

© 2010 DAMIEN CHERIAN MATHEW

COMPUTATIONAL STUDIES OF ORIGINS OF LIFE SCENARIOS

BY

DAMIEN CHERIAN MATHEW

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biophysics & Computational Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Zaida (Zan) Luthey-Schulten, Chair
Professor Anthony Crofts
Professor Thom Dunning
Professor Klaus Schulten

Abstract

Understanding the origins of life on Earth is one of the most intriguing problems facing science today. In the research presented here, we apply computational methods to explore origins of life scenarios. In particular, we focus on the origins of the genetic code and the intersection between geochemistry and a primordial “biochemistry” in which mononucleotides could form short oligonucleotide chains. We also apply quantum chemical methods to a modern biochemical reaction, the charging of tRNA by an aminoacyl-tRNA synthetase, in order to shed light on the possible chemistry one may want to consider in problems relating to the origins of life.

The question of how codons came to be associated with specific amino acids in the present form of the genetic code is one fundamental part of gaining insight into the origins of life. Carl Woese and coworkers designed a series of experiments to test associations between amino acids and nucleobases that may have played a role in establishing the genetic code. Through these experiments it was found that a property of amino acids called the polar requirement (PR) is correlated to the organization of the codon table. No other property of amino acids has been found that correlates with the codon table as well as PR, indicating that PR is uniquely related to the modern genetic code. Using molecular dynamics simulations of amino acids in solutions of water and dimethylpyridine used to experimentally measure PR, we show that variations in the partitioning between the two phases as described by radial distribution functions correlate well with the measured PRs. Partition coefficients based on probability densities of the amino acids in each phase have the linear behavior with base concentration as suggested by the PR experiments.

We also investigate the possible roles of inorganic mineral surfaces in catalysis and stabilization of reactions essential for early forms of replicating systems that could have evolved into biochemical processes we know today. We study a proposed origins of life scenario involving the clay montmorillonite, as well as a generalized form of a charged surface, and their catalytic role in forming oligonucleotides from activated mononucleotides. Clay and mineral surfaces are important for concentrating the reactants and for promoting nucleotide polymerization reactions. Using classical molecular dynamics methods we provide atomic details of reactant conformations prior to polynucleotide formation, lending insight into previously reported experimental observations of this phenomenon. The simulations clarify the catalytic role of metal ions, demonstrate that reactions leading to correct linkages take place primarily in the interlayer, and explain the observed sequence selectivity in the elongation of the chain. The study comparing reaction probabilities involving L- and D- chiral forms of the reactants has found enhancement of homochiral over heterochiral products when catalyzed by montmorillonite.

Finally, we shift our perspective on the problem of the origins of life, by considering a modern biological reaction which is essential to all forms of life today: the charging of tRNA with correct amino acids according to their anticodons. These reactions are performed by amino-acyl tRNA synthetases (AARSs), and are essential for enforcing the genetic code. While studies involving the PR and code optimality apply to a more error-prone epoch of early biology, possibly forming “statistical proteins” whose sequence is de-

terminated probabilistically by a loose mechanism of assignment of amino acids based on (possibly) PR, the mechanisms that charge tRNA today are highly refined to charge only the correct amino acid to a tRNA, and are thus essential for the high-fidelity translation mechanism present in all living cells. To gain some insight into how the charging reaction may have come about, we apply quantum chemical methods to a problem of modern biology to gain a further understanding of the mechanisms behind biochemical reactions.

To my wife, Yuhua.

Acknowledgements

I would like to thank my wife, Yuhua, for her patience and support during my time in Graduate School. At the time I write this, the vast majority of our married life together has been stressful and emotionally trying years of the graduate school experience. Throughout, she has been extremely supportive, and I am very grateful to have her in my life.

My parents, Cherian and Marie, and sister, Teresa, have also been a source of moral support. I couldn't have gotten as far as I have without their confidence in me.

I am also grateful to the friends I have met while living in Champaign-Urbana who added humor and fellowship to my life while here. Neil Parthun, Heather Valdez Stillings, and many others associated with the Independent Media Center and the UIUC Graduate Employees Organization. It has been a pleasure and a privilege to get to know these people. I can't over stress how important it was for me to feel a sense of community outside of school-related work, and those friendships provided that.

This work was supported by NSF-FIBR # SBCSF0526747. Needless to say, without financial support I would never have been able to attend graduate school, much less pursue the fascinating subject area of the origins of life.

State of the art computing resources were also essential, and for that I thank the National Center for Supercomputing Applications and the Texas Advanced Computing Center. I was allocated computing time at these excellent high performance computing centers through Teragrid LRAC# TG-MCA03S027. I also relied heavily on the outstanding staff (Mike Hallock and Chuck Walbaum) and computational resources available through the U of I School of Chemical Sciences which received their funding from NSF CRIF 05-41659.

I am grateful to Nigel Goldenfeld and his student Tom Butler for their insightful and helpful discussions pertaining to the research described in Chapter 2.

Beth Lindquist and Taras Pogorelov provided very helpful discussions and suggestions regarding computational methods in quantum chemistry, which were invaluable in the work described in the last chapter of this dissertation.

The members of the Luthey-Schulten group, including Ke Chen, Alexis Black, Anurag Sethi, have helped create a supportive, intellectually stimulating and challenging environment. Alexis has also provided MR trajectories of the GluRS complex in both the pre- and post-transfer states, and these have provided valuable information about potential reactant configurations for use in Chapter 4.

Andrew Magis and Elijah Roberts were always of great help whenever I had computer-related questions, and Elijah Roberts was very helpful with getting the proper L^AT_EX formatting for this document.

Li Li has been most helpful for his knowledge of biochemistry and chemistry, as well as his ability to explain such things to someone trained only in physics.

Conversations with John Eargle were always enlightening, and helped rekindle an appreciation for science that can often be lost in the minutiae of research.

David Grych is a summer undergraduate research assistant with the Luthey-Schulten group, and has been helpful in contributing to the work in Chapter 4.

Cindy Dodds of the Center for Biophysics and Computational Biology office has helped tremendously in dealing with administrative matters pertaining to my enrollment, and has also been a good friend.

The administrative staff of the Inorganic/Materials/Physical Chemistry office (Beth Myler, Connie Knight, Theresa Struss and Sandy Pijanowski) have also been of great help in navigating administrative issues, scheduling, printing services. I am very grateful for their kind and friendly assistance.

I am highly indebted to my advisor, Zana Luthey-Schulten, for her patience and guidance through my career as a graduate student.

Table of Contents

List of Tables	ix
List of Figures	x
List of Abbreviations	xiii
Chapter 1 Introduction	1
Chapter 2 The Amino Acid Polar Requirement and Implications for the Origins and Structure of the Genetic Code	3
2.1 Background	3
2.2 A Microscopic Model of Partition Chromatography and its Relation to the Amino Acid Polar Requirement	7
2.3 Radial Distribution Function as Metric of Amino Acid Partitioning	8
2.4 Probabilistic Occupancy Mapping Provides a Glimpse of the Amino Acid Polar Requirement	11
2.5 Genetic Code Optimality and the Polar Requirement	14
2.6 Interactions Between Amino Acids and Nucleo-bases in Solution	17
2.7 Arginine-Adenine Associations	18
2.8 Arginine Associations with Guanine	19
2.9 Conclusions	21
2.10 Methods	22
Chapter 3 The Catalytic Role of Mineral Surfaces in the Formation of Oligonucleotides from Activated Mononucleotides	26
3.1 Background	26
3.2 Role of Confinement in Catalysis	30
3.3 Sequence Selectivity in Dimer Formation	34
3.4 Influence of Chirality on Catalytic Efficiency	35
3.5 Generalized Forms of Positive Surfaces and Their Potential Role in Polynucleotide Formation	36
3.6 Promotion of Nucleotide Oligomerization and Associations with Amino Acids	37
3.6.1 Selectivity in Oligonucleotide Formation	37
3.6.2 Promotion and Selectivity of Oligopeptide Formation	38
3.7 Conclusions	40

3.8	Methods	42
Chapter 4	Reaction Pathway for Charging tRNA in an Aminoacyl tRNA Synthetase	47
4.1	Background	47
4.2	Reaction Pathway: Uncatalyzed	48
4.3	Reaction Pathway: Catalyzed	49
4.4	Conclusions	52
4.5	Methods	53
Bibliography	54
Appendix A	Supporting Information	61
Curriculum Vitae	62

