Applications of Numerical Linear Algebra to Protein Structural Analysis: the Case of Methionine-Aromatic Motifs

by

David Sebastian Weber

B.Sc., Simon Fraser University, 2016

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Chemistry
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2018

Approval

Name:	**David Sebastian Weber**
Degree:	**Master of Science (Chemistry)**
Title:	**Applications of Numerical Linear Algebra to Protein Structural Analysis: the Case of Methionine-Aromatic Motifs**
Examining Committee:	**Chair:** Corina Andreoiu Associate Professor
	Jeffrey J. Warren Senior Supervisor Assistant Professor
	Loren Kaake Supervisor Assistant Professor
	Vance E. Williams Supervisor Associate Professor
	David Sivak Internal Examiner Assistant Professor Department of Physics
Date Defended:	**December 7, 2018**

Abstract

Linear algebraic algorithms uncovered a preponderance of protein methionine residues interacting with two or more aromatic residues. The average geometric relationship between transition metals and methionine-aromatic interactions in PDB crystal structures was assessed with a nearest neighbours-like algorithm (“Met-aromatic”) developed for finding and classifying methionine-aromatic interactions. Here, we assumed that a methionine could interact with one to six midpoints between aromatic carbon atoms in any of phenylalanine, tyrosine, and tryptophan; an integer we termed “Order of Interaction.” Serendipitously, an oversight in Met-aromatic led to the discovery of a significant number of interactions of order exceeding VI, suggesting a large number of methionine residues interacting with two or more aromatics. This was termed the “Bridging Interaction.” Herein, the methionine-aromatic and bridging interactions are discussed in light of their possible redox roles.

Keywords: Sulfur-aromatic interactions, numerical linear algebra, bioinformatics, algorithms, Protein Data Bank

Dedication

I dedicate this thesis to the Python Software Foundation. An explanation follows.

My experiences as both a high school student and undergraduate student were unremarkable. I felt privileged to attend university and complete a degree program, but oftentimes I felt a void in my education. Throughout my undergraduate career I felt there was a preponderance for rote memorization over learning to apply concepts. While this model may indeed work for many students, I felt this model of education failed to inspire me. This was especially the case for my mathematics pre-requisite classes. All too often I encountered midterm problems like:

Evaluate the following:

$$\int_0^2 \frac{2+x}{2\cos(x)}dx \tag{1}$$

Or:

Evaluate the following:

$$\begin{pmatrix} 1 & 2 \\ 3 & 3 \end{pmatrix} \begin{pmatrix} a & 2 \\ 1 & 1 \end{pmatrix} \tag{2}$$

These sorts of problems were certainly a good means of practicing my ability to perform calculus and linear algebra but I never felt I was truly inspired beyond finding their solutions. I forgot how to work these mathematical disciplines soon after completing these pre-requisites. While I certainly can't speak for all students, I feel many follow a similar trajectory.

Fast forward about a year, I discovered the Python programming language. This was a particularly pivotal moment of my life as suddenly I had a workspace in which I could begin exploring mathematics dynamically; that is, seeing equations "work" in various constructs such as loops. Suddenly I could build worlds consisting of mathematical operations and conditional statements. Suddenly the significance of operations such as matrix multiplications (which initially meant little to me) became apparent. Suddenly the need to learn how to evaluate equation 2 had some significance.

Having the ability to explore mathematics dynamically and see direct applications, I felt I quickly started seeing the bigger picture. The answers to questions like "why is Graph Theory important?" became readily apparent. Additionally, becoming proficient in a programming language allowed me to view Nature from a different perspective. I began to ask myself questions such as "Does there exist a relationship between peoples' moods and the colors they wear?" or "Could a person's choice of dessert be predicted programmatically?" These seemingly mundane questions provided constant practice. Over time, I started seeing relationships between seemingly independent entities such as transit routes and protein structure; the ability of which proved to be practical for much of the research presented in this thesis.

Additionally, my exposure to programming allowed me to pick up many other practical skills, such as user interface design, machine learning and automation - skills which have real life applications and are especially in demand in today's job market. Python also quickly paved the way for two other programming languages: MATLAB and C++, in addition to a markup language: LaTeX.

Upon joining the Warren lab as a graduate student, I didn't view Protein Data Bank files as mundane text files containing 3-dimensional coordinates. Instead I viewed PDB files as opportunities. Opportunities to develop novel mathematics from which I could derive geometric relationships. 3-dimensional coordinates were not arbitrary numbers, but rather nodes in a larger network which could be explored computationally. It was only because of open source projects like Python that I feel I have had a successful and productive graduate school career and this is why I feel dedicating this thesis to the Python Software Foundation is appropriate.

Acknowledgements

I am grateful to my graduate supervisor, Dr. Jeff Warren, for providing me with the opportunity to work in his lab. Jeff took me under his wing despite my skill set which differs greatly from many other students in his lab. Jeff saw opportunity in me even when I saw little opportunity in myself. Second, Jeff provided me with an ideal working environment which wasn't littered with micromanagement and criticisms, but rather an environment where I was free to approach problems using my own strategies, where creativity was encouraged, and where help was always available if need be. I also feel that Jeff showed me what it's like to work for a great supervisor - a skill that I will carry forward in the event that I end up in a management position. Last, Jeff's lab has set me up to do well in the future: few labs expose students to both the software development environment and the modern laboratory environment. Many thanks Jeff!

I am grateful to my family for putting a roof over my head throughout my graduate career. Time that could have been spent negotiating with roommates, scouring rental properties, etc., was instead spent on a deeper cause: learning and practicing my programming skills. None of the research presented in this thesis could be possible without my family's assistance and for that I will be forever thankful!

I would also like to acknowledge both the SFU Department of Chemistry and the KEY Big Data Initiative, both of which have provided funding for my work.

With special mention to both Drs. Loren Kaake and Vance Williams, who offered to serve on my committee.

Last but certainly not least I would like to acknowledge my lab members. Thanks for putting up with my antics!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Protein Structures

Proteins are biomolecules that perform many tasks within living organisms such as catalysis, transport, stimulus response, recognition and DNA replication. The fundamental units of these macromolecules are amino acids. There are 20 canonical amino acids, but several non-canonical examples are known in natural systems. [1] Protein three-dimensional structures are classified into four distinct levels. The primary structure refers to the linear chain of amino acids held together by peptide bonds, where for n amino acids there are n - 1 peptide bonds. Protein secondary structure refers to the next level of complexity above primary structure. Here, a chain of amino acids can form a simple and regular 3-dimensional structure, usually through self-interactions via hydrogen bonds. The most common secondary structures are the α-helix and the β-pleated sheet. The α-helix refers to the right-handed screw formed as a result of hydrogen bonding between a backbone carbonyl group and the amino group located 3-4 residues downstream. The β-sheet is a pleated sheet of β-strands connected laterally by at least two hydrogen bonds.

In the next level of structure, protein tertiary structure, a secondary structure is rendered more compact through additional folding. Typically, folding is initiated through hydrophobic interactions between different portions of the primary structure. The final protein structure is locked together through interactions such as disulfide bonds, hydrogen bonds, and salt bridges (among others, as described below). Finally, quaternary structure refers to the aggregation of two or more tertiary structures via hydrophobic interactions, hydrogen bonds, salt bridges and other non-covalent interactions that hold together individual polypeptide chains in a higher order, multi-component structure. One common example of a protein quaternary structure is the grouping of four hemoglobin tertiary structures into the functional tetramer. [1]

Different types of interactions lock proteins in distinct 3-dimensional conformations. For example, hydrogen bonding occurs between a hydrogen atom bound to an electronegative atom such as nitrogen, oxygen, or fluorine and another atom containing at least one lone pair of electrons. The partial positive charge on the hydrogen atom is directly attracted to the partial negative charge on the lone pair. There are many examples of hydrogen bonding in proteins: one example is the serine-serine dimer, where a hydrogen bond forms between the oxygen on the terminal OH group of one serine and a hydrogen on the terminal OH group of the other serine residue. Hydrogen bonds also are important features in α-helices and in β-sheets.

In protein tertiary structure, salt bridges frequently contribute to structural stability. Salt bridges are a combination of the two non-covalent hydrogen and ionic bonds. Salt bridges commonly arise between the residues aspartic acid and lysine. An example here includes the interaction between the negatively charged aspartic acid carboxylate (R-COO$^-$) and the positively charged lysine ammonium (R-NH$_3^+$).

In protein tertiary and quaternary structure, there are two additional interactions worth noting in the context of this thesis: cation-π and aromatic interactions. Cation-π interactions are non-covalent attractive interactions between cations and the electron rich faces of conjugated systems. These interactions play a role in protein structure but also in molecular recognition. As an example, the positively charged acetylcholine neurotransmitter binds to a tryptophan π face via a cation-π interaction in the nAChR (nicotinic acetylcholine) receptor. [3] In Fig. 1.1, **C**, a very simple cation-π interaction is shown. But what about aromatic interactions? Strictly speaking, aromatic interactions refer to the interactions of two (or more) conjugated π systems (Fig. 1.1, **D**) in one of several geometries. Aromatic interactions, much like their cation-π kin, serve roles in both molecular recognition and structural stability. A recent paper [4] reports the X-ray crystal structures of the antide-

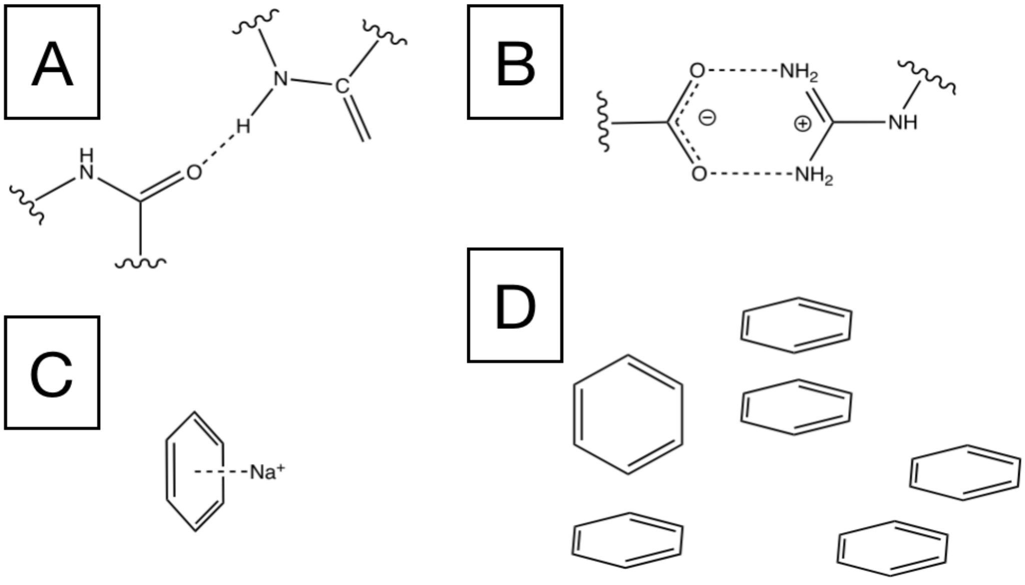


Figure 1.1. Four examples of non-covalent interactions in proteins. [A]: A hydrogen bond. [B]: A salt bridge. [C]: A cation-π interaction. [D]: Aromatic interactions - (*Left*) Edge on- (*Middle*) Stacked- (*Right*) Offset stacked geometry.

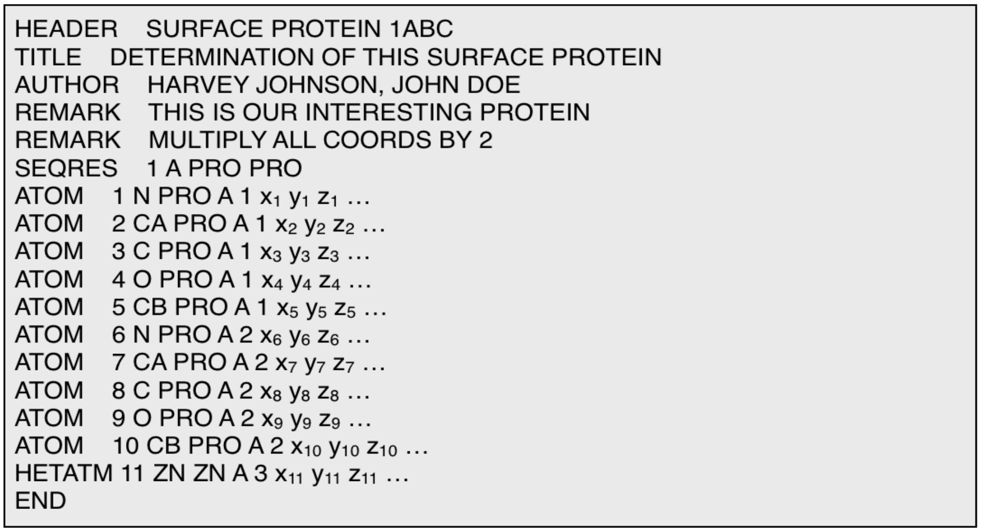
pressants sertraline, fluvoxamine, and paroxetine bound to thermostable variants of the serotonin transporter. In the transporter, all three antidepressants bind via the formation of an offset aromatic interaction. In these aromatic interactions, a pair of aromatic rings has parallel planes and laterally offset centroids (Fig. 1.1, **D**, right).

1.2 The Protein Data Bank

The Protein Data Bank (PDB) is a crystallographic database containing 3-dimensional structural data for many large biological molecules, mainly proteins and nucleic acids. The PDB's beginnings can be traced to 1969 when Edgar Meyer at Texas A&M University began to write computer programs for the storage of such data, with the sponsorship of Walter Hamilton (Brookhaven National Laboratory). [5] The PDB has since seen tremendous growth. The earliest statistics available show 13 structures in the database in 1976, and there are over 140,000 available as of mid-2018. [6]

For the purposes of this thesis, the format of PDB structural files is worth describing. The PDB stores all structural data in the unique .pdb file format. Each PDB entry typically has one associated file with all structural data. Some older structures are replaced with newer data of higher quality, though the original files are still available. A PDB file begins with lines corresponding to HEADER, TITLE and AUTHOR records (Fig. 1.2). These lines provide basic information such as the names of researchers who determined the protein structure and structure determination methods, among others. The REMARK records are lines allocated solely for the purpose of including short, "free form" text. The SEQRES records provide the overall amino acid sequence for the PDB entry. The ATOM records provide 3-dimensional information corresponding to each atom for the protein given in the PDB entry.

Of particular importance to this thesis is the PDB's use of associative arrays to store this structural data. An associative array refers to a collection of key / value pairs where a unique key (such as a student number) identifies a unique value (such as a student's name and gender). Here, unique keys identifying atoms are mapped to $\mathbb{R}^3$ vectors describing individual atomic positions. These keys are organized in SMCRA format (Structure / Model / Chain / Residue / Atom) indicating to which hierarchy each atom in the protein belongs (Fig. 1.3). The HETATM records describe the identity and coordinates of heteroatoms in a protein (atoms that are not directly part of the protein molecule, such as embedded metal ions or other cofactors). The format for each HETATM is identical to that of the ATOM records.



HEADER SURFACE PROTEIN 1ABC
TITLE DETERMINATION OF THIS SURFACE PROTEIN
AUTHOR HARVEY JOHNSON, JOHN DOE
REMARK THIS IS OUR INTERESTING PROTEIN
REMARK MULTIPLY ALL COORDS BY 2
SEQRES 1 A PRO PRO
ATOM 1 N PRO A 1 x_1 y_1 z_1 …
ATOM 2 CA PRO A 1 x_2 y_2 z_2 …
ATOM 3 C PRO A 1 x_3 y_3 z_3 …
ATOM 4 O PRO A 1 x_4 y_4 z_4 …
ATOM 5 CB PRO A 1 x_5 y_5 z_5 …
ATOM 6 N PRO A 2 x_6 y_6 z_6 …
ATOM 7 CA PRO A 2 x_7 y_7 z_7 …
ATOM 8 C PRO A 2 x_8 y_8 z_8 …
ATOM 9 O PRO A 2 x_9 y_9 z_9 …
ATOM 10 CB PRO A 2 x_{10} y_{10} z_{10} …
HETATM 11 ZN ZN A 3 x_{11} y_{11} z_{11} …
END

Figure 1.2. A schematic representation of the PDB file format for the dummy file 1abc.pdb. Researchers "Harvey Johnson" and "John Doe" have crystallized a new metalloprotein consisting of two proline residues and a Zn heteroatom. These researchers have formatted their data in a format compatible with Protein Data Bank standards.

$$\mathbf{S}\text{tructure} = \begin{cases} \mathbf{M}\text{odel 1} & \begin{cases} \mathbf{C}\text{hain A} & \begin{cases} \mathbf{R}\text{esidue} \quad R_1 \begin{cases} \mathbf{A}\text{tom} \quad n+1: & \begin{pmatrix} x_1 & y_1 & z_1 \end{pmatrix} \\ \text{Atom} \quad n+2: & \begin{pmatrix} x_2 & y_2 & z_2 \end{pmatrix} \\ \quad \cdots & \end{cases} \\ \text{Residue} \quad R_2 \quad \cdots \\ \text{Residue} \quad R_3 \quad \cdots \\ \quad \cdots \end{cases} \\ \text{Chain B} & \cdots \\ \text{Chain C} & \cdots \\ \quad \cdots & \end{cases} \\ \text{Model 2} & \cdots \\ \quad \cdots & \end{cases}$$

Figure 1.3. A schematic representation of key organization according to SMCRA hierarchy for coordinates in the PDB.

An example of all of the above features of a .pdb file format is shown in Fig. 1.2. Here, the researchers have organized their data into a format compatible with PDB guidelines. The `SEQRES` record states that the overall protein sequence (i.e., the primary structure) is a PRO-PRO dimer. The `ATOM` records then provide 3-dimensional coordinate data for the PRO-PRO dimer in SMCRA format. Lastly, there is a zinc heteroatom in the protein structure.

1.3 A brief review of the linear algebra used in this thesis

Section 1.2 described associative arrays whose keys are mapped to unique values in the context of PDB structural data. For the PDB, these values refer specifically to 3-dimensional coordinates. Linear algebra is well suited for working with such coordinates. Specifically, the representations of vectors and matrices are important for the work presented in this thesis. Almost all of the work presented here is done from a linear algebraic perspective. Familiarity with linear algebra and computer programming are assumed. The interested reader is directed toward reference [7] for a good overview of linear algebra and reference [8] for an introduction to computer programming. The following paragraphs outline some essential nomenclature and concepts in the context of this thesis.

Vectors are objects that have both magnitude and direction. In the physical sciences, examples of vectors include force and electric or magnetic fields. In this thesis, vectors are used almost exclusively in a geometric context to describe atomic bonds, lone-pair direction and imaginary line segments between chemical moieties. Vector addition and subtraction (equation 1.1) arise frequently in mapping operations, specifically in operations where vectors are moved around in space, such as to the origin of a frame:

$$\vec{u} \pm \vec{v} = (u_1 \pm v_1, u_2 \pm v_2, u_3 \pm v_3) \; . \tag{1.1}$$

The norm, or Euclidean norm, of a vector refers to its magnitude. The norm of a vector is commonly encountered in this thesis for the purpose of determining the physical distance between two atoms. For example, take vector $\vec{u}$ to describe an atom A and vector $\vec{v}$ to describe an atom B. The distance between the two atoms can be found using equation 1.2:

$$\|\vec{u} - \vec{v}\| = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + (u_3 - v_3)^2} \; . \tag{1.2}$$

The dot product arises commonly in more complex operations such as determining the angle between two vectors. The dot product can be found using equation 1.3:

$$\vec{u} \cdot \vec{v} = u_1 v_1 + u_2 v_2 + u_3 v_3 \; . \tag{1.3}$$

The following Chapters discuss the angle between a vector parallel to a methionine δ-sulfur lone pair and an imaginary vector that approximates lone pairs on the sulfur (which are not visible in X-ray structural data). The angle between two vectors can be using equation 1.4:

$$\theta = \cos^{-1}\left(\frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\|\|\vec{v}\|}\right) \; . \tag{1.4}$$

A matrix (**M**, equation 1.5) is a rectangular array of elements (e.g., numbers) and can be thought of as a set of stacked vectors. A square matrix contains an equal number of columns and rows. Associated with every n-dimensional square matrix **M** is an n-dimensional parallelotope. The determinant of $\mathbf{M}$ yields the signed n-dimensional volume of this parallelotope. For example, consider the following definition of $\mathbf{M}$ (where both vectors are linearly independent):

$$\mathbf{M} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} , \tag{1.5}$$

the origin of the frame and the two vectors in $\mathbf{M}$ define three of four vertices of a parallelogram (a 2-parallelotope) in the case of a 2 × 2 matrix. To compute the area of this parallelogram:

$$\begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} = (a_{1,1})(a_{2,2}) - (a_{2,1})(a_{1,2}) , \tag{1.6}$$

the vertical lines in equation 1.6 indicate that the determinant of the 2 × 2 matrix is being taken. The 3-parallelotope associated with a set of three linearly independent vectors is termed the parallelepiped. The work in this thesis relies on the use of the cross product of a set of vectors (in the case of a 3 × 3 matrix) for interpolating the positions of lone pairs, for example. The cross product of two vectors is a third vector orthogonal to the plane containing the two vectors. For example, take a water molecule (of tetrahedral geometry, $C_{2\mathrm{v}}$ symmetry - i.e. a molecule with two mirror planes and a twofold symmetry axis) whose origin is centered at the oxygen atom. Two vectors, $\vec{u}$ and $\vec{v}$, describe the position of the hydrogen atoms, which can be resolved experimentally using X-ray diffraction and related techniques. The lone pairs of electrons are less straightforward to treat. The question is how to find the vector, $\vec{w}$, orthogonal to the plane containing the two hydrogen atoms. This is accomplished by computing the cross product of $\vec{u}$ and $\vec{v}$, as shown in equation (1.8). Here, $\vec{w}$ is coplanar with the lone-pairs, and is thus a crucial element for interpolating lone pair positions in a water molecule.

$$\vec{w} = \vec{u} \times \vec{v} = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} , \tag{1.7}$$

$$\therefore \vec{w} = \hat{\imath}(u_2v_3 - u_3v_2) - \hat{\jmath}(u_1v_3 - u_3v_1) + \hat{k}(u_1v_2 - u_2v_1)\ . \tag{1.8}$$

The vector $\vec{w}$ is important for two reasons. First, $\vec{w}$ specifies the plane containing the vectors $\vec{u}$ and $\vec{v}$. For the work described here, this result is critical for any studies involving studying the orientation of geometric elements relative to a flat aromatic ring. Second, norm $\vec{w}$ (or $\|\vec{w}\|$) yields the area of the quadrilateral whose edges (2 of 4) are the vectors $\vec{v}$ and $\vec{u}$. Correspondingly, the area enclosed by the triangle with vertices (0, 0, 0), $\vec{u}$, and $\vec{v}$ is simply equivalent to $\frac{1}{2}\|\vec{w}\|$. This result is summarized in equation 1.9 (and is used extensively in Chapter 3):

$$a = \frac{1}{2}\|\vec{u} \times \vec{v}\|\ . \tag{1.9}$$

Next, linear transformation matrices are described. These concepts are regularly covered in Inorganic Chemistry courses (e.g., CHEM 230 and 332 at SFU). The best way to describe linear transformation matrices is through an example. One such example is the rotation matrix, which rotates a vector by some angle. Consider the vector $\vec{k} = \begin{bmatrix} a & b & c \end{bmatrix}^{\mathrm{T}}$. $\vec{k}$ can be rotated by $\frac{\pi}{2}$ about the z axis by computing Rot(z, $\frac{\pi}{2}$)$\vec{k}$:

$$\mathrm{Rot}(z, \frac{\pi}{2}) = \begin{bmatrix} \cos(\frac{\pi}{2}) & -\sin(\frac{\pi}{2}) & 0 \\ \sin(\frac{\pi}{2}) & \cos(\frac{\pi}{2}) & 0 \\ 0 & 0 & 1 \end{bmatrix}\ . \tag{1.10}$$

The multiplication operation and result are given in equation 1.11:

$$a \begin{bmatrix} \cos(\frac{\pi}{2}) \\ \sin(\frac{\pi}{2}) \\ 0 \end{bmatrix} + b \begin{bmatrix} -\sin(\frac{\pi}{2}) \\ \cos(\frac{\pi}{2}) \\ 0 \end{bmatrix} + c \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = a \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + b \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} + c \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -b \\ a \\ c \end{bmatrix}\ . \tag{1.11}$$

Transformation matrices are used extensively in the computer graphics industry for modeling the motions of humanoid figures and the like. One transformation matrix used almost exclusively for rendering such motions is the homogeneous transformation matrix, which both rotates and translates a vector in 3-dimensional space. The homogeneous transformation is an example of an *affine transformation*, a transformation that preserves points, straight lines, and planes. There are many forms of the homogeneous transform. For brevity, the homogeneous transformation matrix $\mathbf{T}$ for both a rotation of θ about the z axis and a translation by a, b, c in the x, y, z direction is given in equation 1.12:

$$\mathbf{T} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 & a \\ \sin(\theta) & \cos(\theta) & 0 & b \\ 0 & 0 & 1 & c \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{1.12}$$

Note that $\mathbf{T}$ is a 4×4 matrix. A w vector has been concatenated to $\mathbf{T}$, where $w = \begin{bmatrix} a & b & c & 1 \end{bmatrix}^{\mathrm{T}}$. The w vector is necessary for the translation component of the homogeneous transform. The operations needed to first rotate $\vec{v}$ by θ about the z axis and then translate $\vec{v}$ by a, b, c in the x, y, z direction are shown in equation 1.13. In this thesis, the homogeneous transformation matrix is used to physically move chemical moieties in 3-dimensional space to generate models for density functional theory calculations, in a process analogous to moving a humanoid figure in a video rendering. This is described in more detail in Chapter 2.

$$\mathbf{T}\vec{v} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 & a \\ \sin(\theta) & \cos(\theta) & 0 & b \\ 0 & 0 & 1 & c \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_x \\ v_y \\ v_z \\ 1 \end{bmatrix}. \tag{1.13}$$

1.4 Networks

In Chapter 4, extensive reference is made to networks. But what exactly is a network and how is a network described mathematically? A network, in its strictest sense, is a relationship between discrete objects. Real-world examples of networks are roads connecting cities or the Translink Skytrain System in the B.C. Lower Mainland. For the latter example, the discrete units (i.e., Skytrain stations) are termed *nodes* and the connections (i.e., rail tracks) are termed *edges*. A network of nodes and edges is commonly mapped to a *finite graph* which is described using an *adjacency matrix*. Consider Fig. 1.4.

