STRUCTURE DETERMINATION AND DESIGN OF BIOMINERAL-ASSOCIATED PROTEINS

by
David L. Masica

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
October, 2009

© 2009 David L. Masica
All Rights Reserved
Abstract

Many organisms produce inorganic materials via protein-influenced crystal growth—a process known as biomineralization. Understanding this process would shed light on hard-tissue formation, guide efforts to develop biomaterials, and provide a deeper understanding of in vivo phase-boundary biophysics. Unfortunately, the structure of biomineral-associated proteins cannot be determined by X-ray crystallography or solution-state nuclear magnetic resonance (NMR); therefore, existing structural models of biomineral-associated proteins are inferential, incomplete, or low resolution. In this thesis I develop both computational and combined computational-experimental approaches to investigate protein structure and sequence determinants in biomineralization. The methodological centerpiece of this thesis is the development of a comprehensive suite of computational structural biology tools (RosettaSurface). This software package includes modules for rigid-body protein-surface docking, protein folding at the biomineral surface, solid-state NMR-biased structure determination of protein adsorbed states, and de novo design of biomineralization systems. These represent the first structure-prediction techniques developed for and applied to biomineralization problems. I use these algorithms to investigate a model system: human-salivary statherin and hydroxyapatite (hydroxyapatite is the primary mineral component of mammalian skeletal and dental tissue). Investigation of this model system culminates in the first-ever reasonably high-resolution structure of a protein adsorbed to a solid surface determined using a combined experimental-
computational approach. Remarkably, I had predicted this structure \textit{a priori} to high accuracy using RosettaSurface in the absence of experimental bias, including the accurate prediction of a molecular recognition motif. The last development of the RosettaSurface algorithm that I present in this thesis is the design module. This algorithm simultaneously optimizes protein fold, orientation, and sequence while adsorbed to a mineral surface. I used that program to design peptides to bind different crystal surfaces of the mineral calcite. We chemically synthesized a set of the designer peptides that were predicted to have a high affinity for calcite, introduced designer peptides to solutions of growing calcite crystals, and observed the resulting morphological changes using a scanning electron microscope. All designed peptides had significant biomineralization activity and we found a dependence on basic amino acid content for sequence-order specificity.

The official readers of this thesis are my thesis advisor, Prof. Jeffrey J. Gray (Chemical and Biomolecular Engineering), and Prof. Jan Hoh (Physiology).
Acknowledgements

The single greatest contribution to the completion of my thesis work was the encouragement of my wife and best friend, Carol Masica. Her support made it possible and her love made it worthwhile. Carol’s work ethic and perseverance have also been an inspiration to me. She created a business and continues to thrive as an entrepreneur in Baltimore—the city we chose so that I would have the opportunity to do my doctoral research at Johns Hopkins.

I couldn’t have asked for a better thesis advisor than Jeff Gray, nor could I have asked for a better research environment than the one he provided. Jeff is effective at passing on important skills for beginning scientists such as critical thinking, writing, and presenting. Jeff provided enough insight and direction to keep me on track, but allowed enough room for failure so that I could develop trouble-shooting skills and enough room for self motivation so that I could own my project. Jeff makes sure his lab is a stimulating environment for research, which includes endless opportunities to attend and participate at national conferences, lab trips (hiking, skiing, gathering, etc.,), and he spares no expense to make sure we are not limited by our equipment.

My mother, father, and brother (Linda, Steve, and Brian Masica, respectively) have contributed greatly to my ability to undertake all that comes with earning a Ph.D. Their work ethic and passion are like nothing I have seen elsewhere, and growing up in that environment provided me the endurance and tenacity I could not have acquired elsewhere, and certainly drew upon regularly during my thesis
work. Their belief in me has always been undoubtedly genuine, and that support gave me the confidence to attack many situations, during my graduate research, with the conviction they required.

The entire faculty and staff of The Johns Hopkins University, Program in Molecular Biophysics goes out of their way to provide an effective and comfortable learning environment. I recognize that first-year talks, proficiency examinations, and yearly thesis reviews take a lot of time to coordinate for the faculty and staff. These sorts of checkpoints serve to benefit the students at the expense of faculty and staff, and few other departments on campus participate in these very beneficial checkpoints. I would like to send a special thanks to Ranice Crosby who somehow manages to undertake a seemingly infinite number of administrator tasks so that students can focus on research; Jan Hoh who trained me to use an atomic force microscope, served as my ad-hoc thesis review advisor, and served as the official second reader of this dissertation; Richard Cone who was fascinating and informative to converse with during my teaching assistantship for him; Doug Barrick who was particularly approachable during my time at Johns Hopkins and who always provided excellent experimental critique of my work; Tamara Hendrickson and Megan Erenwerth for their help with peptide synthesis; those who proctored my graduate board oral (Mario Amzel, Karen Fleming, Tom Woolf, Jon Lorsch, Marc Ostermeier); those that oversaw my first-year seminar (Richard Cone and Robert Schleif); and those that sat on my dissertation defense (Jan Hoh, David Draper, David Shortle, and Tom Woolf).
Arvind Sivasubramanian, Kosta Makrodimitis, and Rahul Bhowmik were Gray lab post-doctoral fellows during my time in Gray lab. As scientific superiors I learned a lot from each of them.

Mike Daily, Aroop Sircar, Monica Berrondo, and Sid Chaudry were all Gray lab graduate students whose work overlapped (temporally) considerably with mine. I had a lot of fun in and out of lab with all of them, and definitely learned a lot from all of them.

I had the privilege to mentor five undergraduate students, as primary research supervisor, during my doctoral work (Eleanor Small, Elizabeth Specht, Sarah Schrier, Eric Kim, and Moon Young “Liza” Lee). Each of these students improved the quality of my doctoral work both scientifically and personally. Eleanor Small was integral in getting the design project (chapter 5) started. Elizabeth Specht and Sarah Schrier did the majority of the experiments presented in chapter 5. Eric Kim prepared several of the figures in Chapter 2. Each is now pursuing a successful career in science or will be in the near future.

A very special thanks to Marc Ostermeier who was gracious enough to donate bench space to me, and his graduate students Jing Liang and Richard Heins took great care to teach me the ropes of molecular biology experiments. Without all of their help, the work completed in chapter 5 of this dissertation would not have been possible.

I am very thankful for all of the fruitful collaborations I have established during my dissertation research. Marc McKee, Yung-Ching Chien, and William
Addison from McGill University in Canada were not only coauthors of mine, but were also participants in many enlightening conversations on the topics covered in this thesis. Wendy Shaw from the Pacific Northwest National Laboratory was a coauthor and excellent collaborator. Gary Drobny from University of Washington was a coauthor and collaborator and always provided insightful reviews for other papers that I was writing.

I was fortunate enough to take part in a successful, performing rock band (Frontwise) while I was earning my Ph.D. Performing, practicing, and recording regularly was an instrumental outlet for me during my thesis work. In particular the people I played with (Rob Levendosky, Stephen “Turk” Turk, Jesse Margiota, Seth Sedgwick, and Jeremy Hamman) are not only excellent musicians but also great friends.

For funding I thank: The Arnold and Mabel Beckman Foundation through a young investigator grant, The National Institutes of Health training grant, and The National Science Foundation’s CAREER Award.
Chapter 1. Introduction

1.1 Background

1.2 Studying biomineralization experimentally

1.3 Studying biomineralization computationally

1.4 Specific aims, challenges, and impact of thesis

1.5 Organization of thesis

Chapter 2. Structure prediction of protein-solid surface interactions reveals a molecular recognition motif of statherin for hydroxyapatite

2.1 Introduction

2.2 Computational methods

2.3 Results

2.4 Discussion

2.5 Conclusion

Chapter 3. Solution- and adsorbed-state structural ensembles predicted for the statherin-HAp system

3.1 Introduction
Chapter 4. **Structure determination method for biomineral-associated protein using combined solid-state NMR and computational structure prediction**

4.1 Introduction
4.2 Materials and methods
4.3 Results
4.4 Discussion

Chapter 5. **De novo design of peptide-calcite biomineralization systems**

5.1 Introduction
5.2 Materials and methods
5.3 Results
5.4 Discussion
5.5 Supplement

Chapter 6. **Conclusion**

6.1 Conclusion
6.2 Contributions to the field of biomineralization
6.3 Future directions
References
Curriculum Vitae
List of Tables

2.1: Rotamer parameters to expand the Rosetta rotamer library .................................................. 30
2.2: Residue-surface interactions .................................................................................................. 38
4.1: Solid-state NMR measurements used in first round of biased structure prediction ............. 88
4.2: Solid-state NMR measurements used in second round of biased structure prediction ........ 92
4.3: Solid-state NMR measurements used during round 1 and 2 and the corresponding predictions from the representative model (figure 4.5) .............................................. 97
5.1: Synthesized peptide sequences ............................................................................................. 109
5.2: Predicted binding affinities for the designed and scrambled variant sequences .................. 120
List of Figures

2.1: Structure of hydroxyapatite and statherin 21
2.2: RosettaSurface flowchart 26
2.3: Crystal unit-cell tracking vectors 27
2.4: Phosphoserine rotamer statistics 29
2.5: Statherin-HAp energy landscape 34
2.6: Distributions of statherin-HAp distances among the 200 top-scoring structures 35
2.7: Statherin-HAp interaction energy distributions among the 200 top-scoring structures for selected residues 37
2.8: Orientation of model 1 39
2.9: The IPOT motif 40
2.10: Selected structural details of model 1 41
2.11: Map of interface FADE scores 42
2.12: Distribution of orientation of statherin on the HAp surface among the 200 top-scoring structures 44
2.13: Top view of the four structures 45
3.1: Flowchart for the RosettaSurface algorithm 55
3.2: Predicted statherin solution-state structure 61
3.3: Predicted binding and structural statistics for statherin adsorbed to HAp 62
3.4: Adsorption induced folding event about statherin's Gly-12 as predicted by RosettaSurface 63
3.5: Adsorption induced folding event in a control peptide predicted by RosettaSurface

3.6: Comparison of RosettaSurface predictions and all previously published solid-state NMR measurements

3.7: Clashes created in statherin’s C-terminal segment when NMR constraints are met

3.8: Predicted binding motifs on three HAp surfaces

3.9: Three adsorption phenomena plotted against residue number for the 100 top-scoring decoys

4.1: Schematic of hexagonal HAp crystal

4.2: Ensemble structure of statherin adsorbed to HAp {001} surface determined from the first round of biased structure prediction

4.3: Error $\epsilon$ versus weight at five different HAp surfaces

4.4: Ensemble structure of statherin adsorbed to HAp {001} surface determined from the final phase of biased structure determination

4.5: The molecular structure of statherin adsorbed to HAp {001} surface

5.1: Calcite

5.2: Net positively charged and negatively charged calcite surface terminations used in design simulations

5.3: Structure predicted by RosettaSurface.Design

5.4: SEM images for calcite crystallized in the presence of 0.45 mg/ml peptide
Introduction

1.1 Background

Biomineralization is a process of crystal nucleation and growth controlled by bioorganic molecules such as proteins (Dove et al., 2003, Mann, 2001, Sigel et al., 2008). Many organisms utilize biomineralization to fabricate the solid inorganic components of biogenic materials. These include a variety of structural and functional materials and the mineral phase of hard tissues. Deleterious biomineralization can result in pathologies such as kidney stones (Dussol et al., 1995, Ryall, 1996), dental calculus, and atherosclerosis (Dorozhkin and Epple, 2002).

Mineral crystallization in biology is complex and arises from diverse and sometimes disputed events. Crystallization in general consists of two distinct events: nucleation and crystal growth. Nucleation is typically the rate-limiting step in crystallization (Auer and Frenkel, 2001). Biological systems have devised several strategies to overcome the energy barrier associated with nucleus formation. One common theme is the use of biomolecules for ion sequestration. These strategies include the active transport of high ion concentrations into vesicles (Kroger and Poulsen, 2008, Marsh et al., 2002), polymer-induced liquid precursors (Gower and Odom, 2000), and highly charged protein matrices (Addadi and Weiner, 1985). Once the critical nucleus has formed, crystal growth occurs. During crystal growth proteins can locate and function at the liquid-solid
surface phase boundary and accelerate (Elhadj et al., 2006), inhibit (Shiraga et al., 1992, Boskey et al., 1993), or shape crystal growth (Sollner et al., 2003, Naka and Chujo, 2001). Some common biogenic materials include calcified minerals, magnetite, and biosilica.

Calcified minerals are the most common in biology (Mann, 2001) and include calcium phosphate, calcium carbonate, and calcium oxalate polymorphs. Hydroxyapatite (HAp; \( \text{Ca}_5(\text{PO}_4)_3(\text{OH}) \)) is the primary mineral component of vertebrate skeletal and dental tissues and the most common calcium phosphate polymorph in healthy biomineralization (Robinson et al., 2004). Pathological biomineralization of HAp can result from disordered crystallization, amorphous mineralization, and impurities (Dorozhkin and Epple, 2002).

The most common calcium carbonate polymorphs used in nature are calcite and aragonite (Weiner and Addadi, 1997). Avian eggshells and molluscan shells are primarily biomineralized calcium carbonate (Silyn-Roberts and Sharp, 1986, Falini et al., 1996). Some crustaceans, corals, and sponges make extensive use of calcite and aragonite for the fabrication of mineralized skeletal tissue (Arias and Fernandez, 2008). The balance organ of the mammalian inner ear (vestibule) relies on the calcite biomineral otoconia for gravity sensitivity; other organisms use vestibular calcite, aragonite, or both (Verpy et al., 1999).

The calcium oxalate polymorphs whewellite (monohydrate) and weddellite (dihydrate) rarely participate in healthy tissue formation. These calcium oxalate polymorphs comprise the primary mineral components of kidney stones and
some gallstones (Dussol et al., 1995, Ryall, 1996). In healthy individuals (nonstone formers), the protein lithostathine is thought to help prevent stone formation (Bernard et al., 1995).

Iron oxides and iron sulfides are used for the formation of magnetic crystals in magnetotactic bacterium (Blakemore, 1982). These magnetic crystals are synthesized and stored in membrane bound organelles and used to bias diffusion parallel to the earth’s magnetic field. This method of chemotaxis is in contrast to the pseudo-random walk of some flagellar bacteria and confers some advantages.

Biosilica is a more common in aquatic fauna than in terrestrial fauna. This biomineral comprises the exoskeleton of two unicellular eukaryotes, the diatom (plankton) and radiolarian (protozoan) (Raven and Waite, 2004, Racki and Cordey, 2000). Certain deep-sea sponges synthesize glassy optical waveguides of biosilica called spicules (Brümmer et al., 2008). These waveguides are used by the sponge to help culminate light that is attenuated by the ocean they inhabit. As part of a symbiotic relationship, algae harness the spicule culminated light for photosynthesis. In turn, the algae provide the sponge with essential metabolites.

These diverse organisms benefit from biomineralization through related processes including the production of proteins that affect inorganic crystal growth and morphology. The goal of this thesis is to examine how proteins exert this influence at the structural and chemical level.
1.2 Studying biomineralization experimentally

Many experimental techniques exist for studying and controlling protein adsorption (Gray, 2004). Experimental characterization of protein adsorption to solid surfaces is subject to a unique set of problems. First, material substrates are often micrometer scale. The relatively high volume to surface area ratio at this size regime results in too few binding interactions to yield a high signal to noise ratio in many characterizations. The large particle size also causes light scattering, rendering some spectroscopic techniques difficult or impossible. Second, protein adsorption can be irreversible, making the calculation of thermodynamic properties difficult or irrelevant (Haynes and Norde, 1994). Many materials are crystalline in the bulk, but can be heterogeneous at the surface. Similarly, protein binding at the material surface can be heterogeneous from the sub-monolayer to the multilayer (Ramsden et al., 1994, Hlady and Buijs, 1996, Calonder et al., 2001). These heterogeneities render experimental techniques that rely on a structural consensus (X-ray crystallography and solution NMR, for instance) intractable. Despite these challenges, many experimental studies have advanced our understanding of protein biophysics at the liquid-solid surface interface. As a result of these advances, novel biomaterials are being engineered. Several experimental techniques and applications are outlined below.
Probing the kinetics and thermodynamics of protein adsorption. Standard pull-down binding assays, surface Plasmon resonance spectroscopy (SPR), and the quartz crystal microbalance (QCM) can determine adsorbed amount and hence equilibrium constants and binding energies (Shen et al., 2001, Schuck, 1997, Goobes et al., 2007b). Isothermal titration calorimetry is useful for the determining thermodynamic entropy, enthalpy, and free energy of adsorption (Goobes et al., 2007b). SPR and QCM can also measure adsorption rate and are therefore useful for studying kinetics (Shen et al., 2001, Schuck, 1997).

Probing structure in biomineralization. There are many useful tools for exploring structural aspects of biomineralization and protein adsorption. These include techniques that reveal the morphological influence of proteins on mineralization and the structure of proteins adsorbed to solid surfaces.

Some techniques probe the bulk atomic structure of the crystal. X-ray diffraction is the most accurate and widely used of these techniques (Sollner et al., 2003), but TEM and infrared spectroscopy are also used (Banfield et al., 2000, Politi et al., 2004). These techniques are particularly useful for probing crystal phase transitions induced in biomineralization.

Several microscopies are useful for directly observing crystal morphology resulting from protein-influenced crystallization. Bright-field microscopy is an easy first-line approach to determining the morphology of crystals that are on the micrometer scale (Geider et al., 1996). Scanning and transmission electron microscopy can reveal distinct crystalline features at the micrometer and
submicrometer scales (Aizenberg et al., 2003, Banfield et al., 2000). These techniques suffer from the requirement of drying samples, and TEM requires sub-100 nanometer thickness, often relegating TEM investigation to crystal edges. Atomic force microscopy (AFM) is a powerful tool for working at the micrometer and nanometer scales. AFM is also powerful in its ability to work in solution, allowing the investigation of biomineralization dynamics (Elhadj et al., 2006). AFM is uniquely suited to image steps that are only one atomic layer thick (Hariadi et al., 2002) and can be used to create movies that capture the dynamics of mineral formation (for example, http://www.iop.org/EJ/mmedia/0957-4484/19/10/105304/).

Some mass spectrometric techniques are useful for probing structure and orientation of adsorbed proteins. For instance, in hydrogen-exchange mass spectrometry, the level of deuteration in a particular region of a protein reflects its exposure to solvent (Zhang and Smith, 1993). Subsequent proteolysis and mass spectrometry can reveal those specific regions. In protein adsorption, residues at the protein-surface interface and those sequestered in a stable protein core will be less deuterated than those exposed to solvent (Sokol et al., 2003). Time-of-flight secondary ion mass spectrometry (ToF-SIMS) can also probe protein orientation and structure at the interface. In the ToF-SIMS experiment, primary ions bombard the sample and fragment the protein (Wagner and Castner, 2001). The resulting fragments (secondary ions) are then analyzed with mass spectrometry. Tightly adsorbed protein fragments will not “take flight” and will therefore be absent from the resulting spectrum.
Attenuated total reflectance Fourier transform infrared (ATR-FTIR), Raman, and sum frequency generation spectroscopy are sensitive to protein-adsorbed state structure (Chittur, 1998, Podstawka et al., 2004, Wang et al., 2002). While these methods cannot isolate domain or region specific structure, they can reveal bulk secondary structure and structural homogeneity. A possibility I have considered is combining isotope-edited IR (Decatur, 2006) (a form of IR that relies on isotope incorporation to determine residue-specific secondary structure) and ATR-FTIR to probe residue specific structure of protein-adsorbed states.

Solid-state NMR (ssNMR) is the most powerful experimental tool for determining high-resolution protein adsorbed-state structure (Goobes et al., 2007b, Goobes et al., 2007a, Raghunathan et al., 2006, Goobes et al., 2006a, Gibson et al., 2006, Gibson et al., 2005, Stayton et al., 2003, Long et al., 2001, Shaw et al., 2000, Long et al., 1998). Solid-state NMR is used to investigate dynamics and protein-intramolecular and protein-surface intermolecular distances. The experimental setup requires the incorporation of site-specific isotopes, which limits the size of the polypeptide under investigation to those that can be chemically synthesized. If protein-surface intermolecular information is desired, the material must also contain an atomic species with spin $\frac{1}{2}$ nuclei. In addition to these limitations, data acquisition is very time consuming. Only two mineral-adsorbed proteins have been subjected to this type of experiment, statherin (Goobes et al., 2007b, Goobes et al., 2007a, Raghunathan et al., 2006, Goobes et al., 2006a, Gibson et al., 2006, Gibson et al., 2005, Stayton et al., 2003, Long et al., 2001, Shaw et al., 2000, Long et al., 1998) and the leucine-rich
amelogenin protein (Shaw et al., 2008, Shaw et al., 2004). To date, less than 15 measurements are published for each; only about five distance or angle measurements are published per study. In contrast, typically 10-15 measurements per residue are acquired using solution-state NMR for protein structure determination.

**Designing novel biomaterials with proteins.** Often, the ultimate goal of scientific investigation is acquiring sufficient mastery to design novel systems. Several methods in protein adsorption and biomineralization now enable this goal. One successful approach is the use of directed evolution techniques. Directed evolution techniques such as cell-surface and phage display can biopan for material-binding peptides (Sarikaya et al., 2003). In these experiments, random peptide loops (up to $10^{11}$ sequences per experiment) are exposed on a viral capsid or bacterial cell surface. Peptide sequences are selected for the affinity to bind a material of interest such as calcite (Gaskin et al., 2000), hydroxyapatite (Gungormus et al., 2008), silica (Naik et al., 2002a), or gold (Brown, 1997). The use of directed evolution techniques in biomineralization has already led to novel biominerals including metal nanoparticles (Naik et al., 2002b) and semiconductor nanowires (Mao et al., 2003).

### 1.3 Studying biomineralization computationally

As in other problems in computational structural biology, studying protein adsorbed states is subject to the challenge of sampling an immense conformation space and the limitations of an approximate energy function. Protein-surface
Simulations that model the protein with entirely atomistic (high resolution) or entirely coarse grained (low resolution) representations address different problems. Next I outline several computational techniques in biomineralization separated by the resolution of molecular representation.

**Low-resolution models of protein adsorption.** Perhaps the lowest-resolution techniques are the colloidal or Brownian dynamics simulations. Colloidal models of adsorption represent proteins as single geometric shapes (typically cylinders or spheres) and the solid surface as a hard boundary. Both the protein and surface inherit some properties such as charge or charge density, hydrophobicity, and size. Colloidal simulations typically ignore intramolecular interactions and are useful for studying intermolecular interactions. The advantage gained by such simplified models is that the simulation can accommodate solutions of proteins rather than just single molecules, which affords the simulation of competitive adsorption and the formation of multilayers (Wijmans and Dickinson, 1999, Pugnaloni et al., 2005). Colloidal scale simulations have been used to predict protein adsorption’s dependence of molecular charge and hydrophobicity (Yuan et al., 2000, May et al., 2000), solution conditions (Ravichandran and Talbot, 2000), and concentration (Ravichandran and Talbot, 2000).

