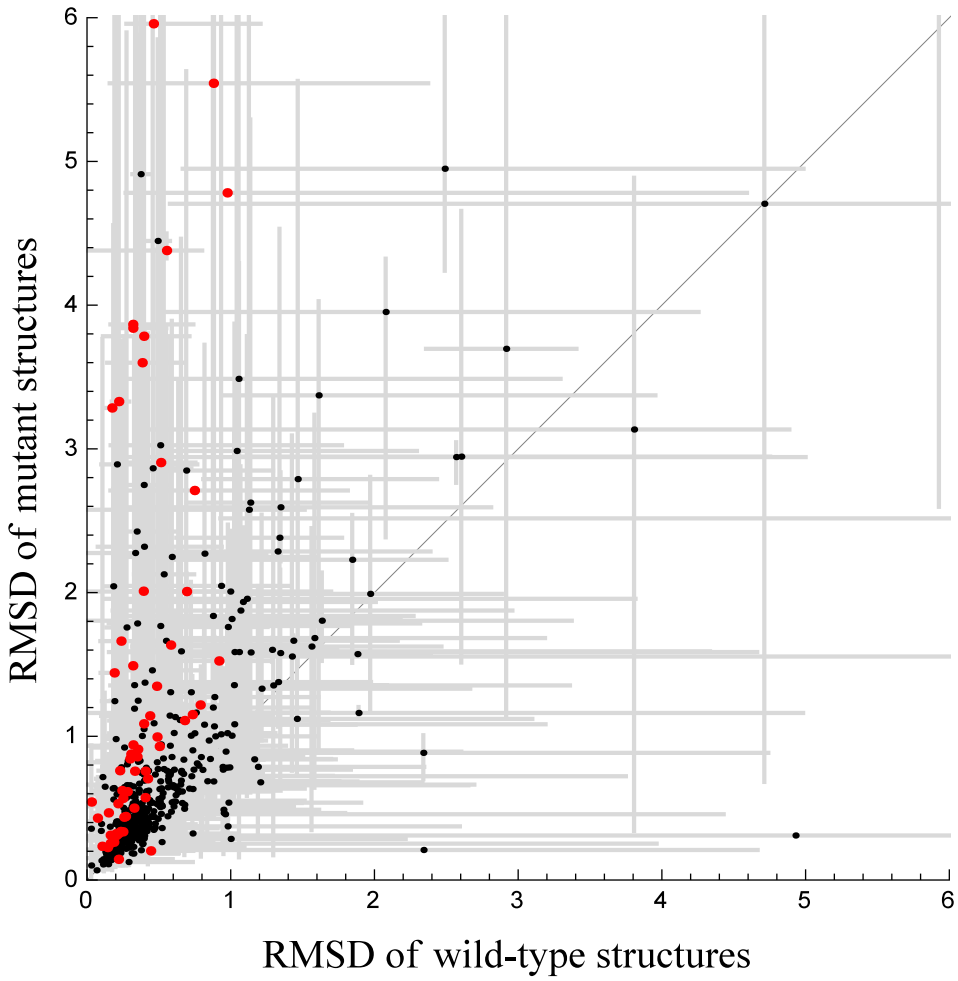


Figure 3.1: Distribution of RMSDs between the reference and the rest of the wild type structures (x-axis) and that between the reference and the rest of mutant structures (y-axis).

Some mutations seem to cause significantly large structural changes. In Fig. 3.1, 70 of total 559 points (each of which represents a protein) are colored red, for which proteins the range of RMSD fluctuations (of the middle 80 percentiles) of mutant structures does not overlap with that of wild type structures, indicating that the fluctuations within the wild type and mutant protein structures are distinctly different.

