PROTEIN FUNCTIONAL SITE PREDICTION USING THE SHORTEST-PATH GRAPH KERNEL METHOD

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By
Malinda Vikum Sanjaka Benaragama Vidanelage

In Partial Fulfillment
for the Degree of
MASTER OF SCIENCE

Major Program:
Software Engineering

April 2013

Fargo, North Dakota

North Dakota State University
Graduate School

Title

PROTEIN FUNCTIONAL SITE PREDICTION USING THE SHORTEST-PATH GRAPH KERNEL METHOD

By

MALINDA VIKUM SANJAKA BENARAGAMA VIDANELAGE

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Changhui Yan
Chair

Dr. Jun Kong

Dr. Juan Li

Dr. Nan Yu

Approved:

June-19-2013	**Dr. Brian Slator**
Date	Department Chair

ABSTRACT

Over the past decade Structural Genomics projects have accumulated structural data for over 75,000 proteins, but the function of most of them are unknown due to limitation of laboratory approaches for discovering the functionality of proteins. Computational methods play key roles to minimize this gap. Graphs are often used to describe and analyze the geometry and physicochemical composition of bimolecular structures such as, chemical compounds and protein functional sites.

In this study, we developed an innovative graph method to represent protein surface based on how amino acid residues contact with each other. Further, we implemented a shortest-path graph kernel method to calculate similarities between the graphs. The nearest-neighbor method was used to compare the similarity of kernel values and predict functional sites of protein structures.

The proposed approach achieved accuracy as high as 77.1% and would provide a useful tool for functional site prediction.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my adviser Dr. Changhui Yan for his continuous encouragements, guidance, and supports to complete this paper successfully. My sincere thanks also go to my committee members, Dr. Juan (Jen) Li, Dr. Jun Kong and Dr. Nan Yu for their willingness to serve as committee members.

Finally, thanks to my wife for her help and suggestions to complete this paper successfully.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1. Problem Statements

Over the past decade Structural Genomics (SG) projects have accumulated structural data for over 75,000 proteins. But the function of many of them are unknown. To analyze these structure data could lead to a better understanding of their function [1]. Most of the early deposited protein structures were annotated with at least some understanding of function, even if the functional residues remained incompletely identified. Structural genomics projects resulted in many structures for which not only functional residues, but also the overall biological function of the molecule is unknown [2]. The major challenge of structural biology is to better understand the function of proteins. However, to determine protein function using experimental methods is costly and time-consuming. Therefore, computational methods are necessary for functional analysis for this rapidly growing data.

1.2. Importance of Functional Site Prediction

Functional sites (e.g. Zinc-binding site, phosphorylation site, DNA binding site) include one or more functional residues that collectively provide desired functionality. Such sites include surface pockets that provide interfaces with ligands or catalytic triads for enzymatic activity, etc. [4]. Being able to predict functional sites in a novel protein will help understand its function. To reveal the structural and functional mechanism of an enzyme molecule, we need to know its active site. In addition, to conduct structure based drug design by targeting an enzyme molecule we also need to know its active site. Identification of the functional site which directly mediate drug interaction is the first step in structure-based drug design (e.g. *Mycobacterium tuberculosis* proteins for structure-based drug design) [6]. The compounds which are bound to the target's active site could interfere with protein function [5]. Further, understanding of an active site, its

geometry and physical and chemical properties are necessary for efficient design of inhibitors of malignant proteins [18].

1.2.1. Enzyme Catalytic Site Prediction

Catalytic sites are defined as "residues that are directly involved in the enzyme-mediated reaction pathway, meaning that catalytic sites represent a small subset of all functional sites" [16]. Catalytic site contains catalytic machinery of the enzyme that performs the catalytic reaction. Historically, catalytic site prediction was based on detecting conservation patterns across a family followed by increasingly powerful sequence-based scoring functions [7] .Those methods depend on information from solved 3D structures, analyzing features such as the geometric arrangements of residues, surface geometry electrostatics, energetics and chemical properties [8]. Recently, catalytic site prediction methods combine features derived from sequence and structure or use sequence data in combination with predicted structure features to improve accuracy [9].

1.2.2. Phosphorylation Site Prediction

Phosphorylation is a post-translational modification on proteins to control and regulate mostly every cellular behavior in eukaryotic cells such as DNA repair [10], environmental stress response [11], metabolism [12], immune response [13], and cellular differentiation [14]. Phosphorylated sites are known to be present often in intrinsically disordered regions of proteins that lack unique tertiary structures, and thus less information is available about the structures of phosphorylated sites. An important challenge is the prediction of phosphorylation sites in protein sequences obtained from mass-scale sequencing of genomes. Phosphorylation sites may aid in the determination of the functions of a protein or even differentiating mechanisms of protein functions in healthy and diseased states.

Historically, low-throughput and recently high-throughput biological techniques have been used to discover the phosphorylation site. However, those methods are associated with some limitations such as, low-throughput techniques are time consuming and expensive to perform. In addition, high-throughput methods such as mass spectrometry method showed that the method cannot identify the protein kinase which is responsible for catalyzing the phosphorylation of the given site [15]. Additionally, mass spectrometry method requires very expensive instruments and specialized training to use the instruments. Due to those limitations of both low-throughput and high-throughput methods, computational methods are becoming of more interest to predict phosphorylation sites in proteins. Computational methods that have been used to predict physphorylation sites differ in many aspects:(1) machine learning technique used, (2) the number of residues surrounding the phosphorylation site that are taken into account, (3) whether the method uses only sequence information or also uses structural information, (4) whether the tool includes models specific to a particular family, and (5) the source(s) of known phosphorylation sites used for training and testing the method.

1.3. Existing Methods for Protein Functional Sites Prediction

1.3.1. Template-Based Methods

Template-based methods construct local structural motifs or patterns that characterize functional sites. The structural templates are defined using sets of inter-residue or inter-atomic distances over a set of functional residues. Template-based methods provide an important way of identifying functional residues. However, those methods do not exploit the power of machine learning techniques [19].

1.3.2. Methods that Explore Residues' Microenvironment

Residue microenvironment-based methods are focused on a single residue or position in the structure and its surrounding small environment. FEATURE is the flagship of the residue microenvironment-based methods [20]. FEATURE models different properties of the neighborhood of the functional sites, including atom/residue type, atom/residue physicochemical properties, chemical groups, and secondary structure information. FEATURE has been used to predict different functional sites of proteins including calcium binding sites, disulfide bond-forming sites, enzyme active sites, ATP-binding sites, zinc binding sites, and etc. [19].

1.3.3. Methods that Explore Residues' Larger Environment

A large group of algorithms and tools have been developed to identify particular classes of larger structural neighborhoods, e.g. surface patches, pockets, cavities or clefts, which provide interfaces to ligands or macromolecular partners.

1.3.4. Evolutionary Trace Methods

Evolutionary trace methods discover evolutionary importance of amino acids by correlating their variations with evolutionary divergences [21]. Several variant s of evolutionary trace methods have been developed, including a weighted evolutionary trace method [22], ConSurf [23], and 3D cluster analyses [24].

1.3.5. Methods Based on Graph Representation

Instead of using atomic coordinates directly, graph-based methods start with transforming protein structures into graphs and then exploit various motif finders, graph similarity measures, machine learning methods to discover functional sites. Previous studies have used graph theory to make local structure comparison [25], to predict calcium binding sites [26], and to extract spatial motifs in protein structure families [27].Despite many successes achieved by graph-based

methods, they suffer from the inability to model actual residue positions and the spatial orientation of structural neighborhoods [20].

1.4. Our Approach

In our research approach, we applied the shortest-path kernel method to compare the similarity between residues' structural environments and use the resulting similarity measures to predict protein functional sites. The shortest-path graph kernel method is faster than other graph kernel methods. Additionally, with shortest-path graph kernel method, assigning weights on edges of a graph can be implemented easily. We used undirected, labeled, weighted graph. We used the nearest-neighbor method to predict functional sites based on the similarity measures given by the shortest-path graph kernel. We explored three variants of the nearest-neighbor method, namely Max, Average and Top 10 average.

We tested three variants of the nearest neighbor method (NNM), namely NNM_AVE, NNM_MAX, and NNM_TOP10, to build predictors for functional site prediction. For a test example, its pairwise similarities to all examples in the training set were calculated using the shortest-path graph kernel. The three NNMs were defined as follows: (1) Let Ave_pos be the average similarity between the test example and all positive examples, and Ave_neg be the average similarity between the test example and all negative examples. Then, the NNM_AVE method predicted the test example to be a functional site if Ave_pos ≥ Ave_neg, and non-functional otherwise. The prediction score for the test example was defined as Ave_pos-Ave_neg; (2) Let Max_pos be the maximum similarity between the test example and all positive examples, and Max_neg be the maximum similarity between the test example and all negative examples. The NNM_AVE method predicted the test example to be a functional site if Max_pos ≥ Max_neg, and non-functional site otherwise. The prediction score for the test example was

defined as Max_pos-Max_neg; (3) Let Top10_pos be the average of the 10 highest similarities between the test example and all positive examples, and Top10_neg be the average of the 10 highest similarities between the test example and all negative examples. In the NNM_TOP10 method, the test example was predicted to be a functional site if Top10_pos ≥ Top10_neg, and non-functional site otherwise. The prediction score for the test example was defined as Top10_pos-Top10_neg. All the predictors were evaluated using leave-one-out cross-validation at protein level, so that when an example was used as the test example, examples from the same proteins were removed from the training set.

2. LITERATURE REVIEW

2.1. Graph

A graph G=<V, E> consists of a set of vertices (also known as nodes) V and a set of edges (also known as arcs) E. An edge connects two vertices u and v; v is said to be adjacent to u. In a directed graph, each edge has a direction from u to v and is written as an ordered pair <u,v> or u->v [29].In an undirected graph, an edge has no direction and is written as an unordered pair {u,v} or u<->v. An undirected graph can be represented by a directed graph if every undirected edge {u,v} is represented by two directed edges <u,v> and <v,u>.

A path in G is a sequence of vertices <v0, v1, v2, ..., vn> such that <vi,vi+1> (or {vi,vi+1}), for each i from 0 to n-1, is an edge in G. The path is simple if no two vertices are identical. The path is a cycle if v0=vn. The path is a simple cycle if v0=vn and no other two vertices are identical.

Graphs are useful for representing networks and maps of roads, railways, airline routes, pipe systems, telephone lines, electrical connections, prerequisites amongst courses, dependencies amongst tasks in a manufacturing system, and many other data. A rooted tree is a special directed graph with a root and that an unrooted tree is a special kind of undirected graph [30].

2.1.1. Undirected Graph

An undirected graph is a graph in which the nodes are connected by *undirected arcs*. An undirected arc is an edge that has no arrow. Both ends of an undirected arc are equivalent; there is no head or tail. Therefore, an edge in an undirected graph is represented as a set rather than an ordered pair.

An Undirected graph is an ordered pair with the following properties:

1. The first component $\mathcal{V}$ is a finite, non-empty set. The elements of $\mathcal{V}$ are called the *vertices* of *G*.

2. The second component $\mathcal{E}$ is a finite set of sets. Each element of $\mathcal{E}$ is a set that is comprised of exactly two (distinct) vertices. The elements of $\mathcal{E}$ are called the *edges* of *G*.

   For example, consider the undirected graph comprised of four vertices and four edges:

$$\begin{aligned} \mathcal{V}_4 &= \{a, b, c, d\} \\ \mathcal{E}_4 &= \{\{a, b\}, \{a, c\}, \{b, c\}, \{c, d\}\} \end{aligned}$$

The graph can be represented graphically as shown in Figure 1[31]. The vertices are represented by appropriately labeled circles, and the edges are represented by lines that connect associated vertices.

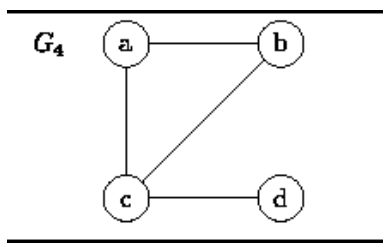


Figure 1: Undirected Graph

2.1.2. Adjacency Matrix

The adjacency matrix, sometimes also called the connection matrix, of a simple graph is a matrix with rows and columns labeled by graph vertices, with a 1 or 0 in position according to weather and are adjacent or not (Figure 2).

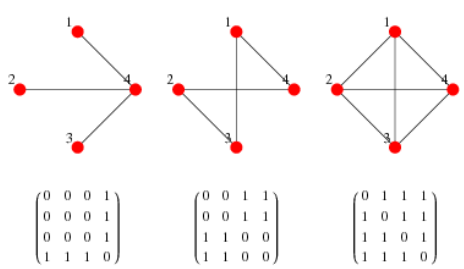
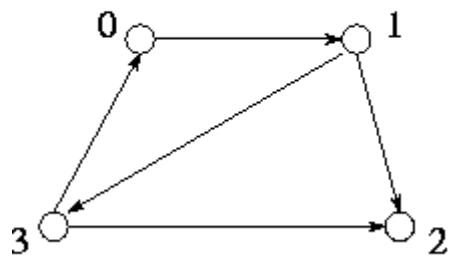


Figure 2: Adjacency Matrix

For a simple graph with no self-loops, the adjacency matrix must have 0s on the diagonal. For an undirected graph, the adjacency matrix is symmetric [32].

2.2. Shortest Distance Path Algorithm

Shortest distance path problem is an important problem in graph theory and has applications in communications, transportation, electronics, and bioinformatics problems.



G

$$A = \begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 2 & 0 & 0 & 0 & 0 \\ 3 & 1 & 0 & 1 & 0 \end{array}$$

Figure 3: Simple Directed Graph, G, and Its Adjacency Matrix, A

The all-pairs shortest-path problem involves finding the shortest path between all pairs of vertices in a graph. A graph $G= (V, E)$ comprises a set V of N vertices, $\{V_i\}$, and a set $E \subseteq V \times V$ of edges connecting vertices in V. In a directed graph, each edge also has a direction, so edges (V_i,V_j) and (V_j,V_i) , $i \neq j$, are distinct. A graph can be represented as an adjacency matrix A in which each element (i,j) represents the edge between element i and j . $A_{ij}=1$ if there is an edge (V_i,V_j) ; otherwise, $A_{ij}=0$ (Figure 3).

2.3. Cross Validation

2.3.1. K-fold Cross-Validation

The K-fold process involves partitioning the data into k separate sets, where k is the number of sets chosen. This division is usually done randomly into k mutually exclusively subset of approximately equal size. This process is illustrated in Figure. 4. A-1 sets are used to train the prediction method, and one set is held out and used to test the prediction method. This process is repeated k times until each subset has been used in the testing set once. In statistical estimation problems, a value of 10 for k is usually chosen.

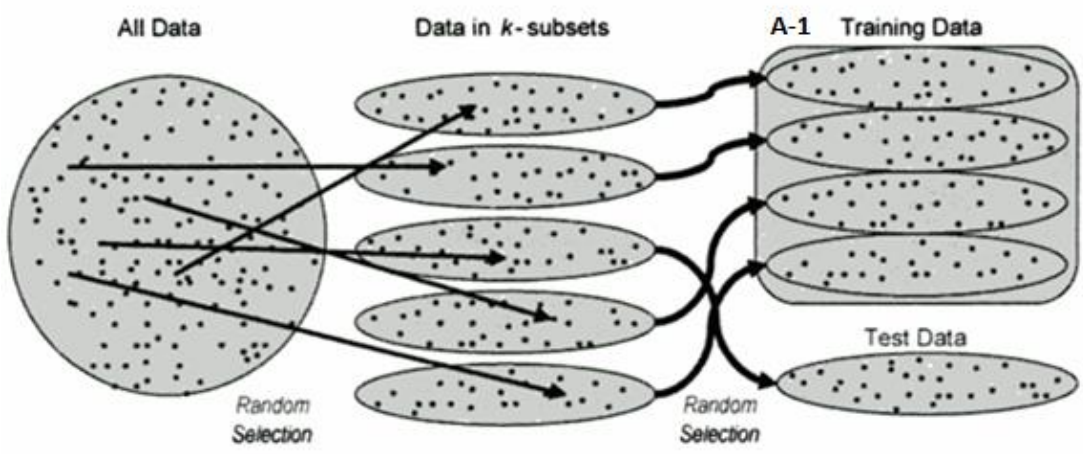


Figure 4: K-Fold Cross- Validation

2.3.2. Leave-One-Out Cross-Validation

The leave-one-out method is a special case of k-folding, where k is equal to the sample size (Figure.5).

Leave-one-out cross validation often works well for continuous-error functions such as the root-mean-square error used in background propagation. It may perform poorly for discontinuous error functions such as misclassification percentage. If a discontinuous error

function is used in the prediction method training, then k-fold cross validation should be used instead of leave-one-out cross-validation [36].

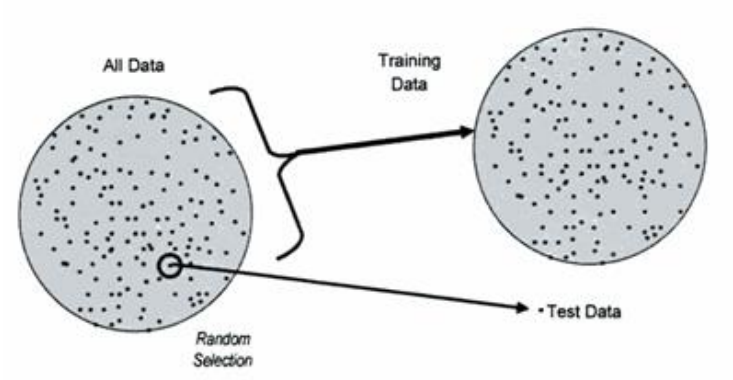


Figure 5: Leave-One-Out Cross-Validation

2.4. True Positive vs. False Positive

We measured the performance of the predictions using the following measurements.

- Accuracy :={(TP +TN) / (TP+TN+FP+FN)}x 100%
- Sensitivity: ={TP / (TP + FN) }x 100%
- Specificity := {TN/(TN+FP)}x 100%

Where, TP is true positive, FP is false positive, TN is true negative, and FN is false negative.

2.5. Percentile

For each protein, we sorted the examples in the order of decreasing prediction scores. We then looked at the ranks of positive examples. The rank of an example was defined as the percentage of examples from the same protein that had higher scores than it. For example, for a given example, if 5% of examples from the same protein had higher prediction scores than it, then its rank was 0.05. Good predictors should assign higher scores to positive examples than to

negatives, thus positive examples should have higher ranks (which correspond to smaller values for ranks) than negative ones.

3. RESEARCH APPROACH

3.1. Problem Statement Overview

Predicting enzyme active-sites in proteins is an important problem not only for protein sciences but also for a variety of practical applications such as drug design. In order to accomplish their biological function, proteins often interact with different types of external molecules such as metal ions, prosthetic groups, and various organic compounds.

Enzymes are a fundamental type of proteins which accelerate chemical processes within a cell, by complexing with the substrate and thus lowering the activation energy of the reaction. Functional residues play various roles in the catalytic process, such as donating electrons or polarizing cofactor bonds. Residues that solely bind substrates, cofactors or metals, are not catalytic residues according to the Catalytic Site Atlas (CSA).

3.2. Research Design

Our goal is to develop effective methods to identify functional residues (active sites) on protein structures. We focus on two types of functional sites, enzyme catalytic sites and Phosphorylation sites.

During our research, we explored various graph kernels methods, including shortest-path graph kernel, random walk, and restricted walk. But, we finally decided to use shortest-path kernel, because other graph kernels were too computationally demanding, making them impractical for large datasets. In the evaluation of our methods, we used balanced datasets that had equal numbers of positive and negative examples. For enzyme catalytic site prediction, a balanced dataset consisting of 201 negative residues and 201 positive residues was chosen from a full dataset that had 20398 negative residues and 201 positive residues. For phosphorylation site prediction, a balanced dataset consisting of 2062 negative residues and 2062 positive residues

was chosen from a full dataset that 139795 negative residues and **2062**-positive residues. The source dataset for enzyme catalytic binding sites was obtained from Catalytic Site Atlas (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/CSA/CSA), the full dataset for phosphorylation site was obtained from http://www.informatics.indiana.edu/predrag/publications.htm.

