Effects of Physicochemical and Functional Constraints on the Sequence and Structure of Proteins

BY

DAVID JIMENEZ-MORALES
B.S., University of Granada, Granada, Spain, 1998
M.S., University Complutense of Madrid, Madrid, Spain, 2004

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Bioinformatics
in the Graduate College of the
University of Illinois at Chicago, 2013

Chicago, Illinois

Defense Committee:

Jie Liang, Chair and Advisor
Yang Dai
Linda Kenney, Microbiology & Immunology
Robert Eisenberg, Rush University Medical Center
Dashuang Shi, George Washington University

This thesis is dedicated to my parents Adriano and Carmen, my sister Esther, my brother Adri, my brother and sister in law Gabriel and Virginia, my nieces and nephew Marina, Paula, Lucia, and German, and of course, to my wife Nieves, my daughter Sara, and my son Diego (who arrived on time!). Because they are my family, the most important thing in life.

ACKNOWLEDGMENTS

I want to thank my advisor Prof Jie Liang for being my mentor, for his support, and for the many valuable lessons that I have learned from him among all these years. As a result, not only I feel that now I am a scientist, but also a better person.

I am also thankful to the faculty members of my thesis committee, Drs Yang Dai, Bob Eisenberg, Linda Kenney, and Dashuang Shi for taking the (every day more expensive) time to serve on my thesis committee. Your encouragement, teachings, and insightful comments are greatly appreciated.

It is completely out of scale the immense gratitude and appreciation that I feel for the person that I consider the co-author of this dissertation: Nieves Lopez Barrera, my closer partner in this journey. Her generosity, sacrifices, kindness, understanding, support, and patience are priceless. I am not sure if I have deserved such treatment, but all what I know is that I am in debt with her for the rest of my life.

I want to thank Jun Tian, our lab technician at Liang's lab, who carried out the experimental part. Without him, we simply could not test our predictions. I thank his patience in our long discussions.

I am also very thankful to my fellow colleagues Yun Xu, Gamze Gurzoy, Meishan Lin, Youfang Cao, Jieling Zhao, Ke Tang, and Marco Maggioni for their scientific comments, discussions, and suggestions, but also for the priceless daily help, for sharing feelings in the good and bad moments, and for being always there when I needed them.

And I cannot –of course– forget my former lab mates Hsiao-mei Lu, Joe Dundas, Larisa Adamian, Zheng Ouyang, Hammad Naveed, Jingzi Li, and Volga Pasupuleti, which kindly opened the door to the lab, took my hand, and guided me when I was (very lost) taking my first steps.

I also want to thank the colleagues from the remaining groups of the Bioinformatics lab, specially to Matt Carson, who has always shown an overwhelming generosity in many different matters. I also thank Damian Roqueiro, Lei Huang, Georgi Genchev, and Adam Carlson, who have always been there to solve questions, provide useful comments, and gave me a hand when I needed it.

I am also very thankful to my collaborators, the outstanding scientists from whom I have had the opportunity to enrich and enlarge my scientific vision in many different aspects. Particularly, I am tremendously thankful to Dr. Kenney for opening the door of her lab to me, letting me attend her group meetings, and very kindly sharing her thoughts, visions, and knowledge. From Dr. Kenney's lab I have to thank Leslie Morgan, an incredibly professional and beautiful person, who has always been available to answer any question I have had. I am also very thankful to Dr. Eisenberg, a scientist full of wisdom, energy, and boundless passion. I am thankful for the many times he kindly invited me to his office (with delicious coffee), shared his knowledge, and made possible a very successful collaboration. I am very thankful to Dr. Daniel Shi, not only because he planted the seed of my last (amazing) research project, but also his trip to Chicago for my thesis defense.

ACKNOWLEDGMENTS (Continued)

I am thankful to many other scientists who have had a vital impact, in one way or another, on the research that I present here. Among them, I must thank Dr. Jacinto Lopez, who took me in an unforgettable walk through the reciprocal space that changed my vision of the process of protein structure determination. I am also very thankful to Dr. Andrew Binkowski, Dr. Andrzej Joachimiak, Dr. Pedro Brugarolas, Dr. Karl Volz, Dr. Jose V Moyano, and Dr. Marcial Escudero, for their very important contributions.

I want to thank Dr. Muhammad Qasim, Dr. Cristian Luciano, and Dr. Silvio Rizzi, wonderful friends that I met at this University, who generously shared vast information that saved me priceless time.

Finally, there are many rewards linked to study a PhD in a different country. It is a life changing opportunity. It is a mind changing experience. It is a inexhaustible source of enrichments. And the opportunity of learning a new language is one of them, although it is a journey full of frustration. And last but not least, I want to thank the people who have been specially patient in helping me with the language that I am using to write this dissertation, specially Barbara Wolfe Boockmeier, essential in my first steps, and my American brother Brian Furey.

PREFACE

Life, with all its forms and mechanisms, is a truly fascinating event full of complexity. My personal aspiration is to understand as much as I can about life and the principles that make it possible. As a biologist, I have created my own vision of life. A "proteocentric" vision. In this vision, proteins are the beginning and the end, the meaning and purpose, the ultimate reason that explains why we are here, in this form, in this manner.

In the research that I have carried out along my PhD studies and summarized in this dissertation, I have had the opportunity of investigating many aspects related to proteins. I now know more than when I started my studies, but I am still seeking answers.

For example, I wonder whether we will be able to understand the amino acid substitutions characterizing every single protein family, to the extent that would make us capable of designing proteins with any desired feature.

I wonder whether we will be able to fully understand the secrets of the cavities where protein enzymes execute their functions, to the extent that would let us identify the function of the protein by just looking at the bunch of amino acids converging at that magical site.

I wonder if we will be capable of anticipating the amino acid in a particular active site that requires a specific post-translational modification, allowing the protein to successfully carry out its function.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

FIGURE **PAGE**

LIST OF ABBREVIATIONS

bbTM	β-barrel Transmembrane Matrix
CharDen	Charge Density
PreLysCar	Predictor of Lysine Carboxylation
FDR	False Discovery Rate
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
MS	Molecular Surface
ASP	Active Site Pocket
K-S	KolmogorovSmirnov
TM	TransMembrane

SUMMARY

Proteins are biological macromolecules that carry out essential functions in living organisms. The sequence and structure of proteins are constrained by numerous factors. In this thesis, we use computational methods to examine the effects of some physicochemical and functional constraints at different levels of the sequence and structure of proteins.

First, we study how the hydrophobic environment of cellular membranes directs the evolution of membrane proteins. We specifically focused on a class of membrane proteins adopting the β-barrel topology. These β-barrel membrane proteins are found in the outer membrane of many important disease-causing bacteria. Their homologs in the eukaryotic mitochondria are also important as they are involved in the onset of apoptosis with implications in cancer, degenerative diseases, and aging. We have estimated the instantaneous substitution rates of amino acids inserted within the membrane of bacterial β-barrel membrane proteins. Scoring matrices were derived from these estimated rates, significantly improving the detection of homologs of β-barrel membrane proteins through database searches. The estimated amino acid substitutions were also used to identify residues that contribute significantly at the protein-protein interface between the subunits of OmpF, a trimeric β-barrel membrane protein. By replacing these residues through site-directed mutagenesis with residues that do not occur in outer membrane proteins, we succeeded in engineering an OmpF mutant with monomeric oligomerization state, instead of the natural trimeric form.

Second, we analyzed the physicochemical features of functional cavities of protein enzymes. Enzymes enable biochemical processes that would be otherwise infeasible. The catalytic reaction usually takes place in a surface pocket on the protein structure where substrates, ions, and amino acids interact. We measured the number density of ionizable residues found in these surface pockets. We found that the number density is extraordinarily large and relatively constant among a variety of enzymes. The catalytic active site of these enzymes is an unusual electrostatic and steric environment in which side chains and reactants are crowded together in a mixture more like an ionic liquid than an ideal infinitely dilute solution. The electrostatics and crowding of reactants and side chains seems likely to be important for catalytic function. Acid and base side chains are reliable markers of catalytic active sites, allowing us to use computational methods to identify with high accuracy the active site of protein enzymes.

Finally, we study lysine carboxylation, a post-translational modification that occurs spontaneously under certain physicochemical conditions, with a critical role in the catalytic mechanism of several important enzymes. We have characterized the signature microenvironment of carboxylation sites and its defining features. We have also developed a computational method for the detection of lysine carboxylation in proteins with available 3D structure, with excellent performance in distinguishing carboxylated lysine residues from a large number of unmodified lysine residues. Accurate prediction allowed us to assess the likely prevalence of lysine carboxylation in the proteome through large scale computations. Our results suggest that about 2% of proteins with more than 200 residues in both prokaryotes and eukaryotes may contain a carboxylated lysine residue, a three-fold increase of what it is currently known. Our results

also suggest that spontaneous post-translational modifications, by switching enzymes on-and-off under appropriate physical-chemical conditions, may frequently serve as an efficient biological machinery for regulation.

CHAPTER 1

INTRODUCTION

Life is a fascinating event occurring in the universe. Any contiguous living system is characterized by some defining features, such as an organized structure, the ability to sustain existence by using the necessary substances, the ability to respond to stimuli, and the fundamental ability to perpetuate in time.

Many macromolecules make life possible as we know it. Among them, proteins are essential, the cornerstone of life. They are characterized by an immense repertoire of functional roles and a remarkable versatility. Consequently, a detailed understanding of protein's nature is necessary toward the comprehension of life itself. In this thesis, we use computational methods to study the effects of physicochemical and functional constraints on different levels of the sequence and structure of proteins.

This introductory chapter is organized as follows. We begin reviewing the basic chemistry, structural organization, and fundamental principles about proteins. Through a brief historic review, we visit the key discoveries that built the foundations of our current knowledge of this complex macromolecule. We also review the most significant achievements in the relatively young discipline of computational biology, with special emphasis on the pillars under which part of this work rely. Finally, we introduce the main aims of the research here presented and its significance, concluding with the outline of the remaining chapters of the dissertation.

1.1 Notes about Proteins

1.1.1 Basic Concepts

Amino acids are the building blocks, the structural units of proteins (Nelson and Cox, 2004). An amino acid is an organic compound that contains a main atom of carbon, known as α-C, with four bonded groups. As its name implies, one of the groups is an amino group (-NH_2) and another is an acidic group (-COOH). Completes the compound an atom of hydrogen and an organic substituent known as the *side chain*.

The side chain confers the physical and chemical properties to amino acids. There are twenty standard amino acids encoded by the universal genetic code of living organisms. Although all different, amino acids can be grouped according to common physicochemical properties shared by their side chains.

Two different molecules of amino acids can form a di-peptide through the formation of a *peptide bond*, which takes place between the amino group of one amino acid and the carboxylic group of the other. The formation of a peptide bond is a dehydration synthesis reaction. Polypeptides are chains of amino acids held together by peptide bonds. One or several polypeptide chains is what we know as *proteins*.

In order to carry out their function, proteins must adopt a particular shape or *fold*. The assembly of the protein is called *folding*. In theory, the number of possible pathways that a given amino acid chain can follow is immense. However, a protein can *select* one in microseconds. How a protein chooses its pathway remains as an intriguing mystery.

The description and understanding of proteins is facilitated by four levels of protein structure commonly defined (Figure 1). **Primary structure** refers to the polypeptide sequence, including disulfide bonds (formed between the thiol groups of cysteine residues).

Secondary structure refers to recurring structurally local patterns of amino acids in proteins. Two types of particularly stable secondary structures occur widely in proteins. These are the *alpha*-helix and *beta*-strand. In the α-helix, the polypeptide chain is tightly wrapped around the imaginary axis drawn longitudinally through the middle of the helix. The α-helix is stabilized by hydrogen bonds between NH and CO groups. The amino acid side chains are projected outward from the helical backbone. In β-strands, the backbone of the polypeptide chain is extended into a zig-zag. Two different polypeptide chains adopting this conformation can be arranged side by side to form a structure stabilized with hydrogen bonds between the two adjacent segments.

Tertiary Structure refers to the overall three-dimensional arrangement of all atoms in a protein. It is usually made up of secondary structural elements and unordered sections.

Finally, since some proteins contain two or more separate polypeptide chains, the arrangement of these protein chains in three-dimensional complexes is referred as the **quaternary structure**.

1.1.2 Main Functions

Proteins take part in nearly every activity and structure of life. Among the most important functions, the capacity of catalyzing chemical reactions is truly remarkable. Protein enzymes

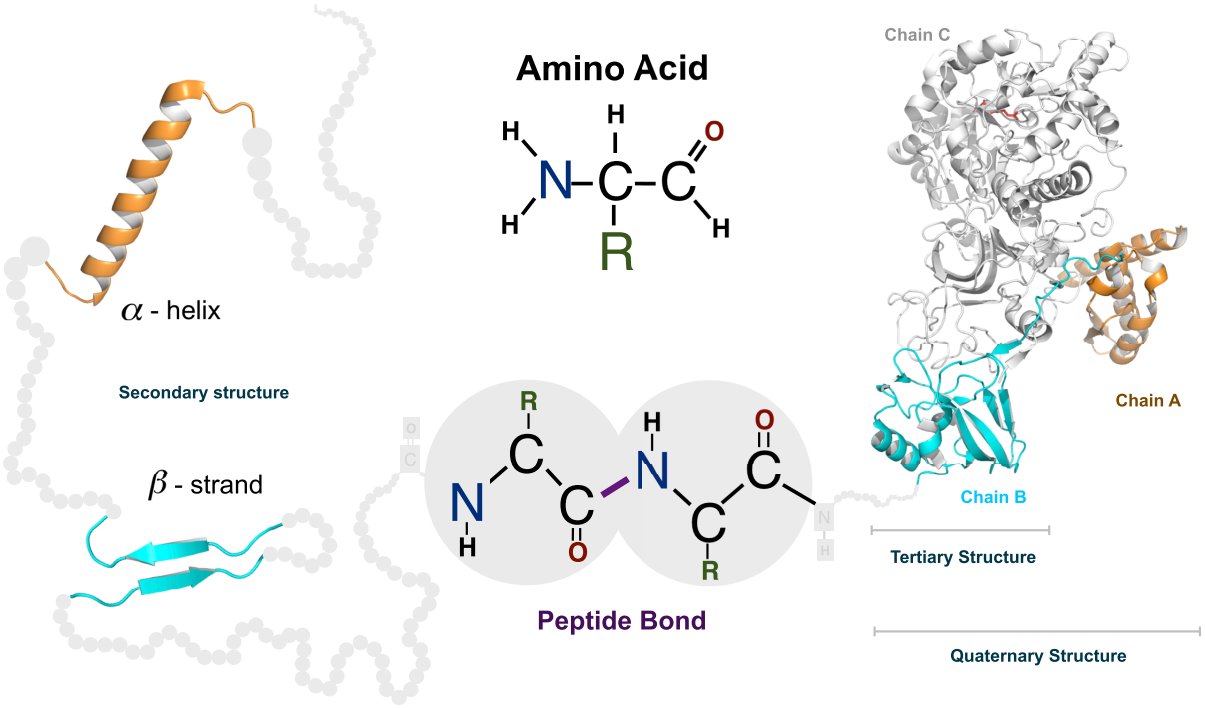

Figure 1. Summary of basic concepts about proteins

make feasible thousands of reactions that would be energetically too expensive otherwise. Enzymes are the biggest and most important group of proteins.

Another important function is transportation or storage. Classical examples of transport proteins are haemoglobin and myoglobin, specialized in the transport of oxygen. In addition, a large number of proteins inserted in biological membranes (membrane proteins) have the function of transportation, from electrons to large macromolecules. Somehow related with the function of transportation is the function of detection and translation of signals, which is essential for living organisms.

The structural function of proteins is also indispensable, involved in the arrangement of different components in living organisms. Collagen, α-keratin, and elastin are examples of proteins involved in the formation of the whole organism. Proteins can also convert chemical energy into mechanical energy. For example, actin and myosin are responsible for muscular motion. In fact, it can be difficult to separate the function as a structural and motor component for these two proteins.

1.1.3 Chemical Modifications of Proteins

The collection of amino acids available in nature is not enough to account for the function of proteins. For example, it is difficult to imagine that the enormous amount of biological reactions requiring protein enzymes as catalysts, is just the result of infinite combinations of only twenty different side chains. An extension of chemistry available for proteins occurs via *post-translational modifications* (Walsh et al., 2005).

After the protein emerges from the ribosome, some amino acids can be covalently modified (Seo and Lee, 2004). These chemical modifications increase the repertoire of functional groups beyond the 20 basic amino acids, which enables new chemistry. In addition, post-translational modifications are key elements in the regulation of protein activity by turning these proteins on and off, and also for the control of the lifetime and location of proteins (Seo and Lee, 2004). According to some estimations, post-translational modifications might increase the molecular variants of the proteins found in cells by two to three order of magnitude (Walsh et al., 2005).

1.1.4 History of Protein Research

The root of the word "protein" derives from the greek "prota", which means "of primary importance". It was named by the Swedish chemist Jöns Jakob Berzelius in 1838, who thought that proteins were the primitive substance made by plants for animal nutrition (Hartley, 1951).

Although the first amino acid was discovered in 1806 (Vauquelin and Robiquet, 1806), it was Franz Hofmeister who in 1902 proposed that polypeptides were amino acids linked by peptide bonds (Hofmeister, 1902). However, the central role of proteins in living organisms was not fully appreciated until 1926, when James B. Sumner, after devising a general crystallization method for enzymes, finally proved that urease was a protein (Sumner, 1926; Simoni et al., 2002).

The 1950s were a decade of great importance in the history of proteins. A key achievement was the sequencing of the first protein by Frederick Sanger in 1951 (Sanger and Tuppy, 1951a; Sanger and Tuppy, 1951b). This proof that proteins have a well-defined chemical composition

was a setback to the general assumption of proteins being somewhat amorphous. The same year, Linus Pauling correctly proposed the alpha helix and beta sheet as the primary structural motifs in protein secondary structure (Pauling and Corey, 1951).

In 1953, Max Perutz took a key step toward the determination of the three-dimentional structure of proteins. He showed that diffracted X-rays from protein crystals could be phased by comparing the patterns from crystals of the protein with and without heavy atoms attached (Perutz, 1953). Before the end of that decade, the structures of myoglobin and haemoglobin were solved by John Kendrew (Kendrew et al., 1958) and Max Perutz (Perutz et al., 1960), respectively. With this significant achievement, a new era of rapid advance began.

The enunciation of the process of protein folding was the next key contribution. In a series of experiments with the ribonuclease, Christian Anfinsen and colleagues concluded that the shape of a protein is determined exclusively by the amino acid sequence, at least for small globular proteins (Anfinsen et al., 1961). This postulate is known as the Anfinsen's dogma (Anfinsen, 1973). In 1969, Cyrus Levinthal reasoned that, given the large number of degrees of freedom that peptide bonds provide to any polypeptide chain, the process of reaching the optimal protein conformation should take an extraordinarily large amount of time. However the protein is always able to find the optimal configuration in microseconds. This is known as the Levinthal's paradox (Levinthal, 1969)

A few decades later, Kurth Wüthrich developed and applied nuclear magnetic resonance (NMR) spectroscopy to determine, with atomic resolution, the structures and dynamics of biological macromolecules and their complexes in solution (Williamson et al., 1985).

1.2 Computational Methods on Protein Sequence and Structure

Computational biology is a multidisciplinary field for the development and application of algorithms and methods to turn biological data into knowledge of biological systems. The combination of computational and experimental information for a better understanding of biological macromolecules has been very successful (Ouzounis and Valencia, 2003). Many different areas have experienced a great advance, such as modelling and dynamics, structural bioinformatics, and protein sequence analysis (Borhani and Shaw, 2012; Ouzounis and Valencia, 2003; Ouzounis, 2012).

The birth of computational biology can be marked in the late 1960s and early 1970s. With the central dogma of molecular biology conceived at that time (Crick, 1970), the types of problems that could be solved using computational models were also formulated. The development of theory and applications on pairwise sequence alignment (Gibbs and McIntyre, 1970; Needleman and Wunsch, 1970), quantification of nucleotide and amino acid substitutions (Clarke, 1970; Epstein, 1967), construction of evolutionary trees (Fitch and Margoliash, 1967), analysis and predictions on primary and secondary structure (Krzywicki and Slonimski, 1967; Pain and Robson, 1970; Ptitsyn, 1969), all of them started in this era. It was also important the consolidation of fundamental theories of evolution, as theory of evolution by gene duplication (Ohno, 1970), neutral evolution (Kimura, 1969; Ohta and Kimura, 1971; Kimura, 1985) and the molecular clock hypothesis (Kimura and Ohta, 1974), which allowed the first construction and phylogenetic analysis of protein families (Wu et al., 1974; Novotn, 1973; Felsenstein,

1978; Felsenstein, 1978; Sattath and Tversky, 1977; Waterman and Smith, 1978; Waterman et al., 1977)

A fundamental development occurred at the end of the 1970s with the creation of public resources to compile, organize, maintain, and distribute biological sequences and structures (Dayhoff, 1979; Bernstein et al., 1977), which have followed exponential growth since their creation. This event triggered the development of key algorithms able to deal with the increasingly large amount of information, such as the Smith-Waterman dynamic programming algorithm to perform optimal local sequence alignments (Smith and Waterman, 1981) and the FASTA algorithms for database search (Lipman and Pearson, 1985; Wilbur and Lipman, 1983).

Finally, numerous computational applications have been very useful, not only in the discovery of sequence motifs (*e.g.*, ATP-binding, zinc-finger, and leucine-zipper motifs), but also studies on specific protein families about their evolution and the relationship between sequence/structure, masterly reviewed somewhere (Ouzounis and Valencia, 2003).

1.3 Motivation and Significance

The changes that affect the evolution of proteins tend to ensure their stability in the physical environment in which they carry out their functions. There are specific solutions that emerge as a consequence of the physicochemical and functional constraints acting on the sequence and structure of proteins. Here we are interested in studying those effects on various levels of the sequence and structure of proteins. For each level, we select different proteins as models. Our ultimate goal is, not only gaining a general understanding, but also to fill specific gaps about the type of protein selected.

We begin by studying the effects of physicochemical and functional constraints on the overall protein sequence and structure. Our focus is on one of the two main folds of membrane proteins, the β-barrel. Biological membranes are the most characteristic example of a hydrophobic environment. β-barrel membrane proteins adopt an interesting topology. They form a concentric barrel of β-sheets. This particular shape exposes half of the side chains to the hydrophobic lipid membrane. The other half is exposed to the more polar environment of the lumen of the pore, which is involved in the function of the poring by enabling the diffusion of molecules (Delcour, 1997; Delcour, 2009).

As consequence of the different constraints directing the evolution of the regions inserted into the outer membrane, only a specific set of amino acid substitutions are allowed. The characterization of the evolutionary pattern of those residues remains as an important task. These membrane proteins are of great importance due to their clinical and technological interest. A proper evolutionary analysis will help in the task of a better detection and increase the tools available for protein engineering design (Liang et al., 2012).

But there are also physicochemical constraints acting in a much reduced dimension. An example are the pockets that emerge in the three-dimensional structure of proteins. Of special interest is the active site of protein enzymes. It is in this cavity where the catalytic reaction occurs. Substrates, ions, and amino acids meet in the active site as components of the biochemical reaction (Fersht, 1998). Despite the unquestionable importance of protein enzymes, there exist important gaps in the understanding on how the catalytic reaction occurs in this space (Benkovic and Hammes-Schiffer, 2003; Hur and Bruice, 2003; Pineda and Schwartz,

2006; Warshel et al., 2006). The physicochemical properties characterizing this critical location are not fully understood. A better physical characterization of the catalytic active site will help to develop more precise theories explaining the elements behind the catalytic power of enzymes. It will also provide, for example, useful information to improve the detection of active sites in three dimensional structures, and ultimately, the function.

The chemistry necessary for enzymes to carry out their functions is limited if we only account for the 20 amino acids (Walsh, 2006). Post-translational modifications enlarge and enrich the chemical repertoire available to proteins (Walsh, 2006). Some of these modifications occur in proteins as consequence of the mediation of an additional enzyme, usually as part of a global regulatory system. Some others have to occur spontaneously. We are interested in the former. Their advantages are clear. For example, it would be convenient for the economy of the system if an enzyme becomes automatically activated when optimal physicochemical conditions are reached.

How is the microenvironment characterizing the residues affected by spontaneous post-translational chemical modifications? We adopt as a model the carboxylation of lysine, a chemical modification that occurs without the mediation of any known additional enzyme and plays important functional roles (Golemi et al., 2001; Dementin et al., 2001; Meulenbroek et al., 2009). There are a variety of experimental complications that makes difficult its detection. The characterization of the microenvironment under which this PTM occurs will facilitate the detection of lysine carboxylation. It will also allow a better evaluation of the real extension of this spontaneous post-traslational modification.

1.4 Thesis Outline / Project Overview

Towards the ultimate goal of a deeper understanding on the effects of physicochemical and functional constraints on the sequence and structure of proteins, the research described here is organized as follows:

In **Chapter 2** we study the effects of physicochemical and functional constraints on the overall sequence variability of proteins inserted into the outer membrane of gram-negative bacteria, mitochondria, and chloroplasts. We characterize the evolutionary pattern of amino acid substitutions at the transmembrane segments of β-barrel membrane proteins. Amino acid substitution rates are estimated at different interfaces of the transmembrane segments using a Bayesian Monte Carlo approach. A tool for the accurate detection of remote homologs of β-barrel membrane proteins is provided by deriving scoring matrices from the estimated substitution rates. We also use the estimated amino acid substitutions as a tool for the design of mutants of OmpF with an improve stability and different oligomerization state. Our goal is to identify key residues with a direct contribution on the oligomerization state of this trimeric porin.

This chapter is partially based on the publications:

- **Jimenez-Morales, D.**, Adamian, L., Liang, J.: Detecting remote homologs using scoring matrices calculated from the estimation of amino acid substitution rates of β-barrel membrane proteins. In *Conf Proc IEEE Eng Med Biol Soc*, 30:1347-1350, 2008.

- Liang, J., Naveed, H., **Jimenez-Morales**, D., Adamian, L., Lin, M.: Computational studies of membrane proteins: Models and predictions for biological understanding. In *Biochim Biophys Acta*, Apr;1818(4):927-41, 2011.
- **Jimenez-Morales, D.**, Liang, J.: Pattern of amino acid substitutions in transmembrane domains of β-barrel membrane proteins for detecting remote homologs in bacteria and mitochondria. In *PLoS ONE*, 6(11):e26400. Epub 2011 Nov 1, 2011.
- Naveed, H., **Jimenez-Morales**, D., Tian, J., Pasupuleti, V.; Kenney, L., Liang, L.: Engineered oligomerization state of OmpF protein through computational design decouples oligomer dissociation from unfolding. In *J Mol Biol*, May 25;419(1-2):89-101, 2012.

Chapter 3 is devoted to study the functional cavities of protein enzymes. We measure the number density of ionizable residues at the catalytic active site. The significance of this physical property as marker of the catalytic active site is evaluated. We also explore the charge density in other pockets (we name them "craters"), and suggest possible roles.

This chapter is partially based on the publication:

- **Jimenez-Morales, D.**, Liang, J., Eisenberg, B.: Ionizable side chains at catalytic active sites of enzymes. In *Eur Biophys J*, May;41(5):449-60, 2012.

Chapter 4 focuses in the study of the physicochemical properties of residues subjected to spontaneous post-translational modifications. We characterize the microenvironment of lysine carboxylation. This spontaneous post-translational modification plays important roles in the catalytic function of some remarkable enzymes. We also develop a Bayesian classifier for the prediction of lysine carboxylation. The predictor is used to search protein databases. We assess

the extent of lysine carboxylation and conclude that it may have a much broader prevalence than is currently known.

This chapter is partially based on the publication:

- **Jimenez-Morales, D.**, Adamian, L., Shi, D., Liang, J.: Unveiling the prevalence of a spontaneous post-translational modification. *In preparation.*

Finally in **Chapter 5**, we outline the main features of this work, highlight the novel aspects, strengths and weaknesses, and provide perspectives on further developments.

CHAPTER 2

ON THE EFFECTS OF PHYSICOCHEMICAL CONSTRAINTS OF BIOLOGICAL MEMBRANES ON THE AMINO ACIDIC SUBSTITUTION PATTERN AT TRANSMEMBRANE SEGMENTS OF β-BARREL MEMBRANE PROTEINS

2.1 Introduction

As one of the two classes of integral membrane proteins, β-barrel membrane proteins are found in the outer membranes of gram negative bacteria, mitochondria, and chloroplasts. Because they are located in the first barrier of bacteria and are in contact with the extracellular environment, they are often key factors providing control of the diffusion, exchange, and transport of ions and organic molecules (Wimley, 2003; Schulz, 2000; Molloy et al., 2000; Benz, 1994; Fischer et al., 1994). They are also involved in the transmission of signals in response to stimuli and, as enzymes, in the maintaining of the stability of the outer membrane (Schulz, 2000; Bishop, 2008). In eukaryotes, mitochondrial outer membrane proteins are part of the mitochondrial permeability transition pore (mtPTP), a major regulator of apoptosis, with important implications in cancer, degenerative diseases, and aging (Wallace, 2005). For example, the voltage-dependent anion channel (VDAC) is considered a promising target for anticancer treatments (Simamura et al., 2008).

β-barrel membrane proteins are also important determinants of bacterial virulence and are promising drug targets (Larbig et al., 2001; Adiga et al., 2009a; Srinivas et al., 2010). As bacterial porins enable the diffusion of hydrophilic antibiotics through outer membranes, mutation of their barrel interior is the basis of a common mechanism of bacterial drug resistance (Delcour, 1997; Delcour, 2009). β-barrel membrane proteins therefore are excellent targets for developing new antibacterial drugs. A promising example is the recent discovery of a new peptidomimetic antibiotic that perturbs the critical LPS transport function of the β-barrel membrane protein LptD (Srinivas et al., 2010).

The architecture and amino acid make-up of β-barrel membrane proteins have been well studied (Schulz, 2000; Ulmschneider and Sansom, 2001; Wimley, 2002; Ujwal et al., 2008). Several methods have been developed for the detection of β-barrel membrane proteins from sequences (Bigelow and Rost, 2009; Randall et al., 2008; Remmert et al., 2009; Bagos et al., 2005). Sequence motifs and antimotifs in transmembrane (TM) regions of β-strands have also been identified, with tyrosine found to play important roles (Jackups et al., 2006). In addition, propensities of residues for different spatial regions and for inter-strand pairwise contact have been quantified (Bishop et al., 2001a; Jackups and Liang, 2005; Jackups et al., 2006). A physical model of energetics based on the estimated propensities of spatial interactions enabled the identification of weakly stable regions in the TM domain, the discovery of general mechanisms of their stabilization, the prediction of oligomerization states, and the delineation protein-protein interaction interfaces (Naveed et al., 2009).

A remaining challenging task is the detection and quantification of evolutionary patterns of residues embedded in the TM region. The amino acid sequences of β-barrel membrane proteins determine how these proteins fold, insert into the membrane, and carry out their biological functions. As evolution proceeds, the set of allowed amino acid substitutions at different positions of the transmembrane segments are constrained by these requirements, which manifest as patterns of substitutions that correlates with the amino acid type, solvent accessibility, secondary structure, depth of lipid buriedness, and side-chain hydrogen bonding states (Overington et al., 1992; Overington et al., 1990). Currently, it is not clear how residues substitute in the outer membrane region of gram negative bacteria. In addition, whether membrane proteins in mitochondria and bacterial outer membrane show the same evolutionary pattern is unknown. Understanding of the evolutionary patterns of β-barrel membrane proteins can help us to identify key features important for their structural and functional integrity. Furthermore, it can aid in the design of mutagenesis studies (Delcour, 2009).

Characterizing amino acid substitutions can also be used to develop scoring matrices specific for β-barrel membrane proteins for sequence alignment, structure prediction, and large scale database searches of remote homologs. Conventional scoring matrices used for database searches are not designed for β-barrel membrane proteins. For example, the PAM (Dayhoff et al., 1978) and BLOSUM matrices (Henikoff and Henikoff, 1992) were derived from large collections of multiple sequence alignments of globular proteins, and are inappropriate for studying membrane proteins (Yu et al., 2003). A number of scoring matrices have been developed for membrane proteins. The PHAT matrices are based on blocks of multi-aligned sequences of transmembrane

segments and hydrophobic segments (Ng et al., 2000). The SLIM matrices are based on models of different background compositions of amino acid residues (Muller et al., 2001). However, they are all derived for studying α-helical membrane proteins.

To capture the pattern of amino acid substitutions of β-barrel membrane proteins, we have estimated substitution rates of amino acids in the transmembrane segments. Our approach was based on a Bayesian Monte Carlo method (Tseng and Liang, 2006). We selected a representative set of eleven proteins with known structures and with pairwise sequence identities below 20%. For each protein, substitution rates were estimated for residues in the transmembrane segments. These estimated rates show characteristic patterns that are unique to β-barrel membrane proteins. From these estimated rates, we derived scoring matrices useful for sequence alignment and for detecting remote homologs of β-barrel membrane proteins. Results of database searches showed that these scoring matrices can significantly improve reliability in detection of β-barrel membrane proteins by eliminating errors of selecting soluble proteins as well as other membrane proteins of similar composition.

A proper evolutionary analysis not only can be helpful in the task of a better detection, but also increase the tools available for protein engineering design on specific β-barrel membrane proteins. Among interesting candidates, OmpF is a porin from *E. coli* whose native oligomerization state is thought to be trimeric (Cowan et al., 1992). However, dimeric structures have also been observed in both *in vitro* and *in vivo* experiments (Visudtiphole et al., 2005; Reid et al., 1988; Surrey et al., 1996; Watanabe, 2002). Similarly, despite the observed trimeric form of porin PhoE, a functional monomeric form of PhoE has been reported in *in vitro* and *in vivo*

studies (de Cock et al., 1990; Van Gelder and Tommassen, 1996). Although the protein-protein interaction (PPI) site for porins is known in many cases, the significance of preferring a particular oligomeric state over the others is not clear (Meng et al., 2009). Understanding the factors that determine the oligomerization state of these proteins will advance our understanding of their biogenesis and function.

A strategy to the study of the stability constraints underlying β-barrel membrane proteins is to characterize their evolutionary patterns. The degree of sequence conservation often correlates with the importance of a particular position of residues in a protein. Slow protein evolution is a consequence of strong purifying selection, which varies among different proteins, or even among different regions within the same protein, due to stability or functional constraints. An interesting possibility is whether the stability and oligomeric properties of the β-barrel membrane proteins can be altered by selecting different amino acid substitutions based on estimated evolutionary patterns of substitutions.

This chapter is organized as follows. We first describe the pattern of amino acid substitutions found for TM segments of bacterial β-barrel membrane proteins. We then discuss how scoring matrices derived from the estimated substitution rates can be used for reliable detection of homologs. Then we show how mitochondria outer membrane proteins can also be detected using these scoring matrices derived from bacterial proteins. We then consider the implications in predicting structures of bacterial and mitochondrial membrane proteins using known structures as templates. We also explore strategies to re-engineer the protein-protein interaction interface in the TM domain of the β-barrel membrane protein OmpF. Several recent studies

have focused on designing new protein-protein interaction interfaces through computational protein re-engineering, and computational *de novo* interface design (Kortemme and Baker, 2004; Ben-Shimon and Eisenstein, 2010; Sharabi et al., 2011; Fleishman et al., 2011). We have re-designed the PPI interface of OmpF by identifying a conserved region in the PPI interface of OmpF and mutating the conserved residues following a substitution pattern different from that observed in β-barrel membrane proteins. Our goal was to obtain stable monomeric forms of OmpF. Mutants engineered through site-directed mutagenesis following our design indeed showed stable monomeric oligomerization states.

2.2 Materials and Methods

2.2.1 Template β-Barrel Membrane Proteins and Homologs

We carried out BLAST searches (Altschul et al., 1990a) using each of the protein sequences of the 20 β-barrel membrane proteins with a solved structure sharing less than 20% pairwise sequence identity as a query against the non-redundant NCBI protein database (Pruitt et al., 2007). For each protein, a multiple sequence alignment was generated using CLUSTALW2 (Larkin et al., 2007). Regions corresponding to the transmembrane segments were extracted to form the *Transmembrane β-Strand Database* (TBSD). Next, using the same query PDB sequences but with only those residues from the transmembrane segments concatenated, we carried out SSEARCH searches (Pearson and Lipman, 1988) against the TBSD database. From the output, two sequences for every interval of 10% sequence identity between 90 and 30% were selected, allowing no more than two gaps in every transmembrane segment. This criterion allows us to avoid the problem of over-representations of proteins in a narrow range of evolutionary distance, and enabled selecting sequences exclusively based on the similarity of the transmembrane fragments. This leads to the exclusion of 9 proteins from the set of 20 β-barrel membrane proteins. The final 11 proteins selected are listed in Table I.

2.2.2 Estimating Amino Acid Residue Substitution Rates

2.2.2.1 The Bayesian Monte Carlo Method

The substitution rates of residues in the transmembrane segments were estimated following the approach of Tseng and Liang (Tseng and Liang, 2006). Briefly, a Bayesian Monte Carlo estimator based on the technique of Markov chain Monte Carlo was used. Estimation is based

TABLE I

TEMPLATE PROTEINS, THEIR COMPOSITION, AND HYDROPHOBICITY INDEX VALUES

	# of Residues and TM Strands			Hydrophobicity Index (GES)		
PDB	TM_{all}/Total/# Strands	TM_{in}	TM_{out}	TM_{all}	TM_{in}	TM_{out}
1A0S	172/413/18	84	87	-0.54	-1.66	0.52
1BXW	84/172/8	42	42	-0.05	-1.76	1.66
1E54	139/332/16	70	69	-0.33	-1.8	1.17
1FEP	206/724/22	102	104	-0.67	-2.25	0.87
1I78	102/297/10	50	51	-0.11	-1.99	1.71
1KMO	217/774/22	108	109	-0.94	-2.6	0.7
1NQE	220/549/22	111	109	-0.87	-2.47	0.77
1QD6	124/240/12	59	64	-0.63	-2.64	1.16
1QJ8	75/148/8	35	40	0.2	-1.02	1.27
2MPR	178/427/16	90	87	-0.75	-2.5	1.04
2OMF	153/340/16	76	77	-0.66	-2.38	1.04
Mean	152/401/16	75	76	-0.49	-2.10	1.08

TM_{all}: number of residues in the TM region; Total: total number of residue in the protein; # Strands: number of β-strands in the TM region; TM_{in}: number of residues in the TM in-facing region; and TM_{out}: number of residues in the TM lipid out-facing region. The hydrophobicity is measured by the GES index (Engelman et al., 1986), with negative values representing polarity and positive values hydrophobicity.

on the selected set of sequences homologous to the template protein and their phylogenetic trees. The entries q_{ij} of the substitution rate matrix Q are substitution rates of amino acid residues for the 20 amino acids at an infinitesimally small time interval. Specifically, we have:

$$Q = \{q_{ij}\} = \begin{pmatrix} - & q_{1,2} & \cdots & q_{1,20} \\ q_{1,2} & - & \cdots & q_{2,20} \\ & & \ddots & \\ q_{1,20} & q_{2,20} & \cdots & - \end{pmatrix},$$

The transition probability matrix of size 20×20 after time t is (Li and Goldman, 1998)

$$P(t) = \{p_{ij}(t)\} = P(0)\exp(Q \cdot t),$$

where $P(0) = I$. Here $p_{ij}(t)$ represents the probability that a residue of type i will mutate into a residue of type j after time t.

Using a Bayesian approach, we describe the instantaneous substitution rate $Q = \{q_{ij}\}$ by a posterior distribution $\pi(Q|\mathcal{S}, T)$, which summarizes the prior information $\pi(Q)$ available on the rates $Q = \{q_{ij}\}$ and the likelihood information $P(\mathcal{S}|T, Q)$ contained in the multiple-alignment $\mathcal{S}$ and the phylogenetic tree T. The posterior distribution $\pi(Q|\mathcal{S}, T)$ can be estimated using Markov chain Monte Carlo as:

$$\pi(Q|\mathcal{S}, T) \propto \int P(\mathcal{S}|T, Q) \cdot \pi(Q) dQ.$$

Further details can be found in (Tseng and Liang, 2006).

In this study, Q takes the form $Q = D/2 \cdot S + S \cdot D/2$, where D is a diagonal matrix with values taken from the amino acid composition of the set of aligned sequences studied, and S is a symmetric matrix with 0 values in diagonal elements, and off-diagonal entries estimated following the model of Adachi et al (Adachi and Hasegawa, 1996). Phylogenetic trees T were obtained using the maximum likelihood method MOLPHY based on the entire length of the protein sequences (Adachi and Hasegawa, 1996) (see Figure 2 for more details).

2.2.2.2 Valid Pairs Correction

Once the initial S matrix was estimated, we made further corrections to account for different occurring frequency of substitutions appearing in the multiple-aligned sequences. We calculate $s'_{ij} = \frac{w}{m} \cdot s_{ij}$, where $w = \sum_1^m \frac{a_i(k) \cdot a_j(k)}{\binom{n}{2}}$. Here m is the total number of columns, $a_i(k)$ and $a_j(k)$ are the number counts of residue i and j in the k-th column of the alignment, respectively, and n is the number of sequences. We calculated the average S' and D matrices for the 11 proteins used, from which the final rate matrix Q is derived (see Figure 5). This is repeated separately for the aligned sequences of TM_{all}, TM_{in}, and TM_{out}.

The Q matrix for each of the region is depicted as a bubble plot, in which the area of the circle for the (i, j)-entry is drawn proportional to the value of q_{ij} (Figure 4). We also created a diagram to represent the pattern of amino acid substitutions. We named it **djm-diagrams** (Figure 3). In this diagram, amino acid residues are grouped according to common physicochemical properties. Any two residues are connected by a line when the substitution values are significant. The thicker the line, the larger the value of substitution.

The scoring matrices at different evolutionary time interval are then derived from the estimated Q matrix. Further details can be found in references (Karlin and Altschul, 1990; Tseng and Liang, 2006). In this study, we use the scoring matrix of evolutionary time of 40 for $bbTM_{all}$ and $bbTM_{out}$, and 36 for $bbTM_{in}$, as they give the best discrimination (see Figure 7).