Out of these 70 proteins, we select those that have five or more wild type and mutant structures. This results in 17 proteins. Two more proteins are further removed since there is only one mutation structure that is accessible under our current procedure. Table 3.1 lists

these proteins: A. PROTEIN (TUBULIN); B. PROTEIN FARNESYLTRANSFERASE SUBUNIT BETA; C. DIHYDROFOLATE REDUCTASE; D. RIBONUCLEOTIDE REDUCTASE R1 PROTEIN; E. CYTOCHROME C OXIDASE POLYPEPTIDE III; F. BETA-SECRETASE 1; G. PROTEIN (S15 RIBOSOMAL PROTEIN); H. TYROSINE-PROTEIN KINASE JAK2; I. GUANINE NUCLEOTIDE-BINDING PROTEIN G(I)/G(S)/G(T) SUBUNIT0 BETA-1; J. TUBULIN ALPHA-1D CHAIN; K. ALPHA ACTIN; L. TERMINAL OXYGENASE COMPONENT OF CARBAZOLE; M. 30S RIBOSOMAL PROTEIN S6; N. REVERSE TRANSCRIPTASE/RNASEH; O. PROTO-ONCOGENE TYROSINE-PROTEIN KINASE SRC. Which correspond to some red points in Fig. 3.1, implying the conformation spaces of mutants are distinctive from (do not overlap with) those of wild types. For each protein, the number of wild type and mutant conformations (or frames) are listed, along with the average (avg) RMSD distances to the reference structures. The pairwise columns represent the mean pairwise RMSD distances within the wildtype and mutant ensembles and that between the two ensembles. For most of these proteins, the mean pairwise distances between wildtype and mutant structures are greater, further confirming that their extents of fluctuations are different. We apply also the ENCORE [22] method to compute the difference between the two sets of ensembles. However, the ENCORE values do not seem to produce a reasonable measure of these ensembles and thus are not included here.

Table 3.1: Proteins whose mutant structures are different from the wild type structures.

NAME	wild types			Between	mutants		
	FrameNum	Avg	Pairwise	Pairwise	Pairwise	FrameNum	Avg
A	58	0.69	1.16	2.08	1.04	40	2.01
B	32	0.32	0.41	0.97	0.15	13	0.94
C	49	0.24	0.36	0.65	0.25	24	0.62
D	16	0.68	0.78	1.03	0.45	15	1.11
E	45	0.1	0.13	0.26	0.15	10	0.23
F	59	0.3	0.43	0.84	0.49	3	0.84
G	88	0.33	0.45	0.79	0.72	3	0.76
H	20	0.79	0.94	1.03	0.28	2	1.22
I	20	0.48	0.65	1.32	0.83	5	1.35
J	27	0.55	0.57	4.31	1.27	29	4.38
K	227	0.17	0.82	6.71	2.81	24	6.69
L	19	0.39	0.50	3.82	0.36	24	3.78
M	98	0.39	0.61	1.95	0.45	8	2.01
N	17	0.51	0.63	3.62	4.69	115	2.90
O	24	0.22	0.26	3.32	5.57	5	3.33

3.3.2 How mutation affects the entropy of a structure?

In this section, we will look into how mutations affect the flexibility of a structure. Does mutation make a structure more flexible or less? To this end, we consider all the "red dot" proteins in Figure 3.1 for which the RMSD fluctuations within the wild type structures and those within the mutant structures are distinctly different. For each of these proteins, we construct a wild type ensemble and a mutant ensemble. We compute the average entropy of these ensembles in the following way. First, we estimate the mean-square fluctuations of each

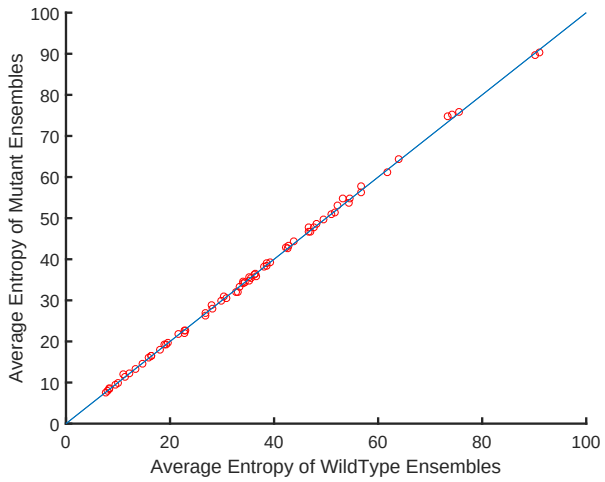


Figure 3.2: The mean entropy of the wild type ensembles and mutant ensembles.

residue by [23, 24]:

$$MSF_i = 1/n_i, \tag{3.4}$$

where n_i is the number of contacts that a residue has with its neighbors. A cutoff distance of 7.3 Å is used when determining if two residues, more precisely their C_α atoms, are in contact (i.e., their separation is less than or equal to the cutoff distance). Once we have the mean square fluctuations, the entropy of the whole structure is computed in the following manner [25],

$$S = \sum_i MSF_i \tag{3.5}$$

The average entropy of an ensemble is then the mean value of the entropies of all the structures in the ensemble.

