8-21-2013

Recurring side-chain triads in monomeric enzymatic structures

Mikhail Osipovich

Follow this and additional works at: http://scholarworks.rit.edu/theses

Recommended Citation

This Thesis is brought to you for free and open access by the Thesis/Dissertation Collections at RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritscholarworks@rit.edu.
Recurring Side-Chain Triads in Monomeric Enzymatic Structures

by

Mikhail Osipovitch

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Bioinformatics.

College of Science
Department of Biological Sciences
School of Life Sciences

Rochester Institute of Technology
Rochester, New York
August 21, 2013
Committee Approval:

Paul A. Craig, Ph.D.

__________________________________________________________________________

(date)

Professor and Head / School of Chemistry and Materials Science
Thesis Project Advisor

Michael V. Osier, Ph.D.

__________________________________________________________________________

(date)

Associate Professor and Bioinformatics Program Director / School of Life Sciences
Committee Member

Gary R. Skuse, Ph.D.

__________________________________________________________________________

(date)

Professor and Associate Head / School of Life Sciences
Committee Member
Abstract

An algorithm was developed to exhaustively screen monomeric enzymatic structures for recurring 3-residue side-chain arrangements. The algorithm was used in the screening of two datasets: 100 enzymatic structures with a recurring 3-residue active site, and 100 enzymatic structures with a unique 3-residue active site. In each structure, the algorithm considered all distinct side-chain triads that can be compiled from the entire complement of residues in a single isolated chain. Increasing chain length demonstrated a logarithmic growth in the number of recurring triads. The distribution of total distances in recurring triads adhered to normality while the distribution of unique triad distances appeared negatively skewed. Analysis of variance indicated that the means of maximum and average total distances are significantly greater in unique triads than in recurring triads. Screening for recurring triads in synthetically generated alternative rotamer structures demonstrated an overall decrease in the percent of recurring triads, as compared to the natural structures.
# Table of Contents

<table>
<thead>
<tr>
<th>1</th>
<th>Introduction</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Materials and Methods</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>2.1 Software and Computational Resources</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>2.2 Structure Pool Compilation and Preliminary Screening</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>2.3 Testing Dataset Compilation</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2.4 Exhaustive Screening for Recurring Side-Chain Triads</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2.4.1 Stage 1: Pre-Processing of PDB Structure Files</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>2.4.2 Stage 2: Data Reduction</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2.4.3 Stage 3: Searching for Recurring Triads</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2.4.4 Stage 4: Results Output</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>2.5 Recurrence Screening in Synthetic Rotamer Structures</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>2.6 Analysis of Results</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>Results and Discussion</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>3.1 Single-Triad Screening Results</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>3.2 Runtime Analysis</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>3.3 Chain Length and the Percent of Recurring Triads</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>3.4 Relative Abundance of Triad Types</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>3.5 Distribution of Total Triad Distances</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>3.6 Chain Length and the Triad Distance Statistics</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>3.7 Analysis of Variance between Means of Distance Statistics</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>3.8 Recurring Triads in Synthetic Rotamer Structures</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>Conclusions and Future Plans</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>Bibliography</td>
<td>35</td>
</tr>
</tbody>
</table>
Introduction

The study of patterns of local structural similarity is important in the understanding of protein structure and enzymatic function. Repeating motifs are a common feature in protein structures and are integral to the biological function of many protein families. The repeating motifs vary in their complexity, structure, and location on the chain. The repeating motifs define entire protein families that are responsible for a verity of critical functions such as protein-protein interaction, mitosis, RNA synthesis, oligomerization of protein chains, nucleic acid and protein recognition.

An extensive list of repeating structural motifs has been identified and described in literature. There are several examples of functionally diverse repeating motifs including the WD repeats, leucine-rich repeats, tetratricopeptide repeats, ankyrin repeats, the leucine zipper, and the zink fingers. The WD repeats are involved in a wide array of functions. These motifs are present in all eukaryotes, and are found in multiple groups of proteins spanning a wide array of important functions such as signal transduction, RNA-processing, transcription regulation, cytoskeleton formation, and the control of various aspects of cell division and metabolism [12]. The leucine-rich repeats provide a versatile structural framework for the formation of protein-protein interactions in a large family of mostly eukaryotic proteins (Figure 1A) [7]. The tetratricopeptide repeat is also a protein-protein interaction motif and is found in a number of functionally diverse proteins that facilitate interactions with other proteins (Figure 1B) [3]. Similarly to the leucine-rich and tetratricopeptide, the ankyrin repeats provide a common structural framework for the interaction with a diverse array of macromolecular targets (Figure 1C). The ankyrin repeats were identified in a large group of proteins including cyclin-dependent kinase inhibitors, transcriptional
regulators, cytoskeletal organizers, and developmental regulators [14]. Another example of repeating motifs containing leucine are the leucine zippers which are found in DNA binding domains in various transcription factors (Figure 1D) [10]. Zink fingers are also found in transcription factors and are responsible for nucleic acid recognition [5]. These examples are large structural motifs, composed of several elements of secondary structure, and present in functionally diverse proteins. In addition, it is common for a protein family to share a structurally similar catalytic site. One such example is the serine protease catalytic triad composed of the amino acids serine, histidine, and aspartate (Figure 1F) [4].

To address the problem of finding similar motifs in protein structures, there exists a variety of computational methods that are able of the detection of patterns of local structural similarity. The programs differ in their manner of representation of the protein structure and in the methods used for the detection of similar spatial arrangements. Some of the classical algorithms introduced in early and mid 1990's, employ such methods as the Ullman graph-isomorphism algorithm, and the geometric hashing techniques. The subgraph-isomorphism algorithm, introduced by J. R. Ullman in 1976, is a computer science problem of detecting the subgraph isomorphism by a brute-force, tree-search enumeration procedure. The algorithm achieves high efficiency by eliminating from the search tree any unnecessary successor nodes, the nodes that are not likely to render a successful search result [16]. The program ASSAM uses the Ullman subraph-isomorphism algorithm for the detection of similar side-chain patterns. In the program, a side chain is represented as a vector between two pseudo-atoms, one representing the backbone atoms and the other representing the functional part of the side chain. The graph representation of a protein structure is then computed with the
side chain vectors as nodes and the intervector distances to other side chain vectors as edges. Given a user-defined query, the main graph can be searched for the isomorphic subgraphs, i.e., only the subgraphs that satisfy the user-defined structural constraints [1] [15].