List of Tables

2.1	Comparison of calculated and experimental [1] PR	12
2.2	$W_{c,c'}$ transition and transversion values for genetic code optimality calculations	15
2.3	Comparison of P_b values for alternate genetic codes observed in living organisms today . . .	17
3.1	Peak $g(r)$ values for hetero- and homo- chiral monomer interaction	36
4.1	Residues from the GluRS catalytic site included at the classical level for calculations of the catalyzed tRNA ^{Glu} charging reaction	50
4.2	Relative configurational energies taken from optimized geometries for reactant and product states of the tRNA-charging reaction	51
4.3	Reaction coordinate values for calculations of pre- and post-transfer states as well as points indicated in Figure 4.2 with and without proton transfer between Glu α -amino and Glu41 (Neutral/Charged, respectively)	52

List of Figures

2.1	Codon table colored by amino acid polar requirement values measured by Woese and coworkers [1]	4
2.2	Representation of a typical paper chromatography experimental measurement. D is the total distance the solvent front has travelled relative to the starting point of the spotted sample. d is the distance travelled by the sample spot relative to the same.	5
2.3	Original experimental polar requirement data [1] for Methionine and Proline. The slope of the line when R_m versus of the mole fraction of water to DMP, when plotted in log-log, defines the PR for each amino acid in these experiments.	5
2.4	Typical snapshots from MD simulations of Left Leucine and Right Aspartate in DMP:H ₂ O solutions at 50% mole fraction water. Note that Asp tends to reside almost exclusively in a water bubble, while Leu tends to reside at the interface of the aqueous and organic phases.	6
2.5	Representation of partition chromatography. Orange dots represent sample molecules, in this example tending to reside mostly in the stationary phase of the chromatogram.	8
2.6	Relative water density as a function of distance from a hydrophilic surface in water-DMP solutions	9
2.7	RDFs for aspartate (a), arginine (b), proline (c) and valine (d) each at three representative water concentrations.	10
2.8	Linear trends in first RDF peak height for aspartate, arginine, valine and proline	11
2.9	Slopes of Peak RDF curves vs experimental PR	11
2.10	Codon table colored by amino acid polar requirement values.	13
2.11	Left Snapshot of asparagine molecules from simulation superimposed on average 10% probability isosurface for water from same simulation. (47% mole fraction water) Right Average 25% probability isosurface for asparagine from same simulation. (Red: Overlap of Asn and water occupancy, Blue: Overall Asn occupancy isosurface).	13
2.12	Trend in the partition coefficient, α , for asparagine with respect to water mole fraction. Slope=-1.003, $R^2=0.98$	14
2.13	Left Angular distribution function for arginine relative to the plane of adenine, depicting strong association between the two molecules. Equation for the calculation of the radial angular distribution function displayed above. Right Snapshot from MD simulation used to calculate $\langle \cos\theta(r) \rangle$ with arrows indicating the base dipole moment, \mathbf{p} , and Arg sidechain orientation, \mathbf{s}	19
2.14	Interaction between arginine and guanine in solution.	20

2.15	Radial distribution function between arginine and guanine sidechain.	20
3.1	Oceanic distributions of the clay mineral montmorillonite measured from deep sea sediments [2] shown as a fraction of the total sediment by color coding. Green X's mark dry-land locations of relatively high purity montmorillonite deposits (taken from the MinDat online mineral data archive: http://www.mindat.org/min-2821.html). The red X marks the location of one of the mid-Atlantic hydrothermal vent fields known as "Lost City."	28
3.2	Left Unactivated AMP adheres to a montmorillonite surface through stacking of the base and interaction with the 2' hydroxy Right 1-MeadpA adhesion orientation. Colored molecules represent structures from the simulations. Structures for both molecules are included below the snapshots. Lower clay sheet added to indicate interactions take place in the interlayer, however the separation distance has been compressed in the figure. (Other monomers, ions and water omitted for clarity.)	30
3.3	Adhesion of activated adenosine dimers (1-MeadpApA) with (Left) and without (Center) confinement in the interlayer. Right Molecular structure. (Other dimers, ions, and water omitted for clarity.)	31
3.4	Trimer formation from 1-MeadpApA + AMP . Radial distribution functions for Left exposed surface and Right Interlayer.	32
3.5	1-MeadpApA + AMP elongation reaction simulated in montmorillonite. (a) Two mechanisms for elongation of a dimer to a trimer that depend on positions Ca^{2+} are shown. (b) Snapshot from simulation of the dominant mechanism. Calcium ion (orange) ($\text{Ca}_{(A)}^{2+}$) mediates interaction of a terminal 3' oxygen (red sphere) with the phosphate (gray sphere) of the incoming AMP (c) Alternate configuration in which $\text{Ca}_{(B)}^{2+}$ brings the reactants in close proximity through 5' - 5' interactions. Both configurations contribute to the g_{max} in Fig. 3.4(Right). The 3' terminal oxygen and 5' AMP phosphorus are within 3.8Å in both cases. Water molecules surrounding the system have been hidden for clarity, but are explicitly present in the simulations.	33
3.6	Radial distribution functions for trimer formation from 1-MeadpApA + AMP in bulk solution without any surface present.	33
3.7	Sequence selectivity for U-A and A-U dimer formation. $g(r)$ s for (Left) AMP to 1-MeadpU and (Right) UMP + 1-MeadpA).	34
3.8	L- (Left) and D- (Right) Stereoisomers of 5' 1-Mead activated Adenosine	36
3.9	Radial distribution function of 5' phosphate oxygens to the 2' oxygens	38
3.10	Radial distribution function of 5' phosphate oxygens to the 3' oxygens	39
3.11	Radial distribution function of backbone amine hydrogens to carboxy oxygens of leu	39
3.12	Radial distribution function of backbone amine hydrogens to backbone carboxy oxygens of glu	40
3.13	Radial distribution function of backbone amine hydrogens of Glu to carboxy oxygens of Leu	41

3.14	Right: Two side views of the montmorillonite structure used. Left: Top view of structure. Silicon atoms represented in yellow, Oxygens in red, Aluminum and Magnesium in green, and Hydrogens in white	44
3.15	Sodium ions (yellow) in an array with equilibrated uracil mononucleotides electrostatically bound. The surface is completely saturated by uracil nucleotides, indicated by the lone uracil mononucleotide adrift in the solution. Waters, ions, and amino acids not shown.	46
4.1	Simplified model and reaction coordinates for the GluRS-tRNA ^{Glu} charging reaction.	48
4.2	Potential energy surface along the two reaction coordinates. Two plausible reaction paths are located. Both exhibit high potential energy barriers. Points labeled on plot (TS, PT1, PT2, and PT3) are used for calculations of the catalyzed reaction.	49
4.3	Left Snapshot from MD simulation including active site residues (all but Glu41 not shown). A76 O2' represented as larger red sphere, and Glu1-AMP carboxyl carbon represented as large cyan sphere. Right Schematic of interactions in MD snapshot, with distances labeled in Angstroms.	53

List of Abbreviations

1-MeadpA – 1-MethylAdenine-phospho-Adenosine
1-MeadpApA – 1-MethylAdenine-phospho-Adenosine-phospho-Adenosine
1-MeadpU – 1-MethylAdenine-phospho-Uracil
Ala – Alanine
AMP – Adenosine Monophosphate
Arg – Arginine
Asp – Aspartate
Cys – Cysteine
DMP – dimethyl pyridine
 $GC^i(x)$ – amino acid metric associated with a codon, x
Gly – Glycine
Glu – Glutamate
 $g(r)$ – Radial pair distribution function
His – Histidine
Ile – Isoleucine
Leu – Leucine
Lys – Lysine
MD – molecular dynamics
Met – Methionine
NMP – Nucleotide Monophosphate
 O_i – Optimality value for a variation of the genetic code, indexed by i
 P_b – Probability of a random genetic code being more optimal than the canonical code
Phe – Phenylalanine
PR – polar requirement
Pro – Proline
RDF – radial distribution function
Ser – Serine
Thr – Threonine
Trp – Tryptophan
Tyr – Tyrosine
UFF – Universal Force Field

UHF – Unrestricted Hartree-Fock

UMP – Uracil Monophosphate

Val – Valine

var – Variance

$W_{a,b}$ – Transition/transversion bias weightings for a codon, a , mutated to codon b