Figure 3.2 shows a scatter plot of the mean entropy of the wild types and that of the mutants. It shows most wild type structures have a lower entropy, which indicates they are relatively more stable. In other words, most mutations destabilize the structures and make them more flexible. However, some mutations seem to have a lower entropy, indicating they become more rigid.

Since the mean values themselves are quite close and may not be enough to determine if the entropy of wild type ensemble and mutant ensemble are distinctly different. We further select a smaller set of proteins to see if their entropy distributions are different. To this end, we

select those proteins whose wild type structures and mutant structures are both greater than or equal to 20.

Figure 3.3 show the distribution plots of entropy for these six example proteins. It is seen that for some of them, the distributions of the entropy are distinctly different, for some others, the distribution are not separable.

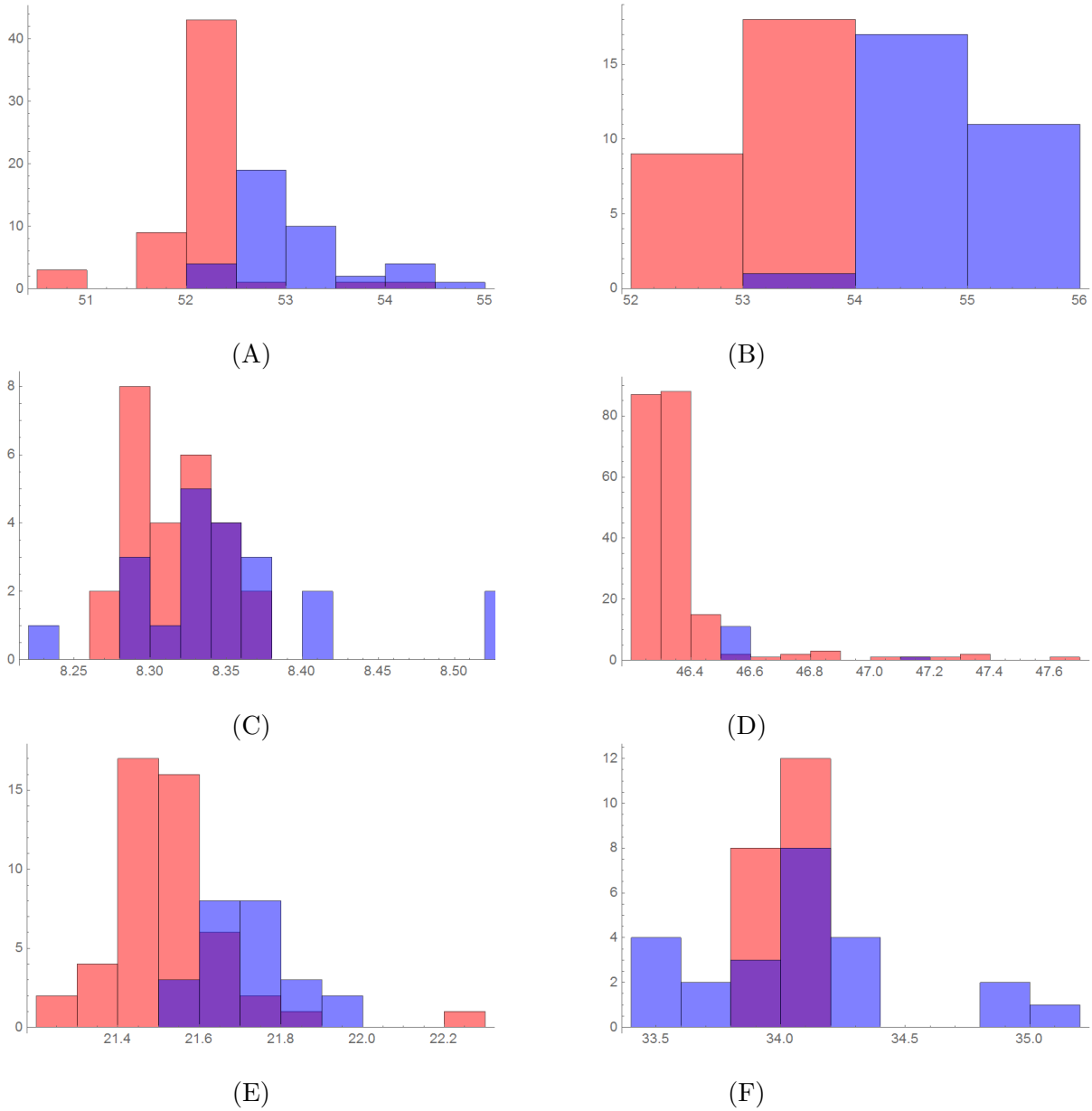


Figure 3.3: The entropy distributions of six proteins.

3.3.3 How mutation affects protein dynamics

Ensembles are commonly used to represent protein conformation heterogeneity and protein dynamics [10, 11, 12, 13, 14]. Best et al. [2] showed that the existing structures in PDB of one protein can capture very well the dynamics of the protein. The principal motions that are encoded in a protein ensemble has been shown to match well with the normal mode motions computed by elastic network model [3]. Mutant structures are often assumed to have the same protein dynamics as wild type structures. They are often mixed with wild type structure in an ensemble, without making any distinction between the two groups. Part of the reason is that there were not that many structures of any protein and a mutant has nearly an identical sequence and it was assumed that mutant structure should have a similar dynamics as the wild type.