**Figure 1: Example of Patterns of Local Structural Similarity.** (A) Leucine-rich repeats of ribonuclease inhibitor. The structure 1DFJ, ribonuclease inhibitor (rainbow) in complex with ribonuclease A (red). (B) Tetratricopeptide repeats of the N-terminal transient receptor potential channel domain of the neutrophil cytosol factor 2 protein, the structure 1WM5. (C) The ankyrin repeat domain of the TRPV4 transient receptor potential channel domain, the structure 4DX1. (D) The leucine zipper repeats (residues
in red) of the bZIP transcription factor in complex with DNA, the structure 1FOS. (E) Zink fingers in the ZIF268 protein-DNA complex, the structure 1AAY. (F) Serine protease catalytic triad in the structure of trypsin, 1A0J.

Another major approach in the detection of local side-chain similarities is derived from geometric hashing, a method for finding two-dimensional objects with a preserved affine structure. A method using the geometric hashing techniques compares local protein motifs based on the spatial relationships between all relevant Cα atoms [5]. Several studies introduced an extension to the comparison between the geometry of different Cα arrangements. These methods utilize a 3D reference frame that is attached to each Cα atom and contains the parameters of translation and rotation. The addition of a reference frame significantly reduces the complexity of computation during the pre-processing and recognition phases of geometric hashing [11][17].

In addition to the classical subgraph-isomorphism and geometric hashing techniques, a more recent algorithm ProBiS detects structurally similar binding sites on protein surfaces by means of local structural alignment. The algorithm compares a user-defined query protein to a database of protein structures and returns the proteins that share local structural similarities to the query protein. The similar structural motifs are represented as local structural alignments with similarity scores indicated as different colors on the surface of the query protein. ProBiS has been successfully utilized in the recognition of protein surface binding sites for other proteins, ligands, and DNA [8].

The described methods are highly efficient and accurate in the identification of local structural similarity patterns. While they can be used for any user-defined query, their focus is primarily on the detection of specific motifs across many protein structures. In the present study, we concentrate on a more holistic view of a single monomeric protein structure by breaking it down into all possible unique 3-residue
arrangements and searching for the recurrences of each of those arrangements within their native structure. In this way, we exhaustively screen a protein chain for similar 3-residue side-chain patterns by taking into account all possible triad arrangements in the chain.

In the proposed algorithm, an amino acid residue is represented by two points in space: a backbone centroid and a side chain centroid. The algorithm detects a recurrence of a structural triad based on a pairwise comparison of the distances between corresponding side chain centroids. A recurrence of a structural triad is defined by the following characteristics: (a) it occurs on the same chain with the triad; (b) it consists of the same residue types as in the triad; (c) the residues in the recurrence are not part of the triad; (d) the corresponding pairwise distances between the side chain centroids in the triad and its recurrence differ by less than a predefined tolerance factor (TF = 2.00 Ångstroms by default). For example, the active site of betaine aldehyde dehydrogenase from cod liver is composed of the residues glutamate, asparagine, and cysteine, and the approximate pairwise distances between their side chain centroids are as follows: 11 Ångstroms from glutamate to asparagine; 8 Ångstroms from asparagine to cysteine; and 7 Ångstroms from cysteine back to glutamate. Given our definition and the TF of 2.00, the recurrence of this active site would appear on the same chain; contain the same residue types, glutamate, asparagine, and cysteine; contain the residues that are not part of the active site; and have the following pairwise distances between the residues: $11 \pm 2.00$ Ångstroms from glutamate to asparagine; $8 \pm 2.00$ Ångstroms from asparagine to cysteine; and $7 \pm 2.00$ Ångstroms from cysteine back to glutamate (Figure 2). The TF parameter allows for a control of levels of selectivity and sensitivity and accounts for atomic coordinate variability in Protein Data Bank (PDB) structure files. The default
value of TF is based on the average resolution of PDB structure files of 2.19 Ångstroms, as calculated from all structure files deposited to PDB.

**Figure 2:** Active Site Triad on Chain A of 1A4S and its Recurrence. The catalytic site of the structure 1A4S, a betaine aldehyde dehydrogenase from cod liver, contains residues glutamate, asparagine, and cysteine. (A) A single chain of 1A4S with all glutamate residues in green, all asparagine residues in blue, and all cysteine residues in red. The rest of the chain is not shown. (B) The catalytic site of 1A4S is composed of residues GLU-263, ASN-166, and CYS-297, and is located in the center of the inner core of the protein chain. (C) The recurrence of the active site triad contains residues GLU-58, ASN-22, and CYS-55, and is located on the surface of the protein. The distances, in Ångstroms, indicated in 1B and 1C are the distances between β-carbon atoms. The distances between corresponding residues differ by less than 2.00 Ångstroms. The distances were calculated and the graphical representation was rendered in the PyMOL molecular graphics system.

The observations made from the preliminary explorations suggested that recurring 3-residue arrangements, composed of either active or structural residues, are fairly common in protein structures. In further investigations we utilized the exhaustive recurrence finding algorithm in the screening of enzymatic structures that have a 3-
residue active site. With the obtained results, a series of analyses was performed to infer potential structural differences between recurring and unique triads, those triads for which a recurrence was not found, in enzymatic structures with a recurring or a unique active site.

Materials and Methods

2.1 Software and Computational Resources. The recurrence finding algorithm and the analysis module were implemented in Java version 1.7. The program was developed and run in the Eclipse Juno Standard Development Kit version 4.2.1. Analysis of distance statistics was performed in Minitab 16 statistical software. All computational work was performed on a Windows computer with 16.0GB RAM and an Intel Core i7-3770 CPU at 3.40GHz.