2.2.2.3 Tool Availability

We have made available a set of tool to perform BLAST searches for β-barrel membrane proteins against the non-redundant NCBI database using the $bbTM$ matrices. The URL is at: http://tanto.bioengr.uic.edu/bbtmst/bbtmstool.php

2.2.3 Experimental Design and Procedures

2.2.3.1 Identification of Evolutionary Conserved Residues in the Protein Protein Interaction Interface of OmpF

Quantification of variability at each position of the transmembrane domains of the OmpF protein was made by collecting homologous sequences from a non-redundant database. A multiple sequence alignment was generated and the entropy of sequence variability is then calculated for every position.

Briefly, using the sequence of OmpF (pdb: 2OMF) as query, a BLAST search was carried out against the non-redundant NCBI protein database (default parameters, except word size = 2, for increased sensitivity) (Altschul et al., 1990a). We selected pairwise sequence alignments of more than 30% sequence identity that cover at least 3/4 of the OmpF sequence. There are 869 alignments that matched these criteria. A simple multiple sequence alignment was then built

based on individual pairwise sequence alignment, *i.e.* by stacking homologous sequences based on their alignment to OmpF.

The information entropy H_k of sequence variations in the unit of bit was then calculated for every column k of the multiple sequence alignment using the formula $H_k = -\sum_{i=1}^{20} f_i \log_2 f_i$. Here f_i is the frequency of amino acid i in the column k. Low entropy indicates that the specific column is very well-conserved. The maximum entropy is $\log_2 20 = 4.32$ bits. For illustration, we show $R_k = \log_2 20 - H_k$, where well-conserved columns show longest bars (Figure 9).

2.2.3.2 Cloning and Strains

OmpF inclusion bodies were expressed from *E. coli* BL21 (DE3) using pET28A vector expression system (Novagen). The ompF signal sequence (residues 1–22) was replaced by a single methionine residue using PCR. The ompF cDNA was amplified by the primer ompF forward and ompF reverse and ligated into *Bam*HI and *Xho*I digested pET28A vector. PrimerX (www.bioinformatics.org/primerx) was used to design the primers for site-directed mutagenesis. Amplification of the product was done using DNA polymerase (PfuTurbo Hotstart - Stratagene) in *E. coli* DH5 cells before sequencing and transformation into *E. coli* BL21 cells for protein expression.

2.2.3.3 Protein Expression and Purification

Transformed cells were grown at $37°C$ in LB medium (LB Broth, Fisher Scientific) containing 50 μg/ml kanamycin. At a D_{600} of 0.6, the culture was induced with 0.5 mM IPTG (isopropyl β-D-thiogalactoside) and grown for 3 hours. The cells were then harvested by centrifugation at 3000 g for 5 minutes. The cell pellet was re-suspended in DPBS 1× (Dulbecco's Phosphate

Buffered Saline). The solution was sonicated briefly to shear DNA and the viscosity of the solution was reduced on ice. After sonication, the inclusion bodies were pelleted and washed two times with DPBS 1×, and re-suspended in 1%(v/v) Triton X-100, 20mM Tris/HCl (pH 8.0), 0.1 mM EDTA (Ethylenediaminetetraacetic acid), and 1 mM DTT (Dithiothreitol) (Visudtiphole et al., 2005). The suspension was then incubated at $37°C$ for 1 hour and washed with DPBS 1×.

2.2.3.4 Protein Refolding and Purification

The inclusion bodies were solubilized in denaturing buffer constituting of 50 mM HCl (pH 8.0) and 8M urea at $55°C$ for 30 minutes. Refolding was performed by a 20× dilution with thorough mixing of the pooled sample, whose concentration had been adjusted to 2 mg/ml. The refolding buffer contained 0.5% (w/v) DG (n-dodecyl-β-D-glucopyranoside), 0.2% (w/v) DM (n-dodecyl-β-D-maltoside) in 50 mM Tris-HCl, 1 mM DTT and 0.1 mM EDTA (pH 8) (Visudtiphole et al., 2005). The mixture was incubated overnight at $37°C$ using Isotemp Hybridization Incubator (Fisher Scientific). After overnight refolding, subsequent degradation of the unfolded protein was induced by the addition of trypsin (trypsin/protein 1:100 w/w) (Baldwin et al., 2011). Final protein was purified and concentrated using Amicon Ultra-0.5 mL Centrifugal Filters 30K (Millipore). All refolded samples used in the analysis were in 0.5% (w/v) OPOE (n-octyl-oligo-oxyethylene), 10 Mm HEPS (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) buffer (pH 7.4).

2.2.3.5 SDS-PAGE

Folding reactions of OmpF were quenched by adding 5× SDS (Sodium dodecyl sulfate) gel-loading buffer (Sambrook et al., 1989) to a final dilution of 1× SDS gel loading buffer. 50 μl of sample was loaded on a 4-20% acrylamide continous gradient precast gel (Bio-Rad) to resolve the folded and unfolded populations (Burgess et al., 2008).

2.2.3.6 Circular Dichroism

Circular dichroism (UV-CD) spectroscopy measurements were recorded on a Jasco J-810 spectrometer (Jasco, Easton, MD) using cuvettes with path lengths of 1 mm for protein concentration of 0.2 mg/mL. The samples contained proteins before trypsin treatment as the yield of all the proteins was similar.Every sample was scanned in the wavelength range 200-250 nm. An average of three scans at 50 nm/min was acquired with a bandwidth of 0.2 nm and a response time of 1 second using three independent protein preparations. Final CD spectrum was then corrected for background by subtraction of spectrum of protein-free samples recorded under the same conditions.

2.2.3.7 Tryptophan Fluorescence Measurements

Wild-type OmpF and all mutants have two tryptophan residues, all located in the transmembrane domain. Trp fluorescence measurements were recorded from 300 to 400 nm on a Fluoromax-3 spectrofluorimeter (Jobin-Yvon Inc., Edison, NJ) using samples containing 0.2 mg/mL protein held in a 1 mm path length cuvette, with an excitation wavelength of 290 nm (Surrey et al., 1996).

2.3 Results

We use a set of 11 β-barrel membrane proteins with known structures sharing less than 20% pairwise sequence identity (Table I). We followed the procedure of Jackups *et al* (Jackups and Liang, 2005) and select the fragments embedded within the outer membrane region. Altogether we have 170 TM-strands. From these, we further derive two additional data sets, one for residues facing the interior of the barrel, and another for residues facing the lipid environment of the outer membrane. These three data sets are termed TM_{all}, TM_{in}, and TM_{out}, respectively. These 11 proteins, their homologous proteins and phylogenetic trees can be found in Figure 2.

The transmembrane segments as a whole have moderate polarity (-0.49 by the GES scale (Engelman et al., 1986)). However, the in-facing residues in TM_{in} are strongly polar (polarity of -2.10) and the out-facing residues in TM_{out} are strongly hydrophobic (+1.08) (Table I).

2.3.1 Pattern of Amino Acid Substitutions in Transmembrane Segments

2.3.1.1 Overall Pattern

The general pattern of amino acid residue substitutions observed for residues in the TM region is shown in Figure 4A (see also Figure 5). Residues with similar physiochemical properties often exchange with each other. V has overall the highest degree of substitutions, and exchanges mostly with L, I, and A. The instantaneous rate of V-I substitution is 194 in the unit of 10^{-4} expected residue changes per 100 site between sequences. The value for V-L is 131. L and I have the next highest overall degree of substitutions. In addition to V, they frequently exchange between themselves (I-L: 44), and substitute with other hydrophobic residues (L-M:16, L-A:10, I-A:3), and the aromatic residue F (L-F:32, I-F:5).

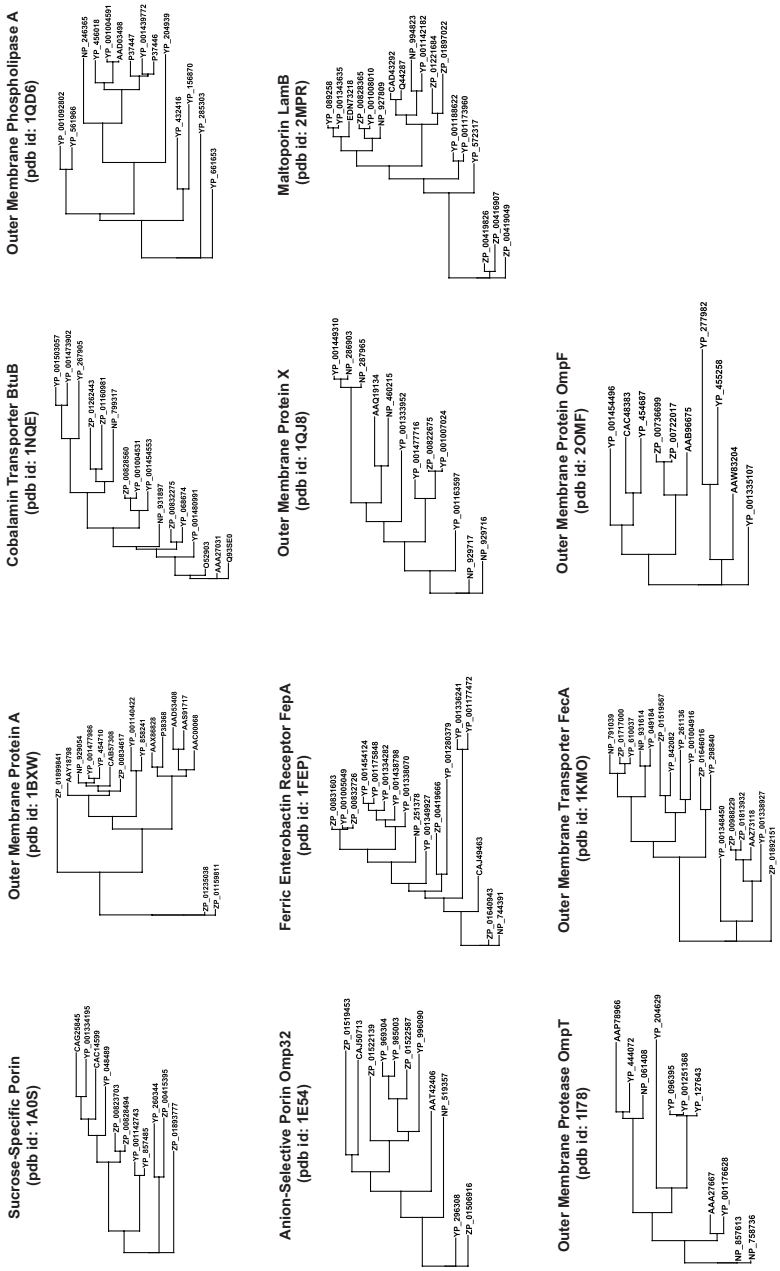

Figure 2. **Proteins used as structural templates to infer substitution rates.** The 11 proteins and their phylogenetic trees (with labelled homologs) that are used to estimate the substitution rates. We obtain one phylogenetic tree for each of the 11 β-barrel membrane proteins, using the multiple sequence alignment for the entire length of the proteins. The same tree was used for the estimation of three independent substitution rate matrices (Q_{all}, Q_{out}, and Q_{in})

Small polar residues S and T substitute mostly between themselves (S-T: 38), with the small residues A (25 for T-A, 18 for S-A) and G (S-G:9, T-G:3). Exchanges also occur with N (T-N:6, S-N:15).

Among large polar residues, Q shows overall lower substitutions, but with a broader number of residue types, *e.g.*, with charged residues E (E-Q:6), H (H-Q:53), R (R-Q:3) and K (K-Q:3), and with polar residues S (S-Q: 3) and N (N-Q: 3). Residue N readily substitutes with polar residues S (S-N:15), T (T-N:6) and Q (Q-N:3), and with the charged residue D (D-N:7).

Aromatic residues most likely substitute among themselves (*e.g.*, Y-F:25, Y-W:9, W-F:3). Residue F has the broadest range of substitutions among aromatics and exchanges with L (F-L:32), V (F-V:8), I (F-I:5), W (F-W:3) and A (F-A:3).

The most abundant residue in the transmembrane segments of β-barrel membrane proteins is residue G. This residue overall experiences little substitutions. The relatively few substitutions are with A (A-G:40), S (S-G:9) and T(T-G:3).

2.3.1.2 Substitution Rate of Residues Facing the Outer Membrane Lipids

The pattern of substitutions for residues facing the outer membrane lipids (TM_{out}) is shown in Figure 3 (see also Figure 4B and Figure 5). The most common substitutions observed are between hydrophobic residues, namely, substitutions among V, I, L, A and F. For example, V has the highest degree in overall substitutions, showing large values with V-I (275), V-L (168), V-A (61) and V-F (22).

L is the most abundant residue in the TM_{out} region. It predominantly substitutes with non-polar residues, *i.e.*, L-V (168), L-I (82), L-A (16) and L-M (16). Other observed exchanges are

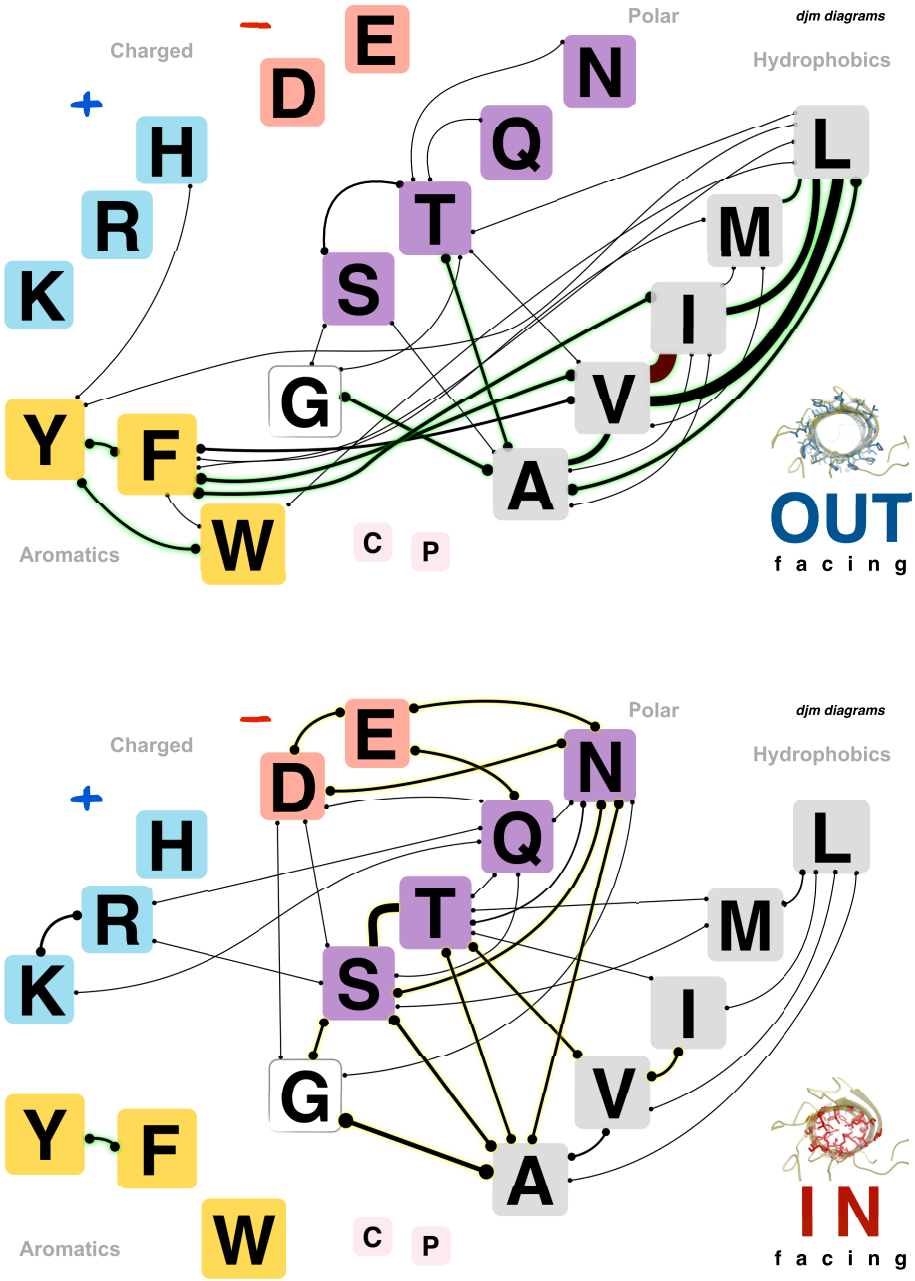

Figure 3. **DJM-diagrams of amino acid substitutions**. Lines represent amino acid substitutions between any two amino acids. The thicker the line, the more frequent substitutions between the two connected amino acids. *Top*, pattern of amino acid substitutions of residues out-facing the lipid membrane. *Down*, pattern of amino acid substitutions of residues in-facing the interior of the barrel. Twice as many amino acid substitutions were found to occur at the lipid facing interface. However, the substitution pattern at the lipid interface is very narrow, indicating a strong selection pressure at this interface for amino acids to maintain the same physicochemical properties. Substitutions in the in-facing region are less common, but with a broader spectrum of substitutions between residues with different physicochemical properties.

with aromatic residues (*e.g.*, L-F:65, L-W:4, L-Y:4) and T (L-T:3). Residue I also exchanges mostly with other non-polar amino acids (I-V:275, I-L:82, I-F:10, I-A:8 and I-M:4).

Residue A has a broad range of substitutions at the TM_{out} interface. It mostly exchanges with other hydrophobic or small amino acids, including V (V-A:61), G (G-A:46), L (L-A:16), F(F-A:7), and I(I-A:5). Notable exceptions are with polar residues T (T-A:22) and S (S-A:6).

Among aromatic residues, Y is well-conserved at the lipid-facing surface of β-barrel membrane proteins. This is reflected by the relatively small number of its substitutions. It is the second most frequent amino acid residue at the lipid interface, and contributes significantly to the formation of the aromatic girdle (Schulz, 1993), a prominent feature of β-barrel membrane proteins. Substitutions of Y with other aromatic residues are most common (Y-F:20, Y-W:18), and to a small degree also with L (Y-L:4) and H (Y-H:3). Aromatic residue F is less abundant compared to Y, but experiences more varieties of substitutions, mostly with hydrophobic residues: F-L (65), F-V (22), F-Y (20), F-I (10), F-A (7), F-W (5) and F-M (4). The aromatic residue W resides mostly in the TM_{out} region. Its pattern of substitution is very restricted and mostly substitutes with Y (W-Y:14).

The predominant polar residue at the TM_{out} interface is T. It substitutes with non-polar amino acids (T-A:17, T-V:6, T-G:4) and the polar residue S (7). In contrast, polar residue S substitutes only with T (7) and A (5). Among ionizable residues, E has low abundance in the TM_{out} interface and low tendency for substitutions.

Finally, the only substitution observed for G in this interface are mainly with A (46), and to a less extent with T (4) and S (3).

2.3.1.3 Substitution Rates of Residues Facing the Interior of the Barrel

The pattern of substitutions for residues facing the interior of the barrel (TM_{in}) differs significantly from that of the TM_{out} region Figure 3 (see also Figure 4C and Figure 5). Small residues S, T, and A experienced most frequent substitutions (S-T:77, S-A:38, and S-N:26). Q and N have a much higher presence at the TM_{in} region, with increased substitutions.

Ionizable residues such as E, R, K and D are more abundant in the TM_{in} interface. Most of them do not substitute with other residues. For example, E is among the most abundant residues in the TM_{in} region. It is well-conserved, substituting mostly only with Q (E-Q:20) and the other negatively charged residue D (E-D:14). Similar patterns are found for the residues R and K, which exchange mostly between themselves (R-K:8) and with polar residue Q (K-Q:5, R-Q:4). The lack of substitutions of ionizable residues suggests that they play a significant role in the function of the β-barrel membrane proteins and are under strong purifying selection pressure.

The pattern of substitutions for hydrophobic residues is somewhat different at this interface. Although V, I, L and M mostly exchange amongst themselves, they also exchanges more frequently with polar residues such as T, in contrast to what is found in the TM_{out} region.

The most abundant residue at this interface is G. Its substitution pattern shows some similarities with G at the lipid interface, although a larger number of substitutions is observed with S (S-G:19).

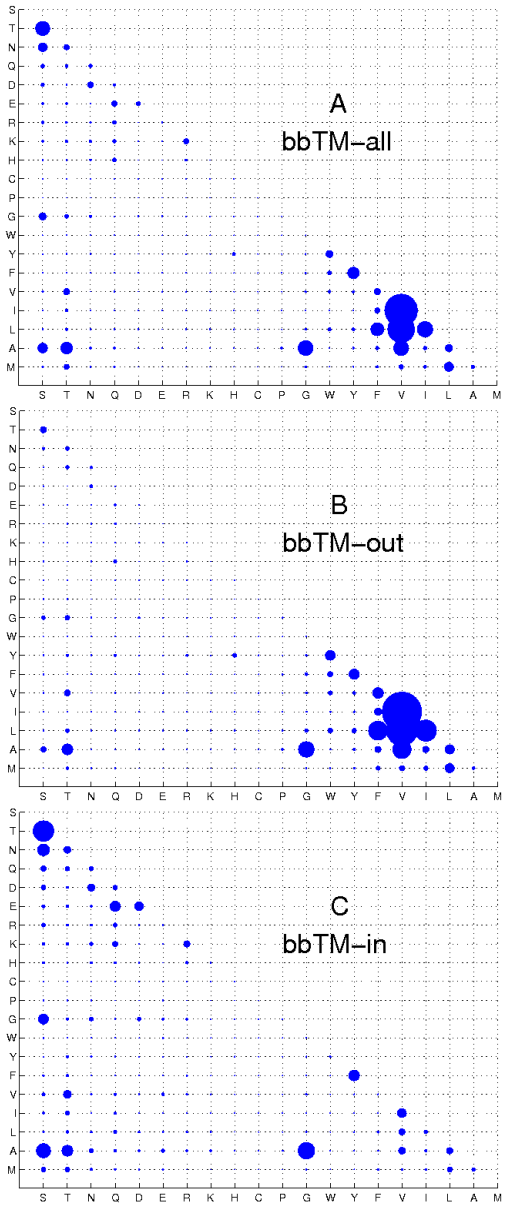

Figure 4. **Bubble plot representation of estimated amino acid substitution rates.** Estimated instantaneous rates of substitution for residues in the TM segments and at different TM interfaces from 11 template β-barrel membrane proteins. The size of the bubble is proportional to the value of the estimated substitution rate. The instantaneous substitution rates (A) for all TM residues (Q_{all}); (B) for residues out-facing the membrane (Q_{out}); and (C) for residues in-facing the membrane (Q_{in}).

A. Q_{all}

```
# Q_all: Substitution rate values for the fragments embedded within the outer membrane region.
# Scaling factor: q_ij x 10^-04
      S      T      N      Q      D      E      R      K      H      C      P      G      W      Y      F      V      I      L      A      M
S -91.699 37.585 14.958  2.972  2.530  0.728  1.435  0.998  0.521  0.133  0.197  9.318  0.049  0.240  0.379  0.618  0.162  0.510 17.641  0.723
T 37.585 -94.888  5.927  2.069  0.438  0.741  0.973  1.355  0.246  0.102  0.161  2.930  0.148  0.566  0.385  7.994  1.537  1.461 24.893  5.379
N 14.958  5.927 -37.543  2.511  6.672  0.487  0.695  1.253  0.817  0.037  0.050  1.394  0.102  0.523  0.146  0.162  0.079  0.212  0.958  0.559
Q  2.972  2.069  2.511 -27.419  1.466  5.979  2.735  2.551  3.045  0.133  0.113  0.307  0.113  0.345  0.104  0.576  0.429  0.444  0.741  0.786
D  2.530  0.438  6.672  1.466 -17.105  3.365  0.174  0.350  0.130  0.030  0.042  0.556  0.060  0.123  0.090  0.151  0.071  0.400  0.360  0.096
E  0.728  0.741  0.487  5.979  3.365 -13.688  0.234  0.234  0.052  0.019  0.127  0.443  0.069  0.132  0.049  0.218  0.040  0.142  0.515  0.114
R  1.435  0.973  0.695  2.735  0.174  0.234 -14.180  4.802  1.003  0.061  0.053  0.300  0.071  0.335  0.251  0.223  0.123  0.361  0.262  0.089
K  0.998  1.355  1.253  2.551  0.350  0.234  4.802 -13.533  0.086  0.029  0.042  0.208  0.120  0.123  0.134  0.100  0.142  0.266  0.598  0.143
H  0.521  0.246  0.817  3.045  0.130  0.052  1.003  0.086 -9.029  0.007  0.013  0.124  0.180  2.050  0.112  0.085  0.056  0.295  0.157  0.049
C  0.133  0.102  0.037  0.133  0.030  0.019  0.061  0.029  0.007 -1.104  0.013  0.055  0.016  0.188  0.061  0.071  0.021  0.052  0.046  0.029
P  0.197  0.161  0.050  0.113  0.042  0.127  0.053  0.042  0.013  0.013 -2.418  0.122  0.022  0.053  0.219  0.042  0.027  0.175  0.931  0.016
G  9.318  2.930  1.394  0.307  0.556  0.443  0.300  0.208  0.124  0.055  0.122 -60.010  0.141  0.305  0.595  0.633  0.303  0.855 40.777  0.643
W  0.049  0.148  0.102  0.113  0.060  0.069  0.071  0.120  0.180  0.016  0.022  0.141 -17.997  9.304  3.245  1.482  0.219  2.228  0.304  0.125
Y  0.240  0.566  0.523  0.345  0.123  0.132  0.335  0.123  2.050  0.188  0.053  0.305  9.304 -43.202 24.672  1.237  0.339  1.682  0.646  0.338
F  0.379  0.385  0.146  0.104  0.090  0.049  0.251  0.134  0.112  0.061  0.219  0.595  3.245 24.672 -78.969  7.825  4.858 31.890  2.568  1.387
V  0.618  7.994  0.162  0.576  0.151  0.218  0.223  0.100  0.085  0.071  0.042  0.633  1.482  1.237  7.825 -391.517 194.115 131.115 41.268  3.601
I  0.162  1.537  0.079  0.429  0.071  0.040  0.123  0.142  0.056  0.021  0.027  0.303  0.219  0.339  4.858 194.115 -251.467 44.434  2.718  1.793
L  0.510  1.461  0.212  0.444  0.400  0.142  0.361  0.266  0.295  0.052  0.175  0.855  2.228  1.682 31.890 131.115  44.434 -242.428  9.930 15.977
A 17.641 24.893  0.958  0.741  0.360  0.515  0.262  0.598  0.157  0.046  0.931 40.777  0.304  0.646  2.568 41.268  2.718  9.930 -147.777  2.465
M  0.723  5.379  0.559  0.786  0.096  0.114  0.089  0.143  0.049  0.029  0.016  0.643  0.125  0.338  1.387  3.601  1.793 15.977  2.465 -34.312
```

B. Q_{out}

```
# Q_out: Substitution rate values for the subset of residues from the transmembrane segments facing the lipid environment
# Scaling factor: q_ij x 10^-04
      S      T      N      Q      D      E      R      K      H      C      P      G      W      Y      F      V      I      L      A      M
S -21.741  7.444  1.569  0.392  0.295  0.075  0.095  0.141  0.236  0.026  0.314  2.878  0.107  0.485  0.459  0.565  0.226  0.585  5.694  0.155
T  7.444 -56.860  3.098  2.569  0.184  0.462  0.531  0.388  0.134  0.101  0.381  4.389  0.192  1.209  0.348  7.530  0.574  3.180 22.296  1.852
N  1.569  3.098 -11.638  1.381  2.049  0.157  0.276  0.263  0.339  0.043  0.036  0.576  0.062  0.534  0.158  0.198  0.081  0.372  0.181  0.264
Q  0.392  2.569  1.381 -12.734  0.390  0.783  0.825  0.421  2.395  0.070  0.168  0.189  0.205  1.428  0.223  0.415  0.131  0.288  0.253  0.206
D  0.295  0.184  2.049  0.390 -5.754  0.297  0.103  0.273  0.056  0.022  0.030  0.618  0.056  0.366  0.100  0.126  0.075  0.540  0.114  0.058
E  0.075  0.462  0.157  0.783  0.297 -3.888  0.104  0.020  0.042  0.011  0.011  0.227  0.074  0.209  0.139  0.357  0.068  0.346  0.483  0.022
R  0.095  0.531  0.276  0.825  0.103  0.104 -5.535  0.546  0.463  0.019  0.031  0.293  0.115  1.108  0.101  0.191  0.153  0.362  0.173  0.045
D  0.141  0.388  0.263  0.421  0.273  0.020  0.546 -3.532  0.092  0.010  0.056  0.121  0.082  0.290  0.196  0.124  0.074  0.267  0.113  0.055
H  0.236  0.134  0.339  2.395  0.056  0.042  0.463  0.092 -8.410  0.017  0.036  0.088  0.312  3.050  0.158  0.132  0.163  0.465  0.180  0.053
C  0.026  0.101  0.043  0.070  0.022  0.011  0.019  0.010  0.017 -1.396  0.029  0.072  0.046  0.340  0.089  0.231  0.065  0.085  0.104  0.019
P  0.314  0.381  0.036  0.168  0.030  0.011  0.031  0.056  0.036  0.029 -3.905  0.452  0.096  0.077  0.484  0.134  0.083  0.470  0.978  0.037
G  2.878  4.389  0.576  0.189  0.618  0.227  0.293  0.121  0.088  0.072  0.452 -62.493  0.234  0.725  1.682  0.933  0.408  2.405 45.843  0.359
W  0.107  0.192  0.062  0.205  0.056  0.074  0.115  0.082  0.312  0.046  0.096  0.234 -34.252 18.012  5.363  3.532  0.550  4.395  0.595  0.223
Y  0.485  1.209  0.534  1.428  0.366  0.209  1.108  0.290  3.050  0.340  0.077  0.725 18.012 -55.681 19.511  1.881  0.600  3.807  1.434  0.615
F  0.459  0.348  0.158  0.223  0.100  0.139  0.101  0.196  0.158  0.089  0.484  1.682  5.363 19.511 -136.508 21.602  9.725 65.440  7.203  3.528
V  0.565  7.530  0.198  0.415  0.126  0.357  0.191  0.124  0.132  0.231  0.134  0.933  3.532  1.881 21.602 -548.101 275.106 168.339 61.303  5.402
I  0.226  0.574  0.081  0.131  0.075  0.068  0.153  0.074  0.163  0.065  0.083  0.408  0.550  0.600  9.725 275.106 -381.698 81.834  7.808  3.972
L  0.585  3.180  0.372  0.288  0.540  0.346  0.362  0.267  0.465  0.085  0.470  2.405  4.395  3.807 65.440 168.339 81.834 -365.387 16.480 15.727
A  5.694 22.296  0.181  0.253  0.114  0.483  0.173  0.113  0.180  0.104  0.978 45.843  0.595  1.434  7.203 61.303  7.808 16.480 -173.182  1.947
M  0.155  1.852  0.264  0.206  0.058  0.022  0.045  0.055  0.053  0.019  0.037  0.359  0.223  0.615  3.528  5.402  3.972 15.727  1.947 -34.540
```

C. Q_{in}

```
# Q_in: Substitution rate values for the subset of residues from the transmembrane segments facing the interior of the barrel
# Unit: q_ij x 10^-04
      S      T      N      Q      D      E      R      K      H      C      P      G      W      Y      F      V      I      L      A      M
S -187.806  76.525 25.615  5.620  4.276  2.347  3.360  1.567  1.069  0.270  0.513 19.043  0.493  0.397  0.681  2.342  0.885  0.796 37.999  4.010
T  76.525 -146.250  9.066  4.168  1.006  1.433  1.399  2.166  0.536  0.130  0.232  1.416  0.204  1.582  1.434 12.145  3.521  1.788 22.943  4.557
N  25.615  9.066 -64.924  3.878  9.932  0.719  0.825  2.727  1.206  0.145  0.225  3.216  0.330  0.763  0.216  0.630  0.192  0.901  3.060  1.279
Q   5.620  4.168  3.878 -56.578  3.627 19.786  3.569  5.042  1.264  0.165  0.180  0.538  0.203  0.584  0.282  1.558  1.307  2.204  1.796  0.809
D   4.276  1.006  9.932  3.627 -41.038 14.389  0.352  0.291  0.207  0.048  0.136  2.966  0.159  0.248  0.383  0.482  0.134  0.893  1.136  0.372
E   2.347  1.433  0.719 19.786 14.389 -47.877  0.356  0.428  0.168  0.220  0.278  1.002  0.443  0.311  0.203  1.886  0.184  0.440  2.496  0.788
R   3.360  1.399  0.825  3.569  0.352  0.356 -23.582  7.702  1.449  0.174  0.076  0.876  0.085  0.302  0.587  0.717  0.220  0.692  0.664  0.178
D   1.567  2.166  2.727  5.042  0.291  0.428  7.702 -23.477  0.272  0.050  0.139  0.398  0.162  0.101  0.117  0.226  0.307  0.561  1.084  0.138
H   1.069  0.536  1.206  1.264  0.207  0.168  1.449  0.272 -7.364  0.012  0.018  0.242  0.040  0.344  0.060  0.075  0.045  0.193  0.134  0.031
C   0.270  0.130  0.145  0.165  0.048  0.220  0.174  0.050  0.012 -1.861  0.007  0.150  0.016  0.181  0.029  0.067  0.019  0.038  0.118  0.020
P   0.513  0.232  0.225  0.180  0.136  0.278  0.076  0.139  0.018  0.007 -3.094  0.234  0.066  0.116  0.053  0.065  0.038  0.087  0.593  0.039
G  19.043  1.416  3.216  0.538  2.966  1.002  0.876  0.398  0.242  0.150  0.234 -83.068  0.331  0.302  0.268  0.616  0.337  0.561 50.062  0.509
W   0.493  0.204  0.330  0.203  0.159  0.443  0.085  0.162  0.040  0.016  0.066  0.331 -3.887  0.737  0.111  0.063  0.052  0.132  0.188  0.071
Y   0.397  1.582  0.763  0.584  0.248  0.311  0.302  0.101  0.344  0.181  0.116  0.302  0.737 -28.516 21.291  0.310  0.246  0.265  0.259  0.178
F   0.681  1.434  0.216  0.282  0.383  0.203  0.587  0.117  0.060  0.029  0.053  0.268  0.111 21.291 -29.099  0.302  0.356  1.117  0.427  1.184
V   2.342 12.145  0.630  1.558  0.482  1.886  0.717  0.226  0.075  0.067  0.065  0.616  0.063  0.310  0.302 -53.401 15.204  7.038  8.938  0.737
I   0.885  3.521  0.192  1.307  0.134  0.184  0.220  0.307  0.045  0.019  0.038  0.337  0.052  0.246  0.356 15.204 -28.079  2.607  1.458  0.967
L   0.796  1.788  0.901  2.204  0.893  0.440  0.692  0.561  0.193  0.038  0.087  0.561  0.132  0.265  1.117  7.038  2.607 -32.970  7.617  5.037
A  37.999 22.943  3.060  1.796  1.136  2.496  0.664  1.084  0.134  0.118  0.593 50.062  0.188  0.259  0.427  8.938  1.458  7.617 -143.667  2.696
M   4.010  4.557  1.279  0.809  0.372  0.788  0.178  0.138  0.031  0.020  0.039  0.509  0.071  0.178  1.184  0.737  0.967  5.037  2.696 -23.602
```

Figure 5. **Estimated instantaneous substitution rates values.** Q_{all}: Instantaneous substitution rate values estimated for residues embedded within the outer membrane region (Q_{all}). The entries q_{ij} of the rate matrix Q are substitution rates of amino acid residues for the 20 amino acids at an infinitesimally small time interval. The values are in the unit of $\times 10^{-4}$ expected residue changes per 100 site between sequences. Estimating Q_{out}: Instantaneous substitution rate values estimated for the subset of residues from the transmembrane segments facing the lipid environment (Q_{out}). Estimating Q_{in}: Instantaneous substitution rate values estimated for the subset of residues from the transmembrane segments facing the interior of the barrel (Q_{in}).

2.3.2 Residues Similar in Substitution Pattern

To identify residues that behave similarly in their patterns of substitutions, we carried out clustering analysis based on the substitution profile of the 20 amino acids. For each amino acid residue, we collected the substitution rates of replacing this residue type with each of the other 19 residue types. These rates form a 19-dimensional vector. As each of the twenty amino acid types has its own vector, we collected a set of twenty vectors and calculated the Euclidean distances between all pairs of vectors. We then carried out single-linkage hierarchical clustering analysis. This is repeated for each interface region and for the entire TM region. The resulting clustering trees are shown in Figure 6.

There is clear grouping of residues in the clustering tree for the TM_{out} region, which correlates well with the physicochemical properties of residues. A tight cluster consisting of ionizable and polar residues (*i.e.*, K, E, R, Q, D, N, H, S), along with infrequently observed residues (C and P) arise naturally. The aromatic residues W and Y are grouped together, and the small residues G and T are also grouped together. The branched hydrophobic residues (I, V, and L) are also found to cluster together, and are all very different from other residues in their behavior of substitution. Aromatic residue F seems to behave differently from Y and W in substitution, and is grouped closer to the hydrophobic amino acids (A, I, L, V) enriched in the lipid-facing interface. Distances are much larger in this interface due to the larger values of substitutions for hydrophobic residues.

The general pattern for the TM_{in} region is different. Residues have overall lower degree of substitutions in this interface, showing closer distances. Hydrophobic residues (L, I, V), which

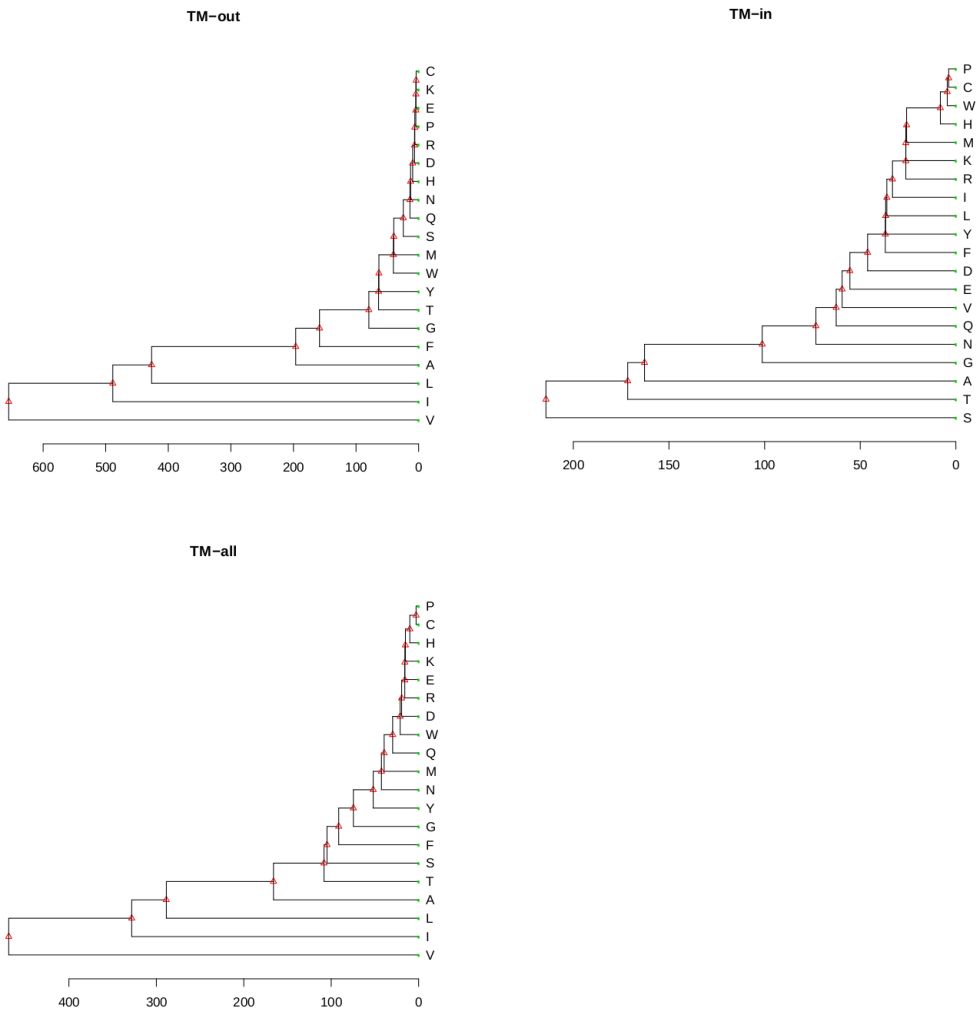

Figure 6. **Clustering trees from amino acid substitutions.** Similarity in substitution pattern for residues in the TM region of β-barrel membrane proteins. Clustering trees showing grouping of residues in the transmembrane regions by similarity in substitution patterns. Residues are clustered by pairwise euclidean distance between the 19-dimensional vectors of instantaneous rates of residue substitutions.

substitute very differently from other residues in the TM_{out} region, are now clustered much closer to other residues. The small residues S, T, A and G, along with N and Q, are grouped together, and show significantly different substitution pattern from other residues.

The hierarchical tree for the TM_{all} interface shows stronger similarities to the tree for the TM_{out} region, reflecting the fact that the substitution pattern in the TM_{out} region dominates. Overall, the hydrophobic residues are found to cluster together (V, I, L, and A). The tight cluster of polar residues and infrequently observed residues are similar to those which are observed in the TM_{out} region.

2.3.3 Detection of Homologs of β-Barrel Membrane Proteins

The estimated amino acid substitution rates can be used to construct scoring matrices for sequence alignment and for large scale database search of homologs of β-barrel membrane proteins. When scoring matrices accurately reflect the evolutionary history of the underlying protein sequences, the detection of homologs usually can be improved significantly (Dayhoff et al., 1978).