United-atom, also known as centroid or pseudo-atom, models are common low-resolution representations in protein modeling. These models reduce system complexity by incorporating groups of atoms into a single particle (united atom). One common protein representation is to consider each protein residue
as a particle. Often, the surface is represented as a two-dimensional lattice of interaction sites. These coarse-grained models have an advantage over colloidal models in that protein-intramolecular interactions are included and protein is modeled with a flexible backbone. United-atom models have been used to calculate adsorption-induced entropy (Liu and Haynes, 2004) and structure (Castells et al., 2002) changes of polymers. These models suffer in that they don’t resolve atomistic contributions to the structures they predict and the structures reflect hypothetical polymers not real biomolecules.

**High-resolution models of protein adsorption.** High-resolution computational approaches in molecular simulation represent molecules as collections of atoms. The resolving power of these techniques facilitates the understanding of residue and atom specific contributions to molecular structure and stability. The cost of this increased resolving power is the great computational burden. Because all-atom simulations are expensive, investigations of large systems require sufficient structural information to generate plausible starting structures. For this reason, all-atom simulations of proteins or protein complexes typically require experimentally acquired atomic coordinates as input. Because the atomic coordinates of biomineral-associated proteins cannot be determined by experiment, all-atom simulations of these systems require making assumptions for which there is little or no empirical evidence.

Perhaps the most common high-resolution modeling technique used in biomineralization is molecular dynamics. During a molecular dynamics
simulation, atoms and molecules interact for a period of time obeying Newton’s laws. All-atom molecular dynamics force fields typically include van der Waals and electrostatics for nonbonded interactions. One benefit of molecular dynamics simulations is that the dynamics simulates kinetic phenomena. Molecular dynamics simulations have been used to investigate the influence of explicit water molecules (Sun et al., 2007, Zhang et al., 2004, Wierzbicki et al., 2007), relative energies, and dynamics of protein adsorption (Raut et al., 2005, Wierzbicki et al., 2007, Zhang et al., 2004, Gerbaud et al., 2000, Huq et al., 2000). The molecular dynamics sampling strategy is often implemented in colloidal and united-atom models as well (Wijmans and Dickinson, 1999).

Another common sampling technique is Monte Carlo. Monte Carlo techniques rely on repeated random sampling and either ranking and/or averaging the resulting ensemble. In molecular biology, molecular conformation, the docked orientation of interacting molecules, or chemical composition can be sampled. Ranking of these randomly generated structures typically relies on empirical, physical, or statistical force fields. Monte Carlo strategies benefit from the ability to sample diverse and unrelated states of a system without the requirement to sample the trajectories joining those states. All-atom Monte Carlo methods have been used to study semiconductor-binding peptide conformation in the absence of the semiconductor surface (Mitternacht et al., 2007) and adsorbed orientation of rigid antibody structures (Zhou et al., 2004). The Monte Carlo sampling strategy is often implemented in colloidal and united-atom models as well (Castells et al., 2002).
The most physically rigorous computational techniques currently employed in biomineralization problems are those that model electron-electron repulsion: the quantum chemistry techniques. These \textit{ab initio} calculations are used to solve Schrödinger’s equation and require no empirical data. Unfortunately the computational expense imparted by these simulations restricts their use to very small systems (typically < 100 atoms). Quantum simulations have been used to investigate the energetics and bonded structure of glycine interacting with small hydroxyapatite clusters (Rimola \textit{et al.}, 2008). One use of quantum techniques is the ability to extrapolate atomic parameters such as charge. Often these parameters are used in computational methods that employ semiempirical force fields (Gerbaud \textit{et al.}, 2000). Quantum mechanics methods can be combined with molecular dynamics or Monte Carlo strategies to access greater length and time scales.

**Multiscale approaches and protein structure prediction.** All-atom methods in macromolecular modeling are too computationally expensive to determine structures \textit{a priori}. Colloidal and united-atom models do not reveal sufficient detail to address many problems such as protein catalysis and design. Multiscale approaches can combine low- and high-resolution strategies to predict all-atom protein structures \textit{a priori} (Ayton \textit{et al.}, 2007). The combination of strategies typically employs a two-step approach: 1) rapidly search diverse conformations while assuming a united-atom representation. 2) Convert united-atom model to all-atom model and search local conformations. Multiscale strategies have been successfully applied to problems such as protein folding (Bradley \textit{et al.}, 2005b)
and docking (Gray et al., 2003a). To date, this type of multiscale method has not been used to predict the structure of a protein adsorbed to a solid surface.

Computational sequence optimization in biomineralization. To date, there is only one published study where novel material-binding peptide sequences were realized by computational methods (Oren et al., 2007). In that study, phage display was used to evolve a set of quartz-binding peptides. Sequence scoring matrices were developed considering peptide sequence and relative affinity for quartz. The scoring matrices were used to rank randomly generated sequences, a subset of which were sequenced and assayed for relative affinity. This approach was able to design novel material binding sequences with predicted relative affinities as assayed by surface plasmon resonance spectroscopy.

1.4 Specific aims, challenges, and impact of thesis

Aim One: develop a computational method to predict the structure of a protein adsorbed to a biomineral surface. Proteins can affect the formation of biominerals by directly binding a crystal face (Sigel et al., 2008, Mann, 2001, Dove et al., 2003). Because the structure of interacting biomolecules influences function and mechanism, determining the structure of biomineral-associated proteins is important for understanding how proteins influence biomineralization. Unfortunately, high-resolution structures of surface-adsorbed proteins cannot be determined by contemporary experimental methods. The absence of large amounts of high-resolution structural data presents a unique challenge for computational structure prediction in biomineralization systems.
On one hand, the absence of training sets and/or benchmarks derived from experimental data makes rigorous algorithmic testing and parameterization difficult. On the other hand, this limitation places special emphasis on the computational investigation of protein-surface interactions, as computation is currently the only means of obtaining atomic coordinates for protein adsorbed states. To that end, my first aim is to develop a structure prediction algorithm capable folding and docking a protein to a biomineral surface. This is the first time a computational structure prediction method is developed to address biomineralization.

**Aim Two: develop a combined computational-experimental method to determine the structure of a protein adsorbed to a biomineral surface.** To further address challenges posed in **Aim One**, I aim to develop a computational method that incorporates minimal, high-resolution ssNMR data. This combined approach considerably improves the plausibility of computational models by reducing conformation space and requiring that all models agree with all known high-resolution experimental data. This is a first time a combined strategy of this type is applied to biomineralization systems.

**Aim Three: Develop a computational method to design biomineralization systems de novo and validate with experiment.** Biogenic materials typically confer several advantages relative to synthetic and geologic counterparts (Sigel *et al.*, 2008, Mann, 2001, Dove *et al.*, 2003). For instance, bioorganic molecules can lower the activation energy of crystal formation, circumventing the requirement for extreme temperature, pressure, or pH (Sigel *et al.*, 2008, Mann, 1988). Also,
biogenic materials often have superior mechanical properties relative to nonbiogenic variants (Jiang et al., 2005). These advantages make biomineralization an attractive alternative to contemporary methods in materials production (Vrieling et al., 1999). Also, the rational design of proteins for specific interactions with such solid-phase materials could allow us to engineer metals, semiconductors, and tissues with nanoscale precision. Currently, our understanding of protein structure and sequence determinants in biomineralization is not sufficiently detailed to fully realize this potential. One method of designing biomolecular interactions to achieve desired phenotypes is de novo design. Computational de novo design has been successfully employed in biological problems including the design of novel protein folds (Kuhlman et al., 2003), protein-protein interactions (Kortemme et al., 2004), protein-DNA interactions (Ashworth et al., 2006), protein-peptide interactions (Sood and Baker, 2006), protein catalysis (Jiang et al., 2008, Murphy et al., 2009, Röthlisberger et al., 2008), and protein folding pathways (Nauli et al., 2001). The final aim of thesis is the development of an algorithm capable of de novo rational design of biomineralization systems. This marks the first time a de novo design strategy is applied to biomineralization.

1.5 Organization of thesis

Chapter 2. We present a novel computational method for modeling the interactions of proteins with solid surfaces using comprehensive sampling and an atomistic description. The approach relies on an all-atom Monte Carlo plus-
minimization search algorithm that rapidly and simultaneously optimizes rigid-body and side-chain conformations. We apply the method to the statherin-HAp system, an evolved protein-surface interaction that is likely to have one or a few specific structural solutions. The algorithm converges on a set of low energy, entropically favorable structures that are consistent with previous experimental results, namely protein-surface intermolecular distances acquired by ssNMR. The simulations isolate particular residues as being primary contributors to the adsorption free energy (hydrogen bonding, van der Waals, and electrostatic energies), in agreement with previous mutagenesis, deletion, and single amino acid experiments. We also report the discovery of a molecular recognition motif where the N-terminal \( \alpha \)-helix of statherin places all four of its basic residues to match the periodicity of open phosphate triad clusters across the \{001\} monoclinic face of the HAp surface. Results suggest new experiments that could further elucidate the structural features of this important biological system.

**Chapter 3.** We develop a multi-scale structure prediction technique to study solution- and adsorbed-state ensembles of biomineralization proteins. The algorithm employs a Metropolis Monte Carlo-plus-minimization strategy that varies all torsional and rigid-body protein degrees of freedom. We applied the technique to fold statherin, starting from a fully extended peptide chain in solution, in the presence of HAp \{001\}, \{010\}, and \{100\} monoclinic crystals. Blind (unbiased) predictions capture experimentally observed macroscopic and high-resolution structural features and show minimal statherin structural change upon adsorption. The dominant structural difference between solution and adsorbed
states is an experimentally observed folding event in statherin’s helical binding domain. While predicted statherin conformers vary slightly at three different HA\(p\) crystal faces, geometric and chemical similarities of the surfaces allow structurally promiscuous binding. Finally, we compare blind predictions with those obtained from simulation biased to satisfy all previously published ssNMR distance and angle measurements (acquired from HA\(p\)-adsorbed statherin). Atomic clashes in these structures suggest a plausible, alternative interpretation of some ssNMR measurements as intermolecular rather than intramolecular. This work demonstrates that a combination of ssNMR and simulation could effectively determine high-resolution protein structures at biomineral interfaces.

**Chapter 4.** We report a method for determining the structure of biomineral-associated proteins. The method combines ssNMR and computational structure prediction in an iterative fashion. In addition, the algorithm is able to identify lattice geometries most compatible with ssNMR constraints. We use this new method to determine the structure of human salivary statherin interacting with HA\(p\). Algorithmic predictions and experiment converge quickly. This is the first structure of a biomineral-adsorbed protein determined with such high accuracy.

**Chapter 5.** We create and test a computational method to design protein-biominalization systems. The algorithm folds a protein from a fully extended structure and simultaneously optimizes the fold, orientation, and sequence of the protein adsorbed to a crystal surface. We used the algorithm to design peptides (16 residues) to modify calcite (CaC\(O_3\)) crystallization. We chemically synthesized six peptides that were predicted to bind two different states of a calcite growth
plane. All six peptides dramatically affected calcite crystal growth (as observed by scanning electron microscopy), and the effects were dependent on the targeted state of the \{001\} growth plane. Additionally, we synthesized and assayed scrambled variants of all six designer peptides to distinguish cases where sequence composition determines the interactions versus cases where sequence order (and presumably structure) plays a role. Scrambled variants of negatively charged peptides also had dramatic effects on calcite crystallization; conversely, scrambled variants of positively charged peptides had a variable effect on crystallization, ranging from dramatic to mild. Special emphasis is often placed on acidic protein residues in calcified tissue mineralization; the work presented here suggests an important role for basic residues as well. In particular, this work implicates a potential role for basic residues in sequence-order specificity for peptide-mineral interactions.

**Chapter 6.** Finally, I summarize my key contributions to the field of biominalization, the impact of those contributions, and provide ideas for future directions in structure determination and design of biominal-associated proteins.
Chapter 2

Structure prediction of protein—solid surface interactions reveals a molecular recognition motif of statherin for hydroxyapatite

Previously published as:


Reprinted with the permission of the publisher. I have abridged the introduction of the original version.

### 2.1 Introduction

In this chapter, we describe a new method for studying protein-surface interactions based on protein structure prediction approaches. The protocol, RosettaSurface, exploits the Rosetta energy function (dominated by van der Waals, solvation, and hydrogen-bonding energies) that has been applied successfully in biological problems (Schueler-Furman *et al.*, 2005) such as protein folding (Bradley *et al.*, 2005a, Bradley *et al.*, 2005b), design (Kuhlman *et al.*, 2003), and docking (Gray *et al.*, 2003b, Daily *et al.*, 2005, Chaudhury *et al.*, 2007). Rosetta employs rapid structure prediction techniques that sample conformation space broadly using discrete fragment and side-chain rotamer libraries, a semiempirical energy function based on physical molecular potentials and
statistics from high-resolution protein structures, and continuous minimization methods. This unique combination of strategies samples extensively while maintaining an all-atom description of the entire system.

We apply our new method to human salivary statherin on hydroxyapatite (HAp). HAp (Figure 2.1a) is a calcium phosphate mineral and is the primary inorganic component of both bone and tooth enamel. Statherin, a 43-residue doubly phosphorylated (Sep2, Sep3; Sep is an abbreviation for phosphorylated serine) human salivary protein (Figure 2.1b), binds HAp in vivo and plays an important role in crystal growth inhibition. This system is an ideal test for RosettaSurface for two reasons. First, since statherin evolved to bind HAp and the adsorption is known to be reversible (Goobes et al., 2006b), one or a few configurations of the statherin-HAp complex are likely to dominate the molecular recognition event. Therefore, we seek the global free-energy minimum conformation of surface-bound statherin. Second, the structure of statherin on HAp has been probed experimentally, so we can assess the accuracy of the predictions and extend existing models. Experimental measurements that give insight into the adsorbed microstructure include fragmentation and substitution experiments to localize the interactions (Raj et al., 1992), intramolecular NMR measurements while adsorbed (Long et al., 1998a, Long et al., 2001), dynamic NMR studies which demonstrate broad regions that interact with HAp (Long et al., 2001), solid-state NMR (ssNMR) measurements that provide distance constraints between side chains and the surface, and thermodynamic measurements on statherin mutants (Gibson et al., 2006, Goobes et al., 2007b).
Figure 2.1: Structure of hydroxyapatite and statherin. (a) Crystal structure of hydroxyapatite, Ca$_5$(PO$_4$)$_3$(OH), {001} plane with lattice parameters a) 9.4 Å, b) 18.8 Å, c) 6.9 Å (Elliott et al., 1973). Coloring (for all HAp figures): Ca, green; P, orange; H, white; O, dark red for phosphate group O atoms that function as hydrogen bond proton acceptors and red for O atoms which can function as either proton donors or acceptors. (b) Statherin structure selected from the set of Goobes et al. (Goobes et al., 2006a) and refined using intramolecular experimental distance constraints (see methods). Sticks depict the two phosphoserine side chains Sep2 and Sep3. Through these studies, statherin is known to have an α-helix near the N-terminus, and several cartoon models have been postulated (Raj et al., 1992,
Our goal is to develop methods for modeling proteins on solid surfaces. To our knowledge, this is the first time a comprehensive suite of tools from the protein structure prediction field has been applied to the protein-surface interaction problem. In this study, we aim to predict the structure and orientation of statherin on a HAp surface, to identify contributions to the binding energy, and to provide a framework for additional directed investigations by ssNMR and mutagenesis experiments. The predicted models give a clearer picture of the statherin-HAp complex and the binding mechanism that validates and extends current experimental hypotheses.

2.2 Computational Methods

**Statherin Model.** For the RosettaSurface simulations presented in this paper, the statherin backbone was held fixed. This assumption allows for more sampling of rigid-body space and tests the extent of agreement with experiment that can be obtained from a fixed-backbone simulation. Typically, Rosetta structure prediction applications, such as RosettaDock, RosettaDesign, and RosettaAbInitio, are parametrized and tested using large data sets of high-resolution protein crystal structures from the Protein Data Bank (PDB) (Berman et al., 2002). Unfortunately, no such benchmark exists for protein-surface simulations, and for this study validation depends on solid-state NMR distance measurements and binding assays.

Candidate low-resolution structures of statherin were predicted previously.
using Rosetta (Gibson et al., 2006, Goobes et al., 2006a). To prepare an input structure for docking, we chose the best-scoring structure from the largest cluster of similar structural models (31 “decoys”). This model agrees with intramolecular atomic distances determined by ssNMR (Goobes et al., 2006a) and circular dichroism studies (Naganagowda et al., 1998) that indicate that some or most of the 15 N-terminal residues are \( \alpha \)-helical. The agreement with experimental data obtained from surface-adsorbed statherin and the fact that Rosetta selects for compact folded structures suggests that the model represents the adsorbed structure of statherin rather than the solution-phase structure, which is suspected to be extended. Furthermore, prior to docking we refined the model with Rosetta (Tsai et al., 2003, Rohl et al., 2004) to obtain a high-resolution, all-atom statherin structure at a local energy minimum using the intramolecular distance constraints measured by ssNMR: Pro33 \( C_\alpha \) to Tyr34 \( C_\alpha \), 3.0-3.25 Å; Pro33 \( C_\alpha \) to Tyr38 N, 5.0-5.8 Å; Tyr34 \( C_\alpha \) to Tyr38 N, 3.5-4.5 Å; and Tyr34 \( C_\alpha \) to Pro23 \( C_\alpha \), 10-11 Å (Goobes et al., 2006a). Rosetta refinement optimizes the structure’s energy via combinatorial rotamer packing and gradient-based minimization of backbone and side-chain torsion angles (Tsai et al., 2003, Rohl et al., 2004).

**Hydroxyapatite Model.** Biological HAp is amorphous on a 100 nm length scale but is approximately crystalline over the much smaller length-scale of the protein (<2 nm) (Raghunathan et al., 2006). Stoichiometric synthetic HAp crystallizes in the monoclinic space group at room temperature and is closely related to the hexagonal (high temperature) HAp crystal structure except for the
loss of inversion symmetry (Raghunathan et al., 2006, Hochrein et al., 2003).

HAp’s fastest growth rate occurs in the direction of the $c$ axis (Figure 2.1a) by deposition of ions in the \{001\} plane (Barnett and Strickland, 1979, Margolis et al., 2006, Simmer and Fincham, 1995, Zhan et al., 2005). The \{001\} plane of HAp is the most stable face under dehydrated and hydrated conditions and is the major experimental cleavage plane (Mkhonto and Leeuw, 2002, Simon et al., 2004). We generated coordinates for HAp ($P2_1/b$ monoclinic) with the \{001\} face exposed using CrystalMaker software (Palmer, 2003). While solid-state NMR can distinguish neither the face for statherin adsorption nor whether statherin binds to a flat surface face, step edge, or defect, both experimental (Long et al., 2001, Hauschka and Carr, 1982, Neves et al., 2002) and simulation studies (Huq et al., 2000, Wright et al., 2004, Zahn and Hochrein, 2003) have assumed \{001\} to be the binding plane and the predominant surface at the enamel-saliva interface (Simmer and Fincham, 1995, Mkhonto and Leeuw, 2002, Zahn and Hochrein, 2003, Radlanski et al., 1989). The lattice parameters of the unit cell (Elliott et al., 1973) are shown in Figure 2.1a.

**Simulation Algorithm.** As with other protein structure prediction problems, the primary scientific challenges for predicting the structure of proteins on solid surfaces are (1) sampling the immense conformational space of the protein on the surface and (2) accurately calculating a potential function to identify the lowest free-energy structure. The algorithm we developed is designed to address these challenges.
Figure 2.2 shows the flowchart of the new RosettaSurface protocol. First, the protein is randomly oriented and positioned away from the surface where there is no intermolecular interaction. The protein is then translated (1) parallel to the surface an integer combination of HAp unit cell vectors $\mathbf{a}$ and $\mathbf{b}$ and (2) normal to the surface to position the protein in glancing contact over the central unit cell (Figure 2.3). The simulation algorithm is a modified version of the all-atom refinement stage of the Monte Carlo plus-minimization (MCM)-based (Li and Scheraga, 1987) RosettaDock algorithm (Gray et al., 2003a). Each cycle of the MCM algorithm begins with a Gaussian distributed random translation of mean 0.1 Å in each Cartesian direction and a Gaussian distributed random rotation of mean 0.05 radians around each Cartesian axis with the rotation origin at the centroid of the interface residues. Next, the interfacial side chains are optimized with a rotamer packing algorithm using an expanded rotamer set (More, 1997, Kuhlman and Baker, 2000) and off-rotamer minimization in torsion space as described by Wang et al. (Wang et al., 2005) During this step, only the residues with side-chain centroid positions less than 7-10 Å (depending on the amino acid size) from the interface are repacked (Gray et al., 2003a, Wang et al., 2005). Finally, the rigid-body translation and rotation are optimized using a gradient-based, iterative Davidon-Fletcher-Powell minimization (Davidon et al., 1963). The minimized structure is compared with the previously accepted structure using the standard Metropolis criterion to determine whether to accept or reject the new position. The MCM procedure, which optimizes both side-chain and rigid-body positions in each iteration, is repeated for 50 cycles, resulting in a final
“decoy” structure.

Figure 2.2: RosettaSurface flowchart.

The best-scoring 200 decoys are retained and compared based on the root-mean-square deviation (rmsd) of the statherin $C_\alpha$ atoms after superposition of the HAp and optimal translation across equivalent crystal surface unit cells (Figure 2.3). Sets of decoys within 2.5 Å rmsd of each other are designated as a cluster using a hierarchical clustering algorithm (Ihaka and Gentleman, 1996). Good clustering indicates convergence of the simulation and favorable conformational entropy.
Figure 2.3: Crystal unit-cell tracking vectors. To account for the symmetry of the surface, the protein is moved to the central unit cell of the surface at the beginning of the simulation and periodically during docking. The \textbf{a} and \textbf{b} vectors are the unit cell crystal vectors parallel to the solid surface. This operation is necessary to calculate rms deviations between structures and to ensure that the surface edge does not affect potential calculations. Surfaces with rotational or inversion symmetry require similar corresponding operations.