Each example (residue) was represented using a graph, which included the amino acid residue corresponding to the example and the residues that it contacted. Two residues were considered contacting if the shortest distance between their atoms was less than the sum of the radii of the corresponding atoms plus 0.5 Å. In the graph representation, each amino acid residue was represented using a node labeled with the 20 PSSM values of the residue. An edge was added between two nodes if the corresponding residues were contacting.

A shortest-path graph kernel was used to calculate the similarity. Briefly, the first step of the shortest-path kernel was to transform original graphs into shortest-path graphs. A shortest-path graph had the same nodes as its original graph, and between each pair of nodes, there was an edge labeled with the shortest distance between the two nodes in the original graph. Then, the shortest-path graph kernel compared all pairs of walks of length 1 from different shortest-path graphs. The comparison of a pair of walks included the comparisons of the involved edges and vertices. Two vertices were compared using a Gaussian kernel and two edges were compared using a Brownian kernel.

We used nearest neighbors method to build classifiers for predicting functional sites using the similarity output by the graph kernel method. All the predictors were evaluated using leave-one-out cross-validation at protein level, so that when an example was used as the test example, examples from the same proteins were removed from the training set. Finally in the

evolution stage, we used various statistical measures such as accuracy, sensitivity, specificity, and percentile ranks to evaluate the performance of the proposed predictors. Prediction of enzyme catalytic site residues

3.3. Prediction of Enzyme Catalytic Site Residues

The Catalytic Site Atlas (CSA) is an online database that collects catalytic residue annotation for enzymes in the PDB [37]. The database consists of two types of annotated sites: an original hand-annotated set containing information extracted from the primary literature (LIT) and a homologous set containing residues inferred by PSI- BLAST [39]. In CSA, enzymes were hierarchically organized based on the Enzyme Commission (EC) number. There are six groups at the first level of the hierarchy, which are EC1 through EC6. We used this link [http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/CSA/CSA_Show_EC_List.pl] to download the list of proteins with known catalytic sites. It was a simple process in which the user needs to enter the EC number into the highlighted text boxes as shown in the Figure 6.

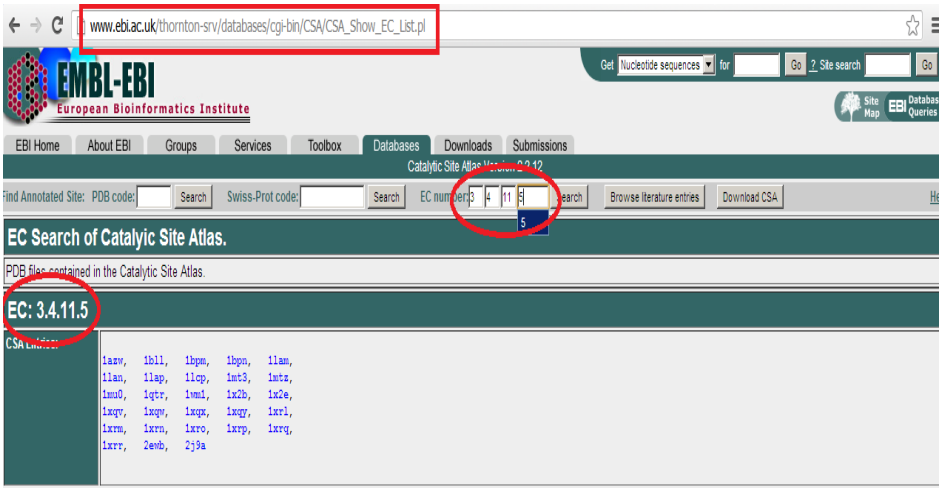


Figure 6: Catalytic Site Atlas Download Site

Then the user click “search” button, so the window shows all PDB IDs which belong to the particular EC group. We then wrote a simple C# code to extract the PDB ID under each group. We examined the number of proteins in each group of the second level. Group EC3.4 had the most proteins at the second level. Thus, we chose EC3.4 as the dataset to test our method. We used program blastclust from the BLAST [18] to remove redundancy so that pairwise similarity between proteins was less than 30%. In the end, 73 proteins were left. There were a total of 201 active catalytic site residues (positive examples) and 20,398 non-catalytic site residues (negative examples) in these proteins. Position-specific scoring matrix (PSSM) of a protein was built by running 4 iterations of PSI-BLAST [18] against the NCBI non-redundant (nr) database. In the PSSM, each residue position was associated with 20 values.

3.3.1. List of Active Residues

Reference: Appendix J

3.3.2. Phosphorylation Site

Proteins with phosphorylation sites were downloaded from http://www.informatics.indiana.edu/predrag/publications.htm. Table 1 shows all protein chains in the dataset. The dataset consists of 679 protein chains with **2062** phosphorylation site residues and 139795 non-phosphorylation resides.

Table 1: Selected Phosphorylation Dataset

1d3v_A	1q46_A	2fmp_A	1z2c_A	1cmf_A	1uss_A	1s70_A	1lkj_A	1neg_A	1z6z_A
1hd7_A	1ebf_A	2bka_A	4nos_A	1php_A	1q2o_A	2fg5_A	1o4x_A	1bla_A	1ni4_A
1th8_B	1m2v_B	1fso_A	1wez_A	1fex_A	1vkh_A	1kx5_B	1kx3_C	1cmz_A	1vyi_A
1tzy_B	1w0j_A	1oxz_A	2f73_A	1pso_E	1oyb_A	1hio_A	1s50_A	1w7b_A	1dfc_A
1x0f_A	1cmi_A	1ghc_A	2b8a_A	1doa_B	1a44_A	1be4_A	1s9j_A	1xd3_A	1gz2_A
1hdi_A	3pmg_A	1h4x_A	1a3w_A	1f60_A	1xfb_A	1hs6_A	1oy2_A	1eo6_A	1w85_A
2b5g_A	2nll_B	1iru_G	1qh4_A	1vjv_A	1rz4_A	1r55_A	1kb9_B	1ok3_A	1hm5_A
2cp6_A	1qk9_A	1g33_A	1ayj_A	2cpe_A	1qkl_A	1rjv_A	1ega_A	1h9f_A	1h9d_B
1jk0_B	1a5e_A	1bd8_A	1my7_A	1u19_A	1ajw_A	1kl9_A	1hlo_A	2ngr_A	1qpg_A
1no8_A	2b3y_A	1df0_A	1ddb_A	1gzd_A	1m6d_A	1okc_A	1cew_I	1xly_A	1e3o_C
1sm2_A	1yat_A	1hhl_A	3pgm_A	3lzt_A	1dn1_A	1qjb_A	1fdj_A	1fw8_A	1pb1_A
1om2_A	1u5r_A	2gj4_A	1ikn_C	1ov3_A	1b89_A	1jey_A	2cqn_A	1lbq_A	1unl_A

Table 1: Selected Phosphorylation Dataset(continued)

1mr3_F	1nw9_B	1ckt_A	1q79_A	1iru_E	1ll2_A	1v6f_A	1ds6_A	1wgf_A	2hf3_A
1ul1_X	1u8f_O	1a5r_A	1rhw_A	1kcm_A	1ukh_A	1g0w_A	1r3s_A	1hwx_A	1eji_A
1hcn_B	1ir3_A	1ig8_A	1gpu_A	1id3_B	1sjq_A	1go4_A	1jh4_A	1mx3_A	1tdq_A
1qsd_A	2btm_A	1cwd_L	1tq4_A	1zs6_A	2trc_P	1tub_A	1usu_A	1vc1_A	1uu3_A
1f7u_A	1eqz_A	1uhg_A	2dmc_A	2cob_A	1s9i_A	1awd_A	1mab_A	1kfu_L	1t2m_A
1ghl_A	1jas_A	1kx5_C	1omw_A	1ajs_A	1ayz_A	1nr7_A	1l0l_H	1k3z_B	2aeb_A
1s3s_G	1mh1_A	1p3m_H	1wnj_A	1hdr_A	1yqa_A	1sph_A	2c2h_A	1iat_A	1ggw_A
1nue_A	1hio_B	1uew_A	1sqn_A	2cq4_A	1psy_A	1u46_A	1gh6_A	1phk_A	1nm1_A
1w8m_A	2fmu_A	1f1g_A	1auz_A	1fw4_A	1z07_A	1ppj_H	1lsg_A	1gji_A	1g0u_D
1xd3_B	1re6_A	1nuy_A	1vjd_A	1twf_B	1pvd_A	1huw_A	1jfi_B	1iru_C	2b1p_A
1q33_A	1g0u_I	1kbh_B	1gt0_C	1hh4_E	1qe3_A	2aqa_A	2b6o_A	1l8b_A	2nzu_L
1mqi_A	1f05_A	1s32_D	1udm_A	1o3x_A	2gjs_A	2axl_A	1ucn_A	1ju5_A	2ayu_A
1nf7_A	1dsy_A	1sqi_A	1uju_A	1rf8_A	1j4n_A	1kx5_A	1e32_A	1nrg_A	1kmq_A
1c44_A	1e9g_A	1beh_A	1dk2_A	1ney_A	1ka5_A	1ud7_A	1t3y_A	1lfb_A	1qhf_A
1tzy_A	1pch_A	2fxu_A	1g62_A	1po5_A	1g9l_A	2f71_A	1x88_A	1oy3_C	1qwt_A
1d2n_A	1mo1_A	2bcg_G	1u3y_A	1one_A	1nvu_S	1qr5_A	1h95_A	1dkf_B	2ayn_A
2lfb_A	1lkk_A	1oqj_A	1iuy_A	1khu_A	1ocs_A	1rw5_A	1nw3_A	1aab_A	1d4b_A
2akz_A	1l8y_A	1ln6_A	2f34_A	1e4u_A	1ikn_D	1hgu_A	1om4_A	2j4z_A	1xmi_A
1iyr_A	2oza_A	1nzp_A	1jk7_A	1nkp_A	1kmt_A	1rho_A	1b66_A	2cwn_A	1t4h_A
1m9m_A	2cui_A	2cpt_A	1ias_A	1mjd_A	1g0u_L	1zok_A	1ng2_A	1cit_A	1vdn_A
1n3k_A	1ef1_C	1bup_A	1nh2_D	1hqs_A	1lfo_A	1t64_A	1tub_B	1xh6_A	1q8g_A
2i47_A	1bfg_A	1axi_A	1ci4_A	1eqz_B	1gky_A	1ptf_A	1cjy_A	1khx_A	1koy_A
2ac0_A	1iu2_A	1wel_A	1m1c_A	1n8p_A	1kx5_D	1x4c_A	1qo5_B	1cvu_A	1r5s_A
1nxk_A	1jdh_A	1olz_A	1mxe_A	4pep_A	1oct_C	1ddj_A	1k8k_A	2fmm_A	1a0r_P
1hm6_A	3psg_A	1x7y_A	1rdq_E	1cb0_A	1xpa_A	1opj_A	2bid_A	1n54_B	2pil_A
1r0w_A	1wg5_A	1jeb_B	1g0u_E	1pin_A	1dgg_A	1yaa_A	3ull_A	1x79_A	1s4b_P
2bcg_Y	1ub1_A	1oey_J	1sms_A	2ad9_A	1q1s_C	1j1b_A	2b0l_A	1h2t_Z	1aoi_C
1jse_A	1ua2_A	2a1l_A	1rwy_A	1sif_A	1ak7_A	1owx_A	1xjd_A	1j3x_A	1rw2_A
1zai_A	1f16_A	1pi1_A	1pme_A	1xw6_A	2c2v_S	1id3_D	2hxm_A	1q1c_A	1st6_A
1u1q_A	1huu_A	1pic_A	1j19_A	1bpo_A	1e31_A	1cm8_A	1ni2_A	1bi9_A	1bd7_A
1k5o_A	1dm5_A	1de4_C	1nd7_A	1r3b_A	2g50_A	1bif_A	1mdy_A	1n0w_A	1qad_A
1liu_A	1gz8_A	2c78_A	1l3k_A	1n0y_A	1ain_A	1sid_A	1w2f_A	1gg2_G	1m4m_A
1a12_A	2c4j_A	1ssu_A	1fot_A	1fmk_A	1sw8_A	1tf7_A	1ef1_A	1omw_G	1srs_A
1hcn_A	1ygp_A	2up1_A	1egw_A	1wk0_A	1nty_A	1jdw_A	1ppj_A	1efc_A	1vbg_A
1ytq_A	1dce_B	1j2m_A	1c3d_A	1ob3_A	1l0b_A	1g8f_A	1yuw_A	1qpc_A	2b5h_A
1rhs_A	1xpc_A	1bxl_A	1wms_A	1dhs_A	1sva_1	2ggm_A	1e5w_A	1bhg_A	1xox_A
3gpd_G	1efv_A	1j8f_A	1yfm_A	1qde_A	1up5_A	1twf_A	1g0u_C	1z0f_A	1yhw_A
1t15_A	1kbl_A	1p7h_L	2b5i_C	1uze_A	2gfs_A	1o6l_A	1pq1_A	1k99_A	1dkg_D
1ydl_A	2ifq_A	1x4n_A	1ctq_A	1p5f_A	1gd0_A	1adt_A	1x5u_A	1umk_A	2cof_A
1w80_A	1vhr_A	1jpa_A	1l6n_A	1byg_A	1byu_A	1efx_A	1i0z_A	1qqd_A	1na7_A
1g83_A	1gl5_A	1qx4_A	1wib_A	1gw5_B	1dsx_A	1f2f_A	1vig_A	1cvj_A	1egx_A
1a81_A	1vg8_A	1p4o_A	2oq1_A	1qly_A	1phr_A	1f68_A	1a5z_A	1hio_D	1amm_A
1u5e_A	1bx4_A	1blx_A	1c0f_A	2cqi_A	1t46_A	1qg3_A	1jr1_A	1fim_A	1qlc_A
1ig4_A	1bf5_A	1w6t_A	1luf_A	1qu6_A	2gst_A	1llc_A	1fhs_A	1oec_A	1bj4_A
1j0x_O	1xkk_A	2if1_A	2f8a_A	1th3_A	1h6v_A	1iru_I	1h7s_A	2shp_A	2dnt_A
1kyf_A	1aww_A	1pa7_A	1kn0_A	1mld_A	2fo0_A	2j7y_A	1qdv_A	1tjx_A	3grs_A
1ldn_A	1e42_A	1kv3_A	1aoa_A	1exb_E	1trn_A	1iyx_A	1fi6_A	1d0n_A	1u5f_A
1g7n_A	1uw2_A	2fym_A	2gdg_A	1fbv_A	1iru_B	1i10_A	1vg0_A	1prx_A	1mcx_A
1hms_A	1ndh_A	2al6_A	1bbz_A	1ftp_A	2hue_C	1iru_2	1p15_A	1yvh_A	1qki_A
2nnq_A	1tad_A	1f8u_A	1pne_A	1lld_A	1ez4_A	1yag_A	1i2m_A	1j7d_B	1awj_A

Table 1: Selected Phosphorylation Dataset(continued)

1u7b_A	1gri_A	1cvs_C	1rv3_A	1fit_A	1d5t_A	2fn4_A	1ds6_B	1fvr_A	1qkm_A
1ej5_A	1o4r_A	1bkl_A	2f9d_A	1k4t_A	1opk_A	5pnt_A	2ch5_A	1bwy_A	1foe_A
1fil_A	1fu6_A	1v18_A	1fgk_A	2h6f_B	1b56_A	1bg1_A	1gcq_A	1us7_B	1xws_A
1qcf_A	1aya_A	1j3d_A	9ldt_A	2fb7_A	1k8k_C	1erj_A	1zww_A	2d4c_A	1oed_E
1yz1_A	1oed_C	1ryh_A	1aui_A	1qy5_A	2b9e_A	1bjt_A	1v04_A	1ul3_A	1taz_A
1mvc_A	2c4k_A	1kqo_A	1gml_A	1tzd_A	1ign_A	1gzk_A	1oed_B	1gp1_A	

3.3.2.1. List of Active Residues

Reference: Appendix I

3.3.3. Balanced Dataset

In our original datasets; the catalytic site dataset consisting of 73 protein chains with 201 active catalytic site residues and 20398 non-active resides and the phosphorylation site dataset consisted of 679 protein chains consisting of 2062 e phosphorylation site residues and 139795 non-active resides. Both datasets were extremely unbalanced. Using such datasets to evaluate prediction methods is problematic. Thus, we constructed balanced datasets which consisted of all the positive residues and equal number of randomly selected negative residues. The balanced dataset for enzyme catalytic sites consisted of 201 active residues and 201 non-active sites, and the balanced dataset for phosphorylation site consisted of 2062 active site residues and 2062 non-active site residues.

3.3.4. Position-Specific Scoring Matrix Calculations (PSSM)

The PSSM shows revolutionary conservation on each residue position of the proteins. Before we calculated the PSSM, we needed to download the PDB files of each corresponding protein sequence [http://ftp.wwpdb.org/pub/pdb/data/biounit/coordinates/all/]. These files are formatted as .gz files; therefore, we needed to decompress the file before using it in the calculations. We used blast-2.2.25+ program and NR database to calculate PSSM for proteins on Microsoft Windows and wrote a simple program process the files in a batch as shown below.

```
Process p = new Process();
p.StartInfo.UseShellExecute = false;
p.StartInfo.RedirectStandardOutput = true;
p.StartInfo.FileName = "C:\\blast-2.2.25+\\bin\\psiblast.exe";
p.StartInfo.Arguments = string.Format("{0}", "-query " + FileNameIN + " -db C:\\blast-2.2.25+\\db\\nr -num_iterations
                        2 -out_ascii_pssm " + FileNameOUT);
p.Start();
```

The full code of the above process is in Appendix C.

The program generated PSSM file for each of the protein sequence. The following example illustrates a line in a PSSM file. In this example, the first column show the sequence index of the amino acid, and “A” is the identify of it, the following 20 values are the PSSM values for this position. At the end of the line, 0.59 is entropy value of associated with this position.

Example: Sample record of .PSSM

```
1 A   5 -2 -2 -2 -1 -1 -2  1 -2 -2 -3 -1 -2 -3 -2  2 -1 -3 -3 -1   77   0   0   0   0   0   0 10   0   0   0   0   0   0   0   0 13   0   0   0   0 0.59 1.#J
```

3.3.5. Calculate Distance between Atoms and Check the Contacting

In our research approach, the number of residues in a graph is decided by the number of neighboring residues that the center residue contacts. So, we needed to calculate the distance between all pairs of residues. The following process [Figure 7] was used to calculate the distance between all pairs of residues.