Three sets of scoring matrices were derived from the estimated substitution rates (see Figure 7): scoring matrices for the whole TM segments ($bbTM_{all}$, for β-barrel Transmembrane Matrices), for residues facing the interior of the barrel ($bbTM_{in}$); and for residues facing the lipid outer membrane ($bbTM_{out}$) (see Figure 7 for details). These scoring matrices were assessed for performance through BLAST searches against several databases using the TM fragments of a set of 20 β-barrel membrane proteins with known structures as templates.

A. $bbTM_{all}$

```
# Lowest score = -24 , Highest score = 24
    A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V   B   Z   X   *
A   6 -11  -7 -10 -19  -9 -10   1 -12  -2  -2  -9  -5  -5  -8  -1   0 -10 -10   0  -8  -4  -6 -24
R -11  10  -7 -11 -13  -4 -10 -12  -5 -12 -11  -1 -12 -11 -14  -6  -8 -14 -11 -12  -9 -11  -8 -24
N  -7  -7  10   0 -13  -4  -8  -7  -6 -12 -12  -5  -7 -12 -14   0  -2 -13 -10 -11   5  -7  -7 -24
D -10 -11   0  11 -13  -5  -2 -10 -10 -13 -11  -9 -12 -14 -14  -4  -8 -14 -14 -12   6  -6  -8 -24
C -16 -13 -14 -14  24 -11 -16 -16 -16 -15 -16 -14 -13 -14 -14 -12 -13 -16 -12 -15 -14 -16 -12 -24
Q  -9  -4  -4  -5 -11  10  -1 -11  -2  -9 -10  -4  -7 -13 -12  -4  -5 -13 -11  -9  -4  -6  -6 -24
E -10 -10  -8  -2 -24  -1  10 -11 -12 -14 -13 -10 -12 -15 -11  -8  -8 -14 -14 -12  -5  -0  -9 -24
G   1 -12  -7 -10 -24 -11 -11   7 -14  -8  -8 -12  -8 -10 -13  -2  -4 -13 -12  -6  -8  -2  -8 -24
H -12  -5  -6 -10 -24  -2 -12 -14  13 -13 -12 -11 -12 -11 -16  -9 -10 -10  -5 -13  -8 -13  -9 -24
I  -2 -12 -12 -13 -17  -9 -14  -8 -13   7   3 -12  -3  -2 -14  -9  -5  -8  -9   5 -12 -11  -7 -24
L  -2 -11 -12 -11 -18 -10 -13  -8 -12   3   6 -12   0   1 -13  -9  -6  -6  -6   3 -11 -10  -6 -24
K  -9  -1  -5  -9 -13  -4 -10 -12 -11 -12 -12  11 -11 -13 -14  -7  -6 -12 -14 -12  -7 -11  -8 -24
M  -5 -12  -7 -12 -13  -7 -12  -8 -12  -3   0 -11  12  -5 -15  -7  -2 -10 -10  -2  -9 -10  -7 -24
F  -5 -11 -12 -14 -13 -13 -15 -10 -11  -2   1 -13  -5   8 -11 -10  -9  -3   1  -1 -13 -12  -7 -24
P  -8 -14 -14 -14 -13 -12 -11 -13 -16 -15 -13 -14 -15 -11  14 -11 -12 -16 -16 -13 -14 -12 -11 -24
S  -1  -6   0  -4 -13  -4  -8  -2  -9  -9  -9  -7  -7 -10 -11   8   2 -15 -12  -7  -2  -5  -6 -24
T   0  -8  -2  -8 -13  -5  -8  -4 -10  -5  -6  -6  -2  -9 -12   2   8 -12 -10  -3  -5  -6  -5 -24
W -10 -14 -13 -14 -22 -13 -14 -13 -10  -8  -6 -12 -10  -3 -16 -15 -12  11  -1  -6 -13 -13 -10 -24
Y -10 -11 -10 -14 -13 -11 -14 -12  -5  -9  -6 -14 -10   1 -15 -12 -10  -1   7  -8 -12 -13  -8 -24
V   0 -12 -11 -12 -14  -9 -12  -6 -13   5   3 -12  -2  -1 -13  -7  -3  -6  -8   5 -11  -9  -6 -24
B  -8  -9   5   6 -14  -4  -5  -8  -8 -12 -11  -7  -9 -13 -14  -2  -5 -13 -12 -11  11  -4  -7 -24
Z  -4 -11  -7  -6 -16  -6  -0  -2 -13 -11 -10 -11 -10 -12 -12  -5  -6 -13 -13  -9  -4  10  -7 -24
X  -6  -8  -7  -8 -12  -6  -9  -8  -9  -7  -6  -8  -7  -7 -11  -6  -5 -10  -8  -6  -7  -7   1 -24
* -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24 -24   1
```

B. $bbTM_{out}$

```
# Lowest score = -16 , Highest score = 25
    A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V   B   Z   X   *
A   6 -12 -11 -13 -12 -11 -10   1 -13  -1  -1 -13  -5  -3  -8  -3   0  -9  -8   0 -12  -4  -6 -16
R -12  14  -7  -9 -12  -5  -8 -10  -6 -12 -12  -4 -13 -13 -14 -11  -8 -12  -8 -12  -8  -9  -8 -16
N -11  -7  13  -1 -12  -3  -8  -9  -7 -13 -12  -7  -8 -12 -14  -3  -3 -14 -10 -12   6  -8  -7 -16
D -13  -9  -1  14 -12  -7  -5  -8 -12 -13 -11  -6 -12 -13 -14  -8 -11 -14 -11 -13   7  -6  -8 -16
C -14 -13 -11 -12  25 -11 -11 -14 -14 -14 -15 -14 -14 -14 -13 -14 -12 -14 -11 -13 -11 -12 -11 -16
Q -11  -5  -3  -7 -10  12  -4 -12  -2 -12 -12  -6  -9 -12 -10  -8  -4 -10  -8 -11  -5  -8  -7 -16
E -10  -8  -8  -5 -12  -4  17 -11 -12 -12 -12 -13 -14 -12 -16 -11  -8 -13 -13 -11  -6   3  -9 -16
G   1 -10  -9  -8 -12 -12 -11   8 -14  -6  -6 -13  -9  -7  -9  -4  -3 -11 -10  -5  -8  -1  -7 -16
H -13  -6  -7 -12 -12  -2 -12 -14  13 -12 -12 -10 -13 -11 -14  -9 -12  -9  -5 -13  -9 -13  -9 -16
I  -1 -12 -13 -13 -13 -12 -12  -6 -12   5   2 -14  -3  -1 -13  -9  -5  -6  -8   3 -13  -9  -7 -16
L  -1 -12 -12 -11 -15 -12 -12  -6 -12   2   4 -13  -2   1 -11  -9  -5  -5  -6   2 -11  -9  -6 -16
K -13  -4  -7  -6 -12  -6 -13 -13 -10 -14 -13  14 -12 -12 -12 -10  -9 -13 -12 -13  -6 -13  -9 -16
M  -5 -13  -8 -12 -12  -9 -14  -9 -13  -3  -2 -12  12  -3 -13 -10  -4  -9  -9  -3 -10 -11  -7 -16
F  -3 -13 -12 -13 -12 -12 -12  -7 -11  -1   1 -12  -3   6 -10 -10  -8  -3  -2   0 -12  -9  -6 -16
P  -8 -14 -14 -14 -12 -10 -16  -9 -14 -13 -11 -12 -14 -10  12  -8  -9 -13 -15 -12 -14 -12 -10 -16
S  -3 -11  -3  -8 -12  -8 -11  -4  -9  -9  -9 -10 -10 -10  -8  11  -1 -12 -11  -8  -5  -7  -7 -16
T   0  -8  -3 -11 -12  -4  -8  -3 -12  -5  -5  -9  -4  -8  -9  -1   9 -11  -8  -3  -7  -5  -5 -16
W  -9 -12 -14 -14 -12 -10 -13 -11  -9  -6  -5 -13  -9  -3 -13 -12 -11   9  -1  -5 -14 -12  -8 -16
Y  -8  -8 -10 -11 -12  -8 -13 -10  -5  -8  -6 -12  -9  -2 -15 -11  -8  -1   6  -7 -10 -11  -7 -16
V   0 -12 -12 -13 -13 -11 -11  -5 -13   3   2 -13  -3   0 -12  -8  -3  -5  -7   4 -12  -8  -6 -16
B -12  -8   6   7 -11  -5  -6  -8  -9 -13 -11  -6 -10 -12 -14  -5  -7 -14 -10 -12  14  -5  -8 -16
Z  -4  -9  -8  -6 -12  -8   3  -1 -13  -9  -9 -13 -11  -9 -12  -7  -5 -12 -11  -8  -5  15  -8 -16
X  -6  -8  -7  -8 -11  -7  -9  -7  -9  -7  -6  -9  -7  -6 -10  -7  -5  -8  -7  -6  -8  -8   1 -16
* -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16 -16   1
```

C. $bbTM_{in}$

```
# Lowest score = -14 , Highest score = 23
    A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V   B   Z   X   *
A   7  -9  -4  -7 -14  -7  -6   1 -12  -6  -3  -8  -5 -10  -9   0   0 -12 -12  -2  -5  -2  -5 -14
R  -9   8  -8 -11 -12  -5 -11 -10  -6 -11  -9  -2 -12  -9 -14  -6  -7 -14 -11  -8  -9 -10  -8 -14
N  -4  -8   8  -1 -12  -4  -8  -6  -6 -10  -8  -5  -6 -11 -11   0  -2 -10  -9  -8   4  -7  -6 -14
D  -7 -11  -1   9 -14  -4   0  -6 -10 -12  -7 -10  -9 -10 -12  -5  -7 -11 -11  -8   4  -3  -7 -14
C -13 -12 -12 -14  23 -12 -11 -14 -14 -14 -14 -14 -14 -14 -14 -12 -13 -14 -11 -12 -13 -12 -11 -14
Q  -7  -5  -4  -4 -12   8   0 -10  -6  -6  -5  -3  -8 -11 -12  -4  -5 -11 -10  -6  -4  -5  -6 -14
E  -6 -11  -8   0 -11   0   8  -9 -12 -11 -10 -10  -8 -12 -11  -7  -7 -10 -11  -6  -4  -0  -7 -14
G   1 -10  -6  -6 -14 -10  -9   6 -13 -11  -9 -11  -9 -13 -13  -2  -5 -12 -13  -8  -6  -1  -8 -14
H -12  -6  -6 -10 -14  -6 -12 -13  14 -13 -10 -10 -14 -13 -14  -8  -9 -13  -9 -12  -8 -12  -9 -14
I  -6 -11 -10 -12 -14  -6 -11 -11 -13  12  -3 -10  -5  -8 -13  -8  -4 -13 -10   3 -11 -11  -7 -14
L  -3  -9  -8  -7 -14  -5 -10  -9 -10  -3  10  -9  -1  -6 -12  -8  -6 -12 -11  -1  -7  -9  -6 -14
K  -8  -2  -5 -10 -14  -3 -10 -11 -10 -10  -9   9 -12 -13 -12  -7  -6 -12 -14 -10  -7 -10  -8 -14
M  -5 -12  -6  -9 -14  -8  -8  -9 -14  -5  -1 -12  12  -5 -13  -4  -3 -12 -11  -6  -7  -8  -7 -14
F -10  -9 -11 -10 -14 -11 -12 -13 -13  -8  -6 -13  -5  11 -13  -9  -7 -11   2  -9 -10 -12  -8 -14
P  -9 -14 -11 -12 -14 -12 -11 -12 -14 -13 -12 -12 -13 -13  16 -10 -11 -11 -12 -12 -11 -11 -10 -14
S   0  -6   0  -5 -12  -4  -7  -2  -8  -8  -8  -7  -4  -9 -10   6   2 -10 -11  -5  -2  -4  -5 -14
T   0  -7  -2  -7 -14  -5  -7  -5  -9  -4  -6  -6  -3  -7 -11   2   7 -12  -7  -1  -4  -6  -5 -14
W -12 -14 -10 -11 -14 -11 -10 -12 -13 -13 -12 -12 -12 -11 -11 -10 -12  13  -7 -13 -10 -11 -10 -14
Y -12 -11  -9 -11 -11 -10 -11 -13  -9 -10 -11 -14 -11   2 -12 -11  -7  -7   9 -10 -10 -12  -8 -14
V  -2  -8  -8  -8 -12  -6  -6  -8 -12   3  -1 -10  -6  -9 -12  -5  -1 -13 -10  11  -8  -7  -6 -14
B  -5  -9   4   4 -13  -4  -4  -6  -8 -11  -7  -7  -7 -10 -11  -2  -4 -10 -10  -8   9  -4  -6 -14
Z  -2 -10  -7  -3 -12  -5  -0  -1 -12 -11  -9 -10  -8 -12 -11  -4  -6 -11 -12  -7  -4   8  -6 -14
X  -5  -8  -6  -7 -11  -6  -7  -8  -9  -7  -6  -8  -7  -8 -10  -5  -5 -10  -8  -6  -6  -6   1 -14
* -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 -14   1
```

Figure 7. **bbTM Scoring Matrices.** A. Scoring matrix $bbTM_{all}$. Scoring matrix at evolutionary time unit of 40 derived from Q_{all}. Note: The last line denotes the lowest score in all columns and rows to conform with format of BLOSUM and PAM for use in programs such as CLUSTALW. B. Scoring matrix $bbTM_{out}$. Scoring matrix at evolutionary time unit of 40 derived from Q_{out}. C. Scoring matrix $bbTM_{in}$. Scoring matrix at evolutionary time unit of 36 derived from Q_{in}.

To obtain objective evaluation, we constructed a "true-positive" data set containing known and predicted β-barrel membrane proteins, as well as a data set of negative controls consisting of randomized sequences of β-barrel and α-helical membrane proteins. We created the first data set by combining 2,130 predicted β-barrel membrane proteins sequences from the PROFTMB database constructed by Bigelow et al (Bigelow et al., 2004), with an additional 1,266 sequences annotated as bacterial outer membrane proteins in the UNIPROT database (Barker et al., 1998). We excluded those proteins with more than 90% identity with the 11 proteins from which we estimated the substitution rates. After removal of redundant sequences, we have a total of 3,079 sequences. The second data set consists of random sequences obtained from fully shuffled sequences of 385 α-helical and β-barrel membrane proteins from different organisms. These random sequences preserve the same amino acid composition as membrane proteins. Assuming none of the randomized sequences resemble a true β-barrel membrane protein, they form a challenging set of "true negatives".

Two additional data set were constructed from the UNIPROT database (Barker et al., 1998). The first data set consists of membrane proteins with a different architecture (non-β-barrel). These were selected based on annotations of "SUBCELLULAR LOCATION: Cell membrane" from *Eukaryota* and *Archaea*. We make the reasonable assumption that β-barrel membrane proteins cannot be found in the cellular membrane of these organisms, and all these membrane proteins are expected to adopt a different three dimensional topology. In total, 10,951 these other-membrane protein sequences were included in the data set (called OMBP, for other MemBrane proteins). The second data set consists of globular protein sequences. We selected

from UNIPROT proteins with annotations that lack the word "membrane". In total, 127,485 globular protein sequences were included in the data set (called GLOBULAR).

We use the concatenated transmembrane segments of 11 proteins from which the scoring matrices were derived, along with an additional 9 β-barrel membrane proteins, as templates to search the databases for homologs. These 20 proteins share less than 20% pairwise sequence identity. Our goal was to detect homologs of β-barrel membrane proteins with accuracy and specificity. We use the simple criterion that resulting hits from BLAST searches using these customized scoring matrices must have e-values smaller than 10^{-1}. e-value measures the statistical significance of matched sequences from database search. It gives the expected total number of hits in a database search one would find by random chance (Altschul et al., 1990a). We therefore set the threshold of e-value to be 10^{-1}. We also require that the alignment must be of a minimum length. Since α-hemolysin has the smallest number of strands in forming a β-barrel membrane, we require that the matched sequence must be at least of the length of about two transmembrane segments. In α-hemolysin, two TM strands form a hair-pin, and seven repeats of hair-pins form the β-barrel membrane protein (Song et al., 1996). Assuming that at least 5 amino acids need to be matched in a TM strand, an hairpin would require 10 amino acids to be matched. We therefore require that an alignment should be no less than 10 residues.

2.3.3.1 Evaluation of Specificity using Random Sequences, Other (Non-β-Barrel) Membrane Proteins and Globular (Non-Membrane) Protein Sequences

We first carry out a test of specificity. Results of BLAST searches against the randomized database are shown in Table II. A perfectly discriminative scoring matrix should not select any sequence from the database of randomized membrane proteins. Search results using the *bbTM* matrices showed excellent specificity, with no sequences retrieved from the random database. Although the default matrix BLOSUM62 used in BLAST searches were designed for soluble proteins and is not suitable for homology detection of membrane proteins, it did not retrieved sequences from the random database.

TABLE II

SPECIFICITY OF SCORING MATRICES: BLAST SEARCHES AGAINST RANDOM SEQUENCES DATABASE

e-value	$bbTM_{all}$	$bbTM_{in}$	$bbTM_{out}$	BLOSUM62	PHAT7573	SLIM161	PAM250
$< 10^{-20}$	0	0	0	0	0	0	0
$< 10^{-10}$	0	0	0	0	0	0	0
$< 10^{-7}$	0	0	0	0	0	0	0
$< 10^{-5}$	0	0	0	0	0	20	0
$< 10^{-4}$	0	0	0	0	0	52	0
$< 10^{-3}$	0	0	0	0	0	142	5
$< 10^{-2}$	0	0	0	0	5	319	42
$< 10^{-1}$	0	0	0	0	45	689	181

Cumulative number of random sequences incorrectly identified as homologs of β-barrel membrane proteins at different *e*-value resulting from BLAST searches against a database of 362 randomized membrane proteins sequences using as queries the concatenated transmembrane segments of 20 template β-barrel membrane proteins.

The scoring matrix PHAT constructed for helical membrane proteins does not work well for β-barrel membrane proteins. It selected a total of 45 random sequences with e-values less than 10^{-1}, five of which with e-value in the range of 10^{-3} to 10^{-2}. That is, 12.4% of the random sequences were mistakenly identified as membrane proteins. The performance of another scoring matrix, SLIM, constructed for helical membrane proteins, also had poor performance: random sequences started to be selected at the significant e-values in the range of 10^{-7} to 10^{-5}, with a total of 689 random sequences selected at the e-values less than 10^{-1}, using 20 proteins as query sequences. Similarly, BLAST searches using the classical PAM250 matrix resulted in 181 random sequences with significant e-values less than 10^{-1}.

When searches were carried out against the data set of other membrane proteins (OMBP) and the data set of globular proteins (GLOBULAR), the matrices *bbTM* and BLOSUM62 showed excellent specificity, with no sequences erroneously identified as β-barrel membrane proteins (Table III). In contrast, varying numbers of other membrane proteins and soluble proteins were erroneously identified as β-barrel membrane proteins when PAM250, PHAT and SLIM matrices were used. Among these, the SLIM matrix resulted in a very large number (1,780) of misidentified non-β-membrane proteins.

We conclude that the scoring matrices PHAT, SLIM, and PAM are not suitable for database search of β-barrel membrane proteins.

2.3.3.2 Detection of Outer Membrane Proteins

Next we performed BLAST searches against the "true-positive" database of outer membrane proteins. Search results are shown in Table IV. *bbTM* matrices retrieved larger numbers of

TABLE III

SPECIFICITY OF SCORING MATRICES: BLAST SEARCHES AGAINST A DATA SET OF MEMBRANE PROTEINS WITH OTHER ARCHITECTURE AND A DATA SET OF GLOBULAR PROTEINS (OMBP/GLOBULAR).

e-value	$bbTM_{all}$	$bbTM_{in}$	$bbTM_{out}$	BLOSUM62	PHAT7573	SLIM161	PAM250
$< 10^{-20}$	0/0	0/0	0/0	0/0	0/0	0/0	0/0
$< 10^{-10}$	0/0	0/0	0/0	0/0	0/0	0/0	0/0
$< 10^{-7}$	0/0	0/0	0/0	0/0	0/0	0/0	0/0
$< 10^{-5}$	0/0	0/0	0/0	0/0	0/0	2/2	0/0
$< 10^{-4}$	0/0	0/0	0/0	0/0	0/0	21/3	0/0
$< 10^{-3}$	0/0	0/0	0/0	0/0	0/1	98/5	1/1
$< 10^{-2}$	0/0	0/0	0/0	0/0	3/6	457/13	3/2
$< 10^{-1}$	0/0	0/0	0/0	0/0	13/26	1780/42	28/31

Cumulative number of sequences of membrane proteins with other architecture and globular protein sequences incorrectly identified as homologs of β-barrel membrane proteins at different e-values resulting from BLAST searches against the OMBP/GLOBULAR data set. The number of sequences part of OMBP is 10,951 (1,061 from *Archaea* and 9,890 from *Eukaryota*). The size of the data set GLOBULAR is 127,485 globular protein sequences (16,814 *Archaea* and 110,671 *Eukaryota*). We used as queries the concatenated transmembrane sequences of the 20 template proteins.

true positives, while maintaining excellent specificity as discussed before. The numbers of true positives retrieved using $bbTM_{all}$, $bbTM_{in}$, and $bbTM_{out}$ are 191, 166, and 245, respectively. Each of these proteins is related to one of the 20 query proteins and shares the same structure. As discussed earlier, the PHAT and SLIM scoring matrices designed for helical membrane proteins are inappropriate for search of β-barrel membrane proteins, as they lack specificity and will select many false positives. The BLOSUM62 matrix performs poorly in detecting β-barrel membrane proteins, with only 5 proteins identified at e-values of $< 10^{-20}$. Altogether, only 126 true positives at e-values $< 10^{-1}$ were identified.

TABLE IV

PERFORMANCE OF *BBTM* MATRICES IN DETECTING HOMOLOGS OF β-BARREL MEMBRANE PROTEIN SEQUENCES FROM THE "TRUE-POSITIVE" DATABASE.

e-value	$bbTM_{all}$	$bbTM_{in}$	$bbTM_{out}$	BLOSUM62	PHAT7573	SLIM161	PAM250
$< 10^{-20}$	49	62	56	5	48	46	8
$< 10^{-10}$	116	106	121	32	121	119	41
$< 10^{-07}$	122	121	129	42	133	130	79
$< 10^{-05}$	128	127	143	83	141	143	102
$< 10^{-04}$	138	131	147	95	148	145	107
$< 10^{-03}$	146	139	168	109	176	170	119
$< 10^{-02}$	153	144	206	120	200	202	136
$< 10^{-01}$	191	166	245	126	272	260	202

Cumulative number of proteins identified as homologs of 20 template β-barrel membrane proteins at different e-values obtained from BLAST searches against the "true-positive" database of 3,079 sequences of β-barrel membrane proteins.

Finally, we also performed BLAST searches against the non-redundant NCBI database (Table V). The *bbTM* matrices retrieved the largest number of hits compared to PHAT or the classical PAM matrices. As discussed earlier, the SLIM matrices suffer from the problem of very low specificity and would erroneously select many false positives. As expected, the BLOSUM62 matrix, despite its excellent specificity, performed poorly: only a small number of sequences were identified from the non-redundant NCBI database.

TABLE V

PERFORMANCE OF *BBTM* MATRICES IN DETECTING HOMOLOGS FROM THE NON-REDUNDANT NCBI PROTEIN SEQUENCE DATABASE.

e-value	$bbTM_{all}$	$bbTM_{in}$	$bbTM_{out}$	BLOSUM62	PHAT7573	SLIM161	PAM250
$< 10^{-20}$	821	934	897	65	605	608	103
$< 10^{-10}$	1556	1579	1977	294	1781	1832	416
$< 10^{-07}$	2020	1879	2211	504	2120	2749	649
$< 10^{-05}$	2201	2135	2327	650	2309	4040	812
$< 10^{-04}$	2262	2212	2377	708	2385	5516	1142
$< 10^{-03}$	2322	2288	2464	856	2477	7495	1475
$< 10^{-02}$	2407	2437	2602	1198	2570	8538	1677
$< 10^{-01}$	2573	2573	2757	1503	2799	9192	1966

Cumulative number of proteins identified as homologs of the 20 template β-barrel membrane proteins at different *e*-value obtained from BLAST searches against the non-redundant NCBI protein database of 13,135,398 sequences.

In summary, the $bbTM_{out}$ and $bbTM_{all}$ matrices have the best performance among all the matrices tested, with the highest number of "true-positives" detected, while maintaining ex-

cellent specificity without erroneously identifying any random sequence, membrane proteins with other architecture, and globular (non-membrane) proteins in our tests at any threshold of e-value. Although the classical BLOSUM62 matrix shows excellent specificity, it has poorer performance in identifying β-barrel homologous proteins. Among membrane protein specific matrices, PHAT retrieves a larger number of true-positive hits, but suffers from the problem of insufficient specificity, as it consistently misidentified random sequences, sequences for other membrane proteins, as well as soluble protein sequences as β-barrel membrane proteins. SLIM shows the poorest performance, as it suffers from generating a significant number of false positives.

2.3.4 Detection of Mitochondria Membrane Proteins

It was estimated that a large number of β-barrel membrane proteins are located at the outer membrane of mitochondria (Wimley, 2003), but only four families have been confirmed to date (Walther et al., 2009b). An interesting question is whether our scoring matrices can be used to detect mitochondria β-barrel proteins. To answer that question, we performed BLAST searches against the non-redundant NCBI database of protein sequences, using transmembrane segments of three different mitochondrial β-barrel membrane proteins as queries. These are the voltage-dependent anion channel (VDAC), the only mitochondrial porin with known structure; the predicted transmembrane segments of TOM40 (Zeth, 2010), the main component of the translocation machinery of mitochondria (Hill et al., 1998); and SAM50, an essential component of the sorting and assembly machinery (Paschen et al., 2003). Using the matrices $bbTM_{all}$, $bbTM_{in}$, and $bbTM_{out}$, we obtained 266, 277 and 269 homologous proteins, respectively, at the

significant level of e-values less than 10^{-20}, and a total of 383, 379 and 388 at e-values less than 10^{-1}. All of these proteins have been verified as mitochondrial proteins by manual inspection of annotations (Table VI).

TABLE VI

PERFORMANCE OF *BBTM* MATRICES IN DETECTING HOMOLOGS OF THE HUMAN MITOCHONDRIAL PROTEINS VDAC, TOM40 AND SAM50

e-value	$bbTM_{all}$	$bbTM_{in}$	$bbTM_{out}$
$< 10^{-20}$	266	277	269
$< 10^{-10}$	335	324	348
$< 10^{-07}$	355	354	360
$< 10^{-05}$	364	361	371
$< 10^{-04}$	369	364	373
$< 10^{-03}$	378	370	381
$< 10^{-02}$	381	376	384
$< 10^{-01}$	383	379	388

Cumulative number of proteins identified as homologs of the human mitochondrial β-barrel membrane proteins VDAC-1 (uniprot:VDAC1_HUMAN), TOM40 (uniprot:TOM40_HUMAN) and SAM50 (uniprot:SAM50_HUMAN), at different e-value obtained from BLAST searches against the non-redundant NCBI database of 13,135,398 sequences. These hits are all confirmed to be mitochondria proteins by manual inspection of annotation.

2.3.5 Implications for Template-Based Structure Prediction of β-Barrel Membrane Proteins

An important implication of our results is that we can now reliably detect remote homologs of β-barrel membrane proteins with known structures at genome scale. This will allow predic-

tion of high quality structural models of β-barrel membrane proteins through template-based modeling (Zhang and Skolnick, 2005). Here we estimate the number of β-barrel membrane proteins in the OMPdb database (Tsirigos et al., 2010) whose TM-structures can be modeled reliably through alignments against the template protein structures using the $bbTM_{out}$ matrix. We found that at the e-value of less than 10^{-1} and with at least 75 amino acids to ensure at least 8 transmembrane strands identified, there are a total of 2,619 protein sequences that can be mapped onto one of the 20 known structures we used (Figure 8). On average, each template can be used to model the structures of 131 membrane protein sequences.

2.3.6 Designing OmpF Mutants with an Alter Oligomerization State

In order to discover residues that contribute significantly to the protein-protein interaction interface of OmpF, we used a computational approach to identify a conserved patch in the protein-protein interaction and we replaced the key residues with substitutions not found in natural β-barrel membrane proteins based on the estimated amino acid substitution rates obtained.

We first describe the highly conserved residues identified by our computational method. We then experimentally assess the contributions of these residues to the protein-protein interaction interface of OmpF.

2.3.6.1 Evolutionarily Conserved Residues in the Protein-Protein Interaction Interface of OmpF

The overall picture of residue conservation of OmpF is in agreement with the pattern described above, *i.e.*, out-facing residues located at the protein-protein interface are found to be

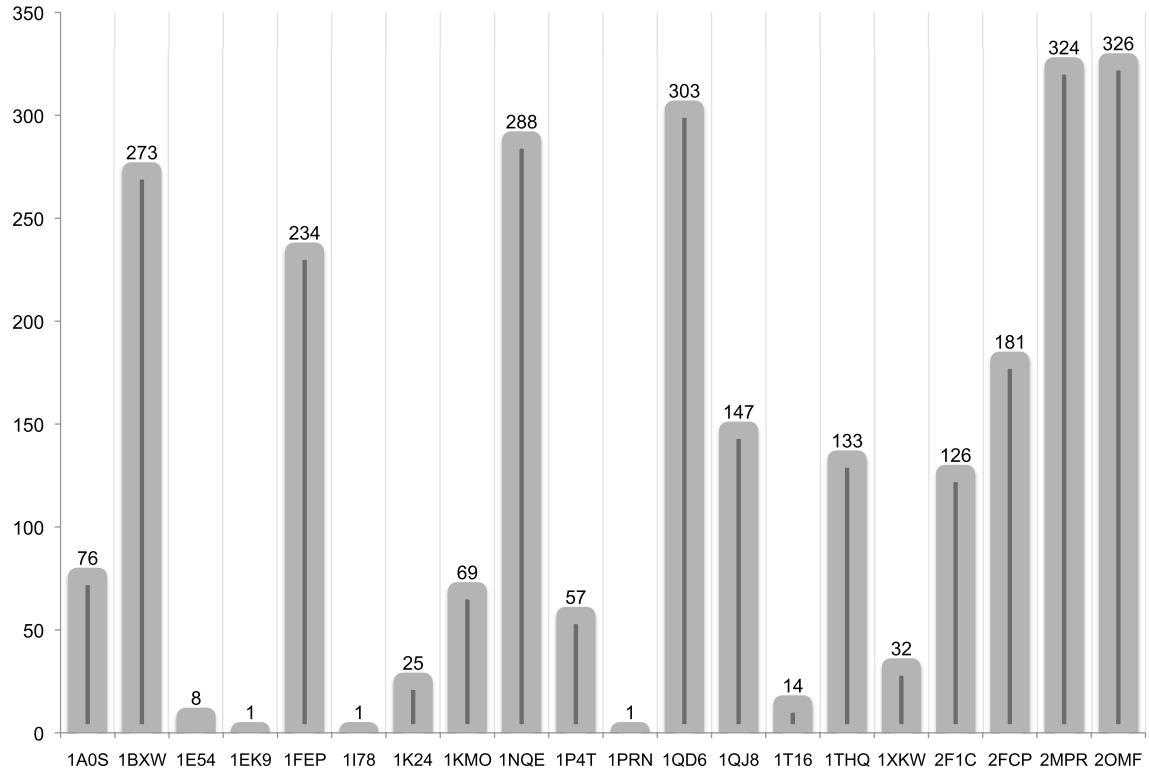

Figure 8. **Number of β-barrel membrane proteins homologous to the 20 proteins with known structures.** There are altogether 2,619 proteins in the OMPDB database of β-barrel membrane proteins, whose TM regions can be mapped onto one of the 20 proteins by using the $bbTM_{out}$ scoring matrix. Structures of the TM regions of these proteins can then be predicted by using template-based structure prediction methods.

better conserved than those facing the lipids. Among them, a group of well-conserved residues are found to form a structurally contiguous surface patch (Figure 9a), including the C-terminal end of the second strand (Gly47 and Thr49) and the N-terminal end of the third strand (Leu55 to Gly59) (Figure 9b-c).

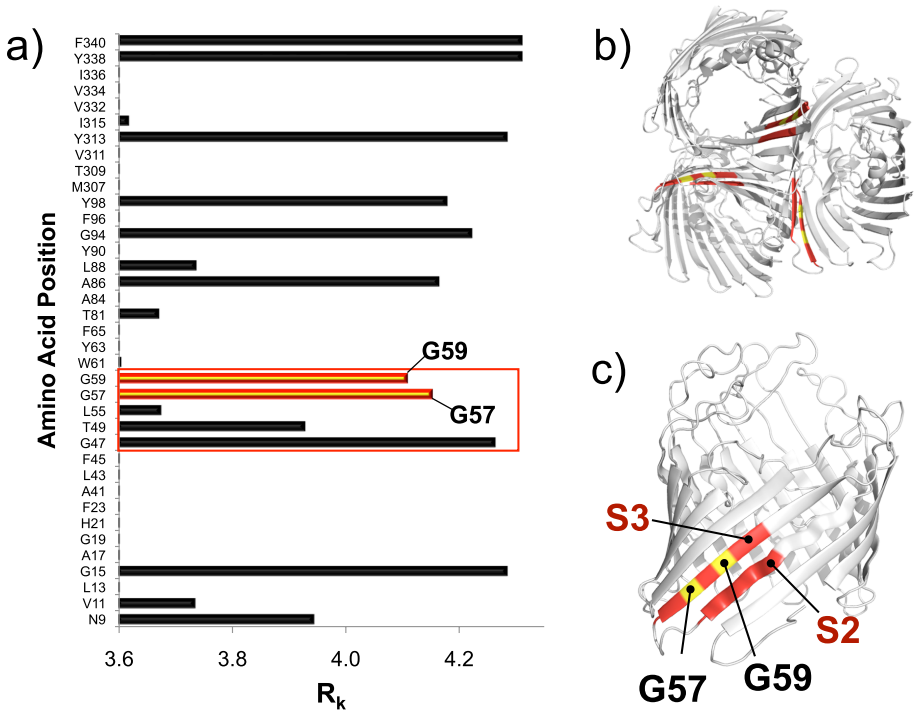

Figure 9. **Evolutionary conserved positions of amino acids are found at the protein-protein interface of OmpF.** a) Conservation entropy (in bits) of transmembrane residues. For illustration, we show $R_k = \log_2 20 - H_k$, where well-conserved columns show longest bars. The red box highlights a well-conserved structurally contiguous surface patch at the oligomerization interface (strands S2 and S3). b) and c) Structural locations of strands S2 and S3 in the trimer and monomer, respectively.

To verify the roles of this well-conserved surface patch in maintaining the stability and promoting oligomerization of OmpF, we designed three mutants: G57A/G59A, G57S/G59S, and G57I/G59L. These substitutions were chosen based on analysis of the evolutionary pattern of amino acid substitutions at the transmembrane segments of β-barrel membrane proteins obtained and described in this work. According to our results, glycine is well-conserved in the out-facing lipid interface. The few substitutions detected for glycine were replacements with serine, threonine and alanine. Long aliphatic and hydrophobic residues (L,V,I,A) were found to readily substitute among themselves in this interface, but not to glycine.

With these observed patterns of substitutions, we hypothesized that substitution of the glycine residues located in the well-conserved surface patch (*i.e.*, G57 and G59) to leucine, valine or isoleucine would affect the stability of the OmpF trimer. In contrast, substitutions of G57 and G59 to serine or alanine would be expected not to have significant impact on the stability of the trimer of OmpF.

2.3.6.2 Secondary and Tertiary Structure of Wild Type and Mutant OmpF Proteins

In order to test to what extent the predicted residues contribute to the overall oligomerization state, we first assessed the folding of wild type and mutant OmpF porins. Wild type and mutant OmpF proteins were expressed in *E. coli* and purified from inclusion bodies under denaturing conditions. Proteins were refolded by dilution of the denaturant into the refolding buffer as reported in (Visudtiphole et al., 2005). SDS-PAGE and Coomassie blue staining indicated that all of the mutants and the wild type protein were successfully expressed and purified.

CD spectra analyses confirmed that the mutants had the same typical β-barrel spectra as wild type OmpF, with a peak around 217 nm (Figure 10).

In addition, the protein samples were free of higher-order aggregates as indicated by the the CD spectra, which approaches ellipticity values close to zero at wavelengths >250 nm. The tertiary structure of the wild and mutant OmpF proteins was analyzed via tryptophan fluorescence spectroscopy. Wild type, the G57A/G59A mutant, G57S/G59S mutant, and G57I/G59L mutant have two tryptophan residues at positions 61 and 214. The emission spectra of the wild type, and mutants G57A/G59A, G57S/G59S, and G57I/G59L were similar (Figure 10), with intensity maxima at approximately 318 nm and an unchanged width of emission spectra. The intensity maxima for these mutant proteins is around 327 nm with a similar width of emission as the wild type protein. Overall, these observations can be interpreted as an indication of an unaltered environment of the tryptophan residues compared to the wild type. Any observed differences in thermal stability and oligomeric state are therefore unlikely to be due to altered protein structures.

2.3.6.3 Engineered Monomeric Oligomerization States

To compare the oligomerization states of the wild type and OmpF mutants, we analyzed these proteins by SDS-PAGE and Coomassie blue staining (Figure 11). Mutants G57I/G59L (Figure 11a, Lane 7), which is designed to alter the PPI interface, indeed was found to exist in a folded monomeric form instead of the trimeric form, in addition to the unfolded form. The wild type (Figure 11a, Lane 2), as well as two mutants G57A/G59A and G57S/G59S (Figure 11a, Lane 3 and 4) were found to be present in the folded trimeric state and an unfolded

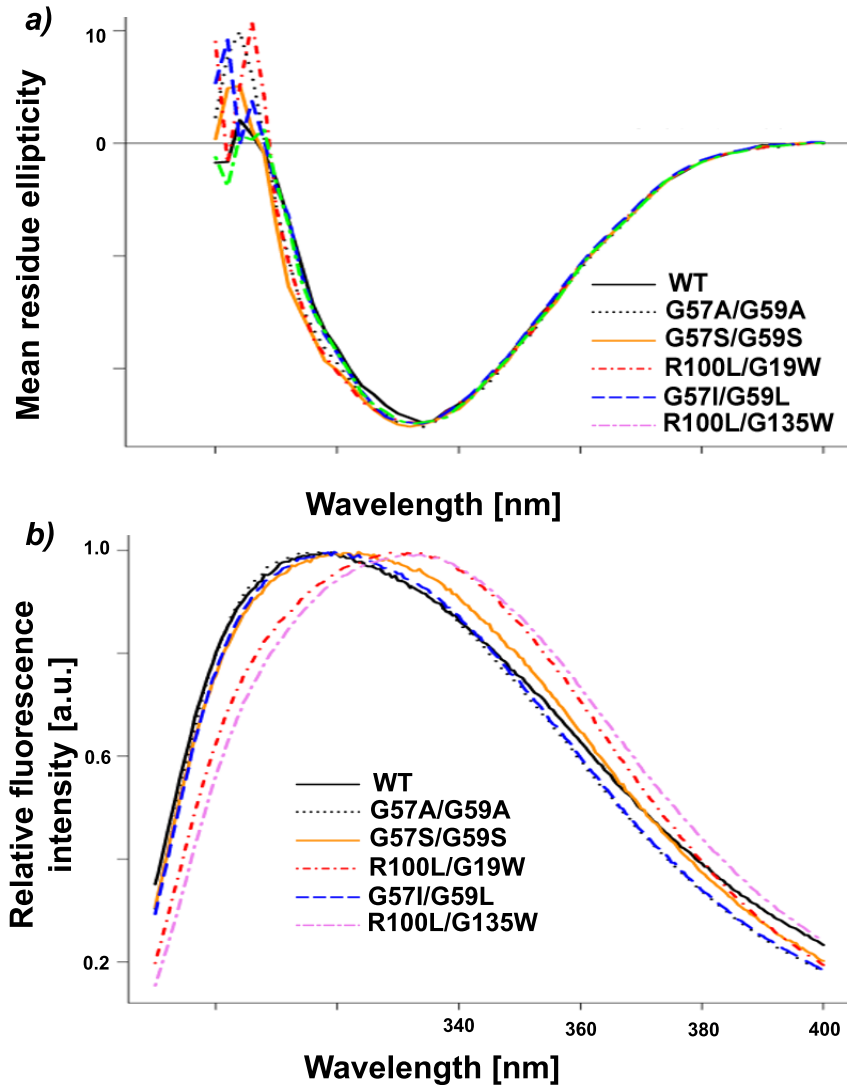

Figure 10. **UV-CD spectra and tryptophan fluorescence emission spectra of wild-type and mutant OmpF.** a) Comparison of UV-CD spectra between wild-type and mutant OmpF protein. Measurements were recorded for protein concentration of 0.2 mg/mL. An average of three scans at 50 nm/min was acquired with a bandwidth of 0.2 nm and a response time of 1 second using three independent protein preparations. Final CD spectrum was then corrected for background by subtraction of spectrum of protein-free samples recorded under the same conditions. Noisy data below 200 nm has been removed. b) Comparison of tryptophan fluorescence emission spectra of wild type and mutant OmpF proteins. Measurements were recorded using samples containing 0.2 mg/mL protein held in a 1 mm path length cuvette, with an excitation wavelength of 290 nm. All mutants and wild-type OmpF protein exhibit similar characteristics in UV-CD and tryptophan fluorescence emission spectra. Thus, secondary structure formation and environment of the tryptophan residues in wild-type and mutant OmpF proteins are similar.

state. This was also in agreement with our experimental design, in which the two substitutions (G57A/G59A and G57S/G59S) follow the substitution pattern observed in β-barrel membrane proteins and were not expected to have an altered the oligomerization state.

Trypsin resistance has been used in previous studies as a useful indicator of protein folding (Baldwin et al., 2011; Surrey et al., 1996). We observed that the folded oligomeric and monomeric bands (Figure 11a) were resistant to trypsin digestion, while the unfolded band disappeared after trypsin treatment (Figure 11b). This indicates that the folded oligomeric and monomeric species are compact and well folded. Moreover, when the samples were heated to $95°C$, both folded oligomeric and monomeric bands disappeared, and the density of the unfolded band increased. Furthermore, mass spectrometric analysis also revealed that only OmpF protein was present in these samples (data not shown).