Chapter 1

Introduction

The quest to understand the origins of life on our planet is perhaps one of the most intriguing problems in science. Unlike cosmology's equally intriguing question of the origins of the universe, origins of life has been an exceptionally speculative field of study. In the modern world, there are complex cellular processes that replicate, transcribe, and translate the information molecules, DNA and RNA, into proteins. Each step is enzymatically controlled: Replication of DNA by DNA polymerases acting in conjunction with helicases, transcription of DNA into mRNA by RNA polymerases, and translation of mRNA into proteins by a nucleoprotein complex, the ribosome, kinetically coupled to protein-tRNA complexes that help maintain the genetic code. If early life on earth relied on archaic reactions without the benefit of protein catalysts, then the following questions arise: what are the origins of the primitive polymers, how did the polymerization reactions of peptides and oligonucleotides become genetically linked, and what are the universal principles that underlie the evolution of functional biomolecules? Progress in developing a detailed understanding of the Earth's early geochemistry, the discovery of nucleobases in rocky meteorites [3], and the production of amino acids in variations of Urey-Miller experiments have opened the field to rigorous scientific study. Still, we have yet to formulate a complete and testable hypothesis for the origins of life. However, computer simulations can provide valuable insight into proposed scenarios. Using computer simulations we have explored potential mechanisms behind the origins of the genetic code, the potential role of inorganic minerals such as clays as catalysts for forming chains of nucleotides, and have demonstrated the ability to utilize quantum chemical methods to elucidate biologically relevant reaction mechanisms.

Chapter 2 – The content of this chapter is derived from two of our previously published papers [4; 5]. We present a characteristic, quantifiable metric for amino acids and show that this metric, the polar requirement, is strongly correlated with the structure of the genetic code. Using molecular dynamics simulations, we show that the polar requirement (PR) results from the partitioning of an amino acid across an aqueous-organic interface. Updated values for the PR are derived from the MD simulation results, and the genetic codes error-tolerance optimality is tested with respect to the PR as a metric of amino acid similarity. The results show a remarkable degree of optimality for the canonical genetic code with respect to PR. Other amino acid metrics are compared to the PR and the genetic code's optimality tested. We find that our PR values are the most highly conserved amino acid metric known, with respect to error-tolerance optimality of the genetic code. We also study the interactions between amino acids and nucleobases in

solution, which may have played a role in the assignment of nucleotide triplets in the genetic code.

Chapter 3 – The chemical reactions that take place in living cells today are nearly without exception catalyzed reactions. The problem of catalysis is, therefore, essential in understanding the origins of biochemistry and life itself. In this chapter, we focus on an inorganic catalyst which promotes the synthesis of nucleotide chains from activated monomers. Positive ions have been shown to play a key role in bringing nucleotides together on surfaces to form polynucleotide chains. We also investigate generalized, positively-charged surfaces (as opposed to negatively-charged clays) to ascertain another possible set of catalytic surfaces for nucleotide polymerization. The contents of this chapter are based in part on previously published work by the author [6].

Chapter 4 – In this chapter, we shift our perspective on the problem of the origins of life. The preceding chapters took a forward-looking view, of how certain features of pre-biological systems may have transitioned into early forms of what might be called life, focussing on an RNA World view and addressing questions pertaining to the origin of the genetic code and the role of geochemistry and minerals. In this chapter, we attempt to address the question of the origin of the genetic code by looking at the primary enforcement agent of the code today, the aminoacyl tRNA synthetase. Gaining insight into the reaction mechanism of the synthetases may help shed light on how the transition from a biological era in which translation was more prone to error into the modern epoch of a high-fidelity translation apparatus with tRNA loaded with amino acids corresponding to their anticodon sequences. We use quantum chemistry methods to help elucidate biomolecular reactions responsible for the charging of tRNA^{Glu} by GluRS.

Chapter 2

The Amino Acid Polar Requirement and Implications for the Origins and Structure of the Genetic Code

2.1 Background

The origin of the genetic code is perhaps one of the most fundamental questions pertaining to the origins of life. In this chapter, we explore the relationship between the canonical genetic code and a property of amino acids, the polar requirement. We devise a computational simulation to capture the physical properties behind the polar requirement. Using these simulations, we establish a set of computationally-derived polar requirement values that are free from many of the experimental errors which were present when the original polar requirement was determined. Using these new polar requirement values, we revisit the relationship between the new polar requirement set and the genetic code, and find that the genetic code is optimized with respect to the polar requirement to a remarkable extent. This last result suggests that understanding the genetic code's evolutionary origins requires a consideration of the physical processes that manifest themselves in the amino acid polar requirement, as those very processes were exploited by whatever mechanism gave rise to the original genetic code.

The polar requirement (PR) is a characteristic property of each amino acid, originally defined by paper chromatographic experiments [1; 7; 8] in aqueous solutions of nucleobases. PR reflects the relative probabilities of an amino acid molecule to dissolve in an organic versus aqueous solvent in a binary mixture of the two. It has been shown that this property is related to the structure of the genetic code [9]. The original experimentally derived PR values are presented in Figure 2.1. Note that amino acids with pyrimidines in the second position of their codons tend to have low PRs while amino acids with high PRs tend to have purines in the middle position.

PR also groups amino acids that are not intuitively related [10]. Asparagine and lysine, for example, have nearly identical PR values, but very dissimilar side chains. The same can be said for glutamine and histidine. These groupings, in addition to being related to the structure of the genetic code, prove to be significant

The contents of this chapter are based in part on work previously published as Damien Mathew and Zaida Luthey-Schulten. "On the physical basis of the amino acid polar requirement," *J Mol Evol*, 66:519-528 (2008), and Tom Butler, Nigel Goldenfeld, Damien Mathew, and Zaida Luthey-Schulten. "Extreme genetic code optimality from a molecular dynamics calculation of amino acid polar requirement," *Phys Rev E*, 79:060901 (2009).

		Middle Position			
		U	C	G	A
First Position	U	Phe 5.0	Ser 7.5	Cys 4.8	Tyr 5.4
				Trp 5.2	
	C	Leu 4.9	Pro 6.6	Arg 9.1	His 8.4
					Gln 8.6
	G	Val 5.6	Ala 7.0	Gly 7.9	Asp 13.0
					Glu 12.5
	A	Ile 4.9	Thr 6.6	Ser 7.5	Asn 10.0
		Met 5.3		Arg 9.1	Lys 10.1
					U
					C

Figure 2.1: Codon table colored by amino acid polar requirement values measured by Woese and coworkers [1]

when the code's error tolerance is taken into account. Recent theoretical studies have demonstrated that the modern genetic code is extremely error tolerant against single base mutations when PR is used as a metric of amino acid similarity [11–15]. No property of amino acids other than PR has been found to yield the same code error tolerance. Is it highly improbable that the genetic code became optimized with respect to PR purely by chance. Though it is clear PR is correlated with the structure of the genetic code, the relationship is at present poorly understood. Furthering our understanding of the microscopic phenomenon behind PR may help advance our understanding of why this relationship exists.

The PR experiments were originally designed to probe stereochemical associations between amino acids and nucleobases in solution. Mixtures of water and dimethylpyridine (DMP) were prepared and used to develop paper chromatographs spotted with an amino acid. Multiple measurements were made for each amino acid, using a range of water:DMP molar ratios from $\sim 80\%$ to $\sim 40\%$ water. When the chromatographic factor R_m , defined in Eq. 2.1 as a function of the standard chromatographic retardation factor R_f , was plotted against the mole fraction water, χ_w , in log-log scale, a linear trend was observed for each amino acid. $R_f = d/D$, where d is the distance the sample has travelled, and D is the distance that the solvent front has travelled, as represented in Figure 2.2. The slope of this line defines PR for an amino acid (Eq. 2.2). Figure 2.3 presents original data from [1] as an example of how the data was plotted and analyzed to obtain the experimental PR data set.

$$R_m = (1 - R_f)/R_f \quad (2.1)$$

$$PR = - \left[\frac{d(\log R_m)}{d(\log \chi_w)} \right] \quad (2.2)$$

Using MD simulations we are able to quantify local differences in water distributions surrounding amino acids. Our MD simulations of amino acids in water-DMP solutions show that high PR amino acids are gen-