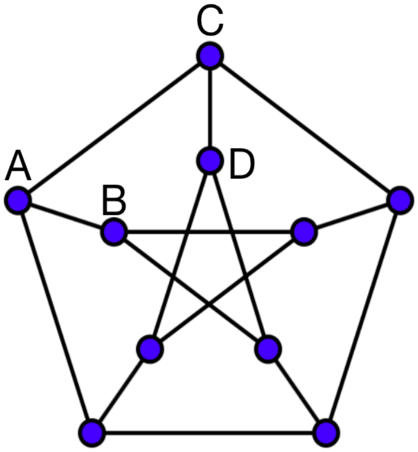


Figure 1.4. The Petersen graph. The Petersen graph contains a total of 10 nodes with 15 edges. The Petersen graph is a member of the cubic family of graphs. All nodes in a cubic graph project connections to three other nodes, that is, all nodes have a degree of 3. Adapted from R. A. Nonenmacher. [9]

In Fig. 1.4, there are 4 nodes labelled A, B, C and D. The connections between these four nodes can be described using adjacency matrix M_A (1.14):

$$M_A = \begin{array}{cc} \begin{array}{cccc} \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} \end{array} & \\ \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} & \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \\ \mathrm{D} \end{array} \end{array} . \tag{1.14}$$

In M_A, the rows and columns correspond to the series of nodes A, B, C, D, and a Boolean true is assigned if a node at a particular column connects to a node at a particular row, and vice versa. More specifically, M_A will always be an $n \times n$ logical matrix with n equivalent to the number of nodes in our network. In this thesis, all networks are assumed to be undirected, and as such, all adjacency matrices M_A will be symmetric. Adjacency matrices are critical because computer architectures do not yet understand graphs and diagrams. Simply stated, a computer cannot “look” at Fig. 1.4 above and determine that nodes A and C are connected, though this limitation may not be the case in a few years time. A computer is, however, well suited for working with linear algebra such as adjacency matrices.

A second matrix that is used extensively in this thesis is the degree matrix. The degree matrix yields information about the number of edges connected to each node. In the case of the subgraph obtained from the Petersen graph, the degree matrix (M_D) is given in equation 1.15. Here, $M_{D(1,1)}$ (i.e., node A) is receiving one connection from node B, one connection from node C, and a connection from an unknown node originating from the bottom left of the Petersen graph. Therefore, $M_{D(1,1)}$ has a degree of 3:

$$M_D = \begin{array}{cccccc} & \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} & \\ \left(\right. & 3 & 0 & 0 & 0 & \left.\right) \mathrm{A} \\ & 0 & 3 & 0 & 0 & \mathrm{B} \\ & 0 & 0 & 3 & 0 & \mathrm{C} \\ & 0 & 0 & 0 & 3 & \mathrm{D} \end{array} . \tag{1.15}$$

In recent years, it has been observed that networks arise frequently in the physical sciences [10] and applications of network science to protein science are well established [11, 12, 13]. In this thesis, proteins are treated as 3-dimensional finite graphs, whose nodes are either atoms or entire amino acids and whose edges are vectors. By simplifying the finite graphs to contain only residues of interest, and assigning connection vectors under specific constraints, we can begin to study the geometric relationships between aromatic residues from a computational point of view, such as how many discrete closely spaced Tyr-Trp residues ("chains") a protein contains and the length of such chains.

The above idea segues into an important mathematical property. We can obtain the Laplacian matrix L (equations 1.14, 1.15) from M_D and M_A as follows:

$$L = M_D - M_A \, , \tag{1.16}$$

Or:

$$L = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \tag{1.17}$$

which yields:

$$L = \begin{bmatrix} 3 & -1 & -1 & 0 \\ -1 & 3 & 0 & 0 \\ -1 & 0 & 3 & -1 \\ 0 & 0 & -1 & 3 \end{bmatrix}. \tag{1.18}$$

The Laplacian matrix L contains critical information about a graph. The algebraic multiplicity of the 0 eigenvalue of L yields the number of disconnected units in a graph. Here, a disconnected unit is defined as a network that is not in any way connected to another network in some graph. The vector of eigenvalues of L is $\vec{\lambda} = \begin{bmatrix} 4.618 & 1.382 & 3.618 & 2.382 \end{bmatrix}^{\mathrm{T}}$. Note that $\vec{\lambda}$ contains no 0 eigenvalues. Therefore, there exist no disconnected components in the Petersen subgraph shown in Fig. 1.4. This important property is applied in Chapter 4. The interested reader is directed towards reference [14] which provides an excellent overview of network theory applications to a wide variety of disciplines including chemistry and physics.

1.5 Research motivation

I hypothesized that the interaction between methionine and the aromatic residues plays a role in electron transfer and I used coordinate data in the PDB alongside Big Data approaches to evaluate this hypothesis. To start with the evaluation, I assumed that a bioinformatics screen of the Protein Data Bank would demonstrate that many methionine-aromatic interaction to redox-active transition metal distances would be short if the methionine-aromatic interaction indeed played a role in mediating electron transfer. To answer this question, I first needed to find methionine-aromatic interactions in proteins. To do so, I developed a novel (and published, [15]) linear algebraic sorting algorithm for finding these pairs. This algorithm, named *Met-aromatic*, is described in **Chapter 2**. I then iterated the Met-aromatic algorithm in parallel with another custom program for finding transition metals over all oxidoreductases present in the Protein Data Bank in 2017 in order to collect these distance data. The details and results of this bioinformatics survey are presented in **Chapter 3**.

A critical oversimplification of the Met-aromatic algorithm serendipitously led to the discovery of a new geometry that involves a Met and more than one aromatic group. We termed these novel motifs “bridging interactions" and carried out a comprehensive screen of all structures in the PDB to further probe the significance of these bridging interactions. As part of this work, I modified the Met-aromatic interaction with bounding conditions to select for only aromatic:methionine:aromatic interactions (our custom nomenclature for a methionine paired with two aromatic residues). These efforts are described in **Chapter 4** where I applied my bounded Met-aromatic variant to 139,948 protein structures in the PDB.

The data science methods employed in this thesis have led to many questions. Why do bridging interactions occur in proteins? Have proteins evolved these interactions for some reason? What are their physical properties, either in redox reactions or in protein structure? My colleagues in the Warren lab are working on experimental probes of these questions to complement the bioinformatics work presented here. To assist the Warren lab in addressing questions about Met-aromatic groups, I developed a computer program for operating a home-built nanosecond fluorescence / transient absorption (TA) spectrometer. The construction notes for this system are presented in **Chapter 5**. My ultimate goal is that the Warren lab will make use of this software to validate the many questions that have arisen as a result of my data analysis using real protein models.

Bibliography

[1] Stubbe, J., van der Donk, WA. Protein Radicals in Enzyme Catalysis. *Chem Rev.* **98**, (1998). 705-762

[2] Branden, C., Tooze, J. (1999) *Introduction to Protein Structure.* New York, NY, Garland Publishing Inc.

[3] Zhong, W., J. P. Gallivan, Y. Zhang, L. Li, H. A. Lester, and D. A. Dougherty. From *Ab Initio* Quantum Mechanics to Molecular Neurobiology: A Cation-π Binding Site in the Nicotinic Receptor. *Proc Natl Acad Sci.* **95**, (1998). 12088-93

[4] Coleman, Jonathan A., and Eric Gouaux. Structural Basis for Recognition of Diverse Antidepressants by the Human Serotonin Transporter. *Nat Structl Mol Biol.* **25**, (2018). 170-75

[5] Meyer, EF. The first years of the Protein Data Bank. *Protein Sci.* **6**, (1997). 1591-97

[6] Anonymous. Hard data: it has been no small feat for the Protein Data Bank to stay relevant for 100,000 structures. *Nature.* **509**, (2014). 260

[7] Kolter, Z. (2008, October 7), Linear Algebra Review and Reference. Retrieved from http://www.cs.cmu.edu/~zkolter/course/15-884/linalg-review.pdf

[8] Miller, B., Ranum, D. (2013), How to Think Like a Computer Scientist. Retrieved from http://interactivepython.org/runestone/static/thinkcspy/index.html

[9] Nonenmacher, R. A. (2008), Heawood graph. Accessed Oct. 6, 2018. Retrieved from https://en.wikipedia.org/wiki/File:Heawood_Graph.svg

[10] Palla, G., Derenyi, I, Farkas, I., Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature.* **435**, (2018). 814-818

[11] Yan, Yan, Shenggui Zhang, and Fang-Xiang Wu. Applications of Graph Theory in Protein Structure Identification. *Proteome Sci.* **9**, (2011). S17

[12] Vishveshwara, Saraswathi, K. V. Brinda, and N. Kannan. PROTEIN STRUCTURE: INSIGHTS FROM GRAPH THEORY. *J Theor and Comput Chem.* **01**, (July 2002). 187-211

[13] Canutescu, Adrian A., Andrew A. Shelenkov, and Roland L. Dunbrack. A Graph-Theory Algorithm for Rapid Protein Side-Chain Prediction. *Protein Sci.* **12**, (2003). 2001-14

[14] Baez, J. (2018) Network Theory. Retrieved from http://math.ucr.edu/home/baez/networks/

[15] D. S. Weber, J. J. Warren. A survey of methionine-aromatic interaction geometries in the oxidoreductase class of enzymes: What could Met-aromatic interactions be doing near metal sites? *J Inorg Biochem.* **186**, (2018). 34-41

Chapter 2

The Met-aromatic algorithm and order of interaction

2.1 Preface

I was previously interested in statistically quantifying the orientations of methionine δ-sulfur lone pairs and proline α-carbon / α-hydrogen bond axes relative to the aromatic rings present on the residues tyrosine, tryptophan, and phenylalanine. This work led to the development of two algorithms for finding such pairs that I named "Pro-aromatic" and "Met-aromatic", respectively. The focus of this thesis is on the Met-aromatic algorithm. However, many of the same concepts are applied in the Pro-aromatic algorithm, and applications to related searches to the PDB should be straightforward. This Chapter describes the mathematical and programming features of the Met-aromatic algorithm. More detailed examples and discussion are given in Chapters 3 and 4.

2.2 What is an "interaction?"

In the 1980s, very early informatics studies [1] of the PDB revealed a tendency of methionine δ-sulfur atoms to localize near the aromatic residues Phe, Tyr, and Trp. This localization was termed an "interaction." Later studies, when many more structures were available, confirmed those early findings and refined the model. [2] Nonetheless, an interaction was still defined simply as a localization between Met-sulfur and any of the atoms of the aromatic residue. The earliest iteration of the Met-aromatic algorithm used this simple model, but analysis of the output data showed many outliers (e.g., instances where Met-sulfur was

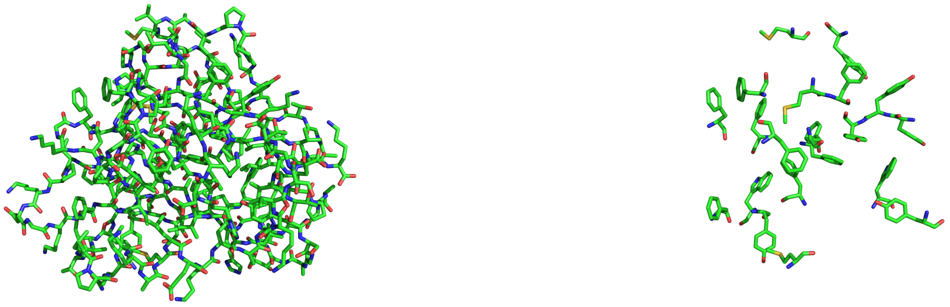


Figure 2.1. Left: Feature space ***P1***. The raw, unrefined set of residues for the PDB entry 1RCY. *Right*: Feature space ***P2***. ***P1*** has been refined down to only those residues containing an aromatic moiety or methionine.

pointed away from the aromatic residue). These results suggested that the definition of "interaction" could be further refined. Consequently, both a distance component (as done in the literature) and an angular component were applied. The latter component can be used to distinguish those interactions where the Met-sulfur points directly toward or away from the aromatic ring. The development of the final, refined algorithm is given in the following sections.

2.3 Creating and refining a feature space

To find methionine-aromatic interactions in a protein, a feature space must first be generated. This $\mathbb{R}^n$ space contains the set of features of interest. In the case of a protein structure, the feature space is simply the $\mathbb{R}^3$ space containing all atomic coordinates in a PDB file (Fig. 2.1, left). However, the focus here is only four residues, so the feature space can be greatly simplified to contain only coordinates for atoms in any of methionine, tyrosine, tryptophan and phenylalanine (Fig. 2.1, right). Further, each of these residues contains atomic coordinates that are unnecessary for this study. Therefore the feature space is narrowed down to contain only methionine SD (δ-sulfur), CE (ϵ-carbon), and CG (γ-carbon) coordinates, in addition to any coordinates describing the positions of the six members of an aromatic ring in the aforementioned residues. This greatly refined feature space is now ready for processing using Met-aromatic.

2.4 The distance condition

The first step in the algorithm ensures that a methionine residue is physically near an aromatic residue. This is the distance condition of Met-aromatic. To apply the distance condition, a vector $\vec{v}$ projects from all methionine δ-sulfur coordinates to midpoints between any two bonded aromatic carbon atoms. Midpoints are pre-computed using a vectorized approach (2.1). In 2.1, the six row vectors describe the positions of aromatic carbon atoms in any of Phe, Tyr or Trp:

$$\begin{bmatrix} x_1^* & y_1^* & z_1^* \\ x_2^* & y_2^* & z_2^* \\ x_3^* & y_3^* & z_3^* \\ x_4^* & y_4^* & z_4^* \\ x_5^* & y_5^* & z_5^* \\ x_6^* & y_6^* & z_6^* \end{bmatrix} = \frac{1}{2}\left\{ \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \\ x_4 & y_4 & z_4 \\ x_5 & y_5 & z_5 \\ x_6 & y_6 & z_6 \end{bmatrix} + \begin{bmatrix} x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \\ x_4 & y_4 & z_4 \\ x_5 & y_5 & z_5 \\ x_6 & y_6 & z_6 \\ x_1 & y_1 & z_1 \end{bmatrix} \right\}, \tag{2.1}$$

therefore $\vec{v} = \begin{bmatrix} x_n^* & y_n^* & z_n^* \end{bmatrix} - \begin{bmatrix} SD_x & SD_y & SD_z \end{bmatrix}$. Midpoints were chosen owing to the maximum of electron density observed in the region of space between two aromatic carbon atoms. There exist $(n)(6m)$ possible vectors $\vec{v}$ for every feature space, where n is the number of methionine residues in our feature space and m the number of aromatic residues. The m is scaled by a factor of 6 owing to the number of possible midpoints in a hexagonal arrangement of aromatic carbon atoms. Aromatic residues distant to a methionine residue are discarded by eliminating residue pairs where the Euclidean norm of $\vec{v}$ exceeds some distance cutoff (e.g., 4.9 Å in Chapter 3 or 6.0 Å in Chapter 4) (Fig. 2.2).

2.5 The angular condition

In contrast to past studies, an angular condition was introduced here. Having refined the feature space to include only properly spaced methionine-aromatic pairs, an angular selection criterion involving sulfur lone pairs and aromatic π systems was imposed. However, one important issue must be addressed before proceeding: the location of the lone pairs on the

Figure 2.2. Left: Feature space ***P2*** from Fig 2.1. *Right*: Feature space ***P3***. ***P2*** has been refined down to only those residues where methionine is physically near an aromatic. Here Met18 is physically near Tyr122 and Met148 is physically near Phe54.

methionine sulfur must be modeled, as Protein Data Bank entries do not contain information about the geometry of lone pairs present on heteroatoms, as is the case for most routine crystallographic data. To circumvent this problem, the position of lone pairs from existing data was estimated. The CE and CG atoms, included in the refinement of the feature space, are bound directly to the SD atom in a methionine residue. Consequently, the positions of the lone pairs are estimated by considering CE and CG atoms to be vertices of a regular tetrahedron with origin at SD. Note that there are alternative means of completing the last two vertices of a tetrahedron given two input vertices and the origin. For completeness, the method used in this thesis is outlined in detail in the following paragraph. A regular tetrahedron with a water-like structure is C_{2v} symmetric. Consider a regular tetrahedron with origin O and vertices A, B, C, and D: the triangle AOB will lie on a plane orthogonal to an equivalent triangle COD. Given that O, A and B are known, the positions of C and D can be computed by rotating $O - A$ and $O - B$ by $\frac{1}{2}\pi$ about the vector $\frac{1}{2}\Big[A + B\Big]$. An identical procedure is used to estimate the positions of lone pairs projecting from an SD coordinate. First the vectors antiparallel to the CE / SD and CG / SD bonds are found:

$$\vec{a'} = \text{SD} - \text{CE}\,, \tag{2.2}$$

$$\vec{g'} = \text{SD} - \text{CG}\,, \tag{2.3}$$

in addition to determining the axis of rotation (the real-valued eigenvector) $\vec{k}$:

$$\vec{k} = \frac{1}{2}(\vec{a} + \vec{g}) \,. \tag{2.4}$$

Note however that the two vectors above, $\vec{a'}$ and $\vec{g'}$ are coplanar with the CG and CE coordinates. Therefore, to find the positions of the lone pairs, the vectors $\vec{a'}$ and $\vec{g'}$ are rotated about a general axis $\vec{k}$ by $\frac{1}{2}\pi$ using Rodrigues' rotation formula. The Rodrigues' rotation formula (commonly used in the study of robotics) requires the computation of a matrix K for the *unit* direction $\hat{k}$,

$$K = \begin{bmatrix} 0 & -\frac{k_z}{||\vec{k}||} & \frac{k_y}{||\vec{k}||} \\ \frac{k_z}{||\vec{k}||} & 0 & -\frac{k_x}{||\vec{k}||} \\ -\frac{k_y}{||\vec{k}||} & \frac{k_x}{||\vec{k}||} & 0 \end{bmatrix} . \tag{2.5}$$

The rotation matrix R is computed from matrix K. R revolves a vector about any general axis by any angle. In this case, vectors are rotated about the line of intersection between the two planes describing all four vertices in a molecule of C_{2v} symmetry, as in equation 2.6

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \sin(\frac{\pi}{2})K + (1 - \cos(\frac{\pi}{2}))K^2 \,. \tag{2.6}$$

The vectors describing the lone pairs are given in equations 2.7 and 2.8

$$\vec{a} = R\vec{a'} \,, \tag{2.7}$$

$$\vec{g} = R\vec{g'} \,. \tag{2.8}$$

Section 2.4 introduced vector $\vec{v}$ in the context of the distance condition, but it also is important in defining the angular condition, specifically in the study of the orientation of lone pairs relative to an aromatic π system. Two angles can be found by mapping $\vec{v}$ to the

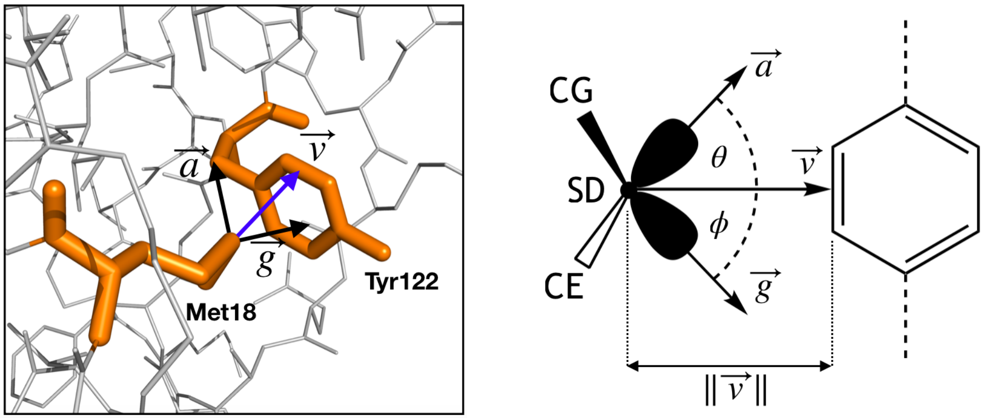


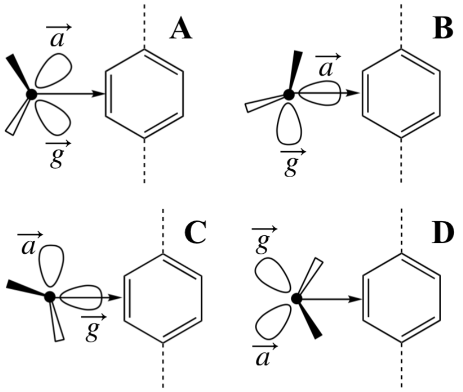
Figure 2.3. Left: Feature space ***P3*** from Fig. 2.2. Here the Met18/Tyr122 interaction has been isolated. The distance cutoff has been set to 6.0 Å. In this case, the Euclidean norm of the first vector $\vec{v}$ is 4.211 Å. Recall that $\vec{v}$ from a methionine SD coordinate to a midpoint between two aromatic carbon atoms. In this case the norm $\leq$ 6.0 Å, meeting the Met-aromatic distance criteria. Here Met-$\theta = 75.766°$ and Met-$\phi = 64.317°$. *Right*: A ChemDraw schematic representation of the geometric elements upon which Met-aromatic constraints are imposed.

origin of the feature space to which the previous $\vec{a'}$ and $\vec{g'}$ vectors were mapped: $\vec{a}\angle\vec{v}$ and $\vec{g}\angle\vec{v}$. These angles are termed Met-θ and Met-ϕ, respectively, which can be found using equations 2.9 and 2.10

$$\text{Met} - \theta = \cos^{-1}\frac{\vec{a}\cdot\vec{v}}{\|\vec{a}\|\|\vec{v}\|} \tag{2.9}$$

$$\text{Met} - \phi = \cos^{-1}\frac{\vec{g}\cdot\vec{v}}{\|\vec{g}\|\|\vec{v}\|} \tag{2.10}$$

The angular condition of Met-aromatic judges that a methionine residue is interacting with a conjugated π system *if and only if* (Met-$\theta \leq 109.5°$ **or** Met-$\phi \leq 109.5°$) where 109.5° refers to the angle between any two vectors projecting from the origin to vertices in a regular tetrahedron. Note that **or** is an inclusive operator (as compared to **xor**, the exclusive **or** operator). This is an important point and means that a methionine residue is also considered interacting if (Met-$\theta \leq 109.5°$ **and** Met-$\phi \leq 109.5°$). The Boolean logic of the angular condition is depicted graphically in Fig. 2.4.



$$\begin{bmatrix} \text{A} & : & \theta \leq c & \phi \leq c & : & \text{True} \\ \text{B} & : & \theta \leq c & \phi > c & : & \text{True} \\ \text{C} & : & \theta > c & \phi \leq c & : & \text{True} \\ \text{D} & : & \theta > c & \phi > c & : & \text{False} \end{bmatrix} \tag{2.11}$$

Figure 2.4. The four methionine poses leading to four possible logic states. Vector $\vec{v}$ is shown as a black arrow pointing to the right in each pose. *Right*: the logic matrix associated with the four poses. Note that poses A-C are deemed interacting pairs according to Met-aromatic criteria. These poses maximize the interaction of lone pairs with a region of electron-rich space between two aromatic carbon atoms. Pose D minimizes this interaction as both lone pairs point away from the aromatic edge. c is equivalent to some arbitrary cutoff. The interaction in feature space ***P3***, Fig. 2.3, is equivalent to pose A.

2.6 Comments on the algorithm

The above paragraphs describe the elements of the Met-aromatic algorithm, but not the function of the algorithm as a whole. The following paragraphs address the related questions about the Met-aromatic code developed in this thesis, i.e., what are some of the algorithm's properties and why are they important in terms of the research presented in this thesis? The architecture of the program is perhaps best laid out as a more compact, human readable form known as pseudocode:

Algorithm 1 MetAromatic algorithm. rp = residue position number. atomID = atomic ID such as CG, SD or CE. resID = residue identifier such as Tyr, Trp or Phe. MET array does not require a resID member as all members are methionine residues.

```
procedure MetAromatic
    MET = [ {$rp_m$, atomID} : {$x_{met}, y_{met}, z_{met}$} ]   ▷ Array of MET CG, SD and CE coordinates
    ARO = [ {$rp_a$, resID, atomID} : {$x_{aro}*, y_{aro}*, z_{aro}*$} ]   ▷ Aromatic carbon midpoints
    RESULTS = [ ]   ▷ Array for storing results
    for $i \in$ groupby(MET[$rp_a$]) do   ▷ groupby() yields subgroups for each MET residue
        $\vec{a}$, $\vec{g}$ = rodrigues($i$.SD, $i$.CG, $i$.CE)
        for $j \in$ ARO do
            $\vec{v} = \begin{bmatrix} j.x_{aro}* & j.y_{aro}* & j.z_{aro}* \end{bmatrix} - i.\text{SD}$   ▷ SD is the origin of all frames in algorithm
            if $\|\vec{v}\| \leq c$ then   ▷ $c$ is some cutoff distance such as 6.0 Å
                Met-$\theta$, Met-$\phi$ = $\cos^{-1}\frac{\vec{a}\cdot\vec{v}}{\|\vec{a}\|\|\vec{v}\|}$, $\cos^{-1}\frac{\vec{g}\cdot\vec{v}}{\|\vec{g}\|\|\vec{v}\|}$
                if Met-$\theta \leq \chi$ or Met-$\phi \leq \chi$ then   ▷ $\chi$ is some cutoff angle such as 109.5°
                    add {$rp_m$, $rp_a$, resID}:{$\|\vec{v}\|$, Met-$\theta$, Met-$\phi$} to RESULTS array
```

Algorithm 1 is the basic scaffold of the program that executes the Met-aromatic algorithm on a protein structure. Note that an analysis of the code wrapping the Met-aromatic algorithm reveals hundreds of lines. This supplementary code is termed overhead and is necessary to connect to the Protein Data Bank, fetch a file, clean up the input data, clean up the output, and export the data, among other tasks. The average time taken for the Met-aromatic algorithm to execute on a protein structure is of quadratic, or $O(n^2)$ time complexity. Time complexity broadly refers to the amount of time an algorithm takes to complete a task as a function of the size n of its input. While the microperformance of the Met-aromatic algorithm is solely of quadratic time complexity, the macroperformance of the entire program wrapping the algorithm is generally linear owing to a host of different variables such as network bit rate (internet speed), background processes in any computer system, and the basic hardware components on which the program is executed. These are important considerations because the Protein Data Bank contains 140,000 protein structures as of 2018. As such, the overhead performance is critical to ensure a data mining operation is completed in a meaningful timeframe. Next one must consider the microperformance of the base algorithm, Met-aromatic, in this case. The choice of algorithm can make the difference between a solution being found in a matter of milliseconds and a matter of hours. As an example one may compare time taken to analyze a protein consisting of n number of amino acids using an $O(n^2)$ algorithm vs. an $O(n!)$ algorithm.