Overall, our results from SDS-PAGE, trypsin digestion, thermal experiments, and mass spectrometric analysis confirmed that the oligomerization state of mutant G57I/G59L was changed from a trimeric form to a monomeric form.

2.3.6.4 Residues G57 and G59 Contribute Significantly towards the Stability of the Oligomerization State

We further assessed the stability of the oligomeric state of wild type and mutant OmpF proteins. In thermal denaturation experiments, wild type OmpF trimers were found to dissociate into monomers at $73(\pm1)°C$. As expected, the double mutants G57A/G59A and G57S/G59S designed through evolutionary analysis did not significantly decrease the temperature of trimer

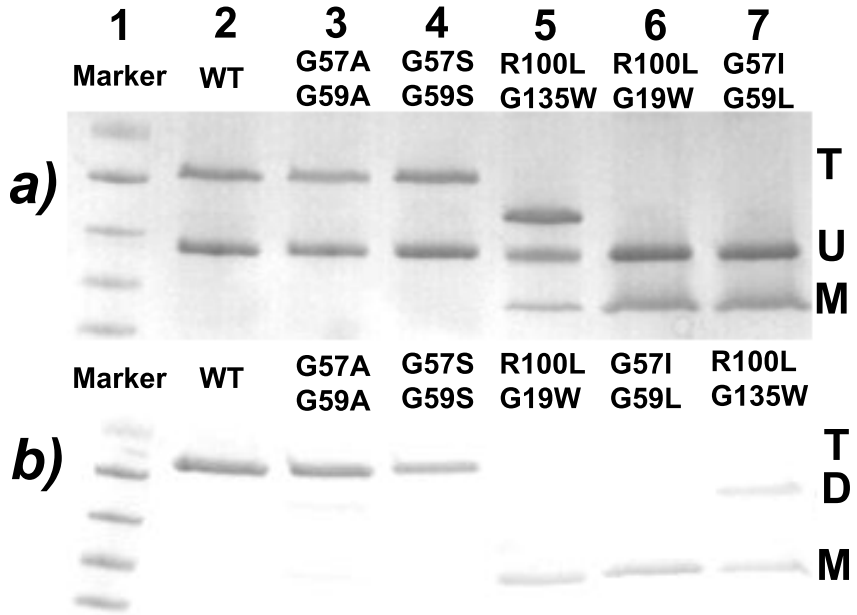

Figure 11. **Oligomerization state of wild type and OmpF mutants.** OmpF proteins were expressed, purified, and refolded from inclusion bodies as described in Methods. Folding reactions were quenched by adding 5× SDS gel-loading buffer to a final dilution of 1× SDS gel loading buffer. 50 μl of sample was loaded on a 4-20% acrylamide continous gradient precast gel to resolve the folded and unfolded populations. a) The wild type and mutants (G57A/G59A and G57S/G59S) were only present in either folded trimeric state (T) or unfolded state (U). The mutant G57I/G59L have folded monomeric (M) and unfolded species. b) After overnight refolding, subsequent degradation of the unfolded protein was induced by the addition of trypsin (trypsin/protein 1:100 w/w). Final protein was purified and concentrated using Centrifugal Filters. The folded oligomeric and monomeric bands were resistant to trypsin digestion, while the unfolded band disappeared after treatment with trypsin.

disassociation. In contrast, G57I/G59L mutant, which was designed with an expected altered oligomerization state, existed primarily in the monomeric form at room temperature.

2.4 Discussion

2.4.1 Patterns of Amino Acid Substitutions at Lipid Interfaces

The estimated substitution rates reveal characteristic patterns common to all β-barrel membrane proteins. For residues facing the interior of the barrel, stronger overall sequence conservation is observed. Residues facing the lipid membrane (TM_{out}) are less conserved and have more substitutions. About twice as many substitutions occur in the TM_{out} region. However, the pattern of substitution in the TM_{out} region is very narrow.

The most frequently observed substitutions in this region are among branched aliphatic or small hydrophobic residues (*i.e.*, V-I, V-I, I-L, A-V or A-L), all with very similar physicochemical properties. Substitutions between aromatic residues (*e.g.*, Y-F and Y-W) are also frequently detected at this interface. Among the aromatics, W has a much larger presence in the TM region (5% in TM_{out}, 2% in TM_{in}) compared to its expected presence in proteins contained in the UNIPROT database (1%, data not shown). It is enriched in the aromatic girdles, and has an overall low substitution rate. W likely plays important roles in maintaining the stability or function of β-barrel membrane proteins.

Substitutions of polar residues frequently occur among themselves, and also with A, G and V. They are likely to be involved in the maintenance of inter-strand polar-polar motifs as described in a previous study (Jackups and Liang, 2005). Some examples of these substitutions can be found in the ferric receptor FepA, the sucrose porin ScrY, the transporters FecA and BtuB, and the ferric hydroxamate uptake receptor FhuA.

With the exception of Glu, ionizable residues in the TM_{out} region are mostly found in the lipid-water interface. They are found in large β-barrel membrane proteins (*e.g.*, ScrY, 18-strands; FepA, 22-strands; BtuB, 22-strands; FhuA, 22-strands; lamB, 18-strands; and OmpF, 16 strands), but not in smaller proteins (*e.g.*, none in OmpA, 8-strands; OmpT, 10-strands; and OmpX with 8-strands).

The overall pattern of substitution of the TM_{out} interface suggests that there exists a rich and specific substitution pattern, reflecting strong selection pressure at this interface for amino acids to maintain the same physicochemical properties. This is perhaps the reason why the $bbTM_{out}$ scoring matrices perform the best in identifying remote homologs of β-barrel membrane proteins.

2.4.2 Physical Basis of the Amino Acid Substitutions in the Transmembrane Region

There are physical constraints on allowed substitutions due to the requirement of folding and stability of β-barrel membrane proteins. For example, the membrane environment and the formation of anti-parallel β-strands are strong constraints that are reflected in the observed substitution pattern.

Anti-parallel strands are arranged with all hydrophobic residues on the side of the barrel facing the lipid interface. Residues L, V, A, F, I and W are frequently found in this interface, which is in agreement with the GES and RW hydrophobicity scales (Engelman et al., 1986; Radzicka and Wolfenden, 1988; Bishop et al., 2001b). Under this constraint, these hydrophobic residues are found to mostly exchange among themselves.

The aromatic girdle represents another structural constraint, where W and Y are enriched. Both W and Y residues at the aromatic girdle are important for the β-barrel stability, as evidenced by their large TM-region propensities and the frequently occurring spatial motifs of non-H-bonded W-Y interactions (Jackups and Liang, 2005). These two residues have very limited substitutions, mostly among themselves. In addition, both aromatic residues may help to facilitate the folding and insertion of the protein into the membrane in a concerted fashion (Huysmans et al., 2010; Kleinschmidt et al., 1999).

The result that abundant G is strongly conserved is consistent with the findings from an earlier study, in which it was shown that the substitution of a residue is only weakly influenced by the composition in amino acids, but strongly depends on the constraints of carrying out biological functions and maintaining structural integrity (Tourasse and Li, 2000). One example of such constraints is the interaction between G and Y on neighboring strands. In an earlier study, G was found to form strong back-bone H-bonds interactions with aromatic residues. This interstrand interaction, called aromatic rescue (Weiss et al., 1991), likely plays an important role stabilizing these membrane proteins (Jackups and Liang, 2005).

The lipid-water interface at the end of the β-strands also imposes additional constraints, which lead to the placement of many polar residues (S, T, Q or N) and ionizable residues.

Since the interior of the barrel is the location where these proteins interact with ions, metabolites, and substrates, amino acids in this interface are under strong selection pressure to carry out specific biological functions. As a consequence, there are limited substitutions for residues in this interface (Figure 4C).

Aromatic residues facing the TM_{in} region show a strong conservation as well. Only exchanges between Y-F (21) are observed in this interface, which suggest a strong structural constraint for these residues to be located in specific parts of the interior of the β-barrel membrane protein, delineating the pathway for substrates across the lumen of the pore or allowing the diffusion of small hydrophobic molecules across the outer membrane (Touw et al., 2010; Hearn et al., 2009).

2.4.3 Performance Evaluation

Although depicting our results in the form of a Receiver Operating Characteristic (ROC) curve is appealing, there are a number of difficulties that prevent us from using an ROC curve. First, the numbers of true positives and true negatives in any of the data set are not known for each of the query sequences. The total number of sequences in the outer membrane database (3,079) is not the same as the number of true positives when we use only the sequences of a small number of known structures as queries. Second, although the data set of shuffled sequences are most likely to be unrelated to the query proteins, one cannot in principle rule out the presence of some sequences that happens to be homologous to the query sequences by random chance. For these reason, the numbers of true negatives are also not known.

2.4.4 Bacterial and Mitochondrial β-Barrel Membrane Proteins

Despite the relatively remote phylogenetic relationship and overall differences, as the proto-mitochondrion probably entered the primitive eukaryotic cell between two and three billion years ago (Wallace, 2005; Tommassen, 2010), our results show that matrices derived from bacterial outer membrane proteins can be used to detect mitochondria outer membrane proteins.

This is consistent with the observation that β-barrel membrane proteins from mitochondria can be readily recognized by the outer membrane insertion machinery of gram negative bacteria (Walther et al., 2010), and bacterial β-barrel membrane proteins can also be recognized and inserted correctly into the outer membrane of mitochondria (Walther et al., 2009a).

Our finding is consistent with a recent hypothesis that no eukaryote-specific signals for the translocation into mitochondria evolved in mitochondrial β-barrel membrane proteins, even though they are now part of eukaryotes. Certain structural elements seems to exist in both mitochondrial and bacterial β-barrel membrane proteins, at least in the TM region, and can be recognized by both insertion machineries (Walther et al., 2009b). The well-conserved pattern of amino acid substitutions seem to be shared between bacteria and mitochondria membrane proteins, as scoring matrices derived from bacterial membrane proteins are very effective in detecting mitochondrial barrel membrane proteins.

2.4.5 Universal Substitution Patterns

The estimated substitution patterns of residues in the TM region of β-barrel membrane proteins are general. In this study, the β-barrel membrane proteins tested in database search for homologs detection are drawn from 19 superfamilies. Despite strong similarity in sequence composition and overall structural similarities, the sequence identity between families is low (<20%). Nevertheless, the scoring matrices can detect remote homologs with excellent specificity and sensitivity. The superfamilies of many of these homologs are not represented by samples from which rates are derived. For example, mitochondria membrane proteins are well detected, which were not used in the estimation of the substitution rates.

Sequences of bacteria and mitochondria are rapidly accumulating from efforts such as metagenomics projects (Liolios et al., 2010). As the chance of the occurrence of false positives increase significantly when a larger number of bacterial genomic sequences are encountered, avoiding incorrect prediction of β-barrel membrane proteins become increasingly important. Existing membrane protein scoring matrices are challenged in this regard. In contrast, the *bbTM* matrices that we developed are well suited for this task, as they have excellent specificity, with no false positives detected in a large scale database search.

2.4.6 Folding Efficiency is Independent of Oligomerization

Previous studies on β-barrel membrane protein folding suggest that OmpF has a lower yield of folded protein than OmpA due to the extra step of oligomerization (Surrey et al., 1996). Our results showed that after eliminating the process of oligomerization by converting OmpF to a monomeric form through amino acid substitutions, the yield of folded protein did not show a significant differences: it changed from 47±6% (12 samples) to 41±4% (6 samples). This suggests that the oligomerization of OmpF does not reduce the folding efficiency of the protein. The minor difference in folding efficiency might be attributed to the amino acid substitutions in the mutant proteins.

2.4.7 Insight into the β-Barrel Folding Process

The highly conserved region around residues Gly57 and Gly59 seems to be important for the oligomerization process of OmpF. Several general features of the glycine residue may explain its role in the process. These are, moderate polarity, the reduced dimension of its side chain, and the flexibility that provides to protein structures (Yan and Sun, 1997).

The moderate polarity explains that Gly is relatively abundant in the lipid protein interface, as we have described above. Given that the most abundant residues at the lipid interface are Ile, Leu, and Val, which are readily substituting between themselves and other residues, it is surprising that Gly was not found to commonly substitute to these residues as well. Instead, Gly substitutes to Ala, which is small and moderately hydrophobic, and Ser, which is small and moderately polar.

The small size of the side chain and the increased flexibility that Gly provides to the protein structure, might be important to allow the interaction of other residues in the oligomerization interface. This region rich in glycine could also facilitate structural changes for the adjustment of different monomers.

When these glycine residues were substituted by long hydrophobic residues, the oligomerization was avoided. This may happen as a consequence of the steric effect of the long side chains, but also because of the stabilization effect of hydrophobic interactions between Ile and Leu with the hydrophobic environment. Our results show that the mutants G57A/G59A and G57S/G59S are able to form trimers. Ala and Ser might still provide enough flexibility to the monomers. In addition, the small size and moderate polarity of both Ala and Ser seem not be sufficient to cause a disruption in the oligomerization process of OmpF.

2.4.8 Relevance of Oligomerization for Porins

The complex and multistep process of biogenesis of porins makes it difficult to fully elucidate the mechanism of oligomerization (Meng et al., 2009). Even though the PPI interfaces of these proteins are known, their importance for the function of porins is largely unexplored.

In this study, we showed that the mutants G57I/G59L can exist in non-preferred oligomeric states (monomer). These constructs can be useful in studying the process and mechanism of oligomerization. It will also be interesting to study the effect of oligomerization on transport from the cytoplasm to the membrane and to assess whether and how different oligomerization states affect the mechanisms of translocation, chaperone interaction, and transporter binding.

2.4.9 Application of Biological Pores in Nano-Biotechnology

A significant number of biological pores are β-barrel membrane proteins (Majd et al., 2010). With strong substrate specificity and multitude of control points, biological nanopores are promising devices for reagentless DNA sequencing, bio-alarm systems, monitoring single-molecule chemical reactions, bio-inspired batteries, and nanotransistors (Majd et al., 2010; Butler et al., 2008; Adiga et al., 2009b; Banerjee et al., 2010). However, these applications have only been explored under controlled laboratory conditions. The limitations of biological nanopores such as lack of stability, non-specific binding, and undesirable oligomeric states are hampering their applications in the uncontrolled environment of the real world, where extreme temperature and denaturing conditions are often encountered. By altering the oligomeric state of β-barrel membrane proteins, we have provided a useful computational addition to the nanobiotechnician's toolbox, with the promise of accelerating efforts for designing novel biological nanopores.

2.5 Conclusions

We have characterized the substitution pattern of residues in the transmembrane segments of β-barrel membrane proteins using a continuous time Markov model of amino acid substitution. We found that residues facing both the lipid environment and the interior of the barrel have characteristic patterns. Despite different evolutionary history for different protein families, their substitution patterns are similar. We also derived scoring matrices from estimated substitution rates. In blind tests including both real β-barrel membrane proteins and random sequences of similar composition as control, our scoring matrices can identify remote homologs with excellent specificity and sensitivity. In addition, we have shown that these scoring matrices can be used to detect mitochondrial outer membrane proteins, suggesting that these two classes of membrane proteins share the same pattern of residue substitution throughout evolution. Our results also imply that the structures of the TM segments of a large number of β-barrel membrane proteins can be predicted reliably based on aligned structural templates. We have also shown that evolutionary analysis of the PPI interface can be used to identify important residues that are required for stable protein-protein interactions in the TM domain. Combining experimental and computational approaches to engineer β-barrel membrane proteins with different oligomerization states and structural properties can facilitate further studies in structural biology of membrane proteins and design of novel nanopores in nanobiotechnology.

CHAPTER 3

ON THE EFFECTS OF PHYSICOCHEMICAL AND FUNCTIONAL CONSTRAINTS ON FUNCTIONAL CAVITIES OF PROTEINS

3.1 Introduction

Enzymes are highly specialized proteins that carry out biochemical reactions with a remarkable speed and specificity. According to the transition state theory (Fersht, 1998), enzymes alter the transition state of a reaction such that the activation energy is significantly decreased, while maintaining an exquisite specificity toward substrates. Although this theory is widely accepted, it has some limitations. For example, it fails for some reactions at high temperatures (Pineda and Schwartz, 2006), it is constrained by classic mechanics (Eyring, 1935), and at times in real reactions some products may appear unexpectedly when applying this theory (Anslyn and Dougherty, 2005). Other theories, which are the subject of broad controversy, attempt to address these limitations. Quantum mechanics tunnelling (Benkovic and Hammes-Schiffer, 2003), low-barrier hydrogen bonds (Neidhart et al., 2001), electrostatic (Warshel et al., 2006), and near-attack conformations (Hur and Bruice, 2003) are the most actively debated. In summary, despite the unquestionable importance of protein enzymes, much is left to learn about them.

The biological function of an enzyme usually takes place in in-foldings (pockets) of the protein called catalytic active sites. Substrate and catalytic side chains are placed together in the catalytic active site. Substrate and side chains form a selective substrate-enzyme complex

in this tiny volume. In this chapter, we investigate the catalytic active site of enzymes combining CASTp and the information of catalytic residues available at the Catalytic Site Atlas database (Porter et al., 2004). CASTp can define and measure the volume of catalytic active sites and determine the number of acid and base side chains in that site (Liang et al., 1998; Dundas et al., 2006). We are motivated by experiments and simulations of ion channels that showed the importance of the acid side chains in calcium channels (Ellinor et al., 1995; Koch et al., 2000; Sather and McCleskey, 2003; Wu et al., 2000; Yang et al., 1993; Boda et al., 2008; Boda et al., 2009; Boda et al., 2010; Gillespie et al., 2009). Models that capture steric exclusion and the special electrostatics of ion channels (and little else) do quite well in describing the selectivity properties of channels and have successfully guided synthesis of artificial selective channels (Miedema et al., 2004; Miedema et al., 2006; Vrouenraets et al., 2006).

Ion channels are specialized proteins with a hole down their middle that allow the movement of specific solutes across otherwise impermeable membranes. Ion channels *catalyze* (Eisenberg, 1990) the selective movement of ions moving through a dielectric barrier (from outside a cell to inside a cell, for example) but they do so without conventional chemistry. The *catalysis* of ion channels does not involve the breaking or making of chemical bonds or the use of chemical energy. The catalytic active sites of ion channel proteins are the selectivity filters of the channel. The selectivity filter distinguishes between ions as the channel protein speeds (*i.e.*, 'catalyzes') their movement across cell membranes (without the hydrolysis of ATP).

Selectivity in three types of membrane proteins comes from charged side chains that face into the pore (Ellinor et al., 1995; Koch et al., 2000; Sather and McCleskey, 2003; Wu et al.,

2000; Yang et al., 1993) and mix with ions in an electrical stew (McCleskey, 2000) in the tiny space of the selectivity filter. *L*-type Ca^{2+} channels ($Ca_V1.n$; n = 1,2, ...) (Boda et al., 2009), voltage activated sodium channels ($Na_V1.n$; n = 1,2, ...) (Boda et al., 2007), and cation selective ryanodine receptors RyRs (Gillespie et al., 2009) can be simulated with success in a wide range of ionic conditions using a model of crowded charges in an implicit solvent (Eisenberg, 2011a). Ion specific properties of bulk electrolytes have been treated in this tradition with some success for a long time (Friedman, 1981; Torrie and Valleau, 1982; Patwardhan and Kumar, 1993; Durand-Vidal et al., 2000; Barthel et al., 1998; Fawcett, 2004; Hansen and McDonald, 2006; Lee, 2008; Kunz, 2009; Li, 2009; Fraenkel, 2010b; Kalyuzhnyi et al., 2010; Vincze et al., 2010; Hnenberger and Reif, 2011),

The idea of catalytic active sites has been important in the history of enzymology (Dixon and Webb, 1979; Kyte, 1995; Segel, 1993) but the idea is not as prominent as it once was, perhaps because the notion of an active site seems rather dim when compared to structures seen in the bright light cast by modern x-ray sources. The phrases 'active site' and 'catalytic active site' are not even in the index of one of the more widely used textbooks of biochemistry (Voet and Voet, 2004).

Here we use the computational power of CASTp to identify and measure all the concavities in enzymes, both pockets and voids. First, we examine these concavities to see if they contain amino acids that participate in the chemical reaction catalyzed by the enzyme. Then, we further examine the concavities that are catalytic active sites to see if they have large densities of acid and base side chains in a small volume, as in calcium and sodium channel proteins.

We find that 573 catalytic active sites of enzymes of known structure and function are easily distinguished by their large numbers of acid and base side chains: Acid and base side chains are reliable markers of catalytic active sites. These enzymes have 4 acid and 5 basic side chains, on the average, in their catalytic active sites. The volume of the catalytic active sites is tiny so the number densities (in chemical units) of acid and base side chains is some 20 molar. In comparison, the number density of solid sodium chloride is 37 molar. The phrase number density is used, as it is in mathematics, to make clear that no assumptions about the properties of the system are made. The number density is simply the number of objects in a volume counted divided by that volume. We fear (and find) that the use of the word 'concentration' causes confusion because 'concentration' is often treated as if it is the (thermodynamic) 'activity', but concentration does not well approximate activity in the ionic solutions found in biology (Eisenberg, 2011; Eisenberg, 2011b).

It seems likely that enzymes use the special properties of such concentrated mixtures of charges to promote catalysis one way or the other, for example, by crowding ions into the special electrostatic environment identified by Warshel (Warshel et al., 2006). We imagine that it will be useful to view catalytic activity of enzymes as a property of an ionic liquid of substrate and (tethered) side chains in the special electrostatic environment of the catalytic active site. Analysis that neglects interactions between ions seems unlikely to be useful, no matter how common in the classical literature of enzymology.

3.2 Methods

3.2.1 Dataset

Catalytic active sites are called that because they contain amino acids known to be directly involved in the catalytic reactions of enzymes. We define the (catalytic) active site pocket as (1) an enclosed space formed in the three-dimensional structure of proteins that also (2) contains the amino acids responsible for the catalytic reaction.

The Catalytic Site Atlas database (CSA) (Porter et al., 2004) annotates a subset of enzymes available in the Protein Data Bank (PDB) (Berman et al., 2002). The database contained 966 entries on June 16, 2011. The CSA classifies side chains using both experimental results and computational predictions. We do not use classifications based on computational predictions. We only use classifications based on experimental results. Redundant sequences are first removed: if a sequence has more than 95% (pairwise) sequence identity to a common sequence (in $> 90\%$ of the length of each sequence), we select just one *at random.* The enzymes are grouped into six main classes according to the chemical reactions they catalyze: EC1, oxidoreductase; EC2, transferases; EC3, hydrolases; EC4, lyases; EC5, Isomerases; EC6, ligases (Tipton, 1994).

3.2.2 Characterization of the Active Site Pocket

Binding sites and catalytic active sites of proteins are often associated with structural pockets and cavities. The CASTp program (Dundas et al., 2006) identifies and measures the pockets and cavities of the experimentally determined structures found in the PDB database (Figure 12). CASTp is based on alpha shape theory of computational geometry. It uses an

analytically exact method to compute the metric properties of voids and pockets on models of macromolecules (Liang et al., 1998).

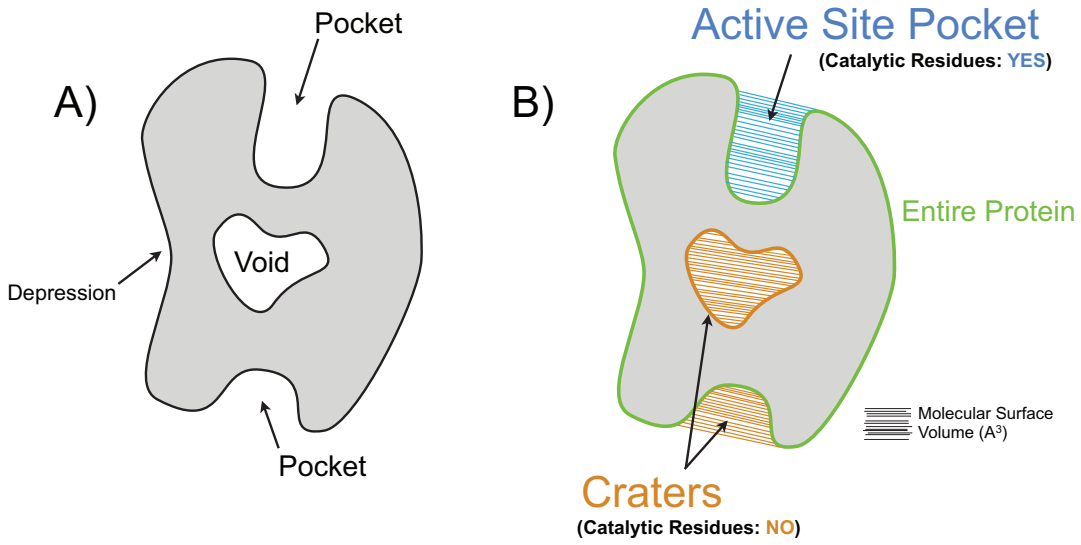

Figure 12. **Sketch of the structural elements calculated and measured in this chapter.** (a). Catalytic active site (left hand panel). The catalytic active site in this example is in a pocket accessible from outside. Most (93 %) of the actives sites in our dataset are accessible. (b) Molecular surface that defines the catalytic active site volume (unit A^3). The volume is reported for both pockets and voids, but not for depressions, where it is difficult to define precisely

We use the Molecular Surface (MS) model (Connolly, 1985) in CASTp to determine metrics (*e.g.*, the volume) of all pockets. The volume of the surface pockets is the measurement of the

space inside the boundary of the pocket that is not occupied by any atom. Details of the pocket geometry calculations can be found in (Edelsbrunner et al., 1998; Liang et al., 1998; Dundas et al., 2006). We use this model to analyze (1) pockets that contain catalytic amino acids, (2) surface pockets (that usually do not contain catalytic amino acids), and (3) interior pockets or voids, whether or not they contain catalytic amino acids. MS represents the protein as a set of intersecting hard spheres (the "atoms"). The outer boundary molecular surface is obtained by tracing the distal edge of a spherical ball that is rolled around the protein molecule (see Figure 12b). This surface is supposed to characterize a spherical solvent molecule rolling around an irregular protein if both were macroscopic uncharged objects. No one knows how to sample the space around an irregular protein the way a solvent or solute molecule actually samples that space in a protein in an ionic solution. Such sampling is needed if the free energy of solvent or solute is to be simulated precisely enough to calculate biological selectivities (Kokubo and Pettitt, 2007; Kokubo et al., 2007; Zhang et al., 2010; Eisenberg, 2010). We use CASTp with MS to measure the catalytic active site because together they provide computer based objective estimates. These estimates are significantly more reproducible than those that require more human judgement.

We are mostly interested in catalytic active sites but first we discuss structural features of the enzyme that do not participate in the catalytic reaction of the enzyme. We call them **craters**. Craters are pockets (1) that do not contain atoms of a catalytic side chain, and also (2) have a volume between 100 and 3,000 $Å^3$. Some craters contain protein-ligand complexes

that do not participate in the substrate chemical reaction. Some do not. The range of volumes for protein-ligand complexes was 100 to 1,694 Å^3 (Saranya and Selvaraj, 2009).

The enzymes surveyed here contain, on average, 48 pockets or voids, most of them with a tiny size (less than 100 Å^3). Of these, 53% are pockets, i.e., accessible from the outside, and 47% are voids (i.e., non-accessible pockets). Craters as we define them are entirely distinct from catalytic active sites. The function of craters is not known despite our speculations later in this chapter.

We define the Active Site Pocket (ASP) as the pocket with a volume between 100 and 3,000 Å^3 that contains the largest number of atoms of the catalytic side chains. The range of minimum and maximum volume of the substrates was from 81 to 768 Å^3. The properties of active sites located in either depressions or convex surfaces are not considered here because volume cannot be measured reliably in those cases.

Figure 13 shows the distribution of the volume of catalytic active sites. Three-quarters of the selected PDB dataset (573 of 759) has active site pockets, as we define ASPs. Figure 14 shows the distribution of amino acids in the entire enzyme; the distribution of amino acids in the active site pockets; and the distribution of amino acids in the catalytic side chains of the selected proteins. The results are consistent with an earlier study (Porter et al., 2004; Dundas et al., 2006; Gutteridge and Thornton, 2005).

One difficulty in measuring the size of surface pockets on proteins is determining the boundary that separates the pocket from the outside solution. In this study, we used the convex hull of the atoms of a protein to define the boundary of the surface pockets. This choice gives an

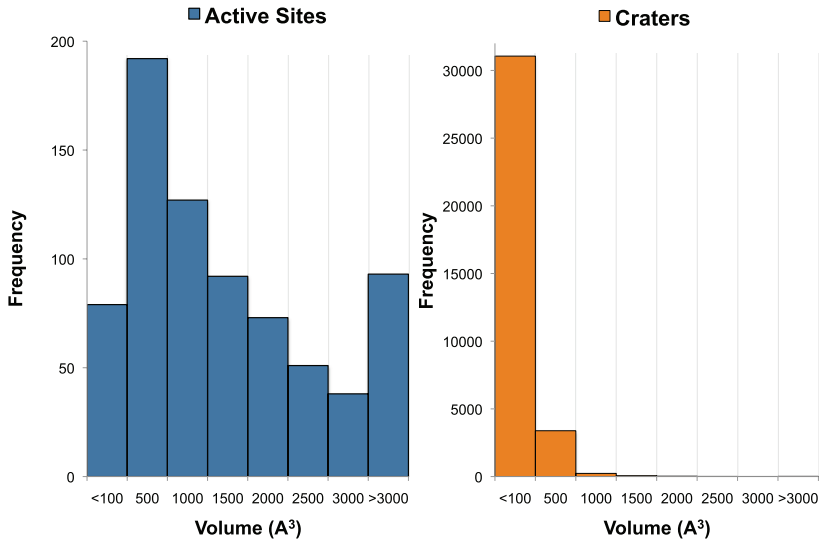

Figure 13. **Volume distribution of the active site.** Histogram of the distribution of the volume of the active site pocket for a set of 759 enzyme structures (unit: A^3). Pockets with volumes between 100 and 3,000 A^3 were used in the determination of the number of acid and base side chains and the calculation of the density of charge

unambiguous measurement, although other definitions may also be possible (Edelsbrunner et al., 1998; Liang et al., 1998). Another difficulty in measuring the size of surface pockets is the significant change of measured volume that is produced by even a small change of the shape of the surface pocket. Here, pocket volume was calculated based on the experimentally determined structure in the conditions in which the structure was measured. The effects of substrate and ion concentrations on structure and active site volume cannot yet be dealt with quantitatively because of the lack of crystallographic data (Otyepka et al., 2007). The experimentally determined structure is a snapshot of the ensemble of conformations a protein adopts, but overall conclusions are well determined estimators of protein properties because they are based

on statistics of the volume measurement gathered from a large number of protein structures. Surface area can be used in our analysis instead of volume without significantly changing our conclusions (data not shown).

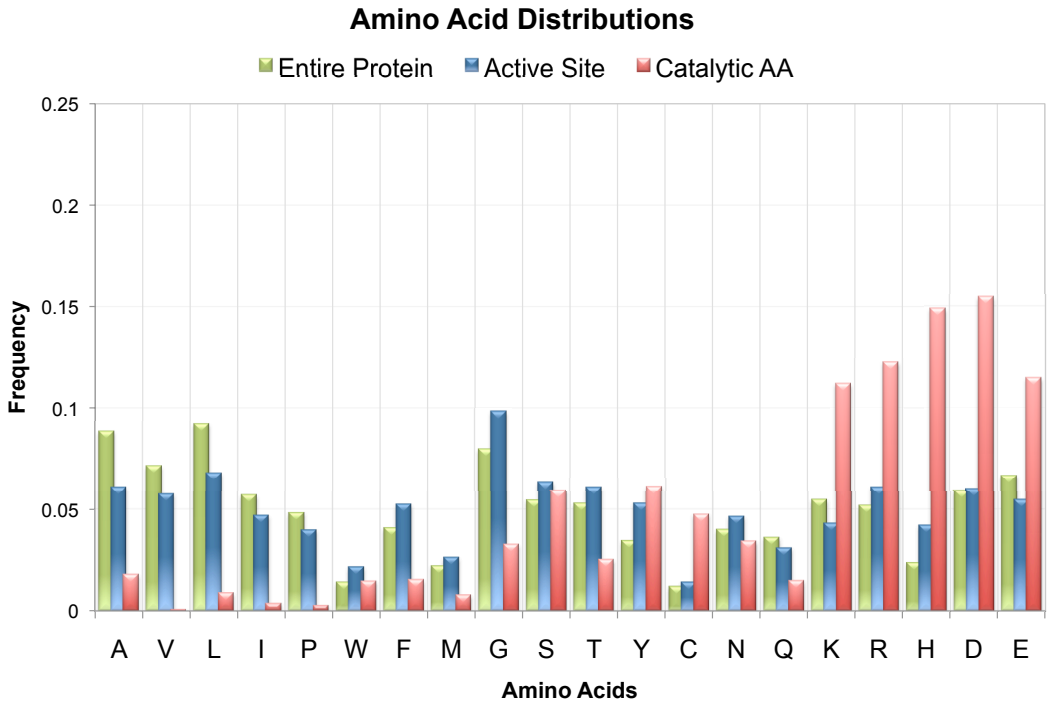

Figure 14. **Amino acid frequencies in our dataset.** Amino acid composition in our dataset for the entire protein (green), all the amino acids in the active site pocket (blue) and only the catalytic amino acids (red). The distribution of amino acids in the entire protein and the catalytic active site are not very different. There is a significant increase of polar charged and uncharged side chains for catalytic residues

3.2.3 Charge Densities (CharDen)

The density of charge (of acid and base side chains) is the key variable that determines a biological function (selectivity) of ion channels (Eisenberg, 2011a), so we are interested in measuring that variable in the catalytic active site of enzymes of known structure.

The calculation of the charge density (CharDen) requires counting the number of acid (negative) and base (positive) side chains and calculation of the volume they surround and occupy. In our calculation, surface exposed atoms are those with non-zero exposed surface area, whether or not the exposed atoms are side-chain atoms or backbone atoms. When counting the number of ionizable residues in an active site, only those with the side chain pointing towards the pocket are considered. We simplify our language by using the word negative to describe the charge of acid side chains and the word positive to describe the charge of basic side chains. Charge density refers to the number density of either acid or basic side chains, or both. We are quite aware that ionization state of the side chains is not known in most cases and is sensitive to the local (and even global) environment.

Number density (objects/m^3) is given in units of molar concentration for easier chemical intuition, e.g., comparison with ionic liquids and solids. No assumptions concerning the activity or activity coefficient are implied. That is why we use the phrase number density. We fear (and find) that "concentration" and activity are often confused, with serious effects when dealing with the ionic mixtures of biology (Eisenberg, 2011; Eisenberg, 2011b). Obviously, acid and base side chains at these number densities are not ideal non-interacting particles with activity coefficients of one. Indeed, it is not clear even how to define the activity coefficient of an

ion in systems this concentrated (Hnenberger and Reif, 2011). The protein creates a special electrostatic environment (Warshel et al., 2006). It also creates a fluid substrate that is more like an ionic liquid than an ideal solution. Theories and simulations that assume ideal properties of reactants or force fields in these conditions are unlikely to be helpful.

3.2.4 Computational Method for Active Site Prediction

We tested whether charge density can be used to predict the active site of proteins. Every active site and crater was defined by a vector with a selected number of features. We used a supervised learning method for the classification of every cavity into either active site or crater. Specifically, we used the library LIBSVM of support vector machine (Chang and Lin, 2011). To prevent over-fitting, we divided each data set in 4/5 for cross-validation (svm-train) and 1/5 for blind test (svm-predict).

3.3 Results

We analyzed the distribution of the different amino acids according to the enzymatic activities of the catalytic active sites (Tipton, 1994) as described in Table VII, noting that some amino acids may be included in two or more classes.

Hydrophobic side chains (*i.e.*, Ala, Val, Leu, Ile) are found less frequently in active site pockets than in the entire protein. Aromatic side chains (Trp, Tyr, Phe), small polar side chains (Ser, Thr), and particularly Glycine, are more common (Figure 14).

The distribution of amino acids responsible for the catalytic reaction is striking. It is very different from the (distribution of the) overall composition of the active site pocket as well as the (distribution of the) amino acids in the whole protein (Figure 14 and Figure 15), as previously reported by (Porter et al., 2004; Gutteridge and Thornton, 2005). The catalytic side chains of transferases (EC2), lyases (EC4) and isomerases (EC5) have a similar distribution of base and acid side chains. However, hydrolases (EC3) have a larger fraction of acid side chains (D and E) and also a larger fraction of histidine. Ligases (EC6) have a larger fraction of base side chains (K, R and H). The evolutionary or chemical reasons for this specialization are not known.

3.3.1 Volumes of Catalytic Active Sites

The mean volume of the catalytic active site pocket is 1072 $Å^3$. The average sequence length of proteins in our dataset is 338 amino acids, with a standard error of the mean of 6.3. The average number of side chains that are part of the active site pocket is 34 $\pm$ 0.77 (n = 573) (mean $\pm$ Standard Error of the Mean).

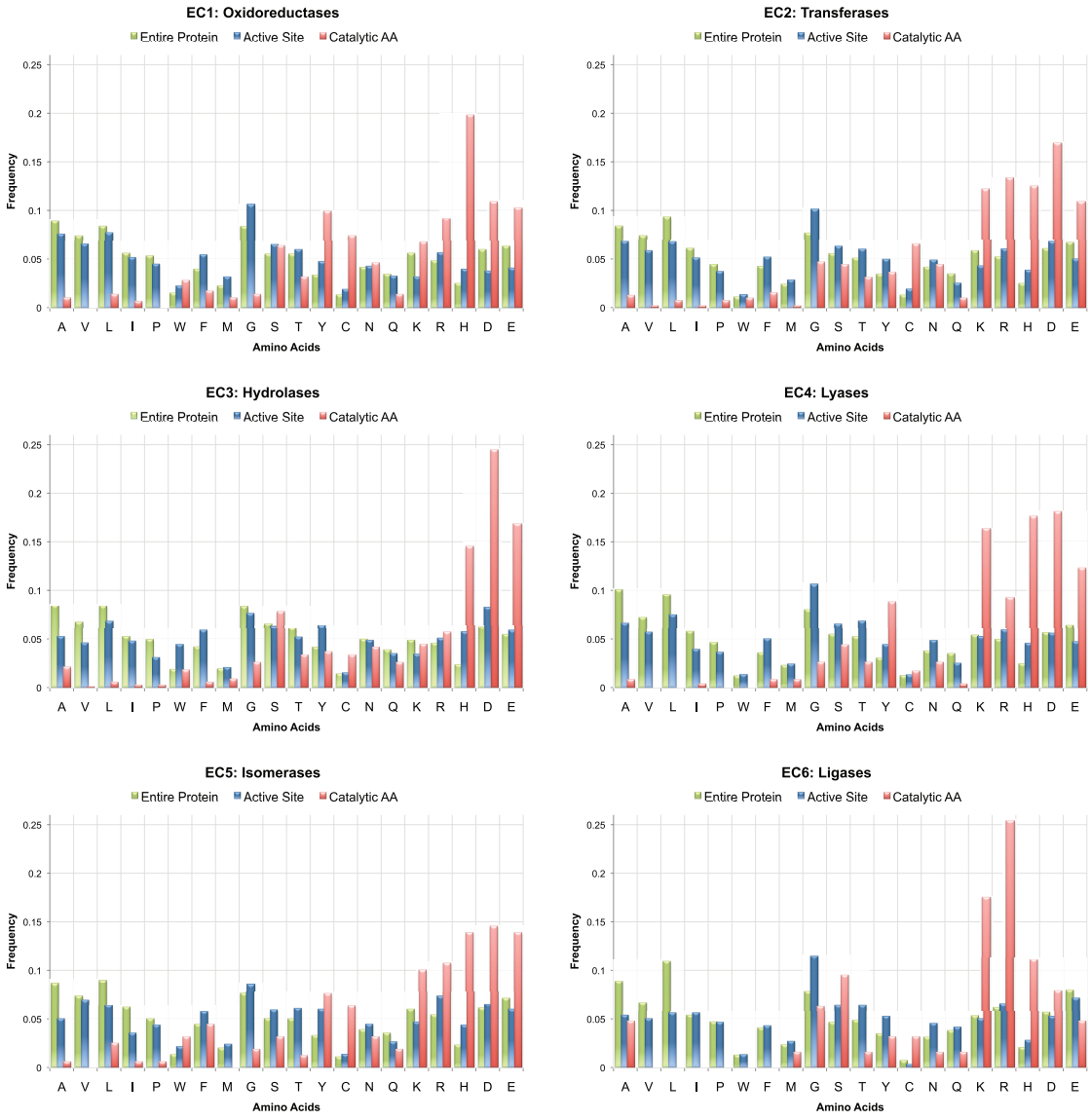

Figure 15. **Amino acid composition grouped by enzymes (EC1-EC6).** All the amino acids in the entire protein, in the catalytic active site pockets and only the catalytic amino acids

TABLE VII

CLUSTER OF AMINO ACIDS.

Group	Amino Acids
Hydrophobic (non-polar, uncharged)	Alanine, Leucine, Isoleucine, Methionine, Phenylalanine, Tryptophan, Tyrosine and Valine
Polar (uncharged)	Serine, Threonine, Asparagine and Glutamine.
Aromatic	Tryptophan, Phenylalanine and Tyrosine.
Basic (positively charged)	Lysine, Arginine and Histidine.
Acidic (negatively charged)	Aspartic and Glutamic acid.
Special cases	Cysteine, Proline and Glycine.

Group of amino acids according to the chemical properties of their side chain.

Different classes of enzymes have somewhat different characteristics. The largest catalytic active sites are found in oxidoreductases (1568 $Å^3$), ligases (1233 $Å^3$) and transferases (1206 $Å^3$). Hydrolases (786 $Å^3$) and isomerases (863 $Å^3$) have the smallest pocket volume. Oxidoreductases (EC1) have the longest sequence length (average 379.7 $\pm$ 18.9, n = 99), and the largest number of amino acids in the catalytic active site (on average 47.6 $\pm$ 2.05, n = 99). Isomerases have the shortest sequence length (296.4 $\pm$ 23.3, n = 43) and isomerases the lowest number of amino acids (29.33 $\pm$ 2.60, n=43).