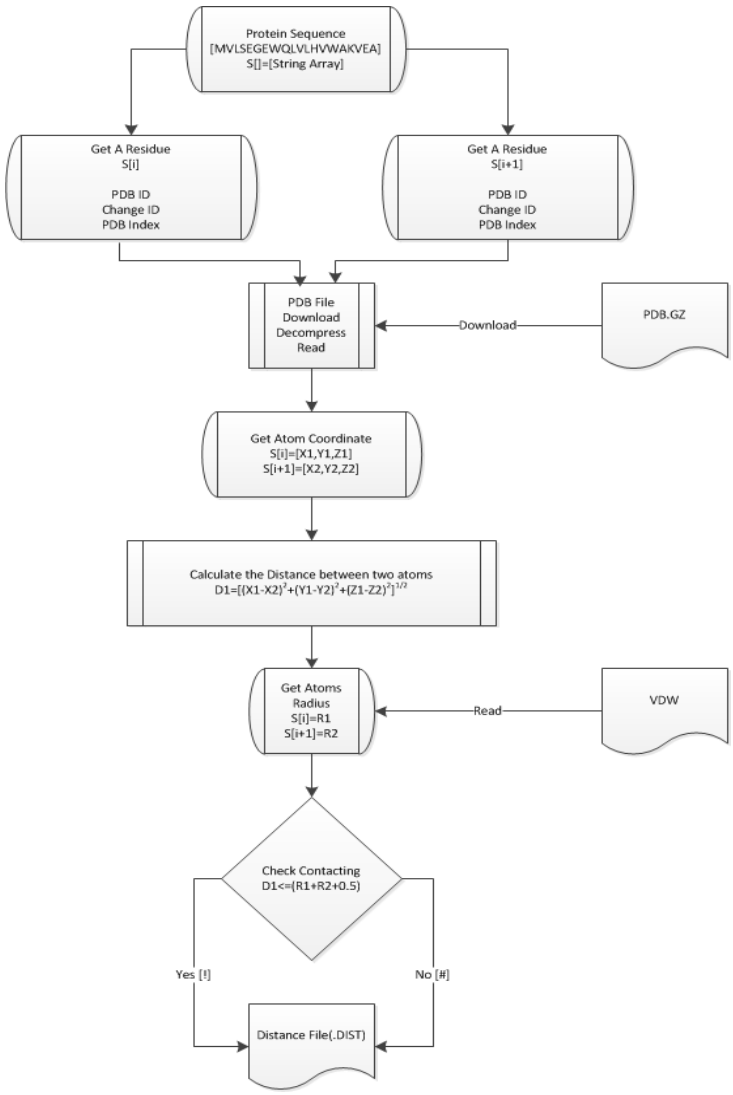


Figure 7: Process of Check Contacting

As shown in Figure 7, first we needed to download the PDB file of each protein chain. Then the PDB file was used to extract the geometric coordinates of the given residue (x, y, z). Next we used these coordinates and applied the following formula to calculate the distance between known atoms.

$$d1=\sqrt{(x1-x2)^2+(y1-y2)^2+(z1-z2)^2}$$

d1 is the distance between any two atoms while (x1, y1, z1) and (x2, y2, z2) are the geometric coordinate of the two atoms.

Van der Waals radii of atoms were used to determine whether two atoms are contacting using the following formula.

$$D1 <= (R1+R2+0.5)$$

Where D1 is the distance between two atoms and R1 and R2 are radius for the two atoms. If the above condition is true when the two atoms contact, In order to implement this calculation process, we wrote a C# program code (see Appendix G). Our program automatically created a distance matrix (.dist) file which included all the atom combinations of a given sequence with information whether the atoms contacted or not.

For example: the residue (atom) in the PDB index 2 position and residue (atom) in PDB index 3 position have contacted each other. So, as noted before in our experiment, these contacted residues were used to create a sub graph.

2 A _ 3 A! : 1.33441
Example of a non-contact residue.
4 A _ 2 A : 4.14432

The above record indicates that the residue in PDB index position 2 and the residue in PDB index position 4 were not contacted.

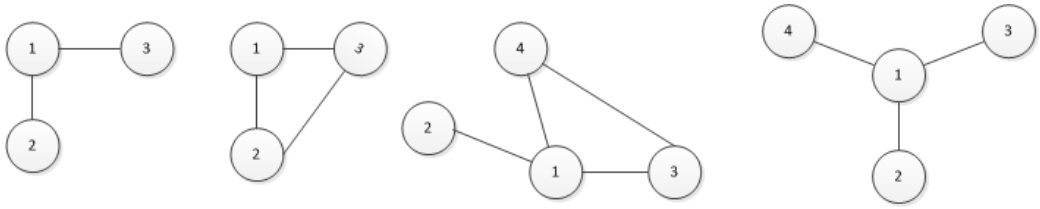
The program we wrote generated the matrix with residues combination. For example, if we had 100 residues on the particular PDB file, the program made 100×100 records as its output.

A part of the sample output file is in Appendix J. We did not attach a full output PDB file because of the large number of records it contains. Other important thing is program running time so we used the Windows operating system with a dual process machine. It took approximately 3 days to complete the distance calculation on a 100 protein chain sequence.

3.3.6. Generate Set of Graphs

Each residue was represented using a graph, which included the amino acid residue corresponding of interest and the residues that it contacted. In the graph representation, each amino acid residue was represented using a node labeled with the 20 PSSM values of the residue. An edge was added between two nodes if the corresponding residues were contacting.

Some examples of graphs and their adjacent matrixes are shown in Figure 8.The first column and first row shows the label of the each residue which makes it easier to identify each residue in a graph. The rest of the rows and columns indicate whether or not these residue contact each other. As a default, we used 0 to indicate same residue. That is why the diagonal of the matrix became 0 while 0 in other positions indicate that these two atoms do not contact. The 1 means those residues contact each other.



	1	2	3
1	0	1	1
2	1	0	0
3	1	0	0

	1	2	3
1	0	1	1
2	1	0	1
3	1	1	0

	1	2	3	4
1	0	1	1	1
2	1	0	0	0
3	1	0	0	1
4	1	0	1	0

	1	2	3	4
1	0	1	1	1
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0

Figure 8: Adjacent Matrix with Undirected Graph

We developed a program to implement the shortest-path graph kernel on the adjacent matrix which converts the original graph into shortest-path graph.

3.3.7. Normalization Labels of Vertices

We normalized the labels of the graph vertices using the linear normalization process. For each attribute, we found the maximum and the minimal values. Then we used the following formula to normalized vertex labels for all the graphs.

Linear Normalization(X1) =(X-Min)/ (Max-Min)

X1 is the normalized value and X is the original value, Min is the minimum value and Max is the maximum value.

3.4. Development of Kernel for Calculating Similarity between Two Graphs

We used the shortest-path graph kernel to compare two graphs as proposed in [40]. The first step of the shortest-path kernel is to transform original graphs into shortest-path graphs. The shortest-path graph has the same nodes as its original graph, and between each pair of nodes, there is an edge labeled with the shortest distance between the two nodes in the original graph. In the current study, the edge label was referred to as the weight of the edge. This transformation can be done using any algorithm that solves the all-pairs-shortest-paths problem. In the current study, the Floyd-Warshall algorithm was used.

Let G_1 and G_2 be two original graphs. They are transformed into shortest-path graphs S_1 (V_1, E_1) and S_2 (V_2, E_2), where V_1 and V_2 are the sets of nodes in S_1 and S_2 respectively, and E_1 and E_2 are the sets of edges in S_1 and S_2 respectively. Then a kernel function is used to calculate the similarity between G1 and G2 by comparing all pairs of edges between S_1 and S_2.

$$K(G_1, G_2) = \sum_{e_1 \in E_1} \sum_{e_2 \in E_2} k_{edge}(e_1, e_2)$$

Where, k_{edge} ( ) is a kernel function for comparing two edges (including the node labels and the edge weight).

Let *e1* be the edge between nodes v_1 and w_1, and e_2 be the edge between nodes v_2 and w_2. Then,

$$k_{edge}(e_1, e_2) = k_{node}(v_1, v_2) * k_{weight}(e_1, e_2) * k_{node}(w_1, w_2)$$

Where, k_{node}( ) is a kernel function for comparing the labels of two nodes, and k_{weight}( ) is a kernel function for comparing the weights of two edges. These two functions are defined in Borgward et al.(2005):

$$k_{node}(v, w) = \exp(-\frac{\| labels(v) - labels(w) \|^2}{2\delta^2})$$

Where, *labels* (*v*) returns the vector of attributes associated with node *v*. Note that K_{node} () is a Gaussian kernel function. $\frac{1}{2\delta^2}$ was set to 72 by trying different values between 32 and 128 with increments of 2.

$$k_{weight}(e_1, e_2) = \max(0, c - | weight(e_1) - weight(e_2) |)$$

Where, *weight* (*e*) returns the weight of edge *e*. K_{weight} ( ) is a Brownian bridge kernel that assigns the highest value to the edges that are identical in length. Constant *c* was set to 2 as in Borgward et al.(2005).

In our application, the labels of vertices are the PSSM values associated with the corresponding residues.

4. EVALUATION OF THE PREDICTORS

We use the shortest-path graph kernel to calculate similarities between all pairs of graph. We used three variants of the nearest neighbor method to predict functional sites using the similarity measures output by the graph kernel.

4.1. Results and Discussion

In our research, we considered 73 enzymes and 679 proteins with phosphorylation sites. The set of enzymes consisted of 201 catalytic site residues and 20398 non-catalytic site residues. The set of phosphorylation site proteins consisted of 2062 phosphorylation sites and 139795 non-phosphorylation site residues. A balanced dataset were obtained for each of the dataset. The balanced dataset had all the functional residues from the original dataset and an equal number of randomly chosen non-functional residues. We implemented the shortest-path graph kernel to calculate the similarity between graphs and use nearest neighbor method to predict functional site. The prediction results are evaluated using TP, FP, accuracy, sensitivity, specificity. Table 2 and 3 show the results for predicting enzyme catalytic sites and phosphorylation sites respectively. Among the three nearest neighbor variants, the NNM_AVE method achieved the best accuracy, with 77.1% for predicting enzyme catalytic sites and 63.8% for predicting phosphorylation sites among the three variants, NNM_TOP10AVE) achieved the best sensitivity (77.6% sensitivity) for catalytic site prediction, but NNM_MAX achieved the best sensitivity (53.t%) for phosphorylation site prediction Overall, NNM_AVE is the best among the three NNM variants.

We exam the location of the FP predictions on the protein structures and found that some of them are very close to a real functional site residue. For example, among the 46 FP that

Table 2: Results for Predicting Enzyme Catalytic Sites

Enzyme catalytic site													
	TP	**TP %**	**FN**	**FN%**	**FP**	**FP%**	**TN**	**TN%**	**Contact**	**Not Contact**	**Accuracy**	**Sensitivity**	**Specificity**
NNM_MAX	150	74.5%	51	25.3%	64	31.8%	137	68.1%	5	59	71.3%	74.5%	68.1%
NNM_AVE	155	77.1%	46	22.8%	46	22.8%	155	77.1%	5	41	**77.1%**	77.1%	77.1%
NNM_TOP10AVE	156	77.6%	45	22.3%	51	25.3%	150	74.6%	5	46	76.1%	77.6%	74.6%

Table 3: Results for Predicting Phosphorylation Sites

Phosphorylation													
	TP	**TP%**	**FN**	**FN%**	**FP**	**FP%**	**TN**	**TN%**	**Contact**	**Not Contact**	**Accuracy**	**Sensitivity**	**Specificity**
NNM_MAX	1104	53.5%	958	46.4%	758	36.7%	1304	50.1%	73	685	58.3%	53.5%	50.1%
NNM_AVE	1054	51.1%	1008	48.8%	482	23.3%	1580	76.6%	54	428	**63.8%**	51.1%	76.6%
NNM_TOP10AVE	1085	52.6%	977	47.3%	667	32.3	1395	67.6%	60	607	60.1%	52.6%	67.6%

NNM_AVE reported, 5 of them were contacting with a real functional sites. Among the 482 FPs that NNM_AVE reported, 54 of them contact with a real functional site residue.

In addition to TP, FP, accuracy, sensitivity, specificity, we also used the percentile ranking to evaluate the performance of the method. The percentile rank of a residue is the percentages of residues of the same protein that has a higher prediction score than it. In the ideal case, all functional sites have higher prediction scores than non-functional residues, therefore should have very low percentile ranks. Table 4 shows the percentile ranks of all enzyme catalytic residues when NNM_AVE was used.

Table 4: Percentile Rank of Enzyme Catalytic Residues

Protein	[Pos]	Percentile	Protein	[Pos]	Percentile	Protein	[Pos]	Percentile
1mpp_A	Pos2	0.0122699386503067	1sca_A	Pos75	0.0136363636363636	1nlu_A	Pos150	0.295302013422819
1mpp_A	Pos1	0.0153374233128834	1sca_A	Pos74	0.0181818181818182	1nlu_A	Pos149	0.486577181208054
1mpp_A	Pos4	0.0306748466257669	1sca_A	Pos76	0.0454545454545455	1nlu_A	Pos148	0.577181208053691
1mpp_A	Pos3	1	1i78_A	Pos78	0.405498281786942	1nlu_A	Pos151	0.996644295302013
1tyf_A	Pos7	0.0182926829268293	1i78_A	Pos81	0.426116838487972	1qtn_A	Pos153	0.0708661417322835
1tyf_A	Pos8	0.317073170731707	1i78_A	Pos77	0.567010309278351	1qtn_A	Pos155	0.15748031496063
1tyf_A	Pos5	0.48780487804878	1i78_A	Pos80	0.632302405498282	1qtn_A	Pos152	0.165354330708661
1tyf_A	Pos9	0.524390243902439	1i78_A	Pos79	0.951890034364261	1qtn_A	Pos154	1
1tyf_A	Pos6	0.75609756097561	1jhf_A	Pos82	0.0494505494505494	1qx3_A	Pos156	0.0241545893719807
1rtf_B	Pos13	0.00458715596330275	1jhf_A	Pos86	0.0659340659340659	1qx3_A	Pos158	0.0917874396135266
1rtf_B	Pos11	0.0275229357798165	1jhf_A	Pos84	0.186813186813187	1qx3_A	Pos157	0.101449275362319
1rtf_B	Pos12	0.0321100917431193	1jhf_A	Pos85	0.697802197802198	2bkr_A	Pos160	0.0820512820512821
1rtf_B	Pos10	0.0596330275229358	1jhf_A	Pos83	0.906593406593407	2bkr_A	Pos161	0.292307692307692
1hr6_A	Pos14	0.0194174757281553	1t7d_A	Pos88	0.0380434782608696	2bkr_A	Pos162	0.697435897435897
1hr6_B	Pos15	0.0298507462686567	1t7d_A	Pos89	0.25	2bkr_A	Pos163	0.764102564102564
1b65_A	Pos18	0.0936454849498328	1t7d_A	Pos87	0.804347826086957	2bkr_A	Pos159	0.979487179487179
1b65_A	Pos19	0.096989966555184	1rgq_A	Pos93	0.0942028985507246	1cqq_A	Pos167	0.00632911392405063
1b65_A	Pos20	0.100334448160535	1rgq_A	Pos92	0.557971014492754	1cqq_A	Pos166	0.139240506329114

Table 4: Percentile Rank of Enzyme Catalytic Residues(continued)

1b65_A	Pos17	0.284280936454849	1rgq_A	Pos90	0.659420289855073	1cqq_A	Pos164	0.253164556962025
1b65_A	Pos16	0.732441471571906	1rgq_A	Pos91	0.77536231884058	1cqq_A	Pos165	0.39873417721519
1r44_A	Pos22	0.0662983425414365	1pxv_A	Pos94	0.0256410256410256	2fqq_A	Pos168	0.0075187969924812
1r44_A	Pos21	0.325966850828729	2bhg_A	Pos97	0.0517241379310345	2fqq_A	Pos169	0.0150375939849624
1iec_A	Pos23	0.122340425531915	2bhg_A	Pos98	0.35632183908046	2fqq_A	Pos170	0.075187969924812
1iec_A	Pos24	0.579787234042553	2bhg_A	Pos96	0.563218390804598	2fqq_A	Pos171	0.233082706766917
1iec_A	Pos26	0.622340425531915	2bhg_A	Pos95	0.649425287356322	1cvr_A	Pos172	0.350785340314136
1iec_A	Pos27	0.627659574468085	1kfu_L	Pos99	0.100763358778626	1cvr_A	Pos175	0.732984293193717
1iec_A	Pos25	0.920212765957447	1kfu_L	Pos102	0.645801526717557	1cvr_A	Pos174	0.824607329842932
1fo6_A	Pos33	0.119047619047619	1kfu_L	Pos100	0.806106870229008	1cvr_A	Pos173	0.905759162303665
1fo6_A	Pos32	0.222222222222222	1kfu_L	Pos101	0.862595419847328	1nln_A	Pos176	0.00571428571428571
1fo6_A	Pos31	0.369047619047619	1lya_B	Pos103	0.224299065420561	1nln_A	Pos179	0.0171428571428571
1fo6_A	Pos29	0.626984126984127	1lya_A	Pos104	0.0823529411764706	1nln_A	Pos177	0.0628571428571429
1fo6_A	Pos28	0.853174603174603	1ge7_A	Pos105	0.392857142857143	1nln_A	Pos178	0.217142857142857
1fo6_A	Pos30	0.924603174603175	1ge7_A	Pos106	0.607142857142857	1gcb_A	Pos182	0.0193704600484262
1cg2_A	Pos34	0.0706214689265537	1qib_A	Pos107	0.00671140939597315	1gcb_A	Pos183	0.690072639225182
1cg2_A	Pos37	0.107344632768362	1tlp_E	Pos108	0.00357142857142857	1gcb_A	Pos180	0.801452784503632
1cg2_A	Pos36	0.11864406779661	1tlp_E	Pos109	0.342857142857143	1gcb_A	Pos181	0.941888619854722
1cg2_A	Pos38	0.367231638418079	1pwv_A	Pos110	0.499259259259259	1s2k_A	Pos184	0.00574712643678161

Table 4: Percentile Rank of Enzyme Catalytic Residues(continued)

1cg2_A	Pos39	0.748587570621469	1lam_A	Pos113	0.0302267002518892	1s2k_A	Pos185	0.0862068965517241
1cg2_A	Pos35	0.858757062146893	1lam_A	Pos111	0.0755667506297229	1r1j_A	Pos186	0.00317460317460317
1aug_A	Pos41	0.0162162162162162	1lam_A	Pos112	0.544080604534005	1r1j_A	Pos187	0.0111111111111111
1aug_A	Pos42	0.145945945945946	1xqw_A	Pos117	0.109848484848485	1r1j_A	Pos188	0.0492063492063492
1aug_A	Pos40	0.72972972972973	1xqw_A	Pos118	0.113636363636364	1r1j_A	Pos189	0.0936507936507937
1l9x_A	Pos44	0.544747081712062	1xqw_A	Pos115	0.121212121212121	1i1i_P	Pos190	0.0179738562091503
1l9x_A	Pos43	0.906614785992218	1xqw_A	Pos116	0.223484848484848	1i1i_P	Pos191	0.911764705882353
1ca0_B	Pos46	0.00854700854700855	1xqw_A	Pos114	0.856060606060606	1slm_A	Pos192	0.00478468899521531
1ca0_B	Pos45	0.0512820512820513	1a16_A	Pos120	0.0131233595800525	1ast_A	Pos193	0.00549450549450549
1ca0_C	Pos47	0.021978021978022	1a16_A	Pos119	0.05249343832021	1ast_A	Pos194	0.604395604395604
1x9y_A	Pos51	0.003125	1amp_A	Pos121	0.129554655870445	1ck7_A	Pos195	0.163478260869565
1x9y_A	Pos50	0.009375	1ei5_A	Pos124	0.651709401709402	1lml_A	Pos196	0.948717948717949
1x9y_A	Pos49	0.18125	1ei5_A	Pos123	0.696581196581197	1eb6_A	Pos197	0.00636942675159236
1x9y_A	Pos48	0.190625	1ei5_A	Pos125	0.837606837606838	1i1e_A	Pos200	0.263660017346054
8pch_A	Pos54	0.574358974358974	1ei5_A	Pos122	0.88034188034188	1i1e_A	Pos198	0.359930615784909
8pch_A	Pos53	0.635897435897436	1fy2_A	Pos128	0.0105263157894737	1i1e_A	Pos199	0.542931483087598
8pch_A	Pos52	0.958974358974359	1fy2_A	Pos127	0.0263157894736842	2fqq_B	Pos201	0.846153846153846
1cmx_A	Pos56	0.0050251256281407	1fy2_A	Pos126	0.105263157894737			
1cmx_A	Pos55	0.0251256281407035	1fy2_A	Pos129	0.126315789473684			

Table4: Percentile Rank of Enzyme Catalytic Residues(continued)

1cmx_A	**Pos58**	**0.0552763819095477**	**1fy2_A**	**Pos130**	**0.342105263157895**			
1cmx_A	**Pos57**	**0.105527638190955**	**1o8a_A**	**Pos135**	**0.209803921568627**			
1azw_A	**Pos60**	**0.0866425992779783**	**1o8a_A**	**Pos132**	**0.5**			
1azw_A	**Pos61**	**0.209386281588448**	**1o8a_A**	**Pos131**	**0.501960784313725**			
1azw_A	**Pos59**	**0.277978339350181**	**1o8a_A**	**Pos133**	**0.543137254901961**			
1xgm_A	**Pos62**	**0.606299212598425**	**1o8a_A**	**Pos134**	**0.782352941176471**			
1itq_A	**Pos64**	**0.666666666666667**	**1ysc_A**	**Pos136**	**0.119791666666667**			
1itq_A	**Pos63**	**0.828478964401295**	**1ysc_A**	**Pos137**	**0.169270833333333**			
1pfq_A	**Pos67**	**0.00296296296296296**	**1bcr_A**	**Pos138**	**0.0577777777777778**			
1pfq_A	**Pos66**	**0.00444444444444444**	**1bcr_A**	**Pos139**	**0.186666666666667**			
1pfq_A	**Pos65**	**0.00740740740740741**	**1bcr_A**	**Pos140**	**0.191111111111111**			
1cbx_A	**Pos68**	**0.393258426966292**	**1bcr_B**	**Pos141**	**0.264285714285714**			
1cbx_A	**Pos69**	**0.936329588014981**	**1lbu_A**	**Pos142**	**0.948186528497409**			
1ybq_A	**Pos70**	**0.0411764705882353**	**1qfm_A**	**Pos143**	**0.00156494522691706**			
2lpr_A	**Pos71**	**0.0114285714285714**	**1qfm_A**	**Pos145**	**0.00782472613458529**			
2lpr_A	**Pos72**	**0.0285714285714286**	**1qfm_A**	**Pos144**	**0.0156494522691706**			
2lpr_A	**Pos73**	**0.102857142857143**	**1hzf_A**	**Pos146**	**0.759124087591241**			
1sca_A	**Pos75**	**0.0136363636363636**	**1hzf_A**	**Pos147**	**0.886861313868613**			

Figures 9-11 showed the percentile distribution for functional site residues when different NNM variants were used. Figures 12-14 showed the percentile distribution for non-functional site residues. These figures showed most of the functional site residues (ranged from 74 to 76) belong to the 0.0 -0.1 percentile group (Figure 9-11). In contrast most of the non-functional residues (ranged from 20 to 21) belong to 0.9- 1.0 percentile group (Figure 12-14). So the results clearly show that functional site residues enriched in the low percentile ranges, which indicated a good prediction performance.