2.7 Physicochemical significance of Met-aromatic

This Chapter has so far described Met-aromatic's mathematical underpinnings, the algorithm as a whole, and briefly discussed the overhead wrapping the algorithm. At this point, the reader may be asking "Why is this algorithm significant?" in the context of a chemistry thesis. The answer ties directly into the physical properties of methionine-aromatic motifs. For this work, the most relevant property is redox reactivity. Others have found that these interactions are important in protein structures, but little mention is made of redox reactions, in which the Warren group is interested. The potential significance is best explained by example. Consider the structure of calmodulin (PDB ID 1CFC). The Met-aromatic al-

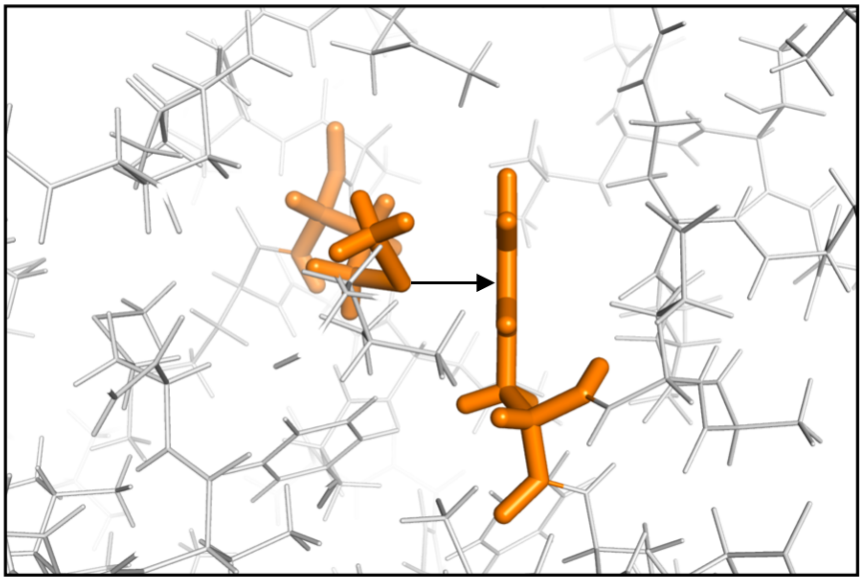


Figure 2.5. A PyMOL rendering for the aromatic interaction between Met72 and Phe12 for calmodulin (PDB ID 1CFC). A vector points from Met72 δ-sulfur to the centroid of the Phe12 aromatic ring.

gorithm reveals an aromatic interaction between Phe12 and Met72 (Fig 2.5). Using this as a starting point, a very basic Density Functional Theory (DFT) calculation can be used to probe how the Met-Phe interaction changes when the Met-Phe distance is changed.

In this computation, a methionine CG-SD-CE moiety and a phenylalanine aromatic ring were isolated. All other coordinate data were discarded using a custom Python program. The Python program then computed a vector $\vec{n}$ normal to the plane containing the aromatic ring and translated the CG-SD-CE structure towards and away from the aromatic plane along the normal vector. The homogenous transformation matrices described in Chapter 1 were used to translate the coordinates of the CG-SD-CE moiety. Here $\hat{n}$ was found from $\vec{n}$. $\hat{n}$ was then scaled by a translational increment t to yield a set of translations. This scaled vector, $t\hat{n}$ served as the w vector $t\begin{bmatrix}\hat{n}_x & \hat{n}_y & \hat{n}_z & 1\end{bmatrix}^{\mathrm{T}}$ which was concatenated into the homogeneous transform $\mathbf{T}$ (2.12):

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 & t\hat{n}_x \\ 0 & 1 & 0 & t\hat{n}_y \\ 0 & 0 & 1 & t\hat{n}_z \\ 0 & 0 & 0 & 1 \end{bmatrix} . \tag{2.12}$$

The resulting transformation for various values of t was obtained using equation 2.13:

$$\mathbf{T}_t = \begin{bmatrix} 1 & 0 & 0 & t\hat{n}_x \\ 0 & 1 & 0 & t\hat{n}_y \\ 0 & 0 & 1 & t\hat{n}_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathrm{CG}_x & \mathrm{SD}_x & \mathrm{CE}_x \\ \mathrm{CG}_y & \mathrm{SD}_y & \mathrm{CE}_y \\ \mathrm{CG}_z & \mathrm{SD}_z & \mathrm{CE}_z \\ 1 & 1 & 1 \end{bmatrix} , \tag{2.13}$$

where $t \in \{t_1, t_2, \ldots, t_n\}$. All $\mathbf{T}_t$ coordinate data was exported in .xyz file format and the resulting .xyz files were visualized using Avogadro. [3] Avogadro was also used to add coordinates for the hydrogen atoms. The resulting system was reduced to a dimethyl sulfide molecule affinely and systematically translated to and from a benzene ring plane (Fig. 2.6). DFT calculations were used to assess the relative energy changes to the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) as a function of Met-Phe distance. The ORCA software suite was used for these calculations at the BP/def2-SVP level of theory. [4] Critically, it was found that translating towards the benzene ring raises the HOMO energy and lowers the LUMO energy. This trend is qualitatively similar to the repulsive portion of the Lennard-Jones model (Fig. 2.8). The Lennard-Jones model provides an approximation for the potential energy between a pair of neutral-state atoms/molecules spaced some distance r apart [6]. $V_{LJ}(r)$ attains a minimum at some distance r_m. Compressing a pair of atoms/molecules a distance below r_m induces a rapid increase in the potential energy of the system. Conversely, separating two atoms/molecules a distance exceeding r_m gradually forces the system to zero potential energy (equation 2.16), Fig. 2.7.

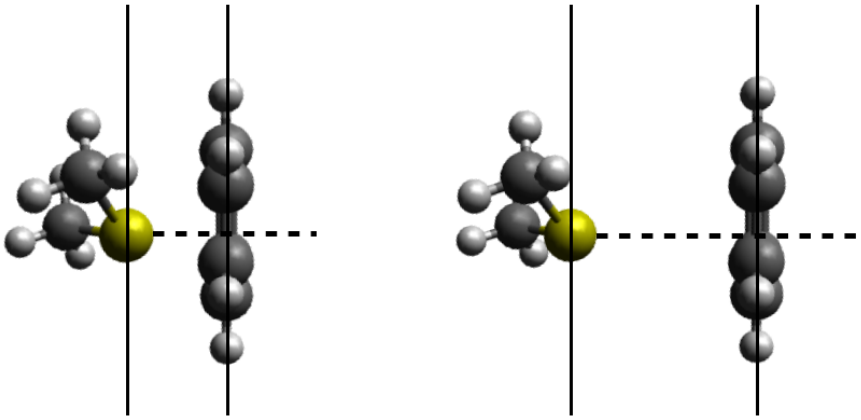


Figure 2.6. An Avogadro rendering of the two extreme ends of our translation series. *Left*: +0.75 Å translation towards the benzene ring. *Right*: -1.00 Å translation away from the benzene ring. The 0.00 Å position refers to the native spacing in the protein. The dashed line is parallel to the w vector in homogeneous transform $\mathbf{T}$.

$$\lim_{r \to 0} \epsilon[(\frac{r_m}{r})^{12} - 2(\frac{r_m}{r})^{6}] = \infty \tag{2.14}$$

$$\lim_{r \to r_m} \epsilon[(\frac{r_m}{r})^{12} - 2(\frac{r_m}{r})^{6}] = -\epsilon \tag{2.15}$$

$$\lim_{r \to \infty} \epsilon[(\frac{r_m}{r})^{12} - 2(\frac{r_m}{r})^{6}] = 0 \tag{2.16}$$

The result presented in Fig. 2.8 fits with expectations from basic chemical principles; forcing two repulsive systems together will always result in an increase in potential energy. However, from a different perspective a question arises: why would proteins evolve the position of two repulsive systems proximal to each other? A methionine-aromatic system meeting the Met-aromatic criteria is not necessarily a favourable conformation. The dispersion forces (or hydrophobic interactions) are thought to add less than 1.5 kcal mol^{-1} [2, 5] to overall protein stability. Such interactions may be the difference between a folded and an unfolded protein under normal conditions. The motivating hypothesis here posits that there could be another, underappreciated, role for Met-aromatic groups. Raising the potential energy of a methionine-tyrosine or methionine-tryptophan pair makes either the tyrosine or tryptophan participant more easily oxidized. This is especially the case where the methionine-aromatic interaction is proximal to a species whose LUMO energy remains

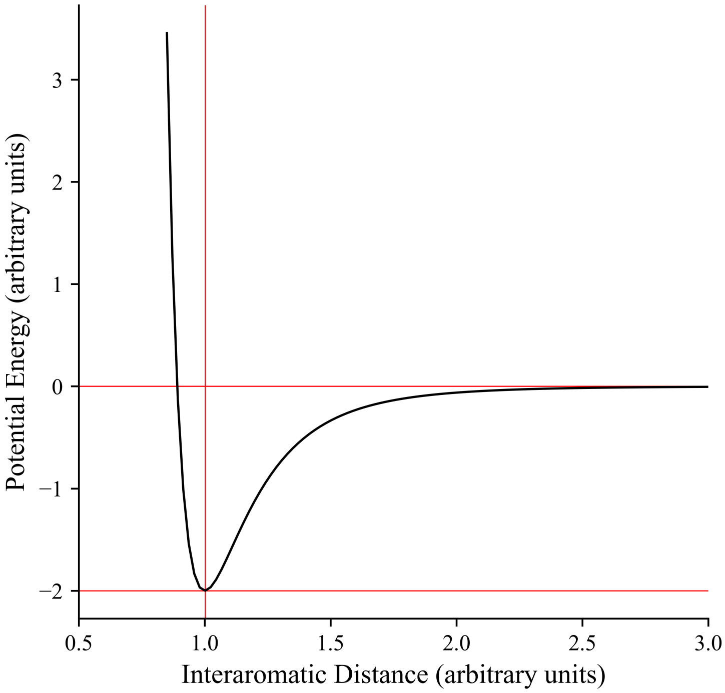


Figure 2.7. An example Lennard-Jones potential plotted using $V_{LJ}(r) = \epsilon[(\frac{r_m}{r})^{12} - 2(\frac{r_m}{r})^6]$ $\text{argmin}_r\, V_{LJ}(r) = r_m = 1.00$. The potential energy between two atoms/molecules is minimized at this distance.

static (regardless of how the translational increment t changes), such as a transition metal. From this perspective, one can begin considering the possibility that proteins have evolved the formation of methionine-aromatic interactions to participate in redox reactions. Chapters 3 and 4 explore this question in more detail.

2.8 Standalone software design and order of interaction

During my time as a graduate student in the Warren lab I had the opportunity to interact and collaborate with colleagues working on experimental protein biochemistry. Parenthetically, one of my early efforts in the Warren lab involved protein expression and purification, but ultimately I found the bioinformatics work presented here to be more tractable. My collaborations and experience in the wet lab taught me important lessons about the importance of user friendliness of any protocol, theoretical or experimental. To greatly simplify the work of students working on experimental aspects of methionine-aromatic interactions in the Warren lab, I spent some time programming a Tkinter user interface that automates

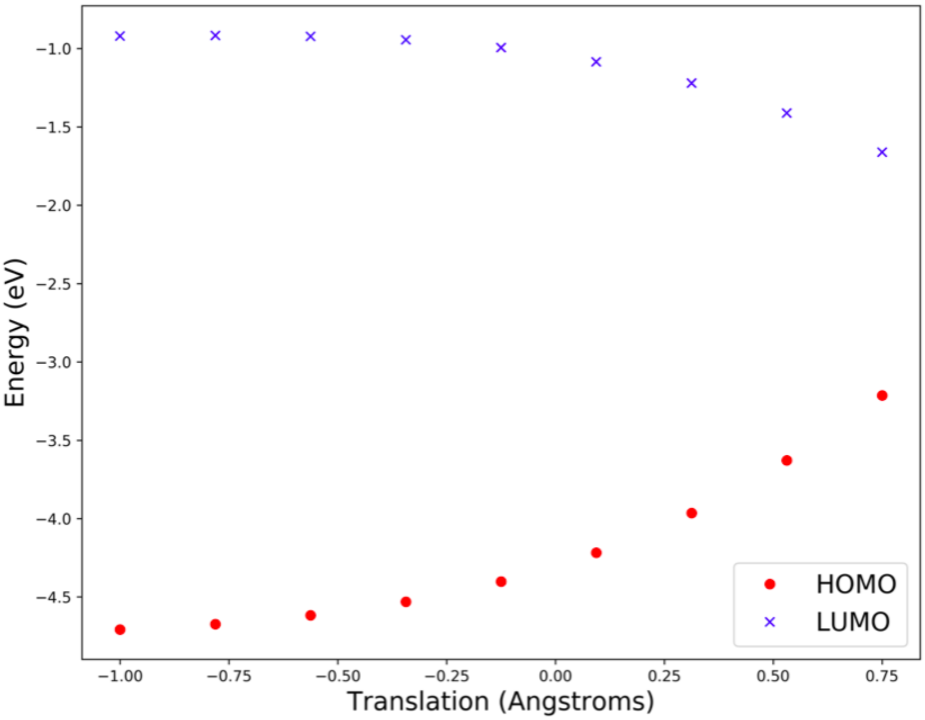
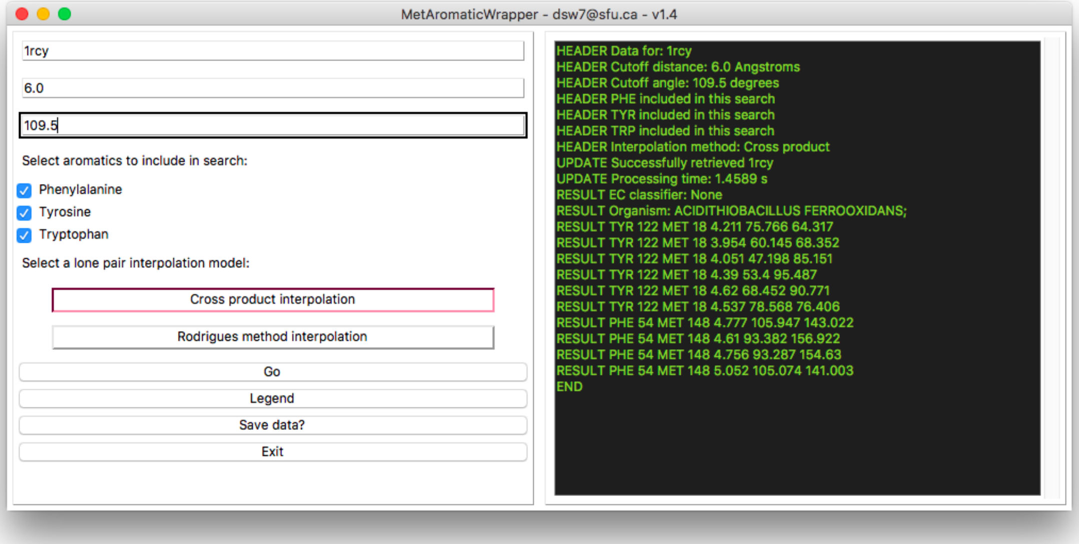


Figure 2.8. Translation of the dimethylsulfide (DMS) approximation in Fig. 2.6 towards the plane containing a benzene molecule. The translation from -1.00 Å to +0.75 Å is in accordance with the scheme depicted in 2.6. An increase in HOMO energy is observed as the DMS approaches the ring plane. The trend presented here is **qualitatively** similar to the repulsive portion of the Lennard-Jones function (see Fig. 2.7).



MetAromaticWrapper - dsw7@sfu.ca - v1.4

1rcy

6.0

109.5

Select aromatics to include in search:

☑ Phenylalanine
☑ Tyrosine
☑ Tryptophan

Select a lone pair interpolation model:

```
HEADER Data for: 1rcy
HEADER Cutoff distance: 6.0 Angstroms
HEADER Cutoff angle: 109.5 degrees
HEADER PHE included in this search
HEADER TYR included in this search
HEADER TRP included in this search
HEADER Interpolation method: Cross product
UPDATE Successfully retrieved 1rcy
UPDATE Processing time: 1.4589 s
RESULT EC classifier: None
RESULT Organism: ACIDITHIOBACILLUS FERROOXIDANS;
RESULT TYR 122 MET 18 4.211 75.766 64.317
RESULT TYR 122 MET 18 3.954 60.145 68.352
RESULT TYR 122 MET 18 4.051 47.198 85.151
RESULT TYR 122 MET 18 4.39 53.4 95.487
RESULT TYR 122 MET 18 4.62 68.452 90.771
RESULT TYR 122 MET 18 4.537 78.568 76.406
RESULT PHE 54 MET 148 4.777 105.947 143.022
RESULT PHE 54 MET 148 4.61 93.382 156.922
RESULT PHE 54 MET 148 4.756 93.287 154.63
RESULT PHE 54 MET 148 5.052 105.074 141.003
END
```

Figure 2.9. Version 1.4 of a Tkinter Met-aromatic user interface wrapper. This particular version has been optimized for use with the MacOSX platform. Here the UI is running the Met-aromatic algorithm on the protein 1RCY from Fig. 2.2. Note that the first `RESULT` line matches exactly the geometric values in Fig. 2.3.

the Met-aromatic algorithm. The user interface and output for analysis of rusticyanin (PDB ID 1RCY) is shown in Figure 2.9).

The results for analysis of rusticyanin (Fig. 2.9) allow for the discussion of one additional point about the Met-aromatic algorithm. Examination of the output in Fig. 2.9 shows six output lines for the Met18/Tyr122 interaction, but only four output lines for the Met148/Phe54 interaction. Recall that a methionine δ-sulfur can project six vectors $\vec{v}$ to six midpoints on a six-membered aromatic ring. However, not all six vectors $\vec{v}$ always meet the conditions of the Met-aromatic algorithm. In the case of the Met148/Phe54 pair, only four of six vectors $\vec{v}$ meet Met-aromatic criteria (Fig. 2.10). The number of vectors $\vec{v}$ projecting from a methionine δ-sulfur and meeting Met-aromatic criteria was termed "order of interaction". Analysis of the order of interaction led to a critical result, which is presented in Chapter 4. The Met-aromatic interaction in Fig. 2.10 is an order-IV interaction, where maximum order of interaction is VI (for six total midpoints on a six-membered ring).

The significance of order of interaction classification lies in the ability to statistically quantify geometries (potentially a computationally complex task), where an order-I inter-

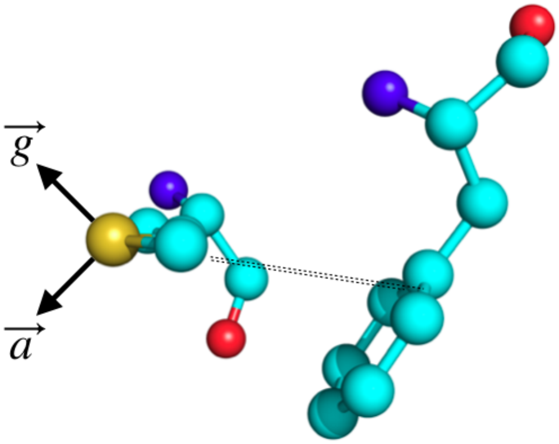


Figure 2.10. 1RCY: Met-ϕ and Met-θ both exceed 109.5° for the two dashed lines projecting from the Met148 δ-sulfur to midpoints between Phe54 CG/CD1 and Phe54 CG/CD2. These two lines do not meet the Met-aromatic angular condition.

action is indicative of a side on lone pair / aromatic interaction, as opposed to an order-VI interaction, which suggests that a lone pair is pointing directly into an aromatic centroid (analogous to a "ball and socket" joint). Interactions of orders I-VI are shown in 2.11. The order of interaction is found by simply counting the number of unique `MET` residue position numbers in the prompt. Formally, this is somewhat similar to cardinality, where in SQL (Structured Query Language) cardinality refers to the "uniqueness" of values in a column in a structured database [7] e.g., Fig. 2.12.

2.9 Conclusions

This Chapter outlined the key aspects of the Met-aromatic algorithm. The work in this thesis starts from the historically narrow definition of "interaction" and expands on that definition to include an angular condition. The approach set out here is applied to two different problems in Chapters 3 and 4. In addition, the development of a simplified user interface will allow other workers to easily analyze protein structures for Met-aromatic interactions. The code also is straightforward to apply to other types of interactions, such as Cysteine-aromatic or Proline-aromatic.

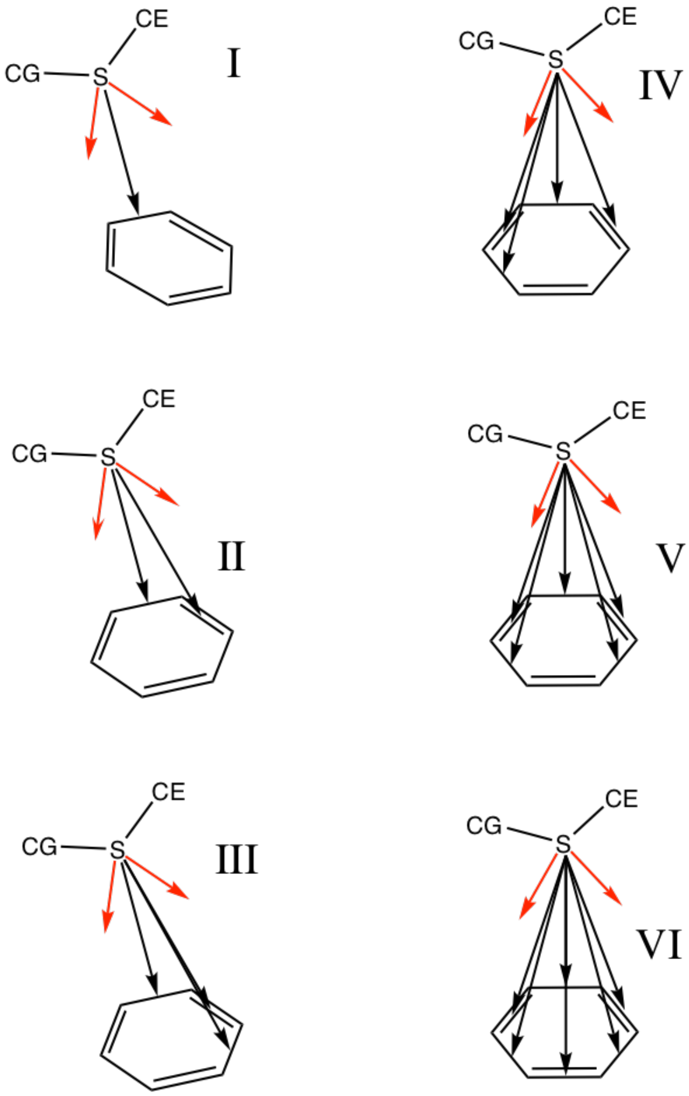
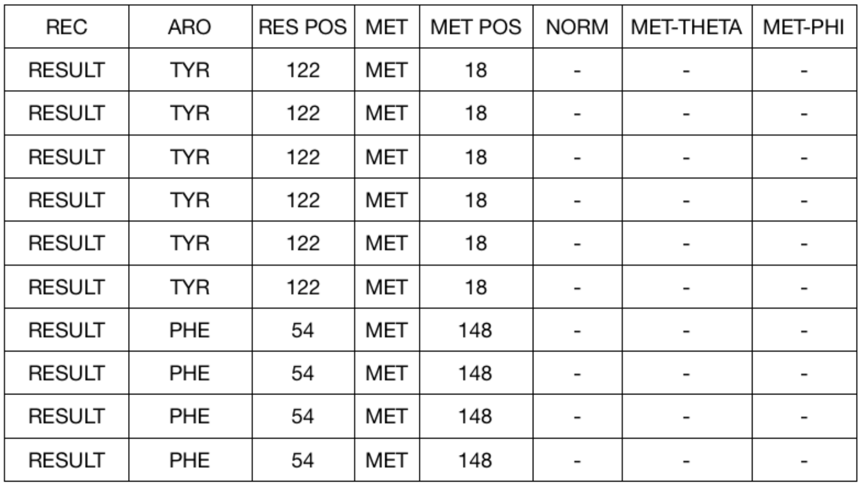


Figure 2.11. The order of interaction classification. An order-I interaction indicates that one vector $\vec{v}$ (shown in black) is shorter than some cutoff distance (i.e. 4.9 Å) and either Met-θ or Met-ϕ is $\leq$ 109.5°. An order-II interaction indicates that two of six vectors meet Met-aromatic criteria. Likewise, an order-VI interaction indicates that all six vectors $\vec{v}$ projecting from a methionine sulfur to aromatic midpoints meet Met-aromatic criteria. Vectors $\vec{a}$ and $\vec{g}$ are shown in red.