3.3.2 Charge Densities at Catalytic Active Sites

We calculated various densities for each pocket (Table VIII), assuming for the purposes of exposition that all acid and base side chains are ionized. We calculated (1) the density of positive charges, (2) the density of negative charges, (3) the density of the absolute value of

charges, namely the total density of acid and base side chains. The mean density for the whole dataset of 573 enzymes is 18.9 ± 0.58 M. The distribution of the total density of charge of catalytic active sites (CharDen) is shown in Figure 16. Isomerases (22.1 M) and hydrolases (22.8 M) have the largest CharDen values. Oxidoreductases have the smallest (12.1 M). For 93% of the proteins in our data set, catalytic active sites have clear connections to the outside through mouth-opening(s) of the pocket(s). These openings are large enough to allow the access of water molecule(s). For the remaining proteins (7%) in our data set, active sites are found to be in voids buried and non-accessible according to our definitions. Since these enzymes do in fact catalyze reactions involving substrates outside the protein, it is likely that the structure of the protein fluctuates to allow substrate and ligand access, as seen in cytochrome P450 (Otyepka et al., 2007; Ludemann et al., 2000; Cojocaru et al., 2011).

3.3.3 Protein Charge Density

We also computed the density of charge of the entire protein. This calculation used the volume of the entire protein (Edelsbrunner et al., 1995). The charge density for the entire protein in our dataset (global charge density) is on average 2.82 M ± 0.03 (n = 573), which is a small dispersion (Figure 16). The value 2.8 M was smaller than we expected considering that 25% of the side chains in proteins are charged.

We find that the positive charge density is always larger than the negative charge density, but for some classes of enzymes the surplus of negative charge is smaller (hydrolases, ligases) than for others (oxidoreductases or lyases).

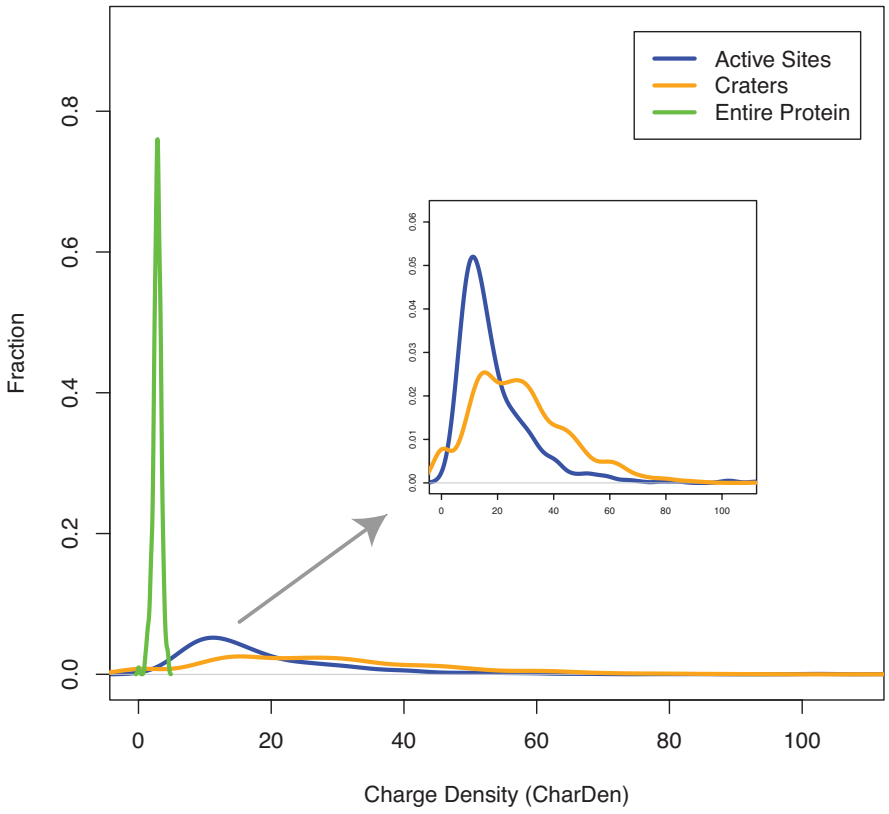

Figure 16. **Density estimation of the CharDen.** Density estimation of the fraction of proteins with a given charge density (CharDen). Catalytic active site, craters and the entire protein CharDen

TABLE VIII

CHARDEN VALUES.

		Catalytic Active Site			**Craters**		
		CD+	CD-	CDt	CD+	CD-	CDt
EC1	Oxidoreductases (n = 98)	7.5	4.6	12.1	16.8	12.1	28.9
EC2	Transferases (n = 126)	9.5	7.2	16.6	16.4	12.5	28.8
EC3	Hydrolases (n = 214)	12.1	10.7	22.8	15.2	11.9	27.1
EC4	Lyases (n = 72)	11.2	7.3	18.5	16.6	11.6	27.8
EC5	Isomerases (n = 43)	12.6	9.5	22.1	16.2	13.5	29.7
EC6	Ligases (n = 20)	9.7	8.3	18	16.2	11.9	28
	Total (n = 573)	10.6	8.3	18.9	16	12.1	28.2

Summary of charge density (CharDen, unit molar) at the catalytic active site, craters, and the entire protein. CD+: Molar positive CharDen; CD−: Molar negative CharDen; CDt: Total (positive + negative) molar CharDen

3.3.4 Craters

Craters (as we have defined them above) have a smaller size (262.2 $Å^3$, 13 amino acids per crater) than catalytic active sites (1072 $Å^3$, 34 amino acids for catalytic active sites) and are mostly (83.5%) accessible from the outside. The volume of craters is largest among ligases and smallest in isomerases. The distribution of the volume of craters is quite different from the distribution of the volume of catalytic active sites (two sample Kolmogorov-Smirnov (K-S) test, p-value = 2.2 x 10^{-16}, Figure 17). The number of amino acids in craters of different types of enzymes ranges from 12.8 (Transferases) to 13.3 (Oxidoreductases and Lygases).

The charge density in craters (28.2 ± 0.34 M) is larger than in catalytic active sites (where it is 18.9 ± 0.58 M: Table VIII). Values in craters are different among the different groups

of enzymes. They vary from 27.1 M (hydrolases) to 29.7 M (isomerases). The distribution of charge density in craters is very different from the distribution in the catalytic active sites (Figure 16; K-S test, p-value = 2.7 x 10^{-15}).

3.3.5 Charge Density on Active Site Prediction

We used a machine learning method to test whether the active site can be predicted based on charge density. We first defined all the cavities, *i.e.,* active sites and craters, by three features: positive CharDen, negative CharDen, and total CharDen. As a result, the classification method obtained an accuracy of 80.7% (525/651). We then included as features for every cavity the pocket volume, amino acid size, total number of acidic residues, total number of basic residues, positive CharDen, negative CharDen, total CharDen, protein CharDen, and number of mouth openings. The accuracy increased to 92.5% (602/651).

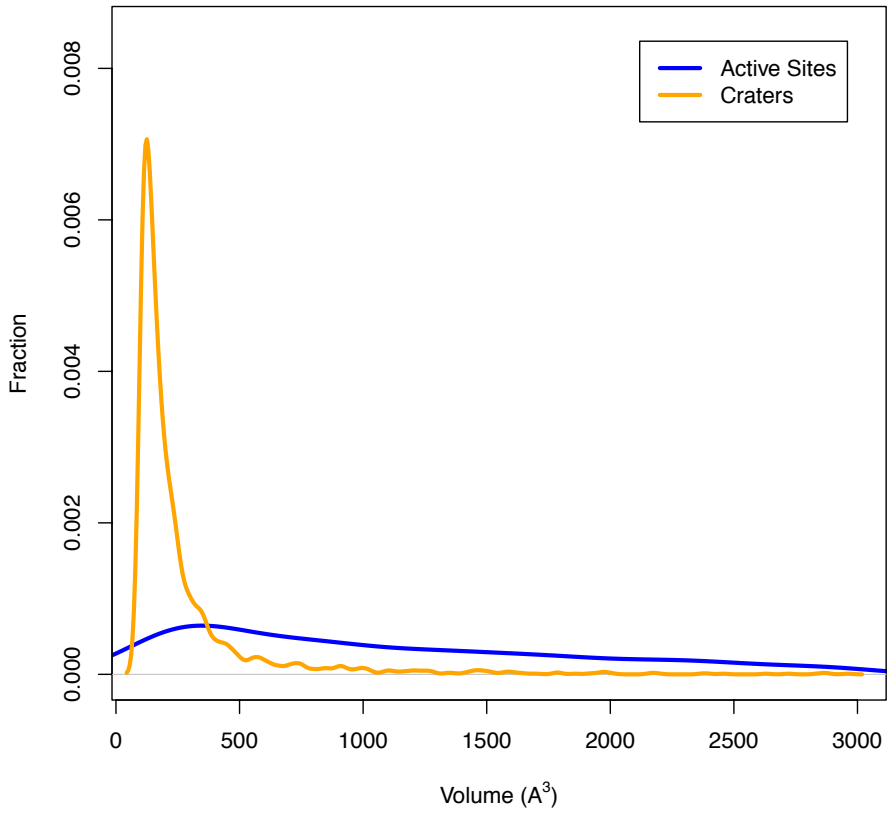

Figure 17. **Density estimation of the volume (A^3) of catalytic active sites and craters.** The charge density in craters (28.2 ± 0.34 M) is larger than in catalytic active sites (where it is 18.9 ± 0.58 M: Table VIII). Values in craters are different among the different groups of enzymes. They vary from 27.1 M (hydrolases) to 29.7 M (isomerases). The distribution of charge density in craters is very different (Fig. 5; K-S test, p-value = 2.7 x 10^{-15}) from the distribution in the catalytic active sites. We do not know why.

3.4 Discussion

A great deal of attention has been paid to the chemical role of acid and base side chains in the catalytic active sites of enzymes, and to the special electrostatic environment of enzymes (Warshel et al., 2006) and channels (Eisenberg, 1996b; Eisenberg, 1996a), but less attention has been paid to the steric effects of excluded volume. Those effects can be substantial when charge densities are high and crowding results. The steric repulsion of finite size ions produces chemical specificity in bulk solution (Friedman, 1981; Torrie and Valleau, 1982; Patwardhan and Kumar, 1993; Durand-Vidal et al., 2000; Barthel et al., 1998; Fawcett, 2004; Hansen and McDonald, 2006; Lee, 2008; Kunz, 2009; Li, 2009; Fraenkel, 2010b; Kalyuzhnyi et al., 2010; Vincze et al., 2010; Hnenberger and Reif, 2011), and some ion channels (Boda et al., 2008; Boda et al., 2009; Boda et al., 2010; Gillespie et al., 2009; Eisenberg, 2011a).

It seems likely to us that the charge densities in catalytic active sites create a special physical environment optimized in some unknown way to help enzymes do their work. The tiny volume of the catalytic active site ensures that even a few acid side chains produce a large density of electric charge. The forces that produce (approximate) electroneutrality ensure that a nearly equal amount of counter charge is near the acid or basic side chains, within a few Debye or Bjerrum lengths. One component of the enzymatic specialization is the electrostatic environment analyzed in detail by Warshel in enzymes (Warshel et al., 2006) and (in significantly less detail) by Eisenberg in channels (Eisenberg, 1996a; Eisenberg, 1996b). In channels, another component of the specialization is the steric effect of crowded charge. In (some types of)

channels, it is the balance between electrostatics and crowding that produce the selectivity that defines channel types.

It seems useful to speculate that enzymes balance electrostatic and steric forces the way some channels do. After all, channels are nearly enzymes (Eisenberg, 1990). The tiny volume surrounding the side chains and counter ions guarantees severe crowding and steric repulsion. In these crowded catalytic active sites, reactants and side chains mix in an environment without much water, very different from the water dominated ionic solutions outside of proteins. The environment does not resemble the infinitely dilute ideal fluid for which the law of mass action is appropriate (Eisenberg, 2011b). The catalytic active site seems more like an ionic liquid (Kornyshev, 2007; Siegler et al., 2010; Spohr and Patey, 2010) than an ideal gas. The ionic liquid of the catalytic active site differs from classical ionic liquids because some of its components are side chains of proteins, 'tethered' to a polypeptide backbone, not free to move into the bulk solution. These charged side chains may have as large a role in the function of proteins (Eisenberg, 1996a; Eisenberg, 1996b) as doping has in transistors (Markowich et al., 1990; Howe and Sodini, 1997; Pierret, 1996; Sze, 1981), although the finite diameter of the side chains adds a strong flavour of chemical selectivity and competition not found in semiconductors (Eisenberg, 2005; Eisenberg, 2012).

3.4.1 Craters

Our main focus has been on the catalytic active sites, but we also found pockets that do not contain catalytic amino acids. We call them "craters". Proteins in our data set contain 4.5 pockets per protein that are large enough for us to analyze (*i.e.*, are larger than 100 $Å^3$ and

are not located in either depressions or convex surfaces). These craters do not contain catalytic residues and are thus not catalytic active sites. Some craters are known to be binding sites for effectors (activators or repressors), *i.e.*, small molecules that change the biological activity of the protein. Craters near the outer surface of a protein are likely to be important in protein-protein interactions because they contain large amounts of permanent (*i.e.*, 'fixed') charge.

Craters seem to us to be atomic-scale ion exchangers, *i.e.*, charged reservoirs of mechanical energy. Ion exchangers are Donnan systems that generate substantial internal osmotic and hydrostatic pressure (Helfferich, 1962; Nonner et al., 2001). The osmotic pressure in craters creates strong mechanical forces in the enzyme. When those forces are unleashed, so they can cause motion, the structure of the enzyme is likely to change, on atomic and also on macroscopic scales. These structural changes can be conformational fluctuations affecting the chemical steps of enzyme catalysis (Bhabha et al., 2011). The osmotic pressure of craters might be one of the forces that drives the conformational changes of enzyme function.

3.4.2 Charge in the Catalytic Active Site: Amount and Role

The large densities of acid and base side chains reported here do not automatically imply a large density of charge. The ionization state of most of these side chains is not known. Direct measurements are needed in our view. Calculations are not reliable given the difficulties in designing force fields and calibrating simulations in the special ionic environment of the catalytic active site, so different from bulk solution. Ionization would, of course, differ from enzyme to enzyme and mutant to mutant. Ionization is expected to depend on the concentrations of reactants and ions near the binding site, as well as in the surrounding baths. Similar charge

interactions were considered long ago (p. 457-463 of (Edsall and Wyman, 1958); p. 117-127 of (Cohen and Edsall, 1943). The special importance of the electrostatic environment was brought to the attention of modern workers by Warshel (Warshel and Russell, 1984), who particularly has emphasized its importance in the active site (Warshel et al., 2006).

Salt bridges are likely to reduce the net charge of catalytic active sites because the negative charge of one acid side chain balances the positive charge of a basic side chain. Specifically, 73% of the catalytic active sites contain at least one acid side chain within 4 Å of a basic side chain (44% of craters).

The leftover charge, not balanced in salt bridges, is still likely to be large. The unbalanced density of side chain charge is still likely to be enough to create densities of ions far beyond those found in bulk solutions. These unbalanced charges are an important source of the special electrostatic environment in active sites (Warshel et al., 2006) we believe.

Large densities of charge obviously have a profound effect on protonation steps of found in many chemical reactions catalyzed by enzymes. Large densities of charge are likely to have other effects beyond shifts in protonation states. The protein creates a charged surface that fits the substrates as a glove fits a hand. Indeed this is a special electrostatic environment.

This special electrostatic structure will have large effects on any step in a chemical reaction that produces changes in charge, or is influenced by the electric field (consider dielectrophoresis (Pohl, 1978)). In addition to these effects, it is possible that the large densities of charge produce special physical constraints on orbitals of electrons in the molecules close to the protein. The permanent (*i.e.*, 'fixed') charge of the protein must enforce a nearly Neumann boundary

condition for the Poisson part of the Schrödinger equation that defines the molecular orbitals of nearby (substrate) electrons.

Whatever the role of the large charge densities in catalysis, their presence produces interactions not present in the law of mass action (Eisenberg, 2011b) used universally in models of enzyme kinetics (Dixon and Webb, 1979; Segel, 1993), with rate constants independent of concentration. That law of mass action is appropriate for an infinitely dilute ideal gas, not for the concentrated solutions (nearly an ionic liquid) in an catalytic active site. 'Everything' interacts with everything else in those conditions. The free energy (that drives a chemical reaction) then depends on the concentrations of all species, not just the concentrations of reactants and products (Pytkowicz, 1979; Hovarth, 1985; Zemaitis et al., 1986; Pitzer, 1995; Barthel et al., 1998; Durand-Vidal et al., 2000; Fawcett, 2004; Kontogeorgis and Folas, 2009; Fraenkel, 2010a). In addition, the flow of reactants is coupled to the concentration (and perhaps flow) of all other species near the catalytic active site. 'Everything' interacts with everything else in the crowded confines of the catalytic active site. Indeed, the singular single file behavior seen in some types of ion channels is an extreme example of nonideal behavior. Ions in such systems clearly do not behave as if they are infinitely dilute with activities independent of other ions. It seems wiser to use mathematics designed to handle interactions in complex fluids (Hyon et al., 2010; Eisenberg, 2010; Liu, 2009; Sheng et al., 2008; Doi, 2009) than mathematics designed to handle infinitely dilute uncharged ideal gases.

Coupling between ions is known to be an inevitable product of nonideal properties of ions in solutions (Pytkowicz, 1979; Hovarth, 1985; Zemaitis et al., 1986; Pitzer, 1995; Barthel et

al., 1998; Durand-Vidal et al., 2000; Fawcett, 2004; Kontogeorgis and Folas, 2009; Fraenkel, 2010a). Ion-ion interactions have not had a prominent role in models of channels, transporters, or enzyme function (Tosteson, 1989; Dixon and Webb, 1979; Segel, 1993). The coupled flows of ions that define transporters (and are characteristic of enzymes) have usually been ascribed entirely to the ion-protein interaction. Perhaps some flows are coupled because of interactions of ions among themselves in the crowded nonideal environments near, if not in the catalytic active sites.

3.4.3 Predictive Power of Charge Density

Our results showed that charge density can be used to predict the active site of proteins with high accuracy. We obtained a 81% accuracy in the identification of the active site by using exclusively positive, negative, and total CharDen. This is a clear indication that the charge density is a defining feature of active site.

3.5 Conclusion

The catalytic active site of enzymes have been defined using a computational geometry program and a database of catalytic residues. These active sites have large numbers of acid and base side chains. The volume of catalytic active sites is small. The number density of acid and base side chains is very high. The active site of proteins can be predicted based on charge density exclusively. The contents of catalytic acid sites do not resemble the infinitely dilute solutions used in classical enzyme kinetics or force fields of modern molecular dynamics. The balance of steric and electrostatic forces in the highly concentrated environment of the catalytic active site is likely to be an evolutionary adaptation that has an important role in enzymatic catalysis. It seems wise to use mathematics designed to handle interactions in complex fluids when studying the catalytic active site of enzymes.

CHAPTER 4

ON THE EFFECTS OF FUNCTIONAL AND PHYSICOCHEMICAL CONSTRAINTS ON FUNCTIONAL RESIDUES SUBJECTED TO SPONTANEOUS POST-TRANSLATIONAL CHEMICAL MODIFICATIONS

4.1 Introduction

The human proteome is more complex than the human genome. Post-translational modifications (PTMs) increase the complexity and functional diversity of proteins beyond that of the genome. According to some estimations, PTMs may enlarge the protein variants by two to three orders of magnitude (Walsh et al., 2005; Eisenhaber and Eisenhaber, 2010). PTMs modulate many important aspects of proteins, including protein life, subcellular location, functional state, protein stability, protein-protein interactions, and protein-DNA interactions (Walsh, 2006; Mann and Jensen, 2003; Jensen, 2004). There are more than 300 PTMs contributing to enlarge the repertoire of chemical functional groups beyond the 20 basic amino acids (Jensen, 2004). Identifying and understanding their biological roles is important for studying cancer, diabetes, metabolic syndromes, and neurological disorders (Jensen, 2004).

Lysine is one of the 15 amino acids known to be subjected to multiple PTMs (Walsh, 2006; Eisenhaber and Eisenhaber, 2010). Methylation, for instance, is a well-known PTM affecting Lys, with a major role in epigenetic inheritance (Greer and Shi, 2012). Another PTM

of Lys, albeit less well-known, is carboxylation. The addition of a carboxyl group to the ε-amino group of the Lys side chain can have remarkable consequences. As a positively ionizable residue, Lys acquires one positive charge at physiological pH. However, the addition of the carboxyl group drastically changes its properties by acquiring negative charge. This modification occurs spontaneously under basic pH conditions involving carbon dioxide in solution (Park and Hausinger, 1995), without the mediation of any other enzyme.

There are two important functions currently known to be associated with this chemical modification. First, Lys carboxylation can be involved in the catalytic function of enzymes by playing a direct role in the catalytic reaction (Golemi et al., 2001; Dementin et al., 2001) or, as catalytic or cocatalytic determinant, by bridging metal ions (Stec, 2012; Meulenbroek et al., 2009). Second, Lys carboxylation may also be involved in the rearrangement of the active site to create the appropriate orientation of different side chains (Lorimer et al., 1976; Wu et al., 2008).

The extent of Lys carboxylation is currently unknown. Unlike other covalent modifications affecting the Lys amino acid (Li et al., 2010), carboxylation has not been fully investigated, partly due to the difficulties associated with the highly unstable nature of this labile PTM. The carboxyl group is spontaneously released in acidic conditions. In addition, mass spectrometry, an experimental procedure commonly used to study PTM by measuring the molecular mass of modified proteins, cannot detect Lys carboxylation (Golemi et al., 2001). The chemical instability results in an incorrect perception of the scope of this PTM.

X-ray crystallography can identify carboxylated Lys residues, but is not exempt of problems. First, the crystallization of a protein with a labile chemical modification is challenging. Second, it can be difficult to distinguish carboxylation from several other post-translational modifications based on the measured electron density map alone. Finally, the most frequent damage produced by third generation synchrotron radiation is the decarboxylation of acidic residues (Ravelli and McSweeney, 2000; Garman, 2010; Borek et al., 2010), which may alter the state of this PTM as well.

Computational methods can be valuable in overcoming these difficulties affecting the detection of spontaneous PTMs, such as Lys carboxylation. However, there are no computational tools currently available for the prediction of Lys carboxylation. In this work, we describe a computational method for this task. Using currently available protein structures with carboxylated Lys residues, our method can detect carboxylated Lys residues with a sensitivity and a specificity of 87% and 99.9%, respectively. It also predicts carboxylation in Lys residues on proteins with and without similarity to proteins where lysine carboxylation is known to occur. We find that some Lys residues annotated as carboxylated are truly non-carboxylated in the functional state of the protein. Finally, we assess the extent of Lys carboxylation in the protein structure database and discuss its implications.

4.2 Methods

4.2.1 Methods Summary

Data sets for training and testing were collected from the Protein Data Bank (Berman et al., 2002). We examined the frequency of amino acids, water molecules, and metal ions on both KCX and LYS sites from protein structures. These frequencies are used by PreLysCar (Predictor of Lysine Carboxylation) to predict KCX sites, which is based on a naïve Bayesian model. For a given Lys residue and the frequency of amino acids in its microenvironment, posterior probabilities of being carboxylated and uncarboxylated are calculated. Lys is classified as carboxylated or uncarboxylated according to the largest posterior probability value. A detailed measure and evaluation of performance was carried out.

4.2.2 Datasets

4.2.2.1 Kcx+ and Lys− Datasets

Any classification method requires, at least, two well-defined classes to be predicted. In the present work, one class consists of Lys residues that are carboxylated (denoted as "KCX" in the Protein Data Bank), and the other Lys residues that are not carboxylated (LYS). To obtain both sets, we downloaded 251 protein structures available (May 2012) at the Protein Data Bank (PDB) (Berman et al., 2002) with at least one subunit containing a carboxylated Lys residue. For each protein structure, only the chain(s) solved with the carboxylated lysine residue (KCX) was selected and performed redundancy reduction by (1) clustering the subunits sharing more than 90% sequence identity using BLASTClust (Altschul et al., 1990b; Wheeler and Bhagwat, 2007) and (2) selecting from each cluster the chain with the best resolution. A

total of 65 structures met this criteria, with an average resolution of 1.99 ± 0.5 Å. As it is explained below, 3 out of 65 were treated as non-carboxylated and further removed, resulting in a final set of 62 protein structures.

KCX residues (62) were included in the data set denoted as KCX+. The remaining Lys residues (21 Lys per protein on average) were included in the data set denoted as LYS−. Since the number of buried Lys residues is limited in proteins and most of the KCX residues were found to be buried, we added to the LYS+ data set a equivalent number of buried Lys residues (62) obtained randomly from the set of high resolution protein structures described below. In total, 1,337 Lys residues were included in the LYS− data set.

4.2.2.1.1 Main Functions of KCX Proteins

About 80% of proteins with a known KCX residue in our data set belong to eubacteria, 6% to eukaryotes, and 3% to archaea. The largest group of enzymes with known carboxylated Lys residues are hydrolases (EC:3). Among them, class D β-lactamases are involved in the resistance to beta-lactam antibiotics (Poirel et al., 2010). Ureases are another class of hydrolases with KCX sites found in many bacterial pathogens, and are important for their clinical detection (Mobley et al., 1995). Hydrolases aryldialkylphosphatases (LeJeune et al., 1998; Seibert and Raushel, 2005a) are of great interest due to their potential use in detoxification of chemical waste and warfare agents (LeJeune et al., 1998).

Among the ligases (EC:6), the UDP-N-acetylmuramoyl-L-alanineD-glutamate ligase is an enzyme involved in the synthesis of the bacterial peptidoglycan (Barreteau et al., 2008). Pyruvate carboxylase is a multifunctional enzyme catalyzing the biotin-dependent production of

oxaloacetate with important roles in gluconeogenesis, lipogenesis, insulin secretion, and other cellular processes (St Maurice et al., 2007).

Among transferases (EC:2), the 5S metalloenzyme chain of the transcarboxylase multienzyme (Hall et al., 2004) requires a carboxylated Lys to coordinate a cobalt ion in the active site. Another transferase with a KCX residue is the N-acetyl-L-ornithine transcarbamoylase (AOTCase). This is an essential enzyme for arginine biosynthesis in several eubacteria (Shi et al., 2006).

The most remarkable lyase (EC:4) known to require Lys carboxylation is the ribulose bisphosphate carboxylase/oxygenase (rubisco), enzyme responsible for creating organic carbon from the inorganic carbon dioxide of the air (Berg et al., 2002).

Alanine racemase is an isomerase (EC:5) that requires a KCX residue to stabilize the active site. This enzyme catalyzes a step involved in bacterial cell wall biosynthesis by the interconversion of alanine enantiomers (Strych et al., 2000). It is also an attractive target for antimicrobial drug development because of the lack of any homologs in eukaryotes (Silverman, 1988).

4.2.2.2 High-Resolution Protein Structures

We used PDBselect (Griep and Hobohm, 2010) to obtain a subset of high-resolution protein structures solved by X-ray crystallography (< 1.5 Å). Each protein subunit was treated independently. Those subunits sharing more than 90% sequence identity were clustered using CD-HIT (Li and Godzik, 2006). The structure with the best resolution was selected from each cluster. Only proteins larger than 200 amino acids were selected, resulting in a final set of 577 structures.

Lys residues located at the surface and buried in the interior of proteins were identified using CASTp, which uses weighted Delaunay triangulation and alpha shape theory to measure the surface accessible residues (Dundas et al., 2006).

4.2.2.3 Redundancy Reduction of the PDB Database

To further explore the incidence of Lys carboxylation on the Protein Data Bank, we first reduced the number of similar proteins. We combined PDBselect (Griep and Hobohm, 2010) and CD-HIT (Li and Godzik, 2006) as described above. As a result, we obtained 14,324 protein structures solved by X-ray crystallography with a resolution equal or larger than 1.5 Å and more than 200 amino acids. A total of 291,434 Lys residues were identified.

4.2.3 Sequence Motif

We investigated whether a sequence motif could be identified at the KCX site. Amino acid subsequences around carboxylated Lys residues (20 amino acids in each direction) were first extracted. We followed two different approaches. First, we used MEME to search for motifs using default parameters (Bailey et al., 2009). No significant motif was found. Second, we measure the sequence conservation at every position by calculating the sequence entropy and create a sequence logo using WebLogo (Crooks et al., 2004). A lack of sequence conservation in the amino acids surrounding the KCX site was also observed (Figure 18).

4.2.4 Measurements in the Microenvironment of KCX and LYS Sites

Every amino acid from both KCX+ and LYS− is defined by the frequency values of the amino acid side chains, water molecules, and metal ions found within 5 Å.

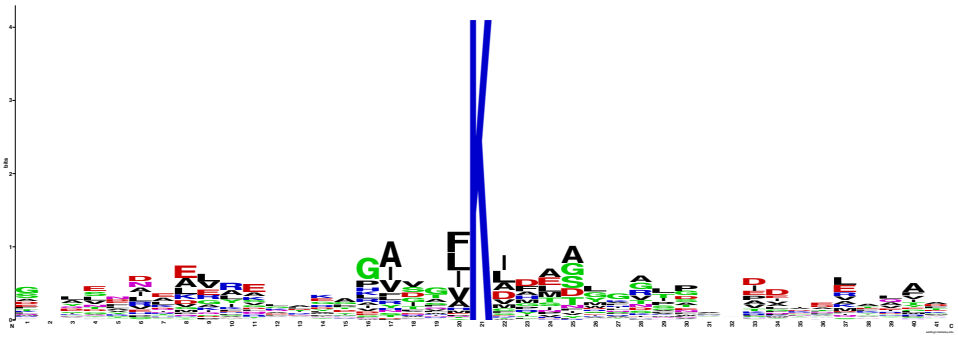

Figure 18. **Sequence conservation at Kcx sites.** A fragment of 40 amino acids was extracted with the KCX residue in the center for each protein sequence. The resulting multiple sequence alignment was used to create a sequence logo (Crooks et al., 2004).

Briefly, euclidean distances were measured between the side chain atoms from the KCX+ and LYS− sets and the remaining amino acids, water molecules, and ions of their respective structures. Any of those components located within a distance of 5 Å was counted in. It is important to notice that the length of the Kcx side chain is larger than Lys side chain due to the presence of the carboxyl group. A bias could arise as a consequence of this extra length, helping the classification method to successfully differentiate between KCX and LYS sites. However, the predictor could potentially fail in identifying carboxylated Lys residues on other structures and, consequently, its real usability compromise. In order to treat both data sets equally, we removed the carboxyl group from KCX residues. Instead, we projected an atom (named pCX of "projected CX") on the tip of the NZ atom of both KCX and LYS side chains, taking as a direction the average direction of all the atoms of the side chain where the pCX is

projected. The pCX atom was also used to measure distances to all the other components of the microenvironment.

Amino acids of the microenvironment were grouped according to physicochemical properties of side chains, *i.e.*, negatively charged (NEG: Asp,Glu), positively charged (POS: Arg,His,Lys), small polar (ST: Ser, Thr), large polar (NQ: Asn, Gln), aromatic (ARO:Trp, Tyr, Phe), and hydrophobic (HYD: Met, Ile, Leu, Val). The metal ions considered were ZN, MG, CO, FE2, FE, NI, and MN. Overall frequencies for the microenvironment of both KCX and LYS sites were finally calculated (Table IX and Table X).

4.2.5 The Bayesian Predictor

There are two main parameters that need to be calculated by our Bayesian method. First, the probability distribution of the features, which we approximate with relative frequencies from the training set ($F_1, F_2, ..., F_n$, see Table IX and Table X). The other is the prior probability, which we arbitrarily selected as a best reasonable guest of the frequency of Lys carboxylation. For any given Lys residue, both the posterior probability of being carboxylated ($p(C_{KCX}|F_1, ..., F_n)$) and the posterior probability of not being carboxylated ($p(C_{LYS}|F_1, ..., F_n)$) are calculated as follow:

$$p(C_{KCX}|F_1, ..., F_n) = \frac{p(C_{KCX})p(F_1, ..., F_n|C_{KCX})}{p(F_1, ..., F_n|C_{KCX}) + p(F_1, ..., F_n|C_{LYS})};$$

and

TABLE IX

FREQUENCIES OF AMINO ACIDS AT KCX SITES

KCX									
BIN	**LEN**	**NEG**	**POS**	**ST**	**NQ**	**ARO**	**HYD**	**ION**	**WAT**
0	-	0.597	0.016	0.145	0.758	0.113	-	0.371	0.274
1	-	0.323	0.194	0.419	0.226	0.403	0.016	0.113	0.371
2	-	0.081	0.161	0.242	0.016	0.258	0.081	0.516	0.226
3	-	-	0.419	0.065	-	0.21	0.258	-	0.081
4	-	-	0.177	0.097	-	0.016	0.274	-	0.032
5	-	-	0.032	0.032	-	-	0.371	-	0.016
6	-	-	-	-	-	-	-	-	-
7	-	-	-	-	-	-	-	-	-
8	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-	-
11	0.032	-	-	-	-	-	-	-	-
12	0.258	-	-	-	-	-	-	-	-
13	0.419	-	-	-	-	-	-	-	-
14	0.194	-	-	-	-	-	-	-	-
15	0.048	-	-	-	-	-	-	-	-
16	0.016	-	-	-	-	-	-	-	-
17	-	-	-	-	-	-	-	-	-
18	-	-	-	-	-	-	-	-	-
19	-	-	-	-	-	-	-	-	-

TABLE X

FREQUENCIES OF AMINO ACIDS AT LYS SITES

LYS									
BIN	LEN	NEG	POSHIS	ST	NQ	ARO	HYD	ION	WAT
0	0	0.218	0.58	0.536	0.58	0.493	0.215	0.991	0.239
1	0.002	0.421	0.313	0.31	0.306	0.344	0.309	0.007	0.156
2	0.028	0.257	0.086	0.102	0.091	0.121	0.224	0.001	0.13
3	0.07	0.086	0.016	0.036	0.022	0.031	0.146	-	0.137
4	0.125	0.016	0.003	0.013	-	0.01	0.07	-	0.095
5	0.162	0.001	0.001	0.002	-	0.001	0.036	-	0.243
6	0.144	-	-	0.001	-	-	0.036	-	-
7	0.109	-	-	-	-	-	0.036	-	-
8	0.105	-	-	-	-	-	0.001	-	-
9	0.064	-	-	-	-	-	-	-	-
10	0.048	-	-	-	-	-	-	-	-
11	0.049	-	-	-	-	-	-	-	-
12	0.037	-	-	-	-	-	-	-	-
13	0.025	-	-	-	-	-	-	-	-
14	0.017	-	-	-	-	-	-	-	-
15	0.01	-	-	-	-	-	-	-	-
16	0.003	-	-	-	-	-	-	-	-
17	0.001	-	-	-	-	-	-	-	-

$$p(C_{LYS}|F_1,...,F_n) = \frac{p(C_{LYS})p(F_1,...,F_n|C_{LYS})}{p(F_1,...,F_n|C_{KCX}) + p(F_1,...,F_n|C_{LYS})}$$

where $p(C_{KCX})$ and $p(C_{LYS})$ are the prior probabilities. The likelihood is calculated as follows:

$$p(F_1,...,F_n|C_{KCX}) = p(F_1|C_{KCX}) \cdot p(F_2|C_{KCX})...p(F_n|C_{KCX})$$

and

$$p(F_1,...,F_n|C_{LYS}) = p(F_1|C_{LYS}) \cdot p(F_2|C_{LYS})...p(F_n|C_{LYS})$$

with $p(F_n|C_{KCX})$ being the probability of the feature n when Lys is carboxylated, and $p(F_n|C_{LYS})$ when Lys is not carboxylated.

The predictor classifies any given Lys according to the largest posterior probability value, *i.e.*, a Lys vector $L_x = (x_1, x_2, ..., x_n)$ will be classified as KCX if $p(C_{KCX}|x_1, x_2, ..., x_n) > p(C_{LYS}|x_1, x_2, ..., x_n)$, and non-carboxylated (LYS) otherwise. We consider the difference between both probability values as a measure of confident.

We developed a tool implementing this method. We named it PreLysCar (Predictor of Lysine Carboxylation).

4.2.6 Measures of Performance

We assessed the performance of the Bayesian classifier by the technique of leave-one-out cross-validation. Briefly, a testing vector is held out from the Kcx+ and Lys+ data sets. The remaining vectors are used for training, from which KCX and LYS frequencies are obtained.

Posterior probabilities for the vector held out are next calculated, both $p(C_{KCX}|F_1, ..., F_n)$, *i.e.*, probability of Lys residue of being carboxylated given the composition of its microenvironment, and $p(C_{LYS}|F_1, ..., F_n)$, *i.e.*, probability of Lys residue of being non-carboxylated given the composition of its microenvironment. The predictor assigns the class "KCX" or "LYS" according to the largest posterior probability value. This operation is repeated for every vector in the entire data set. As a result, the four different possible outcomes of a binary prediction method are obtained, *i.e.*, total number of "True Positives" (TP, KCX correctly predicted as KCX), "True Negatives" (TN, LYS correctly predicted as LYS), "False Positives" (FP, LYS incorrectly predicted as KCX) and "False Negatives" (FN, KCX incorrectly predicted as LYS). Finally, common statistical measures of performance are calculated, *i.e.*, sensitivity (SEN), specificity (SPE), accuracy (ACC), positive predictive value (PPV), negative predictive value (NPV), and Matthews Correlation Coefficient (MCC).

$SEN = \frac{TP}{TP+FN}$;

$SPE = \frac{TN}{TN+FP}$;

$ACC = \frac{TP+TN}{TP+TN+FP+FN}$;

$PPV = \frac{TP}{TP+FP}$;

$NPV = \frac{TN}{TN+FN}$;

$MCC = \frac{((TP*TN)-(FP*FN))}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$;

4.2.7 Electron Density Maps and Remodelling

Electron density maps 2Fo-F and Fo-F were downloaded from the Electron Density Server (Kleywegt et al., 2004). Remodelling of Lys residues predicted as KCX were performed with COOT (Emsley and Cowtan, 2004) and refined with Refmac (Vagin et al., 2004).

4.3 Results

4.3.1 Proteins with Known Carboxylated Lys Residues

We first constructed a data set of protein structures with Lys residues resolved as carboxylated. We identified a total of 251 protein structures in the Protein Data Bank (Berman et al., 2002) with at least one subunit containing a carboxylated Lys residue (denoted as "KCX" in the PDB database). To reduce the redundancy of multiple structures for the same protein, we selected the structure with the highest resolution from those sharing more than 90% sequence identity. A final set of 62 proteins was selected, with an average resolution of 1.99 Å. The 90% sequence identity cut-off was set to ensure that enough redundancy was removed (from 251 to 62), and at the same time to maintain a sufficient amount of data. For example, from 24 structures of class D-β-lactamases available in the Protein Data Bank, only 6 structures were included in our data set. All of them are different oxacillinases-type-β-lactamases (OXAs) (Poirel et al., 2010), *i.e.* OXA-1, OXA-2, OXA-10, OXA-24, OXA-46, and OXA-48 (4.3.1)

From this set of KCX proteins, we constructed a data set consisting of 62 KCX sites, one per protein (denoted as KCX+). In most cases, there is biological understanding of the role of the KCX residue in the protein function. Details of each enzyme, its EC number, and the role of the KCX residue can be found in 4.3.1. Major roles of the KCX residue were found to be catalytic and structural, all of them directly or indirectly part of the active site of the protein.