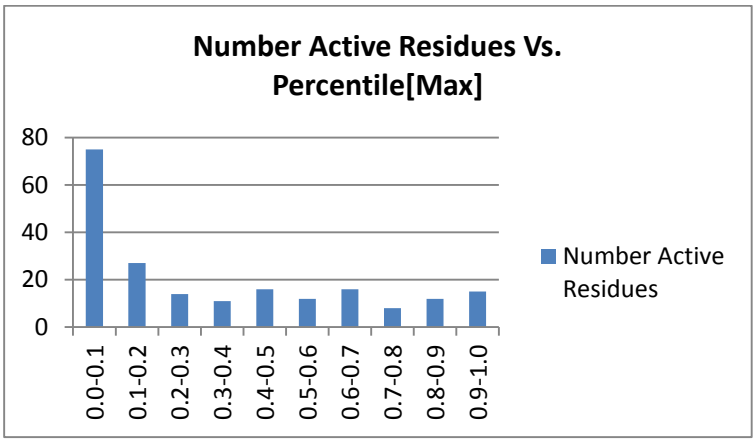


Figure 9: Distribution of Percentile for Functional Residues when NNM_MAX was used

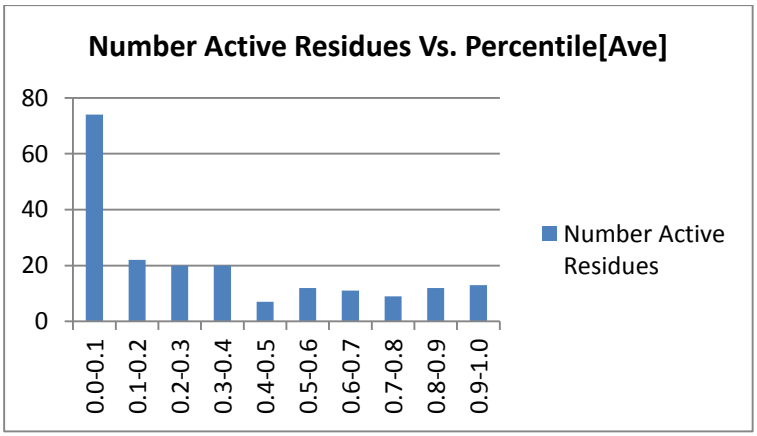


Figure 10: Distribution of Percentile for Functional Residues when NNM_AVE was used

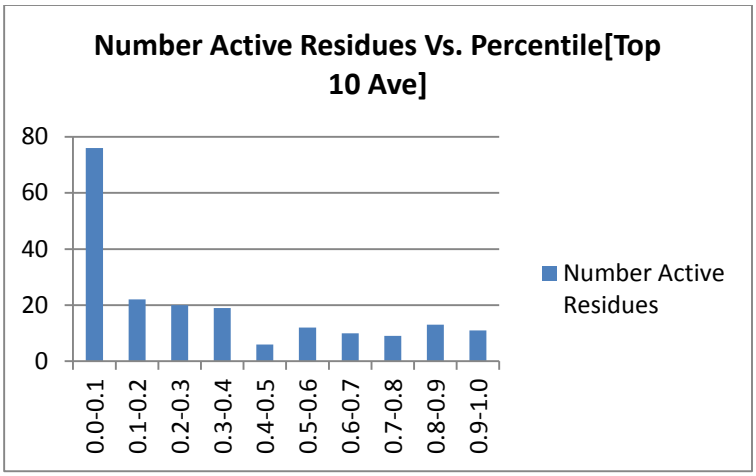


Figure 11: Distribution of Percentile for Functional Residues when NNM_TOP10AVE was used

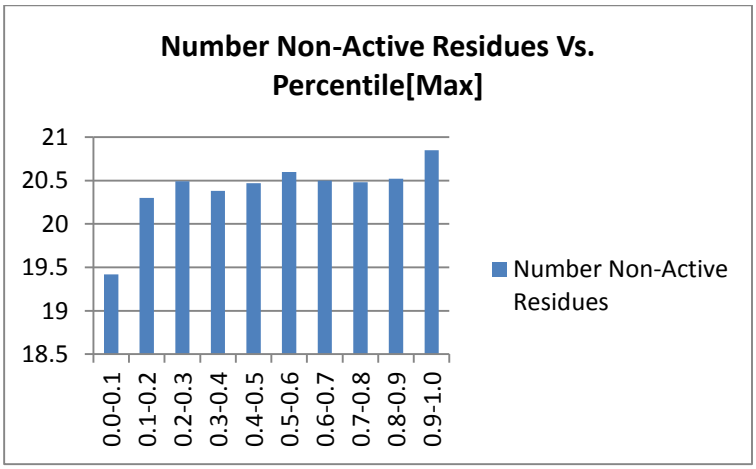


Figure 12: Distribution of Percentile for Non-Functional Residues when NNM_MAX was used

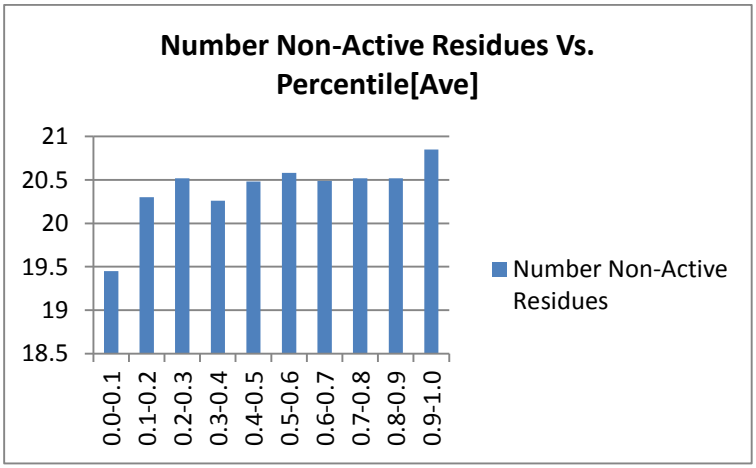


Figure 13: Distribution of Percentile for Non-Functional Residues when NNM_AVE was used

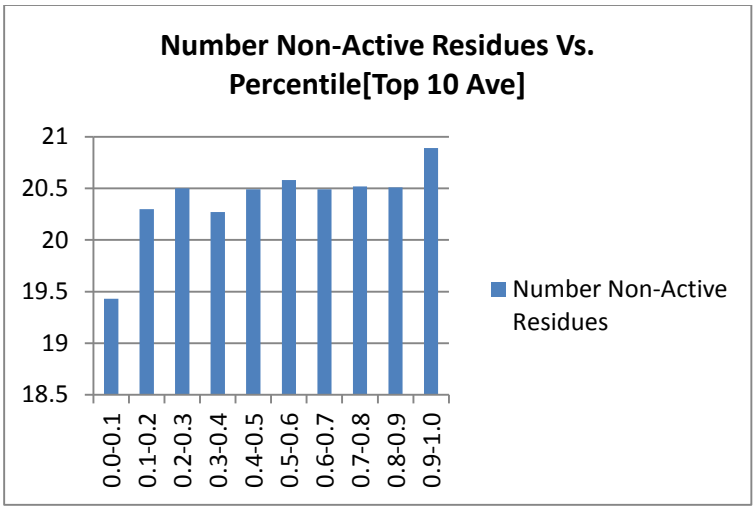


Figure 14: Distribution of Percentile for Non-Functional Residues when NNM_TOP10AVE

In summary, among the three variants of NNM, the NNM_AVE achieved the best results in predicting functional site residues. The enrichment analysis show that in general positive examples that correspond to functional site residues have higher prediction scores, which mean lower percentile ranks, than non-functional site residues.

5. CONCLUSIONS

In this project, we aimed to develop computational methods for predicting protein functional sites. In our approach, each residue on the protein structure was represented using graph, whose nodes were labeled with PSSM vectors that show the evolutionary pressure on each residue position. We develop a shortest-path graph kernel method to compare the similarity between graphs and used three variants of the nearest neighbor method to build classifiers for predicting functional sites. We evaluated the methods using two datasets, enzyme catalytic sites and phosphorylation sites. Leave-one-out cross-validation show that the NNM_AVE achieved 77.1% accuracy in predicting enzyme catalytic sites and 63.8% accuracy in predicting phosphorylation sites. The results also showed that in general, the classification methods assigned higher prediction scores to positive examples, which correspond to functional site residues. Percentile analysis showed that most positive examples have a percentile rank in the range of 0.0-0.1, which indicates very good prediction performance.

6. REFERENCES

[1]. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. (2000) The protein data bank. *Nucleic Acids Res*., 28 (1), 235-242.

[2]. Burley, S. (2000) An overview of structural genomics. *Structural Genomic Supplement*, 932-934.

[3]. Karin M. V.; Judith D. C.; Komandur E. R.; Michael E. W. (2012)Text mining improves prediction of protein functional sites. *PLoS ONE*, 7(2), e32171.

[4]. Fuxiao X.; Predrag R. (2011) Computational methods for identification of functional residues in protein structures. *Current Protein and Peptide Science*, 12, 456-469.

[5]. Laurie A.T.; Jackson R.M. (2006) Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Current Protein Peptide Science*, 7,395–406.

[6]. Chim N.; Habel J.E.; Johnston J.M.; Krieger I.; Miallau L. (2011) The TB structural genomics consortium: a decade of progress. *Tuberculosis (Edinb)*, 91, 155–172.

[7]. Casari,G. (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol*., 2, 171–178.

[8]. Ondrechen, M.J. (2001) THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Nat. Acad. Sci*. USA, 98, 12473–12478.

[9]. Fischer, J.D. (2008) Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, 24, 613–620.

[10]. Wood, C.D. (2009) Nuclear localization of p38MAPKin response to DNA damage. *Int. J. Biol. Sci*., 5, 428–437.

[11]. Wang, Y. Y. (2010) Hydrogen peroxide stress stimulates phosphorylation of FoxO1 in rat aortic endothelial cells. *Acta Pharmacol. Sin*, 31, 160–164.

[12]. Bu, Y. H. (2010) Insulin receptor substrate 1 regulates the cellular differentiation and the matrix metallo peptidase expression of preosteoblastic cells. *J. Endocrinol*., 206, 271–277.

[13]. Kim, S. H.; Lee, C. E. (2011) Counter-regulation mechanism of IL-4 and IFN-± signal transduction through cytosolic retention of the pY-STAT6: pY-STAT2:p48 complex. Eur. *J. Immunol*., 41, 461–472.

[14]. Lian, I. (2010) The role of YAP transcription co-activator in regulating stem cell self-renewal and differentiation. *Genes Dev*., 24, 1106–1118.

[15]. Huttlin, E.L. (2010) A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell*, 143, 1174–1189.

[16]. Swati P.; Amar R.; Dennis R. L. (2007) Prediction of enzyme catalytic sites from sequence using neural networks. Computational intelligence and bioinformatics and computational biology symposium.

[17]. Fuxiao, X.; Predrag, R. (2011) Computational methods for identification of functional residues in protein structures. *Current Protein and Peptide Science*, 12, 456-469.

[18]. Liu, F.; Kovalevsky, A.Y.; Louis, J.M.; Boross, P.I.; Wang, Y.F.; Harrison, R.W.; Weber, I.T.(2006) Mechanism of drug resistance revealed by the crystal structure of the unliganded HIV-1 protease with F53L mutation. *Journal of Molecular Biology*, 358:1191-1199.

[19]. Fuxiao, X.; Predrag, R. (2011) Computational methods for identification of functional residues in protein structures. *Current Protein and Peptide Science*, 12, 456-469.

[20]. Bagley, S.C.; Altman, R.B. (1995) Characterizing the microenvironment surrounding protein sites. *Protein Sci*., 4 (4), 622-635.

[21]. Lichtarge, O.; Bourne, H.; Cohen, H. (1996) Evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257, 342–358.

[22]. Landgraf, R.; Fischer, D.; Eisenberg, D. (1999) Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng*., 12 (11), 943-951.

[23]. Armon, A.; Graur, D.; Ben-Tal, N. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol*., 307 (1), 447-463.

[24]. Landgraf, R.; Xenarios, I.; Eisenberg, D. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol*., 307 (5), 1487-1502.

[25]. Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. (2003) A new bioinformatics approach to detect common 3D sites in protein structures. *Proteins*, 52:137-145.

[26]. Deng, H.; Chen, G.; Yang, W.; Yang, J.J. (2006) Predicting calcium binding sites in proteins - a graph theory and geometry approach. *Proteins*, 64 (1), 34-42.

[27]. Huan, J.; Bandyopadhyay, D.; Wang, W.; Snoeyink, J.; Prins, J.; Tropsha, (2005) A comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *J. Comput. Biol*., 12 (6), 657-671.

[28]. Vacic, V.; Iakoucheva, L.M.; Lonardi, S.; Radivojac, P. (2010) Graphlet kernels for prediction of functional residues in protein structures. *J. Comput. Biol*., 17(1): 55-72.

[29]. Allison, L. (2013) Graphs. http://www.allisons.org/ll/AlgDS/Graph.

[30]. David, J.; Minh, N.; Nathann, C. (2011) Algorithmic Graph Theory.

[31]. Thulasiraman, K. (1992) Graph Theory. http://www.cs.ou.edu/~thulasi/Misc2/graph_theory_chp_7.pdf.

[32]. Weisstein, E. W. (1999) MathWorld. http://www.mathworld.wolfram.com/AdjacencyMatrix.htm.

[33]. James, B. O.; Kamesh, M. K.;Williamson, M. (2010) A faster algorithm for the single source shortest path problem with few distinct positive lengths. *J. of Discrete Algorithms*, 8, (2), 189-198.

[34]. Andrew, F.; Osmar, R. Za¨ıane. (2008) Estimating true and false positive rates in higher dimensional problems and its data mining applications. Proceeding of the IEEE international conference on data mining workshop.

[35]. Schultzkie, L. (1998) Algebra.

http://www.regentsprep.org/Regents/math/ALGEBRA/AD6/quartiles.htm.

[36]. Kevin, L. P.; Paul, E. K. (2005) Artificial neural networks: An introduction. The international society for optical engineering. Belington, Washington, USA.

[37]. Porter, C.T.; Bartlett, G.J.; Thornton, J.M. (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, 32: D129–133.

[38]. Khersonsky, O.; Tawfik, D.S. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev. Biochem.*, 79: 471–505.

[39]. Altschul, S.F.; Madden, T.L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.(1997) Gapped BLAST and PSIBLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25: 3389–3402.

[40]. Borgwardt, K.M.; Schonauer, S.; Vishwanathan, S.V.N.; Smola, A.J.; Kriegel, H.P. (2005) Protein function prediction via graph kernels. *Bioinformatics*, 21, 47.