Protein Energy Function. The Rosetta energy function has been optimized using several distinct structure prediction applications including folding, docking, and design (Rohl \textit{et al.}, 2004). In brief, the all-atom protein energy function includes van der Waals interactions using a modified Lennard-Jones potential (Gray \textit{et al.}, 2003a); solvation using a pairwise Gaussian solvent-exclusion model (Lazaridis and Karplus, 1999); hydrogen-bonding energies using an orientation-dependent function derived from high-resolution protein structures and quantum calculations (Kortemme \textit{et al.}, 2003, Morozov \textit{et al.}, 2004); side-chain internal energies using a rotamer probability term (Dunbrack Jr and Cohen, 1997); residue-residue pair interactions derived statistically from a database of
protein structures (Wang et al., 2005); and an electrostatic term with a distance-dependent dielectric constant and a cutoff of 5.5 Å (Warshel et al., 1984). The energy function includes the solvent entropy explicitly but neglects the conformational entropy of the protein. The energy terms are weighted as in protein-small-molecule docking studies (Meiler and Baker, 2006).

Since statherin contains two phosphoserines (Sep) crucial for the interaction with HAp (Figure 2.1a), we added and parameterized a phosphorylated serine residue for the Rosetta package. Lennard-Jones parameters and partial charges for electrostatics are obtained from CHARMM27 nucleic acid parameters (Foloppe and MacKerell, 2000, MacKerell and Banavali, 2000). Sep phosphate oxygen atoms assume an sp$^3$ hybridization state and function as proton acceptors in the hydrogen bond calculation. The rotamer database for the side-chain packing algorithm requires appropriate sets of side-chain torsion angles (Dunbrack Jr and Cohen, 1997). Therefore, to sample the rotamer conformations with the appropriate probability, we analyzed all of the 118 protein structures with Sep residues in the Protein Data Bank (Berman et al., 2002). Figure 2.4 shows histograms of the observed Sep $\chi$ angle frequencies, and Table 2.1 lists the average $\chi$ angles and standard deviations used for the side-chain packing algorithm. Due to the small number of structures, $\chi_1$, $\chi_2$, and $\chi_3$ are assumed to be independent.
Figure 2.4: Phosphoserine rotamer statistics. Statistics for the three side-chain torsion angles ($\chi_1, \chi_2, \chi_3$) of phosphoserine (Sep) collected from 118 proteins in the Protein Data Bank.
<table>
<thead>
<tr>
<th>Residue</th>
<th>$\chi_1$</th>
<th>$\sigma_1$</th>
<th>$\chi_2$</th>
<th>$\sigma_2$</th>
<th>$\chi_3$</th>
<th>$\sigma_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep</td>
<td>53.8</td>
<td>24.8</td>
<td>98.9</td>
<td>18</td>
<td>61.2</td>
<td>23.9</td>
</tr>
<tr>
<td>Sep</td>
<td>-62.0</td>
<td>28.5</td>
<td>-89.6</td>
<td>24.3</td>
<td>-60.2</td>
<td>22.6</td>
</tr>
<tr>
<td>Sep</td>
<td>-178.0</td>
<td>25.4</td>
<td>-170.2</td>
<td>25.4</td>
<td>-178.6</td>
<td>22.7</td>
</tr>
</tbody>
</table>

Table 2.1: Rotamer parameters to expand the Rosetta rotamer library. Average angles for each side-chain torsional angle ($\chi_i$) for phosphoserine with standard deviations ($\sigma_i$), in degrees. Each $\chi$ angle is assumed to be independent.

Protein-Surface Energy Function. The force field parameters for HAp are chosen as follows. The Lennard-Jones well depth, $\varepsilon$, and the equilibrium internuclear separation, $\sigma$, are taken from the CHARMM22 all-hydrogen parameter file for proteins (MacKerell Jr et al., 1998, Mackerell et al., 2004) and CHARMM27 for nucleic acids (Foloppe and MacKerell, 2000, MacKerell and Banavali, 2000). For purposes of weighting the score function contributions, the Lennard-Jones interactions are partitioned into those interactions which are net attractive ($r > 0.89 \sigma_{ij}$) and those that are net repulsive and cause clashes in the structure ($r < 0.89 \sigma_{ij}$). As in previous Rosetta simulations (Gray et al., 2003a), the repulsive Lennard-Jones term is linearly extrapolated at small radii ($r < 0.6 \sigma_{ij}$) to remove the singularity and improve minimization performance. For the hydrogen-bonding calculation, the HAp phosphate oxygen atoms ($sp^3$ hybridization) are designated as proton acceptors, and the hydroxyl oxygen atoms function as both proton acceptors and donors (Zahn and Hochrein, 2003). The Gaussian solvent-exclusion model for the solvation free energy is extended to the phosphate group atoms with empirical parameters from micelle...
measurements (Lazaridis et al., 2005). The partial atomic charges have been derived from quantum-chemical calculations and dynamical simulated annealing optimization (Hauptmann et al., 2003). A distance-dependent dielectric constant (Warshel et al., 1984) approximates the varying dielectric constant in the system. Electrostatic energies are smoothly truncated at 8.0 Å and capped at a maximum value matching that at the sum of the van der Waals radii of the interacting atoms. The full-atom energy of the protein-surface interaction is a linear combination of attractive and repulsive Lennard-Jones interactions ($E_{\text{att}}$ and $E_{\text{rep}}$), solvation ($E_{\text{sol}}$), hydrogen bonding ($E_{\text{hb}}$), and electrostatics ($E_{\text{coul}}$):

$$E_{p-s} = w_{\text{att}} E_{\text{att}} + w_{\text{rep}} E_{\text{rep}} + w_{\text{sol}} E_{\text{sol}} + w_{\text{hb}} E_{\text{hb}} + w_{\text{coul}} E_{\text{coul}}$$

Weights for combining the energies have been fitted to experimental data of multiple ligand molecules (updated set since the original paper (Meiler and Baker, 2006), personal communication, Jens Meiler) $w_{\text{att}} = 0.8$, $w_{\text{rep}} = 0.4$, $w_{\text{hb}} = 2.0$, $w_{\text{sol}} = 0.6$, $w_{\text{coul}} = 0.25$.

**Structure Evaluation.** Candidate decoy structures are inspected using a combination of external postprocessing tools to ensure structures display “protein-like” features. Of particular interest are interfacial shape complementarity, interfacial voids that are inaccessible to solvent, and a completely satisfied hydrogen bond network.

Shape complementarity was assessed using the Fast Atomic Density Evaluator (FADE) (Kuhn et al., 1992, Mitchell et al., 2001). FADE creates a lattice of points at the molecular interface about which each molecule is proximally
determined to present a crevice, protrusion, or flat surface by calculating a
density exponent $d$ characterizing the rate of increase in atomic density, $\rho$, as a
function of radius, $r$, where $\rho \approx r^d$. Then, $d < 2.8$ is defined as a protrusion, $d >$
2.8, a crevice, and $d \approx 2.8$, as flat. Density exponents are converted to a score
using the following equation $s = (d1 - d0)(d2 - d0)$, where $d1$ and $d2$ are the density
exponents of molecules 1 and 2, and $d0$ is the density exponent for a volume of
perfectly packed spheres (3.0 in theory, in practice set to 2.8 to reflect average
efficiency of protein packing). A negative value for $s$ represents good shape
complementarity and is obtained only when an intermolecular crevice and
protrusion meet at an interface.

Solvent accessibility was calculated using NACCESS version 2.1.1 (Hubbard
and Thornton, 1993). Since Rosetta models solvent implicitly, NACCESS helps
determine the probability of finding a water molecule at an interfacial cavity and
satisfying a hydrogen bond. NACCESS uses the “rolling ball” method (Lee and
Richards, 1971) with a probe radius of 1.4 Å. Protein and surface atomic radii
were set to RosettaSurface van der Waals radii, and all hydrogen atoms were
included in the solvent accessibility calculations.

Contour plots of shape complementarity and solvent accessibility are
superimposed on the molecular coordinates and viewed using PyMOL (DeLano,
2002) to reveal regions of the interface where poor solvent accessibility and poor
shape complementarity are coincident. Regions displaying poor shape
complementarity should be accessible to solvent to avoid vacuum at an interface.
Similarly, solvent accessibility plots, hydrogen bond networks, and molecular coordinates are superimposed to ensure that all hydrogen bond donors and acceptors that are not involved in inter- or intra-molecular hydrogen bonds can be satisfied by solvent.

2.3 Results

The RosettaSurface simulation generated 80,000 decoy structures in approximately 150 CPU-days on a cluster of 2.4 GHz Linux processors. Considering that each decoy completed 50 MCM cycles, $4 \times 10^6$ minima were sampled in the rigid-body conformation space. Furthermore, $\sim 10^3$ conformations were sampled at each rigid-body position during the side-chain packing steps.

Figure 2.5 shows scores of the top quartile of the 80,000 decoys as a function of the distance from an arbitrary starting position. The sampling is extensive, and several local minima emerge. Of the 200 top-scoring decoy structures, 171 comprise four clusters using a cluster radius of 2.5 Å rmsd. Clustering statistics gathered periodically during the simulations indicate that the main clusters had not changed significantly during the last 40% of the run. The four largest clusters contain 138, 17, 9, and 7 members, and members of the four largest clusters are indicated with colors in Figure 2.5. The largest cluster includes the best scoring structure. The clustering indicates that the simulation converges on several related structures, predicting specific protein-surface orientations with favorable conformational entropies about the global energy minima.

Nine residues (Sep2, Sep3, Lys6, Arg9, Arg10, Arg13, Tyr16, Gln32, and Pro33)
appear repeatedly at the interface in the top-scoring structures. Figure 2.6 shows the distribution of several statherin atom-HAp atom distances among the 200 top-scoring structures for these nine residues plus residues for which distances have been measured experimentally (Lys6, Phe7, Phe14). These histograms reveal a broad similarity among structures within clusters as well as between clusters. Distances to Lys6, Arg10, Arg13, and Gln32 are all under 5Å and narrowly distributed. Arg9 is similar, except that structures in cluster 2 show larger distances. The Sep residues and Pro33 have broader distributions from 4 to 8 Å and a few decoys with larger distances.

![Figure 2.5: Statherin-HAp energy landscape.](image)

Scores of the top quartile of the 80,000 decoys are plotted as a function of the distance (rmsd) from an arbitrary starting position. The top-scoring 200 decoys are grouped into clusters indicated by color.
Figure 2.6: Distributions of statherin-HAp distances among the 200 top-scoring structures. Distances are measured from the nearest P atom of HAp to the P in Sep, the N\textsubscript{e} in Lys, the center of the aromatic ring in Phe, the nearest N\textsubscript{h} in Arg, the hydroxyl O in Tyr, the N\textsubscript{e} in Gln, and the N in Pro. For Lys6, the distance to the nearest and next-nearest HAp P is shown. Grey regions indicate distances measured experimentally by solid-state NMR (Gibson et al., 2005, Gibson et al., 2006, Raghunathan et al., 2006), and bar colors denote decoy clusters. Stars indicate bins containing model 1.
For Lys6, the nearest and next-nearest N\textsubscript{\textgamma}-P distance are both plotted to allow comparison with models proposed from ssNMR measurements (grey regions in Figure 2.6). Both distances are within the experimental range for most decoys; however the decoys show more equal nearest and next-nearest distances compared to that proposed by Gibson \textit{et al.} (Gibson \textit{et al.}, 2005). The Phe residues have also been experimentally probed. The simulation decoys show broad distributions, with most of the Phe7 aromatic rings being over 6.5 Å from the surface in agreement with the ssNMR measurements (Gibson \textit{et al.}, 2006). However, the structures have Phe14 distances a few angstroms above the experimental range in all models.

Figure 2.7 shows distributions of the protein-surface intramolecular energies over the 200 top-scoring models for each of the residues near the interface. The basic residues (particularly Arg13) dominate the energetics, although there is considerable variation in individual energies between models. The Sep residues both have bimodal distributions with some models having large energies and some having almost negligible energies. For example, the cluster 2 models have excellent Sep3 energies of 6-8 kcal/mol that perhaps are achieved at the expense of Sep2 energies, which are limited to 1-3 kcal/mol.

In a typical Rosetta simulation, a single structure with both the lowest score (global free energy minimum) and best clustering (most favorable conformational entropy) suggests accurate results (Gray \textit{et al.}, 2003a). In this case, there is such a structure that scores better than any other and clusters with 138 of the 200 top-scoring decoys. While many of the models are plausible by energy, the size
difference of the top clusters coupled with extensive sampling carried out suggests the relevance of this single structure. With limited experimental constraints available, it is impossible to determine which model(s) is (are) likely to be experimentally relevant. To simplify the analysis we focus on that single, dominant structure and refer to it as model 1; later we return to the other models to identify similar and distinguishing characteristics.

**Figure 2.7:** Statherin-HAp interaction energy distributions among the 200 top-scoring structures for selected residues. Bar colors denote decoy clusters, and stars indicate bins containing model 1.

Figure 2.8 shows top and side views of the global orientation of model 1. The system buries 849 Å^2 of statherin’s 4343 Å^2 of solvent-exposed surface area, and the N-terminal helix is pointed from N-to-C in a direction about 30° to the right of the negative a axis (Figure 2.1a). In Figure 2.8b, the semitransparent surface surrounding the cartoon (backbone) representation of statherin represents the
van der Waals surface of the protein and reveals the interfacial shape complementarity. To highlight the global charge distribution, Figure 2.8b also shows the electrostatic potential of isolated statherin superimposed on model 1 as calculated by PyMOL’s vacuum electrostatics module (DeLano, 2002). The HAp surface bears a net negative charge and complements the positively charged binding surface of model 1.

In Figures 2.6 and 2.7, stars mark the bins containing model 1 measurements. Table 2.2 further details the dominant energy terms and percent buried surface area for selected residues at the statherin-HAp interface in model 1. Hydrogen-bonding energy is the largest favorable term followed by van der Waals and then electrostatic energies, offsetting the unfavorable desolvation penalty. The individual residue energies that underlie the molecular recognition phenomenon are noted in detail below.

<table>
<thead>
<tr>
<th>amino acid</th>
<th>van der Waals</th>
<th>solvation</th>
<th>hydrogen bonding</th>
<th>coulomb</th>
<th>total</th>
<th>% desolvated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep2</td>
<td>−1.9</td>
<td>1.2</td>
<td>−2.5</td>
<td>−0.4</td>
<td>−3.5</td>
<td>66</td>
</tr>
<tr>
<td>Sep3</td>
<td>−0.9</td>
<td>0.4</td>
<td>0.0</td>
<td>−1.3</td>
<td>−1.8</td>
<td>71</td>
</tr>
<tr>
<td>Lys6</td>
<td>−3.1</td>
<td>4.8</td>
<td>−5.7</td>
<td>−3.0</td>
<td>−6.9</td>
<td>93</td>
</tr>
<tr>
<td>Arg9</td>
<td>−2.8</td>
<td>3.5</td>
<td>−7.1</td>
<td>−1.4</td>
<td>−7.8</td>
<td>43</td>
</tr>
<tr>
<td>Arg10</td>
<td>−3.9</td>
<td>4.6</td>
<td>−5.5</td>
<td>−1.9</td>
<td>−6.6</td>
<td>93</td>
</tr>
<tr>
<td>Arg13</td>
<td>−4.4</td>
<td>5.7</td>
<td>−10.5</td>
<td>−2.1</td>
<td>−11.2</td>
<td>91</td>
</tr>
<tr>
<td>Tyr16</td>
<td>−1.8</td>
<td>1.0</td>
<td>−2.1</td>
<td>−0.1</td>
<td>−2.2</td>
<td>57</td>
</tr>
<tr>
<td>Gln32</td>
<td>−2.2</td>
<td>1.6</td>
<td>−2.7</td>
<td>−0.1</td>
<td>−3.4</td>
<td>85</td>
</tr>
<tr>
<td>Pro33</td>
<td>−0.8</td>
<td>0.3</td>
<td>0.0</td>
<td>−0.1</td>
<td>−0.7</td>
<td>86</td>
</tr>
<tr>
<td>total</td>
<td>−21.7</td>
<td>23.2</td>
<td>−36.1</td>
<td>−10.5</td>
<td>−45.0</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.2: Residue-surface interactions.** Energies (kcal/mol) and percentages of the residues’ solvent-accessible surface areas which are buried by Adsorption for the Best Statherin-HAp Structure (Model 1).
Figure 2.8: Orientation of model 1. (a) Top view. (b) Side view showing the shape complementarity and charge distribution of statherin (calculated in isolation by PyMOL’s vacuum electrostatics module).

The RosettaSurface simulations reveal a promising molecular recognition motif for statherin binding on HAp. In the left pane of Figure 2.9, white circles encapsulate an atom cluster we denote as the Interstice of the Phosphate-Oxygen Triad (IPOT). IPOTs occur wherever three phosphate oxygens meet on the HAp surface, forming a negative charge center with a trivalent hydrogen bond capacity. The pocket has a depth of \( \sim 1.6 \text{ Å} \) and a diameter of \( \sim 2.1 \text{ Å} \) (inset).
In Figure 2.9, the white parallelograms (left) show the periodicity of the IPOT motif replicated across the HAp surface; the dimensions measure approximately 9.4 Å by 9.4 Å in a regular parallelogram with an acute angle of 60°.

Figure 2.9: The IPOT motif. Left: Repetition of the IPOT pattern the HAp surface (white parallelograms). Inset: Close-up view of an IPOT. Right: Four basic amino acids which comprise the IPOT recognition motif.

Remarkably, in model 1 this geometry matches perfectly that displayed by the four basic amino acids of statherin (Figure 2.9, right). All four basic amino acids, Lys6, Arg9, Arg10, and Arg13, fall directly into an IPOT. These residues provide the largest contribution to every favorable energy term (van der Waals, electrostatics, and hydrogen bonding) and cumulatively the largest contribution to the total binding energy (Table 2.2). Lys6 is positioned such that all polar side-chain hydrogens are donated to HAp phosphate oxygens (Figure 2.10a). Arg9 and Arg10 both form hydrogen bonds to HAp phosphate oxygens with two of their three side-chain nitrogen atoms (Figure 2.10a,b), similar to the “arginine fork” observed in the recognition of nucleic acid backbone phosphates (Calnan et al.,
Arg13 exhibits the strongest hydrogen bond energy and the strongest van der Waals interaction, combining to create the strongest overall interaction (-11.2 kcal/mol). All hydrogen bond donors and acceptors in the interface are fulfilled explicitly or accessible to solvent, except for one donor from the N$_i$ of Arg13. Lys6, Arg10, and Arg13 are over 90% desolvated; Arg9 retains about half of its solvent accessibility (Table 2.2). Figure 2.11 shows the shape complementarity as measured using FADE; the basic amino acids fit well into the IPOT pockets and score very highly.

**Figure 2.10: Selected structural details of model 1.** (a) N-terminal and (b) C-terminal view of basic, aromatic, and phosphoserine residues with hydrogen bonding HAp atoms in color. (c) Sep2 makes hydrogen bonds and Sep3 chelates a Ca$^{2+}$ ion (green).
Figure 2.11: Map of interface FADE scores. The four basic residues of the IPOT recognition motif (outlined in dots) and the two phosphoserines show shape complementarity typical of protein-protein interfaces.

Previous studies (Raj et al., 1992) have suggested a special HAp recognition role for statherin’s N-terminal Sep residues. Considering that the adsorption of Sep residue phosphates to the growing HAp surface would closely mimic the deposition of phosphate ions from solution, Sep residues could be partially responsible for statherin’s crystal growth inhibitory capabilities. In model 1, both of the phosphoserines (Sep2, Sep3) interact strongly with the HAp surface (Table 2.2, Figure 2.7). Figure 2.10c shows the direct interaction of both phosphoserine residues with the HAp surface. Sep3 is striking in that its tetrahedral phosphate coordinates a HAp Ca$^{2+}$ atom in a fashion similar to that of phosphate groups in the bulk of the HAp crystal. The favorable portion of this interaction is dominated almost entirely by electrostatics and to a lesser extent by van der Waals forces (Table 2.2). The interaction of Sep2 is mediated in a different and rather intricate manner. Sep2 is the only residue in which an intermolecular hydrogen bond pair has statherin as the proton acceptor and HAp as the donor.
One of the Sep2 phosphate oxygens accepts a hydrogen bond from a HAp hydroxyl atom in a moderately favorable interaction. The Sep2-HAp interaction is further reinforced by a moderate attractive Lennard-Jones component and a weak electrostatic interaction.

Tyr16 at the end of the N-terminal helix and Gln32 in the C-terminus (Figure 2.10a,b) are moderately strong donors and interact with moderate attractive van der Waals forces (Table 2.2). Pro33’s only favorable energy term is a relatively weak attractive van der Waals interaction. This prediction agrees with experimental observations that the C-terminus interacts with HAp more strongly than the central region but weaker than the N-terminus of statherin (Raj et al., 1992, Long et al., 2001). Gln32 interacts with an IPOT, and its binding might help constrain the global orientation of the statherin. For all three residues, Tyr16, Gln32, and Pro33, the electrostatic contribution to the total binding energy is negligible compared to hydrogen bond contribution.

The difference in interface quality between the N-terminal region and the rest of the protein is also shown in the FADE scores, which measure shape complementarity. For reference, the FADE scores in the set of protein-protein interfaces from the docking benchmark set 2.0 (Mintseris et al., 2005) average -0.103 (0.038 per lattice point. In model 1 of statherin on hydroxyapatite, the average FADE score for the entire interface of model 1 is poor, at -0.014 per lattice point. However, the score for the N-terminal region is -0.062 per lattice point, almost within the standard deviation range of protein-protein interfaces.

Many of the structures in the set of 200 top-scoring decoys share recognition
features with model 1. As seen in the histograms of distances and energies, most models contain the IPOT-recognition motif with each of the four basic residues binding an IPOT pocket. Exceptions are the cluster 2 structures, in which Arg9 does not make contact, and the cluster 4 structures, in which Arg13 and Arg9 share a single IPOT in a less-optimal fashion. Figure 2.12 shows the orientational distribution of the top 200 decoys. Cluster 2 and 3 structures are rotated in the HAp {001} plane by 60° and 90°, respectively, relative to cluster 1. In these different orientations, it can be difficult to make energetically favorable interactions with all of the surface-binding groups, and there is a distribution of better and less-optimal interactions (Figure 2.7). Figure 2.13 shows a top-view comparison of the four model structures from each of the four largest clusters (models 1-4).

![Graph showing distribution of orientation](image)

**Figure 2.12:** Distribution of orientation of statherin on the HAp surface among the 200 top-scoring structures. Angle of the Lys6 Cα to Arg13 Cα relative to the line along the long axis of the IPOT diamond as depicted in Figure 2.9.
Figure 2.13: Top view of the four structures. Models are the lowest-scoring decoys from each of the four largest clusters. Top left, model 1; top right, model 2; bottom left, model 3; bottom right, model 4.