REC	ARO	RES POS	MET	MET POS	NORM	MET-THETA	MET-PHI
RESULT	TYR	122	MET	18	-	-	-
RESULT	TYR	122	MET	18	-	-	-
RESULT	TYR	122	MET	18	-	-	-
RESULT	TYR	122	MET	18	-	-	-
RESULT	TYR	122	MET	18	-	-	-
RESULT	TYR	122	MET	18	-	-	-
RESULT	PHE	54	MET	148	-	-	-
RESULT	PHE	54	MET	148	-	-	-
RESULT	PHE	54	MET	148	-	-	-
RESULT	PHE	54	MET	148	-	-	-

Figure 2.12. A tabular output obtained from the user interface in Fig. 2.9 for 1RCY. A table is the simplest example of a structured database. There are a total of six `MET POS` lines containing the residue position number 18 (an order-VI interaction) and a total of four `MET POS` lines containing the residue position number 148 (an order-IV interaction).

Bibliography

[1] Reid, KSC., Lindley, PF., Thornton, JM. *FEBS Lett.* **190**, (1985). 209-213

[2] Valley, CC., Cembran, A., Perlmutter, JD., Lewis, AK., Labello, NP., Gao, J., Sachs, JN. *J Biol Chem.* **287**, (2012). 34979-34991

[3] Hanwell, MD., Curtis, DE., Lonie, DC., Vandermeersch, T., Zurek, E., Hutchison, GR. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J Cheminform.* **4**, (2012). 17

[4] Neese, F. The ORCA program system. *WIREs Comput Mol Sci.* **2**, (2011). 73-78

[5] Waters, ML. *Biopolymers.* **76**, (2004). 435-445

[6] Lennard-Jones, J. *Proc R Soc Lond A.* **738**, (1924). 463

[7] Crooks, E. (2013, September 26). Why low cardinality indexes negatively impact performance. Retrieved from https://www.ibm.com/developerworks/data/library/techarticle/dm-1309cardinal/

Chapter 3

Surveying the Protein Data Bank: What Could Met-Aromatic Interactions be Doing Near Metal Sites?

3.1 Preface

This chapter describes work published in: *J. Inorg. Biochem.* **2018**, *186*, 34-41 (DOI: 10.1016/j.jinorgbio.2018.05.008). This work was initially motivated by the idea that methionine (Met)-aromatic interactions could play a role in protein electron transfer reactions. The initial survey started with oxidoreductase enzymes (International Union of Biochemistry and Molecular Biology classifier EC 1), which play roles in many different redox reactions. In addition to collecting the methionine-aromatic interactions, coordinates and identities of transition metals also were collected. The geometric relationships between these metal redox centres and Met-aromatic interactions were evaluated to address the hypothesis that Nature evolved the physical placement of a methionine-aromatic interaction proximal to a metal centre to facilitate or prevent electron transfer.

3.2 Introduction

The intramolecular forces that hold proteins in their native states and promote specific peptide or ligand binding interactions have long been of great interest. [1] Fundamental concepts in protein folding have been established and several modern theoretical treatments can describe features of protein structure at early times, and at later times, during the folding process. [2, 3] In addition, the last 20 years have seen the emergence of a wealth of data about protein structure. [4] Since 1997, yearly submissions to the Protein Data Bank (PDB) have increased by almost a factor of 10 and the total number of structures available has increased 20-fold. The growth of this database has allowed for the development of new ideas about how protein structures form and function. [3, 6, 7] For instance, analysis of PDB data reveals a great many trends in spatial arrangement of amino acids relative to each other, including H-bonding, salt bridges, π stacking, cation π, and other interactions of aromatic residues. Of particular interest in this thesis are interactions of the aromatic amino acids phenylalanine (Phe), tyrosine (Tyr) and tryptophan (Trp). Some notable examples involve interactions between one of those residues and proline (Pro) [8], alanine (Ala), [9] or methionine (Met). [10]. Here, we use a statistical analysis of structural data to explore whether or not Met-aromatic interactions could be involved in metalloprotein redox reactions.

The aromatic amino acids Tyr and Trp play central roles in protein electron-transfer chemistry. [11, 12] Canonical examples include the functions of photosystem II, [13] DNA photolyase, [14] and ribonucleotide reductase. [15] Oxidized amino acids, such as Tyr and Trp, also are discussed in the context of oxidative stress. [16] The biological redox reactions required in many of the above examples can involve hole/electron transport exceeding distances over 20 Å. Single-step ET, even at high driving force, cannot deliver holes/electrons fast enough to active sites that require the input of electrons to catalyze reactions on timescales shorter than milliseconds. [17] A common bypass to single step ET is the presence of intermediate redox moieties, often Tyr or Trp, that can promote multistep ET (or hopping). [11, 18] Along with those functional roles, a recent proposal suggests that chains of closely spaced Tyr/Trp residues could act as hole conduits stretching from metallocofactors to protein surfaces. [19] In such a case, hole transport through these Tyr/Trp chains is proposed to provide a protective mechanism when metal sites are activated in the absence of substrate (or in cases of other oxidative malfunction). This idea is supported by theory and experiment. [20] All of the above examples are central to biological functions. In this context, it is remarkable that two residues (Tyr and Trp) can carry out such an array of functions in vivo. This also leads to the natural question of how the microenvironment of an amino acid aids in steering electrons through peptides.

Pioneering work, when fewer than 100 structures were in the PDB, demonstrated a propensity for Met to localize its sulfur atom near the aromatic ring of Phe, Tyr, or Trp in a survey of X-ray structures. [10, 21] Since those initial studies, this localization has been formally termed an “interaction,” where sulfur-aromatic distances of under about 5 Å are considered the most significant. More recent analyses of PDB data showed that about $\frac{1}{3}$ of all structures contain at least one Met-aromatic interaction. [22] Other informatics surveys of PDB data support the structural importance of Met-aromatic interactions (e.g., in membrane proteins [23]) and highlight that other residues, such as histidine, can interact with Met. [24] Related surveys are available for Cys-aromatic interactions. [25, 26] Other literature supports the idea that Met-aromatic motifs are associated with protein stability/folding. [27, 28] Structurally, the Met-aromatic interaction is estimated to add an additional stabilization energy of 1-1.5 kcal mol^{-1} in comparison to a purely hydrophobic interaction. [22, 25]

The structural roles of the methionine aromatic interaction are well established, but the single electron redox chemistry of Met is much less clear. [29] The most common Met redox reactions involve 2 electron oxidation to Met-sulfoxide, which has been implicated as an antioxidant and a marker for oxidative stress. [30, 31, 32] However, when Met sulfur atoms interact with an aromatic residue, they tend to be less prone to such oxidation. [33] A contrasting report highlights how the oxidation state of a Met-sulfur can influence its interactions with other aromatic amino acids, with oxidized Met sulfoxide interacting more strongly. [34] Computational studies of sulfur-aromatic interactions in amino acid fragments [35] and in other model compounds [36] demonstrate the importance of sulfur-aromatic interactions in modifying the redox behavior of both sulfur and the aromatic residue. In particular, studies of models containing Met and an aromatic ring within close proximity underscore the attractive two-center, three-electron bond that forms between an oxidized sulfur and a nearby π system. [36] In this study, a survey of protein structural data was undertaken to address the question of the importance of Met-aromatic interactions in modulating electron transfer (ET) reactions in metalloproteins. The work presented here analyzes PDB data for selected redox proteins to identify and classify Met-aromatic interactions and their relationship with metal sites.

3.3 Overview of experimental methods

A Met-aromatic wrapper (Chapter 2) was written in the Python programming language (Python 3.5). The Python programming language was chosen owing to its general simplicity and readability, in addition to the extensive number of libraries and other open source projects affiliated with the Python Software Foundation. In contrast to the user interface presented in Chapter 2, all scripting/data engineering for this project was done at a high level using the Python Pandas Data Analysis Library (`www.pandas.pydata.org`) and to a lesser extent using the Python Numpy package (`www.numpy.org`). All plots were constructed using the Matplotlib Python 2D plotting library (`https://matplotlib.org`).

The Pandas Met-aromatic wrapper iterated over 12,186 oxidoreductase structures (available 15 January 2017) obtained from the Protein Data Bank. The script was equipped with a method of fetching data corresponding to the list of PDB codes from the PDB server (`www.rcsb.org/pdb`) via FTP. The imported data were stripped to include only the 'A' labelled chain and the first model in multi-model entries. This simplified feature space was further refined and subsequently processed using the Met-aromatic algorithm. Transition metal identities and coordinates were collected alongside methionine-aromatic interaction data ($\|\vec{v}\|$, angles, etc). The output data of the Met-aromatic algorithm were banked locally as .csv files for each iteration. In retrospect, a more efficient approach to this problem would have been to dynamically upload files to a SQL database. Once the directory was loaded with the Met-aromatic output data and transition metals, those data were compressed into one master .csv file (identical to a SQL Table) and this master .csv file was used for all downstream analysis.

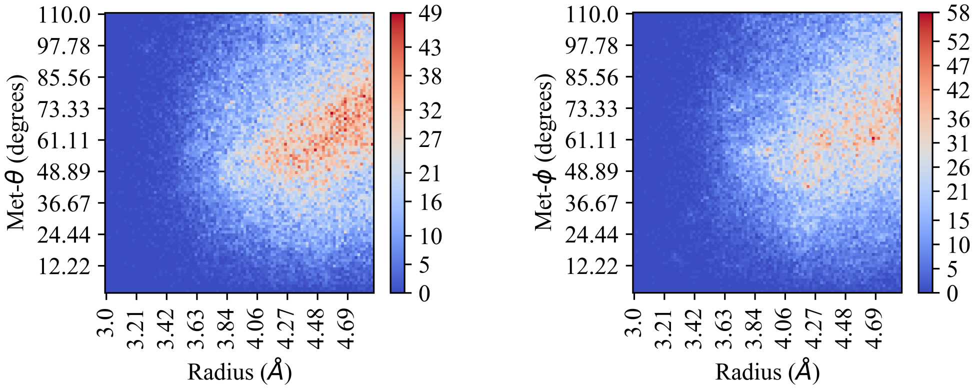


Figure 3.1. Heat maps relating the angle between a Met δ-sulfur lone pair and the vector $\vec{v}$ to the magnitude of $\vec{v}$ (the distance between a Met δ-sulfur and a midpoint between two aromatic carbon atoms). *Left*: The y-axis depicts the angle Met-θ (angle between vectors $\vec{a}$ and $\vec{v}$). *Right*: The y-axis depicts the angle Met-ϕ (angle between vectors $\vec{g}$ and $\vec{v}$). The z-axis depicts counts for vectors $\vec{v}$ between specific distance / angle values.

3.4 Benchmarking incidences and distributions of Met-aromatic geometries in oxidoreductases

Out of 12,186 oxidoreductases, a total of 52,618 interactions meeting Met-aromatic criteria were found. Of this total, 14,272 (27%) were determined to be Met-Tyr interactions, 29,455 (56%) Met-Phe interactions and 8,891 (17%) Met-Trp interactions. Overall, the high incidence of Met-aromatic interactions, with an average of >1 interaction per protein, is consistent with similar literature surveys of the PDB. [22] Most interactions had a methionine-aromatic distance of over 4 Å and a lone pair to $\vec{v}$ angular preference of about 60° (Fig. 3.1). The latter is consistent with other reported angular preferences, i.e., 30° to 60° using similar mathematical methods. [22] Interestingly, as the sulfur-to-ring distance decreased, so did the angle, from about 70° to 50° (Fig. 3.1).

3.5 Order of interaction

A novel observation that emerged during analysis of the Met-aromatic data was the propensity for methionine-aromatic interactions to preferentially assume an "order-II" interaction geometry (see Chapter 1). The next highest incidence involves order-III interaction, defined by three aromatic midpoints. These interactions position at least one sulfur lone pair prox-

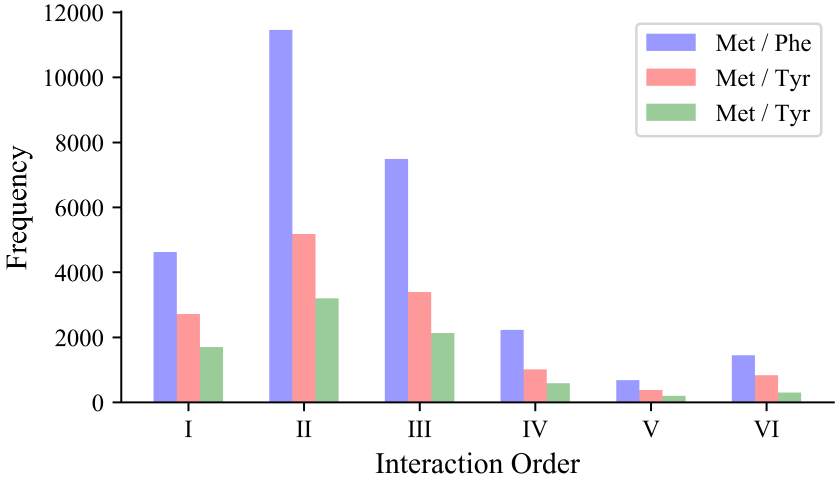


Figure 3.2. A breakdown of counts for all oxidoreductase methionine-aromatic interactions classified by interaction order.

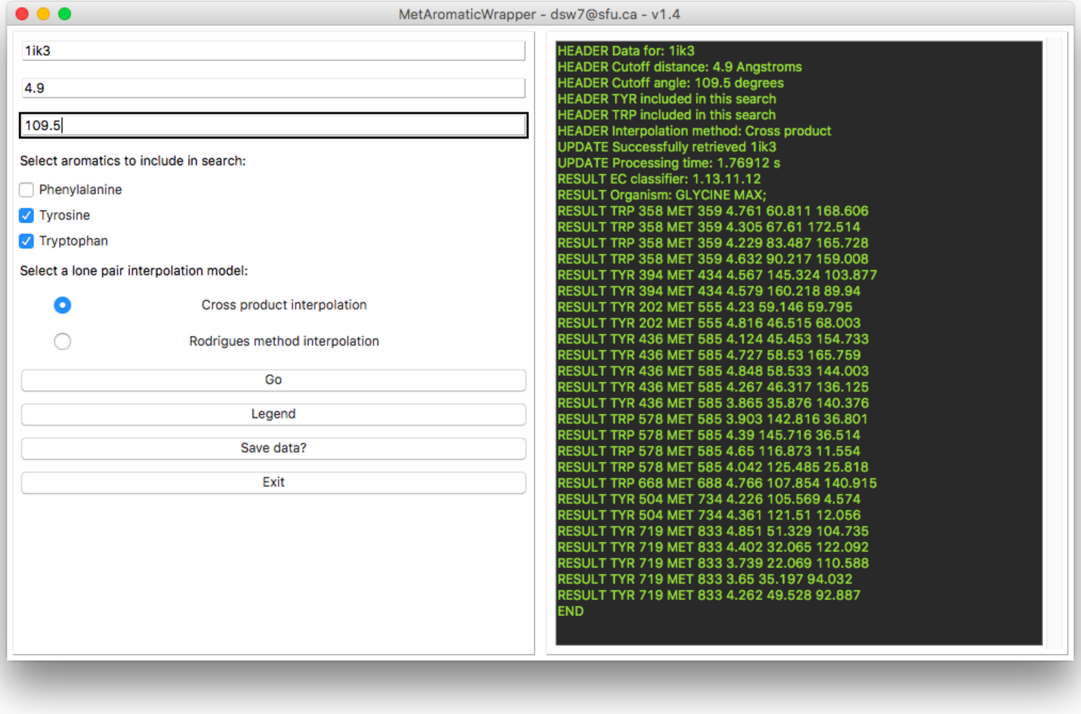
imal to the aromatic plane. All counts separated by interaction order are shown in Fig. 3.2.

3.6 The bridging interaction

In Chapter 2, Section 2.6 the pseudocode describing Met-aromatic (**Algorithm 1**) was presented. In an attempt to save on computational costs, it was assumed that: "There could exist no more than six vectors $\vec{v}$ between a methionine δ-sulfur and the midpoints of a six-membered aromatic ring for some short cutoff distance c." Implicit in that assumption was that Nature would impose this condition upon the Met-aromatic algorithm owing to the fact that a six-membered aromatic system could have a total of six midpoints between adjacent carbon atoms. However, this assumption proved to be incorrect. As an example (Fig. 3.3), analysis of lipoxygenase from soybeans (PDB ID 1IK3) using the Met-aromatic user interface presented in Chapter 2, Section 2.8 was carried out.

In Fig. 3.3., there exist seven unique instances of the methionine residue. An analysis of the cardinality of the `MET POS` attribute reveals what was an unexpected and puzzling observation.

The results in Fig. 3.4 suggest that there exists an order-9 interaction. Indeed, in Fig. 3.3, Met-aromatic found that 9 vectors $\vec{v}$ project from Met585. This result initially suggested



1ik3

4.9

109.5

Select aromatics to include in search:

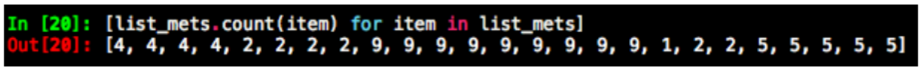
□ Phenylalanine
☑ Tyrosine
☑ Tryptophan

Select a lone pair interpolation model:

● Cross product interpolation
○ Rodrigues method interpolation

```
HEADER Data for: 1ik3
HEADER Cutoff distance: 4.9 Angstroms
HEADER Cutoff angle: 109.5 degrees
HEADER TYR included in this search
HEADER TRP included in this search
HEADER Interpolation method: Cross product
UPDATE Successfully retrieved 1ik3
UPDATE Processing time: 1.76912 s
RESULT EC classifier: 1.13.11.12
RESULT Organism: GLYCINE MAX;
RESULT TRP 358 MET 359 4.761 60.811 168.606
RESULT TRP 358 MET 359 4.305 67.61 172.514
RESULT TRP 358 MET 359 4.229 83.487 165.728
RESULT TRP 358 MET 359 4.632 90.217 159.008
RESULT TYR 394 MET 434 4.567 145.324 103.877
RESULT TYR 394 MET 434 4.579 160.218 89.94
RESULT TYR 202 MET 555 4.23 59.146 59.795
RESULT TYR 202 MET 555 4.816 46.515 68.003
RESULT TYR 436 MET 585 4.124 45.453 154.733
RESULT TYR 436 MET 585 4.727 58.53 165.759
RESULT TYR 436 MET 585 4.848 58.533 144.003
RESULT TYR 436 MET 585 4.267 46.317 136.125
RESULT TYR 436 MET 585 3.865 35.876 140.376
RESULT TRP 578 MET 585 3.903 142.816 36.801
RESULT TRP 578 MET 585 4.39 145.716 36.514
RESULT TRP 578 MET 585 4.65 116.873 11.554
RESULT TRP 578 MET 585 4.042 125.485 25.818
RESULT TRP 668 MET 688 4.766 107.854 140.915
RESULT TYR 504 MET 734 4.226 105.569 4.574
RESULT TYR 504 MET 734 4.361 121.51 12.056
RESULT TYR 719 MET 833 4.851 51.329 104.735
RESULT TYR 719 MET 833 4.402 32.065 122.092
RESULT TYR 719 MET 833 3.739 22.069 110.588
RESULT TYR 719 MET 833 3.65 35.197 94.032
RESULT TYR 719 MET 833 4.262 49.528 92.887
END
```

Figure 3.3. The PDB entry 1IK3 is analyzed using the Tkinter Met-aromatic GUI, v1.4 (the fourth version developed). A cutoff distance c of 4.9 Å and a cutoff angle of 109.5° (i.e. Met-θ or Met-$\phi \leq 109.5°$) were chosen. Note that here phenylalanine was excluded from the search.



```
In [20]: [list_mets.count(item) for item in list_mets]
Out[20]: [4, 4, 4, 4, 2, 2, 2, 2, 9, 9, 9, 9, 9, 9, 9, 9, 9, 1, 2, 2, 5, 5, 5, 5, 5]
```

Figure 3.4. The "Wow!" line. `list_mets` is a Python list containing the residue positions of methionines meeting Met-aromatic criteria for an aromatic interaction. A count of each residue position in `MET POS` was expected to yield a maximum of 6 (order-VI), but here we have 9. This finding ended up revealing an important new methionine-aromatic geometry. I named this the "Wow!" line based on a narrowband radio signal received by the Ohio State University's Big Ear radio telescope on August 15th, 1977. The signal was found to originate from the Sagittarius constellation and was believed to be evidence of extraterrestrial life. The scientist who found the signal named it the "Wow! signal."

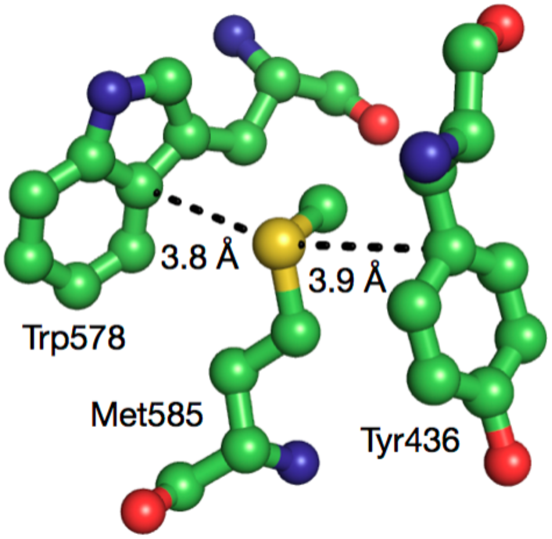


Figure 3.5. The bridging interaction. The presence of two (or more) aromatic residues proximal a methionine residue can yield an interaction of order exceeding VI. In this case, we have a Trp578 : Met585 : Tyr436 bridge from the lipoxygenase 1IK3. This data matches data in the prompt (Fig. 3.3).

a critical error in the algorithm; there are only six midpoints on a six-membered ring. One must realize, however, that Met-aromatic is a naïve algorithm in that it does not care for how many vectors $\vec{v}$ project from a methionine δ-sulfur as long as these vectors meet both angular and distance conditions. Consequently, the only instance where the order of interaction can be greater than VI is if there are two (or more) closely associated aromatic groups. This was termed a "bridging interaction" (Fig. 3.5).

The bridging interaction shown in Fig. 3.5 was termed a dual-bridging interaction (two aromatics paired with one methionine residue). Of 52,618 methionine-aromatic interactions meeting Met-aromatic, it was found that 34% (17,748) involved two different aromatics pairing with one methionine residue. Likewise, triple- and quadruple bridging interactions were identified, but at much lower frequencies of 4,515 (9%) and 1,080 (2%). A similar screen of 4,900 randomly selected PDB structures (including some oxidoreductases) was performed and it was found that ca. 35% of all identified Met-aromatic interactions were involved in a dual-bridging interaction. The importance of such bridging interactions, and what role (i.e., structural, redox) they could play in macromolecules, is not yet clear, but the incidence exceeding 30% suggests this is a motif deserving of further research. The bridging interaction is explored in greater detail in Chapter 4.

3.7 Implications for theoretical analyses

Recent computational analyses of sulfur-aromatic motifs have included interaction geometries where the sulfur atom is positioned directly over the aromatic ring, though other geometries have been probed as well. [22, 35] The observation that order I, II, and III interaction types predominate in proteins suggest that future studies could explore these arrangements. Likewise, the high number of bridging interactions found could be an interesting motif to explore using theory. Bridging interactions that involve redox-active and redox-inactive aromatic groups were identified, so investigation could focus on bridging interaction electronic structure and/or protein stability contribution.

3.8 Statistics of distances between transition metals and aromatic interactions in metal-containing oxidoreductases

The mean distance between any metal ion (structural or catalytic) and any methionine-aromatic (Phe, Tyr and Trp) interaction was determined to be 33 ± 42 Å (one standard deviation). Here, the mean distance was assessed without the removal of any very long-range interactions. Counts under a 60 Å cutoff value comprise 91% of the total counts and the mode of this distribution was found between 17.0 and 18.0 Å. The mean of the distribution with the 60 Å cutoff was found to be 23 ± 11 Å. Those residues that are far from metal sites are less likely to play any direct roles in redox reactions involving a metal site, so the high average value for the metal to methionine-aromatic distance supports the idea that many methionine-aromatic interactions play a role in stabilizing protein structures among other potential roles.

The distances between metal sites and specific methionine-aromatic motifs (of any interaction order) in the subset of oxidoreductases that contain a metal ion were calculated as one metric to assess the potential importance of methionine-aromatic interactions in redox reactions. Metals are often associated with radical reactions of amino acids. A total of

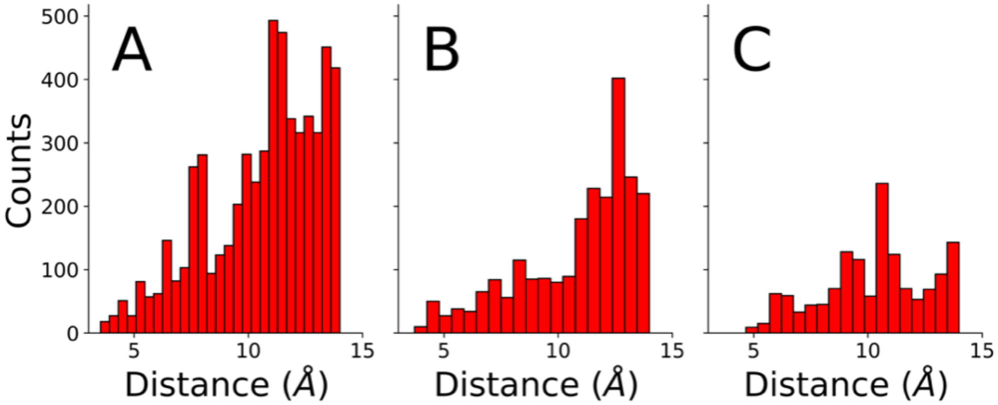


Figure 3.6. A breakdown of counts for distances between all metal ions and Met-aromatic interactions in oxidoreductases, where A = Met-Phe; B = Met-Tyr; C = Met-Trp. The optimal bin width was chosen based on the Freedman-Diaconis rule. An arbitrary cutoff of 14 Å was selected.