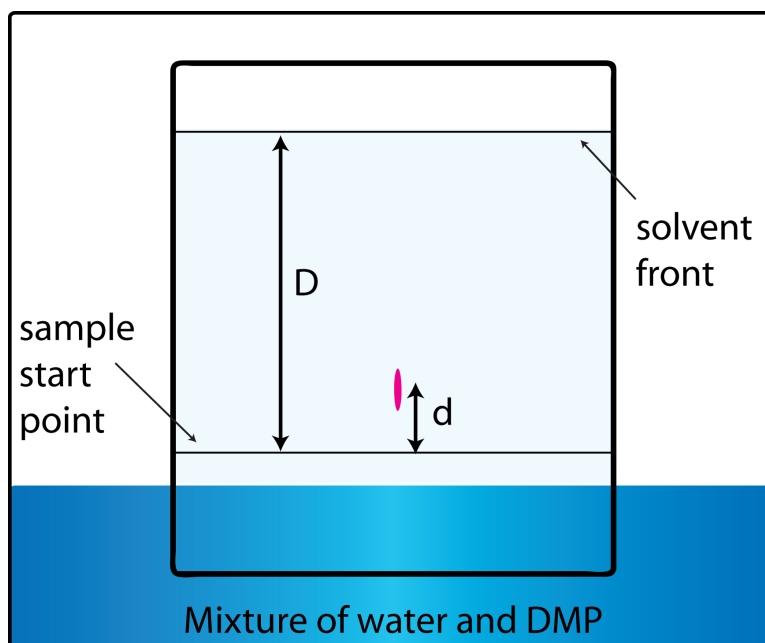


Figure 2.2: Representation of a typical paper chromatography experimental measurement. D is the total distance the solvent front has travelled relative to the starting point of the spotted sample. d is the distance travelled by the sample spot relative to the same.

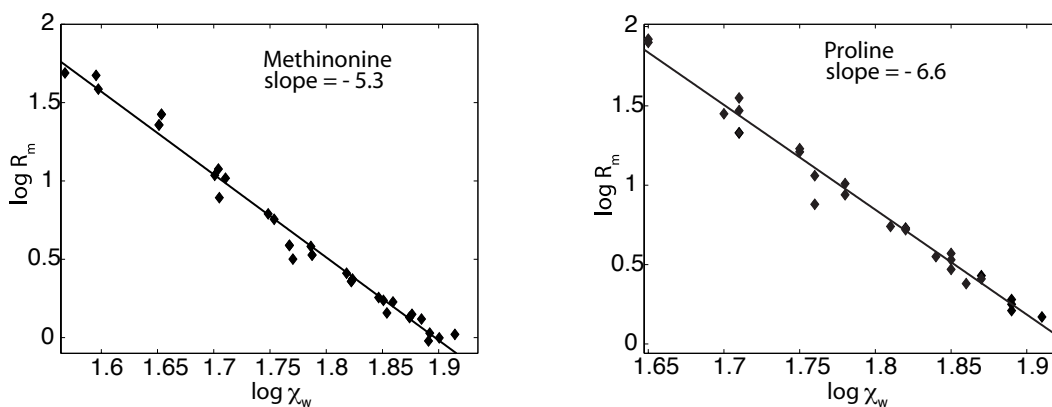


Figure 2.3: Original experimental polar requirement data [1] for Methionine and Proline. The slope of the line when R_m versus of the mole fraction of water to DMP, when plotted in log-log, defines the PR for each amino acid in these experiments.

erally completely surrounded by water. Low PR amino acids generally straddle an aqueous/organic interface with water predominantly surrounding the backbone groups and the side chains oriented towards the organic phase. Figure 2.4 gives a comparison between leucine and aspartate in solution, with all water and DMP molecules within 5\AA displayed. The Woese experiments found leucine's PR to be 4.9 and aspartate's to be 13. The differences in local solvent environment between the two amino acids is clearly visible. It can be inferred that leucine would partition more extensively into an organic phase than would aspartate. Amino

acids will partition into a chromatographic mobile phase as a function of water content and in a manner determined by the nature of the side chain (hydrophobic side chains will be effected by low water content to a lesser degree than hydrophilic side chains).

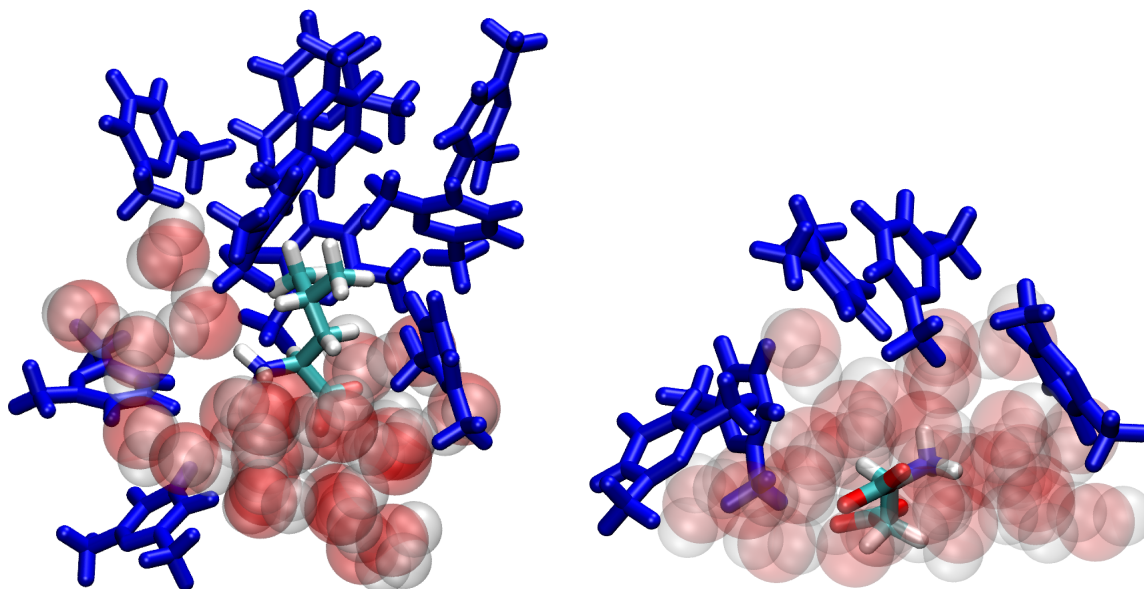


Figure 2.4: Typical snapshots from MD simulations of **Left** Leucine and **Right** Aspartate in DMP:H₂O solutions at 50% mole fraction water. Note that Asp tends to reside almost exclusively in a water bubble, while Leu tends to reside at the interface of the aqueous and organic phases.

We have designed two computational approaches to examine the local environment around amino acid side chains in binary solution using molecular dynamics (MD) simulations (see Methods). For a statistical mechanical treatment, radial distribution functions (RDFs) are used to quantify the microscopic differences in the water structure surrounding each of the amino acid side chains at the infinite dilute limit. We chose the most distant atom of the side chain as a reference for the calculation of RDFs and use the peak height of the first solvation peak as a measurement of maximum local water density [16]. This value relates to the average velocity of the amino acid in an experimental chromatogram, as described in the Methods section of this chapter. We show that the calculated trends on RDF peak height correlate well with the chromatographically observed PR values.

A more direct method of calculating PR involves determining amino acid partitioning by monitoring the spatial probabilities of occupancy for amino acid molecules in the aqueous and organic phases. The ratio of these two probabilities averaged over an MD trajectory should yield a direct measurement of the partition coefficient, α , as defined in equation 2.7. To improve average probabilities for this purpose, we use multiple amino acid molecules solvated in water-DMP solutions. As published results for water RDFs relative to 32 glycine molecules have previously been published [17], we use 32 amino acid molecules for each of these simulations. It is important to note that increasing the number of zwitterionic amino acids to a water-DMP solution can have a salting out effect on the miscibility of the two liquids. The increased concentration of amino acids results in a greater ionic strength of the aqueous phase. At these ionic strengths, the degree

of phase separation between the water and the DMP is strikingly different, which makes direct comparison between our results obtained by probabilistic spatial mapping and the infinite-dilute limit calculations using RDFs difficult.

Of the three components of a water-amino acid-DMP mixture, the fastest reorganization times are expected to be in the amino acid and water components. For our study, we ignore drift and currents of the two phases by averaging over time. We attempt to recover an approximate power law dependence of R_m on water content as observed experimentally.