Since no proteins with more than one carboxylated Lys residue in the same chain are known to exist, we selected the remaining Lys residues from the same 62 proteins as control. A total

TABLE XI

PROTEINS WITH KNOWN CARBOXYLATED LYS RESIDUES INCLUDED IN OUR DATA SET

Group	**EC Number**	**Protein Name**	**PDB_chain**	**KCX role**	**Specie**
Transferase	2.7.11.22	Human Cyclin Dependent Kinase 2	1H01_A	Not provided	*Homo sapiens*
Transferase	2.1.3.9	Transcarbamylase	3KZN_A	Not provided	*Xanthomonas campestris*
Transferase	2.1.3.1	Transcarboxylase 5S Subunit	1RQB_A	Metal Ion Center	*Propionibacterium shermanii*
Hydrolase	—	3-Methyladenine DNA Glycosylase	1PU6_B	Hydrogen bonding?	*Helicobacter pylori*
Hydrolase	3.5.4.2	Adenine Deaminase	2ICS_A	Ion Metal Center	*Enterococcus faecalis*
Hydrolase	3.5.2.5	Allantoinase	3E74_B	Metal Ion Center	*Escherichia coli*
Hydrolase	—	Amidohydrolase	3OVG_A	Metal Ion Center	*Mycoplasma synoviae*
Hydrolase	3.1.8.1	Aryldialkylphosphatase	2VC7_B	Metal Ion Center	*Sulfolobus solfataricus*
Hydrolase	3.1.8.1	Aryldialkylphosphatase	2OB3_B	Metal Ion Center	*Brevundimonas diminuta*
Hydrolase	—	Beta-Lactamase Regulatory Protein	3Q7V_B	Switch	*Staphylococcus aureus*
Hydrolase	—	D-Hydantoinase	1K1D_C	Metal Ion Center	*G. stearothermophilus*
Hydrolase	3.5.2.2	D-Hydantoinase	1NFG_C	Metal Ion Center	*Burkholderia pickettii*
Hydrolase	3.5.2.2	D-Hydantoinase	2FVK_C	Metal Ion Center	*Saccharomyces kluyveri*
Hydrolase	3.5.2.2	D-Hydantoinase	3DC8_B	Metal Ion Center	*Sinorhizobium meliloti*
Hydrolase	3.5.2.2	D-Hydantoinase	1GKR_D	Metal Ion Center	*Arthobacter aurescens*
Hydrolase	3.5.2.2	D-Hydantoinase	1GKP_C	Metal Ion Center	*Thermus sp*
Hydrolase	3.5.2.2	D-Hydantoinase	2FTW_A	Metal Ion Center	*Dictyostelium discoideum*
Hydrolase	—	Dihydroorotase	2OGJ_F	Metal Ion Center	*Agrobacterium fabrum*
Hydrolase	3.5.2.3	Dihydroorotase	3PNU_B	Metal Ion Center	*Campylobacter jejuni*
Hydrolase	3.5.2.3	Dihydroorotase	3JZE_D	Metal Ion Center	*Salmonella enterica*
Hydrolase	3.5.2.3	Dihydroorotase	2Z26_B	Metal Ion Center	*Escherichia coli*
Hydrolase	3.4.19.-	Isoaspartyl Dipeptidase	1ONW_A	Metal Ion Center	*Escherichia coli*
Hydrolase	—	L-Lys, L-Arg Carboxypeptidase	3MTW_A	Metal Ion Center	*Caulobacter crescentus*
Hydrolase	—	Lactonase Lmo2620	3PNZ_A	Metal Ion Center	*Listeria monocytogenes*
Hydrolase	—	Metal-Dependent Hydrolase	3ICJ_A	Metal Ion Center	*Pyrococcus furiosus*
Hydrolase	—	Organophosphorus hydrolase	3GTX_A	Metal Ion Center	*Deinococcus radiodurans*
Hydrolase	3.5.2.6	Oxa-1 Class D Beta-Lactamase	3ISG_B	Catalytic	*Escherichia coli*
Hydrolase	3.5.2.6	Oxa-10 Class D Beta-Lactamase	2X02_B	Catalytic	*Pseudomonas aeruginosa*
Hydrolase	3.5.2.6	Oxa-2 Class D Beta-Lactamase	1K38_B	Catalytic	*Salmonella typhimurium*
Hydrolase	3.5.2.6	Oxa-24 Class D Beta-Lactamase	3G4P_A	Catalytic	*Acinetobacter baumannii*
Hydrolase	3.5.2.6	Oxa-46 Beta-Lactamase	3IF6_C	Catalytic	*Pseudomonas aeruginosa*
Hydrolase	3.5.2.6	Oxa-48 Class D Beta-Lactamase	3HBR_A	Catalytic	*Klebsiella pneumoniae*

Kcx proteins. Continue from previous page

Group	EC Number	Protein Name	PDB_chain	KCX role	Specie
Hydrolase	3.5.-.-	Phosphotriesterase, Lactonase	3OJG_A	Metal Ion Center	*Geobacillus kaustophilus*
Hydrolase	—	Putative Amidohydrolase	3MKV_H	Metal Ion Center	Unidentified
Hydrolase	—	Putative Amidohydrolase	3N2C_P	Metal Ion Center	Unidentified
Hydrolase	—	Sgx9359b	3DUG_G	Metal Ion Center	Environmental sample
Hydrolase	3.5.1.5	Urease	1EJX_C	Metal Ion Center	*Klebsiella aerogenes*
Hydrolase	3.5.1.5	Urease	3QGK_O	Metal Ion Center	*Helicobacter mustelae*
Hydrolase	3.5.1.5	Urease	1E9Z_B	Metal Ion Center	*Helicobacter pylori*
Hydrolase	3.5.1.5	Urease	4UBP_C	Metal Ion Center	*Bacillus pasteurii*
Hydrolase	3.5.1.5	Urease	3LA4_A	Metal Ion Center	*Canavalia ensiformis*
Hydrolase	—	Yp_805737.1	2QPX_A	Metal Ion Center	*Lactobacillus casei*
Lyase	4.1.1.39	Rubisco	1BWV_C	Metal Ion Center	*Galdieria partita*
Lyase	4.1.1.39	Rubisco	1WDD_A	Metal Ion Center	*Oryza sativa*
Lyase	4.1.1.39	Rubisco	3KDN_H	Metal Ion Center	*Thermococcus kodakarensis*
Lyase	—	Rubisco-Like Protein	3NWR_A	Metal Ion Center?	*Burkholderia fungorum*
Lyase	—	Uv Damage Endonuclease	2J6V_A	Unknown	*Thermus thermophilus*
Isomerase	5.1.1.1	Alanine Racemase	1VFS_A	Hydrogen bonding	*Streptomyces lavendulae*
Isomerase	5.1.1.1	Alanine Racemase	3S46_A	Hydrogen bonding	*Streptococcus pneumoniae*
Isomerase	5.1.1.1	Alanine Racemase	1XQL_A	Hydrogen bonding	*Geobacillus stearothermophilus*
Isomerase	5.1.1.1	Alanine Racemase	1RCQ_A	Hydrogen bonding	*Pseudomonas aeruginosa*
Isomerase	5.1.1.1	Alanine Racemase	2ODO_C	Hydrogen bonding	*Pseudomonas fluorescens*
Isomerase	5.1.1.1	Alanine Racemase	2RJH_C	Hydrogen bonding	*Escherichia coli*
Isomerase	5.3.2.-	Rubisco-like protein, enolase	2OEM_B	Metal Ion Center	*Geobacillus kaustophilus*
Ligase	6.3.2.12	Folc Bifunctional Protein	1W78_A	Water coordination	*Escherichia coli*
Ligase	6.3.2.17	Folylpolyglutamate Synthase	2GC5_A	Metal Ion Center	*Lactobacillus casei*
Ligase	6.3.2.9	MurD inhibitor	2X5O_A	Metal Ion Center	*Escherichia coli*
Ligase	6.3.2.13	Mure Ligase	2XJA_A	Metal Ion Center	*Mycobacterium tuberculosis*
Ligase	6.4.1.1	Pyruvate Carboxylase	2QF7_A	Metal Ion Center	*Rhizobium etli*
Ligase	6.4.1.1	Pyruvate Carboxylase Protein	3BG3_C	Metal Ion Center	*Homo sapiens*
Ligase	6.3.2.13	UDP-N-acetylmuramoyl-l-alanine	1E8C_B	Metal Ion Center	*Escherichia coli*
Unknown	—	Putative Dihydroorotase	2GWN_A	Metal Ion Center	*Caulobacter vibrioides*

of 1,337 LYS sites were included in the data set of uncarboxylated Lys residues (denoted as LYS−).

4.3.2 Signature Microenvironment of KCX sites

We next examined the microenvironment of both KCX and LYS sites by determining the frequencies of various amino acids, metal ions, and water molecules found within 5 Å from the KCX and LYS side chains (Figure 19, Table IX and Table X). We did not include the carboxyl group of the KCX residue. Instead, we projected an atom (named pCX of "projected CarboXyl group") onto the tip of the NZ atom of both KCX and LYS side chains, taking as a direction the average direction of the remaining atoms of the side chain. The pCX atom was also used to measure distances to all the other components of the microenvironment. Considering that most of the noncovalent bonding interactions occur in less than 4 Å (Hibbert and Emsley, 1991; Harris and Mildvan, 1999; Kumar and Nussinov, 2002), and since we are not adding oxygen atoms to the pCX, we used 5 Å as a distance cut-off to ensure that enough spatially proximal atoms are found in both KCX and LYS microenvironments.

4.3.2.1 Compactness

A defining feature of KCX sites is the large number of atoms of residues, water molecules, and ions found in the proximity of the KCX side chain. On average, 12.9 ± 1.1 of these are found within 5 Å of the KCX side chain. Amino acids part of KCX sites are extended along the primary sequence of the protein, *i.e.*, from the first to the last position in the sequence, amino acids part of the KCX site occupy large fragments of the protein. We observed that many of these fragments tend to be located towards the protein N-terminus (Figure 20).

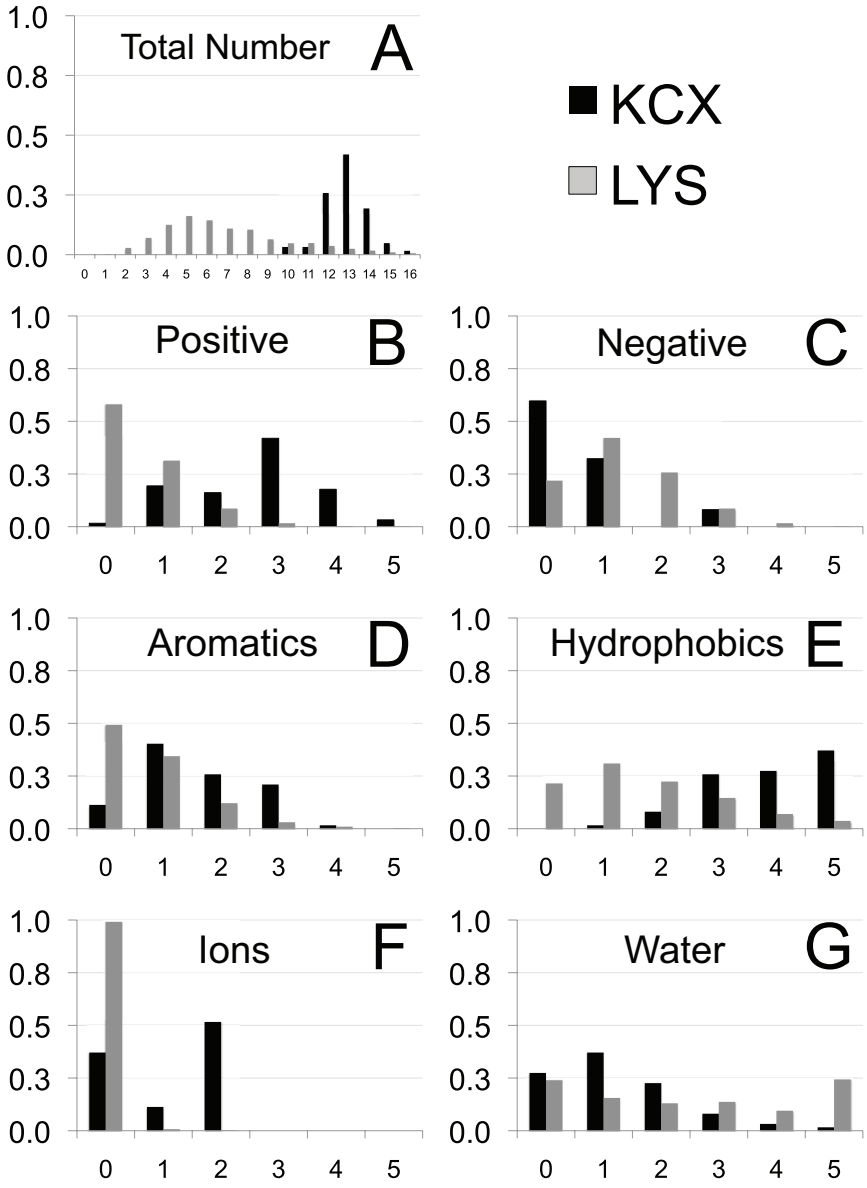

Figure 19. **Physicochemical microenvironment of KCX and LYS sites.** Frequency of amino acids, metal ions, and water molecules found within 5 Å from the side chain of KCX (black bars) and LYS (grey bars) residues. The amino acids found in the microenvironments are grouped according to their main physicochemical properties, *i.e.,* positively ionizable (R, K, H), negatively ionizable (D, E), aromatic (W, F, Y), and hydrophobic (I, L, V, M) residues. The x-axis represents the number of amino acids around the KCX and LYS sites, while the y-axis represents the frequencies of each count.

The number of residues surrounding LYS sites was found to be much reduced (6.9 ± 3.0) (Figure 19A). In fact, only 18% of LYS sites showed more than 9, in contrast to 100% of KCX sites. This difference is in part due to the buried nature of the carboxylated Lys residue. All KCX residues were found buried and not accessible from the surface. In contrast, only 11% of the LYS residues are buried.

As consequence of the buried nature of KCX sites, two other significant differences between LYS and KCX sites emerged. First, hydrophobic residues (Met, Ile, Leu, Val) were more likely to be found in the microenvironment of KCX sites (Figure 19E). The number of hydrophobic residues found within 5 Å was 4.1 ± 1.4, which is larger than that for LYS sites (1.7 ± 1.4). Second, although water molecules are not uniformly resolved in X-ray structures, in general, we found that the amount of water molecules also varies between both data sets, with at least 3 or more water molecules found in 46% of the LYS sites, but only in 13% of KCX sites. The number of water molecules can aid in the identification of Lys residues that are at the surface in contact with the solvent (Figure 19G).

4.3.2.2 Aromatic Residues

Aromatic residues (Phe, Trp, and Tyr) were found in 89% and 51% of KCX and LYS sites, respectively (Figure 19D). Aromatic residues are present in 73% of the microenvironment of buried LYS sites. The 16% increased of hydrophobic residues in the KCX site with respect to other buried LYS sites suggests that aromatic residues may be important for the stabilization of the KCX site.

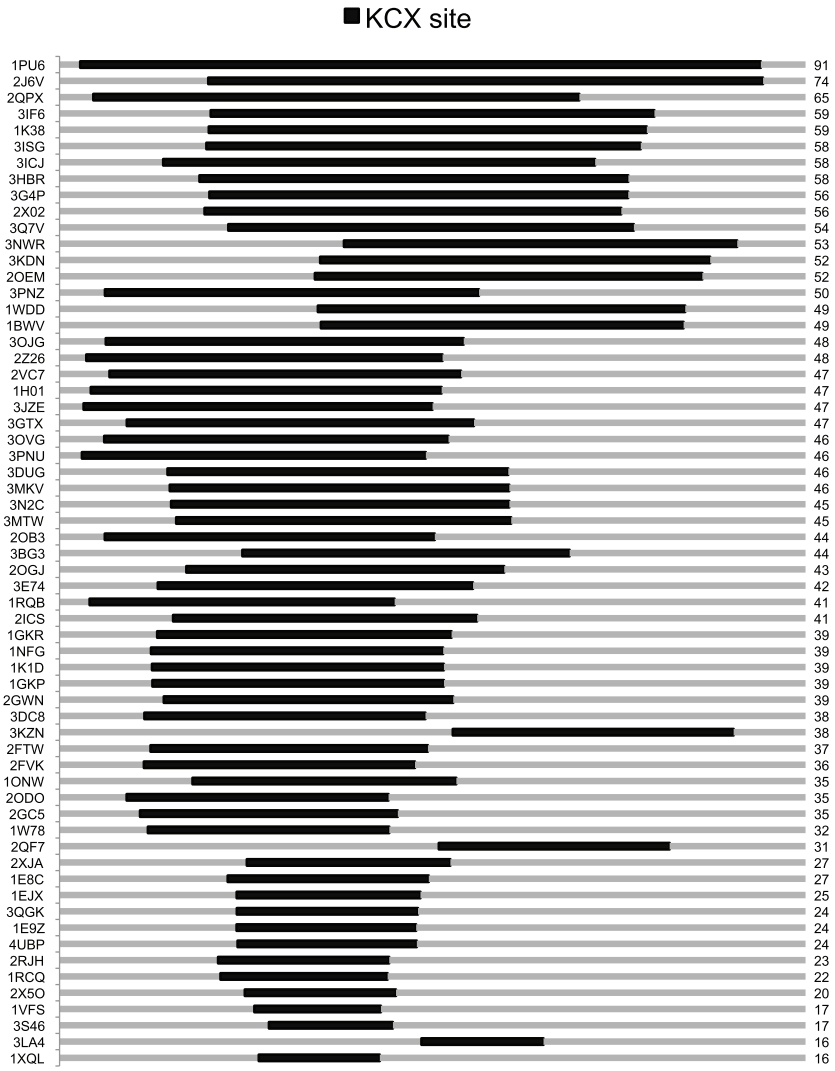

Figure 20. **Dispersion of amino acids part of the KCX site on protein sequences.** Grey bars represent protein sequences, which are scaled for comparison. Black portions represent the fragment from the first to the last amino acid found in every KCX site. Axis show the PDB identification (PDB-ID) of proteins and the percentage of extension that KCX sites occupy on the protein sequences.

4.3.2.3 Ionizable Residues

Lys is a positively ionizable residue. In proteins, positively charged Lys residues are expected to be balanced by negative charge. The enrichment of negatively charged residues (Asp and Glu) near LYS sites confirms this expectation. About 78% of the LYS sites are found to have one or more negatively charged amino acids. In contrast, positively ionizable residues (Arg, Lys, and His) are less likely to be found in the microenvironment of the LYS site, with only 40% of them presenting positively ionizable residues (Figure 19B-C).

Thus, the overall distribution of ionizable residues in the microenvironment of KCX sites is significantly different from that of LYS sites. After carboxylation, the modified Lys residue acquires negative charge. This drastic change in the electrostatic properties of the modified Lys residue is reflected in the composition of ionizable residues found in its microenvironment. About 95% of the ionizable residues found within 5 Å of the KCX side chain were positively charged, in contrast to the 40% of the KCX sites with one or more negatively charged residues (Figure 19B-C).

4.3.2.4 Polar Residues

Significant differences are also observed in the distribution of polar residues in KCX and LYS sites. Long polar residues (Asn and Gln) are absent in 76% and 59% of the KCX and LYS sites, respectively. However, small polar residues (Ser and Thr) are likely to be found within 5 Å of the KCX side chain. About 85% of them showed at least one small polar residue, compared to 46% of LYS sites.

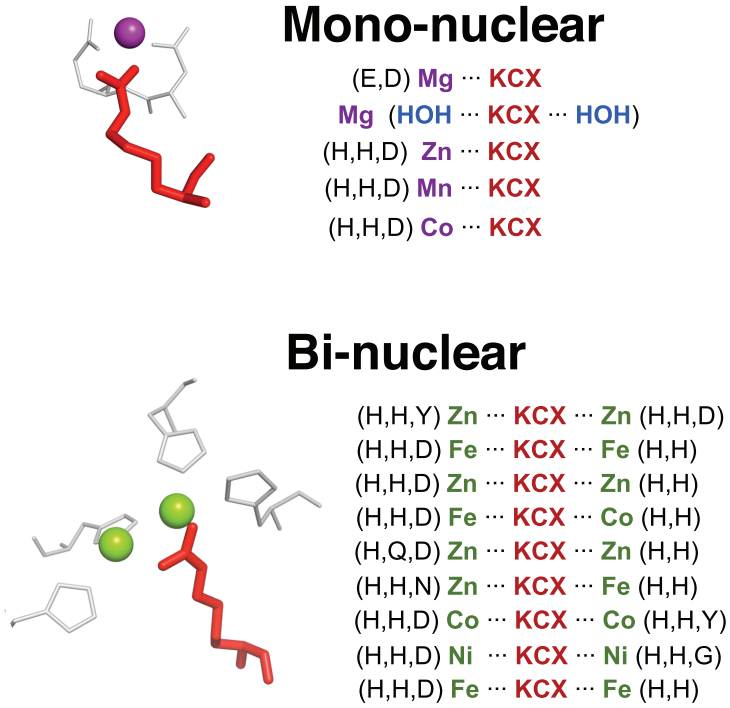

Figure 21. **Schematic representations of metal ion centers involving carboxylated Lys residues.** In parenthesis are the side chains of the residues interacting with the metal ion. For example, (H,H,D)Zn represents side chains of two His and one Asp interacting with a Zn ion. KCX side chains can either interact with one ion (*e.g.* (H,H,D) Zn···KCX), or can bridge two metal ions (*e.g.* (H,H,D)Zn···KCX···Zn(H,H).

4.3.2.5 Metal Ion Centers

The carboxyl group of the KCX residue is commonly found in contact with positively charged ion(s). In 41 out of 62 protein structures, carboxylated Lys residues interact with metal ions in either single or multiple metal ion centers. The majority (33) of the metal centers are binuclear, with 8 mononuclear centers (Figure 21). Overall, we found there are six different metal ions that can interact with KCX: Zn^{2+}, Mg^{2+}, Co^{2+}, Fe^{3+}, Ni^{2+}, and Mn^{2+}. Binuclear centers can contain either the same or different pairs of metal ions. 4.3.2.5 lists the PDB identification, type of metal ions bound, schematic representation of the metal center, protein name, and species source.

Among the six ions, Ni^{2+} and Fe^{3+} are found only in binuclear metal centers. The divalent ions Zn^{2+} and Co^{2+} are found in both mono and binuclear metal centers, while Mg^{2+} and Mn^{2+} are found in mononuclear centers. His and Asp are almost invariably present in all analyzed KCX-containing metal-binding sites. Some binuclear sites contain unusual zinc ligands such as an amide peptide backbone carbonyl (e.g., Gly in ureases), or an hydroxyl group from Tyr (*e.g*, adenine deaminase from *E. faecali*, metal-dependent hydrolase from *L. casei*, and organophosphorus hydrolase from *D. radiodurans*).

Mononuclear metal ions usually form catalytic complexes, with the KCX residue acting as an oxygen donor (Auld, 2001). Binuclear metal sites belong to the co-catalytic type of metal-binding sites (Auld, 2001). They usually contain two metal ions in close proximity bridged by the side chain moiety of the KCX residue. Residues Asp, Glu, or His have also been found carrying out similar roles in other proteins (Auld, 2001). Residues interacting with metal ions

TABLE XII

METAL ION CENTERS RELATED TO KCX SITES

PDB_Chain	Ion	Metal Center Motif	Protein	Specie
2ICS_A	Zn-Zn	(H,H,Y)Zn···KCX···Zn(H,H,D)	Adenine deaminase	*Enterococcus faecali*
3E74_B	Fe-Fe	(H,H,D)Fe···KCX···Fe(H,H)	Allantoinase	*Escherichia coli*
3OVG_A	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Amidohydrolase	*Mycoplasma synoviae*
3MKV_H	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Amidohydrolase	unknown
3DUG_G	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Arginine carboxypeptidase	unknown
2VC7_B	Co-Fe	(H,H,D)Fe···KCX···Co(H,H)	Aryldialkylphosphatase	*Sulfolobus solfataricus*
1NFG_C	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	D-hydantoinase	*Ralstonia pickettii*
1K1D_C	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	D-hydantoinase	*Geobacillus stearothermophilus*
1GKP_C	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	D-hydantoinase	*Thermus sp.*
3PNU_B	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Dihydroorotase	*Campylobacter jejuni*
2Z26_B	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Dihydroorotase	*Escherichia coli*
2OGJ_F	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Dihydroorotase	*Agrobacterium tumefaciens*
2GWN_A	Zn-Zn	(H,Q,D)Zn···KCX···Zn(H,H)	Dihydroorotase	*Porphyromonas gingivalis*
3JZE_D	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Dihydroorotase	*Salmonella enterica*
3DC8_B	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Dihydropyrimidinase	*Sinorhizobium meliloti*
2FVK_C	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Dihydropyrimidinase	*Lachancea kluyveri*
2FTW_A	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Dihydropyrimidine amidohydrolase	*Dictyostelium discoideum*
2J6V_A	Mn(3x)	KCX not in metal center	UV damage endonuclease	*Thermus thermophilus*
2OEM_B	Mg	(E,D)Mg···KCX	Enolase, Rubisco-Like protein	*Geobacillus kaustophilus*
1W78_A	Mg-Mg	(H_2O,HOH)Mg(HOH···KCX···HOH)	FOLC	*Escherichia coli*
1GKR_D	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Hydantoinase	*Arthobacter aurescens*
3ICJ_A	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Hydrolase, metal-dependent	*Pyrococcus furiosus*
2QPX_A	Zn-Zn	(H,H,Y)Zn···KCX···Zn(H,H,D)	Hydrolase, metal-dependent	*Lactobacillus casei*
1ONW_A	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Isoaspartyl dipeptidase	*Escherichia coli*
3OJG_A	Zn-Fe	(H,H,N)Zn···KCX···Fe(H,H)	Lactonase	*Geobacillus kaustophilus*
3PNZ_A	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	Lactonase Lmo2620	*Listeria monocytogenes*
3MTW_A	Zn-Zn	(H,H,D)Zn···KCX···Zn(H,H)	L-Arginine carboxypeptidase	*Caulobacter vibrioides*

Metal Ion Centers. Continue from previous page

PDB_Chain	Ion	Metal Center Motif	Protein	Specie
2XJA_A	Mg	Mg(HOH$\cdots$KCX$\cdots$HOH)	MurE ligase	*Mycobacterium tuberculosis*
3GTX_A	Co-Co	(H,H,D)Co$\cdots$KCX$\cdots$Co(H,H,Y)	Organophosphorus hydrolase	*Deinococcus radiodurans*
2OB3_B	Zn-Zn	(H,H,D)Zn$\cdots$KCX$\cdots$Zn(H,H)	Phosphotriesterase mutant	*Brevundimonas diminuta*
3N2C_P	Zn-Zn	(H,H,D)Zn$\cdots$KCX$\cdots$Zn(H,H)	Prolidase	unknown
2QF7_A	Zn	(H,H,D)Zn$\cdots$KCX	Pyruvate carboxylase	*Rhizobium etli CFN 42*
3BG3_C	Mn	(H,H,D)Mn$\cdots$KCX	Pyruvate carboxylase	*Homo sapiens*
3KDN_H	Mg	(E,D)Mg$\cdots$KCX	RuBisCo	*Thermococcus kodakarensis*
1BWV_C	Mg	(E,D)Mg$\cdots$KCX	RuBisCo	*Galdieria partita*
1WDD_A	Mg	(E,D)Mg$\cdots$KCX	Rubisco	*Oryza sativa*
1RQB_A	Co	(H,H,D)Co$\cdots$KCX	Transcarboxylase 5S subunit	*Propionibacterium freudenreichii*
4UBP_C	Ni-Ni	(H,H,D)Ni$\cdots$KCX$\cdots$Ni(H,H,G)	Urease	*Bacillus pasteurii*
3LA4_A	Ni-Ni	(H,H,D)Ni$\cdots$KCX$\cdots$Ni(H,H,G)	Urease	*Canavalia ensiformis*
1EJX_C	Ni-Ni	(H,H,D)Ni$\cdots$KCX$\cdots$Ni(H,H,G)	Urease	*Klebsiella aerogenes*
1E9Z_B	Ni-Ni	(H,H,D)Ni$\cdots$KCX$\cdots$Ni(H,H,G)	Urease	*Helicobacter pylori*
3QGK_O	Fe-Fe	(H,H,D)Fe$\cdots$KCX$\cdots$Fe(H,H)	Urease	*Helicobacter mustelae*

often come from regions distributed along the entire length of the protein (Auld, 2001), as we clearly observed for KCX sites (Figure 20), which suggests that metal sites may be important not only for its catalytic function, but also to the overall folding of the protein.

In summary, we have characterized the physicochemical environment of the KCX site, which is likely to be associated with the carboxylation event. The relatively large number of positively ionizable residues around the positively charged Lys side chain can be rationalized if the Lys residue becomes carboxylated, and consequently negatively charged. In addition, hydrophobic residues revealed the buried nature of the KCX site. Likewise, the significant presence of aromatic residues suggests their important roles for the functional stability of the site.

4.3.3 Predictor of Lysine Carboxylation (PreLysCar)

4.3.3.1 The Bayesian Classifier

Since a sequence motif cannot be detected, the use of structural information is essential to identifying KCX sites. The components of the microenvironment of both KCX and LYS were used as features for our prediction method, which is based on a naïve Bayesian model. The "naïve" assumption is that all such structural features can be treated effectively as independent. It is known that Bayesian classifiers can achieve excellent results even if the independence of the features is questionable (Zhang, 2004).

We implemented a Bayesian classifier called PreLysCar for Predictor of Lysine Carboxylation (see Methods for details). The effectiveness of the predictor was assessed by performing leave-one-out cross validation tests at different prior probabilities (Table XIII), which is a parameter of the Bayesian model. The final Kcx+ data set consisted of 62 KCX sites and the Lys− of

1,337 LYS sites. The performance of the Bayesian model was evaluated with the technique of leave-one-out cross-validation explained above, under various prior probabilities (Table XIII).

1. Prior = 0.5, an unrealistic assumption of both KCX and LYS having the same probability of occurrence. Under this prior probability value, large number of correctly predicted KCX sites should be expected (high sensitivity), but also a large number of LYS sites incorrectly classified as KCX (low specificity). However, this was not the outcome and 41 false positives out of 1,337 were obtained. Both sensitivity and specificity were above the 90%, with an accuracy of 97%.

2. Prior = 0.05, probability of finding a KCX residue from the total number of Lys amino acids in our data set (62 KCX sites out of 1,399 KCX+LYS sites). Such small prior probability value should improve the specificity by reducing the number of false positives, but worsen the sensitivity with less KCX correctly predicted. As expected, the number of false positives was reduced (from 41 to 14 out of 1,337), but the number of KCX sites correctly predicted was slightly affected (58 to 56). As a result, both sensitivity and specificity of 0.90 and 0.99, respectively, which translates in a better accuracy (0.99).

3. Prior = 0.009, a generous approximation to the probability of finding a KCX residue in the entire Lys universe. As expected, the very low probability value causes a decrease in the number of correctly predicted KCX sites, but not in a significant number (54 out of 62 KCX sites), which maintains a high sensitivity (0.87). The number of correctly predicted LYS residues was very high (1,333 out of 1,337), which results in a specificity of 0.999.

The Matthews correlation coefficient, an indicator of performance when the classes have different sizes, was 0.90, which reflects the overall excellent reliability of our predictor.

TABLE XIII

PERFORMANCE OF THE BAYESIAN CLASSIFIER (LEAVE-ONE-OUT CROSS VALIDATION TEST) AT DIFFERENT PRIOR PROBABILITIES

Prior	TP	TN	FP	FN	SEN	SPF	ACU	PPV	NPV	MCC
0.5	58/62	1296/1337	41/1337	4/62	0.94	0.97	0.97	0.59	1.00	0.73
0.05	56/62	1323/1337	14/1337	6/62	0.90	0.99	0.99	0.80	1.00	0.84
0.009	54/62	1333/1337	4/1337	8/62	0.87	1.00	0.99	0.93	0.99	0.90

The KCX+ (positive) data set consists of 62 KCX sites. The LYS− (negative) data set 1,337 LYS sites. Labels: True Positive (TP); True Negative (TN); False Positive (FP); False Negative (FN); Sensitivity (SEN); Specificity (SPF); Accuracy (ACU); Positive Predictive Value (PPV); Negative Predictive Value (NPV); Matthews Correlation Coefficient (MCC).

4.3.3.2 Blind Predictions: Search to the Protein Data Bank

We applied PRELYSCAR to a subset of protein structures (>200 residues) solved by X-ray crystallography in the PDB database consisting of 14,324 protein chains after redundancy removal. A total of 291,434 Lys residues were found in these proteins, and their microenvironments were determined (see Methods). PRELYSCAR predicted that 598 protein structures contain a carboxylated Lys residue (denoted as "predicted KCX", pKCX).

For a subset of proteins, there is overwhelming evidence supporting our predictions. Some Lys residues predicted to be carboxylated that are known to be carboxylated, had structures

that were solved with the Lys residue uncarboxylated. Generally, this is a consequence of either the effect of an inhibitor, or experimental conditions that prevented carboxylation. PreLysCar also identified in multimeric proteins, known to have a carboxylated Lys residue, subunits in which an equivalent Lys was resolved uncarboxylated.

For example, the phosphotriesterase from *Pseudomonas diminuta* contains Lys169, which was predicted to be carboxylated (PDB ID:1PSC). The structure was described with this Lys residue as carboxylated and bridging two atoms of cadmium at the active site of this holoenzyme (Benning et al., 1995). However, it appears as uncarboxylated in the atomic coordinates available in the PDB. Another example is Lys84 of carbapenemase OXA-24 (PDB ID:2JC7). This enzyme belongs to the family of class D β-lactamases, which is well known to have a carboxylated Lys as part of the active site (Poirel et al., 2010). The electron density map showed a clear cloud of electrons at the tip of Lys84. In this case, a carboxyl group could be built with confidence. The structure of the same protein was subsequently solved again several years later, and a carboxyl group was indeed added to Lys84 (Bou et al., 2010). See below for more related examples and extended descriptions.

Among other predicted KCX proteins, the composition and structural arrangement of residues at KCX sites show the defining signature features of known KCX sites described above. For example, Lys residues predicted to be carboxylated were found buried in highly populated cavities, with three or more His residues located at the mouth of the cavity, and a large number of hydrophobic and aromatic residues exist as part of the microenvironment, all within 5 Å from the pKCX side chain.

A number of these pKCX proteins were found to share 90 to 20% sequence identity to known KCX proteins from different species, indicating that they are likely orthologs. Therefore, the carboxalytion event may also occur in these homologous proteins. For example, Lys278 from the uncharacterized metal-dependent hydrolase of *Pyrococcus horikoshii* (PDB ID:3IGH) was predicted as carboxylated. This protein shows a 70% sequence identity with another metal-dependent hydrolase from *Pyrococcus furiosus* (PDB ID:3ICJ), where Lys294 is known to be carboxylated. Futhermore, both microenvironments show high sequence and structural similarity. Another example is the alanine racemase from *Enterococcus faecalis* (PKCX132, PDB ID:3E5P), which shares 50% sequence identity with the alanine racemase from *Streptococcus pneumoniae* (KCX129, PDB ID:3S46). An extended description for these and many other selected pKCX proteins can be found below.

Sequence similarity was also discovered along short fragments between pKCX and KCX proteins. Although the overall protein topology was found to be different for some of them, these fragments share local structural similarity, generally consisting in alpha helices. These fragments are enriched in the region where both KCX and pKCX are found. Some examples include the putative oxidoreductase from *Erwinia carotovora atroseptica* (PDB ID:2P2S) and the class II fructose-biphosphate aldolase from *Helicobacter pylori* (PDB ID:3C4U) (see below for details).

The uncharacterized protein YML079W from *Saccharomyces cerevisiae* (PDB ID:1XE7) is a protein predicted to have a carboxylated Lys with no sequence similarity detected to any known KCX protein. Lys147 is found buried in a central location of the protein structure. At

least four His and one Glu residues are in positions that resembles a typical KCX site. This protein, of unknown function, was solved at 1.75 Å and pH 5.6, which could have prevented carboxylation. Despite the high resolution, the side chain of Lys147 was modeled in two different conformations, which may provide an indication of an unstable state of Lys147 (see below for more predictions).

Next, we extend the description for a selection of proteins predicted to have a carboxylated Lys residue.

Phosphotriesterase from *Pseudomonas diminuta* (pdb id:1psc). Lys169 was predicted as carboxylated. Phosphotriesterase detoxifies paraoxon and parathion, pesticides widely used, in addition to various mammalian acetylcholinesterase inhibitors (Benning et al., 1995). Lys169 was resolved as carboxylated and bridging two atoms of cadmium at the active site of this holoenzyme (Benning et al., 1995). However, it does not appear as carboxylated in the atomic coordinates.

Carbapenemase OXA-24 from *Acinetobacter baumannii* (2jc7). Lys84 was predicted as carboxylated. Carbapenemase OXA-24, a class D β-lactamase, was isolated from a multi-resistant epidemic clinical strain of *Acinetobacter baumannii* (Santillana et al., 2007). This family of anti-β-lactams is well-known to have a carboxylated lysine as part of the active site. The electron density map shows a clear cloud of electrons at the tip of Lys84. A carboxyl group could be built with confidence. We found that the structure of the same protein was solved years later and the carboxyl group was added to Lys84 (Bou et al., 2010), which verify our prediction.

Amidohydrolase Sgx9260c from environmental sample (3feq). Lys188 was predicted as carboxylated. The structure of Sgx9260c was solved along with Sgx9260b (PDB ID:3MKV), an homologous protein from the same super-family. Both proteins share a 98% sequence identity and their active sites are almost identical, which is typical of type I binuclear metal centers in the amidohydrolase superfamily (Seibert and Raushel, 2005b). The active site consists in a zinc-binding site composed of six conserved residues, including a carboxylated lysine. While Sgx9260b (PDB ID:3MKV) was solved with the carboxylated Lys at position 191, Sgx9260c (PDB ID:3FEQ) lack the post-translational modification at the equivalent position Lys188, despite our prediction. This is because the structure was solved in the presence of an inhibitor and consequently the metal-binding site of Sgx9260c was only partially occupied by Zn^{2+} and partially filled with water (Xiang et al., 2010).

Uncharacterized metal-dependent hydrolase from *Pyrococcus horikoshii* (3igh). Lys278 was predicted as carboxylated. This protein shows a 70% sequence identity with another metal-dependent hydrolase from *Pyrococcus furiosus* (PDB ID:3ICJ), where Lys294 was found to be carboxylated and bridging two Zn ions. Both pKCX and KCX sites show a similar microenvironment with one Glu and 5 His residues characteristic of KCX metal ion centers. The predicted pKCX site contains a sulfate ion, which could prevent carboxylation.

Human dihydropyrimidinase (2vr2). Lys159 from the human dihydropyrimidinase was predicted as carboxylated. The electron density map shows high density between the tip of the Lys159 and a zinc ion. This protein shares 60% sequence identity with the dihydropyrimidinase from *Dictyostelium discoideum* (PDB ID:2FTW), which exhibits a KCX residue at the active

site (KCX158). Dihydropyrimidinases catalyses the reversible hydrolytic ring-opening of cyclic diamides. The active site of dyhydropyrimidinases is characterized for a carboxylated lysine residue involved in bridging a binuclear zinc center.

Alanine racemase from *Enterococcus faecalis* (3e5p). Lys132 was predicted as carboxylated. This racemase shares 50% sequence identity with the alanine racemase from *Streptococcus pneumoniae* (PDB ID:3S46), where Lys129 was resolved as carboxylated and hydrogen bonded to the neighboring Arg residue (Im et al., 2011). There are strong envidences for the presence of such carboxylated residue in alanine racemases (Morollo et al., 1999; LeMagueres et al., 2003).

Allantoinase from *Bacillus halodurans* (3hm7). Lys150 was predicted as carboxylated. This protein shares 40% sequence identity with the allantoinase of *Escherichia coli* (KCX146, PDB ID:3E74 (Kim et al., 2009)). Allantoinases are involved in the final step of the biogenesis and degradation of ureides by catalyzing the conversion of (S)-allantoin into allantoate (Kim et al., 2009). In the allantoinase of *E. coli*, lys146 is carboxylated and bridging two iron metal ions, with one of the ions coordinated in a DHH$\cdots$KCX$\cdots$HH motif. In the case of *Bacillus halodurans* (PDB ID:3HM7), a zinc atom is present in structural region similar to *E. coli* protein, with a similar motif DHH coordinating the Zn ion. The predicted KCX residue is in the proximity, in addition to other His residues.

Rubisco-like enzymes. PreLysCar identified several protein enzymes with different degrees of similarity to Rubisco proteins that are known to have a carboxylated lysine residue. For example, Lys189 of the type III Rubisco from the hyperthermophilic archaeon *Thermococcus*

kodakarensis (Tk-Rubisco, PDB ID:1GEH) was predicted as carboxylated. Although this residue was resolved as carboxylated (Kitano et al., 2001), it was modeled without the carboxyl group. The same protein was years later solved with the carboxyl group on it (pdb id:3a12) (Nishitani et al., 2010). Lys186 of the Rubisco from *Pyrococcus horikoshii* (PDB ID:2CWX) was also predicted as carboxylated. This Rubisco shows a 40% sequence identity with Tk-Rubisco. Both microenvironment are very similar.

Carboxypeptidase (2qs8). Lys194 of this carboxypeptidase from the amidohydrolase superfamily (unknown source) was predicted to be carboxylated. The protein shares a 36% sequence identity with the ZN-dependent arginine carboxypeptidase (PDB-ID:3DUG), which was found to have a carboxylated Lys residue in the equivalent position (KCX182). Both structures were published in the same paper (Xiang et al., 2010). Despite the low sequence identity, both proteins share the same fold. A binuclear metal center coordinated by Asp and five His residues was expected, with Zn ions bridged by a carboxylated lysine (Xiang et al., 2010). However, they were not able to obtain the metal ions in that position. The authors emphasized the difficulties in capturing this post-translational modification through purification and crystallization. Our prediction confirms their expectations.

Imidazolonepropionase enzymes. KCX residues were predicted in a group of imidazolonepropionase enzymes from three different organisms: *Agrobacterium tumefaciens* (Lys155, PDB ID:2GOK), *Bacillus subtilis* (Lys149, textscpdb id:2bb0), and from an unknown environmental sample (Lys 141, PDB ID:2OOF). These three hydrolases share around 40% sequence identity among themselves. The predicted KCX residues are located in a similar position in

the three proteins, *i.e.*, buried in a cavity, which is also located in a central region of the protein. Each of these enzymes shared an overall 20 to 30% sequence identity with two other different hydrolases that are known to have carboxylated Lys residues, the D-hydantoinase from Burkholderia pickettii (PDB-ID:1NFG) and the amidohydrolase PDB-ID:3MKV) from an unknown environmental sample (Xiang et al., 2010). Despite the remote homology, all these proteins show similar topology of the active site and overall protein fold.

Putative oxidoreductase from *Erwinia carotovora atroseptica* (2p2s). Lys96 was predicted as carboxylated. This Lys residue is buried within the protein, in the center of a pocket. The microenvironment and the disposition of the residues is similar to other KCX sites. Four His and one Asp residues are within a distance that could allow the formation of a metal ion center. Lys96 was solved forming hydrogen bonds with 3 water molecules and Asp178, which was modelled in two possible conformations, despite the high-resolution of the structure (1.25 Å).

This putative oxidoreductase shows high sequence similarity with a large number of annotated oxidoreductases. Structurally, two main protein domains are identified by PFAM that belong as well to the oxidoreductase family. Interestingly, we detected remote sequence similarity in a fragment of 75 residues between this oxidoreductase and the rubisco-like protein from *Burkholderia fungorum* (PDB ID:3NWR). The fragment covers both the predicted PKCX96 of the oxidoreductase and the carboxylated Lys (Kcx195) of the rubisco-like protein. Although the topology of both proteins is different, the fragments share local structural similarity, con-

sisting of two alpha helices and a region part of the pocket where both Kcx195 and pKcx96 are located.

Class II fructose-biphosphate aldolase from *Helicobacter pylori* (3c4u). Lys 251 was predicted as pKCX. The composition and structural organization of the microenvironment resemble the KCX sites described. The Lys predicted as carboxylated is buried in the center of a cavity with one Glu and four His residues within a good distance for metal binding. The protein was solved forming hydrogen bonding with Gln, Asn, and Glu. A sodium ion was found on the tip of the Lys residue, although a negative peak shows in the electron density map for Na. A Zn ion was also found in the proximities of the pKCX coordinated by two His residues.