The abundance of structures for both wild type and mutant in our data set makes it possible for us to test if such an assumption is valid. That is, when using structures to represent protein dynamics, is it OK to include mutation structures?

To this end, we divide the 559 proteins in our data set into six groups based on the percentage of mutant structures. The composition of six groups are: 1) the first set 100 proteins with mutants ratio 0%-9%, 2) the second set 100 proteins with mutants ratio 9%-17.5%, 3) the third 100 proteins with mutants ratio 17.5%-27.2%, 4) the fourth 100 proteins with mutants ratio 27.2%-40%, 5) the fifth 100 proteins with mutants ratio 40%-59.2%, 6) the last 59 proteins with mutants ratio 59.2%-88.6%.

For each protein in each group, to measure the extent to which principal motions of a structure ensemble are affected by the amount of mutant structures present in the ensemble, we compute the principal motions with (both wild types and mutants) and without the mutants (wild types only) and then calculate the overlaps between the corresponding principal components (i.e., PC1 vs. PC1, PC2 vs. PC2).

Fig. 3.4 shows how principal component 1 (PC1) and principal component 2 (PC2) are affected when different percentages of mutants are present in the ensemble. Based on the figure, the following observations are made:

- At a low percentage, the overlap between PC1s are high, mostly 90% and above.
- As the percentage of mutants increases, the overlap deteriorates.
- For those proteins whose PC1 overlaps are greater than 0.9, we look at their PC2 overlaps. First, at a low percentage (of mutants), PC2 matches well also (i.e., having a high overlap).
- At high percentage of mutants, not only does the number of proteins with high PC1 overlap decrease, even for those with high PC1 overlaps, PC2 overlap decreases as well.

These observations imply that the presence of a large percentage of mutants in the ensemble may alter the dynamics represented by the ensemble.