2.2 A Microscopic Model of Partition Chromatography and its Relation to the Amino Acid Polar Requirement

Partition chromatography techniques such as the paper chromatographic method used in the original PR experiments generally assume a two phase model. In this model, a stationary liquid phase is taken to be a thin film of fluid that does not move relative to the substrate. This stationary phase remains immobile due to adhesive forces with the substrate, and thus this phase is immediately adjacent to the wetted surface. The second “mobile” phase moves at the same rate as the advancing solvent front, and consists of a liquid that is largely depleted of molecules with an affinity to adhere to the substrate. When a mixture of two solvents is used in such an experiment, the species with the higher affinity for the substrate will dominate the stationary phase and the mobile phase will be enriched in the species with less affinity for the substrate. In this model, the average rate that a sample would move up the chromatogram is a function of its partitioning between these two phases. The partitioning is related to the frequency at which the sample would transfer back and forth between the two phases and the relative amount of time spent in each phase. If a sample were to only reside in the mobile phase, it would move at the same rate as the advancing solvent front. Alternatively, if the sample resided solely in the stationary phase, it would not move up the chromatogram at all. In reality, a sample would be stepping on and off of the moving walkway of the mobile phase, and would have an average velocity that reflected the amount of time spent in the mobile phase. Brownian diffusion would account for the spreading of a spotted sample over time, but the movement of the center of the spot would be dictated by the mechanism described above. Figure 2.5 gives an representation of the microscopic view of our model. Here, the orange spots represent molecules or clusters of the sample. A_O represents the cross-section area of the organic (mobile) phase, and A_W is the cross-section area of the aqueous (stationary) phase.

Considering a partition chromatography experiment from a more microscopic perspective, the pure two-phase model breaks down. Having two homogeneous phases in direct contact would require an unphysically sharp change in chemical potential at the interface, assuming the two solvents were at least partially miscible. Additionally, basic fluid mechanics tells us that laminar fluid flow relative to stationary solid surface results in a linear velocity profile, where the flow velocity approaches zero relative velocity near the surface and would increase linearly as a function of distance from the surface up to a maximum flow rate at the edge of the fluid film furthest from the surface.

We clarified this microscopic picture by carrying out a simulation of the water-DMP mixtures in a box with a sheet of fixed water molecules arranged uniformly in a plane. This sheet represents a generic hydrophilic surface. In this laminar flow model we explain why, as a measure of local solvent environment,

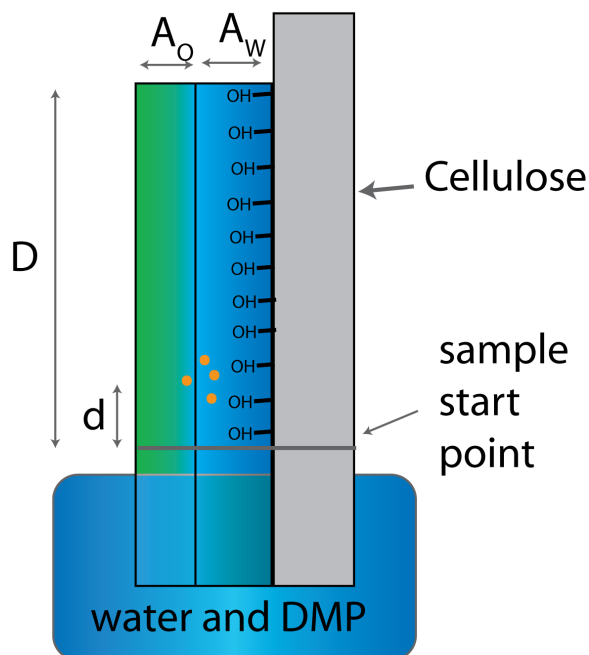


Figure 2.5: Representation of partition chromatography. Orange dots represent sample molecules, in this example tending to reside mostly in the stationary phase of the chromatogram.

an RDF is related to the chromatographic retardation factor. On a microscopic scale, there is a blending between the two phases at their interface with some water molecules in the “mobile phase” (based on the miscibility limit) and likewise for DMP in the “stationary” phase. The spatial dependence of the water concentration relative to a hydrophilic surface is shown for several bulk molar concentrations of water in Figure 2.6 below. The microscopic model revealed a linear dependence of max RDF with bulk concentration of water, and we argued how it together with the assumption of laminar flow can explain the correlation the max RDF data to PR and the power-law behavior of R_m .

2.3 Radial Distribution Function as Metric of Amino Acid Partitioning

Figure 2.7 gives four sets of representative RDFs for aspartate (a), arginine (b), proline (c), and valine (d). The general features of these four RDFs are representative of those observed for all amino acids. The first peak in each of these plots represents the aqueous solvation layer of the side chains, which generally occurs between $\sim 2.5\text{\AA}$ to 4\AA . Backbone contributions begin to appear beyond this peak with little or no overlap for most amino acids. Asp’s RDF (Fig. 2.7a) exhibits features representative of all the charged amino acids, with a strong first peak within 3\AA due to side chain solvation followed by a relatively low and broad second peak centered at $\sim 5\text{\AA}$ or greater (depending of the distance of the reference atom from the backbone) comprising a convolution of the second solvation layer of the side chain with the first solvation layers of the backbone groups. RDFs for amino acids with polar but uncharged side chains tend to have smaller, but still sharp first solvation peaks and are similar to that of Arg (Fig. 2.7b), whose side chain is charged but the charge is distributed over several atoms. Amino acids with particularly short side chains whose solvation

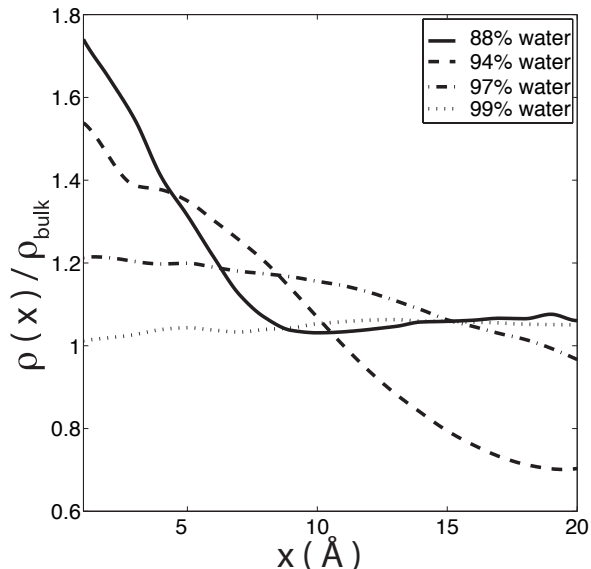


Figure 2.6: Relative water density as a function of distance from a hydrophilic surface in water-DMP solutions

centers are close to the backbone groups tend to have first solvation peaks which exhibit some overlap with the backbone solvation peaks, however even in these cases a distinct peak is discernible for both the side chain and backbone solvation. For Proline (Fig. 2.7c) the side chain solvation peak is observed near 3.7\AA and the backbone influence is centered near 5\AA . Even in the case of Pro, the side chain solvation leads to a distinct feature separable from the backbone influence. Alanine and glycine are exceptions due to the absence of a side chain (Gly) or a side chain short enough to be enveloped within the backbone solvation layer (Ala). Amino acids with non-polar side chains (Fig. 2.7d) tend to have smaller and broader first peaks which can partially overlap with the solvation peaks of the backbone. In all cases, except for Ala and Gly, side chain solvation leads to a distinct and separate peak from that of the backbone. For each amino acid there is a clear inverse relationship of first solvation peak height with water content. With decreasing water mole fractions, a linearly increasing relative local water density is observed surrounding the amino acid side chains. This linear RDF trend implies a power-law behavior in equilibrium solute distance from the support as a function of bulk mole fraction water to be consistent with our microscopic model.

In all cases there is a noticeable change in first solvation peak with water concentration, with the rate of change being the largest for amino acids with the highest reported PRs. The maximum of the first RDF peaks vary linearly with mole fraction water as shown in Fig. 2.8 for Asp, Arg, Val, and Pro. We calculated PR from our RDFs by making a linear least squares fit of these slopes with the corresponding experimental PR (Fig. 2.9). The high degree of correlation from our fit ($R^2 = 0.92$) confirms that the side chain RDF trends are related to PR. When we shift and rescale our calculated slope values using the best fit parameters, we obtain a set of calculated PR values.