54,324 metal-to-aromatic (CE2) distances were enumerated. Of those, 30,827 (57%) were for metal/Met-Phe interactions, 13,740 (25%) for metal/Met-Tyr interactions, and 9,757 (18%) for metal/Met-Trp containing interactions. It should be noted that these proportions are similar, but not identical, to the proportions found in the benchmarking section (Section 3.4). This result is attributed to the observation that some proteins contain multiple metal centers (such as iron-sulfur clusters) thus increasing the frequency of certain metal/methionine-aromatic distance counts. A selection of interactions between metal ions and methionine-aromatic motifs at distances under 14 Å is shown in Fig. 3.6.

Excluding counts with distances above 60 Å, mean distances of 23.3±11.2, 24.4±11.9 and 24.2 ± 11.5 Å were found for metal- Met-Phe, Met-Tyr, and Met-Trp distances, respectively. Most of the Met-aromatic interactions were found to be located over 20 Å from metal sites. Consequently, it is unlikely that the majority of Met-aromatic interactions play a direct role in single electron redox chemistry involving metals, as single-step electron tunneling is prohibitively slow at such distances. [11, 17] Even the methionine-aromatic interactions that involve redox-active Trp and Tyr are mostly located far from metal sites. The evolutionary placement of methionine-aromatic interactions near the surface of proteins (far from metal sites) may play a more general protective role for biomolecules against oxidative stress. For example, Aledo *et al.* concluded that the interaction of a methionine sulfur with an aromatic amino acid increases the resistance of that sulfur to the formation of Met-sulfoxide. [33] However, the original question still remains: could the methionine-aromatic interactions located near metal sites be redox-active?

3.9 Statistics of distances between specific transition metals and Met-aromatic interactions

To gain more insight into potential redox reactivity of methionine-aromatic motifs, raw counts (Fig. 3.6) were divided into counts for different metal ions. Counts for Fe, Cu, and Zn proteins were highest in all cases, comprising over 90% of the total hits (Fig. 3.7). Fe proteins were also broken into heme and non-heme sub-groups and the results are similar to those shown for Fe in Fig. 3.7. Methionine-aromatic interactions are slightly closer to heme-iron than to non-heme iron, but the difference is minimal.

For all transition metal ions, the overall distribution of metal to Met-aromatic interaction distance was very similar. However, there are some notable differences a between these data and the total counts shown in Fig. 3.6. The differences are most apparent when comparing redox-active and redox-inactive metals. For example, for redox-active Fe and Cu, there are a greater proportion of methionine-aromatic interactions that are within 15 Å of the metal site; for Met-Trp interactions 16.1% of Fe and 23.9% of Cu metal to methionine-aromatic distances are below 15 Å, whereas for Zn, 12.8% of the interactions are under 15 Å.

The above analysis shows that methionine-aromatic motifs are positioned nearer to redox-active metals versus redox-inactive metals. There are still many interactions far from metal sites, with roles that are possibly structural or protective from oxidative stress, as described above. However, the higher incidences of methionine-aromatic interactions close to redox-active metals is intriguing. On one hand, localization of methionine-aromatic interactions could maintain local protein structure and shield the Met from sulfoxide formation. This is one likely case for Met-Phe motifs, since Phe is redox-inactive. In addition, shielding Met from oxidation could be another strategy used to steer electrons on functional [11] or protective [19] ET pathways. On the other hand, Met-Trp and Met-Tyr groups could be involved in single-electron redox reactions. For example, the bridging interaction shown in

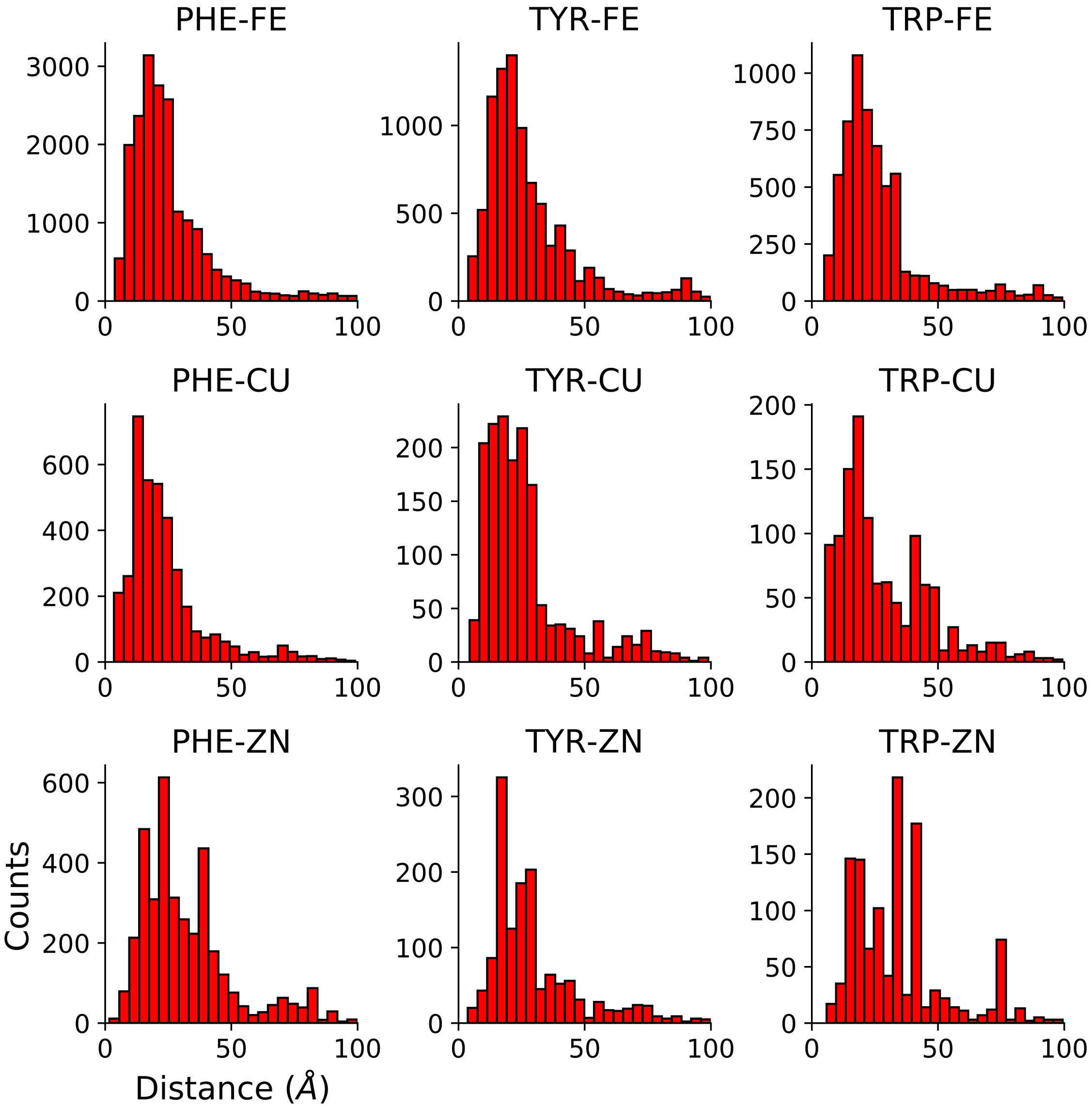


Figure 3.7. A breakdown of counts for Met-aromatic to metal distances for Fe, Cu, and Zn. A cutoff of 100 Å was imposed on all plots. A total of 25 bins were used for each distribution.

Fig. 3.5 is part of a chain of Tyr and Trp that extend from the lipoxygenase active site to the surface of the protein. Such chains have been proposed as protective electron conduits in metalloproteins. [19] This could be an example of a redox role for bridging Met-aromatic interactions.

Given the prevalence of methionine-aromatic interactions, it is remarkable that their redox reactions are not more widely explored. Thioethers positioned near aromatic groups can be more easily oxidized, which is one way that Met could plausibly participate in 1 electron redox reactions. [38, 39] Placement of a Met near Tyr or Trp is expected to affect the reduction potential(s) of both residues; in principle, the reduction potential of Tyr or Trp in a Met-aromatic motif could be modified in either direction. On one hand, the hydrophobic effect of placing a Met near an aromatic could raise aromatic (Tyr or Trp) potentials. On the other hand, a more delocalized electronic structure of methionine-aromatic groups could make the entire motif easier to oxidize. Ultimately, redox reactions involving methionine-aromatic interactions are subject to the same requirements of any ET system; potentials that are too low lead to buildup of radicals between a terminal donor/acceptor pair (and resulting side reactions), and potentials that are too high will inhibit or block ET.

Another facet of these motifs is that an oxidized methionine-aromatic can form a 2-center 3-electron bond, and redox reactions associated with bond making and bond breaking are typically associated with greater nuclear reorganization than are pure ET reactions. [40] Consequently, redox reactions involving methionine-aromatic motifs could be hindered by large reorganization energies. Ultimately, the protein redox chemistry involving methionine-aromatic is underexplored, both experimentally and theoretically. The survey presented here suggests that redox reactivity of methionine-aromatic motifs could be important and warrants greater investigation. We are presently following up this work by designing and investigating proteins that have only one methionine-aromatic interaction.

3.10 Challenges and extensions for PDB surveys

The analysis here presents new perspectives about methionine-aromatic interactions. Different interaction geometries, termed order of interaction, and methionine-aromatic bridging were identified. However, there were some challenges in analyzing these data that are worth pointing out. First, there is some degree of redundancy in these analyses. First, the entire oxidoreductase (EC 1) class of enzymes, including multiple structures of the same enzyme (often with minor differences, e.g., binding of substrate or point mutations) were analyzed. Second, there exists some natural heterogeneity in X-ray data, arising from many factors, including data quality and crystallization conditions. Likewise, resolution is improving with time, but the average resolution available in the PDB (2.2 Å) is not yet atomic resolution.

The completeness and format of the PDB entry data are also important factors. Some structures were discarded as unreadable because of incomplete or unknown delimiters. This was more problematic for older data, as formats are now standardized. In addition, some spurious examples were noted, such as interactions being near metal sites where such metal sites were artifacts of crystallization conditions. Nonetheless, these examples are rare and unlikely to strongly affect the statistics presented here. Ultimately, the available data set is now sufficiently large such that many of these rare examples are statistically insignificant. Still, such examples are important to note for the sake of completeness.

3.11 Examples of redox activity of Met-Trp motifs

Included in this data are some known redox-active methionine-aromatic motifs and some that are potentially important. In this section a few examples of such methionine-aromatic interactions are outlined. An example "hit" involves Trp191 in cytochrome *c* peroxidase (CcP), which interacts with Met230 and Met231. These two Met residues interact with the Trp191 pyrrole ring in the structure [41] of the ferric protein (Fig. 3.8). Met231 forms an order-III interaction and its closest distance (3.6 Å) is to a pyrrole ring carbon on Trp. In structures of oxidized (Compound I) CcP, [42, 43] Met230 and Trp191 are slightly closer together (by approximately 0.2 Å). Notably, mutations to Met230 and Met231 can alter [44] or destroy [45] the ability of CcP to produce a radical at Trp191.

Horse heart myoglobin (1YMB) contains two Trp residues (Trp7 and Trp14). Studies of Trp fluorescence [46] show that the fluorescence of both of these sites is quenched by the heme cofactor. Trp14 is about 10 Å from the heme edge and Trp7 is about 14 Å

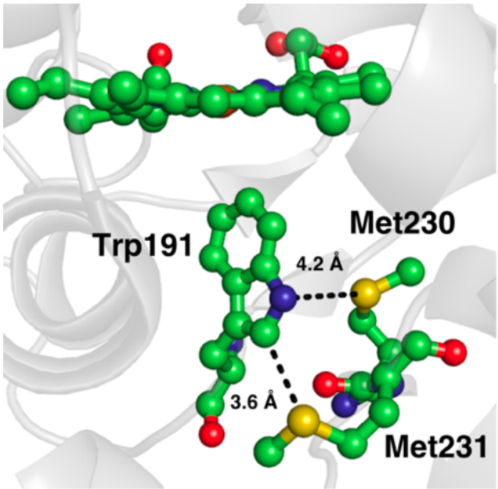


Figure 3.8. Structure of yeast cytochrome *c* peroxidase (PDB ID 2CYP) showing the interaction between Trp191, Met230, and Met231. The closest atomic distance is denoted by a dashed line.

from the heme edge. An advanced 2D UV spectroscopic investigation of the mechanism of fluorescence quenching demonstrated that Trp14 fluorescence is quenched almost exclusively via electron transfer, while Trp7 fluorescence is quenched via energy transfer. [47] As noted in that work, the distinct modes of reactivity are noteworthy. Passing 1YMB into the Met-aromatic algorithm (Fig. 3.3) yields an order-I Met131:Trp14 interaction (Met-ϕ angle $\sim 55°$). The distinct environments of the two Trp residues could be one reason for their different modes of reactivity.

Nitrite reductase (NiR) forms both Met-Tyr and Met-Trp motifs near the catalytic type I Cu site (Met141, Trp144, Tyr203). Mutation of Trp144 and/or Tyr203 results in little change to the redox properties of NiR, but a decrease in the rate of ET from the natural pseudoazurin reductant to the active site. [48] This behavior was probed using serial crystallography experiments carried out during turnover. [49] These experiments demonstrated that Met141 can change conformations during catalysis, with differing interactions between Trp144 and Tyr203. In addition, Tyr203 also can change conformation during turnover. In this case, where driving forces are low and aromatic radical formation is unlikely, the methionine-aromatic motif may play a more dynamic role in promoting conformational change, or enhancing electronic coupling, to promote ET (i.e., from pseudoazurin to the type I copper or from type I to type II copper).

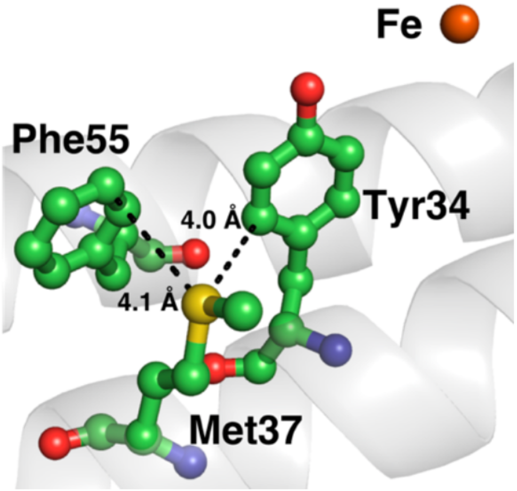


Figure 3.9. Structure of human ferritin (PDB ID 4Y08) showing the interaction between Met37 with Phe55 and Tyr34. The closest sulfur-aromatic atomic distances are shown. The orange sphere is an iron atom located in the ferritin ferroxidase site.

Ferritin, an essential iron storage protein in many organisms, has a highly conserved Tyr near the essential ferrioxidase site. In the human protein, Tyr34 forms an order-II interaction with Met37 (Fig. 3.9) and a nearby Phe55 forms an order-VI interaction with Met37. All three of these residues are largely conserved across many organisms. The involvement of Tyr34 in reactions of the ferrioxidase site has long been known [50], but only recently was it proposed to play a redox role, [51] so this could be another example of a redox-active bridging Met-aromatic interaction. However, we note that there is still controversy about the exact iron-loading mechanism in ferritin, including crucial redox reactions. [52]

In some heme degrading proteins, Met-Tyr interactions are observed near the heme edge. For example, the heme degrading protein ChuS from Escherichia coli O157:H7 has a bridging Met241 situated between the heme substrate and Tyr315. [53] The distance to the nearest aromatic midpoint in both heme and Tyr315 are both about 5 Å. This Tyr may be part of a hole transport chain [19] to the protein's surface, via Tyr297. Likewise, in human heme oxygenase, Met206 interacts with Tyr202. [54] Tyr202 is near a chain of other Tyr residues that extend to the heme surface, another potential example of a hole transfer pathway. Tyr202 also is implicated in protein-protein interactions of heme oxygenases. [55]

3.12 Conclusions and future work

In this Chapter, over 50,000 Met-aromatic interactions were identified in oxidoreductase structures retrieved from the Protein Data Bank. Almost half of those interactions involve the redox-active amino acids Trp or Tyr. During the course of this work, a new descriptor for Met-aromatic interactions called “order of interaction” was developed. This metric describes how much of an aromatic ring interacts with a nearby group. For Met-aromatic groups, the order-II and order-III interactions dominate. These geometries should be taken into account in future computational work. In addition, many bridging interactions involving one Met and two (or more) aromatic amino acids were found. To the best of our knowledge, these have not been discussed in the literature. These are interesting motifs that are deserving of further study.

What could Met-aromatic interactions be doing near metal sites? Met-Phe motifs make up the majority of Met-aromatic interactions. Met-Phe are also found near metals, leading to the conclusion that they possibly shield Met from oxidation or stabilize local structures. In addition, a great many of the Met-aromatic hits were far from oxidoreductase metal sites, consistent with known biological roles as structural stabilizers or as protectants from oxidation. However, methionine-aromatic interactions are found nearer to redox-active metal sites (Fe, Cu) than redox-inactive (Zn) sites. Such interactions could be involved in redox reactions or in protecting structurally important Met from oxidation. The methionine-aromatic motif has long been known in the literature, but this work suggests that its function(s) could be more diverse. Future work on redox reactions in metalloproteins should consider these widespread aromatic interactions. Likewise, related surveys should focus on how such interactions, including bridges, play roles in substrate- or cofactor-binding and in the assembly of protein interfaces.

Bibliography

[1] H. A. Scheraga, S. J. Leach, R. A. Scott, G. Nemethy. *Discuss. Faraday Soc.* **40**, (1965). 268-277

[2] L. Mirny, E. Shakhnovich. *Annu. Rev. Biophys. Biomol. Struct.* **30**, (2001). 361-396

[3] C. M. Dobson, A. Sali, M. Karplus. *Angew. Chem. Int. Ed.* **37**, (1998). 868-893

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

[5] B. Chakrabarty, N. Parekh. *Nucleic Acids Res.* **44**, (2016). W375-W382

[6] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, S. Pietrokovski. *J. Mol. Biol.* **344**, (2004). 1135-1146

[7] A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia. *J Mol Biol.* **247**, (1995). 536-540

[8] N. J. Zondlo. *Acc Chem Res.* **46**, (2013). 1039-1049

[9] M. Brandl, M. S. Weiss, A. Jabs, J. Sühnel, R. Hilgenfeld. *J Mol Biol.* **307**, (2001). 357-377

[10] K. S. C. Reid, P. F. Lindley, J. M. Thornton. *FEBS Lett.* **190**, (1985). 209-213

[11] J. J. Warren, M. E. Ener, A. Vlcek Jr., J. R. Winkler, H. B. Gray. *Coord Chem Rev.* **256**, (2012). 2478-2487

[12] J. Stubbe, W. A. van der Donk. *Chem Rev.* **98**, (1998). 705-762

[13] F. Rappaport, B. A. Diner. *Coord Chem Rev.* **252**, (2008). 259-272

[14] A. Lukacs, A. P. M. Eker, M. Byrdin, K. Brettel, M. H. Vos. *J Am Chem Soc.* **130**, (2008). 14394-14395

[15] E. C. Minnihan, D. G. Nocera, J. Stubbe. *Acc Chem Res.* **46**, (2013). 2524-2535

[16] J. W. Heinecke. *Free Radical Bio Med.* **32**, (2002). 1090-1101

[17] H. B. Gray, J. R. Winkler. *Q Rev Biophys.* **36**, (2003). 341-372

[18] B. Giese, M. Graber, M. Cordes. *Curr Opin Chem Biol.* **12**, (2008). 755-759

[19] H. B. Gray, J. R. Winkler. *Proc Natl Acad Sci U.S.A.* **112**, (2015). 10920-10925

[20] N. F. Polizzi, A. Migliore, M. J. Therien, D. N. Beratan. *Proc Natl Acad Sci U.S.A.* **112**, (2015). 10821-10822

[21] R. S. Morgan, C. E. Tatsch, R. H. Gushard, J. M. Mcadon, P. K. Warme. *Int J Pept Protein Res.* **11**, (1978). 209-217

[22] C. C. Valley, A. Cembran, J. D. Perlmutter, A. K. Lewis, N. P. Labello, J. Gao, J. N. Sachs. *J Biol Chem.* **287**, (2012). 34979-34991

[23] A. Cordomí, J. C. Gómez-Tamayo, V. Gigoux, D. Fourmy. *Trends Pharmacol Sci.* **34**, (2013). 320-331

[24] D. Pal, P. Chakrabarti. *J Biomol Struct Dyn.* **19**, (2001). 115-128

[25] G. Duan, V. H. Smith, D. F. Weaver *Mol Phys.* **99**, (2001). 1689-1699

[26] D. Pal, P. Chakrabarti. *J Biomol Struct Dyn.* **15**, (1998). 1059-1072

[27] M. L. Waters. *Biopolymers.* **76**, (2004). 435-445

[28] C. D. Tatko, M. L. Waters. *Protein Sci.* **13**, (2004). 2515-2522

[29] G. Kim, S. J. Weiss, R. L. Levine. *Biochim Biophys Acta.* **1840**, (2014). 901-905

[30] R. L. Levine, L. Mosoni, B. S. Berlett, E. R. Stadtman. *Proc Natl Acad Sci U.S.A.* **93**, (1996). 15036-15040

[31] R. L. Levine, J. Moskovitz, E. R. Stadtman. *IUBMB Life.* **50**, (2000). 301-307

[32] R. L. Levine, B. S. Berlett, J. Moskovitz, L. Mosoni, E. R. Stadtman. *Mech Ageing Dev.* **107**, (1999). 323-332

[33] J. C. Aledo, F. R. Cantón, F. J. Veredas. *Sci Rep.* **5**, (2015). 16955

[34] A. K. Lewis, K. Dunleavy, T. L. Senkow, C. Her, B. T. Horn, M. A. Jersett, R. Mahling, M. R. McCarthy, G. T. Perell, C. C. Valley, C. B. Karim, J. Gao, W. C. K. Pomerantz, D. D. Thomas, A. Cembran, A. Hinderliter, J. N. Sachs. *Nat Chem Biol.* **12**, (2016). 860-866

[35] E. A. Orabi, English, A. M. *Isr J Chem.* **56**, (2016). 872-885

[36] C. H. Hendon, D. R. Carbery, A. Walsh. *Chem Sci.* **5**, (2014). 1390-1395

[37] I. Z. Siemion. *Biosystems.* **32**, (1994). 163-170

[38] N. P.-A. Monney, T. Bally, G. S. Bhagavathy, R. S. Glass. *Org Lett.* **15**, (2013). 4932-4935

[39] W. J. Chung, M. Ammam, N. E. Gruhn, G. S. Nichol, W. P. Singh, G. S. Wilson, R. S. Glass. *Org Lett.* **11**, (2009). 397-400

[40] J. M. Mayer. *J Phys Chem Lett.* **2**, (2011). 1481-1489

[41] B. C. Finzel, T. L. Poulos, J. Kraut. *J Biol Chem.* **259**, (1984). 13027-13036

[42] G. Chreifi, E. L. Baxter, T. Doukov, A. E. Cohen, S. E. McPhillips, J. Song, Y. T. Meharenna, S. M. Soltis, T. L. Poulos. *Proc Natl Acad Sci U.S.A.* **113**, (2016). 1226-1231

[43] C. M. Casadei, A. Gumiero, C. L. Metcalfe, E. J. Murphy, J. Basran, M. G. Concilio, S. C. M. Teixeira, T. E. Schrader, A. J. Fielding, A. Ostermann, M. P. Blakeley, E. L. Raven, P. C. E. Moody. *Science.* **345**, (2014). 193-197

[44] L. A. Fishel, M. F. Farnum, J. M. Mauro, M. A. Miller, J. Kraut, Y. Liu, X. L. Tan, C. P. Scholes. *Biochemistry.* **30**, (1991). 1986-1996

[45] T. P. Barrows, B. Bhaskar, T. L. Poulos. *Biochemistry.* **43**, (2004). 8826-8834

[46] R. M. Hochstrasser, D. K. Negus. *Proc Natl Acad Sci U.S.A.* **81**, (1984). 4399-4403

[47] C. Consani, G. Auböck, F. van Mourik, M. Chergui. *Science.* **339**, (2013). 1586-1589

[48] K. Yamaguchi, K. Shuta, S. Suzuki. *Bichem Biophys.Res Commun.* **336**, (2005). 210-214

[49] S. Horrell, S. V. Antonyuk, R. R. Eady, S. S. Hasnain, M. A. Hough, R. W. Strange. *IUCrJ.* **3**, (2016). 271-281

[50] G. S. Waldo, J. Ling, J. Sanders-Loehr, E. C. Theil. *Science.* **259**, (1993). 796-798

[51] K. H. Ebrahimi, P. Hagedoorn, W. R. Hagen. *Chem Bio Chem.* **14**, (2013). 1123-1133

[52] W. R. Hagen, P.-L. Hagedoorn, K. H. Ebrahimi. *Metallomics.* **9**, (2017). 595-605

[53] M. D. L. Suits, N. Jaffer, Z. Jia. *J Biol Chem.* **281**, (2006). 36776-36782

[54] C. M. Bianchetti, L. Yi, S. W. Ragsdale, G. N. Phillips. *J Biol Chem.* **282**, (2007). 37624-37631

[55] A. L. M. Spencer, I. Bagai, D. F. Becker, E. R. P. Zuiderweg, S. W. Ragsdale. *J Biol Chem.* **289**, (2014). 29836-29858

Chapter 4

The Bridging Interaction

4.1 Preface

In this Chapter, a computational geometry survey of the Protein Data Bank is discussed, with an emphasis on the "bridging interactions" described in Chapter 3. Here, the motivation is to establish a broad idea of how bridging interaction geometries relate to different parts of the protein scaffold. To do so, bridge anatomy based on aromatic type (Phe, Tyr, Trp) was first enumerated. Next, the relationship between bridges, metal centers and the protein surface was investigated using a scalene triangle model. Finally, the incidences of bridge superimposition onto larger tyrosine/tryptophan chains was studied.