2.2 Structure Pool Compilation and Preliminary Screening. The version 2.2.12 of the Catalytic Site Atlas (CSA) library was used for the identification of enzymatic structures that have a 3-residue catalytic site [9]. Corresponding structure files were downloaded from the Protein Data Bank (PDB) [2]. The resulting pool contained 4,669 PDB structures. Preceding the final exhaustive version of the recurrence finding algorithm that is described in the later sections, a preliminary single-triad version was developed and utilized in the screening of a protein chain for the recurrences of a single triad. The triad could either be defined by a 3-residue active site from the CSA library or assigned by randomly selecting three residues from the chain. The single-triad version was used to search the downloaded enzymatic structures for the recurrences of their 3-residue active sites. Additionally, the single-triad version was used to screen the same pool of structures for the recurrences of randomly selected 3-residue arrangements. The algorithm used a PDB structure file as initial input, isolated a single chain from the
structure, obtained a triad definition from either the CSA library or the random assignment, selected only the residues of the same types as the residues in the triad definition, converted the Cartesian coordinates of the selected residues from All Atom Representation to Reduced Double Centroid Representation, and screened the converted residues for the structural recurrences of the defined triad. All these core routines were kept unchanged for the preliminary single-triad and the final exhaustive versions of the algorithm and are described in depth in the following sections.

2.3 Testing Dataset Compilation. The preliminary screening of the initial structure pool was used to separate the structures into two groups: the structures with a recurring active site and the structures with a unique active site, i.e., no active site recurrences were detected. From each group, 100 structures were randomly selected for the two final datasets: the Recurring Active Site (RAS) dataset (Figure 3A), and the Unique Active Site (UAS) dataset (Figure 3B).

2.4 Exhaustive Screening for Recurring Side-Chain Triads. Each structure from the two datasets was used in the exhaustive screening for the recurrences of all possible triads that can be compiled from the complement of residues in a single chain (Panel 1). In the exhaustive recurrence finding, no information was supplied from the Catalytic Site Atlas library for the triad definitions, instead, the algorithm relied on a pre-compiled list of 1540 triad types. A single triad type represented a combination of three amino acid types arranged alphabetically. The 20 amino acid types yielded 1540 unique triad types. A single triad type could contain two or three of the same amino acid types (e.g., {Ala, Ala, Ala} or {Ala, Ala, Val}); however, given the presence of the triad type {Ala, Ala, Val}, there could not be a triad type {Val, Ala, Ala}. In this way, the triad types
list represented all possible unique combinations of three amino acid types. The following sections describe the four major stages of the recurrence finding algorithm.
**Figure 3:** Example Structures of the Recurring Active Site (RAS) and the Unique Active Site (UAS) Datasets. (A) The chain A of the structure 1LEH, a leucine dehydrogenase, of the RAS dataset with a recurring 3-residue active site (red). Three active site recurrences (blue) were found by the single-triad version of the recurrence finding algorithm. Distances, in Ångstroms, between side chain centroids are shown. All corresponding distances in the active site and the recurrences differ by less than 2.00 Ångstroms. (B) The monomeric structure 1CPN, a jellyroll beta-sandwich protein, of the UAS dataset with a unique 3-residue active site (red). Distances, in Ångstroms, between side chain centroids are shown. The single-triad version of the recurrence finding algorithm did not detect a recurrence of the active site triad.

**2.4.1 Stage 1: Pre-Processing of PDB Structure Files.** The algorithm begins by obtaining 3D coordinates from a PDB file and isolating a single chain from the structure. For each triad type from the pre-compiled list, the routine selects the chain-specific residues: only the ones that correspond to the triad type. For example, for the triad type \{Ala, Ala, Ala\}, only the alanines from the chain are selected; for the triad type \{Ala, Ala, Val\}, only the alanines and valines are selected; for the triad type \{Ala, Leu, Val\}, only the alanines, leucines, and valines are selected. The chain-specific residues corresponding to the triad type are then separated into lists according to their types and each list is then ordered according to the residue number. From the three ordered lists, the algorithm compiles a list of chain-specific triads, i.e., all unique 3-residue arrangements that can be compiled from the total number of residues that correspond to the triad type. In this manner, the algorithm systematically selects all possible chain-specific triads for each of the triad types from the pre-compiled list, placing no distance restriction on the selection of chain-specific triads.
Panel 1: Overview of the Exhaustive Recurrence Finding Algorithm

Outer loop: For each triad type:

- Full PDB structure in AAR
- Isolate chain A and select the chain-specific residues that correspond to the current triad type, e.g., only the alanines, histidines, and leucines.
- Convert from All Atom to Double Centroid Reduced Representation:

<table>
<thead>
<tr>
<th>AAR (8 points in space)</th>
<th>DCRR (2 points in space)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X Y Z</td>
<td>X Y Z</td>
</tr>
<tr>
<td>1 N 31.660 -0.621 3.899</td>
<td>1 BBC 32.357 -10.278 2.683</td>
</tr>
<tr>
<td>2 CA 31.444 -10.590 2.793</td>
<td></td>
</tr>
<tr>
<td>3 C 32.769 -10.896 2.111</td>
<td></td>
</tr>
<tr>
<td>4 O 33.596 -10.006 1.929</td>
<td></td>
</tr>
<tr>
<td>5 CB 30.425 -10.012 1.776</td>
<td></td>
</tr>
<tr>
<td>6 CG 29.899 -11.055 0.838</td>
<td></td>
</tr>
<tr>
<td>7 OD1 30.540 -11.427 -0.150</td>
<td></td>
</tr>
<tr>
<td>8 ND2 28.702 -11.559 1.139</td>
<td></td>
</tr>
</tbody>
</table>

- Separate chain specific residues according to the type and sort in descending order of their numbers in chain, e.g., for the current triad type Ala His Leu:
  - Ala: 1, His: 57, Leu: 21
  - Ala: 3, His: 143, Leu: 79
  - Ala: 194, His: 159, Leu: 81

Inner loop: For each combination of 3 chain-specific residues:

- Let the current chain-specific triad be Ala - 1 His - 57 Leu - 21
- Then the search space are all other triads from the chain-specific residues:
  - Ala - 3 His - 143 Leu - 19
  - Ala - 3 His - 143 Leu - 81
  - Ala - 3 His - 159 Leu - 79
  - Ala - 194 His - 159 Leu - 81

Perform the search for recurrences of the current chain specific triad: other triads in the search space in which the corresponding distances differ by less than a Tolerance Factor of 2.00 Angstroms (see Panel 2).