A comparison of the calculated and experimental PR values is given in Table 2.1. While there is generally a single digit percent difference between the two, the computed PR values capture a slightly different structure in the codon table than does the experimental set. The most significant exception is tyrosine with a

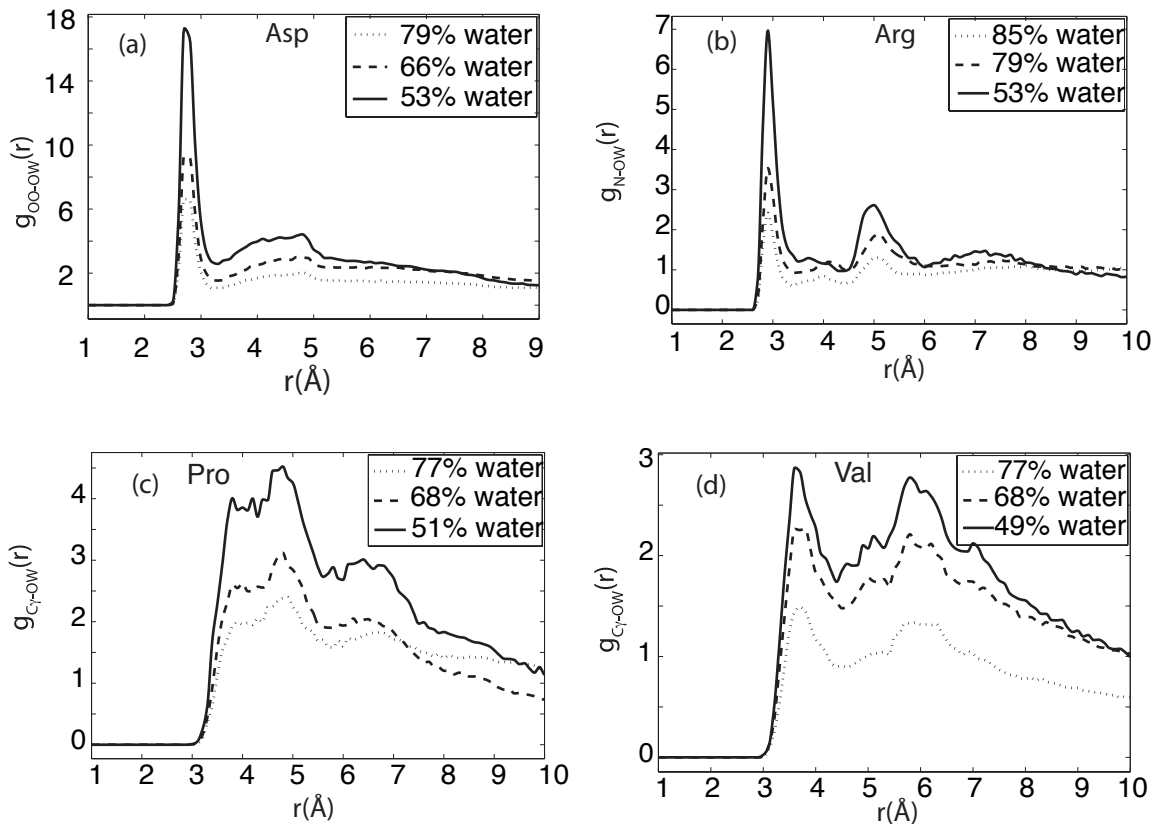


Figure 2.7: RDFs for aspartate (a), arginine (b), proline (c) and valine (d) each at three representative water concentrations.

35% difference. Tyr, which has a low experimental PR value has strong RDFs similar in form to that of Arg. It is worth noting that the calculated value for Tyr is more consistent with the PR values for amino acids that have adenine in their middle codon position. Additionally, it has been documented that tyrosine is an essential residue in cellulose binding domains [18]. Molecular modeling has confirmed that van der Waals interactions between tyrosine and cellulose are responsible for the induced fit leading to carbohydrate binding module activity [19]. The unique interactions between tyrosine and cellulose would retard the motion of tyrosine and lead to a systematic experimental error depressing the observed PR. When Tyr is excluded from the least squares fit, a much stronger correlation is observed between RDF rates of change and the experimental PR ($R^2 = 0.98$).

A qualitative ordering of the genetic code based on PR can be seen in Figure 2.10, with amino acids with pyrimidines in the second position of their codons tend to have low PRs while amino acids with high PRs tend to have purines in the middle position.

Both histidine and glutamine have multiple solvation centers in their side chains. A simple average of the RDFs for each solvation center was used to determine trends for these amino acids. The experimental PR values paired histidine and glutamine with similar values as well as asparagine and lysine. The computed PR values widen the PR difference between both of these pairs, though they are still relatively close. A

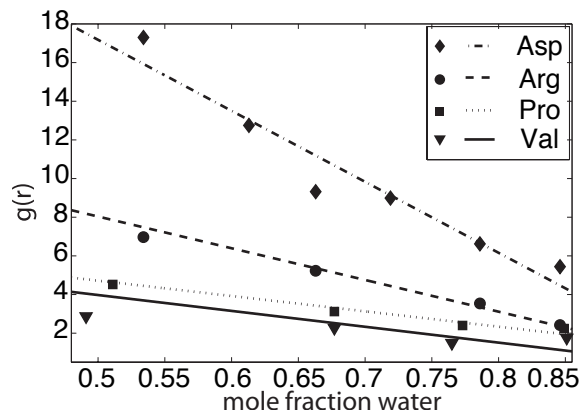


Figure 2.8: Linear trends in first RDF peak height for aspartate, arginine, valine and proline

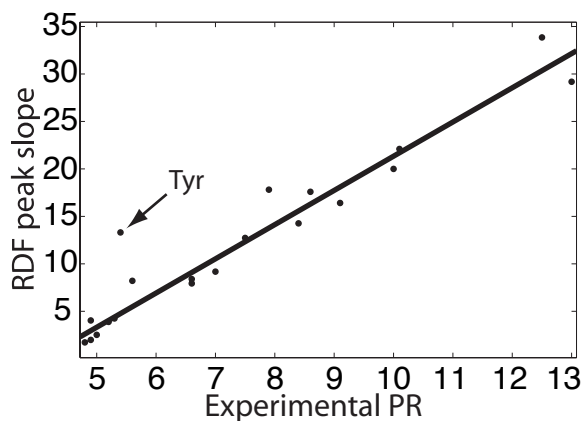


Figure 2.9: Slopes of Peak RDF curves vs experimental PR

weighted average reflecting the non-equilibrant solvation properties of the centers may have resulted in a closer matching of the PR values for these amino acids.

The calculated PR values for glycine also shows some deviation from the experimentally observed value. Use of the α -hydrogens of glycine as the reference atoms for the RDFs for this amino acid resulted in significant overlap of the backbone solvation layers with the solvation of these atoms and subsequently a high level of noise.

2.4 Probabilistic Occupancy Mapping Provides a Glimpse of the Amino Acid Polar Requirement

Figure 2.11 depicts results from a simulation of 32 asparagines in a 47% mole fraction water in water-DMP solution. The image on the left depicts the 10% isosurface for probabilities of water occupancy, regions of the simulated system that contained a water molecule at least 10% of the time ($P_i^{Water} > 0.10$) of the trajectories. The right image depicts occupancy probability isosurfaces for the amino acid, $P_i^{AA} > 0.25$

Table 2.1: Comparison of calculated and experimental [1] PR

Amino Acid	PR (exp)	PR (calc)	% Diff
Ala	7.0	6.5	7.9
Arg	9.1	8.6	6.1
Asn	10.0	9.6	4.1
Asp	13.0	12.2	6.0
Cys	4.8	4.3	10.7
Gln	8.6	8.9	3.4
Glu	12.5	13.6	8.4
Gly	7.9	9.0	12.6
His	8.4	7.9	5.7
Ile	4.9	5.0	1.6
Leu	4.9	4.4	11.0
Lys	10.1	10.2	1.0
Met	5.3	5.0	5.0
Phe	5.0	4.5	9.6
Pro	6.6	6.1	7.7
Ser	7.5	7.5	0.1
Thr	6.6	6.2	5.6
Trp	5.2	4.9	5.2
Tyr	5.4	7.7	34.6
Val	5.6	6.2	10.0

(Blue), and the convolution of the amino acid probability with that of water, $P_i^{AA} \cdot P_i^{Water}$ (Red). With respect to the amino acid molecules, we assume that over the sampled time window a near complete set of physically probable distributions of the molecules should be represented in the trajectory. It is interesting to note that the probabilistic isosurface for the amino acid averaged over 3ns closely mirrors the distribution of the amino acids from a single randomly chosen snapshot from the simulation shown on the left. It can be concluded that the amino acid, at this concentration, does not drift significantly from its most probable occupancy space during the 3ns trajectory.