The Chapter is best started with a rigorous definition of the "bridging interaction." *A bridging interaction is any set of n number of vectors $\vec{v}$ meeting Met-aromatic criteria, where $j \leq n < \infty$ and where n number of vectors $\vec{v}$ project to j number of aromatic residues, and where $j \geq 2$.* Note that n (equivalent to the order of interaction) can be an integer as low as 2. In other words, a bridging interaction does not have to be of order exceeding VI. Instead, the principle requirement is that one methionine projects at least to two different aromatics. In this Chapter, the following notation is used for bridging interactions:

$$\text{A} : \text{MET} : \text{B}\,. \tag{4.1}$$

Here, A and B can be any residue $\in$ {PHE, TYR, TRP}. A critical property of the bridging interaction is that it can be treated as a network. A more detailed description of networks is given in Chapter 1. Here, properties of networks are exploited for differentiating the bridging interaction from a simple aromatic interaction. As an example, consider the following: Let A : MET : B be a bridging interaction in a hypothetical wild type protein. A point mutation is introduced into the protein such that B is replaced with some non-aromatic residue C to yield A : MET/C. This simple example highlights a key computational challenge, namely how a computer can differentiate between the interactions in the wild type and mutant cases? The approach used here is to assign an adjacency matrix to each cluster, where the nodes are the residues and the edges vectors $\vec{v}$. We have 4.2 and 4.3:

$$G_{\text{A:MET:B}} = \begin{array}{cccc} \text{A} & \text{MET} & \text{B} & \\ \left(\begin{matrix} 0 \\ 1 \\ 0 \end{matrix}\right. & \begin{matrix} 1 \\ 0 \\ 1 \end{matrix} & \left.\begin{matrix} 0 \\ 1 \\ 0 \end{matrix}\right) & \begin{matrix} \text{A} \\ \text{MET} \\ \text{B} \end{matrix} \end{array} \tag{4.2}$$

$$G_{\text{A:MET/C}} = \begin{array}{cccc} \text{A} & \text{MET} & \text{C} & \\ \left(\begin{matrix} 0 \\ 1 \\ 0 \end{matrix}\right. & \begin{matrix} 1 \\ 0 \\ 0 \end{matrix} & \left.\begin{matrix} 0 \\ 0 \\ 0 \end{matrix}\right) & \begin{matrix} \text{A} \\ \text{MET} \\ \text{C} \end{matrix} \end{array} \tag{4.3}$$

$$.$$

Note that $G_{\text{A:MET/C (2, 3)}}$ and $G_{\text{A:MET/C (3, 2)}} = 0$ indicating no connection between MET and C. Here Met-aromatic does not find an interaction between the residues MET and C as C was previously defined as being non-aromatic. The degree matrices 4.4, 4.5 can be found from 4.2 and 4.3 by taking the sum of each column:

$$D_{\text{A:MET:B}} = \begin{array}{ccc} \text{A} & \text{MET} & \text{B} & \\ \end{array}$$

$$D_{\text{A:MET:B}} = \begin{array}{c c} \begin{array}{ccc} \text{A} & \text{MET} & \text{B} \end{array} & \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} & \begin{array}{c} \text{A} \\ \text{MET} \\ \text{B} \end{array} \end{array} \tag{4.4}$$

$$D_{\text{A:MET/C}} = \begin{array}{c c} \begin{array}{ccc} \text{A} & \text{MET} & \text{C} \end{array} & \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \begin{array}{c} \text{A} \\ \text{MET} \\ \text{C} \end{array} \end{array} \tag{4.5}$$

.

A degree of 2 at $D_{\text{A:MET:B (MET, MET)}}$ indicates that a methionine residue bridges 2 aromatic residues, yielding a 2-bridge. In general, the degree matrix D for an N-bridge will have a column space dimension of $\mathbb{R}^{\text{N+1}}$ and will have $D_{\text{(MET, MET)}}$ = N.

4.2 Introduction

Proteins are held in their native structures by a great many intramolecular interactions. A handful of examples of such interactions include hydrogen bonding, hydrophobic interactions, salt bridges, and $\pi - \pi$ interactions. [1] The development of the Protein Data Bank has allowed the investigation and elucidation of other interactions that are important for protein structure. One established example are the cation-π interactions. [2] Recently, it was suggested that proteins can be modeled and understood on the basis of these non-covalent interactions, rather than traditional secondary and tertiary structures. [3] In many ways, this framework better emphasizes cross-domain interactions that are essential for structure, as well as those other interactions that give rise to function (e.g., substrate binding, protein-protein interactions, electron flow). Of interest in this Chapter are intraprotein interactions of methionine (Met) with the aromatic residues tyrosine, tryptophan, and phenylalanine.

In the early 1980's, protein structure surveys identified a propensity of the Met δ-sulfur to localize near the aromatic amino acids Phe, Tyr, and Trp. [4] This localization of Met near aromatic residues was termed an "interaction." The first survey included 36 proteins (of only about 195 available at the time), so it is remarkable that the metrics from that small survey have stood the test of time. A more recent survey [6] showed how common these methionine-aromatic interactions are in biomolecules, with many proteins having one or more interactions. In that work, the methionine-aromatic interaction was proposed to contribute up to 1 to 1.5 kcal mol^{-1} to protein stability in a manner that is not purely hydrophobic. A natural question that arises from the observation of the large number of methionine-aromatic interactions is about their specific biological roles. Should these methionine-aromatic motifs be thought of as hydrophobic structural features, based on work with model proteins? Or do these motifs serve other roles? One perspective, based on work with small peptide models, suggests that methionine-aromatic groups provide stability that is purely hydrophobic. [7] This view contrasts with the above idea that a methionine-aromatic group adds stabilization beyond that of hydrophobic interaction. In addition to structural roles for methionine-aromatic motifs, we recently proposed that those involving Tyr and Trp could be involved in redox reactions. [8] During the course of that work, we uncovered a surprising propensity for two aromatic residues to interact with a single Met. We termed these motifs bridging interactions. Herein, we expand on that work and demonstrate that the bridging interaction is ubiquitous in protein systems.

4.3 Methods

All programming was done in the Python 3.5 programming language (https://www.python.org) using PEP8 guidelines. The BioPython package (https://biopython.org) was used to connect to the RCSB Protein Data Bank (https://www.rcsb.org) and fetch a recent index of all PDB codes. The BioPython package was then used to import .pdb files corresponding to those codes present in the recent index. The Python Pandas Data Analysis Library, v0.23.0 (https://pandas.pydata.org) was used to augment imported .pdb data into a dataframe, where only those chains labelled with chain delimiter "A" (a protein can consist of multiple chains) and only the first model in multi-model entries were isolated (A PDB entry can contain multiple models corresponding to crystal structures for different conformations). Data were then further refined to include coordinates corresponding solely to Met, Tyr, Trp and Phe residues. Data was passed off to Met-aromatic for finding and classifying methionine-aromatic pairs. All low-level computation was performed at the C level using the Python Numpy library (`http://www.numpy.org`). The approach here was almost identical to that presented in Chapter 3 with a few exceptions. Multiple versions of Met-aromatic exist as of December 2018, written in Pandas, pure Python and even C++.

Data for the methionine-aromatic pairs obtained from the Met-aromatic algorithm were dynamically banked into a SQL database using Python's pyodbc ODBC bridge (`https://github.com/mkleehammer/pyodbc`). Each line in the SQL database described the geometry of one vector $\vec{v}$ and contained additional information such as the PDB code, methionine residue position number, the aromatic residue position number, Met-θ angle, Met-ϕ angle, $\|\vec{v}\|$, and the aromatic residue identity. To find bridging interactions, a split / apply / combine procedure was performed on individual groupby objects in the SQL database where grouping was done by PDB code. A programmatic augmentation of the above definition of the bridging interaction was applied to each groupby object, where an interaction was deemed "bridging" if any number of vectors $\vec{v}$ projected from one methionine residue to exactly two aromatic residues of any identity (Condition 2, Table 4.1). Only dual bridging interactions were examined in this study; any methionine residue projecting vectors $\vec{v}$ to three or more aromatics was passed.

Note that the lower bound j is conveniently identical to the degree of a methionine residue 2-bridge (see 4.4). This trend continues for bridges of degree greater than 2, i.e.:

Table 4.1: One can find bridges of specific type by imposing lower and upper bounds on Met-aromatic, where these bounds refer to the number of vectors $\vec{v}$ projecting from a methionine to j number of aromatics. As an example, dual bridging interactions can be isolated by excluding any interactions of order less than II (a minimum of two vectors $\vec{v}$ are required to "touch" two aromatics) and order greater than XII. Degree matrices can be obtained from j to programmatically select in specific bridges (4.6, 4.7).

	Proximal aromatics	Lower bound (min n)	Upper bound (max n)
Condition 1	1	1	6
Condition 2	2	2	12
Condition 3	3	3	18
Condition 4	4	4	24
Condition t	j	j	$6j$

$$D_{\text{MET:[A, B, C]}} = \begin{array}{cc} \begin{array}{cccc} \text{MET} & \text{A} & \text{B} & \text{C} \end{array} & \\ \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} & \begin{array}{c} \text{MET} \\ \text{A} \\ \text{B} \\ \text{C} \end{array} \end{array} \tag{4.6}$$

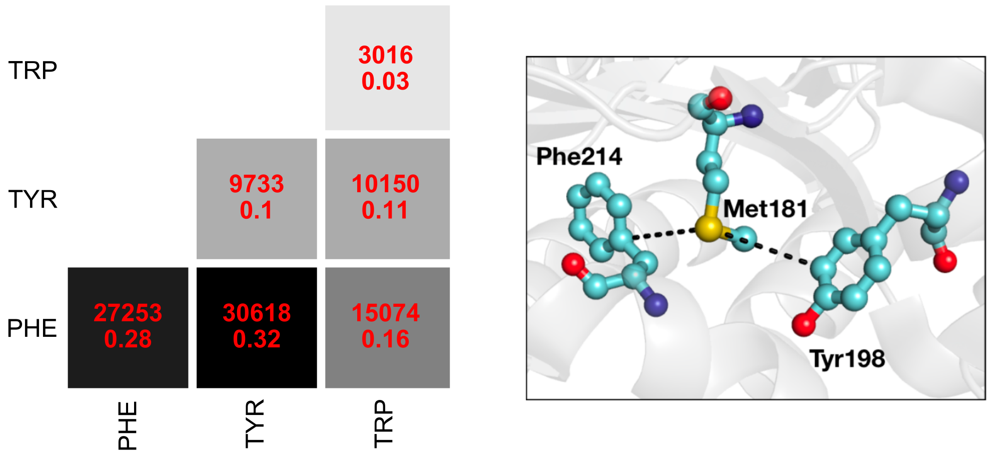
,

for a 3-bridge where aromatics A, B, C surround MET, or:

$$D_{\text{MET:[A, B, C, D]}} = \begin{array}{cc} \begin{array}{ccccc} \text{MET} & \text{A} & \text{B} & \text{C} & \text{D} \end{array} & \\ \begin{pmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} & \begin{array}{c} \text{MET} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \end{array} \end{array} \tag{4.7}$$

,

for a 4-bridge where aromatics A, B, C, D surround MET.



	PHE	TYR	TRP
TRP			3016 0.03
TYR		9733 0.1	10150 0.11
PHE	27253 0.28	30618 0.32	15074 0.16

Figure 4.1. Left: A correlation depicting breakdowns between aromatic members in bridging interactions. Each pixel contains the raw number of interactions (UPPER) and the proportion relative to the total number of bridging interactions found in this study (95,844) (LOWER). *Right*: An example of the most frequent bridging interaction, Phe : Met : Tyr. The PyMOL rendering [5] shows Phe214 : Met181 : Tyr198 in the core domain of inositol polyphosphate multikinase (PBD ID 6C8A). The degree matrix corresponding to this discrete bridge is shown in equation (4.8). Closest contacts for heavy atoms are highlighted with black dashed lines.

4.4 Overall counts for bridging interactions

The modified version of the Met-aromatic algorithm was wrapped in a custom looping script, and iterated over an index of 139,948 structures in the Protein Data Bank (index collected from the PDB on 22 May 2018). Of this number, a total of 95,964 (69%) PDB entries contain at least one aromatic interaction (inclusive of bridges) under a cut-off distance of 6.0 Å and cutoff angles of 109.5°. An analysis of entries containing at least one aromatic interaction revealed a total of 115,030 bridging interactions, that is 115,030 cases where a methionine has exactly two aromatic neighbours. A secondary analysis of the 115,030 bridging interactions using the NetworkX Python library revealed that many bridges were of the form A : MET : B : MET : C : MET : D [repeating] which prompted me to isolate only those bridges where a corresponding adjacency matrix had a column space of $\mathbb{R}^3$ and a degree of 1 for each of A and B (4.4). This reduced the number of counts to 95,844 (Fig. 4.1). A strong preference for sulfur-aromatic distances > 5.0 Å is observed. This observation is consistent with similar bioinformatics studies. [4]

$$D_{\text{Phe214:Met181:Tyr198}} = \begin{array}{c} \begin{array}{ccc} \text{Phe214} & \text{Met181} & \text{Tyr198} \end{array} \\ \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{array}\right) \end{array} \begin{array}{l} \\ \text{Phe214} \\ \text{Met181} \\ \text{Tyr198} \end{array} \tag{4.8}$$

4.5 Survey of bridging geometries

As one metric for assaying whether bridging interactions could play a role in redox reactions, a survey of bridging interaction positions relative to the protein surface and embedded metal centers was carried out. Here, the research question was whether bridging interactions were directly in, or near, a path between a metal and the protein surface, owing to the observation that many biological processes involve transferring oxidizing equivalents away from active sites and to surfaces sites for scavenging. [9] First, all PDB entries containing both a metal atom and a bridging interaction were isolated (8,743 entries). The atomic surface coordinates were found for any PDB entries meeting this initial condition. All surface coordinates were obtained using the PyMOL Application Programming Interface, where a custom Python 2.7 script was executed directly through the PyMOL command interpreter. The surface coordinates were found by taking advantage of PyMOL's FindSurfaceResidues machinery, where the underlying code determines which atoms are exposed based on solvent exposure. Any atom with solvent exposure exceeding 2.5 Å^2 was considered a surface coordinate and all such coordinates were stored in an array. Next, the minimum distance from the protein surface to a bridging interaction methionine δ-sulfur was found iteratively. The surface coordinate in the array meeting this condition was given the label **SF** and the corresponding δ-sulfur coordinate was given the label **SD**. The third coordinate, **MT**, was used to describe the Cartesian position of the metal center. The coordinates **SF**, **SD**, and **MT** complete the vertices of a scalene triangle (Fig. 4.2). To assess the arrangement of a metal, the protein surface, and a bridging interaction, distributions of the face lengths and the areas a of all scalene triangles were generated. The face lengths were found by computing the Euclidean norm of the three vectors enclosing the scalene triangle. The areas were found by computing the norm of the cross product of any two of three vectors and scaling the norm by $\frac{1}{2}$ (equation 4.9). In general, an area a approaching 0 would indicate a very closely spaced surface / bridge / metal triad, however isolated cases could be linearly arranged **SF** / **SD** / **MT** coordinates. It was for this reason that both areas and face lengths were analyzed.

$$a = \frac{1}{2} \left\| \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ p_x & p_y & p_z \\ q_x & q_y & q_z \end{vmatrix} \right\| \,, \tag{4.9}$$

where:

$$p = \mathbf{SF} - \mathbf{SD} \,, \tag{4.10}$$

$$q = \mathbf{MT} - \mathbf{SD} \,. \tag{4.11}$$

This study was done in two groups: the first group consisted of a screen of A : MET : B bridges, where one (or both) of A or B were non redox-active PHE residues. This served as the control group owing to the observation that PHE is an aromatic incapable of electron transfer. Second, an experimental group was set up. Here, C : MET : D bridges were isolated, where both of C and D were any of TYR/TRP. Statistics were collected for the shape of the scalene triangles for both groups.

The shape of the scalene triangle distributions (areas, face lengths) is roughly identical for both bridges not expected to play a role in redox and those bridges that could indeed serve a redox role. A significant finding, however, is the high incidence of extremely short **SF** / **SD** vectors. Both the TYR/TRP bridge case (Fig. 4.3, bottom) and the PHE containing bridge case (control group) (Fig. 4.3, top) demonstrate that the majority of **SF** coordinates were also identified as being the **SD** coordinates, thus also explaining why both the a distributions have a maximum at 0 Å^2. The next closest maximum for the **SF** / **SD** distribution is located at approximately 1.7 Å for both cases which is consistent with the carbon-sulfur bond length. This indicates that in the majority of cases, the bridging methio-

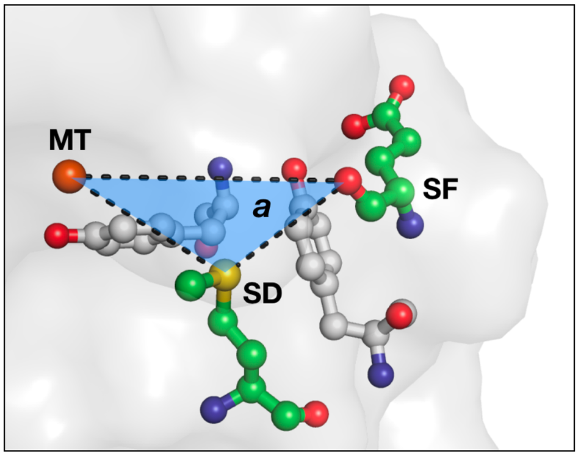


Figure 4.2. The triangular relationship between a surface residue (coordinate **SF**), a bridging interaction (coordinate **SD**) and a metal (coordinate **MT**) for the protein 1A7E. Here **MT** refers to an Fe atom. **SF**: Glu69. **SD**: Met62 bridged by Tyr67 and Tyr114 (shown in gray). The scalene triangle has area a.

nine SD coordinate is located directly on the protein surface with the next most common surface coordinate being either methionine CE or CG.

As of now, we are unsure whether the scalene triangle algorithm is artificially skewing the results. After all, the algorithm first finds a methionine **SD** and then iteratively finds the closest solvent exposed atomic coordinate. The finding that bridges reside near the protein surface in many cases is valid, however concern arises when thinking about how the distributions would change if the algorithm iteratively found the minimum distance between a solvent exposed atomic coordinate and a metal coordinate, **MT**.

4.6 Network analysis and bridge superimposability

In this last part of this Chapter, the superimposition of bridging interactions onto longer closely spaced Tyr/Trp aromatic chains is investigated. The goal in this effort is to assess how Met-aromatic bridges are related to longer chains of closely spaced Tyr and Trp residues [9]. This problem was approached from a graph- and network theory approach, which can be explained using the following example. Note that in this case it is assumed that there is only one bridge and one aromatic chain per protein. Consider Fig. 4.4 which shows five residues located on some arbitrary space, or more succinctly, {A, B, C, D, E} $\in F$. These residues in F are nodes which will be connected using a series of edges.

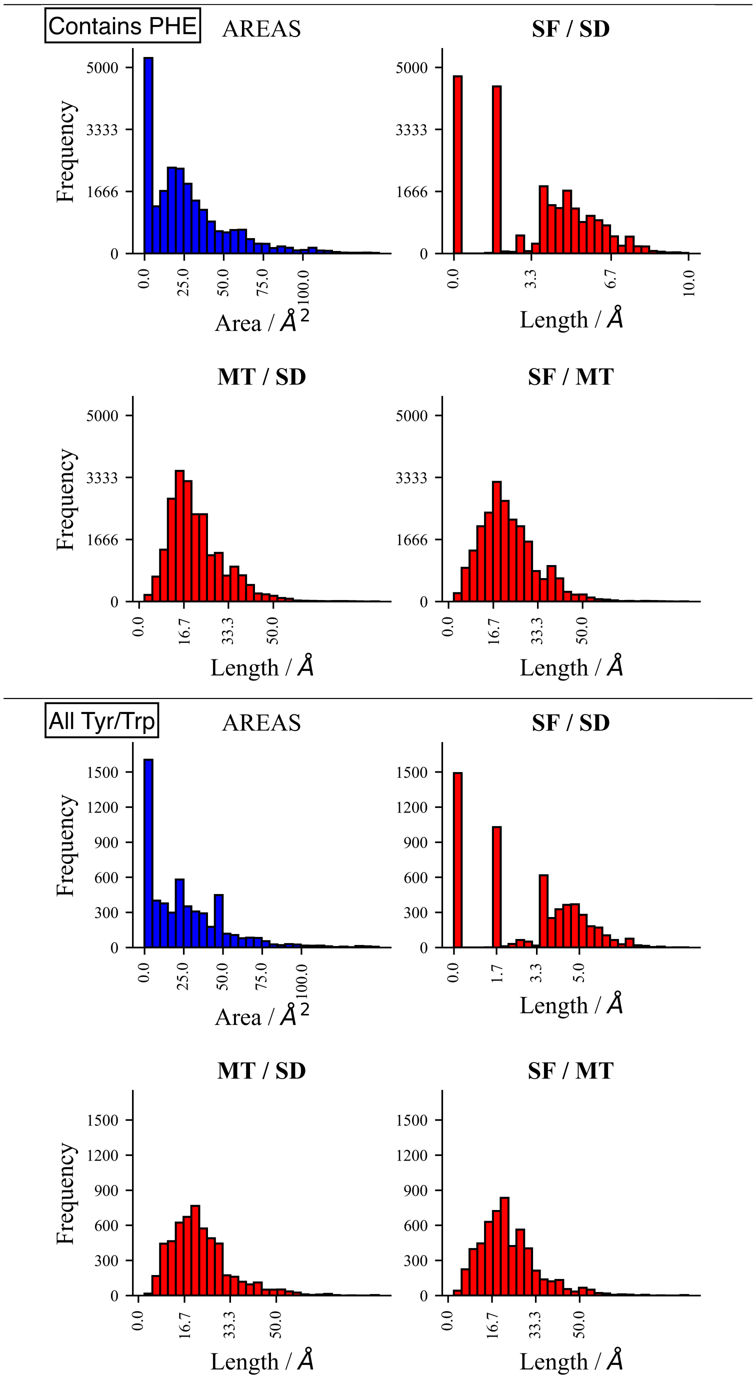


Figure 4.3. Top: Scalene triangle face length and area distributions for all bridges where A = PHE or B = PHE in a bridge of form A : MET : B. Such bridges are not expected to participate in redox reactions. *Bottom*: Scalene triangle face length and area distributions for all bridges where A = (TYR or TRP) and B = (TYR or TRP) in a bridge of form A : MET : B. Such bridges could participate in redox reactions. A total of 30 bins were used for all distributions.

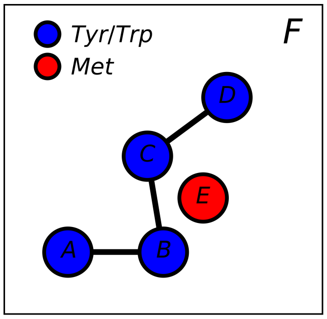


Figure 4.4. A two-dimensional feature space F. This 2-space contains a total of five nodes: four aromatics [A, B, C, D] and one methionine [E]. The edges here are connection vectors between aromatic residues of norm below some cutoff distance.

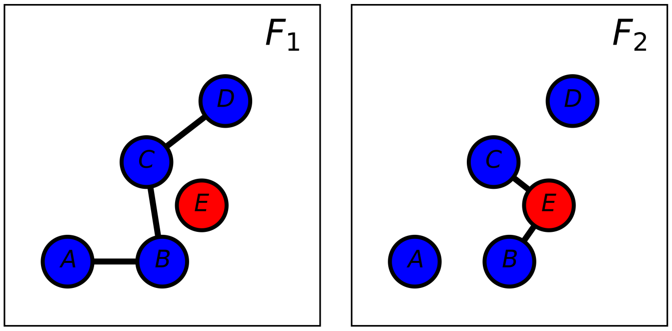


Figure 4.5. Two feature spaces, F_1 and F_2, of varying edge type stemming from the original feature space F. As above, BLUE markers represent aromatics. RED markers represent methionine.

The feature space F from Fig. 4.4 can be modified into two additional feature spaces: F_1 and F_2. The two feature spaces are identical except for the identity of their edges. In F_1, the edges are vectors of norm ≤ 7.4 Å connecting Tyr/Trp residues. These edges are chosen as they ensure that Tyr/Trp residues are proximal to each other, thus allowing for electron transfer. In F_2, the edges are 2-bridging interactions. This is shown in Fig. 4.5.