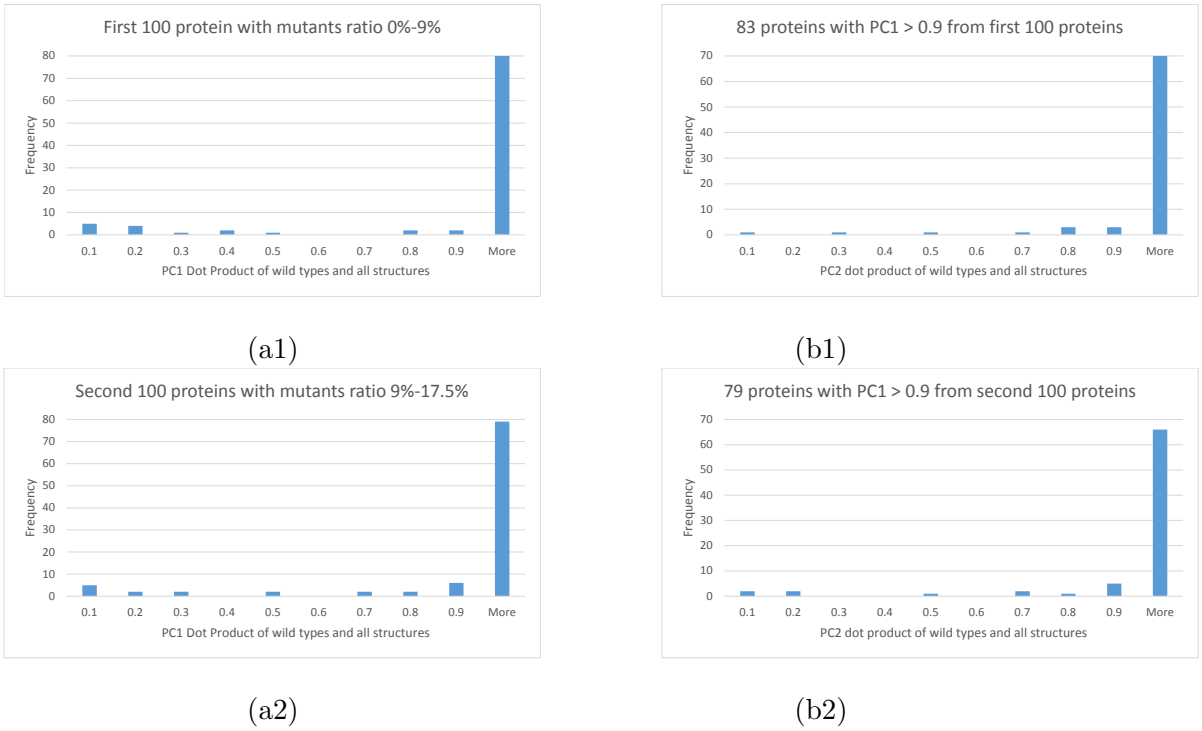
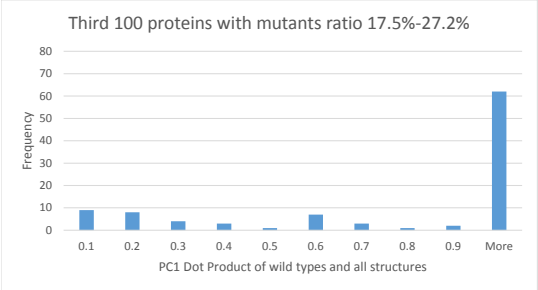
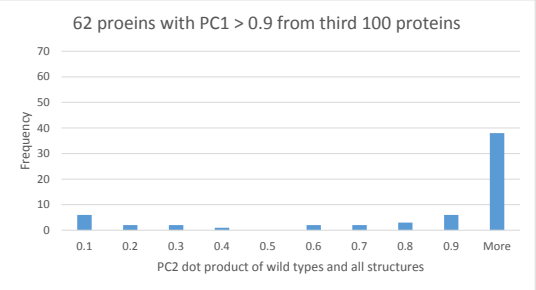


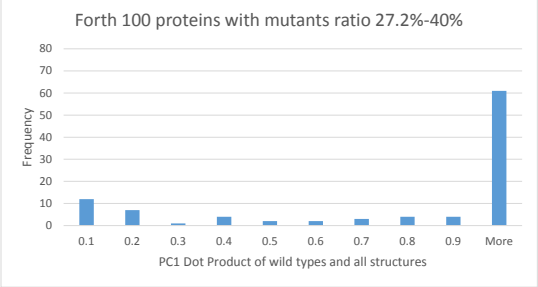
Figure 3.4: How principal components (PC1: left column, PC2: right column) are affected by the inclusion of mutant structures in the ensembles? Different rows show the extent of changes when different percentage of mutant structures are present in the ensembles.