2.4 Discussion

Many studies have shown diverse face and phase recognition of an inorganic solid by proteins (Jia et al., 1996, Fujisawa and Kuboki, 1991, DeOliveira and Laursen, 1997, Belcher et al., 1996). Molecular recognition of a protein-solid surface system may differ from that of a protein-protein or protein-small molecule system. Proteins draw from the chemical complexity of 20 unique amino acids that can be arranged in numerous combinations in primary and tertiary space, and protein topography is geometrically complex, usually not flat or repetitive. Therefore, a protein-protein recognition motif is rarely a repetitive
pattern, but rather a complex arrangement of residues of side-chain lengths, hydrophobicities, pKa’s, and ionic strengths exquisitely complemented by the evolved binding partners. In contrast, solid crystal surfaces typically present a pattern whose chemistry and geometry is repetitive on the scale of a unit cell. Therefore, one may expect a surface-binding protein to possess a complementary pattern in the tertiary arrangement of its side chains (Mann, 1988). Furthermore, biomaterials such as silica, calcite, and HAp typically terminate ionically and yield a net surface charge. If a protein is to complement the charge at the interface, it must bear an array of proximal like-charges in its binding domain. Such arrangements have been observed in antifreeze proteins (Liou et al., 2000) and osteocalcin (Flade et al., 1999).

Statherin’s array of proximal basic residues of one lysine and four arginines in residues 6-13 (KFLRRIGR) with no intervening acidic residues would be expected to destabilize the secondary and tertiary structure. However, residues 2-12 are measured to be helical by NMR (Long et al., 2001), and Rosetta predicted residues 3-15 as helical (Goobes et al., 2006a). When the above sequence is threaded on an α-helical backbone, the basic residues are positioned perfectly to mimic the periodicity of acidic phosphate clusters on the HAp surface. In the RosettaSurface models, this lattice-matching provides a significant contribution to the free energy of adsorption. While the thermodynamic contribution of the basic residues has been measured (Goobes et al., 2007b), to our knowledge this is the first structural identification of this molecular recognition motif. Most previous studies have focused on the Sep residues, although it has been
previously noted that the Sep residues can be substituted by similarly charged Asp residues and still retain function (Raj et al., 1992).

In previous Rosetta studies, alternate force fields (choice of energy functions, parametrizations, and weightings) have been tested using known protein structures from the Protein Data Bank (Berman et al., 2002). After using unvalidated selections of force field parameters from the literature, it is encouraging that the results are broadly consistent with known experimental data. Early experiments on statherin fragments (Raj et al., 1992) and more recent dynamic NMR measurements (Long et al., 2001) identify the 15 N-terminal residues as the primary contributors to the HAp adsorption free energy; in fact, the 15-mer is sufficient for α-helix formation, HAp binding, and inhibition of HAp growth (Raj et al., 1992). Similarly, in the final models, the N-terminal residues dominate the interface and contribute the most to the binding energy while the intervening and C-terminal regions contribute weakly. Single amino-acid experiments have shown that Sep, Arg, Lys, Asp, Glu, Tyr, and Gln interact strongly with HAp (Moreno et al., 1984, Koutsopoulos and Dalas, 2000a, Koutsopoulos and Dalas, 2000b, Koutsopoulos and Dalas, 2000c, Kresak et al., 1977, Tanaka et al., 1989); these residue types comprise the majority of the 15-residue N-terminal fragment. Furthermore, Lys was shown to inhibit HAp crystal growth (Koutsopoulos and Dalas, 2000a), which is thought to be the primary function of statherin in vivo, further implicating Lys’ role in HAp binding. Finally, calorimetric data of mutated statherin fragments verify the importance of the basic amino acids (Lys6, Arg9, Arg10, Arg13) in HAp binding (Goobes et al.,
RosettaSurface both samples and selects for decoys that are in agreement with these experimental results.

The current experimental picture of the statherin-HAp system is one of low resolution, and many of the predicted structures broadly match this picture despite differences up to a 20 Å rmsd. Despite the fact that hydroxyapatite’s {001} surface does not exhibit any formal rotational symmetry, the existence of different clusters of low-energy structures rotated 60° and 90° reveals similar features on the surface across different rotations. New measurements of the residues with broader distance distributions (e.g., Sep2, Sep3, Tyr16, and Pro33) would help to discriminate among the current models, but it will remain difficult to distinguish between the set of top structures due to the pseudo-rotational symmetry of the surface.

The simple model predicted here captures an encouraging amount of the previous experimental data, but there are many aspects of the system still not addressed fully. The statherin backbone conformation was determined by de novo predictions and held fixed during docking. While the structure agrees with solid-state NMR measurements of backbone distances in the adsorbed state, there are several angular measurements that are not consistent with the predicted backbone (Goobes et al., 2006a). Statherin is known to undergo a structural transition upon adsorption, so it is reasonable that accurate backbone structures are only achievable in the presence of the hydroxyapatite. Simultaneous backbone and rigid-body optimization of statherin was beyond the scope of this study but would be an important goal for future studies.
flexibility might resolve the suboptimal interactions of some side-chain groups (e.g., Sep3 in model 1) while retaining the optimal interactions of other groups. One of the currently dominant clusters might be better at achieving a higher degree of electrostatic and shape complementarity, especially beyond the N-terminal region, resolving the remaining uncertainties in the protein orientation. Also, some unraveling of the end of the first helix might allow the structure to meet the experimentally measured Phe14-HAp distance and allow solvent access to the one unsatisfied hydrogen-bonding group in Arg13 while maintaining the IPOT recognition motif.

Additional remaining questions about the statherin-hydroxyapatite system concern the HAp surface. In this study, we assumed that statherin binds to a clean {001} monoclinic face of HAp terminated in a balance of Ca\(^{2+}\) and phosphate groups, as expected in many in vitro experiments. However, the crystallographic binding face of biological HAp for statherin has not been unambiguously determined, and crystallographic faces (\{010\}, \{100\}, etc.), lattice types (hexagonal rather than monoclinic), terminations, and step edges or defects may play a role in statherin binding. For example, the deposition of one additional Ca\(^{2+}\) atom at the appropriate location would provide better electrostatic and van der Waals interactions for one of the Sep residues in model 1. Future simulations on this system would include some of these competing scenarios to address these open questions.

The RosettaSurface method developed here has several advantages over previous approaches. The multistart MCM algorithm is an efficient and parallel
method to sample many local minima in the free energy landscape of the protein-surface system. The energy function, while needing validation for protein-surface interactions, has been broadly successful in a variety of protein structure prediction problems. The description is atomically detailed allowing direct examination of structural features and a detailed decomposition of energetic components. The reversible, evolved statherin-hydroxyapatite system revealed a recognition motif, but this is not expected in the general case: prediction of many protein-surface systems could be confounded by the nonequilibrium nature of the adsorption process and the existence of multiple conformational states on the surface. Furthermore, lateral protein-protein interactions are likely to play a role, perhaps even for the statherin-HAp system (Goobes et al., 2006b), requiring additional representations and sampling. Therefore, development of a general tool for protein-surface interactions remains a challenging problem.

2.5 Conclusion

RosettaSurface simulations have captured some important features of the statherin-hydroxyapatite system. Importantly, it supported previous experimental measurements, assigned residue-specific energy contributions to the known HAp interaction domain of statherin, and identified the IPOT recognition motif. Knowledge of natural biomaterial recognition will help to rationally design novel protein-surface interactions and harness the medical and engineering potential of such systems. Protein-surface interactions present a
unique challenge to the structural biology community in that no current experimental method affords the ability to solve the molecular structures. Molecular modeling currently provides the only means to investigate the protein-surface interface with atomic resolution. Combinations of experimental and computational approaches are promising routes to reveal the molecular structure of proteins at solid surfaces.
Solution- and adsorbed-state structural ensembles predicted for the statherin-hydroxyapatite system

Previously published as:


Reprinted with the permission of the publisher. I have abridged the introduction of the original version.

3.1 Introduction

In the previous chapter, we modified the Rosetta structure prediction suite (Berrondo *et al.*, 2008, Bradley *et al.*, 2005b, Daily *et al.*, 2005, Gray *et al.*, 2003a, Kuhlman *et al.*, 2003, Sivasubramanian *et al.*, 2008, Meiler and Baker, 2006) to dock proteins to solid surfaces (RosettaSurface) and predicted a set of statherin conformers bound to the {001} monoclinic face of HAp (Makrodimitris *et al.*, 2007). While that work represented a significant improvement in protein structure prediction at solid surfaces, simplifications were necessary to make the system tractable. Specifically, statherin had been prefolded in implicit solvent using Rosetta *ab initio* (Goobes *et al.*, 2006a) and refined using four ssNMR intramolecular distance constraints (acquired from statherin bound to HAp); the docking simulation that ensued assumed a fixed backbone and included only a single crystal face of HAp. One critique of that approach is that the number of
protein conformations satisfying only four intramolecular distance constraints is expected to be large, and selecting only one conformation neglects the effect of HAp on statherin’s backbone conformation. Some of these effects might include surface catalyzed structural transitions and conformational selection by HAp for statherin. Similarly, specificity/promiscuity cannot be thoroughly tested by simulating adsorption to a single HAp face, and it is not known which face(s) are biologically relevant for statherin adsorption. Also, a fixed-backbone simulation requires advanced structural knowledge, an advantage not common in protein-surface studies.

In this chapter we present the first structure-prediction-based algorithm capable of folding a protein in the presence of a solid surface. The algorithm can rapidly generate and energy-minimize protein folds and docked orientations, representing a significant improvement in the conformational space accessible during simulation. Because surfaces are suspected to catalyze structural transitions, this RosettaSurface protocol generates large ensembles in both the solution and adsorbed states. With this protocol we address the following questions: (i) How similar are the solution- and adsorbed-state structures? (ii) How does the protein structure differ at three different HAp surfaces ({001}, {010}, and {100})? (iii) Is the RosettaSurface algorithm robust enough to capture known structural features without any bias beginning from a fully extended peptide chain in solution? Finally, we add ssNMR constraints to the RosettaSurface simulation. The outcome of constrained simulation helps interpret the RosettaSurface and ssNMR structural data and provides proof of
concept for a combined technique to aid in adsorbed-state structure determination.

3.2 Materials and Methods

Statherin. We built an extended molecular structure of human salivary statherin using PyMol (DeLano, 2002). The atomic parameter set is the same as in Makrodimitris et al (Makrodimitris et al., 2007), except in this study the \( \gamma \) oxygen atom of phosphoserine (Sep) is modeled as an esterified oxygen (Foloppe and MacKerrell, 2000).

Hydroxyapatite. We built monoclinic hydroxyapatite crystals using CrystalMaker (Palmer, 2005). Appropriate cuts were made to expose flat \{001\}, \{010\}, and \{100\} surfaces with approximately neutral (mixed charge) terminations (see figure 3.6). The atomic parameter set is the same as in Makrodimitris et al (Makrodimitris et al., 2007), except in this study all HAp surface atoms are included when calculating the solvent exclusion term (Lazaridis and Karplus, 1999).

The Algorithm. The algorithm developed for this study (see flow chart in figure 3.1) is based on the multiscale Metropolis Monte Carlo-plus-minimization (MCM) implementation of the Rosetta docking and folding strategies (Bradley et al., 2005b, Gray et al., 2003a, Simons et al., 1997). The protocol implements an initial, fast, low-resolution step to generate diverse protein folds. Subsequent high-resolution steps were optimized to create mostly clash-free protein-like structures in a short amount of time. The protocol energy minimizes the
solution-state protein structure before adsorbing the protein to the solid surface, thus generating large ensembles of candidate structures (decoys) in both the solution and adsorbed states.

**Figure 3.1:** Flowchart for the RosettaSurface algorithm. *CG min* = conjugate gradient minimization (9); *line min* = line minimization. *small move* = perturbations of randomly selected (Φ,Ψ) pairs. *shear move* = perturbation of randomly selected Φ angle with compensatory perturbation of equal and opposite magnitude applied to previous Ψ angle. *crank move* = globally nonperturbing fragment insertion followed (Φ,Ψ) perturbation at residues adjacent to, and two residues not adjacent to, the insertion window.
There are a total of 5 “refinement cycles” (see below) implemented to generate each solution- and adsorbed-state decoy; for each decoy, adsorption occurs at random between refinement cycles one and five. This choice allows for approximately equal sampling at all stages of refinement in both the solution and adsorbed states when a large number of decoys are generated.

The creation of each decoy begins with the rapid collapse of a fully extended protein molecule using a fragment assembly (FA) protocol. The FA protocol is the same as in Simons et al. (Simons et al., 1997), except in this study we removed the radius of gyration term to allow sampling of extended structures. After FA, the backbone is optimized with six “shear” and ~250 “small” moves (Rohl et al., 2004) coupled with conjugate-gradient (Fletcher, 1963) and line minimization respectively. At all stages thus far the structural model is comprised of an all-atom protein backbone and a single pseudo atom to represent each side chain (Simons et al., 1997).

Next, all-atom (including hydrogen) side chains are built from a backbone-dependent rotamer library (Dunbrack Jr and Cohen, 1997) using a simulated annealing protocol (Kuhlman and Baker, 2000), and the protein undergoes \( n \leq 5 \) cycles of backbone refinement (see figure 3.1, right). Each cycle of refinement includes “outer-“ and “inner-loop optimization”. During outer-loop optimization, a sequence of perturbing small-, shear-, and crank-moves (Rohl et al., 2004) are applied to the protein backbone, each followed by conjugate-gradient minimization (Fletcher, 1963). Next, during inner-loop optimization, the same perturbing moves are applied except each is followed by line minimization along
the initial gradient. For each cycle, the outer loop is implemented five times; for each outer loop, the inner loop is implemented five times.

After $n$ refinement cycles an adsorbed-state complex is formed, by introducing the surface in a random orientation and bringing the protein and surface into contact. The adsorbed-state protein undergoes $5 - n$ cycles of refinement at the interface. Refinement is the same as in the previous paragraph, except each cycle ends with a modified version of RosettaDock’s (Gray et al., 2003a) high-resolution sequence. In this study, a sequence of small rigid-body perturbations, side-chain repacking, and gradient-based minimization in rigid-body space is repeated six times; side-chain repacking is combinatorial, rather than individual, every third time. Hence, the protein undergoes simultaneous backbone, side-chain, and rigid-body optimization on the surface.

Solution-state coordinates are output immediately prior to the formation of the adsorbed state. After adsorbed-state refinement is complete, RosettaSurface implements one round of coupled rotamer packing and minimization (Wang et al., 2005) at all side-chain positions and outputs the adsorbed-state coordinates.

The all-atom energy function is similar to that used previously (Makrodimitris et al., 2007) with the following modifications: (i) To account for the affect of the highly charged binding domain on folding, protein self electrostatics are included using a simple distance-dependent-dielectric model (Warshel et al., 1984). (ii) Because statistically derived energy terms and coefficients (weights) had little effect on final decoy discrimination, both adsorbed- and solution-state decoys
were scored using a nonweighted physical potential comprising van der Waals (Gray et al., 2003a), electrostatic (Warshel et al., 1984), hydrogen bond (Kortemme et al., 2003), and implicit solvent (Lazaridis and Karplus, 1999) interactions. (iii) For ssNMR constrained simulations, a harmonic potential with a spring constant of 10 Rosetta Energy Units (REU)/Å (or 0°) was used to bias predictions toward satisfying all 15 published ssNMR distance and angle measurements (Shaw et al., 2000, Long et al., 2001, Gibson et al., 2005, Gibson et al., 2006, Goobes et al., 2006a, Raghunathan et al., 2006).

**Structural analysis.** The structural designations “Helix” and “Turn” were assigned using the DSSP (Kabsch and Sander, 1983) definitions; classification relied on Rosetta’s hydrogen-bond function (Kortemme et al., 2003) rather than the generalized hydrogen-bond function implemented by the DSSP package. The structural designation “Other” indicates that hydrogen bonding was either long range or absent at that residue. Side-chain solvent-accessible surface area was calculated using Naccess (Hubbard and Thornton, 1993) with van der Waals radii adjusted to match those in the Rosetta force field. Residues were considered “bound” if they had a nonzero contribution to total adsorption energy.

### 3.3 Results

Here we apply the RosettaSurface protocol to statherin interacting with the HAp {001} monoclinic surface. We also demonstrate the algorithmic portability by adsorbing a control peptide to the HAp {001} surface. In addition, we run two NMR-biased simulations (adsorbing statherin to HAp’s {001} face) to assess the
quality of unbiased RosettaSurface predictions and to help infer structure from ssNMR measurements. Finally, we apply the RosettaSurface protocol to two other HAp surfaces ({010} and {100}) to probe statherin specificity.

In all simulations we generated $10^5$ decoys for both the solution and adsorbed state. Approximately $3.5 \times 10^4$ conformers were sampled for each of the $10^5$ decoys generated; therefore, about $3.5 \times 10^9$ conformers were assessed by the RosettaSurface energy function in each “run.” For all runs the top scoring 0.1% (100 decoys) from each state was selected for further analysis. RosettaSurface generation of every $10^4$ decoys produced structurally similar top-scoring (i.e., converged) decoys. We generated $10^5$ decoys to enrich top-scoring decoys for subsequent statistical analysis.

Because the middle segment of statherin is thought to be unstructured, analysis is often carried out with respect to three individual statherin segments (Naganagowda et al., 1998, Raj et al., 1992). Here we adopt a similar convention when applicable: residues 1-15 referred to as N-terminal, 16-29 as middle, and 30-43 as C-terminal.

**Statherin in implicit solvent.** Figure 3.2 details the predicted structure of statherin in solution. Figure 3.2a shows the population of three secondary structure motifs at each residue, averaged over the 100 top-scoring decoys from each of three runs. The predicted structure of the N-terminal segment is mostly helical from residues 4-12, with a “frayed” helical motif from residues 13-15; residues 1 and 2 are mostly unstructured. The C-terminal segment exhibits
moderate helical structure, which initiates at Pro36 and fluctuates between turn- and helix-like hydrogen-bond configurations. The middle segment is more difficult to characterize due to structural dispersion. The Ramachandran plot (figure 3.2b) shows that Rosetta predicts a mostly polyproline II (PPII) and β-turn structure in the middle segment, with moderate right- and left-handed helical structure. The occurrence of torsion angles indicative of canonical secondary structure (figure 3.2b) coupled with a lack of persistent local hydrogen-bonded secondary structure (figure 3.2a) may arise from the abundance of proline residues (seven from residues 20-36). Coordinates for the 10 top-scoring decoys (figure 3.2c) show regular helical structure in the N-terminal segment, a structurally dispersed middle segment, and partially structured C-terminal segment.

Previous experimental measurements on solvated statherin support these predictions. Circular dichroism (CD) spectra of a solvated statherin fragment comprising the N-terminal segment display a significant population of helix (Chen et al., 2008, Raj et al., 1992). NMR experiments on full-length statherin in the helix-stabilizing solvent 2,2,2-trifluoroethanol (TFE) suggest a PPII type secondary structure for statherin’s middle segment (Naganagowda et al., 1998). CD experiments on a statherin fragment comprising the middle segment found a β-turn structure in TFE and a PPII structure in phosphate buffered saline (Raj et al., 1992). CD spectra of a statherin fragment comprising the C-terminal segment show turn-like secondary structure (Raj et al., 1992), while NMR experiments on
full-length statherin in TFE predict a helical structure from residues 36-43 (Naganagowda et al., 1998).

Figure 3.2: Predicted statherin solution-state structure. (a) Distributions of three secondary structure motifs averaged for three independent simulations (error bars represent standard deviation from the mean, for clarity, shown only for helix). (b) Ramachandran plot for 100 top-scoring structures shows regions of populated \((\phi, \psi)\) space at each residue (blue to red = N- to C-terminus). (c) Coordinates for 10 top-scoring solution-state decoys superimposed about the N-Terminal binding domain.
Structure and binding of statherin at the HAp {001} surface. Figure 3.3a shows the change in secondary structure upon adsorption (upper panel), average change in side-chain solvent-accessible surface area (Å²) upon adsorption (middle panel), and the frequency that each residue adsorbed to HAp in the 100 top-scoring decoys (lower panel); all are plotted against residue number.

Figure 3.3: Predicted binding and structural statistics for statherin adsorbed to HAp. (a) Three adsorption phenomena plotted against residue for the 100 top-scoring decoys adsorbed to the {001} surface. The upper panel shows the average change in secondary structure upon adsorption. The middle panel shows the average change in solvent-accessible surface area (Å²) upon adsorption. The lower panel shows the adsorption frequency. Background shading (blue, green, and red) indicates the defined statherin segments (N-terminal, middle, and C-terminal respectively). Statherin amino-acid (one-letter code) sequence plotted along upper x axis. (b) Histogram showing distribution of N-terminal helicity for 1000 top-scoring and (c) 1000 randomly selected decoys.

In the N-Terminal binding domain, the dominant structural difference between states occurs between residues 12 and 14 (upper panel, figure 3.3a).
RosettaSurface predicts that Gly12 is a helical cap in the solution state, a common role of glycine residues at the carboxy terminus of helices (Aurora and Rose, 1998). The frayed helix motif at residues 12-14 is stabilized upon adsorption (Figure 3.4).

![Figure 3.4: Adsorption induced folding event about statherin's Gly-12 as predicted by RosettaSurface. (a) RosettaSurface predicts that the statherin solution-state N-terminal 15-mer is partially stabilized by an electrostatic network (dashed lines) among its ionic residues (shown in stick representation). (b) Superposition of representative solution-(gray) and adsorbed-state (teal) decoys depicting predicted adsorption induced folding event.

In solution (figure 3.4a), we find that electrostatics cause the binding residues to orient in a manner that complements the chemistry and geometry of the HAp surface (lattice matching), with the exception of Arg13. Arg13 is the only ionic residue in the binding domain without a residue of complementary charge.
nearby. Therefore, Arg13 has a higher tendency to position its side chain away from Arg9 and Arg10 and thus away from the binding interface. Upon adsorption (figure 3.4b), however, HAp shields local protein electrostatics and facilities a conformational change extending the helix, placing the positively charged arginine side chain in contact with negatively charged HAp phosphates. NMR measurements support an adsorption induced shift towards helical φ angles at glycine 12 (Shaw et al., 2000).