APPENDIX A. AUTOMATION DOWNLOADING PDB FILE

```
public static void DownloadPDB(string file)
        {
        try{
            string uri = "ftp://ftp.wwpdb.org/pub/pdb/data/structures/all/mmCIF/" + file;
          Uri serverUri = new Uri(uri);
          if (serverUri.Scheme != Uri.UriSchemeFtp)
          {return;}
FtpWebRequestreqFTP;
reqFTP = (FtpWebRequest)FtpWebRequest.Create(new
Uri("ftp://ftp.wwpdb.org/pub/pdb/data/biounit/coordinates/all/" + file));
          reqFTP.KeepAlive = false;
          reqFTP.Method = WebRequestMethods.Ftp.DownloadFile;
          reqFTP.UseBinary = true;
          reqFTP.Proxy = null;
          reqFTP.UsePassive = false;
          FtpWebResponse response = (FtpWebResponse)reqFTP.GetResponse();
          Stream responseStream = response.GetResponseStream();
          FileStreamwriteStream = new FileStream(file, FileMode.Create);
          int Length = 2048;       Byte[] buffer = new Byte[Length];
intbytesRead = responseStream.Read(buffer, 0, Length);
          while (bytesRead> 0)
          { writeStream.Write(buffer, 0, bytesRead);
         bytesRead = responseStream.Read(buffer, 0, Length);  }
          writeStream.Close();
          response.Close();
        }
        catch (WebExceptionwEx)
        {
          StreamWriter err = File.AppendText("Error.txt")  ;
          Console.WriteLine(wEx.Message);
          //      MessageBox.Show(wEx.Message, "Download Error");
          err.WriteLine(file.ToString());
          err.Close();
        }
        catch (Exception ex{
          StreamWriter err = File.AppendText("Error.txt");
          Console.WriteLine(ex.Message);
                err.WriteLine(file.ToString());
                err.Close();
              }

        }
```

APPENDIX B. AUTOMATION DECOMPRESSING PDB FILES

```
public static void Decompress(FileInfo fi)
        {
        // Get the stream of the source file.
        try
        {
          if (fi.Exists)
          {
             using (FileStreaminFile = fi.OpenRead())
             {
                // Get original file extension, for example
                // "doc" from report.doc.gz.
                stringcurFile = fi.FullName;
                string origName = curFile.Remove(curFile.Length -
                       fi.Extension.Length);

                //Create the decompressed file.
                using (FileStreamoutFile = File.Create(origName))
                {

                         using (GZipStream Decompress = new GZipStream(inFile,
                          CompressionMode.Decompress))
                   {
                       // Copy the decompression stream
                       // into the output file.
                       Decompress.CopyTo(outFile);
                       Console.WriteLine("Decompressed: {0}", fi.Name);
                   }
                }
             }
          }
        }
        catch (Exception)
        {
          throw;
        }
```

APPENDIX C. AUTOMATION PSSM FILE GENERATION

```
using System;
using System.Collections.Generic;
using System.ComponentModel;
using System.Data;
using System.Linq;
using System.Text;
using System.Diagnostics;
using System.IO;
using System.Text.RegularExpressions;
using System.Net;
using System.IO.Compression;
namespace PSSM_Calculation1
{
  class Program
{
      static void Main(string[] args)
      {
         StreamReaderFinalSummaryLast = new StreamReader("SeqFile.txt");
         string strFinalSummaryLast = FinalSummaryLast.ReadToEnd();
         FinalSummaryLast.Close();      //    StreamWriterNewFile;
         char[] xx = new char[] { '\n' };
         char[] pipe = new char[] { '|' };
         char[] grether = new char[] { '>' };
         char[] colon = new char[] { ':' };
         char[] comor = new char[] { ',' };
         char[] Minus = new char[] { '-' };
         char[] Semicolon = new char[] { ';' };
         string[] strFinalSummaryLastList = strFinalSummaryLast.ToString().Trim().Split(grether,
                                                StringSplitOptions.RemoveEmptyEntries);
         foreach (var item in strFinalSummaryLastList)
         {
            if (!string.IsNullOrWhiteSpace(item ) )
            {
               string FileNameIN = item.Trim().Substring(0,6) +".txt";
               string FileNameOUT =item.Trim().Substring(0,6) +".pssm";
               StreamWriterNewFile = new StreamWriter(FileNameIN);
               NewFile.WriteLine(">" + item.Trim());
```

```
                NewFile.Close();
                Console.WriteLine("File Created :" + FileNameIN );
                try
                {
                    // Start the child process.
                    Process p = new Process();
                    // Redirect the output stream of the child process.
                    p.StartInfo.UseShellExecute = false;
                    p.StartInfo.RedirectStandardOutput = true;
p.StartInfo.FileName = "C:\\blast-2.2.25+\\bin\\psiblast.exe";
p.StartInfo.Arguments = string.Format("{0}", "-query " + FileNameIN + " -db C:\\blast-
                                2.2.25+\\db\\nr -num_iterations 2 -out_ascii_pssm " + FileNameOUT);
                    p.Start();
                    // Do not wait for the child process to exit before
                    // reading to the end of its redirected stream.
                    // p.WaitForExit();
                    // Read the output stream first and then wait.
                   System.Threading.Thread.Sleep(50000);
                   string output = p.StandardOutput.ReadToEnd();
                    p.WaitForExit();
                  //  return output;
                  Console.WriteLine(output);
                }
                catch (Exception ex)
                {
                    Console.WriteLine(ex.Message.ToString());
                }
            }
        }
        Console.WriteLine("Compltetd");
        Console.ReadLine();
    }
  }
}
```

APPENDIX D. SAMPLE PART OF PSSM FILE

Last position-specific scoring matrix computed, weighted observed percentages rounded down, information per position, and relative weight of gapless real matches to pseudocounts

```
    A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
 1 A   5 -2 -2 -2 -1 -1 -2  1 -2 -2 -3 -1 -2 -3 -2  2 -1 -3 -3 -1  77  0  0  0  0  0  0 10  0  0  0  0  0  0  0 13  0  0  0  0 0.59 1.#J
 2 I   0 -3 -4 -4 -2 -3 -3 -4 -4  4  1 -3  0 -1 -3 -2 -1 -3 -2  4   7  0  0  0  0  0  0  0  0 39  7  0  0  0  0  2  0  0  0 46 0.51 1.#J
 3 K  -2  4 -1 -2  2  1  1 -3  3 -3 -3  3 -2 -1 -2 -1  0 -2  1 -3   0 27  0  0  7  2  7  0 12  0  0 25  0  3  0  2  8  0  8  0 0.35 1.#J
 4 K  -2  3 -1 -2 -4  1  1 -3  1 -1 -2  5  0 -3 -2 -1 -2 -4 -2 -2   0 14  0  0  0  3  5  0  4  6  2 63  3  0  0  0  0  0  0  0 0.60 1.#J
 5 A   2 -2 -1 -2 -2 -1  1 -2 -2  0 -1 -1 -1 -3 -2  2  2 -3 -2  1  31  0  0  0  0  0 10  0  0  6  0  3  0  0  0 15 19  0  0 17 0.23 1.#J
 6 H  -3 -2 -1 -3 -5  1 -2 -4 10 -5 -3 -2 -3 -3  2 -2 -3 -4  0 -5   0  0  0  0  0  5  0  0 83  0  2  0  0  0  9  0  0  0  0  0 1.74 1.#J
 7 I  -1  0 -3 -3 -2 -2 -2 -4 -3  3  1  1  2 -1 -3 -2  0 -3 -2  3   0  5  0  0  0  0  0  0  0 21 12 15  9  0  0  2  3  0  0 34 0.28 1.#J
 8 E  -3 -2 -2  0 -5  2  7 -4 -2 -5 -5 -1 -4 -5 -3 -2 -2 -5 -4 -4   0  0  0  0  0  5 95  0  0  0  0  0  0  0  0  0  0  0  0  0 1.50 1.#J
 9 K   1  0 -1 -2 -3  0 -1 -2 -2 -4 -4  6 -3 -4 -2  1  0 -4 -3 -3  15  0  0  0  0  0  0  0  0  0 70  0  0  0  9  6  0  0  0 0.77 1.#J
10 D  -4 -3  0  8 -5 -2  0 -3 -3 -5 -6 -2 -5 -5 -3 -2 -3 -6 -5 -5   0  0  0 100  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 1.89 1.#J
11 F  -4 -4 -5 -5 -4 -5 -5 -5 -3  2  0 -5 -1  8 -5 -3 -3 -1  1 -2   0  0  0  0  0  0  0  0  0 11  5  0  0 82  0  2  0  0  0  0 1.38 1.#J
12 I  -1 -4 -4 -5  1 -4 -4 -5 -4  6  2 -4  0 -1 -4 -3 -2 -4 -3  3   4  0  0  0  3  0  0  0  0 66 13  0  0  0  0  0  0  0  0 15 0.78 1.#J
13 A   4 -3 -3 -3  6 -2 -2 -2 -3  0  0 -2 -1 -3 -1  1  0 -4 -3  0  46  0  0  0 19  0  0  0  0  6  8  0  0  0  2 12  4  0  0  3 0.48 1.#J
14 F   0 -4 -4 -4 -3 -4 -4 -2 -3 -1  1 -4 -1  7 -4 -1 -3 -1  1 -2   8  0  0  0  0  0  0  4  0  0 16  0  0 67  0  6  0  0  0  0 0.95 1.#J
15 C  -2 -1 -3 -4  8 -2 -3 -4  2 -2 -1  1  3  0 -4 -2 -2 -2  4 -2   0  3  0  0 39  0  0  0  7  0  6 10 13  2  0  0  0  0 22  0 0.87 1.#J
16 S   2 -3 -1 -2 -3 -2 -2 -2 -3 -4 -4 -2 -3 -4 -3  6  0 -5 -3 -3  14  0  0  0  0  0  0  0  0  0  0  0  0  0  0 86  0  0  0  0 1.12 1.#J
17 S  -1 -2 -1 -2  2 -2 -2 -2 -3 -1 -2 -2  0 -3 -3  5  3 -4 -3 -2   0  0  0  0  5  0  0  0  0  5  3  0  2  0  0 59 26  0  0  0 0.68 1.#J
18 T  -2 -3 -2 -3 -3 -2 -3 -2 -3 -2 -2 -2  2 -1 -3  0  6 -4 -3  0   0  0  0  0  0  0  0  2  0  0  0  0  7  3  0  3 78  0  0  7 0.97 1.#J
19 P  -3 -4 -4 -3 -5 -3 -2 -4 -4 -5 -5 -3 -4 -6  8 -2 -1 -6 -5 -4   0  0  0  0  0  0  2  0  0  0  0  0  0  0 93  2  3  0  0  0 2.49 1.#J
20 D  -3 -3  3  5 -5 -2 -1  1  6 -5 -5 -2 -4 -4 -3 -2 -3 -4  0 -5   0  0 13 42  0  0  0 13 28  0  0  0  0  0  0  0  0  4  0 0.95 1.#J
21 N  -3 -2  6  1 -4 -2 -2 -3  2 -3 -4 -2 -3 -3 -3 -1  3 -3  2 -2   0  0 46  8  0  0  0  0  6  0  0  0  0  0  0  0 27  0 11  3 0.76 1.#J
22 V   0 -3 -4 -4 -2 -3 -4 -4 -4  2  0 -1  0  0 -4 -3 -2 -4 -2  5   5  0  0  0  0  0  0  0  0 11  4  5  1  5  0  0  0  0  0 68 0.70 1.#J
23 S   1 -3 -1 -2 -3 -2 -2 -2 -3 -4 -4 -2 -3 -4 -3  6  0 -5 -4 -4   8  0  0  0  0  0  0  0  0  0  0  0  0  0  0 92  0  0  0  0 1.30 1.#J
24 W  -5 -5 -6 -6 -5 -4 -5 -5 -3 -4 -1 -5 -3 -1 -6 -5 -5 12  5 -4   0  0  0  0  0  0  0  0  0  0  5  0  0  0  0  0  0 71 24  0 2.94 1.#J
25 R  -3  8 -2 -3 -5 -1  0 -4 -2 -5 -4  1 -3 -5 -4 -2 -3 -5 -4 -5   0 91  0  0  0  0  5  0  0  0  0  3  0  0  0  2  0  0  0  0 1.68 1.#J
26 H  -3 -2  3  4 -1  0 -1 -3  7 -3 -2 -2 -3 -3 -3 -1 -1 -4 -2 -2   0  0 19 29  2  4  0  0 32  0  5  0  0  0  0  3  2  0  0  4 0.73 1.#J
27 P  -1 -1  1  0 -3  0  0 -1 -1 -1 -3  0 -2 -4  5  1  0 -4 -3  0   0  2  9  3  0  3  5  3  2  5  0  6  0  0 36 16  3  0  0  8 0.58 1.#J
28 T  -1 -1 -1  0 -3  2  1 -3 -2  1 -1  1  1 -3 -2 -1  3 -3 -1  0   2  2  1  6  0 10 10  0  0 12  2 11  6  0  0  0 30  0  2  6 0.21 1.#J
29 M  -2  2  2  0 -4  1  2 -2  0 -3 -2  3  1 -2 -2  0  0 -3  0 -2   0 11 12  3  0  3 13  1  2  0  3 32  5  2  0  4  4  0  4  1 0.29 1.#J
30 G  -2 -4 -2 -3 -5 -4 -4  7 -4 -6 -6 -4 -5 -5 -4 -2 -4 -5 -5 -5   0  0  0  0  0  0  0 100  0  0  0  0  0  0  0  0  0  0  0  0 1.96 1.#J
31 S  -1 -3 -1 -2  1 -2 -2 -2 -3 -4 -4 -2 -3 -4  0  6  1 -5 -4 -3   0  0  0  0  4  0  0  0  0  0  0  0  0  0  4 87  5  0  0  0 1.16 1.#J
32 V  -2 -3 -4 -5 -3 -3 -4 -4  0  2  3 -4  0  3 -4 -3 -2  5  2  2   0  0  0  0  0  0  0  0  3 11 32  0  0 14  0  0  0 11  8 22 0.55 1.#J
33 F  -4 -5 -5 -6 -4 -5 -5 -5 -3 -2  0 -5 -2  8 -6 -4 -4 -1  3 -3   0  0  0  0  0  0  0  0  0  0  6  0  0 89  0  0  0  0  5  0 1.87 1.#J
34 I  -3 -5 -5 -5 -3 -4 -5 -6 -5  7  0 -4 -1 -2 -5 -4 -1 -4 -3  2   0  0  0  0  0  0  0  0  0 86  0  0  0  0  0  0  4  0  0 10 1.26 1.#J
35 G  -1 -1 -1 -1 -3  3  0 -1 -2  1 -2  0  2 -3 -2  2  2 -3 -3 -1   2  0  1  3  0 24  4  3  0 10  0  4  8  0  0 20 20  0  0  1 0.29 1.#J
36 R   0  2 -1 -1 -3  2  1 -2  2 -3 -1  2 -2 -2 -2  1  0 -2  3 -2   6 14  0  3  0 15  8  2  6  0  5 15  0  0  0  9  5  0 13  0 0.25 1.#J
37 L  -3 -4 -5 -5 -3 -4 -4 -5 -4  3  5 -4  1  1 -4 -4 -2 -3 -2  1   0  0  0  0  0  0  0  0  0 18 72  0  0  4  0  0  0  0  0  6 0.86 1.#J
38 I  -2 -4 -4 -5  2 -4 -4 -5 -5  6  0 -4  0 -2 -4 -3  0 -4 -3  4   0  0  0  0  5  0  0  0  0 56  0  0  0  0  0  0  5  0  0 34 0.89 1.#J
39 E  -2 -1  0  4 -4  0  3 -3  1 -3 -3  1 -3 -3 -2  0  1 -4  0 -3   1  2  2 31  0  0 25  0  4  0  2  9  0  0  0  7 12  0  5  0 0.48 1.#J
40 H  -2 -3 -1 -4  7 -1 -3 -4  3  0 -1 -1 -2 -1 -4 -2  0 -2  3  3   0  0  3  0 28  3  0  0  8  5  1  4  0  0  0  0  6  0 14 28 0.55 1.#J
41 M  -4 -4 -5 -5 -4 -4 -5 -5 -3  1  1 -4  4  7 -5 -4 -3 -1  1 -1   0  0  0  0  0  0  0  0  0  8  7  0 14 68  0  0  0  0  0  2 1.26 1.#J
42 Q  -1  3  2 -2  3  5  0 -2 -1 -1 -3  2 -2 -4 -3  0 -2 -4 -3 -3   2 15 13  0  8 38  0  1  0  5  0 13  0  0  0  5  0  0  0  0 0.54 1.#J
43 E  -2  1  1  0 -4  2  3 -3 -1 -3 -3  4 -1 -4 -2  1  1 -4 -3 -3   0  3  6  4  0  8 24  0  0  2  0 35  2  0  0  9  8  0  0  0 0.47 1.#J
44 Y  -3 -3 -3 -2  2 -3 -3 -4  4 -3 -1 -3 -2  1 -4  1 -3  0  7 -3   0  0  0  3  5  0  0  0 13  0  7  0  0  3  0 11  0  0 58  0 1.09 1.#J
45 A   5 -3 -2 -3  2 -2 -2 -2 -3 -3 -3 -2 -3 -4 -2  3 -1 -4 -3 -2  65  0  0  0  4  0  0  0  0  0  0  0  0  0  0 30  0  0  0  0 0.87 1.#J
46 A  -1 -1 -3 -4  6 -3 -3 -4  5 -4 -3 -1 -3  0 -4 -1 -2  9  1 -4   4  4  0  0 18  0  0  0 20  0  0  5  0  3  0  4  3 38  3  0 1.37 1.#J
47 S  -1  0  0 -1  8 -1  0 -3 -3 -3 -3  0 -3 -4 -3  2  0 -4 -3 -1   0  5  5  2 41  2  8  0  0  0  0  6  0  0  0 22  5  0  0  4 0.76 1.#J


 K        Lambda
Standard Ungapped    0.1357     0.3244
Standard Gapped      0.0410     0.2670
PSI Ungapped         0.1357     0.3244
PSI Gapped           0.0410     0.2670
```

APPENDIX E. OUTPUT SAMPLE PART OF RASA FILE

```
REM Relative accessibilities read from external file "standard.data"

REM  File of summed (Sum) and % (per.) accessibilities for

REM RES _ NUM      All-atoms   Total-Side   Main-Chain    Non-polar    All polar
REM              ABS   REL    ABS   REL    ABS   REL    ABS   REL    ABS   REL
RES ARG K  15   244.06 102.2 203.16 101.0  40.90 109.0  79.44 102.1 164.62 102.3
RES ASN K  16   107.79  74.9 107.00 100.7    .79   2.1  24.00  51.9  83.79  85.7
RES TRP K  17   177.23  71.1 176.57  83.6    .65   1.7 152.52  80.4  24.71  41.4
RES VAL K  18    95.17  62.8  95.02  83.1    .16    .4  95.02  82.3    .16    .4
RES PRO K  19    67.16  49.3  66.95  55.8    .20   1.3  67.16  55.5    .00    .0
RES THR K  20    47.56  34.1  46.70  45.9    .86   2.3  15.26  20.2  32.30  50.8
RES ALA K  21    59.57  55.2  50.68  73.0   8.89  23.1  51.39  72.0   8.18  22.4
RES GLN K  22   124.27  69.6 116.81  82.8   7.46  19.9  35.06  67.1  89.21  70.6
RES LEU K  23   118.19  66.2 117.71  83.4    .48   1.3 118.04  82.9    .15    .4
RES TRP K  24   168.62  67.6 159.43  75.5   9.18  24.1 146.46  77.2  22.16  37.1RES GLY
K  25    46.02  57.5  35.30 109.2  10.72  22.4  39.46 105.1   6.56  15.4
RES ALA K  26    51.41  47.6  50.63  72.9    .78   2.0  50.72  71.1    .69   1.9
RES VAL K  27    90.62  59.8  89.61  78.4   1.01   2.7  89.61  77.6   1.01   2.8
RES GLY K  28    44.55  55.6  35.60 110.1   8.95  18.7  40.01 106.6   4.54  10.7
RES ALA K  29    54.37  50.4  52.03  75.0   2.34   6.1  52.03  72.9   2.34   6.4
RES VAL K  30    97.16  64.2  97.13  85.0    .03    .1  97.13  84.1    .03    .1
RES GLY K  31    31.12  38.9  29.23  90.4   1.89   4.0  30.76  81.9    .36    .9
RES LEU K  32   118.13  66.1 114.13  80.9   4.00  10.7 114.13  80.2   4.00  11.0
RES VAL K  33   117.15  77.4 100.88  88.3  16.27  43.8 100.88  87.4  16.27  45.2
RES SER K  34    80.27  68.9  63.94  81.9  16.34  42.5  38.18  78.6  42.09  61.9
RES ALA K  35    88.51  82.0  59.21  85.3  29.30  76.0  61.30  85.9  27.21  74.4
RES THR K  36   126.13  90.6  76.21  74.9  49.92 132.9  87.25 115.2  38.88  61.2
END  Absolute sums over single chains surface
CHAIN  1 K     2155.1       1943.9       211.1       1585.8       569.3
END  Absolute sums over all chains
TOTAL          2155.1       1944.0       211.1       1585.8       569.3
```

APPENDIX F. READ THE RASA FILE

```
class RASA
  {
     string Residue3;
     public string Residue31
     {
        get { return Residue3; }
        set { Residue3 = value; }
     }
     string PDBINDEX;
     public string PDBINDEX1
     {
        get { return PDBINDEX; }
        set { PDBINDEX = value; }
     }
     string ChanID;
     public string ChanID1
     {
        get { return ChanID; }
        set { ChanID = value; }   }
     string Value;
     public string Value1
     {
        get { return Value; }
        set { Value = value; }
     }

     public List<RASA> Read(string PDB, string ChanID)
     {
        List<RASA>ObjlstRasa = new List<RASA>();
        RASA ObjTempRasa;
        StreamReaderFileRasa = new StreamReader(@"RASA\\" + PDB);
        char[] Space = new char[]{' ','\t'};
        string Line = "";
        while ((Line =FileRasa.ReadLine())!= null   )
        {
           if (Line.StartsWith("RES") )
           {
              string[] LineS = Line.Split(Space, StringSplitOptions.RemoveEmptyEntries);
```

```
            if (LineS[2] ==ChanID.Trim()  )
            {
                ObjTempRasa = new RASA();
               ObjTempRasa.Residue31 = LineS[1].Trim().ToString();
               ObjTempRasa.PDBINDEX1= LineS[3].Trim().ToString();
               ObjTempRasa.ChanID1 = LineS[2].Trim().ToString();
               ObjTempRasa.Value1= LineS[6].Trim().ToString();
               ObjlstRasa.Add(ObjTempRasa);
            }

         }
      }
      FileRasa.Close();     return ObjlstRasa;

   }
   public string getValue(string PDBID, string ChainID, List<RASA>lstRasa  )
   {
      string valueR ="0.000";
      foreach (var item in lstRasa )
      {
         if (item.ChanID1 ==ChainID.Trim() && item.PDBINDEX1 ==PDBID.Trim()      )
         {
            valueR = item.Value1.Trim();
            break;
         }
      }
      return valueR;
   }
}
```