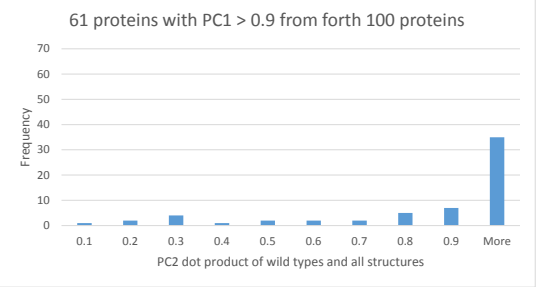
(a3)



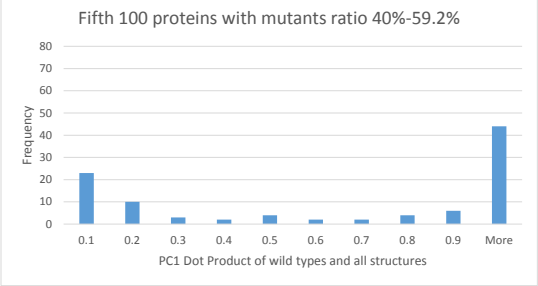
(b3)



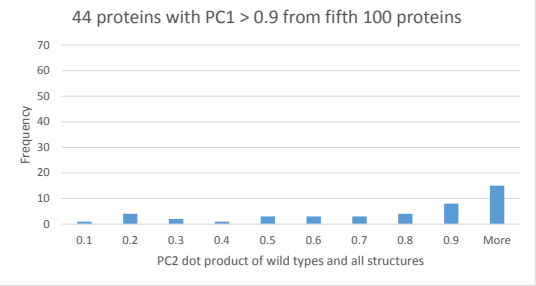
(a4)



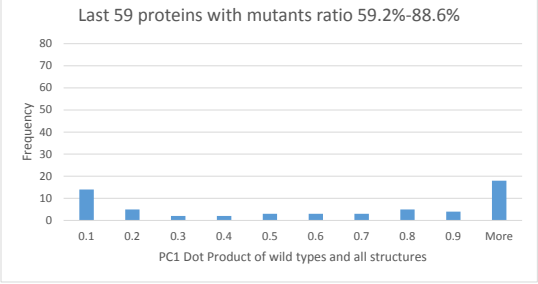
(b4)



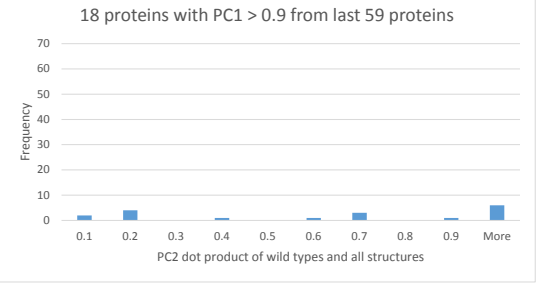
(a5)



(b5)



(a6)



(b6)

Figure 3.4: (continued)

3.3.4 Validation

Realizing there is a possibility that the above result may be due not to the increasing percentage of mutants present in the ensemble, but to the increasing percentage of new structures included into the ensemble, we carry out the following test.

We select 25 proteins that has the largest number of wild type and mutant conformations (at least 84 wild type frames and 24 mutant frames).

For each of these proteins, we start a wild type ensemble using 50 randomly chosen wild type structures. We then gradually add to this ensemble the same amount (10 conformations at a time) of either more wild type structures or mutant structures, as long as there are still structures left to be added. The above procedures are repeated 10 times and the results are averaged. If the changes in PC overlaps seen in Figure 3.4 are purely because an increasing amount of new structures are added to the ensemble and thus alter its dynamics, we should see no difference, i.e., adding more wild type structures or mutant structures has a similar impact on the original wild type ensemble. However, if dynamics is altered more significantly by the inclusion of mutants, we should see a clear difference.