Save results to tab-delimited output files storing results for each triad type in all screened chains

triad_type_file
- each line contains:
  - Types and numbers in chain-specific triad, distances between the residues, number of recurrences of the chain triad

cumulative_recurrence_data_file
- each line contains:
  - Triad type, number of triad type residues, number of total chain-specific triads, number of recurring triads, total number of recurrences for the triad type

1. Ala Ala Ala
2. Ala Ala Arg
...
Panel 1: Overview of the Exhaustive Recurrence Finding Algorithm. The high-level description explains the logistics of the implemented algorithm in finding the recurrences of side-chain triads. The only necessary input information is a PDB structure file. The algorithm outputs the findings in tab-delimited files that are used in further analysis.

2.4.2 Stage 2: Data Reduction. Once the residues are isolated, separated, ordered, and the list chain-specific triads is generated, the algorithm proceeds by converting the Cartesian coordinates in All Atom Representation (AAR) to Double Centroid Reduced Representation (DCRR) [13]. This step significantly decreases the number of distance calculations that are necessary during the search for recurrences. Given the AAR, an amino acid residue is defined by four sets of Cartesian coordinates for the four backbone atoms and an appropriate number of coordinates for the side chain atoms. In DCRR, the representation is reduced to two points in space, regardless of the side chain structure: backbone centroid (BBC) and side chain centroid (SDC). The conversion is achieved by obtaining the averages between the corresponding coordinates of the backbone and side chain atoms.

2.4.3 Stage 3: Searching for Recurring Triads. After the conversion of coordinates from AAR to DCRR, the algorithm proceeds with the screening for recurrences of each chain-specific triad. The search space for possible recurrences is formed by all other chain-specific triads. The recurrence-search subroutine performs the distance comparisons between the residues of the current triad and all other triads in the search space (Panel 2). In the process, the distance between a pair of residues is calculated as the vector magnitude between the two points in space. A triad recurrence is registered only if all three pairwise distances differ by less than 2.00 Angstroms. With the exception of a glycine, the algorithm calculates and compares the distances between
the side chain centroid coordinates of the two corresponding residues. In the case of
glycine, the distance to the glycine's backbone centroid is used instead.

Panel 2: Distance Comparisons in Triad Recurrence Finding.

The current chain-specific triad (upper left) from the example in Panel 1 consists of the three residues: Ala-1, His-57, and Leu-21. The small dots represent all other chain-specific residues that are of the same type as the triad type: alanines (red), histidines (green), and leucines (blue). Ala*, Leu*, and His* in the center compose the registered recurrence of the current triad. All three corresponding distances differ by less than the Tolerance Factor (TF) of 2.00 Ångstroms.

A distance between two side chain centroids is calculated as the magnitude of a vector between two points in space. The washers represent the cross sections of hollow spheres that outline the distance constraints: the inner bound of the sphere is the distance between the two residues in the current triad minus the TF; the outer bound is the distance plus the TF; and the dashed middle circle is the exact distance. The search process is a cascading event and the search order is alphabetical in accordance to the ordering in each triad type. The algorithm begins with an alanine by casting a sphere around it in search for a leucine that satisfies the distance constraint. It then casts another sphere around the leucine to search for a histidine. Finally, the algorithm looks to close the triad if the distance constraint is satisfied from the histidine back to the alanine that the search had started with. The search is abandoned if the distance constraint is not satisfied at any step of the cascade, in that case the algorithm proceeds to the next possible recurrence of the triad type.

Panel 2: Distance Comparisons in Triad Recurrence Finding. The panel describes the core subroutine of the recurrence finding algorithm. The method calculates and compares distances between residues in a triad and its possible recurrence. The search is performed in the alphabetic order, in this case, starting with the alanine and finishing with the leucine. The search is satisfied when all corresponding pairwise distances are found to differ by no more than the Tolerance Factor of 2.00 Ångstroms.
2.4.4 Stage 4: Results Output. Once the search for recurrences in completed, the algorithm reports the results to tab-delimited files which are then used in the downstream analysis. A file is created for each triad type with each line containing the following information: the types and numbers of chain-specific residues in the triad; the number of recurrences for that triad; and the corresponding distances between the residues in that triad. The result are the 1540 such files for each triad type. Each of these files is separated by a special identifier for each screened structure. In addition, a file with the cumulative recurrence data is created to store the information on the aggregating counts of recurring triads and their corresponding recurrences. Each line in the file contains the following information: the triad type; the size of the current structure; the numbers of chain-specific residues corresponding to the triad type; total unique chain-specific triads; total recurring chain-specific triads; and the total number of recurrences that were found for the triad type.

2.5 Recurrence Screening in Synthetic Rotamer Structures. In addition to the two datasets of natural structures, the exhaustive recurrence finding algorithm was employed in programmatically generated alternative rotamer structures. A set of 100 alternative structures was generated based on four structures from the dataset of enzymes with a recurring active site. The four structures varied in chain length from 104 to 374 residues. Each alternative rotamer was obtained by randomizing the side chain centroid coordinate of each residue in the structure without disturbing their respective backbone centroid coordinates. For each residue, the algorithm randomly generated a new set of side chain coordinates while keeping the distance from the new side chain centroid to its backbone centroid constant. No control for steric effects was utilized in the generation of the new side chain coordinate with the only restriction being the
distance to the relative backbone centroid. The percentages of recurring triads obtained in the rotamers were compared to those of the corresponding natural structures.

2.6 Analysis of Results. A separate module was developed to analyze the results generated by the exhaustive recurrence finding algorithm. For the results from each dataset, the analysis module processed the 1540 individual result files for each triad type. The module calculated basic descriptive distance statistics including the minimum, maximum, and average of the total triad distances for recurring and unique triads. In addition, the analysis module calculated the distribution of distances in recurring and unique triads across 30 intervals of 10 Ångstroms each. A simple ANOVA analysis was performed in determination of the variance between the means of distance statistics to the 95% confidence interval. The results of the ANOVA analysis were verified with Tukey's and Fisher's multiple comparison methods.