In Fig. 4.5, (F_1), a chain of connected aromatic amino acids A-B-C-D is shown, with E (the hypothetical Met) neglected. The algorithm analyzing F_1 can only "see" the A-B-C-D chain. In Fig. 4.5, (F_2), a C : E : B bridge is shown and the algorithm analyzing F_2 is "blind" to the residues A and D. To the human eye, it is clear that the bridging interaction C : E : B superimposes onto A-B-C-D, but the problem is slightly more complex from a computational perspective. First, note that the cardinality of $\{A, B, C, D\}$ is 4. Next, consider two adjacency matrices, G_1 (4.12) and G_2 (4.13) of dimension $\mathbb{R}^4$ corresponding to two feature spaces and F_1 and F_2:

$$G_1 = \begin{array}{cc} \begin{array}{cccc} \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} \end{array} & \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} & \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \\ \mathrm{D} \end{array} \end{array} , \tag{4.12}$$

$$G_2 = \begin{array}{cc} \begin{array}{cccc} \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} \end{array} & \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} & \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \\ \mathrm{D} \end{array} \end{array} . \tag{4.13}$$

The product of G_1 and G_2 yields critical information about the type of superimposition present in our protein of interest and is given the variable S:

$$S = \begin{cases} 0 & \text{if } \mathrm{Tr}(G_1 G_2) = 0 \\ \neg 0 & \text{if } \mathrm{Tr}(G_1 G_2) > 0 \end{cases} . \tag{4.14}$$

If $S = 0$ then the algorithm determines that a chain in F_1 shares no nodes with a bridge in F_2 (4.14). In Fig 4.5, $S \neq 0$. Here the diagonal of $G_1 G_2$ follows (4.15):

$$G_1 G_2 = \begin{array}{cc} \begin{array}{cccc} \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} \end{array} & \\ \begin{pmatrix} 0 & & & \\ & 1 & & \\ & & 1 & \\ & & & 0 \end{pmatrix} & \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \\ \mathrm{D} \end{array} \end{array} \tag{4.15}$$

,

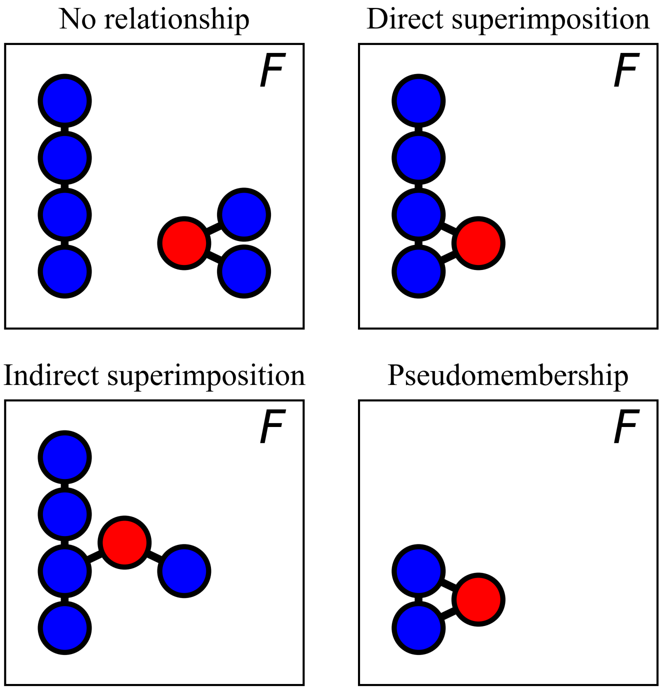


Figure 4.6. Four feature spaces F for four different geometries we discovered in the initial study. RED CIRCLE: the methionine residue. Each methionine residue is bound to two aromatics (BLUE CIRCLES).The case of no relationship and direct superimposition (TOP ROW) was previously described. Here two new geometries are presented: *indirect superimposition* and *pseudomembership.*

where $\text{diag}(G_1G_2)$ is TRUE at the indices corresponding to B and C residues. This finding suggests that the C : E : B bridge superimposes onto the aromatic chain A-B-C-D. The above method was used in an initial network analysis of 2,611 randomly selected protein structures. Feature spaces F_1 and F_2 were computed for each structure. Importantly, it was assumed that there could only exist two types of geometries: a geometry where a bridge and an aromatic chain bore no relationship whatsoever (*no relationship*) and a geometry where both ends of a 2-bridging interaction were also members of a longer aromatic chain(i.e. in the above example, bridging residues **B** and **C** are also part of the longer chain A-**B**-**C**-D) (*direct superimposition*). An analysis of the data obtained from the screen of 2,611 structures revealed that the network analysis problem was more complex. First, there existed four unique geometries (see fig. 4.6). Second, a protein structure could have two or more closely spaced aromatic chains. The initial network analysis algorithm failed to account for these unique cases and generated erroneous data.

Still assuming every protein contained only one aromatic chain, S was rewritten as in (4.16):

$$S = \begin{cases} \text{NR} & \text{if } \mathrm{Tr}(G_1G_2) = 0 \ \& \ G_1G_2 = 0_{K_{n,n}} \\ \text{IS} & \text{if } \mathrm{Tr}(G_1G_2) = 0 \ \& \ G_1G_2 \neq 0_{K_{n,n}} \\ \text{DS} & \text{if } \mathrm{Tr}(G_1G_2) \neq 0 \ \& \ \mathrm{Nullity}(G_1G_2) \neq 0 \\ \text{2C} & \text{if } \mathrm{Tr}(G_1G_2) \neq 0 \ \& \ \mathrm{Nullity}(G_1G_2) = 0 \end{cases} . \tag{4.16}$$

The condition S is named as any of: NR = *no relationship*, IS = *indirect superimposition*, DS = *direct superimposition*, 2C = *pseudomembership*, as opposed to assigning S some integer (4.14). The value of n is equal to the dimension of G_1G_2 and $0_{K_{n,n}}$ = the null (zero) matrix of dimension equivalent to G_1G_2. In the DS and 2C cases, $\mathrm{Tr}(G_1G_2)$ will always be non-zero. The only way to further separate DS from 2C cases is to use the Rank-Nullity theorem to find the nullity of G_1G_2, where a non-zero nullity indicates a direct superimposition geometry.

S must be computed for all bridge-chain combinations. Consider the following example: a protein structure contains the chains A-B-C and D-E-F. The protein structure also contains the bridges B : MET : G and H : MET : I. Here, a total of 4 conditions S must be determined:

1. $\{\text{A, B, C}\} \cap \{\text{B, MET, G}\} = \{\text{B}\} \therefore S = \text{IS}$
2. $\{\text{A, B, C}\} \cap \{\text{H, MET, I}\} = \varnothing \therefore S = \text{NR}$
3. $\{\text{D, E, F}\} \cap \{\text{B, MET, G}\} = \varnothing \therefore S = \text{NR}$
4. $\{\text{D, E, F}\} \cap \{\text{H, MET, I}\} = \varnothing \therefore S = \text{NR}$

There exists no relationship in 3 of the 4 bridge-chain combinations. The residue B, however, is an element in both A-B-C and B : MET : G indicating an indirect superimposition geometry.

To identify disconnected aromatic chains in a protein structure, the adjacencies G_1 for all feature spaces F_1 are found programmatically. The discrete Laplacian L_{F_1} (4.17) is then found from G_1 (4.17), where D_{F_1} refers to the degree matrix corresponding to adjacency matrix G_1.

$$L_{F_1} = D_{F_1} - G_1 \tag{4.17}$$

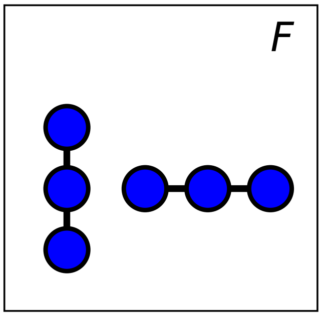


Figure 4.7. A feature space F containing six aromatic chains, or two 3-chains. This condition occurred in a minority of protein structures surveyed.

L_{F_1} is programmatically eigendecomposed (LAPACK routine) to yield all eigenvalues (4.18):

$$L_{F_1}(\vec{v}) = \lambda\vec{v}\,. \tag{4.18}$$

The number of disconnected units in F_1 is equivalent to the algebraic multiplicity of the $\lambda = 0$ eigenvalue. From here, the geometry S could be found for each disconnected chain. A brief example below demonstrates how to compute the number of disconnected chains in a protein structure. Let F be the 2-dimensional feature space containing two disconnected chains: A-B-C and D-E-F (Fig. 4.7).

The adjacency matrix A_F (4.19) follows:

$$A_F = \begin{array}{c} \begin{array}{cccccc} \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} & \mathrm{E} & \mathrm{F} \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array} \begin{array}{c} \\ \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \\ \mathrm{D} \\ \mathrm{E} \\ \mathrm{F} \end{array}\,, \tag{4.19}$$

From which the degree matrix D_F (4.20) is obtained (by summing the columns in A_F):

$$D_F = \begin{array}{cccccccc} & \text{A} & \text{B} & \text{C} & \text{D} & \text{E} & \text{F} & \\ \left(\right. & 1 & 0 & 0 & 0 & 0 & 0 & \left.\right) \text{A} \\ & 0 & 2 & 0 & 0 & 0 & 0 & \text{B} \\ & 0 & 0 & 1 & 0 & 0 & 0 & \text{C} \\ & 0 & 0 & 0 & 1 & 0 & 0 & \text{D} \\ & 0 & 0 & 0 & 0 & 2 & 0 & \text{E} \\ & 0 & 0 & 0 & 0 & 0 & 1 & \text{F} \end{array} . \tag{4.20}$$

Next, the discrete Laplacian matrix L_F (4.21) is found by subtracting A_F from D_F:

$$L_F = D_F - A_F = \begin{array}{ccccccc} \text{A} & \text{B} & \text{C} & \text{D} & \text{E} & \text{F} & \\ 1 & -1 & 0 & 0 & 0 & 0 & \text{A} \\ -1 & 2 & -1 & 0 & 0 & 0 & \text{B} \\ 0 & -1 & 1 & 0 & 0 & 0 & \text{C} \\ 0 & 0 & 0 & 1 & -1 & 0 & \text{D} \\ 0 & 0 & 0 & -1 & 2 & -1 & \text{E} \\ 0 & 0 & 0 & 0 & -1 & 1 & \text{F} \end{array} . \tag{4.21}$$

The eigendecomposition yielding the vector of eigenvalues of L_F is performed programmatically (4.22):

$$\vec{\lambda} = \begin{pmatrix} 3 & 1 & 0 & 3 & 1 & 0 \end{pmatrix}^{\mathrm{T}} . \tag{4.22}$$

The algebraic multiplicity of 0 in $\vec{\lambda}$ (or the dimension of the null space of L_F) is 2. Therefore Fig. 4.7 contains 2 disconnected amino acid chains for which S is found. This approach still does not address the condition of multiple bridging interactions in a protein, which is found in certain cases. At this point, it should be more clear how the linear algebraic analysis of these protein networks can rapidly grow to an extremely complicated problem. In an attempt to simplify this problem, much of the network analysis presented here was carried out using the NetworkX Python library. NetworkX is a library commonly used for social media network analysis, bioinformatics, tracking, and others. [10]

The initial screen of 2,611 PDB structures revealed the above complexities. That subset of structures was pre-selected to contain only those proteins with both a bridge and at least one chain of closely spaced aromatics. A second analysis of all proteins in the PDB using the refined definition of S (4.16) was carried out. Here, any set of Tyr/Trp residues spaced ≤ 7.4 Å apart [23] were considered as connecting nodes in F_1. In F_2, Tyr/Trp pairs in 2-bridges (Table 4.1, Condition 2) were considered nodes connected via an edge, where the edge identity was the 2-bridging interaction (see Fig. 4.5, right). The constraints passed into the Met-aromatic algorithm were as follows: $\|\vec{v}\| \leq 6.0$ Å and (Met-$\theta \leq 109.5°$ or Met-$\phi \leq 109.5°$).

A total of 144,682 PDB files were analyzed in the secondary screen. 20,784 PDB files contained at least one bridging interaction and at least one aromatic chain. These candidates were passed off to the above algorithm for computing geometries S. A total of 119,038 bridge/chain pairs were analyzed in the 20,784 candidate files. Here, a bridge/chain pair is defined as the relationship between one chain and one bridge. Therefore, a total of $n \times m$ bridge/chain pairs were analyzed for each protein, where n refers to the number of bridges in a given protein and m the number of chains. 99,166 bridge/chain pairs were of $S =$ NR, indicating no relationship. However, in 19,872 bridge/chain pairs, S was any of DS, IS, or 2C conditions, indicating some close geometric relationship between bridges and Tyr/Trp chains. All results of this analysis are presented graphically in Fig. 4.8.

4.7 Studying bridging interactions in proteins

To date, several papers have described computational geometry studies of methionine aromatic interactions in proteins. [4, 6, 8] These works mention that methionine residues are within some cutoff distance of more than just one aromatic residue, but do not go any further. This may be, at least in part, due to the complexity of quantifying the incidences of such chain-like structures, which is an abstract problem that draws from more than just protein science. Herein, graph theory approaches were used to model bridging interactions,

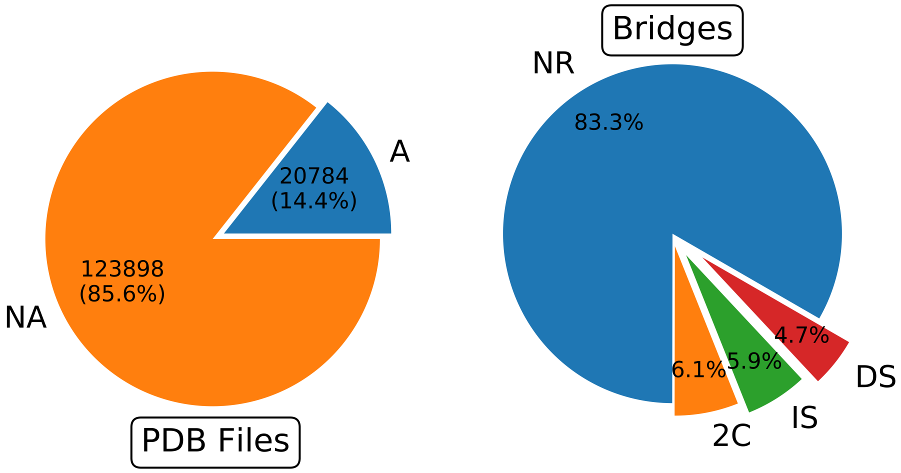


Figure 4.8. Left. NA: PDB files lacking either a closely spaced aromatic chain or a bridging interaction. A: PDB files containing at least one aromatic chain and at least one bridge. *Right.* Counts for geometries S in 20,784 PDB files. Wedges are labelled in line with equation 4.16 and Fig. 4.6.

where the inspiration for their use came from use in related research [13, 14, 15] in addition to their extensive usage in molecular modeling software.

4.8 Incidences of bridging interactions

Based on the survey described in this Chapter, we can propose that bridging interactions play several possible roles in proteins. First, the large number of bridging interactions across all proteins is a good indicator that these interactions serve a ubiquitous role in protein structural stabilization. Indeed, several clinically relevant examples underscore this observation. In von Willebrand disease (a bleeding disorder), substitution of Met239 in wild-type von Willebrand Factor A1-glycoprotein Ibα complex (PDB ID 1SQ0) with a valine residue results in nonspecific loss of hydrophobic contact. [16] For comparison, processing the 1SQ0 'B' chain with the Met-aromatic algorithm returns a Phe199 : Met239 : Phe201 bridging interaction.

An analysis of Fig. 4.1 reveals that 76% of bridging interactions contain exactly one redox-inactive phenylalanine residue with the opposing residue being any of Phe, Tyr or Trp. In addition, 24% (Fig. 4.1) of bridging interactions comprised solely of the redox-active residues Tyr and/or Trp. A possibility is that here, methionine residues are stabilizing closely spaced tyrosine/tryptophan residues in order to facilitate electron transfer. This idea is worthy of closer examination and experimental work, some of which is presently being carried out in the Warren laboratory.

4.9 Computational geometry and the scalene triangle

Another approach to delineating the potential roles of Met-aromatic bridges is to assess their location with respect to protein cofactors and surfaces. In this case, we were interested in determining whether the geometry of the scalene triangle model varied significantly between those bridges containing a Phe and those bridges whose residues were redox-active. Overall, a key finding here is that a significant number of bridging interactions reside at or very near to protein surfaces, regardless of redox-active or inactive terminal residues, with the **SF** coordinate being found on the methionine residue containing the bridging **SD** coordinate. (Fig. 4.2). This is especially interesting given that methionine is normally considered a hydrophobic residue that is located deep inside a protein vs. being located near the protein surface. The immediate possibility here is that bridging interactions serve a role in protein docking or recognition. Indeed, aromatic interactions are known to serve a role in molecular recognition, [17, 18, 19] and several examples highlight the presence of bridging interactions on protein surfaces.

The Met-aromatic algorithm was applied to the human estrogen receptor (PDB ID 1HCP), a protein capable of DNA recognition. [20] That analysis revealed a Phe31 : Met72 : Tyr19 bridge located on the protein's surface. Likewise, analysis of the EcoRI DNA-endonuclease recognition complex (PDB ID 1ERI) [21] yields two bridging interactions near the protein surface: Phe156 : Met157 : Phe163 and Phe168 : Met251 : Trp246. In another example, unrelated to DNA recognition, Met143 on the chemotaxis receptor recognition protein (PDB ID 1BC5) [22] is located on the surface of the protein and bridges Tyr184 and Tyr205, with an additional tryptophan nearby (Trp148). The localization of bridges at protein surfaces was unexpected and suggests potential roles at protein interfaces (i.e., protein-protein or protein-solvent). As such, the bridging interaction's role in molecular recognition warrants further investigation.

We were further interested in determining whether the bridging interaction had some geometric relationship to a metal. The average **MT / SD** and **MT / SF** distances were roughly in the 17 Å range. While there could be discrete examples of chemistry involving embedded cofactors and Met-aromatic bridges, the analysis presented here suggests that those instances are likely isolated.

4.10 Bridging interaction network analysis

Out of a total 119,038 bridge/chain pairs analyzed, a total of 19,962 consisted of an Aro : Met : Aro bridge serving as a member of a larger redox-active aromatic chain (where Aro is strictly any of Tyr or Trp). The most immediate conclusion that can be drawn is that here, the methionine is serving a structural role. The proposal from Gray and Winkler ([9]) that extended aromatic chains can serve protective roles provided motivation to approach this problem from a different perspective. For example, work by Cordes *et al.* [12] demonstrated that in a model analogous to a bridging interaction of form AroA : I : AroB, electron transfer was highly dependent on the identity of the side chain of I, with electron transfer from AroA to AroB occurring either via superexchange (electron transfer directly from AroA to AroB) or via hopping through I. The hopping process is about 20-30 times faster than superexchange. As such, the presence of an aromatic-methionine-aromatic bridge in a longer aromatic chain could have a dual protein stabilizing and electron transfer regulatory role. The bioinformatics work presented here suggests that Met-aromatic groups, especially bridges, could be involved in analogous long-range electron transfer reactions. This is a thought-provoking dataset and deserves closer inspection from an experimental perspective.

4.11 Conclusion

In this Chapter, extensive use of graph- and network-theory methods to study the dual bridging interaction in protein crystal structures in the PDB was made. Work started by enumerating bridges of differing aromatic residue type, resulting in the finding that the Phe : Met : Tyr bridge was most common in proteins. The first conclusion is that the majority of bridges serve a structural role. A minority of bridges, however, were comprised solely of the redox-active residues Tyr/Trp. Bridges whose aromatic members were strictly any of the redox-active Tyr/Trp residues were isolated and their relationship to longer closely spaced Tyr/Trp chains was assessed. It was found that in approximately 17% of cases, a dual bridging interaction had some geometric relationship to a closely spaced Tyr/Trp chain. This indicates that at least some bridging interactions could serve a role in modulating electron flow in biomolecules, perhaps via direct hopping through methionine or by superexchange. To further understand whether this could be the case, the geometric relationships between redox-inactive bridges and those bridges which are capable of redox activity to both metal sites and protein surfaces were analyzed using a scalene triangle model. No difference between the overall dimensions of the scalene triangle for redox-active and inactive bridges was found. However, a strong propensity for bridging interactions to localize on protein surfaces was noted. This finding was surprising given that methionine residues are normally buried deep within the protein volume, and this finding deserves a more thorough analysis, perhaps in the context of molecular recognition.

Bibliography

[1] Branden, C., and Tooze, J. (1991) Introduction to Protein Structure. Garland Pub.

[2] Ma, J. C., and Dougherty, D. A. (1997) The Cation π-Interaction. *Chem Rev.* **97**. 1303-1324

[3] Chakrabarty, B., and Parekh, N. (2016) NAPS: Network Analysis of Protein Structures. *Nucleic Acids Res.* **44**. W375-W382

[4] Reid K.S.C., Lindley P.F., and Thornton J.M. (1985) Sulphur-aromatic interactions in proteins. *FEBS Lett.*190. 209-213

[5] The PyMOL Molecular Graphics System, Version 1.7.x, Schrödinger, LLC.

[6] Valley, C. C., Cembran, A., Perlmutter, J. D., Lewis, A. K., Labello, N. P., Gao, J., and Sachs, J. N. (2012) The Methionine-aromatic Motif Plays a Unique Role in Stabilizing Protein Structure. *J Biol Chem.* **287**. 34979-34991

[7] Tatko, C. D., and Waters, M. L. (2009) Investigation of the nature of the methionine-π interaction in β-hairpin peptide model systems. *Protein Sci.* **13**. 2515-2522

[8] Weber, D. S., and Warren, J. J. (2018) A survey of methionine-aromatic interaction geometries in the oxidoreductase class of enzymes: What could Met-aromatic interactions be doing near metal sites? *J Inorg Biochem.* **186**. 34-41

[9] Gray, H. B., and Winkler, J. R. (2015) Hole hopping through tyrosine/tryptophan chains protects proteins from oxidative damage. *Proc Natl Acad Sci USA.* **112**. 10920-10925

[10] Hagberg, A., Swart, P., S Chult, D., (2008) Exploring network structure, dynamics, and function using networkx. Conference: SCIPY 08; August 21, 2008; Pasadena, CA, USA.

[11] Zauhar, R. J., Colbert, C. L., Morgan, R. S., Welsh, W. J. (2000) Evidence for a strong sulfur-aromatic interaction derived from crystallographic data. *Biopolymers.* **53**. 233-248

[12] Cordes, M, Kottgen, A., Jasper, C., Jacques, O., Boudebous, H., Giese, B. (2008) Influence of Amino Acid Side Chains on Long-Distance Electron Transfer in Peptides: Hopping via "Stepping Stones". *Agnew Chem.* **47**. 3461-3463

[13] Yan, Y., Zhang, S., Wu, F. X. (2011) Applications of Graph Theory in Protein Structure Identification. *Proteome Sci.* **9**. 1

[14] Vishveshwara, S., Brinda, K. V., Kannan, N. (2002) Protein Structure: Insights From Graph Theory. *J Theorl Comput Chem.* **1**. 187-211

[15] Canutescu, A. A., Shelenkov, A., A., Dunbrack, R. L. (2003) A Graph-Theory Algorithm for Rapid Protein Side-Chain Prediction. *Protein Sci.* **12**. 2001-14

[16] Dumas, J. J., Kumar, R., McDonagh, T., Sullivan, F., Stahl, M. L., Somers, W. S., Mosyak, L. (2004) Crystal Structure of the Wild-Type von Willebrand Factor A1-Glycoprotein Ibα Complex Reveals Conformation Differences with a Complex Bearing von Willebrand Disease Mutations. *J Biol Chem.* **279**. 23327-34.

[17] Hunter, C. A. (1994) Meldola Lecture. The Role of Aromatic Interactions in Molecular Recognition. *Chem Soc Rev.* **23**. 101-9

[18] Meyer, E. A., Castellano, R. K., Diederich, F. (2003) Interactions with Aromatic Rings in Chemical and Biological Recognition. *Agnew Chem.* **42**. 1210-1250

[19] Muehldorf, A. V., Van Engen, D., Warner, J. C., Hamilton, A. D. (1988) Aromatic-Aromatic Interactions in Molecular Recognition: A Family of Artificial Receptors for Thymine That Shows Both Face-to-Face and Edge-to-Face Orientations. *J Am Chem Soc.* **110**. 6561-6562

[20] Schwabe, J. W., Chapman, L., Finch, J. T., Rhodes, D., and Neuhaus, D. (1993) DNA Recognition by the Oestrogen Receptor: From Solution to the Crystal. *Structure.* **1**. 187-204

[21] Kim, Y. C., Grable, J. C., Love, R., Greene, P. J., Rosenberg, J. M. (1990) Refinement of Eco RI Endonuclease Crystal Structure: A Revised Protein Chain Tracing. *Science.* **249**. 1307-9

[22] Djordjevic, S., Stock, A. M. (1998) Chemotaxis Receptor Recognition by Protein Methyltransferase CheR. *Nat Struct Mol Biol.* **5**. 446-50

[23] Reece, S.Y., Seyedsayamdost, M. R. (2017) Long-range proton-coupled electron transfer in the Escherichia coli class Ia ribonucleotide reductase. *Essays Biochem.* **61**. 281-292

Chapter 5

Warren Lab Fluorescence/TA Spectrometer

5.1 Preface

In Chapters 2-4, the Met-aromatic algorithm and its application to the Protein Data Bank were discussed. The resultant findings of the Met-aromatic algorithm yielded more questions than answers, however. For example, could the bridging interaction described in Chapter 4 have some functional significance? After all, a survey of protein crystal structures revealed a total of 115,030 bridging interactions! (or 0.82 2-bridges per protein). These are the sorts of questions that data-driven approaches fail to answer. Instead, the answers to these questions lie in raw experimentation.

To address some of the questions raised through the bioinformatics work described in this thesis, other researchers in the Warren lab are investigating electron flow in artificial protein systems. An example of one such system is shown in Fig. 5.1. Photosensitizer-modified proteins have been used to investigate and rationalize intraprotein electron transfer reactions for almost 30 years. [1] Of special importance to the Warren lab are those systems that have been used to probe intraprotein redox reactions involving tyrosine and tryptophan ([2] and [3]). In either case, nanosecond time-resolved emission and absorbance spectroscopy are used to monitor the redox reactions of interest. A brief overview of the relevant photochemical processes, and their relation to the work described in this Chapter, are given below.

In the hypothetical system shown in Fig. 5.1, a Ru^{3+} oxidant can be generated, which triggers follow-up redox reactions involving the Met:Trp and the Cu. By investigating the

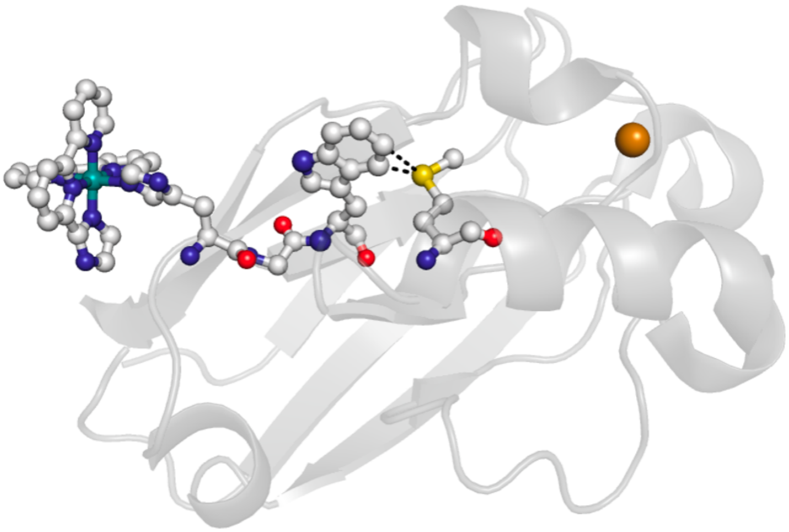


Figure 5.1. A PyMOL-generated model of *Pseudomonas aeruginosa* azurin labelled with Ru(II)(2,2'-bipyridyl)$_2$(imidazole) at histidine 107. The native Cu ion in azurin is shown as an orange sphere, and a proposed Met:Trp interaction is shown between the Ru and Cu sites.

kinetics of these reactions, proposals can be made about the role(s) of the Met:Trp moiety. The Ru(III) is generated using the flash-quench technique (e.g., Fig. 5.2) [4]. A pulse from a nanosecond laser is used to excite the Ru(II) label. The electronically excited state, $^*\text{Ru}^{2+}$, can transfer an electron to an exogenous acceptor (e.g., $\text{Ru(NH}_3)_6\text{Cl}_3$) to yield Ru^{3+}. Intramolecular electron transfer events follow and can be monitored using the appropriate hardware. My role in this effort was to engineer software (and some associated hardware) for the operation of a nanosecond transient absorption/fluorescence spectrometer. In this Chapter, the underpinnings of this system are described.