APPENDIX G. CALCULATE THE DISTANCE BETWEEN GIVEN TWO ATOMS AND CHECK CONTACT

```
public double CalculateDistance(clsPDB Atom1, clsPDB Atom2)
    {
      double Distance = 0.0;
      try
      {
        doubleDx = Convert.ToDouble(Atom1.X1) - Convert.ToDouble(Atom2.X1);
        doubleDy = Convert.ToDouble(Atom1.Y1) - Convert.ToDouble(Atom2.Y1);
        doubleDz = Convert.ToDouble(Atom1.Z1) - Convert.ToDouble(Atom2.Z1);
        Distance = Math.Sqrt(Dx*Dx+Dy*Dy+Dz*Dz);
      }
      catch (Exception Ex)
      {
        Console.WriteLine(Ex.Message.ToString());
        throw;
      }
      return Distance;
    }
    public Boolean checkContacting(clsPDB Atom1, clsPDB Atom2, double Distance,
                                   clsListVDWObjListVDW)
    {
      Boolean check = false;
      clsVDWobjVDW = new clsVDW();
      double R1 = objVDW.GetRaduis(Atom1.Residue3char1, Atom1.Atom1, ObjListVDW);
      double R2 = objVDW.GetRaduis(Atom2.Residue3char1, Atom2.Atom1, ObjListVDW);
      if (Distance <=(R1+R2+0.5))
      {
        check = true;
      }
      return check;
    }
    public string AddPrepix(Boolean contact, double distance)
    {
      string distance1 = "";
      if (contact) { distance1 = "!" + distance; }
      else { distance1 = "#" + distance; }
      return distance1;
    }
```

APPENDIX H. SAMPLE PART OF DISTANCE OUTPUT FILE

```
2    A _    3    A!   : 1.33441
2    A _    51   A!   : 2.73817
2    A _    50   A!   : 3.46362
2    A _    49   A    : 3.91860
2    A _    4    A    : 4.14432
2    A _    5    A    : 5.86273
2    A _    48   A    : 7.02426
2    A _    32   A    : 7.93030
2    A _    33   A    : 8.68827
2    A _    47   A    : 9.37570
2    A _    6    A    : 9.64237
2    A _    31   A    : 9.86820
2    A _    30   A    : 10.92820
2    A _    46   A    : 12.00734
2    A _    34   A    : 12.85975
2    A _    25   A    : 13.17666
2    A _    7    A    : 13.17978
2    A _    26   A    : 13.48462
2    A _    29   A    : 13.79507
2    A _    9    A    : 14.35912
2    A _    8    A    : 14.44288
2    A _    35   A    : 14.86449
2    A _    45   A    : 15.43867
2    A _    37   A    : 15.73409
2    A _    22   A    : 16.34917
2    A _    10   A    : 16.85862
2    A _    21   A    : 17.01219
2    A _    27   A    : 17.22173
2    A _    28   A    : 17.28770
2    A _    39   A    : 17.62282
2    A _    44   A    : 17.66178
2    A _    12   A    : 17.69294
2    A _    36   A    : 17.82309
2    A _    24   A    : 18.07109
```

APPENDIX I. ACTIVE RESIDUE LIST (PHOSPHORYLATION PROTEIN SEQUENCE)

Protein	Chan	Index ID																		
1D3V	A	230	1OY2	A	29	3PGM	A	194	1USU	A	334	1HUW	A	51	1OY3	C	276	1CIT	A	306
1Q46	A	51	1OY2	A	154	3LZT	A	24	1VC1	A	59	1HUW	A	106	1OY3	C	311	1CIT	A	316
2FMP	A	44	1EO6	A	10	3LZT	A	50	1UU3	A	160	1JFI	B	205	1QWT	A	385	1VDN	A	145
2FMP	A	55	1EO6	A	39	1DN1	A	142	1TUB	A	439	1IRU	C	13	1QWT	A	386	1N3K	A	25
1Z2C	A	26	1W85	A	283	1DN1	A	146	1F7U	A	15	1IRU	C	75	1QWT	A	396	1N3K	A	104
1CMF	A	81	2B5G	A	146	1DN1	A	158	1EQZ	A	1	2B1P	A	167	1QWT	A	398	1N3K	A	116
1CMF	A	101	2B5G	A	149	1DN1	A	345	1EQZ	A	19	1Q33	A	68	1QWT	A	402	1EF1	C	576
1USS	A	174	2NLL	B	338	1QJB	A	58	1UHG	A	68	1G0U	I	24	1QWT	A	405	1BUP	A	153
1S70	A	42	1IRU	G	242	1QJB	A	184	1UHG	A	240	1KBH	B	53	1D2N	A	577	1NH2	D	104
1S70	A	48	1QH4	A	164	1FDJ	A	1035	1UHG	A	344	1GT0	C	107	1MO1	A	46	1HQS	A	104
1LKJ	A	111	1QH4	A	285	1FW8	A	38	2DMC	A	46	1HH4	E	401	2BCG	G	230	1LFO	A	56
1NEG	A	66	1VJV	A	470	1FW8	A	42	2COB	A	16	1HH4	E	415	1U3Y	A	335	1T64	A	39
1Z6Z	A	210	1RZ4	A	216	1FW8	A	58	2COB	A	24	1HH4	E	474	1ONE	A	9	1TUB	B	174
1HD7	A	290	1R55	A	269	1FW8	A	82	2COB	A	28	1QE3	A	189	1ONE	A	103	1XH6	A	139
1EBF	A	237	1KB9	B	141	1PB1	A	113	2COB	A	32	2AQA	A	36	1ONE	A	118	1XH6	A	338
2BKA	A	45	1OK3	A	108	1OM2	A	85	1S9I	A	216	2B6O	A	229	1NVU	S	1043	1TUB	B	392
4NOS	A	234	1HM5	A	184	1OM2	A	88	1S9I	A	222	2B6O	A	231	1QR5	A	46	1TUB	B	430
1PHP	A	183	2CP6	A	22	1U5R	A	181	1AWD	A	9	2B6O	A	235	1H95	A	52	1Q8G	A	3
1Q2O	A	143	2CP6	A	27	2GJ4	A	14	1MAB	A	33	1L8B	A	53	1DKF	B	369	1Q8G	A	24
2FG5	A	35	2CP6	A	31	1IKN	C	335	1KFU	L	369	2NZU	L	12	2AYN	A	147	1Q8G	A	156
1O4X	A	110	1QK9	A	88	1OV3	A	208	1T2M	A	82	2NZU	L	46	2LFB	A	59	2I47	A	382
1BLA	A	73	1QK9	A	90	1B89	A	1494	1GHL	A	24	1MQI	A	150	1LKK	A	158	1BFG	A	64
1NI4	A	203	1G33	A	72	1JEY	A	51	1GHL	A	50	1MQI	A	184	1LKK	A	162	1AXI	A	51
1NI4	A	264	1G33	A	78	1JEY	A	222	1JAS	A	120	1F05	A	237	1LKK	A	194	1AXI	A	106
1NI4	A	266	1AYJ	A	8	2CQN	A	787	1KX5	C	1	1S32	D	1233	1LKK	A	213	1CI4	A	4
1NI4	A	271	2CPE	A	358	1LBQ	A	102	1KX5	C	19	1UDM	A	122	1OQJ	A	129	1EQZ	B	36
1TH8	B	58	1QKL	A	802	1UNL	A	159	1OMW	A	29	1O3X	A	236	1IUY	A	61	1GKY	A	148
1M2V	B	178	1RJV	A	79	1MR3	F	147	1AJS	A	65	2GJS	A	214	1KHU	A	462	1PTF	A	46
1FSO	A	101	1EGA	A	37	1NW9	B	144	1AYZ	A	120	2AXL	A	110	1KHU	A	463	1CJY	A	228
1FSO	A	115	1H9F	A	54	1NW9	B	183	1NR7	A	170	1UCN	A	44	1KHU	A	465	1CJY	A	431
1FSO	A	174	1H9F	A	56	1NW9	B	196	1L0L	H	48	1UCN	A	120	1OCS	A	83	1KHX	A	464
1WEZ	A	37	1H9F	A	57	1CKT	A	34	1K3Z	B	276	1UCN	A	122	1RW5	A	135	1KHX	A	465
1FEX	A	23	1H9D	B	10	1Q79	A	24	2AEB	A	230	1UCN	A	125	1RW5	A	166	1KHX	A	467

1FEX	A	25	1JK0	B	55	1IRU	E	16	1S3S	G	272	1JU5	A	41	1RW5	A	179	1KOY	A	257
1VKH	A	9	1JK0	B	169	1IRU	E	56	1MH1	A	71	2AYU	A	140	1NW3	A	297	2AC0	A	99
1KX5	B	1	1A5E	A	7	1LL2	A	44	1P3M	H	1433	1NF7	A	160	1AAB	A	34	1IU2	A	13
1KX5	B	47	1A5E	A	8	1V6F	A	59	1WNJ	A	118	1DSY	A	226	1D4B	A	17	1IU2	A	82
1KX3	C	19	1A5E	A	140	1V6F	A	78	1HDR	A	222	1SQI	A	250	2AKZ	A	78	1WEL	A	415
1CMZ	A	151	1A5E	A	152	1V6F	A	89	1YQA	A	174	1UJU	A	49	2AKZ	A	79	1WEL	A	420
1VYI	A	210	1BD8	A	66	1DS6	A	71	1SPH	A	12	1UJU	A	52	1L8Y	A	6	1WEL	A	422
1VYI	A	271	1BD8	A	76	1WGF	A	3	2C2H	A	71	1RF8	A	2	1LN6	A	334	1WEL	A	424
1TZY	B	36	1MY7	A	276	2HF3	A	239	1IAT	A	184	1RF8	A	15	1LN6	A	338	1M1C	A	580
1W0J	A	33	1U19	A	334	2HF3	A	323	1GGW	A	2	1RF8	A	28	1LN6	A	343	1N8P	A	39
1OXZ	A	236	1U19	A	338	1UL1	X	187	1GGW	A	6	1RF8	A	30	2F34	A	162	1KX5	D	11
2F73	A	56	1U19	A	343	1U8F	O	83	1NUE	A	44	1J4N	A	249	1E4U	A	71	1KX5	D	33
1PSO	E	68	1AJW	A	101	1A5R	A	2	1NUE	A	120	1KX5	A	10	1IKN	D	283	1X4C	A	94
1OYB	A	352	1AJW	A	115	1A5R	A	9	1HIO	B	36	1KX5	A	28	1IKN	D	288	1X4C	A	96
1HIO	A	19	1AJW	A	174	1RHW	A	88	1UEW	A	107	1E32	A	352	1HGU	A	51	1X4C	A	100
1S50	A	150	1KL9	A	48	1KCM	A	166	1SQN	A	793	1NRG	A	165	1HGU	A	106	1QO5	B	35
1S50	A	163	1HLO	A	10	1UKH	A	129	2CQ4	A	133	1NRG	A	241	1HGU	A	150	1CVU	A	451
1S50	A	168	2NGR	A	71	1G0W	A	163	1PSY	A	38	1KMQ	A	26	1OM4	A	374	1R5S	A	5
1W7B	A	26	1QPG	A	3	1G0W	A	164	1U46	A	149	1C44	A	1	2J4Z	A	369	1R5S	A	12
1DFC	A	1039	1QPG	A	35	1G0W	A	285	1GH6	A	106	1E9G	A	265	1XMI	A	660	1R5S	A	29
1X0F	A	190	1QPG	A	109	1R3S	A	61	1GH6	A	112	1BEH	A	52	1IYR	A	37	1R5S	A	32
1CMI	A	88	1QPG	A	113	1HWX	A	170	1PHK	A	30	1BEH	A	54	2OZA	A	328	1R5S	A	56
1GHC	A	70	1QPG	A	129	1EJI	A	35	1PHK	A	81	1BEH	A	153	1NZP	A	246	1R5S	A	64
2B8A	A	107	1QPG	A	153	1HCN	B	66	1NM1	A	239	1DK2	A	44	1JK7	A	42	1R5S	A	75
1DOA	B	101	1QPG	A	390	1HCN	B	96	1NM1	A	323	1DK2	A	55	1JK7	A	48	1R5S	A	78
1DOA	B	115	1QPG	A	396	1IR3	A	1035	1W8M	A	21	1NEY	A	96	1NKP	A	920	1R5S	A	80
1DOA	B	174	1QPG	A	412	1IR3	A	1037	2FMU	A	45	1NEY	A	100	1KMT	A	101	1R5S	A	118
1A44	A	51	1NO8	A	182	1IG8	A	158	1F1G	A	38	1KA5	A	46	1KMT	A	115	1NXK	A	272
1A44	A	152	2B3Y	A	138	1GPU	A	335	1F1G	A	98	1UD7	A	57	1KMT	A	174	1NXK	A	328
1BE4	A	44	1DF0	A	369	1ID3	B	64	1F1G	A	111	1UD7	A	65	1RHO	A	101	1JDH	A	191
1BE4	A	120	1DDB	A	61	1SJQ	A	98	1F1G	A	116	1T3Y	A	115	1RHO	A	115	1OLZ	A	174
1BE4	A	122	1DDB	A	64	1SJQ	A	99	1AUZ	A	57	1LFB	A	59	1RHO	A	174	1MXE	A	81
1BE4	A	125	1DDB	A	78	1GO4	A	170	1FW4	A	81	1QHF	A	11	1B66	A	18	1MXE	A	101
1S9J	A	212	1GZD	A	184	1GO4	A	178	1FW4	A	101	1QHF	A	115	2CWN	A	237	4PEP	A	68
1S9J	A	218	1M6D	A	94	1GO4	A	195	1Z07	A	124	1QHF	A	126	1T4H	A	378	1OCT	C	107
1XD3	A	75	1OKC	A	41	1JH4	A	53	1PPJ	H	48	1QHF	A	127	1M9M	A	141	1DDJ	A	578
1XD3	A	130	1OKC	A	126	1MX3	A	300	1LSG	A	25	1QHF	A	184	2CUI	A	47	1K8K	A	418
1GZ2	A	61	1CEW	I	80	1TDQ	A	224	1LSG	A	51	1QHF	A	196	2CPT	A	109	2FMM	A	175

1GZ2	A	67	1XLY	A	166	1QSD	A	94	1GJI	A	266	1TZY	A	19	1IAS	A	187	1A0R	P	73
1HDI	A	135	1E3O	C	107	2BTM	A	212	1G0U	D	16	1PCH	A	46	1IAS	A	189	1HM6	A	5
1HDI	A	202	1SM2	A	565	1CWD	L	36	1XD3	B	57	2FXU	A	239	1IAS	A	191	3PSG	A	68
1HDI	A	389	1YAT	A	44	1CWD	L	40	1XD3	B	65	2FXU	A	323	1MJD	A	74	1X7Y	A	250
3PMG	A	116	1HHL	A	24	1CWD	L	72	1RE6	A	88	1G62	A	174	1MJD	A	90	1RDQ	E	14
1H4X	A	58	1HHL	A	50	1CWD	L	91	1NUY	A	1207	1G62	A	175	1MJD	A	110	1RDQ	E	34
1A3W	A	213	3PGM	A	11	1TQ4	A	84	1VJD	A	135	1PO5	A	128	1MJD	A	115	1RDQ	E	139
1F60	A	18	3PGM	A	115	1ZS6	A	61	1VJD	A	202	1G9L	A	99	1G0U	L	161	1RDQ	E	338
1F60	A	163	3PGM	A	126	2TRC	P	73	1VJD	A	389	2F71	A	50	1ZOK	A	232	1CB0	A	183
1XFB	A	39	3PGM	A	127	1TUB	A	6	1TWF	B	919	2F71	A	295	1ZOK	A	301	1XPA	A	196
1HS6	A	415	3PGM	A	182	1TUB	A	48	1PVD	A	223	1X88	A	61	1NG2	A	208	1OPJ	A	465
2BID	A	66	1QJB	A	215	1S50	A	199	1OY2	A	35	1N0W	A	315	2SHP	A	327	1BBZ	A	52
2BID	A	80	1QJB	A	232	1RJV	A	83	1CMZ	A	201	1F2F	A	213	2FXU	A	166	1FTP	A	20
1N54	B	13	1E31	A	34	1NTY	A	1486	1P7H	L	585	1VIG	A	11	2FXU	A	218	1B89	A	1477
1N54	B	18	1E31	A	117	1H9F	A	35	1P7H	L	599	1J1B	A	216	2FXU	A	294	2HUE	C	51
2PIL	A	68	1CM8	A	183	1H9F	A	52	1DN1	A	107	1NW9	B	153	1NI2	A	146	1IRU	2	57
1R0W	A	660	1Z2C	A	19	1JDW	A	417	1DN1	A	346	1CVJ	A	140	2AD9	A	127	1MJD	A	70
1WG5	A	9	1NI2	A	235	1PPJ	A	347	1DN1	A	574	1EGX	A	39	2DNT	A	39	1P15	A	580
1JEB	B	44	1BI9	A	101	1EFC	A	382	2B5I	C	139	1A81	A	131	1KYF	A	807	1K8K	A	231
1JEB	B	80	1BI9	A	104	1VBG	A	456	1UZE	A	111	1VG8	A	1183	2SHP	A	62	1YVH	A	337
1G0U	E	16	1BD7	A	99	1YTQ	A	117	2GFS	A	180	1P4O	A	950	1AWW	A	15	1QKI	A	401
1PIN	A	108	1XH6	A	195	1DCE	B	3	2GFS	A	263	1P4O	A	1131	1PA7	A	124	1QKI	A	507
1DGG	A	422	1XH6	A	197	1K3Z	B	254	1RWY	A	82	1P4O	A	1135	1KN0	A	104	2NNQ	A	19
1YAA	A	388	1XH6	A	201	2CP6	A	9	1ZAI	A	64	1P4O	A	1136	1MLD	A	32	2NNQ	A	128
3ULL	A	63	1K5O	A	4	1OKC	A	125	1ZAI	A	234	1P4O	A	1250	1FMK	A	436	1TAD	A	142
1X79	A	236	1ST6	A	604	1UA2	A	170	1ZAI	A	240	1P4O	A	1251	1FMK	A	527	1F8U	A	133
1S4B	P	643	1G0U	D	55	1J2M	A	17	1CB0	A	188	1X4C	A	84	2FO0	A	134	1PNE	A	128
2BCG	Y	174	1DM5	A	6	1C3D	A	38	1O6L	A	309	1X4C	A	97	2J7Y	A	443	1LLD	A	227
1UB1	A	165	1DE4	C	657	1DDB	A	58	2F34	A	198	2OQ1	A	128	1HM6	A	21	1EZ4	A	238
1UB1	A	167	1ONE	A	220	1JU5	A	42	1UU3	A	245	1QLY	A	8	1HM6	A	207	1YAG	A	53
1OEY	J	315	2BID	A	61	1OB3	A	14	1NKP	A	905	1KX5	B	51	1QDV	A	116	1I2M	A	147
1SMS	A	55	1DF0	A	370	1OB3	A	163	1NKP	A	947	1AIN	A	180	1EJI	A	34	1CMF	A	99
1SMS	A	169	1RDQ	E	195	1L0B	A	1646	1F1G	A	131	1PHR	A	131	1TJX	A	364	1CMF	A	138
2AD9	A	140	1RDQ	E	197	1UKH	A	255	1PQ1	A	115	1PHR	A	132	3GRS	A	21	1J7D	B	76
2AD9	A	141	1RDQ	E	201	1UKH	A	258	2B1P	A	293	1F68	A	734	1LDN	A	238	1AWJ	A	29
1Q1S	C	105	1PPJ	H	50	1G8F	A	126	1K99	A	16	1A5Z	A	237	1VJD	A	195	1U7B	A	249
1J1B	A	215	1PVD	A	353	1YUW	A	477	1G0W	A	282	1HIO	D	51	2UP1	A	167	1GRI	A	160
1J1B	A	219	1ND7	A	822	1QPC	A	501	1G0W	A	289	2OQ1	A	250	1E42	A	737	1GRI	A	209