With the exception of the abovementioned structural difference between states, predicted solution- and adsorbed-state structures are similar. In our previous study (Makrodimitris et al., 2007), we found that a helical fold for the N-terminal segment facilitates a lattice-matching arrangement of binding residues. Here we find that lattice-matching conformers also dominate the solution-state energy minima, and that a combination of van der Waals, electrostatic, and hydrogen-bond interactions cause unfolded states to score poorly relative to folded states at the interface and in solution.

The structural similarity between solution- and adsorbed-states might come from insufficient sampling; however, analysis to this point had been restricted to decoys in the top-scoring 0.1%. We compared the 1000 top-scoring decoys with 1000 randomly selected decoys to see if diverse structures were being sampled at the interface but not being selected for by the RosettaSurface energy function. Figures 3.3b and 3.3c show the distribution of helical content for the N-terminal binding domain, in the context of full-length statherin, for the 1000 top-scoring and 1000 randomly selected adsorbed-state decoys respectively. More than 55%
percent of the top-scoring 1000 (figure 3.3b) decoys have greater than 70% helicity in the adsorbed-state N-terminal binding domain, while only 30% of the randomly chosen decoys meet this criteria (figure 3.3c). Also, 2% of the randomly chosen decoys have a completely unfolded (i.e., 0% helicity) adsorbed-state N-terminal binding domain, whereas completely unfolded decoys are absent in the top 1000. Thus, sampling includes both nonhelical and helical states, but helical states are selected by the RosettaSurface energy function.

Figure 3.3a also shows the predicted binding pattern of HAp-adsorbed statherin at each residue (middle and lower panels). The middle panel in figure 3.3a shows the change in solvent accessibility upon adsorption and the lower panel shows the adsorption frequency of each residue. In general, high binding frequency at a residue should correspond to a decrease in solvent accessibility relative to the solution state; however, the correlation can vary based on changes in protein fold. RosettaSurface predicts that the acidic residues Sep2 and Sep3 and the basic residues Lys6, Arg9, Arg10, and Arg13 adsorb frequently. Strong electrostatic interactions locate these residues at the surface, and in the case of the basic residues there is also a large energetic contribution from hydrogen bonding with HAp surface phosphates. Interestingly, Glu5 adsorbs significantly more often than Glu4. The fact that neighboring glutamate residues bind the surface disproportionately suggests some orientational specificity. This specificity may arise from statherin’s fold coupled with strong binding of its Sep and basic residues. Sep2, Sep3, Lys6, Arg9, Arg10, and Arg13 have been implicated in binding by mutagenesis (Goobes et al., 2007b, Raj et al., 1992),
deletion (Raj et al., 1992), and ssNMR experiments (Raghunathan et al., 2006). An important experiment might investigate the absence/presence of preferential adsorption for one of the neighboring glutamate residues at positions 4 and 5.

The middle and C-terminal segments (figure 3.3a) show minimal change in secondary structure (figure 3.3a, upper panel) and minimal binding (figure 3.3a, middle and lower panels). The combined effect of strong N-terminal binding, a semi-rigid proline-rich middle segment, and the energetic cost of desolvating the HAp surface enables only weak interaction of middle and C-terminal segments. Aside from a lone glutamate residue at position 26, there are no charged side chains downstream of Arg13, and therefore little electrostatic compensation for HAp desolvation. It has been experimentally demonstrated that the middle and C-terminal segments have little or no affinity for HAp (Raj et al., 1992).

**Adsorption of a control peptide.** Although unfolded states were significantly populated at the time of adsorption (figure 3.3c), those states did not significantly populate the top-scoring decoys (figure 3.3b). As a further test of the algorithm’s ability to capture extended states on the surface, we sought a control peptide of known conformation whose ionizable side-chain positioning would require adsorption induced rearrangement to complement HAp surface electrostatics. One candidate is the 17 residue peptide designed by Marqusee et al (Marqusee et al., 1989). This polyalanine peptide was designed to fold into a helix by placing three glutamate-lysine pairs at i and i + 4 positions in its primary sequence; the peptide was determined to be helical using CD spectroscopy. We
simulated the adsorption of this control peptide to the \{001\} surface of HAp using the same protocol.

RosettaSurface predicts a mostly helical fold for the control peptide in solution, staggering its ion pairs about the circumference of its helix. A significant unfolding event is required if the ionizable side chains are to compliment the surface. Figure 3.5a shows that the control peptide undergoes drastic conformational change upon adsorption accompanied by strong binding of the ionic residues.

Figure 3.5: Adsorption induced folding event in a control peptide predicted by RosettaSurface. (a) Three adsorption phenomena plotted against residue number (see Fig. 2a caption for complete description of these three phenomena) show strong HAp binding via the control peptide's ionic residues that is facilitated by a considerable unfolding event in the peptide's central region. (b) Representative solution-state (left) and HAp-absorbed models for control peptide.

Adsorption stabilizes the otherwise frayed termini and the central portion unfolds, placing ionizable residues in contact with the HAp surface. The surface
induced/selected unfolding event is depicted structurally in figure 3.5b. This control shows that the RosettaSurface protocol developed for this study can select for gross structural differences between solution and adsorbed states.

**Comparison with NMR data.** We compared RosettaSurface decoys with all published high-resolution ssNMR distance and angle measurements (acquired from HAp-adsorbed statherin). The histograms in figures 3.6a-c show this comparison separated into angle, intra, and intermolecular distance measurements for statherin adsorbed to the HAp \{001\} surface. RosettaSurface predictions agree well with the 15 published ssNMR measurements, however, the following measurements disagreed: the long-range intramolecular distances Pro23-Pro33 and Pro23-Tyr34, and the intermolecular distance between Phe14 and the nearest HAp phosphorus. These distances vary between top-scoring models in our simulation and no single measurement is predicted.

RosettaSurface predictions agreed with measured distances and angles indicative of helical structure in the N-terminal binding domain such as the \( \phi \) angles of Leu8 and Gly12 and distance between Sep3 & Phe7 and Leu8 & Gly12. The intermolecular distance from the Lys6 side-chain nitrogen to the closest HAp phosphorus is in close agreement with experiment as is the short-range Pro33-Tyr34 distance measurement. Moderate agreement was observed at the \( \phi \) and \( \psi \) angle of Tyr34; with that \( \psi \) angle showing some dispersion. RosettaSurface predictions were close to the measured Tyr34-Tyr38 and Pro33-Tyr38 distance
FIGURE 3.6  Comparison of RosettaSurface predictions and all previously published solid-state NMR measurements. Histograms from unbiased (a, b, and c) and NMR-biased (a', b', and c') simulations for statherin adsorbed to the 001 monoclinic face of HAp. (a and a') Statherin torsion angles, (b and b') statherin intramolecular distances, and (c and c') statherin-HAp intermolecular distances. Red histograms show the distributions for the 100 top-scoring decoys and gray bars show the NMR measurement (width depicts experimental error).
measurements. However, the following measurements disagreed: the long-range intramolecular distances Pro23-Pro33 and Pro23-Tyr34, and the intermolecular distance between Phe14 and the nearest HAp phosphorus. These distances vary between top-scoring models in our simulation and no single measurement is predicted.

To investigate the abovementioned discrepancies we added a harmonic potential to the RosettaSurface energy function to bias the simulation toward satisfying all ssNMR distance and angle measurements. The histograms in figures 3.6a†-c† show the increased accuracy, with respect to ssNMR measurements, that resulted from biasing the simulation. NMR-biased simulation satisfies the long-range distance measurements between Pro23 & Pro33 and Pro23 & Tyr34. To determine why these long-range distance measurements disagreed with unbiased RosettaSurface simulation, we analyzed the residue-specific energies for both biased and unbiased simulations. In general, decoys generated from biased simulations had poor repulsive van der Waals scores and scored worse than decoys obtained from unbiased simulations. Satisfying the long-range distance measurements between residues 23 & 33 and 23 & 34 always produced steric clashes among intervening residues. These steric clashes likely excluded decoys satisfying long-range distance measurements from being enriched in the 100 top-scoring decoys from unbiased simulations. Figure 3.7a shows a representative structure highlighting clashes caused by the long-range distance constraint between Pro23 and Tyr34.
FIGURE 3.7  Clashes created in statherin’s C-terminal segment when NMR constraints are met between (a) Pro33-Tyr38 and (b) Pro23-Tyr34. Residues that contain clashing atoms (red spheres) and isotopically labeled residues are shown in stick representation. Labels indicate distance (Å).

A surprising consistency between NMR biased and unbiased simulation was the predicted ~6.5 Å Pro33-Tyr38 distance. The experimentally measured distance of ~5.5 Å was rarely captured despite the potential function bias to meet that distance constraint. The Pro33-Tyr38 and Tyr34-Tyr38 distance constraints always produced clashes. Clashes were so severe between residues 33 and 38 that even with bias only 31 of the 100 top-scoring decoys satisfied the constraint. However, in 59 of the 100 top-scoring decoys the Pro33-Tyr38 distance is between 6-7 Å; these decoys achieved good atomic packing without creating steric clash. Figure 3.7b shows some clashes created when satisfying the Pro33-Tyr38 distance measurement.
Inclusion of ssNMR bias led to a significant adsorption-induced folding event in the C-terminal segment (data not shown). As much as 55% of the population shifted from turn to helix between residues 31 and 38 during biased predictions; no such transition results from unbiased predictions (upper panel, figure 3.3a). Based on the Pro23-Pro33, Pro23-Tyr34, Pro33 & Tyr38, and Tyr34-Tyr38 ssNMR measured distances, Goobes et al. (Goobes et al., 2006a) concluded that a significant C-terminal folding event accompanied HAp adsorption.

In the absence of a biasing potential, Phe14 tends to adsorb infrequently relative to the rest of the binding domain. It is difficult to determine whether this departure from ssNMR experiment arises from insufficient sampling and/or a deficiency in the energy function. To see if we could eliminate the energy function as a possibility we ran a second biased simulation, where the protein-surface intermolecular constraints were only enforced to form the initial adsorbed-state complex. Subsequent moves did not bias the relative rigid-body positions of the protein and surface. Therefore, top-scoring decoys will likely display enriched binding of Phe14 if the energy function favors the sampled adsorbed-state conformations. We found that even though the initial adsorbed-state complex satisfied the HAp-Phe14 distance constraint, subsequent moves often removed Phe14 from the surface. Therefore, even when conformations with Phe14 adsorbed are sampled, the energy function does not favor the adsorption of Phe14. Although this suggests the energy function is deficient, a second possibility is that Phe14 adsorbs but the correct statherin conformations were not found despite extensive sampling of adsorbed-state conformations.
Another possibility is that the Phe14 ssNMR measured distance is not relevant to the predictions made here. Phe14 ssNMR distance measurements were acquired in the context of a statherin molecule truncated at residue 15, leaving a negatively charged carboxyl group adjacent to Phe14 that was not present in our models.

There is a moderate difference in binding with respect to the three surfaces, most notably between the \{001\} and \{100\} surfaces (lower panels, figures 3.2a and 3.9b respectively). Residues Sep2, Sep3, Lys6, Arg9, Arg10, and Arg13 bind with approximately equal frequency, between 48% and 69%, at the \{001\} surface. Conversely, at the \{100\} surface, Arg13 only adsorbed 38% of the time (58% at the \{001\} surface), and Sep2 & Sep3 each adsorbed 93% of the time (~68% at the \{001\} surface). Furthermore, Asp1 bound approximately as or more frequently than Arg9, Arg10, or Arg13 at both the \{010\} and \{100\} surfaces (lower panel, figures 3.9a and 3.9b respectively). And in the case of the \{100\} surface, Glu4 binds more frequently than Arg10 or Arg13.

**Statherin adsorption at the \{010\} and \{100\} faces of HAp.** It is not known to which face(s) of HAp statherin binds. To test whether RosettaSurface could distinguish surfaces relevant to statherin recognition, we applied our protocol to dock statherin to two additional HAp surfaces, \{010\} and \{100\}. Atomically flat, mixed-charge terminations were used at each surface. Simulation methodology was identical at all three surfaces.
Surprisingly, similar statherin conformers adsorbed to three HAp faces (\{001\} face in figure 2a & \{010\} and \{100\} faces in figures 3.9a & 3.9b respectively). In all three runs the adsorbed state is mostly helical and the N-terminal acidic and basic residues bind the charged HAp surface.

The most significant difference in statherin’s conformation at the three surfaces is the more pronounced folding event about Gly12 at the \{001\} (figure 2a) surface compared to the \{010\} and \{100\} surfaces (figures 3.9a and 3.9b respectively). One reason folding may differ slightly at each surface is the geometry of the binding moieties replicated across the three surfaces. In figure 3.8a, the white parallelogram shows the interstice of the phosphate-oxygen triad (IPOT) motif, the motif hypothesized to be responsible for binding statherin’s basic residues at the HAp \{001\} face in our previous, fixed backbone docking study (Makrodimitris et al., 2007b). The IPOT motif is a periodic replication of open phosphate clusters, which expose hydrogen-bond-accepting oxygen atoms and afford favorable van der Waals and electrostatic interactions. Similar motifs are shown for the \{010\} and \{100\} surfaces in figures 3.8b and 3.8c respectively. Amongst these motifs, the IPOT on the \{001\} surface has unique dimensions (a parallelogram with 9.42 Å sides and 16.2 Å diagonal) when compared to the motifs on the \{010\} and \{100\} surfaces (both are rectangles, 9.42 by 6.88 Å and an 11.64 Å diagonal). The motif dimension at the \{001\} surface is more easily complemented by a statherin binding domain in the extended helix fold (~16Å in length). This correlation may account for the elevated binding of Arg13 at the
Figure 3.8: Predicted binding motifs on three HAp surfaces. (a) The IPOT motif (38) of the \{001\} monoclinic face of HAp and similar motifs on the (b) \{010\} and (c) \{100\} monoclinic faces of HAp.

FIGURE 3.9  Three adsorption phenomena plotted against residue number for the 100 top-scoring decoys at the \{010\} (a) and \{100\} (b) surfaces. See figure 2a caption for a complete description of these three phenomena.
{001} HAp face, whereas there is elevated binding of Asp1 and/or Glu4 at the {010} and {100} faces.

3.4 Discussion

**Structural differences in the solution- and adsorbed states.** Protein adsorption has historically been associated with unfolding and/or structural rearrangement (Haynes and Norde, 1994). This notion may have arisen, to some extent, from the fact that protein-surface interactions are often studied using stock-room proteins such as lysozyme, albumin, and fibrinogen that have not evolved to bind specifically to the surfaces studied in those experiments (Ball and Jones, 1995, Buijs et al., 2000, Kim et al., 2002, Steadman et al., 1992, Xia et al., 2002). While gross unfolding events may be expected to accompany nonspecific adsorption owing to an increase in entropy, stable, well-defined adsorbed-state folds have been experimentally inferred in several evolved systems including antifreeze proteins binding ice (Sicheri and Yang, 1995, Leinala et al., 2002, Ko et al., 2003) and HAp and calcite biomineralization systems (Reyes-Grajeda et al., 2004, Dowd et al., 2003, Frazao et al., 2005, Hoang et al., 2003a). The findings in this study suggest that the binding domain of statherin has evolved such that the solution and adsorbed states are similar, and hence little conformational rearrangement upon adsorption is necessary. This finding is supported by CD experiments that suggest the N-terminal segment of statherin to be stable and partially helical in solution (Chen et al., 2008, Raj et al., 1992) and ssNMR measurements that predict that the adsorbed state is also helical (Long et al., 2001, Shaw et al., 2000).
The only significant structural difference between statherin’s predicted solution and HAp-bound states is the extension of the helical conformation in the binding domain of the HAp-bound states. This prediction agrees with ssNMR measurements (Shaw et al., 2000) and addresses an apparent conflict between macro and microscopic studies on statherin’s solution and adsorbed states. Naganagowda et al. (Naganagowda et al., 1998) observed a single sharp $^{31}$P NMR signal at Sep2 and Sep3, and Shaw et al. (Shaw et al., 2000) measured φ angles at Leu8 and Gly12, $73^\circ$ and $80^\circ$ respectively, in the solution state. Conversely, two independent CD spectra (Chen et al., 2008, Raj et al., 1992) recorded on statherin’s N-terminal binding domain indicate that helical conformers are populated in solution. Because ssNMR measurements predict a helical structure for statherin’s N-terminal binding domain in the adsorbed state (Long et al., 2001, Shaw et al., 2000), disagreements with respect to the solution state have led some to report large-scale HAp induced folding while others suggest little or no difference between the states. Based on our findings, we suggest an intermediate model where statherin solution states are partially prestructured for adsorption. We find that Asp1 and Sep2 are not part of any regular secondary structure and that a folding event occurs near Gly12. These models agree with CD data and suggest the macroscopic structure of the N-terminal binding domain is partially helical in solution.

**Structural differences at the three HAp surfaces.** The mean score at the {001}, {010}, and {100} faces is $-77 \pm 8$, $-77 \pm 7$, and $-77 \pm 8$ REU (see Materials and Methods in the Supplementary Material) respectively. These small differences in
score suggest promiscuous binding at three HAp surfaces. Similarly, predicted structures at the three HAp faces agree approximately equally well with ssNMR experimental measurements (data not shown for HAp {010} and {100} surfaces). The chemistry and geometry at the three HAp surfaces is similar, and small changes in statherin’s backbone and side-chain torsion angles can complement the three HAp surfaces in similar ways with small energetic barriers. But, because the simulation of all physiological conditions during biomineralization is not feasible (protein concentration, protein-protein interactions, salt concentration, simultaneous crystallization of all relevant faces, etc.,) the relevant free energy of adsorption cannot be calculated. It is therefore difficult to assess the effects such small structural perturbations might impart on specificity.

The vast knowledge obtained from studying crystallized protein-protein complexes from the Protein Data Bank (Berman et al., 2002) makes it tempting to assume proteins generally interact in a highly specific manner. But there may be considerable differences in the way proteins behave at phase boundaries compared to globular protein-protein association. For instance, biomineral inhibition could be accomplished by adsorbing protein to all or several faces of a growing biomineral, the relative rates depending on the intrinsic energy of each exposed face. Also, the presence of an adsorbate protein may be sufficient to inhibit biomineralization, regardless of whether a single structure dominates the free energy minimum. With an increased number of ssNMR measurements it may be possible to constrain biased simulations to determine the relevant
binding face(s). For instance, the arrangement of ionic side chains in constrained structures may electrostatically repel the relative calcium and phosphate geometries at some HAp surfaces, potentially identifying such faces as incompatible with statherin binding.

**Ability to predict adsorbed-state structures de novo.** Results from unbiased simulations agree remarkably well with experiment considering the conformational space associated with folding a protein at an interface. Also, adsorption of a control peptide resulted in an anticipated unfolding event between solution and adsorbed states. The largest departure from experiment was the absence of predicted structures in agreement with ssNMR measured long-range Pro23-Pro33 and Pro23-Tyr34 distances and the intermolecular HAp-Phe14 distance.

Proline rich segments, such as the statherin segment comprising residues 23-34, often give rise to structural dispersion due to torsional constraints imposed by proline’s imide bond and an absence of backbone hydrogen-bond donors (Williamson, 1994). Given that the magnitude of the long-range distance measurements span the HAp unit cell (long-range distance measurements are 8.0-11.5 Å; HAp unit cell $a = 9.4$ Å), one possible explanation for this discrepancy is that the ssNMR experiments detected protein-protein intermolecular interactions from adjacently adsorbed statherin molecules. Also, the isotopic labeling scheme at Pro23 included fluorine (Goobes et al., 2006a), and fluorine labeled proline residues can change and/or stabilize protein folds via electronegative inductive effects (Renner et al., 2001). Rosetta samples proline
cis- and trans-isomers and alternate ring puckers; therefore, the absence of these proline conformers during decoy generation was not a cause of steric clash.

Another possible explanation for discrepancies between ssNMR measurements and our predictions is the RosettaSurface energy function and representation of the system. RosettaSurface accounts for solvent implicitly rather than explicitly representing individual water molecules and salt ions, and water molecules and salt ions can influence protein adsorption (Castells et al., 2002). Furthermore, the EEF-1 implicit solvent model (Lazaridis and Karplus, 1999) employed by RosettaSurface was not originally developed to capture solvent effects at interfaces. Finally, the charge density at the statherin-HAp interface is significant, and it has not been demonstrated that the distant-dependent-dielectric method applied here can correctly account for electrostatic effects at such an interface.

Unbiased RosettaSurface simulation did not predict the C-terminal folding event that biased simulation predicted. Structures predicted from biased simulation place Pro36 in the middle of an α-helix. Proline residues are rare in the middle of α-helices and always produce a kink of ~20-30° or greater (Barlow and Thornton, 1988). Unbiased simulation predicts that the C-terminal α-helix initiates at Pro36 (figure 1a), in agreement with solution-state NMR measurements for statherin in TFE (Naganagowda et al., 1998).

It is difficult to infer detailed protein structure from minimal experimental constraints. RosettaSurface-predicted structures present a valid concern for
interpreting ssNMR long-range distance measurements as intramolecular, suggest alternative interpretations consistent with the data, and propose experiments that could determine the correct interpretation(s). Similarly, RosettaSurface predicted Phe14 adsorption only after ssNMR bias was added to the simulation, highlighting the benefit experimental measurements can impart in computational structure prediction. The synergy between ssNMR and RosettaSurface may be particularly useful since adsorbed-state structures cannot be determined via alternative methods.

**Prospects for a combined solid-state NMR RosettaSurface method.** The concluding remarks of Goobes *et al.* (Goobes *et al.*, 2007a) in their 2007 review on ssNMR spectroscopy and protein-surface interactions highlight the importance of the methods developed for this study: “Current protein structure prediction programs cannot predict the folded state of a protein in the presence of another macromolecule or surface starting from an unfolded or random coil conformation. The ultimate goal of computational technique development would be to assert the experimental observation of a transition from an unfolded state to an active state upon exposure of the potential energy of the surface.” For the work presented in this study, we developed a program capable of accomplishing such a goal.

Future RosettaSurface NMR collaborations on new systems may take the following form: educate initial placement of isotopically labeled amino-acid pairs by analysis of unbiased RosettaSurface output, considering areas of high and low confidence. The initial NMR measurements would in turn confirm or refute high-
confidence predictions and inform low-confidence predictions for subsequent rounds of biased simulations. Given the results of this work, it may be possible to converge on plausible structures with only a few iterations of a combined method. We wish not to understate the difficulty of such a collaboration, but to emphasize the unique capability it possess to address the difficult and important problem of macromolecular structure prediction at interfaces.
Structure determination method for biomineral-associated protein using combined solid-state NMR and computational structure prediction

4.1 Introduction

The structure of biomineral-associated proteins cannot be determined by X-ray crystallography or solution NMR; as a result, high-resolution aspects of protein biomineralization are not well understood. While many experimental methods exist for studying proteins adsorbed to solid surfaces, most can only resolve macroscopic features. Solid-state NMR (ssNMR) is uniquely suited for determining the distance between pairs of isotopically labeled atoms at the protein-surface interface (Goobes et al., 2007a). High-resolution ssNMR data exists for at least two protein-biomineral complexes. In solution-state NMR, 10-15 measurements are typically acquired at each residue during protein structure determination; this is not tractable by ssNMR methods.