As mentioned previously, the macroscopic partition coefficient, α , is directly related to the chromatographic factor, R_m (Eq. 2.6). It is therefore anticipated that a plot of $\log \alpha$, defined in Eq. 2.7 in terms of probability densities, should display a linear behavior with \log mole fraction water as observed in the PR experiments. Such linear trends are observed for glycine, proline, and asparagine. For each of these amino acids, slopes of order one were observed. All slopes were consistently below the corresponding experimental PR values, and no trend was observed in the steepness of the slopes corresponding to PR differences. For asparagine (Fig. 2.12) a slope of -1.003 is calculated. Experimentally, asparagine has a slope (from PR) of -10.0.

A number of assumptions were made in determining the probabilistic occupancy which would have to be modified to obtain better agreement with the experiments. One assumption was that the time-scale averaged over, the last 3ns of the simulation, was shorter than the characteristic lifetime of a domain boundary

		Middle Position			
First Position		U	C	G	A
	U	Phe 4.5	Ser 7.5	Cys 4.3	Tyr 7.7
				Trp 4.9	
	C	Leu 4.4	Pro 6.1	Arg 8.6	His 7.9
					Gln 8.9
	G	Val 6.2	Ala 6.5	Gly 9.0	Asp 12.2
					Glu 13.6
	A	Ile 5.0	Thr 6.2	Ser 7.5	Asn 9.6
		Met 5.0		Arg 8.6	Lys 10.2

Figure 2.10: Codon table colored by amino acid polar requirement values.

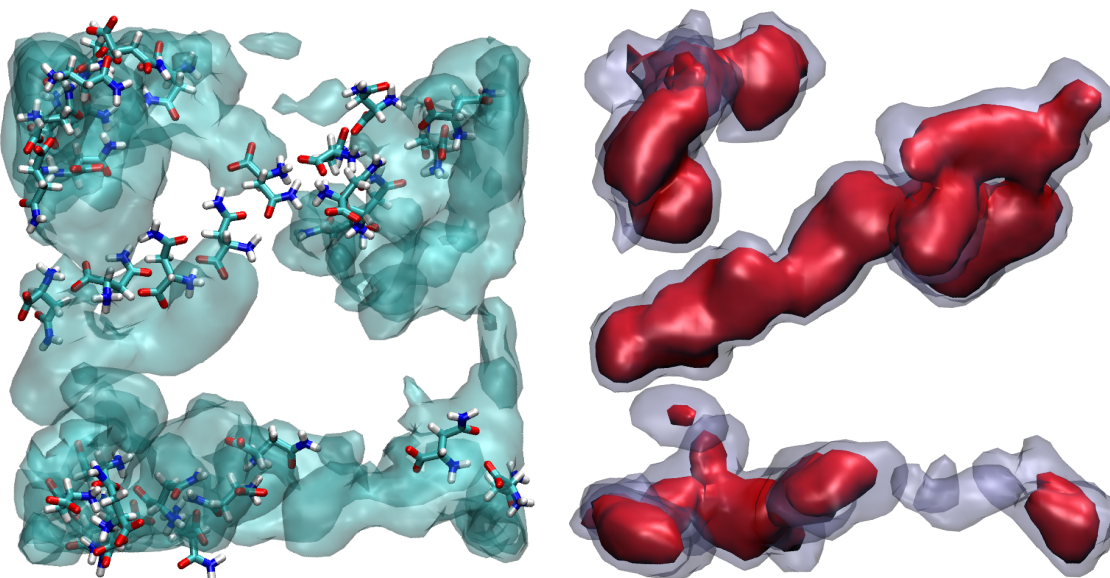


Figure 2.11: **Left** Snapshot of asparagine molecules from simulation superimposed on average 10% probability isosurface for water from same simulation. (47% mole fraction water) **Right** Average 25% probability isosurface for asparagine from same simulation. (Red: Overlap of Asn and water occupancy, Blue: Overall Asn occupancy isosurface).

between the organic and aqueous phases. In fact, there are fluctuations in domain boundaries over the ~ 1 ns timeframe, and so an appropriate reference was missing which introduces a degree of error to our calculations. Also, we approximate in this bulk model that the “mobile” chromatographic phase as comprised solely of DMP molecules. As discussed in our model used for RDF studies, neither phase is of pure water or DMP, and the interfacial region between the two are not captured in the model used in this study. A ratio of the amino acid probability overlaps with water and DMP is clearly an oversimplification. We have also explored water-DMP mixtures with varying concentrations of NaCl, and have found that ionic strength

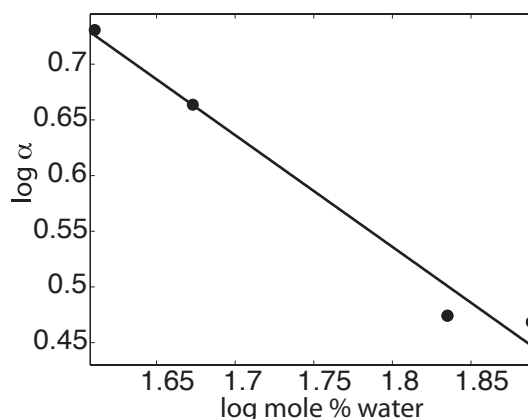


Figure 2.12: Trend in the partition coefficient, α , for asparagine with respect to water mole fraction. Slope=-1.003, $R^2=0.98$

can have a drastic effect on the degree of phase separation. Charged amino acids would increase the ionic strength of the aqueous phase, which will have an effect on the degree of phase separation. A more accurate determination of the partitioning effect would employ a frame-by-frame calculation of the occupancy overlap or a consensus score using a larger cell size prior to time averaging, although the reorganization times would vary with the amino acid's contribution to the ionic strength.

2.5 Genetic Code Optimality and the Polar Requirement

Aside from the qualitatively obvious correlation between the PR and the order of the genetic code (rough ordering of amino acids based by PR based on the middle position base in their codons), as we have mentioned in the introduction to this chapter, a number of quantitative studies have demonstrated the remarkable relationship between the structure of the genetic code and PR. These findings are significant, in that they suggest the origin of the genetic code was not the “frozen accident” suggested by Francis Crick [20] but rather followed a process of selection likely based on the chemical properties of the amino acids and the propensities to partition across an interface defined by an aqueous-organic liquid suspension or possibly even an inorganic surface in contact with an aqueous environment. While Crick's argument in favor of a frozen accident was compelling given what was known about the genetic code in 1968, more recent studies have demonstrated that the structure of the genetic code is, to a very high probability, not formed by a random assignment of codon blocks to amino acids. Building from earlier work using Monte Carlo methods [9] others have found that the canonical genetic code presents a high degree of optimization with respect to a number of properties of the amino acids [12; 13]. In particular, the work by Freeland and Hurst makes the eye-catching statement that the genetic code is “one in a million” when the PR as reported from the Woese experiments is used as a metric of amino acid similarity.