Results and Discussion

3.1 Single-Triad Screening Results. The preliminary single-triad version was employed to screen the initial pool of 4,669 enzymatic structures for the recurrences of their 3-residue active sites. As a result, 1,541 (33%) of those structures were found to have an active site with on average of 3.7 recurrences. In addition, the preliminary version was used in the screening of the same structure pool for recurrences of randomly selected 3-residue sites. In the random selection of triads, no constraints on the distances between the selected residues were placed; the results were taken as averages of 20 iterations through the structure pool selecting a different random triad in each structure at each iteration. In each structure, the random triads were composed of any three residues from a single isolated chain except for the residues involved in catalysis. As a result, 81.1±0.6% of the screened structures were found to have on
average of 17.5 recurrences of a randomly selected triad. These findings indicated that active site triads recur on a lower rate as compared to the triads that are not involved in catalytic activity. To explore the possible differences in structures with and without a recurring active site, the final exhaustive version of the recurrence finding algorithm was used in the screening of two datasets of 100 enzymatic structures each: the Recurring Active Site (RAS) dataset, in which at least one recurrence of the active site was found; and the unique active site (UAS) dataset, in which no active site recurrences were found (Figure 4).

(A) RAS Dataset: 100 Enzymatic Structures Containing a Recurring 3-Residue Active Site

2aae,2aad,1leh,1ldy,1fyl,1ldf,2qfr,2aat,2qfp,11cp,2uyu,1pvs,1dcy,3tdt,1lbf,2ldx,1lya,2uyv,2ocq,3dha,2oci,2ock,1ldiz,1plr,1bcb,2cer,1h69,1h66,1yah,3dih,1ya1,1h3i,1ye3,1b8a,1j1a,1jlt,2jhf,2jhg,2cj1,2ewb,1fp8,1dos,1fp9,6adh,1pod,1poq,1hl1,2pnd,3bir,1b15,1txg,4bir,1nhw,1n70,1jji,1f3l,1wng,1hse,1h17,3cvt,3cvs,1ahg,1hkh,3cwa,2p5f,1cwy,1aco,2p5c,2z30,2d61,3cwa,2d66,2d65,2d64,2p6d,1j1u,2p6k,1q1j,1sfv,2z57,2z58,2z55,2p07,2p05,1adg,1adf,1adb,2z2y,1ma0,1q8b,1q8b,2d7y,2d7z,2r2n,1wqw,1qbg,1ayp,3cnj,2b5w,1db4

(B) UAS Dataset: 100 Enzymatic Structures Containing a Unique 3-Residue Active Site

1dji,1w7w,1y7i,1fe5,1u5c,3b3g,1vip,2itk,3df9,1cpn,1kko,1pjs,1rtu,2z2i,1gwu,2qgb,1ley,1bc2,2cjy,1n7k,2gn4,1dhf,1gcl,2f5m,1fqr,1mf4,1odc,1djh,2dy3,1jgg,1jr4,1r70,1r44,2h9i,1iyh,1h5g,1jif,1eim,1rym,1g7x,2cbu,1fbb,2vte,1cvh,2dtp,2z5z,1ucj,1ttm,2o0t,1gx4,1eet,3dc,1h11,3exb,3bp2,1k7f,1zai,1grk,1av,1h5h,2whd,1beq,2yyy,1t3c,1eal,2pc,6atj,1zsn,1rza,2ckr,2nux,3f4v,1lxy,1xxw,3fuj,2egz,1r0m,2vm3,1gim,1ugq,1rgh,1adi,1wal,3fuk,1g53,1g3q,2bia,3ii0,1de5,2nuy,1bjj,1gms,2idz,3knf,1oz6,1gmp,1rge,3gke,1fkl,1act

*Figure 4:* Datasets of Enzymatic Structures for the Exhaustive Recurrence Finding Algorithm. The structures in each dataset were selected randomly from an overall pool of 4,669 PDB structure files of enzymes with a 3-residue active site. (A) 100 enzymatic structures containing a 3-residue active site with at least one recurrence. (B) 100 enzymatic structures containing a 3-residue active site with no recurrences.
3.2 Runtime Analysis. For the two datasets in Figure 4, the structures were randomly selected from the two groups of the original pool. In the RAS dataset, the structures ranged from 104 to 532 residues in chain length. In the UAS dataset, the structures ranged from 96 to 624 residues. The time for the algorithm to complete the screening of a single structure was measured. The increasing chain length caused an exponential growth in the number of total triads in the structure, therefore, an exponential growth in the runtime. The smallest structure of 92 residues took just over 12 seconds to complete while the largest structure took 72.4 hours (Figure 5).

![Figure 5: Growth in Algorithm Runtime with Increasing Chain Length](image)

Figure 5: Growth in Algorithm Runtime with Increasing Chain Length. The runtime was measured for a small subset of the two datasets of enzymatic structures. The increasing chain length demonstrated an exponential growth in runtime. The smallest structure of 92 residues took just over 12 seconds to complete while the largest structure took 72.4 hours. An exponential trend line was fitted, and the line equation was used in the estimation of total runtime for the two 100-structure datasets.

3.3 Chain Length and the Percent of Recurring Triads. Apart from the growth in runtime of the algorithm, the increasing chain length clearly demonstrated a positive correlation to the percent of recurring triads in a structure (Figure 6). Between the two datasets, the smallest chains showed just below 40% of recurring triads while the largest
ones fell above 90% of recurring triads. The structures that were registered with the highest percent of recurring triads were the chain A's of two homologous structures from the UAS dataset, 2Z5Z and 3GC1. The two structures showed 97.3% of recurring triads. The positive correlation between the chain length and the percent of recurring triads does not come as a surprise. With the growing chain length, the number of possible unique 3-residue arrangements grows exponentially causing a dramatic increase in the search space for each chain-specific triad. Enzymatic structures organize tightly into highly specific and intricately ordered 3D folds, creating a large collection of individual amino acid residues concentrated in a very limited 3D space. Our data shows that as the structures become larger, these tightly packed clusters create more possibility for a greater number of small, structurally similar subsets. Additionally, the tertiary structures of larger chains possess a greater number of repetitive secondary structure elements that can cause a bias towards the repeating of smaller arrangements of side chains.
Figure 6: Chain Length and the Percent of Recurring Triads. In each structure, the percent of recurring triads was calculated with the total number of unique 3-residue arrangements compiled from the entire residue complement in a single chain, and the number of those 3-residue arrangements that were found to recur at least once on the same chain. In the two datasets, RAS (blue, 100 structures with a recurring active site) and UAS (orange, 100 structures with a unique active site), larger structures showed greater percentages of recurring triads. In the RAS dataset, the smallest structures tested were of 104 residues and were calculated as 51.2% of recurring triads, the largest were of 532 residues and of 96.5% of recurring triads. In UAS, the smallest structures were of 96 residues showed between 43.6% and 47.5% of recurring triads, and the largest ones were of 624 residues and showed 93.9% of recurring triads. The structures that showed the largest percent were both from the UAS dataset, composed of 595 residues each, and showed 97.3% of recurring triads.