Recently, combined solution NMR-computational structure prediction methods have drastically reduced the amount of NMR data necessary to determine high-resolution protein structures in solution (Cavalli et al., 2007, Shen et al., 2008, Gong et al., 2007). In particular, the Rosetta structure prediction method was combined with chemical shift (Shen et al., 2008), nuclear overhauser effect (Bowers et al., 2000), or residual dipolar coupling (Rohl and Baker, 2002)
NMR data. In all three cases, sets of protein structures predicted by Rosetta were at or near atomic-level accuracy. The use of high-resolution experimental data biases sampling to relevant conformation space and helps account for inaccuracies in computational energy functions. An approach combining high-resolution experiment and structure prediction has the potential to elucidate biomineral-associated protein structure to unprecedented resolution.

In chapter 3, I developed Rosetta to predict the fold and orientation of a protein on a biomineral surface (RosettaSurface). In addition, chapter 3 included structure prediction biased by ssNMR measurements. In this chapter I present an improved version of ssNMR-biased RosettaSurface structure prediction (RosettaSurface.NMR). RosettaSurface.NMR folds a protein on a biomineral surface using ssNMR data to bias structure prediction and has improved sampling of protein-surface intermolecular constraints.

I demonstrate the improved algorithm using iterative ssNMR-RosettaSurface.NMR. I use the iterative method to determine the structure of human salivary statherin adsorbed to hydroxyapatite (see 2.1 for background on statherin and hydroxyapatite). I perform two iterations (rounds) of biased predictions. In round 1, I use previously published ssNMR measurements to predict the structure of statherin bound to a single hydroxyapatite (HAp) crystal surface. Structures from round 1 predictions guide the placement of isotopic labels for six new ssNMR measurements. These new ssNMR measurements bias a second round of RosettaSurface.NMR predictions for statherin adsorbed to five HAp crystal surfaces. Results from round 2 biased structure prediction suggest
preferential adsorption at a one of the HAp crystal surfaces. Finally, I perform more extensive sampling at that crystal surface to produce the final structures.

4.2 Materials and methods

All experiments, including sample preparation and ssNMR, were performed as previously published (Gibson et al., 2005, Goobes et al., 2006a, Long et al., 2001). All ssNMR measurements were acquired in Gary Drobny’s lab at the University of Washington; I took no part in those experiments.

The algorithm developed here is based on RosettaSurface (section 3.2). Briefly, each execution of the RosettaSurface algorithm folds a protein from a fully extended conformation in solution using a united-atom model. Then, a high-resolution (all-atom) representation of the peptide is minimized in solution and adsorbed to a biomineral surface in a random orientation. The protein’s fold and orientation are optimized on the surface resulting in one candidate adsorbed-state structure.

Development of RosettaSurface.NMR included two modifications: First, the full-atom energy ($E$) used for decoy discrimination is a linear combination of attractive and repulsive Lennard-Jones interactions ($E_{\text{att}}$ and $E_{\text{rep}}$), solvation ($E_{\text{sol}}$), hydrogen bonding ($E_{\text{hb}}$), electrostatics ($E_{\text{coul}}$), and a term to enforce ssNMR constraints ($E_{\text{constraint}}$):

$$E = E_{\text{att}} + E_{\text{rep}} + E_{\text{sol}} + E_{\text{coul}} + E_{\text{hbond}} + E_{\text{constraint}}$$
where \( E_{\text{constraint}} = k \varepsilon^2 \). The biasing term \( \varepsilon \) is equal to \( \sqrt{\sum_{i=1}^{n}(\Delta x_i)^2} \), a harmonic potential, where \( \Delta x_i \) is the difference between the \( i^{th} \) experimentally determined angle or distance constraint and the corresponding predicted angle or distance, and \( n \) is the number of biasing constraints used during structure prediction. During biased structure prediction the spring constant, \( k \), can adjust the strength of the bias toward experimental measurements; the units of the weight are Rosetta Energy Units (REU)/Å\(^2\) (or \(^2\)).

A second modification to RosettaSurface is that immediately following the formation of the adsorbed-state complex (see Figure 3.1), the orientation is rapidly optimized in an attempt to satisfy all experimentally determined protein-surface intermolecular contacts. The rapid optimization consists of 100 attempted random moves of a Gaussian distributed translation of mean 0.1 Å in each Cartesian direction and a Gaussian distributed random rotation of mean \( \sim 17^\circ \) around each Cartesian axis. Each move that reduces the magnitude of \( \sum_{i=1}^{n}(\Delta x_i) \) (for \( n \) measured protein-surface intermolecular constraints) is accepted, otherwise the complex is returned to its previous state (i.e., zero temperature Monte Carlo search with only \( E_{\text{constraint}} \) considered). During this intermolecular constraint optimization, the surface is represented as a plane to avoid calculating every interatomic distance. After the 100 attempts, the protein-surface system proceeds to adsorbed-state refinement and surface atoms are returned to their explicit representations.
For all predictions, $10^5$ candidate HAp-adsorbed statherin structures were generated using RosettaSurface.NMR. Each candidate structure is assessed by the error, $\varepsilon = \sum_{i=1}^{n}(\Delta x_i)$ (for all $n$ ssNMR measurements). Structures with the least error are used for further analysis and model representation.

The 43-residue molecular structure of human salivary statherin and all HAp crystal surface coordinates were generated as described in section 3.2. In total, five HAp surfaces were used in this study (see Figure 4.1): {001}, {010}, two differentially terminated {100} faces ({100}-T1 and {100}-T2), and {101}. The {100}-T2 surface terminates with a higher density of calcium ions than the {100}-T1 surface. All other HAp surfaces were cut to expose approximately neutral (mixed charged) surfaces.

**Figure 4.1: Schematic of hexagonal HAp.** The five crystal surfaces (two differentially terminated surfaces at the {100} crystal plane) used for random-weight biased RosettaSurface.NMR structure prediction. Note: the indices refer to a hexagonal set of basis vectors and the third vector index is omitted.
These faces were chosen because they are expressed in stable HAp crystals and have a range of surface geometries (Astala and Stott, 2008).

### 4.3 Results

I performed two rounds of ssNMR-biased structure prediction, beginning first with published ssNMR measurements (Table 1).

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Label</th>
<th>Distance (Å) or Angle (°)</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>pS_3 φ</td>
<td>C_2-C_3</td>
<td>-60±10°</td>
<td>(Long et al., 2001)</td>
</tr>
<tr>
<td>pS_3-F_7</td>
<td>C_1-N_1</td>
<td>4.3±0.2 Å</td>
<td>(Long et al., 2001)</td>
</tr>
<tr>
<td>K_7-HAp</td>
<td>N_1-P_9</td>
<td>4.0±0.5 Å</td>
<td>(Gibson et al., 2005)</td>
</tr>
<tr>
<td>F_14-HAp</td>
<td>C_6-P_11</td>
<td>&gt; 6 Å</td>
<td>(Gibson et al., 2006)</td>
</tr>
<tr>
<td>L_8 φ</td>
<td>C_7-C_8</td>
<td>60±9°</td>
<td>(Long et al., 2001)</td>
</tr>
<tr>
<td>L_8-G_12</td>
<td>C_11-N_12</td>
<td>4.8±0.4 Å</td>
<td>(Long et al., 2001)</td>
</tr>
<tr>
<td>G_13 φ</td>
<td>C_11-C_12</td>
<td>73±3.6°</td>
<td>(Long et al., 2001)</td>
</tr>
<tr>
<td>F_14-HAp</td>
<td>C_6-P_11</td>
<td>6±1.0 Å</td>
<td>(Gibson et al., 2006)</td>
</tr>
<tr>
<td>P_23-P_33</td>
<td>C_1-C_2</td>
<td>10.5±1.0 Å or 8.8±0.8 Å</td>
<td>(Goobes et al., 2006a)</td>
</tr>
<tr>
<td>P_23-Y_34</td>
<td>C_1-C_2</td>
<td>10.5±1.0 Å or 8.8±0.8 Å</td>
<td>(Goobes et al., 2006a)</td>
</tr>
<tr>
<td>P_33-Y_34</td>
<td>C_11-C_12</td>
<td>3.12±0.13 Å</td>
<td>(Goobes et al., 2006a)</td>
</tr>
<tr>
<td>P_33-Y_38</td>
<td>C_11-N_12</td>
<td>5.3±0.5 Å</td>
<td>(Goobes et al., 2006a)</td>
</tr>
<tr>
<td>Y_34 φ</td>
<td>C_15-C_24</td>
<td>75±15°</td>
<td>(Goobes et al., 2006a)</td>
</tr>
<tr>
<td>Y_34 ψ</td>
<td>C_15-C_24</td>
<td>40±10°</td>
<td>(Goobes et al., 2006a)</td>
</tr>
<tr>
<td>Y_38-Y_48</td>
<td>C_15-N_12</td>
<td>4.0±0.5 Å</td>
<td>(Goobes et al., 2006a)</td>
</tr>
</tbody>
</table>

**Table 4.1:** Solid-state NMR measurements used in first round of biased structure prediction. The residue for which the Measurement was acquired, the placement of the isotopic Label, and the determined Distance (Å) or Angle (°). Angle measurements were determined using $^{13}C'$ of that residue and $^{13}C'$ of the preceding residue. The Phe14 measurement was removed from the previously published data set owing to uncertainties in the experimental construct (personal communication with Gary Drobny).
While the biologically relevant HAp face(s) for statherin binding is not definitely known, the \{001\} face is one potential binding surface (see section 2.2). To simplify calculations in the first round, I predicted statherin’s structure only at the HAp \{001\} surface.

**Round 1:** Figure 4.2 shows protein intramolecular (figure 4.2a) and protein-surface intermolecular contacts (figure 4.2b) for the 100 structures with the least error (smallest $\varepsilon$) from round 1 predictions (see Materials and methods). Predicted contacts are useful for suggesting measurements for subsequent rounds of ssNMR and for indicating the predicted structure of the ensemble. During round 1 predictions the weight, $k$, (see Materials and methods) was set to 10 REU/Å² ($^{02}$).

In Figure 4.2a the dense regions of $i$ to $i + 4$ contacts in the N-terminal domain suggest helical secondary structure from residues 3 to 14, with some helical fraying from residues 11 to 14. The $i$ to $i + 4$ contacts between residues 31 and 39 also suggest helical structure. There are few contacts made between residues that are greater than five residues apart in primary sequence, suggesting little tertiary structure. However, there are a significant number of contacts made between residues 25 and 32, and some residues within that segment. These contacts are expected from the ssNMR measurements used to bias structure prediction.

In figure 4.2b, high-frequency residue-surface contacts exist only in statherin’s N-terminal domain, indicating that HAp binds exclusively to that
domain. Asp1, Sep3 and to a lesser extent Sep2 are all located at the interface (Sep is an abbreviation for phosphorylated serine). Isotopic labeling of phosphoserine residues is avoided because it is difficult to distinguish between ssNMR signal from HAp surface phosphates and Sep side-chain phosphates.

Figure 4.2: Ensemble structure of statherin adsorbed to HAp {001} surface determined from the first round of biased structure prediction. (A) Pair-wise statherin intramolecular residue-residue contacts and (B) pair-wise residue-surface distances for the 100 structures with smallest $\varepsilon$ (see Materials and methods). An intramolecular residue-residue contact is declared if two residues have an inter-residue atomic pair within 4 Å. A residue-surface distance reflects the closest atomic contact for that residue and the closest surface atom. Note for comparing contacts with ssNMR measurements: The intra or intermolecular atomic pairs that constitute a contact are not necessarily the same atoms that were isotopically labeled for ssNMR measurements.
Glu5 is predicted to adsorb frequently whereas Glu4 is predicted to point away from the surface. The Glu5 side-chain carboxyl oxygens are predicted ~2.5 Å from a HAp calcium atom; Glu4 is predicted to be ~8 Å from the HAp surface. Arg9, Arg10, and to a lesser extent Arg13 are also predicted to bind HAp. The C-terminal domain does not bind, even though there is a negatively charged glutamic acid residue in that region (Glu26). This may be surprising considering the high affinity of acidic amino acids for HAp (Addadi and Weiner, 1985, Fu et al., 2005, Gotliv et al., 2005, Tsukamoto et al., 2004).

Round 2: Considering the difficulty in acquiring ssNMR distance measurements at the protein-biomineral interface, we restrict round 2 to six new ssNMR measurements. The six measurements should confirm or refute predictions from round 1 biased simulation. The measurements include intermolecular distances from HAp phosphate phosphorous atoms to residues Glu4, Glu5, Arg9, Arg10, Glu26, and Pro28. The orientation of these residues at the interface should be revealing about the bound orientation of the statherin-HAp complex, and will address predictions including the disproportionate binding of neighboring acidic residues, binding if basic residues, and lack of binding in the C-terminal domain.

Table 4.2 shows the new set of distances measured by ssNMR. Remarkably, all measurements are in agreement with round 1 predictions. This agreement suggests convergence between ssNMR and RosettaSurface.NMR, at least in the ssNMR measured regions.
**Table 4.2: Solid-state NMR measurements used in second round of biased structure prediction.** The residue for which the Measurement was acquired, the placement of the isotopic Label, and the determined Distance (Å).

Next I performed round 2 predictions using all ssNMR measurements from Tables 4.1 and 4.2. I performed round 2 of ssNMR-biased structure prediction at five HAp crystal faces: \{001\}, \{010\}, \{100\}-T1, \{100\}-T2, and \{101\} (see Materials and methods). At each surface, the predictions were divided into 11 runs of \~9,100 candidate structures. For each of the 11 runs at each surface, the weight, \(k\), was set to a number between 0 and 1 evenly divisible by 0.1 (i.e., 0, 0.1, 0.2 .... 1). I employed this approach to see if predictions preferentially agreed with ssNMR experiment at a particular statherin-HAp interface. This result would manifest as a prediction at one or more HAp surfaces exhibiting smaller error (\(\varepsilon\)) versus weight relative to predictions at the other HAp surfaces. This approach has the potential to probe surface specificity at protein-surface interfaces.

**Table 4.2**

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Label</th>
<th>ssNMR Distance (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E(_4)-HAp</td>
<td>13C(^{\delta-31}P)</td>
<td>&gt; 6</td>
</tr>
<tr>
<td>E(_5)-HAp</td>
<td>13C(^{\delta-31}P)</td>
<td>4.2</td>
</tr>
<tr>
<td>R(_9)-HAp</td>
<td>13C(^{\kappa-31}P)</td>
<td>4.5</td>
</tr>
<tr>
<td>R(_{10})-HAp</td>
<td>13C(^{\kappa-31}P)</td>
<td>4.4</td>
</tr>
<tr>
<td>E(_{26})-HAp</td>
<td>13C(^{\delta-31}P)</td>
<td>&gt; 6</td>
</tr>
<tr>
<td>P(_{28})-HAp</td>
<td>13C(^{\prime-31}P)</td>
<td>&gt; 6</td>
</tr>
</tbody>
</table>

Figure 4.3 shows the error (\(\varepsilon\)) versus weight for structures resulting from round 2 ssNMR-biased predictions. As expected, \(\varepsilon\) decreases with increasing
weight at all five HAp surfaces. At all spring constants, $\varepsilon$ is smallest at the {001} surface. That is, it is easiest to for RosettaSurface to create structures matching the experimental constraints when statherin binds the {001} face, suggesting preferential adsorption at this surface. The geometry of phosphate clusters on the {001} surface is unique. The error for the {010} and {100}-T1 faces track very closely, perhaps because the geometry of phosphate clusters on the {010} and {100}-T1 faces is similar (see Figure 3.8). Convergence between prediction and experiment was slowest at the {101} or {100}-T2 surface; the {101} surface has a unique geometry and the {100}-T2 surface has a high calcium concentration. It is difficult to assess the significance in the differences between each face in figure 4.3.

![Figure 4.3: Error $\varepsilon$ versus weight at five different HAp surfaces.](image-url)
However, RosettaSurface.NMR predictions do preferentially agree with ssNMR data at the \{001\} surface; therefore, I restrict the rest of my investigation to the predictions at the \{001\} surface.

For the last phase of biased RosettaSurface.NMR structure determination, I generated $10^5$ structures of HAp \{001\}-bound statherin using all ssNMR measurements (tables 4.1 and 4.2). I used a weight ($k$) of 10 REU/Å$^2$ (or \(\text{o}^2\)) to encourage the generation of predicted structures that satisfy all ssNMR constraints. Figure 4.4 shows ensemble structural statistics for the 100 structures with smallest $\epsilon$ from this final phase of structure determination. In figure 4.4a, the residue-surface contact map shows fewer, more populated bins compared with predictions from round 1 (figure 4.2b). This increased resolution shows the benefit incurred from biasing RosettaSurface.NMR predictions with the additional ssNMR measurements in table 4.2. In particular, Glu5, Lys6, Arg9, Arg10, and Arg13 are located at the HAp surface, and Glu4, Phe7, Leu8, Ile11, and Gly12 have defined positions \(\sim\)8 Å from the surface. Figure 4.4b shows a stable helix from residue 4-11 and less populated helix from residue 34-39. The ensemble shows turn and extended structure for statherin’s middle segment (see chapter 3.2 for a definition of these basic secondary structural elements).
4.4: Ensemble structure of statherin adsorbed to HAp {001} surface determined from the final phase of biased structure determination. (A) Pair-wise residue-surface distances and (B) distributions of three basic secondary structure motifs for the 100 lowest error predicted structures.

Figure 4.5 shows a representative structural model from the 100 structures with smallest $\varepsilon$ from the final phase of round 2 structure determination. This structure was chosen because it represents the dominant ensemble structure (figure 4.4) and closely matches ssNMR measurements (tables 4.1 and 4.2). Figure 4.5a shows a global view of the statherin-HAp complex. For that model, figures 4.5b-d show predicted distances and angles for which biasing was applied. These figures are
Figure 4.5: The molecular structure of statherin adsorbed to HAp {001} surface. (A) Representative structure from the final phase of round 2 biased predictions. Opacity represents statherin’s molecular shape, cartoons represent secondary structure, and sticks are shown for amino acids that are known from experiment to interact with HAp. Predicted distance and angle measurements at constrained atoms for the N- (B and C) and C-terminal domains (D). For comparison with ssNMR measurements see tables 1 and 2. In panel (D), three measurements are color coded for clarity.
useful for directly showing the extent of biasing at specific regions of the complex and the structure that was predicted in those regions. Table 4.3 shows all experimental measurements from rounds 1 and 2 and the corresponding predictions from the representative model (figure 4.5).

Table 4.3: Solid-state NMR measurements used during round 1 and 2 and the corresponding predictions from the representative model (figure 4.5). The residue for which the Measurement was acquired, the placement of the isotopic Label, and the determined Distance (Å) or Angle (°) and corresponding prediction from the Representative Model. Angle measurements were determined using $^{13}\text{C}'$ of that residue and $^{13}\text{C}'$ of the preceding residue.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Label</th>
<th>Distance (Å) or Angle (°)</th>
<th>Rep. Model</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>pS3 $\phi$</td>
<td>$^{13}\text{C}'$-$^{13}\text{C}'$</td>
<td>$-60\pm10^0$</td>
<td>$-60.1^0$</td>
<td>(Long et al., 2001)</td>
</tr>
<tr>
<td>pS3,F3</td>
<td>$^{13}\text{C}'$-$^{15}\text{N}$</td>
<td>$4.3\pm0.2$ Å</td>
<td>$4.5$ Å</td>
<td>(Long et al., 2001)</td>
</tr>
<tr>
<td>E3-HAp</td>
<td>$^{13}\text{C}'$-$^{13}\text{P}$</td>
<td>&gt; 6 Å</td>
<td>$13.1$ Å</td>
<td>New</td>
</tr>
<tr>
<td>E3-HAp</td>
<td>$^{13}\text{C}'$-$^{13}\text{P}$</td>
<td>4.2 Å</td>
<td>4.8 Å</td>
<td>New</td>
</tr>
<tr>
<td>K3-HAp</td>
<td>$^{13}\text{N}$-$^{13}\text{P}$</td>
<td>$4.0\pm0.5$ Å</td>
<td>$3.6$ Å</td>
<td>(Gibson et al., 2005)</td>
</tr>
<tr>
<td>F3-HAp</td>
<td>$^{13}\text{C}'$-$^{13}\text{P}$</td>
<td>&gt; 6 Å</td>
<td>$10.9$ Å</td>
<td>(Gibson et al., 2006)</td>
</tr>
<tr>
<td>I4 $\phi$</td>
<td>$^{13}\text{C}'$-$^{13}\text{C}'$</td>
<td>$60\pm9^0$</td>
<td>$60.2^0$</td>
<td>(Long et al., 2001)</td>
</tr>
<tr>
<td>I4-G12</td>
<td>$^{13}\text{C}'$-$^{15}\text{N}$</td>
<td>$4.8\pm0.4$ Å</td>
<td>$3.89$ Å</td>
<td>(Long et al., 2001)</td>
</tr>
<tr>
<td>R3-HAp</td>
<td>$^{13}\text{C}'$-$^{13}\text{P}$</td>
<td>4.5 Å</td>
<td>4.4 Å</td>
<td>New</td>
</tr>
<tr>
<td>R3-HAp</td>
<td>$^{13}\text{C}'$-$^{13}\text{P}$</td>
<td>4.4 Å</td>
<td>4.4 Å</td>
<td>New</td>
</tr>
<tr>
<td>G12 $\phi$</td>
<td>$^{13}\text{C}'$-$^{12}\text{C}'$</td>
<td>$73\pm3.6^0$</td>
<td>$73.2^0$</td>
<td>(Long et al., 2001)</td>
</tr>
<tr>
<td>F34-HAp</td>
<td>$^{13}\text{C}'$-$^{13}\text{P}$</td>
<td>$6\pm1.0$ Å</td>
<td>NA</td>
<td>(Gibson et al., 2006)</td>
</tr>
<tr>
<td>P34-P33</td>
<td>$^{13}\text{C}'$-$^{13}\text{C}'$</td>
<td>$10.5\pm1.0$ Å or $9.6\pm0.9$ Å</td>
<td>$10.9$ Å</td>
<td>(Goobes et al., 2006a)</td>
</tr>
<tr>
<td>P34-P33</td>
<td>$^{13}\text{C}'$-$^{13}\text{C}'$</td>
<td>$3.12\pm0.13$ Å</td>
<td>$3.1$ Å</td>
<td>(Goobes et al., 2006a)</td>
</tr>
<tr>
<td>P34-P33</td>
<td>$^{13}\text{C}'$-$^{15}\text{N}$</td>
<td>$5.3\pm0.5$ Å</td>
<td>$6.6$ Å</td>
<td>(Goobes et al., 2006a)</td>
</tr>
<tr>
<td>Y34 $\phi$</td>
<td>$^{13}\text{C}'$-$^{13}\text{C}'$</td>
<td>$75\pm15$°</td>
<td>$74.8$°</td>
<td>(Goobes et al., 2006a)</td>
</tr>
<tr>
<td>Y34 $\psi$</td>
<td>$^{13}\text{C}'$-$^{13}\text{C}'$</td>
<td>$40\pm10$°</td>
<td>$40.3$°</td>
<td>(Goobes et al., 2006a)</td>
</tr>
<tr>
<td>Y34 $\psi$</td>
<td>$^{13}\text{C}'$-$^{13}\text{C}'$</td>
<td>$4.0\pm0.5$ Å</td>
<td>$4.2$ Å</td>
<td>(Goobes et al., 2006a)</td>
</tr>
</tbody>
</table>
4.4 Discussion

Proteins can affect the formation of biominerals by directly binding a crystal face (Elhadj et al., 2006, Shiraga et al., 1992, Boskey et al., 1993, Sollner et al., 2003, Naka and Chujo, 2001). Because the structure of interacting biomolecules influences function and mechanism, determining the structure of biomineral-associated proteins is paramount for understanding how proteins influence biomineralization. In this chapter I devised and implemented a method that provides the most detailed description of solid-surface adsorbed protein structure to date. This method could provide answers to key questions posed in chapters 2 and 3. Are surfactant proteins structured in the same way that solvated proteins or protein-protein complexes are? Can proteins recognize specific crystal surfaces, or are surfactant proteins promiscuous? Determining structures of ~10-20 biomineral-associated proteins using the method developed in this chapter could be useful for addressing questions of structure and specificity.