We detected a 27% sequence identity in a fragment of 80 residues between this aldolase and two different KCX proteins, the OXA 10 class D beta-lactamase from *Pseudomonas aeruginosa* (PDB ID:2X02) and the transcarboxylase from *Propionibacterium shermanii* (PDB ID:1RQB). Although the overall topology between the proteins is different, the fragments share local structural similarity, which forms several alpha helices and a region part of the pocket where the KCX and pKCX are located.

Tagatose 6-phosphate kinase from *Escherichia coli* (2fiq). Lys279 was predicted as carboxylated. The microenvironment of this lysine shows the pattern that we have described, with the pKCX residue situated in the center of the TIM-barrel domain of the protein. Three His and a negatively charged residue are within 5 Å from the amino terminal group of the pKCX. The protein structure was solved at 2.23 Å resolution and pH 6.5. At the moment of

writing this article, there is no structural and functional characterization of the mechanism for tagatose 6-phosphate kinases.

Mannose 6-Phosphate Isomerase from *Bacillus subtilis* (1qwr). Lys96 was predicted as KCX. This protein does not share sequence similarity with any protein with a known carboxylated Lys residue. This monomeric enzyme catalyzes the interconversion of fructose 6-phosphate and mannose-6-phosphate (Yeom et al., 2009). It uses zinc as cofactor ligand, which appears to play a role in substrate binding and in maintaining the architecture of the active site (Sagurthi et al., 2009). The residue Lys96 lies near a zinc ion in the protein structure. The structure was solved at 1.8 Å and pH of 5.5, which might have prevented carboxylation. Despite the high-resolution, the electron density for the side chain of this amino acid is poor. A known KCX protein was solved with one atom of Zn, the multifunctional pyruvate carboxylase from *Rhizobium etli* (PDB ID:2QF7 (St Maurice et al., 2007)), where the Zn atom was found coordinated by 2 His and 1 Asp residues, instead of Glu observed for the mannose 6-phosphate isomerase.

Mandelate racemase from *Roseobacter denitrificans* (3tcs) and *Roseovarius nubinhibens* (3u4f). Lys145 is predicted in both proteins to be carboxylated. These two racemases from the muconate lactonizing protein family share a 85% sequence identity between themselves. No sequence similarity was detected to any protein with a known carboxylated Lys residue. However, the microenvironment of both Lys is similar to carboxylated lysine residues described. In both pKCX sites, an atom of Mg is present, although in different conditions. In one of the structures is bind to a D-alanine, while in the other is bind to a nucleotide (guani-

dine). In addition, the structure of *Roseovarius nubinhibens* (PDB ID:3U4F) was solved using selenomethionine, which may potentially prevent carboxylation on Lys145.

Glycerol dehydrogenase (GlyDH) from *Bacillus stearothermophilus* (1jpu) and *Sinorhizobium meliloti* (3uhj). Lys277 and Lys298 were predicted as carboxylated in the glycerol dehydrogenase (GlyDH) from *Bacillus stearothermophilus* (PDB ID:1JPU) and *Sinorhizobium meliloti* (PDB ID:3UHJ), respectively. Both proteins share a 40% sequence identity, in addition to an overall similar structural topology. These proteins do not share sequence similarity with any protein with a known carboxylated Lys residue. Both Lys residues occupy a equivalent position within the structure. The composition of the microenvironment is similar as well, with at least five His and a Glu residue within a distance that would allow the formation of a ion metal center.

The GlyDH from *B. stearothermophilus* was originally indicated as part of the iron-containing alcohol dehydrogenase family, but curiously, based on the strict dependency on zinc for the protein activity, the entire group was renamed to "family III metal-dependent polyol dehydrogenases". GlyDH from *B. stearothermophilus* (PDB ID:1JPU) was solved at 1.8 Å resolution (pH not available) with two zinc ions solved in the same area where the pKCX is found. The GlyDH and *Sinorhizobium meliloti* (PDB ID:3UHJ) was solved at 2.34 Å, pH of 5.5, and only one of the two Zn ions resolved.

YML079w from *Saccharomyces cerevisiae* (1xe7). Lys147 was predicted as carboxylated. It is found buried in a central location of the protein structure. No sequence similarity with any protein with a known carboxylated Lys residue was identified. At least four His and

one Glu residues are in positions that resembles a typical KCX site. This protein, of unknown function, was solved at 1.75 Å and pH 5.6, which could have prevented carboxylation. Despite the high resolution, the side chain of Lys147 was modeled in two different conformations, which may provide an indication of an unstable state of Lys147.

ArnB (PmrH) aminotransferase from *Salmonella typhimurium* (1mdo). Lys188 was predicted as carboxylated. This residue is buried in the active site, buried at the center of the protein. No sequence similarity was detected with any protein with a known carboxylated Lys residue. The composition of the microenvironment is similar to the KCX sites described. The protein was solved at 1.7 Å and pH 5.5. An electron density feature was found to extend from the ligand, which is in contact with Lys188 and that could not be explained (Noland et al., 2002). In addition, Lys188 was proposed to act as a general base to abstract the -proton, but also involved in a role as a general acid or base in several steps of the enzymatic mechanism.

4.3.3.3 Correcting Mis-Modeled KCX Residues

There are three proteins solved with a carboxylated Lys residue that PreLysCar classified as LYS. These are the KCX residues of the TdcF from *Escherichia coli*, alpha-l-fucosidase from *Thermotoga maritima*, and class C *beta*-lactamase from Enterobacter cloacae. If our predictions are correct, these three KCX residues should be considered to be uncarboxylated.

4.3.3.3.1 TdcF from *Escherichia coli* (2uyn)

TdcF is a member of the highly conserved family YjgF/YER057c/UK114, widespread in nature. The biological function is not known (Burman et al., 2007). Kcx58 of TdcF (pdb-id:2uyn) is likely to be a non-carboxylated Lys residue. First, the authors suggested that the

KCX modification could be an artefact with no biological importance (Burman et al., 2007). Second, the ambiguity of the electron density map at the tip of Lys58 compelled the authors to model the carboxylated Lys in two different positions. Third, known KCX sites are mostly buried, but Kcx58 is on the surface and is quite accessible to solvent. Finally, no evidence exists for this modification in any other TdcF structure (Burman et al., 2007). Therefore, we believe that Lys58 was likely resolved incorrectly as carboxylated. As a consequence, it was eliminated from the KCX+ data set.

4.3.3.3.2 Putative α-l-fucosidase of *Thermotoga maritima* (1hl9)

α-l-fucosidases catalyze the removal of non-reducing terminal l-fucose residues (Sulzenbacher et al., 2004). Lys338 was resolved with a carboxyl group, although several reasons suggest that this residue is likely non-carboxylated. First, Lys338 showed extra density in chain A, but not in chain B. Second, the remaining atomic coordinates solved in the same publication in different conditions for the same protein, showed low resolution for the side chain of Lys338 (Sulzenbacher et al., 2004). Third, although known KCX sites are found mostly buried, Lys338 is at the surface and quite accessible to solvent. Four, the structure of the same protein was solved years later and Lys338 was not modelled carboxylated (Wu et al., 2010). Finally, fucosidase is a lysosomal enzyme, which operates at pH 5. Given the lability of the carboxyl group under acidic environments, it is unlikely that Lys338 could become carboxylated. Since Lys residues can also be modified by other PTMs, which is difficult to distinguish based on the electron density alone, other PTM cannot be ruled out. Therefore, we believe that Lys338 is mis-

modeled and is likely uncarboxylated. As consequence, it was eliminated from the KCX+ data set.

4.3.3.3.3 Class C β-lactamase of Enterobacter cloacae (2p9v)

This serine-dependent enzyme is involved in the hydrolysis of the β-lactam ring of the β-lactam antibiotic (Bush et al., 1995). Lys315, resolved as carboxylated but predicted as non-carboxylated, adopted an unusual conformation. A carbamate cross-linking between Lys315 and Ser64 occurred as consequence of the action of an inhibitor (Wyrembak et al., 2007). However, carboxylation does not occur in functional proteins of the class C of β-lactamases, as it occurs only in class D (Poirel et al., 2010). There are 66 solved structures of class C β-lactamase available at the PDB database sharing more than 98% sequence identity with 2P9V and none of them were resolved with a KCX residue. Carboxylation occurred in this class C β-lactamase as a consequence of an inhibitor of Lys315, but does not occur in functional states of the class C β-lactamase. Therefore, we believe that PreLysCar correctly classified Lys315 as uncarboxylated. Consequently, the protein was removed from the KCX+ set.

4.3.3.4 Test on High-Resolution Protein Structures (<1.5Å)

We selected a set of 577 high-resolution protein structures (< 1.5 Å) each with more than 200 amino acids (see Methods). The high-resolution allowed us to make the reasonable assumption of a less likely mis-modeled carboxylated Lys. We tested the performance of PRESLYSCAR on this set. In total, 12 out of $9,775$ Lys residues were misclassified, which reflects a specificity of 99.88% for our predictions.

4.3.4 Biochemical Function of Proteins With Predicted Kcx Groups

PreLysCar predicted 598 protein chains out of 14,324 with a carboxylated Lys residue. Among them, 58% are from eubacteria, 34% from eukaryotes, and 8% from archaea. We were interested on studying their enzymatic functions. We used the EC number available in the PDB when available, *i.e.* the classification scheme according to the Enzyme Commission (Tipton, 1994), to describe the biochemical function of enzymes predicted to have a KCX group.

A total of 344 have their EC numbers annotated, with 220 unique EC identifiers. We find that 215 out of the 220 EC numbers are chemical reactions catalyzed by enzymes previously not known to require a Lys residue to be carboxylated. Among them, oxidoreductases (EC:1), which are not present in the current set of enzymes with KCX sites, appear in large numbers (45 out of 220). Hydrolases (EC:3) and transferases (EC:2) also have significant number of pKCX sites (58 and 54 respectively). Lyases (EC:4), isomerases (EC:5), and ligases (EC:6) have 27, 18, and 18 enzymes with carboxylated Lys residues, respectively, with an enriched number of reactions where the carboxylation of Lys might play some roles. Overall, our results suggest that KCX may be far more prevalent than is currently known.

4.4 Discussion

We developed a computational method capable of identifying Lys residues that undergo spontaneously carboxylation. However, these Lys residues may exist in an uncarboxylated form in X-ray protein structures due to experimental conditions such as acidity, availability of the required metal ions, presence of inhibitors, or intrinsic lability of the carboxyl group.

The diverse group of proteins with a carboxylated Lys residue is consistent with the fact that currently there are no sequence motifs known for amino acids surrounding the KCX site (Figure 18). However, the structural conservation of the KCX site that we have identified made it possible to predict KCX sites. The task is, however, not exempt of complications. First, the microenvironment of non-carboxylated Lys residues buried in the protein often resembles that of the KCX site. Second, the correct classification of KCX residues is more challenging for those KCX residues that are not part of metal ion centers, *i.e.* either involved in hydrogen bond interactions with other residues of the active site, or directly involved in the catalytic reaction, due to a larger variability of amino acids.

Finally, the protein conformational changes that occur before and after the carboxylation of the Lys residue may have an impact on the local environment of the affected Lys residue. If the protein was crystallized in a non-carboxylated state, the microenvironment of a positively ionizable residue could be different from the carboxylated Lys residue, and consequently, more difficult to detect.

In spite of all these difficulties, PreLysCar generally works well in distinguishing Lys residues that become carboxylated from those that do not.

4.4.1 Estimation of the Prevalence of Lys Carboxylation

The development of a useful and reliable computational prediction tool requires a balance in the trade-off between sensitivity and specificity. Our emphasis has focused on minimizing the number of false positives, *i.e.* LYS incorrectly predicted as KCX sites, while maintaining the best possible sensitivity (large number of true positive KCX sites).

We assessed the extent of Lys carboxylation using PreLysCar. In addition to the tests carried out using the training data set, we further evaluated the false positive rate in a set of high-resolution protein structures (less than 1.5Å). High resolved protein structures allow us to make the reasonable assumption of a less likely mis-modeled carboxylated Lys. PreLysCar misclassified 12 Lys residues as carboxylated in 577 protein structures, *i.e.* it has a false positive rate of 2.08%.

PreLysCar predicted 598 proteins with a pKCX site in 14,324 protein chains with more than 200 residues and more than 1.5 Å resolution. Based on the calculated false positive rate (2.08%), even before applying PreLysCar, we should expect about 298 out of the 14,324 proteins be incorrect predictions. But since PreLysCar predicted 598 pKCX proteins, this implies that the remaining 300 predicted KCX are expected to be correct. According to this conservative estimation, about 2% of proteins with more than 200 amino acids in the PDB database may potentially be subjected to spontaneous Lys carboxylation, which implies a much broader prevalence of KCX than it is currently known.

4.4.2 Implications

Taking into consideration the importance of post-translational modifications and how they contribute to enrich the functional variability of proteins, we believe that spontaneous PTMs, such as the one here studied, can have a much broader implication than is currently known, both in eukaryotes and prokaryotes. Spontaneous PTMs can switch enzymes on and off under the appropriate physicochemical conditions, which may provide yet another elegant and efficient biological mechanism of regulation. As shown by our study of Lys carboxylation, computational analysis may help to overcome experimental challenges in detecting spontaneous PTMs due to a large number of experimental complications.

4.5 Conclusion

We have developed a method for the prediction of Lys carboxylation. This PTM is difficult to detect with common experimental techniques such as mass spectrometry and X-ray crystallography. We have characterized the KCX microenvironment and its defining features. The computational method Predictor of Lysine Carboxylation (PreLysCar) developed here can provide useful predictions with excellent performance (87% sensitivity and 99.9% specificity). PreLysCar predicted KCX residues on proteins within different ranges of sequence identity with homologous proteins known to have KCX residues. It also unveiled carboxylation sites on proteins without similarities with any known cases. In addition, a few Lys sites mis-modeled as carboxylated in the Protein Data Bank have been also uncovered. Our results indicate that the scope of this spontaneous PTM, which does not require the action of additional enzymes, might be larger than expected. We estimate that about 2.1% of the protein structures available in the PDB database with more than 200 residues contain a carboxylated Lys residue, which implies a three-fold increase of what it is currently known. Spontaneous post-translational modifications of functional residues under certain physical-chemical are straightforward and may provide an efficient mechanism of biological regulation.

CHAPTER 5

CONCLUSIONS

In this thesis, we have developed computational methods to examine the effects of several physicochemical and functional constraints on different levels of the sequence and structure of proteins.

5.1 On β-Barrel Membrane Proteins

We have characterized the substitution patterns of residues in different interfaces of the transmembrane segments of β-barrel membrane proteins. We found that residues facing both the lipid environment and the interior of the barrel have characteristic patterns, which differ from each other. We also derived scoring matrices from the estimated substitution rates. Our blind tests showed that our scoring matrices can identify remote homologs with excellent specificity and sensitivity. In addition, we have shown that these scoring matrices can be used to detect mitochondrial outer membrane proteins, suggesting that these two classes of membrane proteins share the same pattern of residue substitution throughout evolution. Our results also imply that the structures of the TM-segments of a large number of β-barrel membrane proteins can be predicted reliably based on aligned structural templates. We developed a web utility to perform database searches using our scoring matrices.

We have also shown how an evolutionary analysis of the protein-protein interaction interface of the trimeric porin OmpF allowed the identification of important residues required to stabilize

the interactions in the protein-protein interface of the TM domain. Combining experimental and computational approaches to engineer β-barrel membrane proteins with different oligomerization states and structural properties can facilitate further studies in structural biology of membrane proteins. It can also help in the design of novel nanopores in nanobiotechnology.

5.1.1 Future Work

The Bayesian Markov chain Monte Carlo method used to estimate amino acid substitution rates is computationally expensive. To overcome this limitation, we selected a set of β-barrel membrane protein families with sufficient phylogenetic variability in order to reduce the number of sequences necessary for the simulation. Although the specificity obtained using the scoring matrices derived from the substitution rate estimations was excellent, the sensitivity obtained in database searches was poor. A valid-pair correction method was further applied, which successfully improved the sensitivity. An interesting next step could be to extent the analysis to the loop regions of β-barrel membrane proteins. It would be also interesting to test the same approach to other protein families, as for example, α-helical membrane proteins.

The characterized pattern of amino acid substitutions was useful as part of a toolkit for protein engineering design. Experimental results showed that wild-type and designed mutants were successfully expressed and purified from inclusion bodies under denaturing conditions (Naveed et al., 2012). They also showed that mutants had the same typical secondary structure spectra as wild-type OmpF. However, there are questions that remain to be answered, for example, if these mutants are correctly folded and functionally active. The *in vivo* expression, a proper re-

constitution avoiding inclusion bodies, in addition to electro-physiological measurements, could provide reliable answers.

5.2 On Functional Cavities of Protein Enzymes

We have defined the catalytic active sites of enzymes using a computational geometry approach applied on the Catalytic Site Atlas (CSA) (Porter et al., 2004), a database of enzyme structures and annotated catalytic residues. We found that these active sites have large numbers of acid and base side chains. We measured the volume of the catalytic active sites using precise geometrical method. For the first time, the volume and number of acid and base side chains have been used to provide the number density of charged residues defining these cavities. We call it *charge density* (CharDen). The charge density of active sites was found to be high and relatively constant among a variety of enzymes. It can also be used to successfully identify the active site of proteins. In addition, we identified other pockets on the surface of the protein (we call them 'craters') with high values of charge density. We proposed that they might be involved in conformational fluctuations affecting the chemical steps of enzyme catalysis.

5.2.1 Future Work

Despite using a computational method based on precise geometrical models, the structure of proteins do not always show structural conformations for the accurate automatic definition of these spaces. Further improvements to address specific geometrical challenges will be helpful. The most reliable data source that we found was limited to 577 protein enzymes (after redundancy reduction). Larger samples of protein enzymes with their corresponding catalytic residues annotated would be also convenient. Finally, how to effectively incorporate the charge

density of active sites in protein function prediction, a field with a considerable need for improvement (Radivojac et al., 2013), remains a challenging task.

5.3 On Lysine Carboxylation

Lysine carboxylation is a post-translational modification that is difficult to detect with common experimental techniques such as mass spectrometry and X-ray crystallography. We have characterized the KCX microenvironment and its defining features. We have also developed a method for the prediction of Lys carboxylation. Our computational method called PreLysCar (Predictor of Lysine Carboxylation) predicted KCX residues on proteins within different ranges of sequence identity to homologous proteins known to have KCX residues. It also unveiled carboxylation sites on proteins without similarities to any known cases. In addition, a few lysine sites mis-modeled as carboxylated in the protein structure database have been also uncovered. Our results indicate that the scope of this spontaneous PTM, which does not require the action of additional enzymes, might be larger than expected. We estimate that about 2% of the protein structures with more than 200 residues available in the PDB database contain a carboxylated lysine residue, which represents a three-fold increase of what it is currently known. We also suggest that spontaneous post-translational modifications affecting functional residues under certain physical-chemical are straightforward and may provide an efficient mechanism of biological regulation.

5.3.1 Future Work

Our computational method obtained excellent results, not only in the training and testing datasets, but also in the blind searches on the PDB database by identifying numerous likely

orthologs of KCX proteins. However, a final test is yet to be performed: the experimental validation of the predictions. The validation would imply a convincing proof of the estimated extension of this spontaneous post-translational modification, and suggest the development of more precise experimental and computation procedures for the accurate detection of this and other post-translational modifications.

REFERENCES

Adachi, J. and Hasegawa: MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Computer Science Monographs of Institute of Statistical Mathematics, 28:1–150, 1996.

Adiga, S. P., Jin, C., Curtiss, L. A., Monteiro-Riviere, N. A., and Narayan, R. J.: Nanoporous membranes for medical and biological applications. Wiley Interdisciplinary Reviews. Nanomedicine and Nanobiotechnology, 1(5):568–581, October 2009. PMID: 20049818.

Adiga, S., Jin, C., Curtiss, L., Monteiro-Riviere, N., and Narayan, R.: Nanoporous membranes for medical and biological applications. Wiley Interdiscip Rev Nanomed Nanobiotechnol, 1(5):568–581, Sep 2009.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J.: Basic local alignment search tool. J Mol Biol, 215(3):403–10, 1990.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J.: Basic local alignment search tool. Journal of Molecular Biology, 215(3):403–410, October 1990. PMID: 2231712.

Anfinsen, C. B.: Principles that govern the folding of protein chains. Science (New York, N.Y.), 181(4096):223–230, July 1973. PMID: 4124164.

Anfinsen, C. B., Haber, E., Sela, M., and White, F H, J.: The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proceedings of the National Academy of Sciences of the United States of America, 47:1309–1314, September 1961. PMID: 13683522.

Anslyn, E. V. and Dougherty, D. A.: Modern Physical Organic Chemistry. University Science, illustrated edition edition, July 2005.

Auld, D. S.: Zinc coordination sphere in biochemical zinc sites. Biometals: an international journal on the role of metal ions in biology, biochemistry, and medicine, 14(3-4):271–313, December 2001. PMID: 11831461.

Bagos, P. G., Liakopoulos, T. D., and Hamodrakas, S. J.: Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. BMC Bioinformatics, 6:7, 2005. PMID: 15647112.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S.: MEME SUITE: tools for motif discovery and searching. Nucleic acids research, 37(Web Server issue):W202–208, July 2009. PMID: 19458158.

Baldwin, V., Bhatia, M., and Luckey, M.: Folding studies of purified lamb protein, the maltoporin from the escherichia coli outer membrane: trimer dissociation can be separated from unfolding. Biochim Biophys Acta, 1808(9):2206–2213, Sep 2011.

Banerjee, A., Mikhailova, E., Cheley, S., Gu, L., Montoya, M., Nagaoka, Y., Gouaux, E., and Bayley, H.: Molecular bases of cyclodextrin adapter interactions with engineered protein nanopores. Proc Natl Acad Sci U S A, 107(18):8165–8170, May 2010.

Barker, W. C., Garavelli, J. S., Haft, D. H., Hunt, L. T., Marzec, C. R., Orcutt, B. C., Srinivasarao, G. Y., Yeh, L. S., Ledley, R. S., Mewes, H. W., Pfeiffer, F., and Tsugita, A.: The PIR-International protein sequence database. Nucleic Acids Res, 26(1):27–32, 1998.

Barreteau, H., Kovac, A., Boniface, A., Sova, M., Gobec, S., and Blanot, D.: Cytoplasmic steps of peptidoglycan biosynthesis. FEMS microbiology reviews, 32(2):168–207, March 2008. PMID: 18266853.

Barthel, J., Krienke, H., and Kunz, W.: Physical Chemistry of Electrolyte Solutions: Modern Aspects. New York, Springer, 1998.

Ben-Shimon, A. and Eisenstein, M.: Computational mapping of anchoring spots on protein surfaces. J Mol Biol, 402(1):259–277, Sep 2010.

Benkovic, S. J. and Hammes-Schiffer, S.: A perspective on enzyme catalysis. Science, 301(5637):1196–1202, August 2003.

Benning, M. M., Kuo, J. M., Raushel, F. M., and Holden, H. M.: Three-dimensional structure of the binuclear metal center of phosphotriesterase. Biochemistry, 34(25):7973–7978, June 1995. PMID: 7794910.

Benz, R.: Permeation of hydrophilic solutes through mitochondrial outer membranes: review on mitochondrial porins. Biochimica Et Biophysica Acta, 1197(2):167–196, June 1994. PMID: 8031826.

Berg, J., Tymoczko, J., and Stryer, L.: Biochemistry. 5th edition. New York: W H Freeman, 2002.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., and Zardecki, C.: The protein data bank. Acta Crystallogr D Biol Crystallogr, 58(Pt 6 No 1):899–907, 2002.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E F, J., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M.: The protein data bank: a computer-based archival file for macromolecular structures. Journal of molecular biology, 112(3):535–542, May 1977. PMID: 875032.

Bhabha, G., Lee, J., Ekiert, D. C., Gam, J., Wilson, I. A., Dyson, H. J., Benkovic, S. J., and Wright, P. E.: A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. Science (New York, N.Y.), 332(6026):234–238, April 2011. PMID: 21474759.

Bigelow, H. and Rost, B.: Online tools for predicting integral membrane proteins. Methods Mol Biol, 528:3–23, 2009.

Bigelow, H. R., Petrey, D. S., Liu, J., Przybylski, D., and Rost, B.: Predicting transmembrane beta-barrels in proteomes. Nucleic Acids Res, 32(8):2566–77, 2004.

Bishop, C. M., Walkenhorst, W. F., and Wimley, W. C.: Folding of beta-sheets in membranes: specificity and promiscuity in peptide model systems. Journal of Molecular Biology, 309(4):975–988, June 2001. PMID: 11399073.

Bishop, C. M., Walkenhorst, W. F., and Wimley, W. C.: Folding of beta-sheets in membranes: specificity and promiscuity in peptide model systems. Journal of Molecular Biology, 309(4):975–988, June 2001. PMID: 11399073.

Bishop, R. E.: Structural biology of membrane-intrinsic beta-barrel enzymes: sentinels of the bacterial outer membrane. Biochim Biophys Acta, 1778(9):1881–96, 2008.

Boda, D., Giri, J., Henderson, D., Eisenberg, B., and Gillespie, D.: Analyzing the components of free energy landscape in a calcium selective ion channel by widoms particle insertion method. Journal of Chemical Physics (submitted), 2010.

Boda, D., Nonner, W., Henderson, D., Eisenberg, B., and Gillespie, D.: Volume exclusion in calcium selective channels. Biophys. J., 94(9):3486–3496, 2008.

Boda, D., Nonner, W., Valisko, M., Henderson, D., Eisenberg, B., and Gillespie, D.: Steric selectivity in na channels arising from protein polarization and mobile side chains. Biophys. J., 93(6):1960–1980, 2007.

Boda, D., Valisko, M., Henderson, D., Eisenberg, B., Gillespie, D., and Nonner, W.: Ionic selectivity in l-type calcium channels by electrostatics and hard-core repulsion. J. Gen. Physiol., 133(5):497–509, 2009.

Borek, D., Cymborowski, M., Machius, M., Minor, W., and Otwinowski, Z.: Diffraction data analysis in the presence of radiation damage. Acta crystallographica. Section D, Biological crystallography, 66(Pt 4):426–436, April 2010. PMID: 20382996.

Borhani, D. W. and Shaw, D. E.: The future of molecular dynamics simulations in drug discovery. Journal of Computer-Aided Molecular Design, 26(1):15–26, January 2012. PMID: 22183577 PMCID: PMC3268975.

Bou, G., Santillana, E., Sheri, A., Beceiro, A., Sampson, J. M., Kalp, M., Bethel, C. R., Distler, A. M., Drawz, S. M., Pagadala, S. R. R., van den Akker, F., Bonomo, R. A., Romero, A., and Buynak, J. D.: Design, synthesis, and crystal structures of 6-alkylidene-2'-substituted penicillanic acid sulfones as potent inhibitors of acinetobacter baumannii OXA-24 carbapenemase. Journal of the American Chemical Society, 132(38):13320–13331, September 2010. PMID: 20822105.

Burgess, N., Dao, T., Stanley, A., and Fleming, K.: Beta-barrel proteins that reside in the escherichia coli outer membrane in vivo demonstrate varied folding behavior in vitro. J Biol Chem, 283(39):26748–26758, Sep 2008.

Burman, J. D., Stevenson, C. E. M., Sawers, R. G., and Lawson, D. M.: The crystal structure of escherichia coli TdcF, a member of the highly conserved YjgF/YER057c/UK114 family. BMC Structural Biology, 7:30, 2007. PMID: 17506874.

Bush, K., Jacoby, G. A., and Medeiros, A. A.: A functional classification scheme for beta-lactamases and its correlation with molecular structure. Antimicrobial agents and chemotherapy, 39(6):1211–1233, June 1995. PMID: 7574506.

Butler, T., Pavlenok, M., Derrington, I., Niederweis, M., and Gundlach, J.: Single-molecule DNA detection with an engineered mspa protein nanopore. Proc Natl Acad Sci U S A, 105(52):20647–20652, Dec 2008.

Chang, C.-C. and Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol., 2(3):27:127:27, May 2011.

Clarke, B.: Selective constraints on amino-acid substitutions during the evolution of proteins. , Published online: 10 October 1970; | doi:10.1038/228159a0, 228(5267):159–160, October 1970.

Cohen, E. J. and Edsall, J.: Proteins, Amino Acids, and Peptides. New York, Reinhold, 1943.

Cojocaru, V., Balali-Mood, K., Sansom, M. S., and Wade, R. C.: Structure and dynamics of the membrane-bound cytochrome p450 2C9. PLoS Comput Biol, 7(8):e1002152, 2011.

Connolly, M. L.: Computation of molecular volume. Journal of the American Chemical Society, 107(5):1118–1124, March 1985.

Cowan, S., Schirmer, T., Rummel, G., Steiert, M., Ghosh, R., Pauptit, R., Jansonius, J., and Rosenbusch, J.: Crystal structures explain functional properties of two e. coli porins. Nature, 358(6389):727–733, Aug 1992.

Crick, F.: Central dogma of molecular biology. Nature, 227(5258):561–563, August 1970. PMID: 4913914.

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E.: WebLogo: a sequence logo generator. Genome research, 14(6):1188–1190, June 2004. PMID: 15173120.

Dayhoff, M. O.: Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, April 1979.

Dayhoff, M., Schwartz, R., and Orcutt, B.: A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure, 5(3):345–352, 1978.

de Cock, H., Hendriks, R., de Vrije, T., and Tommassen, J.: Assembly of an in vitro synthesized escherichia coli outer membrane porin into its stable trimeric configuration. J Biol Chem, 265(8):4646–4651, Mar 1990.

Delcour, A. H.: Function and modulation of bacterial porins: insights from electrophysiology. FEMS Microbiology Letters, 151(2):115–123, 1997.

Delcour, A. H.: Outer membrane permeability and antibiotic resistance. Biochimica Et Biophysica Acta, 1794(5):808–816, May 2009. PMID: 19100346.

Dementin, S., Bouhss, A., Auger, G., Parquet, C., Mengin-Lecreulx, D., Dideberg, O., van Heijenoort, J., and Blanot, D.: Evidence of a functional requirement for a carbamoylated lysine residue in MurD, MurE and MurF synthetases as established by chemical rescue experiments. European Journal of Biochemistry / FEBS, 268(22):5800–5807, November 2001. PMID: 11722566.

Dixon, M. and Webb, E. C.: Enzymes. New York, Academic Press, 1979.

Doi, M.: Gel dynamics. Journal of the Physical Society of Japan, 78:052001, 2009.

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., and Liang, J.: CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res, 34(suppl 2):W116–W118, 2006.

Durand-Vidal, S., Simonin, J., and Turq, P.: Electrolytes at Interfaces. Boston, Kluwer, 2000.

Edelsbrunner, H., Facello, M., Fu, P., and Liang, J.: Measuring proteins and voids in proteins. In Proceedings of the Twenty-Eighth Hawaii International Conference on System Sciences, 1995, volume 5, pages 256–264 vol.5. IEEE, January 1995.

Edelsbrunner, H., Facello, M., and Liang, J.: On the definition and the construction of pockets in macromolecules. Discrete Applied Mathematics, 88:83–102, 1998.

Edsall, J. and Wyman, J.: Biophysical Chemistry. NY, Academic Press, 1958.

Eisenberg, B.: Lifes solutions are not ideal. Posted on arXiv.org with Paper ID arXiv:1105.0184v1, 2011.

Eisenberg, B.: Living transistors: a physicists view of ion channels. available on http://arxiv.org/ as q-bio/0506016v2 24 pages, 2005.

Eisenberg, B.: Multiple scales in the simulation of ion channels and proteins. The Journal of Physical Chemistry C, 114(48):20719–20733, 2010.

Eisenberg, B.: Crowded charges in ion channels. In Advances in Chemical Physics, pages 77–223 also available at http:\arix.org as arXiv 1009.1786v1. John Wiley & Sons, Inc., 2011.

Eisenberg, B.: Mass action in ionic solutions. Chemical Physics Letters, 511, 2011.

Eisenberg, B.: Ions in fluctuating channels: Transistors alive. Fluctuations and Noise Letters, (in the press):Earlier version 'Living Transistors: a Physicists View of Ion Channels' available on http://arxiv.org/ as q–bio/0506016v2, 2012.

Eisenberg, R.: Channels as enzymes: Oxymoron and tautology. Journal of Membrane Biology, 115:112. Available on arXiv as http://arxiv.org/abs/1112.2363, 1990.

Eisenberg, R.: Atomic biology, electrostatics and ionic channels. In New Developments and Theoretical Studies of Proteins, ed. R. Elber, volume 7, pages 269–357. Published in the Physics ArXiv as arXiv:0807.0715. Philadelphia, World Scientific, 1996.

Eisenberg, R.: Computing the field in proteins and channels. J. Membrane Biol., 150:125. Also available on http:\arxiv.org as arXiv 1009.2857, 1996.

Eisenhaber, B. and Eisenhaber, F.: Prediction of posttranslational modification of proteins from their amino acid sequence. Methods in molecular biology (Clifton, N.J.), 609:365–384, 2010. PMID: 20221930.

Ellinor, P., Yang, J., Sather, W., Zhang, J., and Tsien, R.: Ca2+ channel selectivity at a single locus for high-affinity ca2+ interactions. Neuron, 15:1121–1132, 1995.

Emsley, P. and Cowtan, K.: Coot: model-building tools for molecular graphics. Acta crystallographica. Section D, Biological crystallography, 60(Pt 12 Pt 1):2126–2132, December 2004. PMID: 15572765.

Engelman, D. M., Steitz, T. A., and Goldman, A.: Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annual Review of Biophysics and Biophysical Chemistry, 15:321–353, 1986. PMID: 3521657.

Epstein, C. J.: Non-randomness of amino-acid changes in the evolution of homologous proteins. Nature, 215(5099):355–359, July 1967. PMID: 4964553.

Eyring, H.: The activated complex and the absolute rate of chemical reactions. Journal of Chemical Physics, 17:67–77, 1935.

Fawcett, W. R.: Liquids, Solutions, and Interfaces: From Classical Macroscopic Descriptions to Modern Microscopic Details. New York, Oxford University Press, 2004.

Felsenstein, J.: The number of evolutionary trees. Systematic Biology, 27(1):27–33, March 1978.

Fersht, A.: Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. W. H. Freeman, 1st edition, September 1998.

Fischer, K., Weber, A., Brink, S., Arbinger, B., Schnemann, D., Borchert, S., Heldt, H. W., Popp, B., Benz, R., and Link, T. A.: Porins from plants. molecular cloning and functional characterization of two new members of the porin family. The Journal of Biological Chemistry, 269(41):25754–25760, October 1994. PMID: 7523392.

Fitch, W. M. and Margoliash, E.: Construction of phylogenetic trees. Science (New York, N.Y.), 155(3760):279–284, January 1967. PMID: 5334057.

Fleishman, S., Whitehead, T., Strauch, E., Corn, J., Qin, S., Zhou, H., Mitchell, J., Demerdash, O., Takeda-Shitaka, M., Terashi, G., Moal, I., Li, X., Bates, P., Zacharias, M., Park, H., Ko, J., Lee, H., Seok, C., Bourquard, T., Bernauer, J., Poupon, A., Aze, J., Soner, S., Ovali, S., Ozbek, P., Tal, N., Haliloglu, T., Hwang, H., Vreven, T., Pierce, B., Weng, Z., Perez-Cano, L., Pons, C., Fernandez-Recio, J., Jiang, F., Yang, F., Gong, X., Cao, L., Xu, X., Liu, B., Wang, P., Li, C., Wang, C., Robert, C., Guharoy, M., Liu, S., Huang, Y., Li, L., Guo, D., Chen, Y., Xiao, Y., London, N., Itzhaki, Z., Schueler-Furman, O., Inbar, Y., Patapov, V., Cohen, M., Schreiber, G., Tsuchiya, Y., Kanamori, E., Standley, D., Nakamura, H., Kinoshita, K., Driggers, C., Hall, R., Morgan, J., Hsu, V., Zhan, J., Yang, Y., Zhou, Y., Kastritis, P., Bonvin, A., Zhang, W., Camacho, C., Kilambi, K., Sircar, A., Gray, J., Ohue, M., Uchikoga, N., Matsuzaki, Y., Ishida, T., Akiyama, Y., Khashan, R., Bush, S., Fouches, D., Tropsha, A., Esquivel-Rodriguez, J., Kihara, D., Stranges, P., Jacak, R., Kuhlman, B., Huang, S., Zou, X., Wodak, S., Janin, J., and Baker, D.: Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol, Sep 2011.

Fraenkel, D.: Monoprotic mineral acids analyzed by the Smaller-Ion shell model of strong electrolyte solutions. The Journal of Physical Chemistry B, 115(3):557–568, 2010.

Fraenkel, D.: Simplified electrostatic model for the thermodynamic excess potentials of binary strong electrolyte solutions with size-dissimilar ions. Molecular Physics, 108(11):1435 – 1466, 2010.

Friedman, H. L.: Electrolyte solutions at equilibrium. Annual Review of Physical Chemistry, 32(1):179–204, 1981.

Garman, E. F.: Radiation damage in macromolecular crystallography: what is it and why should we care? Acta Crystallographica Section D: Biological Crystallography, 66(Pt 4):339–351, April 2010. PMID: 20382986 PMCID: PMC2852297.

Gibbs, A. J. and McIntyre, G. A.: The diagram, a method for comparing sequences. its use with amino acid and nucleotide sequences. European journal of biochemistry / FEBS, 16(1):1–11, September 1970. PMID: 5456129.

Gillespie, D., Giri, J., and Fill, M.: Reinterpreting the anomalous mole fraction effect. the ryanodine receptor case study. Biophyiscal Journal, 97(8):pp. 2212 – 2221, 2009.

Golemi, D., Maveyraud, L., Vakulenko, S., Samama, J. P., and Mobashery, S.: Critical involvement of a carbamylated lysine in catalytic function of class d beta-lactamases. Proceedings of the National Academy of Sciences of the United States of America, 98(25):14280–14285, December 2001. PMID: 11724923.

Greer, E. L. and Shi, Y.: Histone methylation: a dynamic mark in health, disease and inheritance. Nature reviews. Genetics, 13(5):343–357, May 2012. PMID: 22473383.

Griep, S. and Hobohm, U.: PDBselect 1992-2009 and PDBfilter-select. Nucleic acids research, 38(Database issue):D318–319, January 2010. PMID: 19783827.

Gutteridge, A. and Thornton, J. M.: Understanding nature's catalytic toolkit. Trends in Biochemical Sciences, 30(11):622–629, November 2005. PMID: 16214343.

Hall, P. R., Zheng, R., Antony, L., Pusztai-Carey, M., Carey, P. R., and Yee, V. C.: Transcarboxylase 5S structures: assembly and catalytic mechanism of a multienzyme complex subunit. The EMBO journal, 23(18):3621–3631, September 2004. PMID: 15329673.

Hansen, J. and McDonald, I. R.: Theory of Simple Liquids. New York, Academic Press, third edition edition, 2006.

Harris, T. K. and Mildvan, A. S.: High-precision measurement of hydrogen bond lengths in proteins by nuclear magnetic resonance methods. Proteins, 35(3):275–282, May 1999. PMID: 10328262.

Hartley, H.: Origin of the word Protein. , Published online: 11 August 1951; | doi:10.1038/168244a0, 168(4267):244–244, August 1951.

Hearn, E. M., Patel, D. R., Lepore, B. W., Indic, M., and van den Berg, B.: Transmembrane passage of hydrophobic compounds through a protein channel wall. Nature, 458(7236):367–370, March 2009.

Helfferich, F.: Ion Exchange. New York, McGraw Hill reprinted by Dover, 1962.

Henikoff, S. and Henikoff, J. G.: Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A, 89(22):10915–9, 1992.

Hibbert, F. and Emsley, J.: Hydrogen bonding and chemical reactivity. In Advances in Physical Organic Chemistry, ed. D. Bethell, volume Volume 26, pages 255–379. Academic Press, 1991.

Hill, K., Model, K., Ryan, M. T., Dietmeier, K., Martin, F., Wagner, R., and Pfanner, N.: Tom40 forms the hydrophilic channel of the mitochondrial import pore for preproteins. Nature, 395(6701):516–521, October 1998.

Hofmeister, F.: On the structure and grouping of the protein bodies. Ergebnisse der Physiologie, 1(759), 1902.

Hovarth, A. L.: Handbook of aqueous electrolyte solutions: physical properties, estimation, and correlation methods. New York, Ellis Horwood,, 1985.

Howe, R. T. and Sodini, C. G.: Microelectronics: an integrated approach. Upper Saddle River, NJ USA, Prentice Hall, 1997.

Hur, S. and Bruice, T. C.: Just a near attack conformer for catalysis (Chorismate to prephenate rearrangements in water, antibody, enzymes, and their mutants). J. Am. Chem. Soc., 125(35):10540–10542, 2003.

Huysmans, G. H. M., Baldwin, S. A., Brockwell, D. J., and Radford, S. E.: The transition state for folding of an outer membrane protein. Proceedings of the National Academy of Sciences of the United States of America, 107(9):4099–4104, March 2010. PMID: 20133664.

Hyon, Y., Kwak, D. Y., and Liu, C.: Energetic variational approach in complex fluids : Maximum dissipation principle. Discrete and Continuous Dynamical Systems (DCDS-A),

26(4: April):1291 – 1304, available at URL: http://www.ima.umn.edu as IMA Preprint Series # 2228, 2010.

Hnenberger, P. H. and Reif, M.: Single-Ion Solvation. Cambridge UK, RSC Publishing, 2011.

Im, H., Sharpe, M. L., Strych, U., Davlieva, M., and Krause, K. L.: The crystal structure of alanine racemase from streptococcus pneumoniae, a target for structure-based drug design. BMC microbiology, 11:116, 2011. PMID: 21612658.