3.4 Relative Abundance of Triad Types. To quantify the expected recurrence rates for each of the 1540 triad types, the relative abundances for each triad type were calculated with respect to the dataset. The relative abundance values were calculated as the percent of the chain-specific triads of a given triad type from the total number of triads from the dataset. The more abundant triad types were expected to have higher rates of recurring chain-specific triads. In agreement with the expectation, increasing relative abundance promoted a steep logarithmic growth in the percentage of recurring triads (Figure 7). In both datasets, the most abundant triad types reached above 98% of recurring triads. The most recurring triad types were composed primarily of amino acids alanine, leucine, valine, and glycine. The least recurring triads were composed primarily of amino acids cysteine, tryptophan, histidine, and methionine. These were the triads of low relative abundance that fell below 20% of recurring triads in the RAS dataset, and below 30% in the UAS dataset.
Figure 7: Relative Abundance of Triad Types and the Percent of Recurring Triads. (A) Enzymatic structures with a recurring active site. (B) Enzymatic structures with a unique active site. The relative abundance values for each triad type were calculated as the percent of all chain-specific triads of the triad type from the total number of triads from the dataset. The relative abundance values were scaled up by 100-fold. Increasing relative abundance promoted a steep logarithmic growth in the percent of recurring triads. Indicated are the 5 topmost and the 5 bottommost triad types in terms of the percent of recurring triads.

3.5 Distribution of Total Triad Distances. To construct the distribution graphs for the total triad distances, the analysis module tallied the numbers of triads that fall into 10-Ångstrom intervals in the range of 0 to 300 Ångstroms. The distribution shapes indicate negative skewness for the distances in unique triads from both datasets, while the distances in recurring triads appear to be normally distributed (Figure 8). The distribution peaks for the recurring triads were observed in the intervals of 90 to 100 and of 80 to 90 Ångstroms in the RAS and the UAS datasets, respectively. The distribution peaks for the unique triads fall into the intervals of 110 to 120 and of 100 to 110 Ångstroms in the RAS and the UAS datasets, respectively. In the RAS structures the population means were calculated as 81.7 Ångstroms in the recurring triads (population size of 774,061,064 triads) and as 98.5 Ångstroms in the unique triads (population size
of 75,038,683 triads). In the UAS structures, the population means were calculated as 85.5 Ångstroms in the recurring triads (population size of 634,623,587 triads) and 99.9 Ångstroms in the unique triads (population size of 57,909,844 triads). Given the negative skewness of the distribution of the total unique triad distances, the majority of observations fall to the right of the population mean. In other words, the residues in the majority of unique triads are located farther apart of each other, as opposed to the residues in recurring triads. The difference in the distribution shapes can also be attributed to the different population sizes of recurring and unique triads. It is possible that the distribution of distances in unique triads approaches normality given large enough population size. Nevertheless, in agreement with the indication of larger distances in unique triads are the results of the further analysis of distance statistics drawn from recurring and unique triads.

Figure 8: Distribution of Total Triad Distances in Populations of Recurring and Unique Triads. (A) Distribution of total triad distances in recurring triads in enzyme structures with a recurring active site. (B) Distribution of total triad distances in recurring triads in enzyme structures with a unique active site.
**Figure 8 Continued:** Distribution of Total Triad Distances in Populations of Recurring and Unique Triads. (C) Distribution of total triad distances in unique triads in enzyme structures with a recurring active site. (D) Distribution of total triad distances in unique triads in enzyme structures with a unique active site. Population means (blue lines) are indicated in Ångstroms. The negatively skewed distribution of unique triad distances indicates that the residues in unique triads tend to be separated by greater distances than the residues in recurring triads which adhere to normal distribution.

### 3.6 Chain Length and the Triad Distance Statistics.

In addition to the distribution of distances in recurring and unique triads, the analysis module registered the minimum, maximum, and average total distances between the residues in the recurring and the unique triads of each screened structure (*Figure 9*). Similar patterns were observed regardless of whether the active site is unique or recurring. With the growing chain length, we observed an increase in the maximum and average total distances between the triad residues. In the smallest structures from either dataset, the maximum total and the average total distances fell around 100 and around 50 Ångstroms, respectively, from either recurring or unique triads. For the larger structures, the maximum distances increased to about 200 Ångstroms in the RAS dataset, and to about 250 Ångstroms in the UAS dataset. The linear trend lines of the growth in
maximum distances revealed a difference in the Y-intercept of around to 10 Ångstroms between the populations of recurring and unique triads. Additionally, on the plots in Figure 4A (the RAS dataset) and Figure 4B (the UAS dataset) the linear trend lines for the maximum distances in recurring and unique triads appear parallel to each other, indicating no increasing separation between the two as the chain length grows. Contrary to the maximum distances, the growth in average total distances appears slightly steeper for the unique triads as compared to the recurring triads. As the structures grow in chain length, the separation between the average total distances in recurring triads and the average total distances in unique triads appears to increase. Between the two datasets, the average total distances from recurring triads range from 50 Ångstroms in the smallest structures to 95 Ångstroms in the largest ones, while the average total distances in unique triads range from 50 Ångstroms to slightly above 140 Ångstroms in the largest ones. The observations in maximum and average total distances suggest that the triads of residues that are positioned closer together have a higher chance of a recurrence. While it is difficult to explain the growing separation between the average total distances, a possible interpretation may be that the spatial density of residues stays relatively constant with growing chain lengths. In other words, larger protein chains still maintain the necessary level of compactness of their residues resulting in faster growth rates for maximum triad distances, the residues at polar ends of the chain, as opposed to the average triad distances, the residues in the inner core. In support to this interpretation, the minimum triad distances appeared to be independent of the chain length. For either the RAS or the UAS dataset, the minimum distances consistently fall between 10 and 15 Ångstroms.
Figure 9: Chain Length versus Minimum, Maximum, and Average Total Triad Distances in Recurring and Unique Triads. (A) Dataset of Enzymatic Structures with a Recurring Active Site. As the linear trend lines indicate, maximum and average total distances between triads increase with the growing chain length while the minimum distances are static. As the chain length increases, the separation between the average total distances also increases.