Little high-resolution structural data is available for protein solid-surface complexes. In the absence of this data, structural models based entirely on computation are difficult to validate. Similarly, it is difficult to build structural models based on minimal high-resolution experimental data without the use of computation. The combined method developed here uses a highly successful structure prediction algorithm to search conformation space deemed relevant by high-resolution experimental data. These techniques appear to be an excellent...
complement, resulting in structural models with increased resolution, certainty, and scope.

Particularly encouraging is the fact that most ssNMR measurements used for biased predictions (tables 4.1 and 4.2) had been predicted using RosettaSurface in the absence of experimental bias (chapter 3). This included predicting the preferential adsorption of Glu5 relative to Glu4, the adsorption of basic residues Lys6, Arg9, and Arg10, the helical structure of the binding domain, and the lack of adsorption in the C-terminal domain. This predictive ability suggests that iterative structure determination using the combined approach could begin with structure prediction rather than ssNMR experiment. Beginning with structure prediction could reduce the amount of ssNMR data required for structure determination by informing the initial placement of isotopic labels.

Despite the improved resolution of statherin’s HAp-adsorbed structure resulting from the combined approach presented here, uncertainties in the model remain. In chapter 3, I suggested that the ssNMR measurements Pro33-Tyr38 and Pro23-Tyr34 may arise from adjacently adsorbed statherin molecules rather than statherin intramolecular interactions. This suggestion reflected the fact that satisfying those measurements always created atomic clashes in RosettaSurface predicted structures. It is still uncertain whether the clashes arise from algorithmic inadequacies such as sampling, the energy function, or system representation, or if the signal was from adjacently adsorbed statherin molecules.
One potential algorithmic improvement would be to account for experimental error in the biasing term. In the biasing term, $k_i \sum_{i=1}^{n} (\Delta x_i)^2$, the spring constant (weight), $k$, could be divided by the experimental error (i.e., $k \sum_{i=1}^{n} \left(\frac{\Delta x_i}{\sigma_i}\right)^2$, where $\sigma$ is the error of the $i^{th}$ experimental measurement).

Accounting for the experimental error in this way would match the “stiffness” of the “spring” to bias more precise measurements more strongly and to allow uncertain constraints to fluctuate more.

With the advent of this combined approach, analysis can now be made at a sufficiently high resolution to begin understanding residue- and atom-specific contributions to the process of biomineralization and hard tissue formation. A test of the success and general applicability of the combined method will require investigation on other protein biomineral systems.
Chapter 5

Rational design of peptide-calcite biomineralization systems

Under my supervision, Elizabeth Specht and Sarah Schrier contributed greatly to the experimental work in this chapter.

5.1 Introduction

The biomineralization of calcified hard tissues appears to rely on acidic protein residues (Addadi and Weiner, 1985, Fu et al., 2005, Gotliv et al., 2005, Tsukamoto et al., 2004) (e.g. aspartate, glutamate, and phosphorylated serine) more than basic residues (e.g. arginine and lysine). This apparent disparity may be surprising considering that minerals are composed of both anions and cations. High acidic amino-acid content imparts an intrinsic disorder in the predicted binding domains of some mineral-associated proteins (Kapon et al., 2008, Amos and Evans, 2009, Buchko et al., 2008). This high uniform charge and resulting disorder are at least part of the reason that some biomineralization proteins exhibit binding promiscuity (Delak et al.). For instance, osteopontin binds hydroxyapatite (Addison et al., 2009), calcite (Chien et al., 2008), and calcium oxalate crystals (Chien et al., 2009); lithostathine is found associated with different mineral types during litholysis (Ryall, 1996, Bernard et al., 1995); and poly-Asp can modify the crystallization of multiple mineral types and phases (Politi et al., 2007, Kasugai et al., 2000, Chien et al., 2009). In vivo, this promiscuity can be advantageous (Delak et al., 2009), and can be controlled by protein expression, matrix mediation, or boundary organization (Mann, 2001).
Because contemporary *in vitro* methods in molecular biology cannot achieve the spatiotemporal control utilized by living cells, the tailoring of biomaterials with nanoscale precision *in vitro* may require the ability to design protein-mineral interactions that have greater specificity.

Several groups have developed successful design strategies in biomineralization problems (Ajikumar *et al.*, 2003, Brown, 1997, Capriotti *et al.*, 2007, DeOliveira and Laursen, 1997, Elhadj *et al.*, 2006, Sarikaya *et al.*, 2003). Those strategies relied on modifying naturally evolved proteins or employing directed evolution techniques. Also, one group has predicted novel material-binding peptide sequences *in silico* (Oren *et al.*, 2007). In that work, bioinformatics techniques were used to analyze peptide sequences previously isolated from directed evolution experiments. Ideally, *de novo* design of biomineralization systems is possible. In the computational *de novo* design of proteins, a protein sequence is optimized to achieve a desired phenotype, such as binding a mineral surface. *De novo* design methods do not rely on evolved sequences and may afford added specificity in designing protein-mineral interactions.

Over the past decade, development of the structure-prediction algorithm Rosetta led to unprecedented achievements in many protein design problems (the RosettaDesign algorithm). These achievements include the design of a novel protein fold (Kuhlman *et al.*, 2003), β-sandwich proteins (Hu *et al.*, 2008), enzyme catalytic sites (Ashworth *et al.*, 2006, Jiang *et al.*, 2008, Röthlisberger *et al.*, 2008), protein-protein interaction specificity (Kortemme *et al.*, 2004), protein-DNA
specificity (Ashworth et al., 2006), protein-peptide specificity (Sood and Baker, 2006), and protein folding pathways (Nauli et al., 2001). In chapters 2-4, I developed Rosetta to predict the structure of proteins adsorbed to biominerals by incorporating the energetics of protein-solid surface interactions (RosettaSurface) (Makrodimitris et al., 2007, Masica and Gray, 2009).

For the present investigation, I sought to further develop principles for the rational design of biomineralization systems, giving special consideration to peptide charge, composition, and sequence. To accomplish this goal I developed RosettaSurface to simultaneously optimize the fold, orientation, and sequence of a protein adsorbed to a crystal surface (RosettaSurface.Design). I used RosettaSurface.Design to design de novo $10^5$ peptide sequences (16 residues) to bind differentially terminated states of a calcite ($\text{CaC}_0_3$) growth plane. From the predicted libraries, we chemically synthesized six sequences and tested their effect on calcite crystallization. In addition, my team and I synthesized scrambled variants of the six designer peptides, and compared the effects of designed and scrambled variant sequences on calcite crystallization. This is, to my knowledge, the first structure-prediction-based algorithm capable of protein design on a solid surface.

5.2 Materials and methods

**Algorithm development and implementation.** The algorithm developed here is based on RosettaSurface (Masica and Gray, 2009). Briefly, each execution of the RosettaSurface algorithm folds a peptide from a fully extended
conformation in solution using a united atom model. Then, a high-resolution (all-atom) representation of the peptide is minimized in solution and adsorbed to a biomineral surface in a random orientation. The protein’s fold and orientation are optimized on the surface resulting in one candidate adsorbed-state structure.

Development of RosettaSurfaceDesign included the following modifications: 1) the initial high-resolution representation of the protein is constructed using a sequence optimization algorithm (design) (Kuhlman and Baker, 2000). 2) The first, third, and fifth cycle of the high-resolution phase end in sequence optimization. During sequence optimization, sequence space is searched using a simulated-annealing Metropolis Monte Carlo procedure that replaces a side chain at a random position with a side chain from a backbone-dependent rotamer library (Dunbrack Jr and Cohen, 1997). There are ~10⁶ substitutions per Monte Carlo cycle.

The force field parameters for calcite were chosen as follows: The Lennard-Jones well depth, $\varepsilon$, and internuclear separation, $\sigma$, were taken from Stockelmann et al. (Stockelmann and Hentschke, 1999). For the hydrogen-bond function, the carbonate oxygen atoms (sp² hybridization) are proton acceptors. The Gaussian solvent-exclusion model for solvation free energy was parameterized from calculated water density profiles (Kerisit and Parker, 2004). Atomic charges were assigned from quantum calculations (Catti, 2001).

During sequence optimization, energetic penalties of 2.0, 2.5, 0.5, and 1.0 kcal/mol were used for the incorporation of Lys, Arg, Asp, and Glu residues,
respectively. The penalties prevented the over design of charged residues at the charged calcite surface and helped ensure sequence diversity.

**Starting materials for simulation.** An extended molecular structure of a 16-residue peptide was constructed using PyMol (DeLano, 2002). In each simulation the initial peptide sequence was 15 alanines (residues 1-15) followed by a tyrosine (residue 16; included for concentration determination in the experimental constructs). Residues 1-15 were designed during decoy generation, while the tyrosine at position 16 was fixed.

Sequences for scrambled variants were generated using a random number generator, using as input, the sequences of the corresponding designer peptide and conserving the tyrosine residue at position 16 for concentration determination in the experimental construct.

Coordinates for calcite surfaces were generated using CrystalMaker software (Palmer, 2003). Appropriate cuts were made to expose \{001\} surfaces with either a net negative or net positive charge.

**Peptide synthesis, purification, and verification.** All peptides were synthesized using standard fmoc (9-fluorenylmethoxycarbonyl) chemistry on a solid Anaspec fmoc-tyr-wang resin using a Protein Technologies Symphony Quartet automated peptide synthesizer (Atherton and Sheppard, 1989). The synthesis program used double coupling and included a capping step after each amino acid coupling to increase the yield of the desired product. All amino acid reagent concentrations were 200 mM. The peptides were cleaved from the dried
resin by exposure to 4 mls of trifluoroacetic acid/anisole/water 95/2.5/2.5 (by volume) for two hours at room temperature with constant agitation. After precipitating the cleavage reaction products with ice-cold diethyl ether and washing with diethyl ether two subsequent times, the crude product was reversed-phase HPLC purified. The samples were injected onto an Agilent ZORBAX C-18 column at 40°C and eluted using a 0-100% gradient of acetonitrile and water, each with 0.1% trifluoroacetic acid. The UV detector measured absorbance at 254 nm and 280 nm, to detect backbone signal and tryptophan residues respectively. The collected peaks were then analyzed using a Finnigan LCQ ion trap mass spectrometer with electrospray ionization to determine the peak with the correct product mass.

Crystal growth experiments. Calcite crystals were grown from saturated Ca(HCO$_3$)$_2$ solution. Solution was prepared by bubbling CO$_2$ gas over 6.7 mM CaCO$_3$ solution for one hour (Mann et al., 1990). 7 µl of solution was placed on a glass cover slip, siliconized with hexamethyldisilazane (Alfa Aesar), and then the cover slip was inverted over a 24-well plate. Cover slips were sealed with vacuum grease allowing each drop to grow in its own microenvironment. Crystals were allowed to nucleate at room temperature for 20 minutes, at which time the seal was broken and either 3µl of water or 3µL of peptide solution was added (DeOliveira and Laursen, 1997). Final peptide concentration was 0.45mg/ml. Crystals were grown overnight. Controls (crystals grown in the absence of modifiers) were confirmed by light microscopy on an Axiovert 200M to ensure
consistency of experiments. All crystallization experiments were done in triplicate a minimum of three independent times (nine times total per peptide).

**Scanning electron microscopy.** Crystals were air dried, sputter-coated with platinum, and viewed with an FEI Quanta ESEM 200, operating at 3 kV. All SEMs presented here were chosen to represent the typical biomineralization activity observed for each corresponding peptide.

### 5.3 Results

The {001} and {012} calcite surfaces arise from successive deposition of alternating negatively and positively charged ions. Both the {001} and {012} surfaces have been implicated as relevant growth planes during calcite crystallization (Weiner and Addadi, 1997, Addadi et al., 2008, Berman et al., 1995); for this work we restricted computational design to peptides targeting the {001} surface. The cartoon in Fig. 5.1a depicts an almost mature calcite crystal, expressing six stable {104} planes and one high-energy {001} plane.

**Figure. 5.1: Calcite.** (A) Sketch of the targeted {001} growth plane and the stable {104} planes of calcite. (B) SEM of calcite native crystal (grown in the absence of peptides) expressing six stable {104} faces.
Fig. 5.1b is a scanning electron microscope (SEM) image of a calcite crystal showing typical rhombohedral morphology.

For this work we targeted two differentially terminated calcite {001} planes. The terminations were chosen to meet two criteria: 1) have a net charge of either negative or positive and 2) be heterocharged, and hence more stable than a homocharged termination. The more stable surface conformers will be more populated during experimental biomineralization. Further, heterocharged surface chemistries select for more heteropolymeric peptide sequences during RosettaSurface.Design implementation. Heteropolymeric sequences scramble in more unique ways, increasing the likelihood that designs and variants will bind differently; i.e., exhibit sequence-order specificity. The two mixed charge surfaces are shown in Fig. 5.2. The calcium density of the net negatively charged surface (figure 5.2a) is twice that of the calcium density of the net positively charged surface (figure 5.2b).

Figure 5.2: Net positively charged (A) and negatively charged (B) calcite surface terminations used in design simulations. Calcium atoms are green; carbon atoms are gray; oxygen atoms are red.
For this study, the design algorithm was supplied with a reduced amino-acid library containing only Ala, Pro, Gly, Glu, Asp, Arg, and Lys. The reduced amino-acid library includes residues that are charged at neutral pH (Asp, Glu, Arg, and Lys) and those that typically play a structural role (Ala, Pro, Gly).

We used RosettaSurface.Design to generate $10^5$ sequences on both the net positively and net negatively charged surfaces (see *Materials and methods*). Six peptides were chosen for chemical synthesis based first on predicted adsorbed-state energy and then on sequence (Table 5.1). Ideally, the binding affinity (adsorbed-state minus solution-state free energy) could be calculated and used as a discriminating parameter.

<table>
<thead>
<tr>
<th>Number</th>
<th>Design Sequence</th>
<th>Scrambled Variant</th>
<th>Net Charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GEAEGEEAAAGEGGAY</td>
<td>EGGAAGAEAEHGAEGY</td>
<td>-5</td>
</tr>
<tr>
<td>2</td>
<td>GEEAAADAAGAAEAGAY</td>
<td>AAAAGAEAGAAEEDY</td>
<td>-5</td>
</tr>
<tr>
<td>3</td>
<td>AKAPKDGRAKEGGAAY</td>
<td>AAGPDAKKARGGEAKY</td>
<td>+2</td>
</tr>
<tr>
<td>4</td>
<td>GAAAAARKAEKGAKAY</td>
<td>AARGKGAAAEEAKAAY</td>
<td>+3</td>
</tr>
<tr>
<td>5</td>
<td>APPRAKAAKAAAAGKY</td>
<td>AAAKAPAAGKPARAY</td>
<td>+4</td>
</tr>
<tr>
<td>6</td>
<td>GPPPPAKAAKKAALKY</td>
<td>AKAPKKAPPPAALKPY</td>
<td>+5</td>
</tr>
</tbody>
</table>

**Table 5.1: Synthesized peptide sequences.** *Number* defines the naming convention throughout, e.g., *design 1, variant 1*, etc. Designed and variant sequences are described using one-letter amino-acid codes. Net charge is the difference in acidic and basic amino-acid content.
Predicting the binding affinity would require predicting the solution-state energy for each of the $10^5$ candidate sequences in each run and is computationally prohibitive. The adsorbed-state energy represents the strength of interaction between the peptide and surface and is computed for each candidate sequence generated by RosettaSurface.Design. We chose sequences of both uniform (negative or positive) and mixed charge to investigate the effects of peptide net charge and relative charge composition on calcite biomineralization.

Figs. 5.3a and 5.3b show the predicted structures of two sequences (designs 1 and 4 respectively) evolved by RosettaSurface.Design to bind the calcite {001} growth plane. In Fig. 5.3a, design 1 binds a positively charged calcite surface by aligning all of its aspartic-acid side chains with adjacent rows of exposed calcium atoms. In Fig. 5.3b, design 4 binds a negatively charged calcite surface by filling adjacent rows of calcium vacancies with its basic residues. All designer peptides chosen for subsequent chemical synthesis recognize calcite in a similar manner.

We grew calcite crystals in the presence of the synthesized designer peptides and observed the resulting crystal morphology with a SEM. All six designer peptides affected calcite crystallization significantly (Fig. 5.4a-f). The modification in crystallization resulting from inoculation with negatively charged design 1 (Fig. 5.4a) produced macro stepping and kinking. The only other net negatively charged design, design 2 (Fig. 5.4b), produced many Ca$_2$CO$_3$ crystals that were spherical. An average reduction in size and an increased occurrence in crystal twinning were observed with designs 1 and 2.
Figure 5.3: Structures predicted by RosettaSurface.Design for (A) design 1 bound to a net positively charged \{001\} calcite surface and (B) design 4 bound to a net negatively charged \{001\} calcite surface.

All of our net positively charged designer peptides (designs 3-6) modified crystallization similar to one another (Figs. 5.4c-f), with design 5 (Fig. 5.4e) being a possible outlier. The distinct feature in these crystals is the heavily stepped corners and edges. These crystals were typically smaller than unmodified crystals and also displayed an increased occurrence in twinning. The biggest difference between crystals grown in the presence of the negatively charged designs and
those grown in the presence of positively charged designs is the location of the steps. Crystals grown in the presence of negatively charged designs have considerable stepping on all faces; the resulting crystals lose most of their rhombohedral morphology. In contrast, crystals grown in the presence of positively charged designs have steps located almost exclusively at crystal edges and corners, leaving much of the rhombohedral morphology intact.

Next we synthesized variants of the designer peptides, whose sequences were each randomly scrambled, and grew calcite crystals in the presence of the scrambled variants (table 1). Fig. 4g shows crystals grown in the presence of variant 1. Crystals grown while inoculated with Design 1 (Fig. 5.4a) and variant 1 (Fig. 5.4g) share a similar morphology, which includes the macro stepping and kinking mentioned above. The other net negatively charged variant, variant 2, produced Ca$_2$CO$_3$ crystals baring almost no gross resemblance to rhombohedral calcite crystals (Fig. 4h). While both design 2 and variant 2 affected Ca$_2$CO$_3$ crystallization (Figs. 5.4b and 5.4h respectively), the effects were generally different.
Figure 5.4: SEM images for calcite crystallized in the presence 0.45 mg/ml peptide. (A) Design 1 (B) variant 1 (C) Design 2 (D) variant 2 (E) Design 3 (F) variant 3 (G) design 4 (H) variant 4 (I) Design 5 (J) variant 5 (K) design 6 (L) variant 6.
Scrambled variants of the net positively charged designer peptides displayed a range of ability to alter Ca\textsubscript{2}CO\textsubscript{3} crystallization. *Variant 3*, which is net positively charged but contains two negatively charged residues, had a similar effect to *design 3*. *Variant 4*, which is positively charged and contains only one negatively charged residue, had little effect on crystallization, in contrast to *design 4*. This difference suggests that sequence order was a partial determinant of activity, not just the sequence composition. *Variant 5* has no negatively charged residues and never created steps, but did create twinned crystals. *Design 5* also caused crystal twinning, though to a lesser extent than *variant 5*. Also, *design 5* created steps, whereas *variant 5* did not. This suggests some sequence-order specificity for *design 5*. *Variant 6*, which contains 5 basic residues and no acidic residues, had little effect on Ca\textsubscript{2}CO\textsubscript{3} crystallization. In contrast, *design 6* had significant effect on Ca\textsubscript{2}CO\textsubscript{3} crystallization, suggesting that sequence order, in addition to composition played a role.

### 5.4 Discussion

Recently, there has been significant progress in the *de novo* design of biological macromolecules (Hu *et al.*, 2008, Jiang *et al.*, 2008, Kortemme *et al.*, 2004, Kuhlman *et al.*, 2003, Nauli *et al.*, 2001, Röthlisberger *et al.*, 2008, Sood and Baker, 2006, Butterfoss and Kuhlman, 2006, Ashworth *et al.*, 2006). Particularly exciting is the design of proteins with novel properties. For instance, RosettaDesign successfully designed a protein fold not observed in nature (Kuhlman *et al.*, 2003) and successfully designed enzymes that catalyze
nonbiological reactions (Röthlisberger et al., 2008). Given the diversity of superior functional and structural materials fabricated in biological systems, the potential for novel materials resulting from the de novo design of biomineralization systems is seemingly limitless. One advantage of de novo design in biomineralization is the potential to target a specific crystal face. This includes the ability to target faces that are not well represented in the stable, fully developed crystal. Directed evolution techniques are useful for evolving novel material binding peptide sequences. But these directed evolution techniques suffer from an inability to design for a specific crystal face, and the experiment is typically limited to crystal faces present in the stable substrate material. The ability to design specific interactions in molecular biology is useful for recovering specific phenotypes.