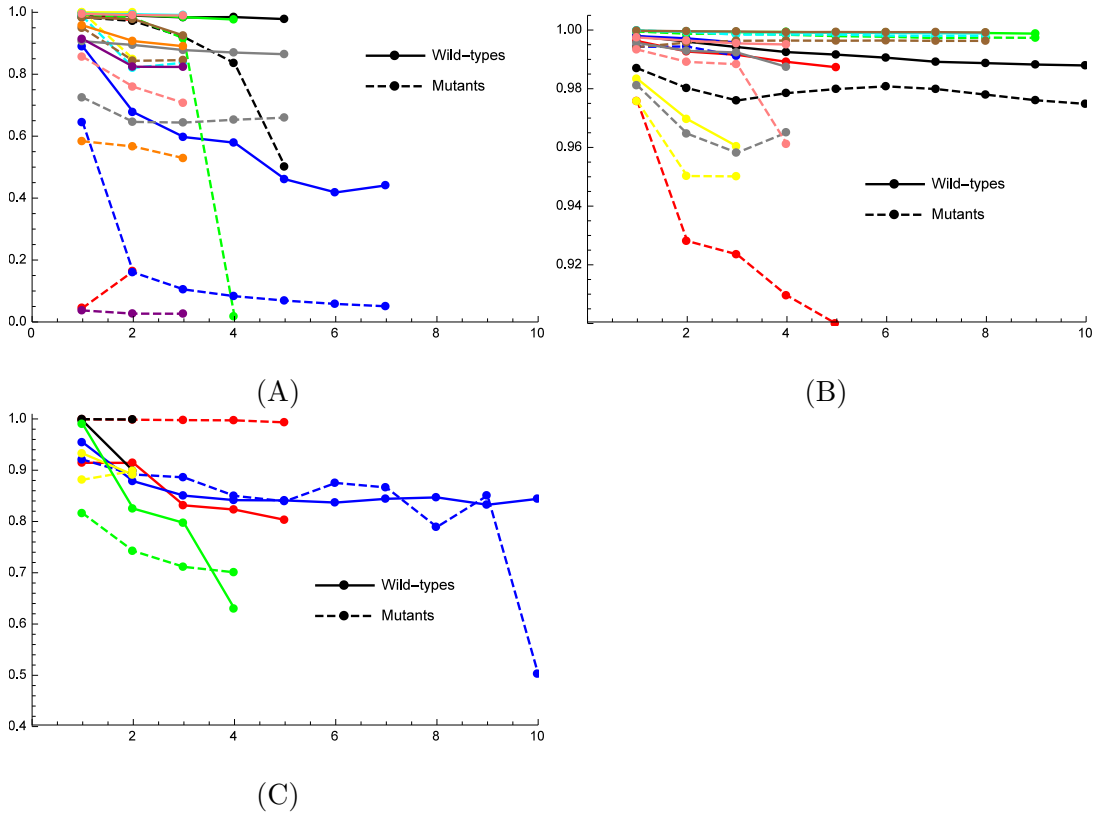


Figure 3.5: Mutants can have a much more pronounced impact on the dynamics of an ensemble. Wild types and mutants of a protein share the same color but individually with sold line and dashed line.

It is seen from Figure 3.5 that **(A, B)** For most proteins, when the same amount of new structures (mutants or more wildtype structures) are included into a wild type ensemble, the dynamics, here overlaps in the principal components, is altered (more) significantly when mutants are included. **(C)** There exists a few cases where the mutants seem to have less effect.

Figure 3.5(A) contains the proteins for which the addition of mutants in the ensemble greatly alter the first PC, as is shown in the significant decreases in overlap.

For the other proteins shown in Figure 3.5(B), the decrease in PC overlaps is smaller (all the overlaps remain greater than 0.9), even though the addition of mutants still brings a larger change to the principal components than the addition of the same number of wild types.

Based on these observations, we would like to make the following recommendation regarding including mutant structures in protein ensembles to represent protein dynamics.

Since a new structure, especially a mutant structure, has the potential to alter the dynamics significantly, one should impose an overlap threshold to prevent the principal motions (PC1, PC2, etc) from being altered too much. One can monitor how much the principal motions (PC1, PC2 etc) are altered when a new structure is considered for inclusion and allow no mutant structure to be added if it causes the principal motions to deviate from the original principal motions beyond the given overlap threshold.

CHAPTER 4. CONCLUSION

In this work we have developed a wild type and mutant structure database. Using data from this database we have analyzed how sequence changes (mutations) affect protein structure and dynamics.

In the future, we will study if excluding mutants in ensembles will produce a better match between normal modes (such as those computed from ANM model [26]) and principal components.

BIBLIOGRAPHY

[1] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–42, 2000.