Figure 9 Continued: Chain Length versus Minimum, Maximum, and Average Total Triad Distances in Recurring and Unique Triads. (B) Dataset of Enzymatic Structures with a Unique Active Site. Similarly to the results for the RAS structures in Figure 9A,
the maximum and average total distances between triads increase with the growing chain length while the minimum distances are static. As the chain length increases, the separation between the average total distances also increases.

3.7 Analysis of Variance between Means of Distance Statistics. The differences in the means of minimum, maximum, and average total distances were assessed in a one-way analysis of variance to the 95% confidence level (Figure 10) and in a multiple comparison analysis using Tukey's and Fisher's methods (Table 1). The corresponding means were compared within each dataset separately and across the two datasets. All results obtained with the ANOVA and the multiple comparison analyses agreed in the determination of significantly different population means. In the RAS dataset, the following means were found to be significantly different: the means of maximum distances in the recurring and the unique triads ($P = 0.004$), the means of average distances in recurring and unique triads ($P = 0.000$). Contrary to maximum and average, the means of minimum distances in the recurring and the unique triads of the RAS dataset were not found to be significantly different ($P = 0.460$). Contrary to the RAS, the comparisons between the recurring and the unique triads within the UAS dataset showed significant differences in means of all three distance statistics: $P = 0.001$ for the means of minimum distances, $P = 0.015$ for the means of maximum distances, and $P = 0.000$ in means of average distances between the residues in recurring and unique triads. In the cross-dataset comparisons between the RAS and the UAS datasets, significantly different means were found only between the minimum distances in recurring triads from the two datasets ($P = 0.000$). All other cross-dataset comparisons rendered means as insignificantly different. The means of total active site distances were not found to significantly different between the two datasets.
Figure 10: Analysis of Variance Between Triad Distance Statistics from Structures of Recurring Active Site (RAS) and Unique Active Site (UAS) Datasets. (A) Minimum triad distances in recurring (population size of 774,061,064) and unique (population size of 57,909,844) triads. (B) Maximum triad distances in recurring and unique triads. (C) Average total triad distances in recurring and unique triads. (D) Mean active site distances. In comparisons within the RAS dataset, to the 95% confidence level the means were found to be significantly different between the maximum and average total distances. In comparisons within the UAS dataset, to the 95% confidence level the means were found to be significantly different in all three distance statistics. In the
cross-dataset comparisons, significantly different means were found in the minimum distances in recurring triads, and in the average total distances also in recurring triads.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pairwise Comparison</th>
<th>Grouping with Tukey's Test</th>
<th>Grouping with Fisher's Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comparison within the RAS dataset</strong></td>
<td>Minimum Recurring</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Minimum Unique</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Maximum Recurring</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>Maximum Unique</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Average Recurring</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>Average Unique</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td><strong>Comparison within the UAS dataset</strong></td>
<td>Minimum Recurring</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Minimum Unique</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>Maximum Recurring</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>Maximum Unique</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Average Recurring</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>Average Unique</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td><strong>Comparison across the RAS and the UAS datasets</strong></td>
<td>Minimum Recurring RAS</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>Minimum Recurring UAS</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Minimum Unique RAS</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Minimum Unique UAS</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Maximum Recurring RAS</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Maximum Recurring UAS</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Maximum Unique RAS</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Maximum Unique UAS</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Average Recurring RAS</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Average Recurring UAS</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Average Unique RAS</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Average Unique UAS</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

**Table 1:** *Grouping of Means by Tukey's and Fisher's Methods.* The pairwise comparisons of distance statistics were performed within the two datasets individually and across the two datasets. The pairwise comparisons that rendered significantly different means are highlighted in bold. Within the Recurring Active Site (RAS) dataset, the maximum and average distances in recurring and unique triads were grouped differently, indicating a significant difference in means. Within the Unique Active Site (UAS) dataset, all three distance statistics were grouped differently by either comparison method. In the cross-dataset comparison, only the minimum distances in recurring triads from the two datasets were assigned into different groups.
The results obtained with the analysis of variance between the means of distance statics support the observations made from the distribution of total triad distances in recurring and unique triads (Figure 8). The negatively skewed distribution of distances in unique triads indicates that the residues in unique triads tend to be separated by greater distances as compared to the residues in recurring triads, the total triad distances of which adhere to normal distribution. In agreement, the analysis of variance revealed significantly greater means of maximum and average total distances between in the unique triads as compared to those in the recurring triads.

3.8 Recurring Triads in Synthetic Rotamer Structures. To infer whether the recurring triads are only common to natural, biologically relevant protein structures, the exhaustive recurrence finding algorithm was used to calculate the percentages of recurring triads in a set of synthetic, alternative rotamer structures. Based on four structures from the RAS dataset, a set of 100 alternative rotamers was generated and supplied into the recurrence finding algorithm. The alternative structures were programmatically obtained by randomizing the side chain coordinates for every residue in the chain while keeping the distance to the corresponding backbone centroid constant. The backbone centroid coordinates were not altered in the generation of rotamers. With the growing chain length, we observed an increase in the percent of recurring triads in the alternative rotamers. In the smallest structure of 104 residues the difference between the mean percentage in the 100 rotamers and the percent in the natural structure is close to 4% (Figure 11A), while in the largest structure of 374 residues the difference is only a fraction of a percent (Figure 11D). The smallest difference is with the second largest structure 2UYU of 274 residues. The natural structure of 2UYU showed 79.2% of recurring triads, a difference of 0.3% as compared to the mean of 78.9% in the
alternative structures (*Figure 11C*). Additionally, the largest number of rotamers with a percentage higher than that of a natural structure was found for the structure of 2UYU, as indicated by a number of peaks that reach higher than the level of the percentage of recurring triads in the natural structure (*Figure 11C*).