5.2 Fluorescence spectroscopy

The Warren lab laser software operates two spectroscopic techniques. The first is the time resolved fluorescence option. Here, a pulsed Nd:YAG laser is used to excite a sample, and fluorescence decay is monitored at some wavelength that is selected using a monochromator. An example of a fluorescence trace is shown in Fig. 5.3. Here the fluorescence trace f is piecewise defined and is a function of time t. A typical Warren lab fluorescence trace is fitted by two splines: the first spline, defined on the region $(-\infty, 0]$, is the pre-pump region.

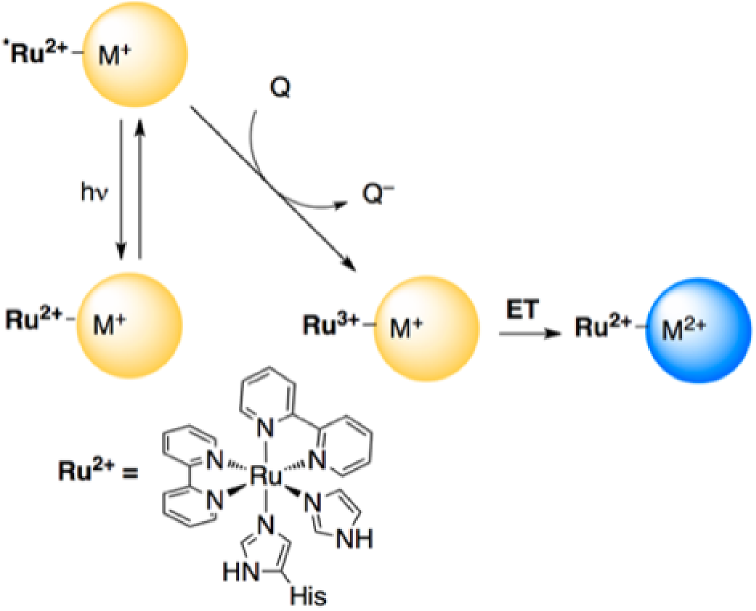


Figure 5.2. Oxidative flash-quench scheme for a Ru-modified protein (shown as yellow and blue circles). Q is any electron acceptor capable of reacting with excited $\mathrm{Ru}^{2+} = (^*\mathrm{Ru}^{2+})$. ET stands for electron transfer.

In this time frame, the sample is sitting idly and the monochromator is collecting some Gaussian noise $\delta(t)$. At time 0, an Nd:YAG pump pulse arrives and instantaneously excites the sample. The second spline, defined on the region $(0, \infty)$, describes the time window in which the sample fluorescence decays exponentially. The function f describing fluorescence can be described using the piecewise relation (5.1).

$$f(t) = \begin{cases} \delta(t), & t \in (-\infty, 0] \\ Ae^{-kt} + \delta(t), & t \in (0, \infty) \end{cases} \tag{5.1}$$

The piecewise function (5.1) only accounts for the dynamics of the fluorescing system. A fluorescence trace also consists of an instrument response function g. g is a mapping between photon flux into the PMT and events detected by an oscilloscope. The instrument response function g is typically approximated using a Gaussian fit where $\mu = 0$ (i.e. the distribution is centred at $t = 0$ (5.2):

$$g(t) \approx \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2} . \tag{5.2}$$

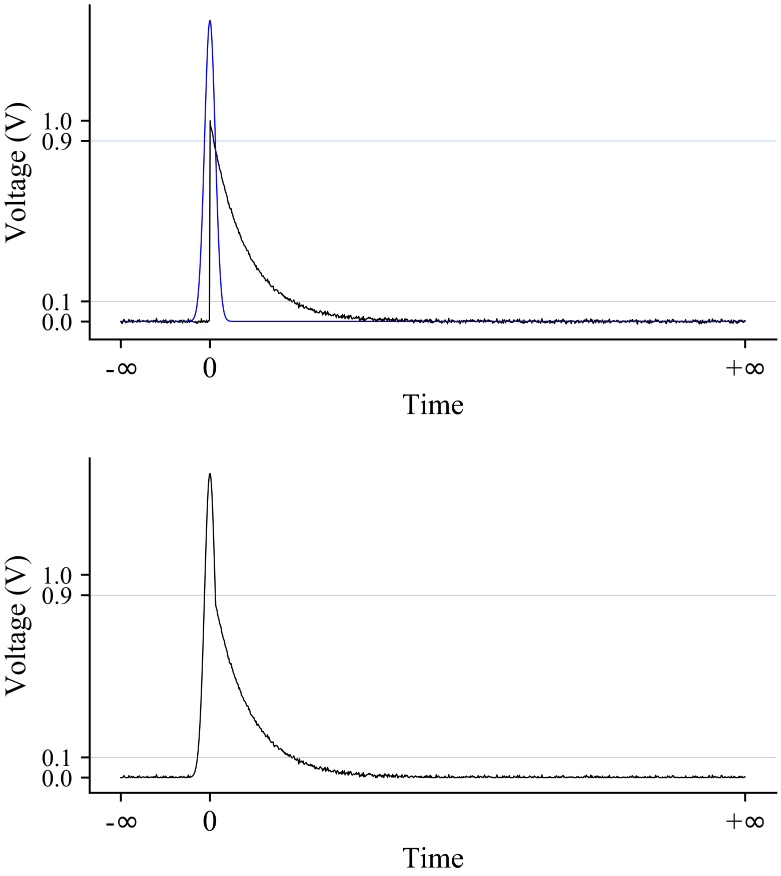


Figure 5.3. A representation of the anatomy of fluorescence traces collected in the Warren lab. *Top*: The black trace (f) depicts the dynamics of a UV-Vis active system. The blue trace depicts the instrument response function g. Here μ is approximated at 0 (see equation 5.2). *Bottom*: A plot of the convolution of f and g: $\int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau$.

The resultant trace is a convolution of f and g (5.3) and an approximation is shown in Fig. 5.3, bottom. The asterisk denotes the convolution operator and should not be interpreted as the product operator.

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau \tag{5.3}$$

A user must ultimately deconvolute $\int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau$ manually by choosing a subset of data through which the second spline of f can be fit such that a decay constant can be obtained.

5.3 Transient absorption spectroscopy

TA control software was developed alongside the aforementioned fluorescence software to allow the study of absorbance events in proteins down to 25 nanosecond resolution. In the TA spectroscopy technique, we use a laser pulse (pump) to excite a sample which in turn generates some non-equilibrium state (such as the promotion of an electron to an antibonding orbital). A probe light is then used to monitor the sample behaviour as the system recovers to an equilibrium state.

5.4 Detecting light

For both fluorescence and TA spectroscopy, we are essentially counting photons of some specific energy, with a particular interest in the rate of change of photon counts as a function of time. In the Warren lab, we collect raw photon counts using a Hamamatsu R928 photomultiplier tube (PMT). PMTs are ultra-sensitive light-detecting vacuum tubes that operate on the principle of the photoelectric (PE) effect. Photons entering the PMT strike a photocathode, generating electrons by way of the PE effect. These electrons are then focused onto a primary dynode stage, resulting in the generation of secondary electrons by way of secondary emission. The stream of secondary electrons strikes a second dynode stage, subsequently generating more electrons that strike a third dynode stage. This electron flow continues over several dynode stages, proportionally amplifying the current produced by input of incident light upwards of millions of times. Our PMT is connected to a FEMTO DHPVA-200 wideband voltage amplifier. This device amplifies the voltage generated by the PMT. Output from the voltage amplifier is then directly routed to a stack of dual-channel PCI digitizer boards. These are essentially oscilloscopes built directly into a PC. The Warren lab uses two digitizer boards: the CS12502 and the CS8422. The CS12502 is used for short timescale experiments, namely those experiments whose events either require monitoring at the nanosecond resolution and/or events which occur on the order of nanoseconds to a few hundred microseconds. The CS8422 has poorer time resolution but allows for monitoring events up to several seconds post flash. The CS12502 and CS8422 digitizer boards serve a

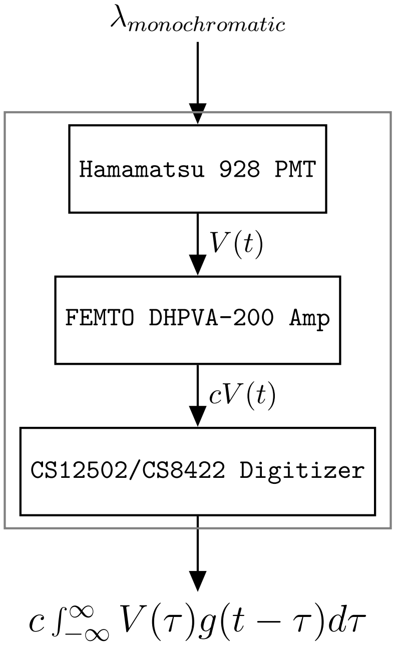

$\lambda_{monochromatic}$
Hamamatsu 928 PMT
$V(t)$
FEMTO DHPVA-200 Amp
$cV(t)$
CS12502/CS8422 Digitizer
$c\int_{-\infty}^{\infty} V(\tau)g(t-\tau)d\tau$

Figure 5.4. A block diagram depicting the transmission of data from an experiment to some convolved function. Photons are collected by the PMT and mapped to some voltage function V(t). The voltage function V(t) is scaled by some constant c imposed by the amplifier. cV(t) is then fitted to the convolution $c\int_{-\infty}^{\infty} V(\tau)g(t-\tau)d\tau$ by way of the digitizer hardware. $f(t)$ can be deconvoluted from $c\int_{-\infty}^{\infty} V(\tau)g(t-\tau)d\tau$ and modified such that a user can obtain a fluorescence decay constant, as an example.

critical role: these devices map fluorescence (or absorbance) data into digital space, subsequently allowing such data to be processed downstream using computational methods. The digitizer boards are also responsible for sampling potentials at a specific frequency. Fig. 5.4 depicts the flow of data in a TA/fluorescence experiment.

5.5 Wavelength selection

Section 5.4 provided a description of digital data collection. How do we select the "type" of photon striking the PMT photocathode however? We use a DK240 monochromator operating on a Czerny-Turner mechanism [5] to select photons of a particular energy. Light being emitted from our sample is collimated into the entry port of the DK240, upon which it hits the entrance slit. The entrance slit permits some number of photons of any energy to pass. Photons passing through the slit are then narrowly selected by way of the internal Czerny-Turner mechanism. A selected number of photons within some very narrow range

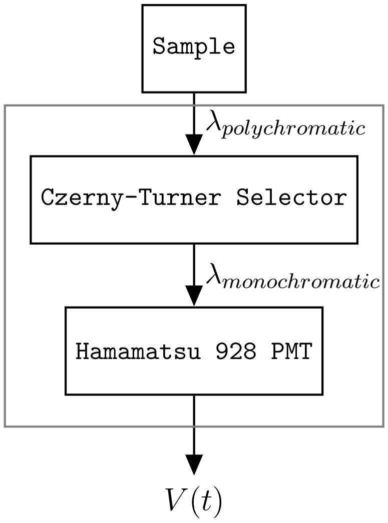


Figure 5.5. A block diagram depicting the transmission of light from sample to PMT.

of wavelengths then pass through the exit slit and into an aluminum enclosure housing our PMT. This is shown in Fig. 5.5.

5.6 Pulse train control

The Warren Lab TA/fluorescence setup is controlled using a master/slave configuration. A Continuum laser is pulsed at a frequency of 10 Hz using a digital delay generator (QC9514). The digital delay generator controls the laser using a 5V TTL (transistor-to-transistor) pulse. The pulsed laser beam is split into two beams using a pickoff mirror, where one beam projects onto a silicon photodiode and the other beam projects directly into a sample holder containing a cuvette. A small change in voltage is observed upon illumination of the Si photodiode. This change in voltage serves as an external trigger that forces the CS12502/CS8422 digitizers to begin data acquisition from the PMT (V(t)). For TA experiments, an additional master is synced to the 10 Hz output from the QC9514. This second master channel flashes the arc lamp at an identical frequency of 10 Hz with a 60 μs offset from the primary master.

5.7 Data acquisition

Data is acquired in 2 steps. First, a matrix of background data is collected. A primary shutter is opened, allowing for the pulsed laser beam to strike the pickoff filter. Part of this beam then triggers data acquisition by striking the Si photodiote. The other part of this beam is blocked from striking the sample cuvette by way of a secondary shutter. The digitizer then collects a background data matrix D_b of dimension $N \times s$ (5.4), where N is a user-defined number of shots and s refers to the number of samples collected by the digitizer in a given timeframe.

$$D_b = \begin{bmatrix} V_1(t_1) & V_2(t_1) & V_3(t_1) & \dots & V_N(t_1) \\ V_1(t_2) & V_2(t_2) & V_3(t_2) & \dots & V_N(t_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ V_1(t_s) & V_2(t_s) & V_3(t_s) & \dots & V_N(t_s) \end{bmatrix} . \tag{5.4}$$

The secondary shutter is then opened, allowing for excitation of the sample. A secondary experimental data matrix D_e is collected. D_e is of dimension equivalent to D_b:

$$D_e = \begin{bmatrix} V_1(t_1) & V_2(t_1) & V_3(t_1) & \dots & V_N(t_1) \\ V_1(t_2) & V_2(t_2) & V_3(t_2) & \dots & V_N(t_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ V_1(t_s) & V_2(t_s) & V_3(t_s) & \dots & V_N(t_s) \end{bmatrix} . \tag{5.5}$$

A background correction is performed by subtracting D_b from D_e (5.6):

$$D_{corr} = D_e - D_b \, . \tag{5.6}$$

Signal averaging is then performed according to equation 5.7, where $V_n(t_s)$ are entries at indices n, s in D_{corr}. Here we take the average of all columns:

$$\overline{D_{corr}} = \frac{1}{N}\begin{bmatrix}\sum_{n=1}^{N} V_n(t_1)\\ \sum_{n=1}^{N} V_n(t_2)\\ \vdots \\ \sum_{n=1}^{N} V_n(t_s)\end{bmatrix}\,. \tag{5.7}$$

The sequence of background corrections $\overline{D_{corr}}$ is performed S number of times. We termed these "groups of shots." The $1 \times s$ $\overline{D_{corr}}$ column vectors for S number of groups are then averaged:

$$\frac{1}{S}(\overline{D_{corr1}} + \overline{D_{corr2}} + \ldots + \overline{D_{corrS}})\,. \tag{5.8}$$

The entries in 5.8 can ultimately be fitted to the convolution (5.3).

5.8 Acquisition Control

Data acquisition is controlled via a shutter open/close sequence. No data acquisition takes place when the primary shutter is closed (Fig. 5.6, left) because the photodiode is not generating a trigger event. The primary shutter opens when prompted by a user, which directs a portion of the laser beam to the photodiode (Fig. 5.6, middle) thus forcing the acquisition of data into matrix D_b. The secondary shutter opens after the Si photodiode has counted N number of "blank" shots (Fig. 5.6, right). At this point, N is reset to 0 and data for matrix D_e is gradually loaded into a buffer. Data is processed using the sequence of equations 5.6, 5.7 upon reaching N number of shots. The entire sequence of events is repeated S number of times. The sequence of operations is identical for a TA experiment with one exception: a third shutter (shutter C, Fig. 5.6) is held open throughout the experiment. This shutter permits light from the arc lamp to reach the sample.

5.9 Software

Operation of the Warren laboratory laser table is for the most part automated. All programming was done in the MATLAB programming language. The layout of the user interface

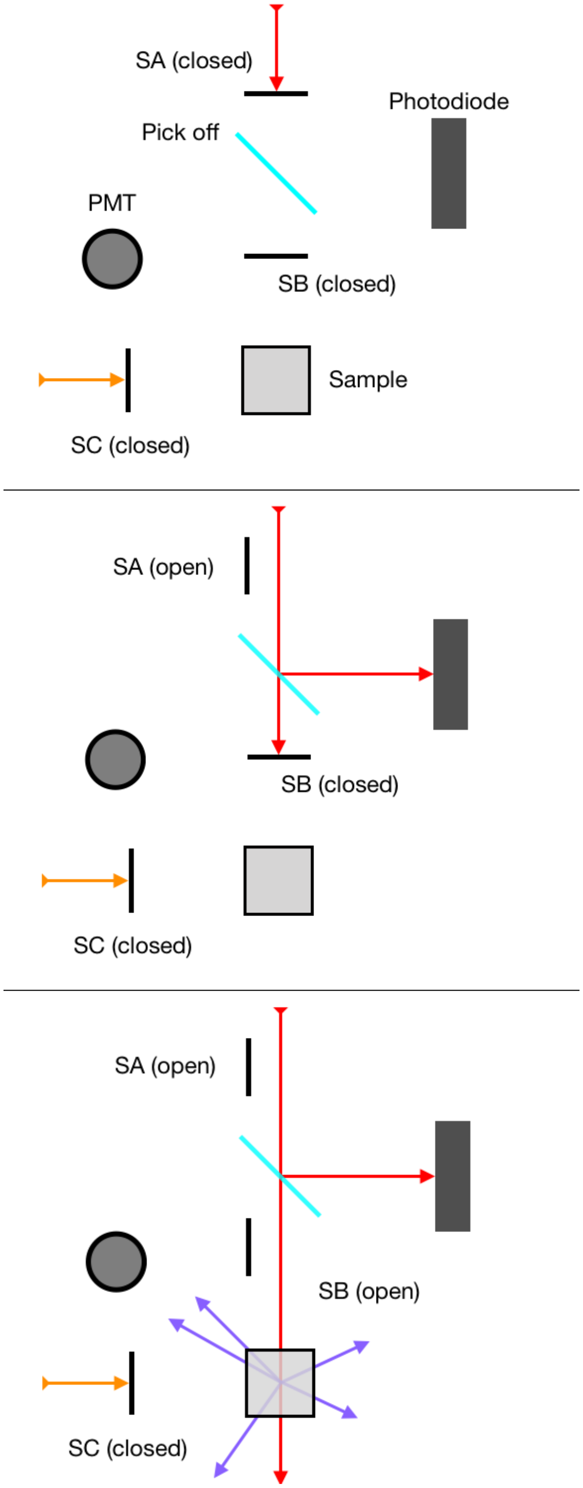
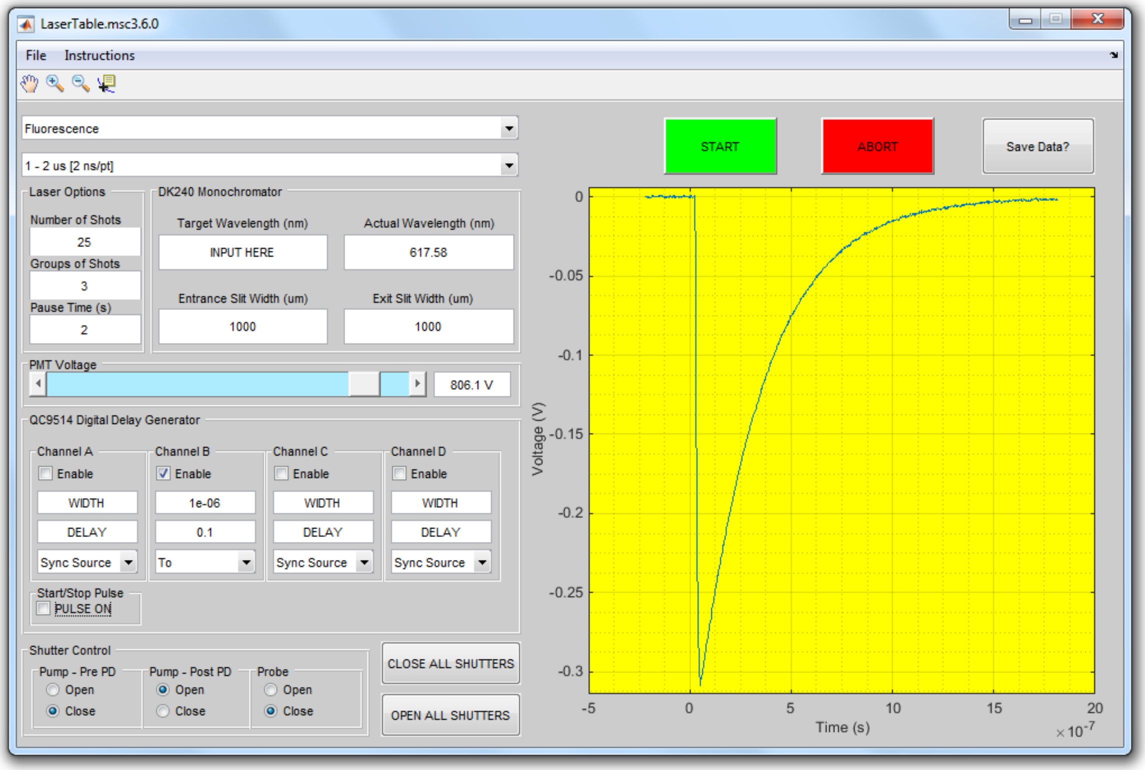


Figure 5.6. The fluorescence shutter sequence. *Top*: All shutters are closed. No data acquisition is taking place. *Middle*: The primary shutter is opened. A blank read is taking place. *Bottom*: The secondary shutter is opened. The laser beam excites the sample and luminescence data is being collected by the PMT.


LaserTable.msc3.6.0

File Instructions

Fluorescence

1 - 2 us [2 ns/pt]

Laser Options

Number of Shots: 25

Groups of Shots: 3

Pause Time (s): 2

DK240 Monochromator

Target Wavelength (nm): INPUT HERE

Actual Wavelength (nm): 617.58

Entrance Slit Width (um): 1000

Exit Slit Width (um): 1000

PMT Voltage: 806.1 V

QC9514 Digital Delay Generator

Channel A	Channel B	Channel C	Channel D
□ Enable	☑ Enable	□ Enable	□ Enable
WIDTH	1e-06	WIDTH	WIDTH
DELAY	0.1	DELAY	DELAY
Sync Source	To	Sync Source	Sync Source

Start/Stop Pulse: □ PULSE ON

Shutter Control

Pump - Pre PD	Pump - Post PD	Probe
○ Open	● Open	○ Open
● Close	○ Close	● Close

Figure 5.7. A screenshot of version 3.6.0 of the Warren laboratory laser software.

is shown in Fig. 5.7. The user interface closely parallels the descriptions in the preceding sections. A user first selects either the fluorescence or TA mode. Next the user selects the timebase (the time interval which being sampled over); 2 μs was used in the figure. Note the sampling rate of 2`ns/pt`. This is the number of nanoseconds between each read (2,000 reads in 2 μs). The value of N is selected in the `Number of Shots` edit box and the S number is chosen in the `Groups of Shots` edit box. Therefore, from Fig. 5.7 the signal average is computed from three matrices D of shape $25 \times (2,000\ ns\ /\ 2\ ns\ pt^{-1})$. Wavelengths and slit widths (see Fig. 5.5) at which an experiment is to be monitored are selected under the `DK240 Monochromator` panel. Voltage needs to be applied to the PMT to detect weak light signals. This is done by selecting an 800 V potential using the `PMT Voltage` slider bar. The `Start/Stop Pulse` button triggers the 10 Hz TTL pulse from the QC9514 DDG thus pulsing the laser. At this stage, a user presses `START` which opens the primary shutter and begins data acquisition.

5.10 Conclusion

The data mining operations that have been described in the previous chapters have raised important questions that can be answered with experiment. To assist in addressing some of these questions, part of this research has been focused on developing software for operating the Warren laboratory fluorescence and transient absorption spectrometer. Here, the ultimate goal is that other graduate students can develop protein models based on some of the geometries found using the Met-aromatic algorithm (such as the bridging interaction) and begin to probe their significance *in vitro*.

Bibliography

[1] Chang, I.J., Gray, H. B., Winkler, J. R. (1991) High-driving-force electron transfer in metalloproteins: intramolecular oxidation of ferrocytochrome *c* by Ru(2,2'-bpy)$_2$(im)(His-33)$^{3+}$. *J Am Chem Soc.* **113**. 7056-7057

[2] Shih, C., Museth, A. K., Abrahamsson, M., Blanco-Rodriguez, A.M., Di Bilio, A.J., Sudhamsu, J., Crane, B.R., Ronayne, K.L., Towrie, M., Vlcek, A Jr., Richards, J. H., Winkler, J. R., Gray, H. B. (2008) Tryptophan-accelerated electron flow through proteins. *Science.* **320**. 1760-1762

[3] Warren, J. J., Herrera, N., Hill, M. G., Winkler, J. R., Gray, H. B. (2013) Electron Flow through Nitrotyrosinate in *Pseudomonas aeruginosa* Azurin. *J Am Chem Soc.* **135**. 11151-11158

[4] Gray, H. B., Winkler, J. R. (2009) Electron Flow through Proteins. *Chem Phys Lett.* **483**. 1-9

[5] Czerny, M., Turner, A. F. (1930) Über den Astigmatismus bei Spiegelspektrometern. *Z Phys.* **61**. 792

Appendix A

Code

All relevant code is available at: https://github.com/dsw7