[2] R. B. Best, K. Lindorff-Larsen, M. A. DePristo, and M. Vendruscolo. Relation between native ensembles and experimental structures of proteins. *Proc Natl Acad Sci U S A*, 103(29):10901–6, 2006.

[3] L. Yang, G. Song, A. Carriquiry, and R. L. Jernigan. Close correspondence between the motions from principal component analysis of multiple hiv-1 protease structures and elastic network modes. *Structure*, 16:321–330, February 2008.

[4] E. Eyal, L.-W. Yang, and I. Bahar. Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics*, 22(21):2619–2627, 2006.

[5] Ahmet Bakan, Lidio M. Meireles, and Ivet Bahar. Prody: Protein dynamics inferred from theory and experiments. *Bioinformatics*, 27(11):1575, 2011.

[6] Lars Skjærven, Shashank Jariwala, Xin-Qiu Yao, and Barry J. Grant. Online interactive analysis of protein structure ensembles with bio3d-web. *Bioinformatics*, 2016. In press.

[7] M. Levitt, C. Sander, and P. S. Stern. The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. *Int. J. Quant. Chem.*, 10:181–199, 1983.

[8] B. Brooks and M. Karplus. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA*, 80(21):6571–6575, November 1983.

[9] N. Go, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA*, 80(12):3696–3700, June 1983.

[10] K. Lindorff-Larsen, R. B. Best, M. A. DePristo, C. M. Dobson, and M. Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433:128–132, January 2005.

[11] B. Richter, J. Gsponer, P. Varnai, X. Salvatella, and M. Vendruscolo. The mumo (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *Journal of Biomolecular Nmr*, 37(2):117–35, 2007.

[12] O. F. Lange, N. A. Lakomek, C. Fares, G. F. Schroder, K. F. Walter, S. Becker, J. Meiler, H. Grubmuller, C. Griesinger, and B. L. de Groot. Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. *Science*, 320(5882):1471–5, 2008.

[13] Vijay Vammi, Tu-Liang Lin, and Guang Song. Enhancing the quality of protein conformation ensembles with relative populations. *Journal of Biomolecular NMR*, 58(3):209–225, 2014.

[14] Vijay Vammi and Guang Song. Ensembles of a small number of conformations with relative populations. *Journal of Biomolecular NMR*, 63(4):341–351, 2015.

[15] T. Kawabata, M. Ota, and K. Nishikawa. The protein mutant database. *Nucl. Acids Res.*, 27:355–357, 1999.

[16] M. Michael Gromiha, J. An, H. Kono, M. Oobatake, H. Uedaira, and A. Sarai. Protherm: Thermodynamic database for proteins and mutants. *Nucl. Acids Res.*, 27:286–288, 1999.

[17] M. Fischer and J. Pleiss. The lipase engineering database: a navigation and analysis tool for protein families. *Nucl. Acids Res.*, 31:319–321, Jan 2003.

[18] L. Whitmore and B. A. Wallace. The peptaibol database: a database for sequences and structures of naturally occurring peptaibols. *Nucl. Acids Res.*, 32:D593–D594, Jan 2004.

[19] A. d'Acierno, A. Facchiano, and A. Marabotti. Galt protein database, a bioinformatics resource for the management and analysis of structural features of a galactosemia-related protein and its mutants. *Genomics, Proteomics & Bioinformatics*, 7:71–76, June 2009.

[20] Vikram Alva, Seung-Zin Nam, Johannes Sding, and Andrei N. Lupas. The mpi bioinformatics toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Research*, 44(W1):W410–W415, 2016.

[21] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.*, 32:922, 1976.

[22] M. Tiberti, E. Papaleo, T. Bengtsen, W. Boomsma, and K. Lindorff-Larsen. Encore: Software for quantitative ensemble comparison. *PLoS Comput Biol*, 2015.

[23] Bertil Halle. Flexibility and packing in proteins. *Proceedings of the National Academy of Sciences*, 99(3):1274–1279, 2002.

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

[25] Michael T. Zimmermann, Sumudu P. Leelananda, Andrzej Kloczkowski, and Robert L. Jernigan. Combining statistical potentials with dynamics-based entropies improves selection from protein decoys and docking poses. *The Journal of Physical Chemistry B*, 116(23):6725–6731, 2012.

[26] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80(1):505–515, January 2001.