As mentioned above, the backbone structures were not altered in the generation of the alternative rotamers. The only constraint in the assignment of alternative side chain coordinate was maintaining the unchanged distance to the corresponding backbone coordinate without taking into account any possible steric effects. The results suggest that a significant decrease in the number of recurring triads can only be achieved by disrupting the structure of the backbone which may be positioned in such a way that it guarantees for the majority of small spatial arrangements to find at least one structural recurrence. Nevertheless, in the four cases of this experiment the means of the recurring triad percentages in the alternative structures were consistently lower than those of the natural structures. These observations may suggest a potential bias towards the recurring small structural side-chain arrangements in biologically relevant structures.
Figure 11: Recurring Triads in Alternative Rotamer Structures. The alternative rotamer structures were obtained programmatically by randomly assigning new side chain centroid coordinates to every residue in the structures. The distances from each new side chain to the corresponding backbone centroid were kept constant. A set of 100 alternative rotamers was based on one of four natural structures ranging in chain length from 104 to 374 residues. The exhaustive recurrence finding algorithm was employed to measure the percentage of recurring triads in each alternative structure. With the growing chain length, the separation between the percentage in natural structure (blue lines) and the mean percentage in rotamers (red lines) decreases. (A) Rotamers based on chain A of the structure 2AAE, 104 residues. The natural structure of 2AAE showed 51.9% of recurring triads, 3.8% higher as compared to the mean of 48.1% in the alternative structures. The rotamer with the highest percentage of recurring triads tied with the natural structure at 51.9%. (B) Rotamers based on chain A of the structure 1FY1, 225 residues. The natural structure of 1FY1 showed close to 83.5% of recurring triads, less than 1% higher as compared to the mean of 82.5% in the alternative structures. (C) Rotamers based on chain A of the structure 2UYU, 274 residues. The natural structure of 2UYU showed 79.2% of recurring triads, a difference of 0.3% difference as compared to the mean of 78.9% in the alternative structures. (D) Rotamers based on chain A of the structure 1AXE, 374 residues. The natural structure of 1AXE showed 90.9% of recurring triads, a difference of 0.4% against the mean of 90.5% in the alternative structures.
Conclusions and Future Plans

Repeating structural motifs are a common feature in protein structures. The study of patterns of local structural similarity is important in structural biology. Discovering small structurally similar protein motifs helps in the understanding of a protein structure and may aid in the discerning of its function. The current study proposes a holistic approach in discovering small, structurally similar side-chain arrangement within a single, monomeric protein structure. To this end, we have developed an exhaustive recurrence finding algorithm to screen a monomeric protein structure in the search of recurring 3-residue side-chain arrangements. The algorithm considers all distinct triads that can be compiled from the entire complement of residues in the chain.

We have utilized the proposed algorithm in the screening of two datasets of structures: 100 enzymes with an active site with at least one recurrence, and 100 enzymes with a structurally unique active site. The results of the screening revealed that increasing chain length promotes a strong growth in the percent of recurring triads. While no appreciable difference was found between the two datasets, the analysis of descriptive statistics revealed significant differences in the means of maximum and average total triad distances from recurring and unique side-chain triads. The means of maximum and average total triad distances in recurring triads were found to be significantly less as compared to the means of maximum and average total distances in unique triads. The findings suggest that small patterns of local structural similarity tend to be composed of side chains that are positioned closer together, as opposed to the side chains of the arrangements that are unique to the chain. Additionally, with the growing chain length we observed an increasing separation of the two distance statistics from
recurring and unique triads. As the protein chains become larger, the distances between the residues in unique triads increase at a higher rate as opposed to the distances between the residues in recurring triads. In agreement with this observation, the constructed distribution of total triad distances revealed that the distances in recurring triads are distributed normally, while the distances in unique triads are negatively skewed.

In addition, our results may suggest a potential bias towards recurring structural triads in biologically relevant structures. In the exhaustive screening of synthetically generated alternative rotamer structures, we observed a slight overall decrease in the percent of recurring triads as compared to the natural structures that the rotamers were based on. The results suggest that a significant decrease in the number of recurring triads can only be achieved by disrupting the underlying backbone structure of the protein chain. Given the compactness and intricate organization of protein structures, the recurrence of such small side chain arrangements may occur purely by chance; however, even the slight overall decrease in the percent of recurring triads in the synthetic structures may hint towards a possible biological trend.

While it is difficult to pinpoint any tangible biological significance or practical implications of the presented results, the current study suggests an alternative viewpoint in the analysis of local similarity patterns in protein structures. In the course of this project, in addition to the two datasets of enzymatic structures and the sets of alternative rotamers, the algorithm was used in the screening on recurring triads in a small set of predicted structures (results not shown). The predicted structures were obtained by homology modeling and were based upon an archaeal bifunctional aldolase/phosphotase enzyme. The quality of predicted structures was assessed via a
Ramachandran plot. While a much larger dataset of predicted structures is necessary to draw any meaningful conclusions, our limited results suggested a negative correlation between the quality of the predicted structure and the percent of recurring triads. As part of the future directions of this study, we plan to expand the dataset of predicted protein structures and utilize the proposed recurrence finding algorithm as a potential, additional means of assessing the quality of a predicted protein structure.

In addition, we plan to expand the algorithm with the ability to screen a protein chain for the recurrences of larger side-chain arrangements to take into account enzymatic structures with active sites of 4 and 5 residues. Especially true to the structures in which the 3-residue active sites recur more than once, it was observed that the active site recurrences appear on the protein surface, rather than in the inner core as mostly the case with the actual catalytic sites. These active site recurrences can potentially be important to the catalytic activity as they may possess limited binding affinity for substrate and may aid the enzyme in finding its substrate in the environment. In future studies we plan to concentrate on the separation of active site recurrences that appear on the surface from those that are found in the inner core of protein structures. To infer a potential significance of the surface recurrences we plan to utilize the hydrophobicity of the residues in the active site recurrences that appear on the surface. An analysis can be performed to discern if the statistical likelihood of a given hydrophobic residue being found on the surface is greater given that the residue is a part of an active site recurrence, as opposed to any other hydrophobic residue that is exposed to the environment.

Finally, we plan to revisit the implementation of the algorithm in an attempt of improving the algorithm's efficiency to reduce the required runtimes. Screening larger
datasets of structures may provide more useful information in analyzing the descriptive
statistics from populations of recurring and unique triads. Higher efficiency can also
allow us to perform experiments in which the effect of changing the tolerance factor on
the number of recurring triads is measured.
Bibliography