Here we developed the first de novo design approach for biomineralization applications. The RosettaSurfaceDesign method designed peptides with sequence-order specificity and allowed us to identify basic residues as being a contributor to that specificity. Our results are in agreement with previous findings, where sequence order did not affect the biomineralization activity of peptides containing acidic residues and no basic residues (Elhadj et al., 2006). In our study, net negatively charged sequences (designed and variant) always produced drastic changes in calcite morphology. In contrast, some net positively charged sequences had little or no effect on calcite crystallization. Basic residues can affect calcite crystallization, but it appears acidic residues can affect crystallization more potently. The results here show a greater dependence on
specific sequence order for net positively charged peptides. Therefore, peptide-surface interactions involving basic residues may require greater optimization, in turn providing added specificity.

Some naturally evolved proteins involved in calcified tissue mineralization bind crystal surfaces via basic residues. Basic residues are essential for the interaction of human-salivary statherin with hydroxyapatite (the primary mineral component of tooth enamel) (Makrodimitris et al., 2007, Masica and Gray, 2009, Goobes et al., 2007b, Raghunathan et al., 2006). We predicted statherin’s basic residues to be important for specific recognition of a phosphate motif on the monoclinic {001} surface of hydroxyapatite (HAp) (Makrodimitris et al., 2007, Masica and Gray, 2009). Also, human lysozyme binds hydroxyapatite via basic residues (Aizawa et al., 1998). Previously unrecognized is the fact that human lysozyme and human statherin, two salivary proteins, bind hydroxyapatite with very similar motifs. These motifs comprise four basic residues oriented in an equilateral parallelogram, ~10 Å on each side; this is the geometry of open phosphate clusters replicated across the HAp {001} surface (Makrodimitris et al., 2007).

Increased specificity may be achieved by requiring greater stability in design constructs. Small peptides are typically flexible and can acquire multiple binding conformations. Increased stability would help overcome some of this intrinsic flexibility and limit the number of accessible binding states. Increased peptide chain length and disulfide bonding are two strategies that may increase stability. Osteocalcin is a well-studied HAp-biomineralization protein that exhibits these
stabilizing factors (Hoang et al., 2003). The structure of this ~43 residue protein is stabilized by a single conserved disulfide bond (Frazao et al., 2005, Hauschka and Carr, 1982). Osteocalcin binds HAp and does so with greater affinity in presence of reducing agents (Hauschka and Wians Jr, 1989). Why would evolution conserve the stable, disulfide bonded structure of osteocalcin, given that its affinity for HAp increases with decreasing structure? One possibility is that the folded conformation of osteocalcin facilitates a specific interaction with HAp. The disulfide-bonded structure of the stable protein can access fewer conformations and should therefore complement fewer HAp surfaces. Considering similar features during de novo design may result in interactions that are more stable and specific. Given the success of RosettaDesign in the designing stable, structured, globular proteins, the design of stable protein-adsorbed states may be possible with RosettaSurface.Design.

The de novo design of biomineralization systems presents some unique challenges. For instance, at the nanoscale, many different surface features may be present during the kinetic process of mineral crystallization. In addition, a crystal face may exhibit some disorder as ions encounter the surface and acquire their lattice positions. The similar chemistry and structure of many of these surface conformers may promote promiscuous protein binding in lieu of specific recognition. In the computational design of solution-state proteins, specificity has been achieved using so-called “negative design” to search for sequences that favor one binding partner over others (Grigoryan et al., 2009). In the case of biomineralization, the sheer number of alternate surface conformer possibilities
makes negative design difficult. The surfaces available to bind are dictated by the relative stabilities of different surface conformers and the kinetic process of crystallization. Thus, future designs will be improved by a greater understanding of the crystallization process, which can help guide the choice of surface conformers used for positive and negative design calculations. Also, increased affinity may result from negative design against the solution state.

Another challenge is relating gross crystal morphology to atomistic phenomena at the protein-surface interface. From SEM images one cannot ascertain in what crystal plane, termination, or orientation a peptide is bound. Similarly, an algorithm that does not capture crystallization dynamics cannot predict crystal morphologies resulting from predicted interactions. These disparities complicate the interpretation of results. For instance, different peptides could produce different crystal morphologies even if they adsorbed to the same crystal face. Alternatively, different peptides could produce the same morphology even while binding different faces or terminations. The crystal morphology resulting from protein-crystal interactions should have some dependence on the energetics of that interaction. This affinity will dictate binding at different stages of crystal growth and influence crystal growth kinetics and thermodynamics. In summary, peptide-induced morphological changes are more complex than the simple inhibition or enhancement of a single crystal face.

One benefit incurred from RosettaSurface.Design is that designed sequences have an increased probability of binding the target face. RosettaSurface.Design is fast and samples ~35,000 conformations and ~10^6 sequences for binding the
target surface. This approach allows the rapid generation of many peptides that are chemically and geometrically compatible with any given surface. In addition, each round of de novo design can inform subsequent rounds of design. This information can bias the algorithm to capture advantageous phenotypes observed in previous designs and avoid those that were deleterious. In this way the user can apply selective pressure to evolve desired traits. In this study, for instance, the reduced amino-acid library and reference energies were chosen based on observations made of naturally evolved systems (see Materials and methods). And, future RosettaSurface.Design predictions for calcite biomineralization systems could consider the determinants of sequence-order specificity gleaned here. For instance, crystallization modification had a greater dependence on sequence order for net positively charged sequences than net negatively charged sequences. This dependence on sequence order suggests a dependence on molecular structure. This relationship, sequence determines structure, is the important link between sequence-order specificity and crystal-face binding specificity. Therefore, the requirement to include basic amino acids in subsequent rounds of RosettaSurface.Design may help achieve a desired specific interaction. Additionally, we suggest negative design, expanded amino-acid libraries, increased chain length, and increased protein stability as potential future directions for improved design of biomineralization proteins.

5.5 Supplement
I used RosettaSurface, as described in chapter 3.2, to estimate binding affinities for the designed and scrambled variant sequences (table 5.2). For each of the 12 peptides (6 designed and 6 scrambled variants), I generated $10^5$ solution- and adsorbed-state structures and calculated the average energy of the low-energy ensemble (100 structures). Here, the binding affinity, $\Delta \Delta G_{\text{Ads}}$, is the difference between average energies for the low-energy ensembles ($\Delta G_{\text{Surf}} - \Delta G_{\text{Sol}}$).

<table>
<thead>
<tr>
<th>Design Sequence</th>
<th>$G_{\text{Surf}}$</th>
<th>$G_{\text{Sol}}$</th>
<th>$\Delta G_{\text{Ads}}$</th>
<th>Net Charge</th>
<th>Scrambled Sequence</th>
<th>$G_{\text{Surf}}$</th>
<th>$G_{\text{Sol}}$</th>
<th>$\Delta G_{\text{Ads}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEAEAGAAGGAEY</td>
<td>-48.8±2.0</td>
<td>-38.8±1.8</td>
<td>-10.0±4.3</td>
<td>-5</td>
<td>EGGAAGAEEGAEY</td>
<td>-48.3±2.1</td>
<td>-28.1±3.6</td>
<td>-20.1±4.2</td>
</tr>
<tr>
<td>GEEADAAGAEEAGY</td>
<td>-47.6±1.4</td>
<td>-31.7±1.4</td>
<td>-15.9±2.0</td>
<td>-5</td>
<td>AAAAGAEAGGAEDY</td>
<td>-32.7±2.1</td>
<td>-20.5±3.7</td>
<td>-11.7±4.3</td>
</tr>
<tr>
<td>AKAPKDRAEGGAAY</td>
<td>-71.1±4.6</td>
<td>-39.9±2.3</td>
<td>-31.2±5.1</td>
<td>+2</td>
<td>AAGPDAAKKARGGAEY</td>
<td>-64.2±3.7</td>
<td>-31.3±2.3</td>
<td>-32.8±4.3</td>
</tr>
<tr>
<td>GAAAAARKAEGAKAY</td>
<td>-85.7±3.8</td>
<td>-51.6±2.0</td>
<td>-34.1±4.3</td>
<td>+3</td>
<td>AARGKAAAEEKAAY</td>
<td>-76.2±4.1</td>
<td>-44.1±3.8</td>
<td>-32.1±5.6</td>
</tr>
<tr>
<td>APPRAAKAAAAAGKY</td>
<td>-94.1±4.9</td>
<td>-60.1±1.8</td>
<td>-34.0±5.2</td>
<td>+4</td>
<td>AAAKAPAAGPKARAY</td>
<td>-82.7±5.1</td>
<td>-51.8±1.6</td>
<td>-30.9±5.3</td>
</tr>
<tr>
<td>GPPPAKAARKAKKY</td>
<td>-89.6±4.4</td>
<td>-57.9±2.2</td>
<td>-31.7±4.9</td>
<td>+5</td>
<td>AKAPKAPPAGAKKY</td>
<td>-86.8±4.0</td>
<td>-51.1±2.0</td>
<td>-35.7±4.5</td>
</tr>
</tbody>
</table>

Table 5.2: Predicted binding affinities for the designed and scrambled variant sequences. For each synthesized peptide I generated $10^5$ structures using RosettaSurface (chapter 3.2). For each peptide, the predicted binding affinity for the respective calcite {001} surface is the difference in averages of the lowest-energy 100 structures in the adsorbed-state ($G_{\text{Surf}}$) and the difference in averages of the lowest-energy 100 structures in the solution-state ($G_{\text{Sol}}$); i.e., $G_{\text{Surf}} - G_{\text{Sol}} = \Delta G_{\text{Ads}}$. Errors for $G_{\text{Surf}}$ and $G_{\text{Sol}}$ are reported as the standard deviation and the error for $\Delta G_{\text{Ads}}$ is the square root of the sum of squares of the error for $G_{\text{Surf}}$ and $G_{\text{Sol}}$.

Except for design 1 (row 1 in table 5.2), predicted affinities for designed and scrambled variant sequences are within error of one another. Variant 1 is the
only variant with significantly lower (more favorable) predicted binding energy than the corresponding designer sequence. *Design 1 and variant 1* have nearly identical predicted interaction energies ($\Delta G_{\text{Surf}}$), but the predicted solution-state energy ($\Delta G_{\text{Sol}}$) of the *design 1* is considerably lower than *variant 1*. Therefore, predictions suggest that an unstable solution-state for *variant 1* is a primary contributor binding.

In general, predicted affinities did not prove useful in interpreting experimental results (figure 5.4). For instance, *variant 6* is predicted to bind with greater affinity than *design 6*, though *variant 6* had little effect on calcite crystallization and *design 6* had significant effect on calcite crystallization. However, these results may be expected. Because the simulation of all experimental conditions is not feasible (protein concentration, protein-protein interactions, salt concentration, simultaneous crystallization of all relevant faces, etc.,) the relevant free energy of adsorption cannot be calculated.
6.1 Conclusion

Consider how the advent of X-ray crystallography and solution-state NMR have advanced our understanding of biomolecules. The structures determined from these methods could be considered the foundation of molecular biology. From the rational design of disease therapeutics to understanding natural strategies in evolved phenotypes, these structures are instrumental for understanding the molecules of life. Now consider where molecular biophysics would be in the absence of those structures. That reflects, to some extent, the current state of affairs in protein biomineralization. Even membrane biophysicists commonly acknowledge the disparity of solved membrane protein structures relative to those available for globular proteins, and they have \( \sim 180 \) unique high-resolution structures (White, 2009)! For those of us studying the biophysics of biomineral-associated proteins, there are zero structures, and this is perhaps the biggest detriment to the field of protein biomineralization. The biomineral community will be well served by continued efforts to resolve atomic details at the protein-biomineral interface.

In section 1.4 I expounded the specific aims of my thesis, they were as follows. **Aim 1** develop a computational method to predict the structure of a protein adsorbed to a biomineral surface. To accomplish this aim I developed the Rosetta structure prediction suite to model both protein-solid surface docking (chapter 2) and protein folding on a solid surface (chapter 3), and employed these methods to predict the structure of statherin bound to hydroxyapatite
Aim 2) develop a combined computational-experimental method to determine the structure of a protein adsorbed to a biomineral surface. To accomplish this aim I further developed RosettaSurface to include solid-state NMR (ssNMR) data and employed that method to determine the structure of statherin bound to HAp (chapter 4). Aim 3) develop a computational method to design biomineralization systems \textit{de novo} and validate with experiment. To accomplish this aim I developed RosettaSurface to simultaneously optimize the fold, orientation, and sequence of biomineral-associated protein. I used that algorithm to design peptides to bind differentially terminated \{001\} calcite surfaces and experimentally assayed the peptides’ ability to alter calcite crystallization.

6.2 My contributions to the field of biomineralization

Structure prediction is a useful tool in biology. As reliable structure prediction algorithms develop they complement or help interpret experiment. This is synergy evident by the recent success in computational protein folding (Bradley \textit{et al.}, 2005b), docking (Gray \textit{et al.}, 2003a), and design (Kuhlman \textit{et al.}, 2003) and that of combined approaches (Cavalli \textit{et al.}, 2007, Shen \textit{et al.}, 2008, Gong \textit{et al.}, 2007). Structure prediction in biomineralization is powerful and unique in that protein-adsorbed state structure determination by experiment is currently impossible. For my thesis I developed the first algorithm capable of adsorbed-state structure prediction. I used this algorithm to accurately predict, \textit{a priori}, the fold and orientation of biomineral-associated protein, a molecular
recognition motif, and the preferential adsorption of one of two neighboring acidic residues located near a charged biomineral surface. The molecular recognition motif I predicted was the first of its kind and subsequently confirmed by ssNMR experiments. This algorithm is the only algorithm with the demonstrated ability to accurately predict high-resolution adsorbed-state protein structure a priori.

The novelty of the software makes it attractive for other, primarily experimental, groups. In work that does not appear in this thesis, I have successfully collaborated with several experimental groups, predicting the structure of natural- and laboratory-evolved material-binding proteins. I have published two studies in collaboration with Marc McKee at McGill University (Chien et al., 2009, Addison et al., 2009). In those studies I used RosettaSurface to predict the structure of differentially phosphorylated fragments of the bone protein osteopontin adsorbed to HAp and calcium oxalate dihydrate.

The development of a combined computational-experimental approach is, in my mind, my most important contribution to the field of biomineralization. While the full potential of this approach is far from realized, it has immediate utility. To that end, I have established collaborations with a significant fraction of the ssNMR spectroscopists working on biomineralization problems, including principal investigators Gary Drobny (University of Washington), Wendy Shaw (Pacific Northwest National Laboratory), Gil Goobes (Bar-Ilan University), and Joanna Long (University of Florida). Having experimental collaborations helps motivate continued algorithmic development and use. The use of combined
approaches also adds validity to structural models. I, with my collaborators, determined the structure of a biomineral-associated protein to unprecedented resolution. We are currently determining the structure of several other biomineral-associated proteins using the combined approach. We are now completing a study with Wendy Shaw where we determined the structure of the leucine-rich amelogenin protein bound to HAp. Also, we have begun to determine the structure of osteonectin and bone sialoprotein adsorbed to HAp.

The ability to design molecular interactions *de novo* not only provides the opportunity to extrapolate nature’s industrial prowess, but also demonstrates our understanding of molecular interactions. When I began this work in 2004, *de novo* design in biomineralization was seemingly out of reach, and I struggled with developing a working system for a long time. In 2009, Skelton *et al.* published a paper on simulating peptide and peptide mutants adsorbing to a titania surface; the final remarks in the discussion read:

*While in this work MD simulations have been used to infer connections with experimental data relating to point mutations, the next step is to go beyond this, e.g., to simulate scrambled strong-binder sequences. This approach would present new opportunities for simulation to connect with the underexplored area of peptide specificity. Understanding how to manipulate the preferential adsorption of peptides will be fundamental to the successful exploitation of these interfaces in the environmentally friendly fabrication of nanostructured materials in an aqueous solution.*

At that time I had just finished developing the design algorithm, designing material-binding sequences, scrambling them, and experimentally assaying their
effect on biomineralization. This marked the development of the first *de novo* design algorithm for biomineralization and the most extensive experimental investigation on the affect of peptide sequence order on biomineralization. The results indicated a potential role for basic residues in sequence-order specificity and may be used to design protein-mineral interactions with greater specificity in the near future.

### 6.3 Future directions

Despite realizing the abovementioned accomplishments, predicting the structure of protein adsorbed states is nascent and far from reliable let alone optimized. The absence of experimental structures makes parameterizing and benchmarking to experimental data impossible. For this reason, biomineral-associated protein structure prediction, while informative on some level, is highly speculative. I believe the continued use and development of RosettaSurface.NMR to be the most promising direction. In the near future we will have determined only two structures using this combined ssNMR-RosettaSurface approach, far too few make any generalizations about protein-adsorbed state structure. A data set of ~10-20 structures determined by iterative ssNMR-RosettaSurface.NMR would be very informative. Furthermore, those structures could serve as a benchmark to improve structure prediction protocols such as RosettaSurface.

RosettaSurface.Design was only tested once, and therefore its capabilities cannot be assessed with much certainty. In chapter 5, I designed peptides to
bind a biomineral surface. Peptides are intrinsically flexible, and can adopt many different binding configurations. To achieve the goal of specific design, I suggest designing more stable biomineral-associated proteins. A small tertiary fold (~40-50 residues) or a helical hairpin (~30-35) cyclized at the termini might both be fruitful second generation strategies. The ultimate goal would be to design a stable, structured protein biomineral complex using RosettaSurface.Design and determine its structure using iterative ssNMR-RosettaSurface.NMR.

Furthermore, the RosettaSurface suite needs to undergo developmental efforts aimed at ease of use. While I attempted to develop it for the scientific community at large, my main motivation was developing software from which I could ascertain scientific knowledge. With moderate development RosettaSurface can become more integrated at the academic level and eventually be incorporated into the Rosetta license for academics and industry.


PERSONAL

Bone February 4, 1978, Minneapolis, MN.

EDUCATION

The Johns Hopkins University

Program in Molecular Biophysics, Ph.D. expected fall 2009

Oakland University

B.A. Physics, 2003

With honors (GPA 3.8)

RESEARCH EXPERIENCE

Graduate Research Assistant, Johns Hopkins University

Advisor: Prof. Jeffrey J. Gray


- Developed the structure biology software Rosetta to account for protein-mineral surface interactions (RosettaSurface). The development led to the first structure-prediction-based software capable of predicting the fold and orientation of a
protein while associated with its target hard tissue (tooth enamel or bone for instance).

- Added functionality to RosettaSurface such that high-resolution experimental information could be used to bias simulation. This was an important step in protein structure determination at interfaces because experimental techniques (such as x-ray crystallography and solution NMR) don’t exist for protein structure determination at interfaces. Now a relatively small amount of experimental data (acquired from solid-state NMR (ssNMR) for instance) can be translated in plausible, high-resolution adsorbed-state structures.

- Established collaborations with several labs from around the world who collect ssNMR data for biomineralization proteins in the adsorbed state; we subsequently determined the first-ever reasonably high-resolution protein adsorbed-state structures resulting from a collaboration of experiment and computation.

- Added sequence optimization (design) capabilities to RosettaSurface. We then experimentally synthesized proteins/peptides that RosettaSurface designed to bind calcite and observed the resultant morphology of the biomineral calcite via scanning electron microscopy.

Undergraduate Research Assistant, Oakland University

Advisor: Prof. Uma D. Venkateswaran
Topic: Properties of carbon nanotubes as a function of pressure and temperature.

- Developed and implemented a program (written in LabView) to monitor resistivity of carbon nanotubes as a function of temperature.
- Studied the mechanical properties of carbon nanotubes as a function of pressure using Raman spectroscopy.

**PUBLICATIONS**


of phosphorylated osteopontin and poly-aspartate peptide showing occlusion by sectoral (compositional) zoning.” *J. Biol. Chem.*:M109.021899.


**SELECTED HONORS AND AWARDS**

- Genentech poster award, Society for Biological Engineering, 1st International Conference on Biomolecular Engineering. (2007)
• Provost-sponsored University Student Research Scholar Award, Oakland University. (2001)

PRESENTATIONS

Talks


Posters (for which Masica was the presenter)


4. Institute for Biophysical Research annual meeting, Baltimore, MD, November 2007 “Protein-Inorganic Material Surface Interactions.”


7. 2002 Sigma Xi annual meeting, Galveston Island, TX, November 2002. “Radial and Tangential Vibrational Modes in HiPCO-Derived Carbon Nanotubes Under Pressure.”

**SKILLS**

**Experiment:**

- Recombinant protein expression and purification (PCR, gel electrophoresis, dialysis, lysis, transformation, etc.).
- Peptide synthesis (FMOC, automated) and purification (HPLC and ESI-mass spectrometry).
- Bright-field, fluorescence, atomic-force, and scanning-electron microscopy.
- Ultraviolet, Raman, and circular-dichroism spectroscopy.
Computer:

- Programming: C++, python, shell, flash, html, javascript.
- Scientific software: Rosetta (developer), NAMD, DSSP, NACCESS, FADE, PyMol, GDB debugger, Origin, GNUplot, MPI.
- Publishing software: PhotoShop, Illustrator, Flash, Dreamweaver, Fireworks, Publisher, Powerpoint.
- OS: Windows, Linux, Mac.

TEACHING EXPERIENCE

- Teaching assistant, Cellular and Molecular Physiology, The Johns Hopkins University, fall 2005.

Mentored as primary research supervisor five Johns Hopkins undergraduate research assistants

1. Eleanor Glifort (2005-2006). Currently a graduate student at Drexel University, PA.
3. Elizabeth Specht (2006-present). Currently a graduate work at the University of California at San Diego.


REFERENCES

1. Prof. Jeffrey J. Gray, Ph.D.
Department of Chemical and Biomolecular Engineering
The Johns Hopkins University
3400 N. Charles St., Baltimore, MD 21218
410.516.5313, email: jgray@jhu.edu

2. Prof. Marc D. McKee, Ph.D.
Faculty of Dentistry
McGill University
3640 University St., Montreal, QC, Canada H3A 2B2
514.398.7203 ext. 00041, email: marc.mckee@mcgill.ca

3. Prof. Jan Hoh, Ph.D.
Department of Physiology
The Johns Hopkins University School of Medicine
735 N. Wolfe St, Baltimore, MD 21205
410.614.3795, email: jhoh@jhmi.edu