Jackups, R., Cheng, S., and Liang, J.: Sequence motifs and antimotifs in beta-barrel membrane proteins from a genome-wide analysis: the Ala-Tyr dichotomy and chaperone binding motifs. J Mol Biol, 363(2):611–23, 2006.

Jackups, R. and Liang, J.: Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. Journal of Molecular Biology, 354(4):979–993, December 2005. PMID: 16277990.

Jensen, O. N.: Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Current opinion in chemical biology, 8(1), February 2004. PMID: 15036154.

Kalyuzhnyi, Y. V., Vlachy, V., and Dill, K. A.: Aqueous alkali halide solutions: can osmotic coefficients be explained on the basis of the ionic sizes alone? Physical Chemistry Chemical Physics, 12(23):6260–6266, 2010.

Karlin, S. and Altschul, S. F.: Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A, 87(6):2264–8, 1990.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C.: A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature, 181(4610):662–666, March 1958. PMID: 13517261.

Kim, K., Kim, M.-I., Chung, J., Ahn, J.-H., and Rhee, S.: Crystal structure of metal-dependent allantoinase from escherichia coli. Journal of molecular biology, 387(5):1067–1074, April 2009. PMID: 19248789.

Kimura, M.: The rate of molecular evolution considered from the standpoint of population genetics. Proceedings of the National Academy of Sciences of the United States of America, 63(4):1181–1188, August 1969. PMID: 5260917.

Kimura, M.: The Neutral Theory of Molecular Evolution. Cambridge University Press, reprint edition, February 1985.

Kimura, M. and Ohta, T.: On some principles governing molecular evolution*. Proceedings of the National Academy of Sciences of the United States of America, 71(7):2848–2852, July 1974. PMID: 4527913 PMCID: PMC388569.

Kitano, K., Maeda, N., Fukui, T., Atomi, H., Imanaka, T., and Miki, K.: Crystal structure of a novel-type archaeal rubisco with pentagonal symmetry. Structure, 9(6):473–481, June 2001.

Kleinschmidt, J. H., den Blaauwen, T., Driessen, A. J., and Tamm, L. K.: Outer membrane protein a of escherichia coli inserts and folds into lipid bilayers by a concerted mechanism. Biochemistry, 38(16):5006–5016, April 1999. PMID: 10213603.

Kleywegt, G. J., Harris, M. R., Zou, J. Y., Taylor, T. C., Whlby, A., and Jones, T. A.: The uppsala Electron-Density server. Acta crystallographica. Section D, Biological crystallography, 60(Pt 12 Pt 1):2240–2249, December 2004. PMID: 15572777.

Koch, S. E., Bodi, I., Schwartz, A., and Varadi, G.: Architecture of ca(2+) channel pore-lining segments revealed by covalent modification of substituted cysteines. J Biol Chem, 275(44):34493–34500, 2000.

Kokubo, H. and Pettitt, B. M.: Preferential solvation in urea solutions at different concentrations: properties from simulation studies. The journal of physical chemistry. B, 111(19):5233–42, 2007.

Kokubo, H., Rosgen, J., Bolen, D. W., and Pettitt, B. M.: Molecular basis of the apparent near ideality of urea solutions. Biophys. J., 93(10):3392–3407, 2007.

Kontogeorgis, G. M. and Folas, G. K.: Thermodynamic Models for Industrial Applications: From Classical and Advanced Mixing Rules to Association Theories. John Wiley & Sons, Ltd, 2009.

Kornyshev, A. A.: Double-Layer in ionic liquids: Paradigm change? J. Phys. Chem. B, 111(20):5545–5557, 2007.

Kortemme, T. and Baker, D.: Computational design of protein-protein interactions. Curr Opin Chem Biol, 8(1):91–97, Feb 2004.

Krzywicki, A. and Slonimski, P. P.: Formal analysis of protein sequences. i. specific long-range constraints in pair associations of amino acids. Journal of theoretical biology, 17(1):136–158, October 1967. PMID: 6055367.

Kumar, S. and Nussinov, R.: Close-range electrostatic interactions in proteins. Chembiochem: a European journal of chemical biology, 3(7):604–617, July 2002. PMID: 12324994.

Kunz, W.: Specific Ion Effects. Singapore, World Scientific, 2009.

Kyte, J.: Mechanism in Protein Chemistry. New York, Garland, 1995.

Larbig, M., Mansouri, E., Freihorst, J., Tmmler, B., Khler, G., Domdey, H., Knapp, B., Hungerer, K. D., Hundt, E., Gabelsberger, J., and von Specht, B. U.: Safety and immunogenicity of an intranasal pseudomonas aeruginosa hybrid outer membrane protein F-I vaccine in human volunteers. Vaccine, 19(17-19):2291–2297, March 2001. PMID: 11257350.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G.: Clustal w and clustal x version 2.0. Bioinformatics, 23(21):2947–8, 2007.

Lee, L. L.: Molecular Thermodynamics of Electrolyte Solutions. Singapore, World Scientific, 2008.

LeJeune, K. E., Wild, J. R., and Russell, A. J.: Nerve agents degraded by enzymatic foams. Nature, 395(6697):27–28, September 1998. PMID: 9738495.

LeMagueres, P., Im, H., Dvorak, A., Strych, U., Benedik, M., and Krause, K. L.: Crystal structure at 1.45 a resolution of alanine racemase from a pathogenic bacterium, pseudomonas aeruginosa, contains both internal and external aldimine forms. Biochemistry, 42(50):14752–14761, December 2003. PMID: 14674749.

Levinthal, C.: How to fold graciously. Mossbauer Spectroscopy in Biological Systems, pages 22–24, 1969.

Li, B.: Continuum electrostatics for ionic solutions with non-uniform ionic sizes. Nonlinearity, 22(4):811, 2009.

Li, W. and Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (Oxford, England), 22(13):1658–1659, July 2006. PMID: 16731699.

Li, Y., Yu, X., Ho, J., Fushman, D., Allewell, N. M., Tuchman, M., and Shi, D.: Reversible post-translational carboxylation modulates the enzymatic activity of N-acetyl-L-ornithine transcarbamylase. Biochemistry, 49(32):6887–6895, August 2010. PMID: 20695527.

Liang, J., Edelsbrunner, H., and Woodward, C.: Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Science, 7:18841897, 1998.

Liang, J., Naveed, H., Jimenez-Morales, D., Adamian, L., and Lin, M.: Computational studies of membrane proteins: Models and predictions for biological understanding. Biochimica Et Biophysica Acta, 1818(4):927–941, April 2012. PMID: 22051023.

Liolios, K., Chen, I. A., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V. M., and Kyrpides, N. C.: The genomes on line database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Research, 38(Database issue):D346–354, January 2010. PMID: 19914934.

Lipman, D. J. and Pearson, W. R.: Rapid and sensitive protein similarity searches. Science (New York, N.Y.), 227(4693):1435–1441, March 1985. PMID: 2983426.

Liu, C.: An introduction of elastic complex fluids: An energetic variational approach. In Multi-scale Phenomena in Complex Fluids: Modeling, Analysis and Numerical Simulations, eds. T. Hou, C. Liu, and J. Liu. Singapore, World Scientific Publishing Company, 2009.

Li, P. and Goldman, N.: Models of molecular evolution and phylogeny. Genome Research, 8(12):1233–1244, December 1998. PMID: 9872979.

Lorimer, G. H., Badger, M. R., and Andrews, T. J.: The activation of ribulose-1,5-bisphosphate carboxylase by carbon dioxide and magnesium ions. equilibria, kinetics, a suggested mechanism, and physiological implications. Biochemistry, 15(3):529–536, February 1976. PMID: 3199.

Ludemann, S. K., Lounnas, V., and Wade, R. C.: How do substrates enter and products exit the buried active site of cytochrome p450cam? 2. steered molecular dynamics and adiabatic mapping of substrate pathways. J Mol Biol, 303(5):813–30, 2000.

Majd, S., Yusko, E., Billeh, Y., Macrae, M., Yang, J., and Mayer, M.: Applications of biological pores in nanomedicine, sensing, and nanoelectronics. Curr Opin Biotechnol, 21(4):439–476, Aug 2010.

Mann, M. and Jensen, O. N.: Proteomic analysis of post-translational modifications. Nature Biotechnology, 21(3):255–261, March 2003.

Markowich, P. A., Ringhofer, C. A., and Schmeiser, C.: Semiconductor Equations. New York, Springer-Verlag, 1990.

McCleskey, E. W.: Ion channel selectivity using an electric stew. Biophys J, 79(4):1691–2, 2000.

Meng, G., Fronzes, R., Chandran, V., Remaut, H., and Waksman, G.: Protein oligomerization in the bacterial outer membrane (review). Mol Membr Biol, 26(3):136–145, Apr 2009.

Meulenbroek, E. M., Paspaleva, K., Thomassen, E. A. J., Abrahams, J. P., Goosen, N., and Pannu, N. S.: Involvement of a carboxylated lysine in UV damage endonuclease. Protein Science: A Publication of the Protein Society, 18(3):549–558, March 2009. PMID: 19241382.

Miedema, H., Meter-Arkema, A., Wierenga, J., Tang, J., Eisenberg, B., Nonner, W., Hektor, H., Gillespie, D., and Meijberg, W.: Permeation properties of an engineered bacterial OmpF porin containing the EEEE-locus of ca2+ channels. Biophys J, 87(5):3137–47, 2004.

Miedema, H., Vrouenraets, M., Wierenga, J., Gillespie, D., Eisenberg, B., Meijberg, W., and Nonner, W.: Ca2+ selectivity of a chemically modified OmpF with reduced pore volume. Biophys J, 91(12):4392–400, 2006.

Mobley, H. L., Island, M. D., and Hausinger, R. P.: Molecular biology of microbial ureases. Microbiological reviews, 59(3):451–480, September 1995. PMID: 7565414.

Molloy, M. P., Herbert, B. R., Slade, M. B., Rabilloud, T., Nouwens, A. S., Williams, K. L., and Gooley, A. A.: Proteomic analysis of the escherichia coli outer membrane. European Journal of Biochemistry / FEBS, 267(10):2871–2881, May 2000. PMID: 10806384.

Morollo, A. A., Petsko, G. A., and Ringe, D.: Structure of a michaelis complex analogue: propionate binds in the substrate carboxylate site of alanine racemase,. Biochemistry, 38(11):3293–3301, March 1999.

Muller, T., Rahmann, S., and Rehmsmeier, M.: Non-symmetric score matrices and the detection of homologous transmembrane proteins. Bioinformatics, 17 Suppl 1:S182–9, 2001.

Naveed, H., Jackups, R., and Liang, J.: Predicting weakly stable regions, oligomerization state, and protein-protein interfaces in transmembrane domains of outer membrane proteins. Proceedings of the National Academy of Sciences of the United States of America, 106(31):12735–12740, August 2009. PMID: 19622743.

Naveed, H., Jimenez-Morales, D., Tian, J., Pasupuleti, V., Kenney, L. J., and Liang, J.: Engineered oligomerization state of OmpF protein through computational design decouples oligomer dissociation from unfolding. Journal of molecular biology, 419(1-2):89–101, May 2012. PMID: 22391420.

Needleman, S. B. and Wunsch, C. D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48(3):443–453, March 1970. PMID: 5420325.

Neidhart, D., Wei, Y., Cassidy, C., Lin, J., Cleland, W. W., and Frey, P. A.: Correlation of low-barrier hydrogen bonding and oxyanion binding in transition state analogue complexes of chymotrypsin. Biochemistry, 40(8):2439–2447, February 2001. PMID: 11327865.

Nelson, D. L. and Cox, M. M.: Lehninger Principles of Biochemistry, Fourth Edition. W. H. Freeman, fourth edition edition, April 2004.

Ng, P. C., Henikoff, J. G., and Henikoff, S.: PHAT: a transmembrane-specific substitution matrix. predicted hydrophobic and transmembrane. Bioinformatics, 16(9):760–6, 2000.

Nishitani, Y., Yoshida, S., Fujihashi, M., Kitagawa, K., Doi, T., Atomi, H., Imanaka, T., and Miki, K.: Structure-based catalytic optimization of a type III rubisco from a hyperthermophile. Journal of Biological Chemistry, 285(50):39339–39347, December 2010.

Noland, B. W., Newman, J. M., Hendle, J., Badger, J., Christopher, J. A., Tresser, J., Buchanan, M. D., Wright, T. A., Rutter, M. E., Sanderson, W. E., Mller-Dieckmann, H. J., Gajiwala, K. S., and Buchanan, S. G.: Structural studies of salmonella typhimurium ArnB (PmrH) aminotransferase: a 4-amino-4-deoxy-l-arabinose lipopolysaccharide-modifying enzyme. Structure (London, England: 1993), 10(11):1569–1580, November 2002. PMID: 12429098.

Nonner, W., Gillespie, D., Henderson, D., and Eisenberg, B.: Ion accumulation in a biological calcium channel: effects of solvent and confining pressure. J Physical Chemistry B, 105:6427–6436, 2001.

Novotn, J.: Genealogy of immunoglobulin polypeptide chains: a consequence of amino acid interactions, conserved in their tertiary structures. Journal of theoretical biology, 41(1):171–180, September 1973. PMID: 4754904.

Ohno, S.: Evolution by gene duplication. Springer-Verlag, 1St edition edition, 1970.

Ohta, T. and Kimura, M.: Functional organization of genetic material as a product of molecular evolution. Nature, 233(5315):118–119, September 1971. PMID: 16063236.

Otyepka, M., Skopalk, J., Anzenbacherov, E., and Anzenbacher, P.: What common structural features and variations of mammalian p450s are known to date? Biochimica Et Biophysica Acta, 1770(3):376–389, March 2007. PMID: 17069978.

Ouzounis, C. A.: Rise and demise of bioinformatics? promise and progress. PLoS Comput Biol, 8(4):e1002487, April 2012.

Ouzounis, C. A. and Valencia, A.: Early bioinformatics: the birth of a discipline–a personal view. Bioinformatics (Oxford, England), 19(17):2176–2190, November 2003. PMID: 14630646.

Overington, J., Donnelly, D., Johnson, M. S., Sali, A., and Blundell, T. L.: Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Science: A Publication of the Protein Society, 1(2):216–226, February 1992. PMID: 1304904.

Overington, J., Johnson, M. S., Sali, A., and Blundell, T. L.: Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proceedings. Biological Sciences / The Royal Society, 241(1301):132–145, August 1990. PMID: 1978340.

Pain, R. H. and Robson, B.: Analysis of the code relating sequence to secondary structure in proteins. , Published online: 04 July 1970; | doi:10.1038/227062a0, 227(5253):62–63, July 1970.

Park, I. S. and Hausinger, R. P.: Requirement of carbon dioxide for in vitro assembly of the urease nickel metallocenter. Science (New York, N.Y.), 267(5201):1156–1158, February 1995. PMID: 7855593.

Paschen, S. A., Waizenegger, T., Stan, T., Preuss, M., Cyrklaff, M., Hell, K., Rapaport, D., and Neupert, W.: Evolutionary conservation of biogenesis of [beta]-barrel membrane proteins. Nature, 426(6968):862–866, December 2003.

Patwardhan, V. S. and Kumar, A.: Thermodynamic properties of aqueous solutions of mixed electrolytes: A new mixing rule. AIChE Journal, 39(4):711–714, 1993.

Pauling, L. and Corey, R. B.: Configurations of polypeptide chains with favored orientations around single bonds. Proceedings of the National Academy of Sciences of the United States of America, 37(11):729–740, November 1951. PMID: 16578412 PMCID: PMC1063460.

Pearson, W. R. and Lipman, D. J.: Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A, 85(8):2444–8, 1988.

Perutz, M. F.: An optical method for finding the molecular orientation in different forms of crystalline haemoglobin; changes in dichroism accompanying oxygenation and reduction. Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain), 141(902):69–71, March 1953. PMID: 13047271.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., and North, A. C.: Structure of haemoglobin: a three-dimensional fourier synthesis at 5.5-a. resolution, obtained by x-ray analysis. Nature, 185(4711):416–422, February 1960. PMID: 18990801.

Pierret, R.: Semiconductor Device Fundamentals. New York, Addison Wesley, 1996.

Pineda, J. R. and Schwartz, S. D.: Protein dynamics and catalysis: the problems of transition state theory and the subtlety of dynamic control. Philosophical Transactions of the Royal Society B: Biological Sciences, 361(1472):1433–1438, August 2006.

Pitzer, K. S.: Thermodynamics. New York, McGraw Hill, 3rd edition, 1995.

Pohl, H. A.: Dielectrophoresis: The Behavior of Neutral Matter in Nonuniform Electric Fields. New York, Cambridge University Press, 1978.

Poirel, L., Naas, T., and Nordmann, P.: Diversity, epidemiology, and genetics of class d beta-lactamases. Antimicrobial agents and chemotherapy, 54(1):24–38, January 2010. PMID: 19721065.

Porter, C. T., Bartlett, G. J., and Thornton, J. M.: The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Research, 32(Database issue):D129–133, January 2004. PMID: 14681376.

Pruitt, K. D., Tatusova, T., and Maglott, D. R.: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 35(Database issue):D61–65, January 2007. PMID: 17130148.

Ptitsyn, O. B.: Statistical analysis of the distribution of amino acid residues among helical and non-helical regions in globular proteins. Journal of molecular biology, 42(3):501–510, June 1969. PMID: 5804157.

Pytkowicz, R. M.: Activity Coefficients in Electrolyte Solutions, volume 1. Boca Raton FL USA, CRC, 1979.

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J. M., Talwalkar, A. S., Repo, S., Souza, M. L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., Fang, H., Gough, J., Koskinen, P., Trnen, P., Nokso-Koivisto, J., Holm, L., Cozzetto, D., Buchan, D. W. A., Bryson, K., Jones, D. T., Limaye, B., Inamdar, H., Datta, A., Manjari, S. K., Joshi, R., Chitale, M., Kihara, D., Lisewski, A. M., Erdin, S., Venner, E., Lichtarge, O., Rentzsch, R., Yang, H., Romero, A. E., Bhat, P., Paccanaro, A., Hamp, T., Kaner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron, M., Hnigschmid, P., Hopf, T. A., Kaufmann, S., Kiening, M., Krompass, D., Landerer, C., Mahlich, Y., Roos, M., Bjrne, J., Salakoski, T., Wong, A., Shatkay, H., Gatzmann, F., Sommer, I., Wass, M. N., Sternberg, M. J. E., kunca, N., Supek, F., Bonjak, M., Panov, P., Deroski, S., muc, T., Kourmpetis, Y. A. I., Dijk, A. D. J. v., Braak, C. J. F. t., Zhou, Y., Gong, Q., Dong, X., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Camillo, B. D., Toppo, S., Lan, L., Djuric, N., Guo, Y., Vucetic, S., Bairoch, A., Linial, M., Babbitt, P. C., Brenner, S. E., Orengo, C., Rost, B., Mooney, S. D., and Friedberg, I.: A large-scale evaluation of computational protein function prediction. Nature Methods, 2013.

Radzicka, A. and Wolfenden, R.: Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemistry, 27(5):1664–1670, March 1988.

Randall, A., Cheng, J., Sweredoski, M., and Baldi, P.: TMBpro: secondary structure, beta-contact and tertiary structure prediction of transmembrane beta-barrel proteins. Bioinformatics (Oxford, England), 24(4):513–520, February 2008. PMID: 18006547.

Ravelli, R. B. and McSweeney, S. M.: The 'fingerprint' that x-rays can leave on structures. Structure (London, England: 1993), 8(3):315–328, March 2000. PMID: 10745008.

Reid, J., Fung, H., Gehring, K., Klebba, P., and Nikaido, H.: Targeting of porin to the outer membrane of escherichia coli. rate of trimer assembly and identification of a dimer intermediate. J Biol Chem, 263(16):7753–7759, Jun 1988.

Remmert, M., Linke, D., Lupas, A. N., and Soding, J.: HHomp–prediction and classification of outer membrane proteins. Nucleic Acids Res, 2009.

Sagurthi, S. R., Gowda, G., Savithri, H. S., and Murthy, M. R. N.: Structures of mannose-6-phosphate isomerase from salmonella typhimurium bound to metal atoms and substrate: implications for catalytic mechanism. Acta crystallographica. Section D, Biological crystallography, 65(Pt 7):724–732, July 2009. PMID: 19564693.

Sambrook, J., Fritsch, E. F., and Maniatis, T.: Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Pr, 2nd edition, December 1989.

Sanger, F. and Tuppy, H.: The amino-acid sequence in the phenylalanyl chain of insulin. 1. the identification of lower peptides from partial hydrolysates. Biochemical Journal, 49(4):463–481, September 1951. PMID: 14886310 PMCID: PMC1197535.

Sanger, F. and Tuppy, H.: The amino-acid sequence in the phenylalanyl chain of insulin. 2. the investigation of peptides from enzymic hydrolysates. Biochemical Journal, 49(4):481–490, September 1951. PMID: 14886311 PMCID: PMC1197536.

Santillana, E., Beceiro, A., Bou, G., and Romero, A.: Crystal structure of the carbapenemase OXA-24 reveals insights into the mechanism of carbapenem hydrolysis. Proceedings of the National Academy of Sciences of the United States of America, 104(13):5354–5359, March 2007. PMID: 17374723.

Saranya, N. and Selvaraj, S.: Variation of protein binding cavity volume and ligand volume in protein-ligand complexes. Bioorganic & Medicinal Chemistry Letters, 19(19):5769–5772, October 2009. PMID: 19706358.

Sather, W. A. and McCleskey, E. W.: Permeation and selectivity in calcium channels. Annu Rev Physiol, 65:133–59, 2003.

Sattath, S. and Tversky, P. A.: Additive similarity trees. Psychometrika, 42(3):319–345, September 1977.

Schulz, G. E.: Bacterial porins: structure and function. Current Opinion in Cell Biology, 5(4):701–707, August 1993. PMID: 8257610.

Schulz, G. E.: beta-Barrel membrane proteins. Current Opinion in Structural Biology, 10(4):443–447, August 2000. PMID: 10981633.

Segel, I. H.: Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems. New York, Wiley: Interscience, enzyme kinetics: behavior and analysis of rapid equilibrium and Steady-State enzyme systems edition, 1993.

Seibert, C. M. and Raushel, F. M.: Structural and catalytic diversity within the amidohydrolase superfamily. Biochemistry, 44(17):6383–6391, May 2005. PMID: 15850372.

Seibert, C. M. and Raushel, F. M.: Structural and catalytic diversity within the amidohydrolase superfamily. Biochemistry, 44(17):6383–6391, May 2005. PMID: 15850372.

Seo, J. and Lee, K.: Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. Journal of Biochemistry and Molecular Biology, 37(1):35–44, January 2004. PMID: 14761301.

Sharabi, O., Yanover, C., Dekel, A., and Shifman, J.: Optimizing energy functions for protein-protein interface design. J Comput Chem, 32(1):23–32, Jan 2011.

Sheng, P., Zhang, J., and Liu, C.: Onsager principle and electrorheological fluid dynamics. Progress of Theoretical Physics Supplement No. 175, pages 131–143, 2008.

Shi, D., Yu, X., Roth, L., Morizono, H., Tuchman, M., and Allewell, N. M.: Structures of n-acetylornithine transcarbamoylase from xanthomonas campestris complexed with substrates and substrate analogs imply mechanisms for substrate binding and catalysis. Proteins, 64(2):532–542, August 2006. PMID: 16741992.

Siegler, W. C., Crank, J. A., Armstrong, D. W., and Synovec, R. E.: Increasing selectivity in comprehensive three-dimensional gas chromatography via an ionic liquid stationary phase column in one dimension. J Chromatogr A, 1217(18):3144–9, 2010.

Silverman, R. B.: The potential use of mechanism-based enzyme inactivators in medicine. Journal of enzyme inhibition, 2(2):73–90, 1988. PMID: 3069967.

Simamura, E., Shimada, H., Hatta, T., and Hirai, K.: Mitochondrial voltage-dependent anion channels (VDACs) as novel pharmacological targets for anti-cancer agents. Journal of Bioenergetics and Biomembranes, 40(3):213–217, June 2008. PMID: 18704666.

Simoni, R. D., Hill, R. L., and Vaughan, M.: Urease, the first crystalline enzyme and the proof that enzymes are proteins: the work of james b. sumner. Journal of Biological Chemistry, 277(35):e23–e23, August 2002.

Smith, T. F. and Waterman, M. S.: Identification of common molecular subsequences. J Mol Biol, 147(1):195–7, 1981.

Song, L., Hobaugh, M. R., Shustak, C., Cheley, S., Bayley, H., and Gouaux, J. E.: Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. Science (New York, N.Y.), 274(5294):1859–1866, December 1996. PMID: 8943190.

Spohr, H. V. and Patey, G. N.: Structural and dynamical properties of ionic liquids: Competing influences of molecular properties. J Chem Phys, 132(15):154504–12, 2010.

Srinivas, N., Jetter, P., Ueberbacher, B. J., Werneburg, M., Zerbe, K., Steinmann, J., Van der Meijden, B., Bernardini, F., Lederer, A., Dias, R. L. A., Misson, P. E., Henze, H., Zumbrunn, J., Gombert, F. O., Obrecht, D., Hunziker, P., Schauer, S., Ziegler, U., Kch, A., Eberl, L., Riedel, K., DeMarco, S. J., and Robinson, J. A.: Peptidomimetic antibiotics target outer-membrane biogenesis in pseudomonas aeruginosa. Science (New York, N.Y.), 327(5968):1010–1013, February 2010. PMID: 20167788.

St Maurice, M., Reinhardt, L., Surinya, K. H., Attwood, P. V., Wallace, J. C., Cleland, W. W., and Rayment, I.: Domain architecture of pyruvate carboxylase, a biotin-dependent multifunctional enzyme. Science (New York, N.Y.), 317(5841):1076–1079, August 2007. PMID: 17717183.

Stec, B.: Structural mechanism of RuBisCO activation by carbamylation of the active site lysine. Proceedings of the National Academy of Sciences of the United States of America, 109(46):18785–18790, November 2012. PMID: 23112176.

Strych, U., Huang, H. C., Krause, K. L., and Benedik, M. J.: Characterization of the alanine racemases from pseudomonas aeruginosa PAO1. Current microbiology, 41(4):290–294, October 2000. PMID: 10977898.

Sulzenbacher, G., Bignon, C., Nishimura, T., Tarling, C. A., Withers, S. G., Henrissat, B., and Bourne, Y.: Crystal structure of thermotoga maritima alpha-L-fucosidase. insights into

the catalytic mechanism and the molecular basis for fucosidosis. The Journal of Biological Chemistry, 279(13):13119–13128, March 2004. PMID: 14715651.

Sumner, J. B.: The isolation and crystallization of the enzyme urease preliminary paper. Journal of Biological Chemistry, 69(2):435–441, August 1926.

Surrey, T., Schmid, A., and Jahnig, F.: Folding and membrane insertion of the trimeric beta-barrel protein ompf. Biochemistry, 35(7):2283–2288, Feb 1996.

Sze, S.: Physics of Semiconductor Devices. New York, John Wiley & Sons, 1981.

Tipton, K.: Enzyme nomenclature. recommendations 1992. European Journal of Biochemistry, 223(1):1–5, 1994.

Tommassen, J.: Assembly of outer-membrane proteins in bacteria and mitochondria. Microbiology (Reading, England), 156(Pt 9):2587–2596, September 2010. PMID: 20616105.

Torrie, G. M. and Valleau, A.: Electrical double layers: 4. limitations of the Gouy-Chapman theory. Journal of Physical Chemistry, 86:3251–3257, 1982.

Tosteson, D.: Membrane Transport: People and Ideas. Bethesda MD, American Physiological Society, 1989.

Tourasse, N. and Li, W.: Selective constraints, amino acid composition, and the rate of protein evolution. Mol Biol Evol, 17(4):664, 656, April 2000.

Touw, D. S., Patel, D. R., and van den Berg, B.: The crystal structure of OprG from pseudomonas aeruginosa, a potential channel for transport of hydrophobic molecules across the outer membrane. PloS One, 5(11):e15016, 2010. PMID: 21124774.

Tseng, Y. Y. and Liang, J.: Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a bayesian monte carlo approach. Mol Biol Evol, 23(2):421–36, 2006.

Tsirigos, K. D., Bagos, P. G., and Hamodrakas, S. J.: OMPdb: a database of -barrel outer membrane proteins from gram-negative bacteria. Nucleic Acids Research, 39(Database):D324–D331, October 2010.

Ujwal, R., Cascio, D., Colletier, J., Faham, S., Zhang, J., Toro, L., Ping, P., and Abramson, J.: The crystal structure of mouse VDAC1 at 2.3 a resolution reveals mechanistic insights into metabolite gating. Proceedings of the National Academy of Sciences of the United States of America, 105(46):17742–17747, November 2008. PMID: 18988731.

Ulmschneider, M. B. and Sansom, M. S.: Amino acid distributions in integral membrane protein structures. Biochimica Et Biophysica Acta, 1512(1):1–14, May 2001. PMID: 11334619.

Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F., and Murshudov, G. N.: REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. Acta crystallographica. Section D, Biological crystallography, 60(Pt 12 Pt 1):2184–2195, December 2004. PMID: 15572771.

Van Gelder, P. and Tommassen, J.: Demonstration of a folded monomeric form of porin phoe of escherichia coli in vivo. J Bacteriol, 178(17):5320–5322, Sep 1996.

Vauquelin, L. and Robiquet, p.: The discovery of a new plant principle in asparagus sativus. Annales de Chimie, 57:88–93, 1806.

Vincze, J., Valisko, M., and Boda, D.: The nonmonotonic concentration dependence of the mean activity coefficient of electrolytes is a result of a balance between solvation and ion-ion correlations. J Chem Phys, 133(15):154507–6, 2010.

Visudtiphole, V., Thomas, M., Chalton, D., and Lakey, J.: Refolding of escherichia coli outer membrane protein f in detergent creates LPS-free trimers and asymmetric dimers. Biochem J, 392(Pt 2):375–381, Dec 2005.

Voet, D. and Voet, J.: Biochemistry. Hoboken, NJ USA, John Wiley, third edition, 2004.

Vrouenraets, M., Wierenga, J., Meijberg, W., and Miedema, H.: Chemical modification of the bacterial porin OmpF: gain of selectivity by volume reduction. Biophys J, 90(4):1202–11, 2006.

Wallace, D. C.: A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. Annual Review of Genetics, 39:359–407, 2005. PMID: 16285865.

Walsh, C.: Posttranslational modification of proteins: expanding nature's inventory. Roberts and Company Publishers, 2006.

Walsh, C. T., Garneau-Tsodikova, S., and Gatto, Gregory J, J.: Protein posttranslational modifications: the chemistry of proteome diversifications. Angewandte Chemie (International Ed. in English), 44(45):7342–7372, December 2005. PMID: 16267872.

Walther, D. M., Bos, M. P., Rapaport, D., and Tommassen, J.: The mitochondrial porin, VDAC, has retained the ability to be assembled in the bacterial outer membrane. Molecular Biology and Evolution, 27(4):887 –895, April 2010.

Walther, D. M., Papic, D., Bos, M. P., Tommassen, J., and Rapaport, D.: Signals in bacterial beta-barrel proteins are functional in eukaryotic cells for targeting to and assembly in mitochondria. Proceedings of the National Academy of Sciences of the United States of America, 106(8):2531–2536, February 2009. PMID: 19181862.

Walther, D. M., Rapaport, D., and Tommassen, J.: Biogenesis of beta-barrel membrane proteins in bacteria and eukaryotes: evolutionary conservation and divergence. Cellular and Molecular Life Sciences: CMLS, 66(17):2789–2804, September 2009. PMID: 19399587.

Warshel, A. and Russell, S.: Calculations of electrostatic interactions in biological systems and in solutions. Quarterly Review of Biophysics, 17:283–422, 1984.

Warshel, A., Sharma, P. K., Kato, M., Xiang, Y., Liu, H., and Olsson, M. H. M.: Electrostatic basis for enzyme catalysis. Chemical Reviews, 106(8):3210–3235, 2006.

Watanabe, Y.: Effect of various mild surfactants on the reassembly of an oligomeric integral membrane protein ompf porin. J Protein Chem, 21(3):169–175, Mar 2002.

Waterman, M. S. and Smith, T. F.: On the similarity of dendrograms. Journal of theoretical biology, 73(4):789–800, August 1978. PMID: 703348.

Waterman, M. S., Smith, T. F., Singh, M., and Beyer, W. A.: Additive evolutionary trees. Journal of theoretical biology, 64(2):199–213, January 1977. PMID: 839800.

Weiss, M. S., Abele, U., Weckesser, J., Welte, W., Schiltz, E., and Schulz, G. E.: Molecular architecture and electrostatic properties of a bacterial porin. Science (New York, N.Y.), 254(5038):1627–1630, December 1991. PMID: 1721242.

Wheeler, D. and Bhagwat, M.: BLAST QuickStart: example-driven web-based BLAST tutorial. Methods in molecular biology (Clifton, N.J.), 395:149–176, 2007. PMID: 17993672.

Wilbur, W. J. and Lipman, D. J.: Rapid similarity searches of nucleic acid and protein data banks. Proceedings of the National Academy of Sciences of the United States of America, 80(3):726–730, February 1983. PMID: 6572363.

Williamson, M. P., Havel, T. F., and Wthrich, K.: Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. Journal of molecular biology, 182(2):295–315, March 1985. PMID: 3839023.

Wimley, W. C.: Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. Protein Science: A Publication of the Protein Society, 11(2):301–312, February 2002. PMID: 11790840.

Wimley, W. C.: The versatile beta-barrel membrane protein. Current Opinion in Structural Biology, 13(4):404–411, August 2003. PMID: 12948769.

Wu, D., Hu, T., Zhang, L., Chen, J., Du, J., Ding, J., Jiang, H., and Shen, X.: Residues asp164 and glu165 at the substrate entryway function potently in substrate orientation of alanine racemase from e. coli: Enzymatic characterization with crystal structure analysis. Protein Science: A Publication of the Protein Society, 17(6):1066–1076, June 2008. PMID: 18434499.

Wu, H.-J., Ho, C.-W., Ko, T.-P., Popat, S. D., Lin, C.-H., and Wang, A. H.-J.: Structural basis of alpha-fucosidase inhibition by iminocyclitols with k(i) values in the micro- to picomolar range. Angewandte Chemie (International ed. in English), 49(2):337–340, 2010. PMID: 19967696.

Wu, T. T., Fitch, W. M., and Margoliash, E.: The information content of protein amino acid sequences. Annual review of biochemistry, 43:539–566, 1974. PMID: 4369316.

Wu, X., Edwards, H. D., and Sather, W. A.: Side chain orientation in the selectivity filter of a voltage-gated ca channel. Journal of Biological Chemistry, 2000.

Wyrembak, P. N., Babaoglu, K., Pelto, R. B., Shoichet, B. K., and Pratt, R. F.: O-aryloxycarbonyl hydroxamates: new beta-lactamase inhibitors that cross-link the active site. Journal of the American Chemical Society, 129(31):9548–9549, August 2007. PMID: 17628063.

Xiang, D. F., Patskovsky, Y., Xu, C., Fedorov, A. A., Fedorov, E. V., Sisco, A. A., Sauder, J. M., Burley, S. K., Almo, S. C., and Raushel, F. M.: Functional identification and structure

determination of two novel prolidases from cog1228 in the amidohydrolase superfamily. Biochemistry, 49(31):6791–6803, August 2010. PMID: 20604542.

Yan, B. X. and Sun, Y. Q.: Glycine residues provide flexibility for enzyme active sites. Journal of Biological Chemistry, 272(6):3190–3194, February 1997.

Yang, J., Ellinor, P., Sather, W., Zhang, J., and Tsien, R.: Molecular determinants of ca2+ selectivity and ion permeation in l-type ca2+ channels. Nature, 366:158–161, 1993.

Yeom, S., Ji, J., Kim, N., Park, C., and Oh, D.: Substrate specificity of a mannose-6-phosphate isomerase from bacillus subtilis and its application in the production of l-ribose. Applied and environmental microbiology, 75(14):4705–4710, July 2009. PMID: 19447949.

Yu, Y. K., Wootton, J. C., and Altschul, S. F.: The compositional adjustment of amino acid substitution matrices. Proc Natl Acad Sci U S A, 100(26):15688–93, 2003.

Zemaitis, Joseph F., J., Clark, D. M., Rafal, M., and Scrivner, N. C.: Handbook of Aqueous Electrolyte Thermodynamics. New York, Design Institute for Physical Property Data, American Institute of Chemical Engineers, 1986.

Zeth, K.: Structure and evolution of mitochondrial outer membrane proteins of beta-barrel topology. Biochimica Et Biophysica Acta, May 2010. PMID: 20450883.

Zhang, C., Raugei, S., Eisenberg, B., and Carloni, P.: Molecular dynamics in physiological solutions: Force fields, alkali metal ions, and ionic strength. Journal of Chemical Theory and Computation, 6(7):2167–2175, 2010.

Zhang, H.: The optimality of naive bayes. Proceedings of the 17th International FLAIRS conference (FLAIRS2004), 2004.

Zhang, Y. and Skolnick, J.: The protein structure prediction problem could be solved using the current PDB library. Proceedings of the National Academy of Sciences of the United States of America, 102(4):1029–1034, January 2005. PMID: 15653774 PMCID: 545829.

VITA

David Jimenez-Morales

Education

University of Illinois at Chicago, Chicago, IL
Ph.D., Bioinformatics, (Aug'06 - May'13)
Dissertation: Effects of Physicochemical and Functional Constraints on the Sequence and Structure of Proteins.

University Complutense of Madrid, Madrid, Spain
M.S., Bioinformatics and Computational Biology, (Jan'03 - June'04)

University Complutense of Madrid, Madrid, Spain
Teaching Certificate, (1999)

University Complutense of Madrid, Madrid, Spain
Microbiology PhD Program and Research Certificate, (1998 - 2000)

University of Granada, Granada, Spain
B.Sc., Fundamental Biology, (2008)

Teaching

University of Illinois at Chicago, Chicago, IL
Instructor
Bioe480, *Introduction to Bioinformatics* (Fall 2010)
Bioe481, *Bioinformatics Lab* (Fall 2012)

Teaching Assistant
Bioe250, *Clinical Problems in Bioengineering* (Spring 2013)
Bioe480, *Introduction to Bioinformatics* (Fall 2010, 2012)
Bioe481, *Bioinformatics Lab* (Fall 2012)
Bioe455, *Intro to Cell and Tissue Engineering* (Spring 2007)
Bioe240, *Modeling Physiological Data & Systems* (Fall 2006)

Awards

Beca Talentia Excellence Grant, Regional Ministry for Innovation, Science and Enterprise. Junta de Andalucia. Spain, (2008 – 2010)

Publications

Peer-reviewed Publications

1. **Jimenez-Morales, D.**, Adamian, L., Shi, D., Liang, J. *Unveiling the prevalence of a spontaneous post-translational modification.* In preparation.
2. **Jimenez-Morales, D.**, Liang, J., Eisenberg, B. *Ionizable side chains at catalytic active sites of enzymes.* Eur Biophys J, May;41(5):449-60, 2012.
3. Naveed, H., **Jimenez-Morales**, D., Tian, J., Pasupuleti, V.; Kenney, L., Liang, L. *Engineered oligomerization state of OmpF protein through computational design decouples oligomer dissociation from unfolding.* J Mol Biol, May 25;419(1-2):89-101, 2012.
4. **Jimenez-Morales, D.**, Liang, J. *Pattern of amino acid substitutions in transmembrane domains of β-barrel membrane proteins for detecting remote homologs in bacteria and mitochondria.* PLoS ONE, 6(11):e26400. Epub 2011 Nov 1, 2011.
5. Liang, J., Naveed, H., **Jimenez-Morales**, D., Adamian, L., Lin, M. *Computational studies of membrane proteins: Models and predictions for biological understanding.* Biochim Biophys Acta, Apr, 1818(4):927-41, 2011.
6. **Jimenez-Morales, D.**, Adamian, L., Liang, J. *Detecting remote homologs using scoring matrices calculated from the estimation of amino acid substitution rates of β-barrel membrane proteins.* Conf Proc IEEE Eng Med Biol Soc, 30:1347-1350, 2011.
7. **Jimenez-Morales, D.** *Substitution Rates of Amino Acid residues in Transmembrane Regions of Beta-Barrel Membrane Proteins.* BioE Student Journal. N1, 2008.

Abstracts

1. **Jimenez-Morales D**, J Liang, and B Eisenberg. *Active Sites of Enzymes are Crowded with Charge.* Biophysical Journal. 100(3): p. 218a. Abstract 1191-Pos and Poster Board B101, 2011.
2. **Jimenez-Morales D**, J. Liang, and B. Eisenberg. *Active Sites of Enzymes are Crowded with Charge.* 6th Annual Midwest Conference on Protein Folding, Assembly, and Molecular Motions. University of Notre Dame, 2011.
3. **Jimenez-Morales D**, Rudong Li, Zhuo Wang, Yingzi Li, Lei Liu, Jie Liang. *Evolutionary Speed of Enzymes Functional Surfaces and their Relationship with Metabolic Fluxes in Networks of Central Carbon Metabolism of Bacteria.* Biophysical Journal 98(3) pp. 741a, 2010.
4. **Jimenez-Morales D**, Adamian L, Jie Liang *Substitution Rates Of Amino Acid Residues In Transmembrane - Extracellular And Periplasmic- Regions Of Beta-Barrel Membrane Proteins.* Biophysical Journal 94(2) pp. 1959, 2008.

5. **Jimenez-Morales D**, Fuentes F, Heredia S, Calvo A, Ramos C. *Susceptibility of Enterococcus faecalis isolated from animals to several groups of antimicrobial agents.* Third European Chemotherapy Congress. Madrid, 10. 7 May, 2000.
6. **Jimenez-Morales D**, Heredia S, Calvo A, Gomez-Luz ML, Prieto J. *Sensibilidad a antimicrobianos de Enterococcus faecalis de procedencia animal.* Spanish Chemotherapy Society Congress. May, 1999.

Books

1. **Coauthor.** *Propiedades adicionales no antibioticas de los macrolidos* Scientific Communication Management Editors. 2001.
2. **Coauthor.** *Gua Setas Autoctonas.* Editorial Doyma. Madrid. 2001.