2B0L	A	215	1R3B	A	34	2B5H	A	59	1DKG	D	199	1RHS	A	164	1L3K	A	167	1CVS	C	154
1H2T	Z	13	1GPU	A	248	1RHS	A	163	2HF3	A	202	1TUB	A	161	1ST6	A	537	1JEB	B	41
1H2T	Z	18	2G50	A	327	1XPC	A	311	2HF3	A	203	1AMM	A	62	1KV3	A	369	1RV3	A	34
1AOI	C	19	1BIF	A	443	1NM1	A	202	1U19	A	335	1AMM	A	65	2AYN	A	135	1FIT	A	145
1JSE	A	24	1MDY	A	115	1NM1	A	203	1U19	A	336	1U5E	A	197	2GFS	A	182	2C4J	A	33
1JSE	A	50	1N0W	A	309	1X0F	A	193	1U19	A	340	1BX4	A	60	2GFS	A	323	1Q8G	A	68
1UA2	A	164	1OPJ	A	413	1QWT	A	404	1G0U	L	163	1OPJ	A	412	1AOA	A	127	1Q8G	A	140
2A1L	A	165	2FXU	A	202	1BXL	A	115	1IRU	E	55	1BLX	A	13	1BUP	A	15	1D5T	A	333
2A1L	A	262	2FXU	A	203	1WMS	A	175	1YDL	A	69	1BLX	A	24	1GZ8	A	15	2FN4	A	66
1RWY	A	72	1QAD	A	10	1DHS	A	95	1IKN	D	291	1TUB	A	272	1GZ8	A	19	1DS6	B	24
1RWY	A	78	1EGA	A	36	1OWX	A	302	2IFQ	A	100	1TUB	A	357	1DN1	A	145	1FVR	A	897
1SIF	A	57	1LIU	A	556	1SVA	1	337	1X4N	A	67	1TUB	A	432	1DN1	A	473	1FVR	A	992
1SIF	A	65	1MXE	A	44	1IAS	A	185	1NXK	A	334	1C0F	A	53	2NGR	A	64	1FVR	A	1048
1AK7	A	3	1MXE	A	79	1IAS	A	186	1NXK	A	338	2CQI	A	50	1EXB	E	120	1FVR	A	1102
1OWX	A	325	1WG5	A	12	1KMQ	A	19	1F60	A	430	1CM8	A	27	1TRN	A	151	1FVR	A	1108
1XJD	A	695	1Q79	A	23	2GGM	A	26	1CTQ	A	35	1CM8	A	185	1IYX	A	281	1FVR	A	1113
1J3X	A	35	1GZ8	A	14	1E5W	A	235	1XFB	A	119	1IR3	A	1158	1FI6	A	15	1QKM	A	488
1RW2	A	19	1GZ8	A	160	1KX3	C	120	1P5F	A	67	1IR3	A	1162	1UKH	A	190	1EJ5	A	50
1RW2	A	21	2C78	A	394	1V6F	A	33	1GD0	A	36	1IR3	A	1163	1UKH	A	259	1O4R	A	72
1RW2	A	22	1L3K	A	138	1BHG	A	274	1ADT	A	195	1V6F	A	90	1UKH	A	357	1LKK	A	192
1ZAI	A	35	1Q8G	A	25	1XOX	A	34	1X5U	A	59	1T46	A	823	1D0N	A	382	1U8F	O	42
1ZAI	A	38	1N0Y	A	44	1XOX	A	117	2B1P	A	297	1QG3	A	1207	1U5F	A	197	1BKL	A	70
1ZAI	A	45	1AIN	A	189	3GPD	G	210	2B1P	A	395	1JR1	A	400	1OKC	A	190	1NI4	A	260
1ZAI	A	353	1SID	A	357	2OZA	A	222	1UMK	A	129	1FIM	A	36	1G7N	A	19	1NI4	A	272
1ZAI	A	355	1LN6	A	335	2OZA	A	226	2COF	A	67	1QLC	A	240	1G7N	A	128	1JDH	A	654
1F16	A	184	1LN6	A	336	2OZA	A	334	1W80	A	807	1IG4	A	52	1UW2	A	8	2F9D	A	86
1RF8	A	22	1LN6	A	340	2OZA	A	338	1IRU	G	160	1BF5	A	701	1UW2	A	12	1Z0F	A	99
1PI1	A	14	1E9G	A	60	1N3K	A	6	1VHR	A	138	1W6T	A	285	2FYM	A	283	1K4T	A	268
1J4N	A	248	1E9G	A	250	1EFV	A	42	1QPC	A	394	1CWD	L	70	3GPD	G	41	1HDI	A	195
1PME	A	185	1W2F	A	311	1L0L	H	50	1SJQ	A	85	1DGG	A	231	2GDG	A	36	1OPK	A	204
1PME	A	190	1GG2	G	52	1UJU	A	42	1JPA	A	808	1DGG	A	386	1FBV	A	371	5PNT	A	131
1XW6	A	33	1R5S	A	76	1UNL	A	14	1XW6	A	32	1LUF	A	754	1IRU	B	23	5PNT	A	132
1BFG	A	112	1M4M	A	34	1HCN	B	97	1L6N	A	132	1N3K	A	108	1IRU	B	97	2CH5	A	205
2C2V	S	277	1EF1	C	558	1J8F	A	63	1BYG	A	416	1QU6	A	106	1IRU	B	120	1TUB	B	36
1ID3	D	39	1A12	A	411	1YFM	A	428	1BYU	A	147	1QU6	A	167	1I10	A	238	1BWY	A	19
1QH4	A	282	2C4J	A	34	1TUB	B	292	1EFX	A	59	2GST	A	32	1VG0	A	302	1MXE	A	138
1QH4	A	289	1SSU	A	50	1TUB	B	376	1I0Z	A	239	1LLC	A	237	1PRX	A	89	1FOE	A	1323
2HXM	A	117	1FOT	A	241	1TUB	B	419	1QQD	A	59	2I47	A	379	1MCX	A	207	1FIL	A	128

2J4Z	A	288	1FMK	A	508	1QDE	A	145	2OZA	A	63	1ZAI	A	203	1W7B	A	24	1FU6	A	48
1EQZ	A	120	1G33	A	82	1KX5	C	120	2G50	A	465	1FHS	A	109	1W7B	A	30	1V18	A	654
1Q1C	A	143	1SW8	A	44	1UP5	A	44	1NA7	A	255	1OEC	A	657	1W7B	A	238	1FGK	A	654
1EBF	A	239	1SW8	A	79	1TWF	A	621	2BID	A	56	1BJ4	A	34	1NM1	A	166	2H6F	B	800
1HM6	A	24	1TF7	A	432	1QPG	A	330	1G83	A	214	1J0X	O	39	1NM1	A	218	1B56	A	22
1ST6	A	324	1MY7	A	254	1QPG	A	391	1GL5	A	205	1XKK	A	998	1NM1	A	294	1B56	A	131
1U1Q	A	138	1EF1	A	235	1U8F	O	211	1CVU	A	460	1XKK	A	1016	1XPC	A	537	1HD7	A	262
1HUU	A	4	1OMW	G	52	1BLA	A	121	1QX4	A	129	2IF1	A	43	1HMS	A	19	1BG1	A	705
1KX5	A	3	1SRS	A	159	1VJV	A	389	1WIB	A	14	2FXU	A	53	1FW4	A	99	1GCQ	A	160
1KX5	A	11	1HCN	A	39	1G0U	C	63	1ID3	B	51	2F8A	A	96	1FW4	A	138	1GCQ	A	209
1OY3	C	254	1KFU	L	370	1Z0F	A	98	1GW5	B	6	1TH3	A	230	1OB3	A	15	1US7	B	298
1PIC	A	11	1YGP	A	10N	1YHW	A	423	1DSX	A	116	1TH3	A	385	1NDH	A	101	1NXK	A	63
1J19	A	235	2UP1	A	138	1FW8	A	259	2F71	A	20	1H6V	A	11	1U1Q	A	167	1XWS	A	218
2B5G	A	10	1M6D	A	82	1FW8	A	320	2F71	A	66	1H6V	A	131	1PME	A	187	1QCF	A	213
1PHP	A	299	3PMG	A	114	1T15	A	1700	2F71	A	152	1IRU	I	111	1PME	A	205	2HF3	A	53
1BPO	A	394	1EGW	A	20	1T15	A	1720	2F71	A	153	1H7S	A	181	2AL6	A	347	1AYA	A	62
1W8M	A	157	1WK0	A	11	1KBL	A	453	2J4Z	A	148	2SHP	A	304	1UNL	A	15	2HF3	A	166
2HF3	A	218	1XFB	A	301	1YZ1	A	15	1VC1	A	29	1VJD	A	1	2F71	A	28	1NG2	A	178
2HF3	A	294	1HS6	A	239	1FW8	A	69	1VC1	A	52	1VJD	A	86	2F71	A	80	1BUP	A	13
1J3D	A	68	1HS6	A	290	1FW8	A	289	1UU3	A	262	1VJD	A	152	2F71	A	104	1BUP	A	16
1J3D	A	75	1OY2	A	65	1U5R	A	38	1EQZ	A	16	1VJD	A	304	2F71	A	190	1BUP	A	40
9LDT	A	237	1OY2	A	130	1U5R	A	174	1UHG	A	103	1TWF	B	50	2F71	A	201	1NH2	D	107
2AKZ	A	43	1OY2	A	149	1U5R	A	260	1UHG	A	269	1TWF	B	182	2F71	A	243	1NH2	D	118
1U46	A	284	1OY2	A	151	2GJ4	A	513	1UHG	A	313	1TWF	B	218	1X88	A	159	1HQS	A	361
2FB7	A	17	1W85	A	88	2GJ4	A	561	1UHG	A	324	1TWF	B	242	1X88	A	179	1QY5	A	187
1QCF	A	416	1W85	A	122	2GJ4	A	812	2DMC	A	11	1TWF	B	493	1X88	A	235	1TUB	B	80
1D3V	A	72	1W85	A	367	1OV3	A	171	2DMC	A	45	1TWF	B	700	1X88	A	240	2B9E	A	228
1D3V	A	137	1IRU	G	86	1PQ1	A	14	2DMC	A	97	1TWF	B	853	1X88	A	314	2B9E	A	357
1D3V	A	199	1OED	E	460	1PQ1	A	145	1S9I	A	248	1TWF	B	869	1OY3	C	240	2B9E	A	396
1D3V	A	302	1VJV	A	178	1JEY	A	78	1S9I	A	372	1TWF	B	1032	1D2N	A	606	2B9E	A	408
1USS	A	177	1VJV	A	230	1JEY	A	96	1MAB	A	56	1TWF	B	1155	1MO1	A	52	2B9E	A	414
1USS	A	244	1VJV	A	403	1JEY	A	144	1MAB	A	63	1PVD	A	196	2BCG	G	72	1BJT	A	691
1S70	A	177	1RZ4	A	175	1LBQ	A	174	1KFU	L	503	1PVD	A	268	2BCG	G	137	1BJT	A	948
1LKJ	A	29	1RZ4	A	212	1LBQ	A	187	1T2M	A	20	1PVD	A	284	2BCG	G	164	1BJT	A	968
1LKJ	A	129	1R55	A	386	1LBQ	A	192	1T2M	A	59	1HUW	A	100	1DKF	B	229	1BJT	A	1024
1Z6Z	A	114	1KB9	B	188	1UNL	A	46	1GHL	A	72	1HUW	A	188	2AYN	A	124	1TUB	B	280
1Z6Z	A	133	1OK3	A	74	1UNL	A	93	1JAS	A	29	1IRU	C	7	2AYN	A	240	1Q8G	A	108
1Z6Z	A	142	1OK3	A	120	1UNL	A	105	1JAS	A	97	1IRU	C	81	2AYN	A	455	1Q8G	A	120

1HD7	A	201	1HM5	A	91	1UNL	A	229	1KX5	C	127	2B1P	A	182	1LKK	A	133	1V04	A	66
1HD7	A	275	1HM5	A	209	1MR3	F	6	1OMW	A	41	1Q33	A	323	1KHU	A	290	1BFG	A	143
1HD7	A	307	1HM5	A	358	1MR3	F	135	1OMW	A	121	1G0U	I	84	1OCS	A	97	1AXI	A	7
1EBF	A	174	1HM5	A	532	1NW9	B	334	1AJS	A	332	1G0U	I	86	1RW5	A	26	1UL3	A	31
4NOS	A	486	2CP6	A	142	1IRU	E	43	1AYZ	A	104	1G0U	I	105A	1RW5	A	163	1EQZ	B	123
1PHP	A	346	2CP6	A	154	1IRU	E	172	1AYZ	A	108	1G0U	I	140	1NW3	A	11	1GKY	A	35
1Q2O	A	138	1QK9	A	4	1LL2	A	69	1NR7	A	62	1HH4	E	324	1AAB	A	45	1CJY	A	206
2FG5	A	69	1QK9	A	40	1LL2	A	173	2AEB	A	5	1HH4	E	362	1D4B	A	21	1CJY	A	278
2FG5	A	148	1QK9	A	58	1V6F	A	8	1S3S	G	262	1QE3	A	291	1D4B	A	113	1CJY	A	285
1BLA	A	137	1G33	A	71	1V6F	A	97	1S3S	G	285	1QE3	A	298	1L8Y	A	79	1CJY	A	573
1NI4	A	101	2CPE	A	443	1WGF	A	2	1P3M	H	1457	1QE3	A	372	1LN6	A	176	1CJY	A	583
1M2V	B	622	2CPE	A	453	1WGF	A	26	1HDR	A	23	1AUI	A	126	2F34	A	122	1KHX	A	296
1WEZ	A	8	1BG1	A	599	2HF3	A	265	1HDR	A	96	1AUI	A	171	2F34	A	152	1KHX	A	306
1WEZ	A	90	1BG1	A	629	1UL1	X	16	1HDR	A	110	1AUI	A	233	2F34	A	163	1KHX	A	359
1WEZ	A	101	1BG1	A	631	1UL1	X	62	1HDR	A	117	1AUI	A	337	2F34	A	193	1KHX	A	458
1VKH	A	232	1EGA	A	155	1UL1	X	157	1HDR	A	192	2B6O	A	63	2F34	A	213	2AC0	A	127
1K8K	C	64	1H9D	B	65	1A5R	A	61	1HDR	A	223	2B6O	A	188	1IKN	D	174	2AC0	A	241
1K8K	C	200	1JK0	B	196	1A5R	A	99	1YQA	A	177	1L8B	A	141	1IKN	D	262	1M1C	A	50
1K8K	C	221	1BD8	A	13	1OED	C	275	1YQA	A	215	1MQI	A	108	1HGU	A	85	1M1C	A	95
1K8K	C	330	1BD8	A	130	1UKH	A	34	1YQA	A	217	1MQI	A	142	1OM4	A	457	1M1C	A	113
1CMZ	A	86	1U19	A	98	1R3S	A	219	1SPH	A	66	1F05	A	47	1OM4	A	585	1M1C	A	327
1CMZ	A	148	1U19	A	127	1HWX	A	204	2C2H	A	89	1UDM	A	29	1OM4	A	622	1N8P	A	24
1CMZ	A	156	1HLO	A	42	1EJI	A	23	2C2H	A	158	1O3X	A	245	1OM4	A	684	1QO5	B	16
1ERJ	A	340	1QPG	A	1	1EJI	A	58	1IAT	A	454	1BXL	A	4	2J4Z	A	245	1QO5	B	21
1ERJ	A	490	1NO8	A	110	1EJI	A	74	1IAT	A	531	1BXL	A	145	2J4Z	A	342	1QO5	B	131
1ERJ	A	581	1NO8	A	141	1EJI	A	206	1NUE	A	70	1BXL	A	154	1XMI	A	478	1QO5	B	275
1ERJ	A	593	2B3Y	A	127	1EJI	A	233	1NUE	A	131	2GJS	A	210	1XMI	A	531	1CVU	A	471
1ERJ	A	647	2B3Y	A	391	1EJI	A	339	1NUE	A	144	1DSY	A	264	1XMI	A	573	1NXK	A	265
2F73	A	4	1DHS	A	28	1EJI	A	477	1HIO	B	91	1SQI	A	306	1XMI	A	589	1OLZ	A	58
1PSO	E	35	1DDB	A	6	1HCN	B	81	1SQN	A	711	1UJU	A	3	1XMI	A	605	1OLZ	A	121
1PSO	E	47	1DDB	A	28	1IR3	A	982	1SQN	A	767	1RF8	A	59	1IYR	A	8	1OLZ	A	163
1PSO	E	110	1DDB	A	117	1IR3	A	1006	1SQN	A	792	1KX5	A	86	1IYR	A	85	4PEP	A	47
1PSO	E	147	1DDB	A	184	1IR3	A	1086	1SQN	A	796	1KX5	A	87	2OZA	A	169	4PEP	A	147
1PSO	E	250	1GZD	A	106	1IR3	A	1090	1SQN	A	846	1E32	A	276	2OZA	A	243	4PEP	A	185
1PSO	E	284	1GZD	A	137	1IR3	A	1189	1SQN	A	847	1NRG	A	205	2OZA	A	265	4PEP	A	250
1OYB	A	76	1GZD	A	532	1IR3	A	1190	1SQN	A	866	1C44	A	65	1NZP	A	245	1DDJ	A	594
1OYB	A	252	1M6D	A	29	1GPU	A	66	2CQ4	A	183	1E9G	A	206	1JK7	A	268	1DDJ	A	653
1S50	A	184	1M6D	A	110	1GPU	A	237	1PSY	A	6	1E9G	A	225	1KMT	A	191	1K8K	A	297

1W7B	A	85	1M6D	A	114	1GPU	A	277	1PSY	A	39	1E9G	A	235	1B66	A	104	1TAZ	A	206
1W7B	A	89	1M6D	A	193	1ID3	B	60	1PSY	A	42	1BEH	A	13	2CWN	A	13	1TAZ	A	290
1W7B	A	134	1OKC	A	21	1GO4	A	16	1PSY	A	113	1BEH	A	185	2CWN	A	47	1TAZ	A	372
1W7B	A	164	1CEW	I	75	1GO4	A	150	1U46	A	245	1NEY	A	16	2CWN	A	112	1TAZ	A	382
1W7B	A	277	1XLY	A	91	1JH4	A	61	1U46	A	337	1NEY	A	202	2CWN	A	187	2FMM	A	128
1ZWW	A	193	1E3O	C	150	1TDQ	A	197	1PHK	A	122	1KA5	A	31	2CWN	A	216	2FMM	A	162
2D4C	A	130	1SM2	A	543	1TDQ	A	203	1NM1	A	338	1KA5	A	52	2CWN	A	275	1A0R	P	119
2D4C	A	220	1SM2	A	553	1TDQ	A	227	1NM1	A	348	1KA5	A	71	1T4H	A	402	1A0R	P	219
1GHC	A	74	1YAT	A	-5	1TDQ	A	266	1W8M	A	51	1KA5	A	82	1M9M	A	102	1HM6	A	46
1DOA	B	191	1HHL	A	72	1QSD	A	98	1F1G	A	23	1T3Y	A	127	1M9M	A	125	1HM6	A	229
1A44	A	5	1HHL	A	100	1RYH	A	2A	1AUZ	A	36	1LFB	A	57	2CUI	A	6	1HM6	A	244
1A44	A	103	3PGM	A	28	1RYH	A	177	1AUZ	A	58	1QHF	A	55	2CUI	A	8	1HM6	A	273
1BE4	A	144	3LZT	A	60	1CWD	L	4	1AUZ	A	83	1PCH	A	20	2CUI	A	16	1MVC	A	336
1XD3	A	92	3LZT	A	72	1TQ4	A	59	1LSG	A	82	2FXU	A	155	2CUI	A	38	1MVC	A	380
1GZ2	A	27	3LZT	A	100	1TQ4	A	80	1LSG	A	86	2FXU	A	271	2CUI	A	71	3PSG	A	46
1GZ2	A	91	1DN1	A	37	1TQ4	A	231	1LSG	A	101	2FXU	A	281	2CPT	A	10	3PSG	A	62
1GZ2	A	100	1DN1	A	149	1TQ4	A	269	1GJI	A	271	2FXU	A	338	2CPT	A	64	3PSG	A	79
1GZ2	A	124	1DN1	A	533	1TQ4	A	382	1G0U	D	169	1G62	A	130	1IAS	A	235	3PSG	A	226
1HDI	A	319	1QJB	A	114	2TRC	P	45	1G0U	D	180E	1G62	A	166	1IAS	A	241	3PSG	A	281
1A3W	A	287	1QJB	A	190	2TRC	P	106	1XD3	B	20	1PO5	A	141	1IAS	A	308	1X7Y	A	47
1F60	A	107	1FDJ	A	1068	1TUB	A	136	1NUY	A	1088	1PO5	A	207	1IAS	A	437	1RDQ	E	159
1F60	A	128	1FDJ	A	1131	1TUB	A	165	1NUY	A	1124	1PO5	A	214	1MJD	A	147	1CB0	A	165
1XFB	A	4	1YZ1	A	9	1USU	A	478	1NUY	A	1237	1PO5	A	483	1ZOK	A	251	1CB0	A	260
1XPA	A	173	1BJT	A	444	3PMG	A	184	1TWF	A	1080	1CVU	A	495	2IF1	A	67	1W7B	A	199
1OPJ	A	248	1BJT	A	513	1KQO	A	21	1TWF	A	1141	1CVU	A	504	2FXU	A	91	1OB3	A	4
1OPJ	A	367	1BJT	A	572	1KQO	A	74	1QPG	A	248	1QX4	A	79	1TH3	A	83	1NDH	A	65
1OPJ	A	368	1BJT	A	835	1RJV	A	105	1QPG	A	365	1QX4	A	93	1TH3	A	235	1U1Q	A	124
1OPJ	A	504	1BJT	A	839	1NTY	A	1440	1U8F	O	229	1GW5	B	121	1TH3	A	369	1PME	A	233
2BID	A	2	1BJT	A	873	1JDW	A	140	1VJV	A	239	1GW5	B	136	1TH3	A	404	1PME	A	317
2BID	A	34	1BJT	A	901	1JDW	A	307	1VJV	A	244	1GW5	B	277	1TH3	A	499	1UNL	A	179
2BID	A	163	1CM8	A	49	1JDW	A	336	1VJV	A	346	2J4Z	A	246	1H6V	A	116	1UNL	A	236
2BID	A	190	1CM8	A	94	1PPJ	A	350	1VJV	A	396	2J4Z	A	320	1H6V	A	127	1UNL	A	285
1R0W	A	635	1CM8	A	103	1PPJ	A	373	1G0U	C	124	1F2F	A	202	1H6V	A	330	1FTP	A	130
1R0W	A	641	1CM8	A	224	1EFC	A	297	1G0U	C	213	1J1B	A	127	1OED	B	220	1B89	A	1211
1WG5	A	103	1CM8	A	301	1EFC	A	334	1Z0F	A	77	1J1B	A	234	1H7S	A	191	1B89	A	1365
1JEB	B	49	1Z2C	A	127	1YTQ	A	149	1Z0F	A	107	3ULL	A	83	2SHP	A	380	1B89	A	1404
1G0U	E	35	1BI9	A	244	1DCE	B	98	1FW8	A	94	1EGX	A	16	2SHP	A	511	1B89	A	1451
1G0U	E	42	1XH6	A	300	1DCE	B	291	1FW8	A	169	1EGX	A	72	2FXU	A	240	2HUE	C	88

1G0U	E	144	1ST6	A	592	2CP6	A	113	1OY2	A	25	1A81	A	28	2FXU	A	279	1IRU	2	83
1G0U	E	211	1ST6	A	614	1OKC	A	23	1P7H	L	612	1VG8	A	1037	2FXU	A	306	1MJD	A	138
1PIN	A	71	1G0U	D	156	1UA2	A	96	1DN1	A	218	1VG8	A	1144	1NI2	A	137	1P15	A	647
1PIN	A	115	1DM5	A	56	1UA2	A	121	1UZE	A	54	1P4O	A	1162	1KN0	A	36	1P15	A	777
1PIN	A	138	1DE4	C	666	1UA2	A	228	1UZE	A	302	1QLY	A	10	1KN0	A	65	1K8K	A	316
1DGG	A	187	1DE4	C	708	1C3D	A	83	1UZE	A	544	1AIN	A	213	1KN0	A	81	1K8K	A	400
1YAA	A	64	1ONE	A	274	1DDB	A	173	2GFS	A	44	1AIN	A	216	1KN0	A	126	1YVH	A	307
1YAA	A	78	1S9J	A	238	1OB3	A	47	2GFS	A	175	1F68	A	787	1KN0	A	171	1QKI	A	249
1YAA	A	81	1DF0	A	428	1OB3	A	229	1RWY	A	3	1A5Z	A	248	1FMK	A	357	1QKI	A	308
1YAA	A	92	1RDQ	E	88	1L0B	A	1604	1RWY	A	104	1A5Z	A	274	1FMK	A	376	1QKI	A	424
1YAA	A	255	1RDQ	E	299	1GML	A	345	1CB0	A	118	1A5Z	A	278	1FMK	A	479	1QKI	A	484
1YAA	A	327	1ND7	A	575	1UKH	A	65	1GZK	A	148	2OQ1	A	200	1HM6	A	39	1QKI	A	503
3ULL	A	42	1ND7	A	889	1UKH	A	228	1GZK	A	313	2OQ1	A	223	1HM6	A	243	1TAD	A	91
1S4B	P	388	1GPU	A	145	1G8F	A	116	1GZK	A	436	1TUB	A	210	1QDV	A	68	1TAD	A	150
1S4B	P	440	1GPU	A	151	2B5H	A	180	1O6L	A	148	1AMM	A	6	1EJI	A	175	1F8U	A	77
1S4B	P	492	1GPU	A	247	1RHS	A	190	1O6L	A	213	1U5E	A	150	3GRS	A	23	1F8U	A	98
1S4B	P	542	1X7Y	A	228	1RHS	A	246	2F34	A	144	1OPJ	A	272	3GRS	A	114	1EZ4	A	201
2BCG	Y	17	1X7Y	A	285	1RHS	A	286	2F34	A	205	1OPJ	A	454	3GRS	A	147	1EZ4	A	280
1UB1	A	87	1N0W	A	225	1XPC	A	431	1UU3	A	180	1OPJ	A	459	3GRS	A	327	1EZ4	A	286
1UB1	A	117	1N0W	A	298	1XPC	A	465	1PQ1	A	109	1OPJ	A	468	1OED	E	286	1YAG	A	69
1UB1	A	182	1OPJ	A	325	1DHS	A	131	1PQ1	A	190	1OPJ	A	475	1L3K	A	124	1YAG	A	91
1OEY	J	263	1OPJ	A	338	1QY5	A	121	1K99	A	42	1OPJ	A	488	1KV3	A	159	1MVC	A	403
1SMS	A	92	1OPJ	A	425	1QY5	A	219	1G0W	A	262	1TUB	A	319	1KV3	A	274	1I2M	A	79
1SMS	A	128	2FXU	A	160	1TZD	A	274	1DKG	D	221	1C0F	A	133	2AYN	A	150	1I2M	A	155
1SMS	A	132	2FXU	A	277	1TZD	A	347	2HF3	A	297	1CM8	A	59	2AYN	A	194	1GP1	A	175
1SMS	A	196	2FXU	A	278	1TZD	A	354	2HF3	A	324	1CM8	A	135	2AYN	A	417	1AWJ	A	47
1Q1S	C	77	2FXU	A	351	1TZD	A	419	1U19	A	160	1CM8	A	143	2AYN	A	435	1U7B	A	239
1Q1S	C	92	1QAD	A	64	1IGN	A	588	1U19	A	193	1V6F	A	110	2GFS	A	307	1RV3	A	82
1Q1S	C	179	1EGA	A	23	1SVA	1	183	1U19	A	243	1T46	A	870	1BUP	A	107	1RV3	A	180
1Q1S	C	194	1MXE	A	28	1SVA	1	191	1IRU	E	14	1T46	A	900	1BUP	A	115	1RV3	A	183
1J1B	A	78	1WG5	A	57	1SVA	1	218	1IKN	D	185	1QG3	A	1199	1BUP	A	134	1RV3	A	205
1J1B	A	368	1Q79	A	34	1IAS	A	491	1IKN	D	273	1JR1	A	32	1GZ8	A	179	1Q8G	A	82
2B0L	A	232	1Q79	A	177	1KMQ	A	37	1NXK	A	195	1JR1	A	110	1GZ8	A	269	1D5T	A	339
1JSE	A	60	1GZ8	A	158	1KMQ	A	175	1NXK	A	317	1JR1	A	233	1DN1	A	157	1D5T	A	387
2C4K	A	264	1SID	A	67	2GGM	A	138	1CTQ	A	2	1FIM	A	98	1DN1	A	191	1DS6	B	146
2C4K	A	301	1SID	A	155	1XOX	A	48	1XFB	A	97	1X7Y	A	34	1DN1	A	212	1FVR	A	860
1UA2	A	70	1SID	A	192	2OZA	A	86	1XFB	A	269	1X7Y	A	90	1DN1	A	264	1FVR	A	954
2A1L	A	14	1SID	A	248	2OZA	A	206	1XFB	A	299	1X7Y	A	113	1DN1	A	554	1FVR	A	976

2A1L	A	80	1LN6	A	108	2OZA	A	315	2B1P	A	49	1X7Y	A	134	2NGR	A	40	1QKM	A	397
1RWY	A	55	1E9G	A	72	1N3K	A	100	2B1P	A	82	1X7Y	A	224	2NGR	A	51	1U8F	O	45
1SIF	A	20	1E9G	A	222	1EFV	A	93	1IRU	G	125	1X7Y	A	246	2NGR	A	72	1U8F	O	49
1J3X	A	46	1JPA	A	727	1EFV	A	171	1IRU	G	158	1BF5	A	634	1EXB	E	72	1U8F	O	320
1J3X	A	58	1JPA	A	887	1UNL	A	77	1IRU	G	198	1W6T	A	256	1EXB	E	95	1BKL	A	59
1RW2	A	137	1EF1	C	526	1UNL	A	181	1VHR	A	23	1W6T	A	423	1TRN	A	59	1NI4	A	89
1ZAI	A	300	1A12	A	188	1HCN	B	28	1QPC	A	263	1DGG	A	260	1TRN	A	94	1NI4	A	127
1F16	A	60	1A12	A	195	1J8F	A	89	1QPC	A	489	1DGG	A	280	1TRN	A	232	1JDH	A	604
1F16	A	163	1A12	A	274	1J8F	A	218	1SJQ	A	72	1DGG	A	447	1IYX	A	248	2F9D	A	42
1RF8	A	133	1A12	A	293	1YFM	A	122	1JPA	A	851	1DGG	A	500	1FI6	A	84	1Z0F	A	136
1PI1	A	99	1A12	A	373	1YFM	A	218	1XW6	A	6	1LUF	A	656	1UKH	A	11	1K4T	A	241
1J4N	A	158	1SSU	A	10	1YFM	A	359	1XW6	A	40	1LUF	A	659	1UKH	A	71	1OPK	A	191
1BFG	A	121	1FOT	A	92	1YFM	A	402	1BYG	A	268	2GST	A	22	1OKC	A	131	2CH5	A	157
2C2V	S	247	1FOT	A	132	1YFM	A	435	1BYU	A	79	2I47	A	236	1UW2	A	37	2CH5	A	191
1QH4	A	322	1FOT	A	318	1YFM	A	459	1BYU	A	98	2I47	A	250	2FYM	A	302	2CH5	A	233
2J4Z	A	384	1FMK	A	440	1TUB	B	314	1EFX	A	85	2I47	A	298	2FYM	A	421	2CH5	A	341
1EQZ	A	59	1FMK	A	453	1QDE	A	160	1EFX	A	118	2I47	A	352	3GPD	G	93	1FOE	A	1304
1CJY	A	310	1FMK	A	457	1UP5	A	5	1EFX	A	123	1ZAI	A	213	3GPD	G	139	1FIL	A	59
1CJY	A	321	1SW8	A	28	1UP5	A	26	1QQD	A	67	1ZAI	A	222	1IRU	B	154	1FGK	A	677
1CJY	A	680	1TF7	A	409	1TWF	A	40	1NA7	A	138	1FHS	A	67	1I10	A	246	1FGK	A	701
1EBF	A	176	1TF7	A	426	1TWF	A	170	1NA7	A	146	1OEC	A	608	1VG0	A	62	2H6F	B	682
1EBF	A	225	1MY7	A	292	1TWF	A	475	1NA7	A	196	1OEC	A	616	1VG0	A	309	2H6F	B	751
1ST6	A	327	1SRS	A	177	1TWF	A	527	1NA7	A	261	1BJ4	A	211	1VG0	A	395	2H6F	B	914
1HUU	A	13	1SRS	A	181	1TWF	A	634	1NA7	A	307	1BJ4	A	269	1VG0	A	570	1HD7	A	45
1KX5	A	6	1YGP	A	37	1TWF	A	682	1G83	A	150	1BJ4	A	286	1MCX	A	84	1HD7	A	184
1KX5	A	58	1YGP	A	343	1TWF	A	703	1GL5	A	232	1XKK	A	764	1MCX	A	156	1HD7	A	264
1KX5	A	80	1YGP	A	369	1TWF	A	831	1GL5	A	244	1XKK	A	801	1MCX	A	230	1US7	B	248
1E31	A	21	3PMG	A	8	1TWF	A	885	1CVU	A	147	1XKK	A	813	1MCX	A	243	1US7	B	331
1E31	A	97	3PMG	A	152	1TWF	A	976	1CVU	A	262	1XKK	A	915	1W7B	A	75	1NXK	A	128
1NXK	A	194	1QCF	A	202	2HF3	A	337	2AKZ	A	130	2FB7	A	32	1QCF	A	479			
1XWS	A	38	1AYA	A	80	2AKZ	A	24	1U46	A	232	2FB7	A	37	1QCF	A	492			
1XWS	A	215	2HF3	A	169	2AKZ	A	56	1U46	A	361	2FB7	A	74	1QCF	A	527			

APPENDIX J. ACTIVE RESIDUE LIST (CSA PROTEIN SEQUENCE)

Protein	Index	Protein	Index	Protein	Index	Protein	Index	Protein	Index	Protein	Index	Protein	Index
1mpp_A	32	1pfq_A	708	1o8a_A	353	1lml_A	265	1pfq_A	630	2fqq_A	285	2bhg_A	182
1mpp_A	35	1pfq_A	740	1o8a_A	354	1eb6_A	129	1kfu_L	262	2fqq_A	286	1kfu_L	99
1mpp_A	75	1cbx_A	127	1o8a_A	384	1i1e_A	267	1kfu_L	286	1cvr_A	152	1kfu_L	105
1mpp_A	215	1cbx_A	270	1o8a_A	513	1i1e_A	369	1lya_B	231	1cvr_A	211	2bkr_A	163
1tyf_A	68	1ybq_A	285	1o8a_A	523	1i1e_A	372	1lya_A	33	1cvr_A	212	1cqq_A	40
1tyf_A	97	2lpr_A	57	1ysc_A	146	2fqq_B	390	1ge7_A	118	1cvr_A	244	1cqq_A	71
1tyf_A	98	2lpr_A	102	1ysc_A	397	1cg2_A	175	1ge7_A	133	1nln_A	54	1xgm_A	187
1tyf_A	122	2lpr_A	193	1bcr_A	53	1cg2_A	176	1qib_A	202	1nln_A	71	1itq_A	152
1tyf_A	171	1sca_A	32	1bcr_A	146	1cg2_A	200	1tlp_E	143	1nln_A	115	1itq_A	288
1rtf_B	57	1sca_A	64	1bcr_A	147	1cg2_A	385	1tlp_E	231	1nln_A	122	1cqq_A	147
1rtf_B	102	1sca_A	221	1bcr_B	397	1aug_A	91	1pwv_A	728	1gcb_A	67	2fqq_A	237
1rtf_B	193	1i78_A	83	1lbu_A	192	1aug_A	144	1lam_A	255	1gcb_A	73	2fqq_A	238
1rtf_B	195	1i78_A	85	1qfm_A	554	1aug_A	168	1lam_A	262	1gcb_A	369	1fo6_A	127
1hr6_A	65	1i78_A	99	1qfm_A	641	1l9x_A	110	1lam_A	336	1gcb_A	392	1fo6_A	146
1hr6_B	73	1i78_A	210	1qfm_A	680	1l9x_A	220	1xqw_A	37	1s2k_A	53	1fo6_A	172
1b65_A	146	1i78_A	212	1hzf_A	991	1ca0_B	57	1xqw_A	105	1s2k_A	136	2bhg_A	46
1b65_A	218	1jhf_A	118	1hzf_A	994	1ca0_B	102	1xqw_A	106	1r1j_A	584	2bhg_A	84
1b65_A	250	1jhf_A	119	1nlu_A	80	1ca0_C	196	1xqw_A	244	1r1j_A	650	2bhg_A	163
1b65_A	288	1jhf_A	127	1nlu_A	84	1x9y_A	237	1xqw_A	271	1r1j_A	711	2bkr_A	102
1b65_A	289	1jhf_A	152	1nlu_A	170	1x9y_A	243	1a16_A	361	1r1j_A	717	2bkr_A	103
1r44_A	71	1jhf_A	156	1nlu_A	287	1x9y_A	340	1a16_A	383	1i1i_P	503	2bkr_A	119
1r44_A	181	1t7d_A	88	1qtn_A	258	1x9y_A	360	1amp_A	151	1i1i_P	613	1azw_A	110
1iec_A	63	1t7d_A	90	1qtn_A	317	8pch_A	19	1ei5_A	62	1slm_A	202	1azw_A	266
1iec_A	134	1t7d_A	145	1qtn_A	350	8pch_A	25	1ei5_A	153	1ast_A	93	1azw_A	294
1iec_A	157	1rgq_A	60	1qtn_A	360	8pch_A	159	1ei5_A	155	1ast_A	149	1fy2_A	157
1iec_A	165	1rgq_A	84	1qx3_A	121	1cmx_A	84	1ei5_A	287	1ck7_A	404	1fy2_A	192
1iec_A	166	1rgq_A	140	1qx3_A	122	1cmx_A	90	1fy2_A	88	1fo6_A	197	1cqq_A	145
1fo6_A	47	1rgq_A	142	1qx3_A	163	1cmx_A	166	1fy2_A	120	1cg2_A	112		
1fo6_A	110	1pxv_A	360	2bkr_A	26	1cmx_A	181	1fy2_A	121	1cg2_A	141		