In each of these studies, randomized codes are generated by reassigning amino acids to the codon blocks currently present in the canonical code. For each randomized code, an optimality was determined by calculating the differences in amino acids that would result for a given point mutation in the codon. For example,

the codon GGG codes for glycine in the canonical code, which has a PR value of 9.0. A point mutation in the third position would still code for glycine, and so the total “error” for this mutation would be zero. However, a mutation in either of the other two position does result in coding for a different amino acid. So, for instance, if GGG were mutated to GAG, the codon would code for glutamate with a PR value of 13.6. The error for this mutation is the absolute value of the difference between the intended amino acid and the one actually coded for (in this case 4.6). For each random code, every possible point mutation is scanned over and the errors are summed to give a measure of the error-tolerance “optimality” for the codon assignment scheme being considered. When Freeland and Hurst claim the canonical code is “one in a million” they mean that it has a higher degree of error tolerance than one million other random codon assignment schemes *when PR is used as the metric*. To arrive at this finding, one additional piece of information needs to be included, and that is the relative likelihood of a mutation that replaces a purine for a purine or pyrimidine with a pyrimidine (a transition error) versus a mutation that changes a pyrimidine to a purine or vice versa (a transversion). Freeland and Hurst note that it has been observed that there is a distinct bias towards transition mutations versus transversions, and reflect this bias as a weighting that weights more heavily error resulting from a transition bias over the error resulting from a transversion. Code optimality, as we have just described, can be expressed mathematically by equation 2.3, where O_i is the optimality value for a randomized code indexed as i , c is the intended codon, c' is the mutated codon, $GC^i(x)$ is the PR value (or other quantifiable amino acid property) associated with the codon table’s assignment for the codon x , and $d^q(a, b)$ is the error between the two return values from the code, $d^q(a, b) = |a - b|^q$. $\langle c, c' \rangle \neq Ter$ indicates that the sum is taken to exclude all point mutations to or from a stop codon. $W_{c,c'}$ is the transition/transversion weighting. Values for $W_{c,c'}$ were taken from [13] and are given in Table 2.2.

$$O_i^{-1} = \sum_{\langle c, c' \rangle \neq Ter} W_{c,c'} d^q [GC^i(c), GC^i(c')] \quad (2.3)$$

Table 2.2: $W_{c,c'}$ transition and transversion values for genetic code optimality calculations

	First position	Second position	Third position
Transitions	1.0	0.5	1.0
Transversions	0.5	0.1	1.0

Other quantifiable metrics for amino acids have also been considered, including hydropathy, molecular volume, isoelectric point, etc. PR was consistently found to be one of the most conserved properties (the canonical code was remarkably optimized to error tolerance with respect to changes in coded amino acid PR with point mutation). Hydropathy was also found to exhibit some degree of conservation, which is not surprising, considering that hydropathy and PR are somewhat related: the effect of hydropathy strongly influences the partition coefficient for transfer between an aqueous-organic interface, which is what PR quantifies.

In collaboration with Nigel Goldenfeld and Tom Butler [4], we have further explored the relationship between PR and error-tolerance optimality of the genetic code. In this study, we examined the genetic

code's optimality following a similar methodology to that used in earlier work [12; 13] using our computed PR values and including an analytical form of bootstrap resampling. To measure the significance of the canonical code's optimality, the probability, P_b , of a random code having a greater optimality than that of the canonical code is calculated by dividing the number of codes with optimality greater than the canonical by the total number of random codes studied ($P_b = N_{O>O_1}/N_{total}$).

An analytical form of bootstrap resampling was used to compute the error in P_b . We are interested in determining the error in estimating the number of codes which are more optimal than the canonical. We consider drawing at random a code more optimal than the canonical as a step in one direction in a one-dimensional random walk. Since the asymmetric one-dimensional random walk is a well known problem in statistical physics, we can easily obtain an error estimate from equation 2.4 assuming $P_b \ll 1$.

$$var [P_b] = var \left[\frac{N_{O>O_1}}{N_{total}} \right] = \frac{P_b(1 - P_b)}{N_{total}} \approx \frac{N_{O>O_1}}{N_{total}^2} \quad (2.4)$$

When $q = 2$ in equation 2.3 $P_b = (19 \pm 4.36) \times 10^{-8}$ when our computationally derived PR values are used. The value obtained using the original experimental PR values, $P_b = (26.5 \pm 1.63) \times 10^{-7}$, was significantly lower than the value from our computationally-derived PR. This difference in optimality scores is partly due to the significantly different value for tyrosine in the new PR set. When the tyrosine value in the original, experimental PR set was replaced with that in the computational PR, the optimality score drops to $P_b = (9.3 \pm 1.0) \times 10^{-7}$. This new value approaches that of the P_b value derived from the complete experimental PR set, however it still presents a higher level of optimality. The remaining difference is attributable to the nature of the computational methodology used in obtaining the new PR set. In the computational method to obtain PR, errors that would manifest themselves in experiments are no longer present: DMP-water ratios are controlled precisely, purity of the materials used and of the paper substrate are not a factor, etc.

We have also explored different q values in equation 2.3. Finding a stronger P_b from using a smaller value of q in the distance metric would suggest that the canonical code evolved in such a way as to be more forgiving of smaller and/or more frequent errors. Conversely, should a stronger P_b be found for a larger value of q , the code may have evolved in a manner that more strictly suppressed larger and perhaps more rare errors arising from point mutations in codons. Finding the q value resulting in the lowest P_b , therefore, tells us something about the process in which the genetic code evolved. Were evolutionary pressures on the code coming from large or small PR errors, or was there little difference in the frequency of occurrence between small or large errors? In calculating P_b values with q ranging from 0.25 to 5, we have observed that the lowest P_b results from $1.0 \leq q \leq 2.0$. P_b values increased sharply outside of this range, suggesting that evolutionary pressures at the time of the genetic code's origin did not push the code to evolve more strict control over large or small PR deviations resulting from point mutation. Having no bias against large PR errors resulting from point mutations suggests that evolutionary pressure must have acted on the early translation mechanism to discard or correct overly flawed protein sequences resulting from point mutations. The mechanism through which the genetic code evolved was apparently not capable of more strongly enforcing against large errors in PR due to point mutation. The code is, of course, optimized against large deviations

in PR due to mutation in codons, however, there is no evidence of strong forcing against these mutations beyond what would be derived from a simple linear to quadratic difference metric.

Not all living organisms use precisely the canonical genetic code, there are a few minor deviations. We have run our calculations on a few of these variants, and presented the P_b results for them in table 2.3. The canonical code remains within a standard deviation and a half of the most optimized code, yeast mitochondrial, and therefore the difference is not taken to be significant. The difference between the canonical and the CDH nuclear code is, likewise, insignificant being well within a single standard deviation. The other two codes studied are significantly less optimized with respect to PR error tolerance than the canonical. From this, we can conclude that the evolutionary pressures that forged the canonical code were no longer relevant when the newer variants were derived. It is likely that at the time that the code variants were created, a robust translational machinery was in place with a high degree of error checking mechanisms built in. It is possible that the evolutionary pressures present at that time were then perhaps lightly forcing certain organisms to adopt their slight variations in order to prevent other organisms or viruses from inserting foreign genetic material. The variations in these codes are likely adaptations from an original (most likely the canonical) code, since they are all exceptionally similar, like dialects of the same language rather than new languages.

Table 2.3: Comparison of P_b values for alternate genetic codes observed in living organisms today

Codon table	P_b
Canonical	$(19 \pm 4.36) \times 10^{-8}$
Yeast mitochondrial	$(11 \pm 3.32) \times 10^{-8}$
CDH nuclear code	$(21 \pm 4.58) \times 10^{-8}$
Ascidian mitochondrial	$(583 \pm 24.15) \times 10^{-8}$
Echinoderm mitochondrial	$(51 \pm 7.14) \times 10^{-8}$

We have also compared the results from our calculations using the computational PR with another amino acid metric, the Grantham polarity. The Grantham polarity has also been suggested to be an important amino acid property with respect to the genetic code [21]. The Grantham polarity was established to be a single quantitative value that could be uniquely assigned to each amino acid and that captured many of each amino acid's properties. These properties included chemical composition, polarity, and molecular volume. An earlier study involving genetic code optimality had found the Grantham polarity to be the property most optimized by the genetic code [22]. Our results using the Grantham polarity yield $P_b = (285 \pm 16.88) \times 10^{-8}$, which is an order of magnitude higher than the P_b derived from using the computational PR. We can thereby conclude that the computational PR reported by us [5] is the strongest metric known to date by which the genetic code has been optimized.

2.6 Interactions Between Amino Acids and Nucleo-bases in Solution

Given the remarkably high level of optimization observed in the canonical genetic code with respect to PR, what can we conclude about the origins of the genetic code? Since PR is related to an amino acid's ability to exploit an interface, it is reasonable to conclude that the original assignments were created via a mechanism

that also presented an interface with the aqueous environment. Two possibilities present themselves for a mechanism linking specific amino acids to nucleotides, given the correlation between PR and the genetic code.

One possible scenario involves having the entire process take place purely in solution, in which a concentrated mixture of nucleotides and water exists and amino acids may partition across the nucleotide/water interface depending both on the nucleotides present and the PR of the amino acid. While the primordial soup was likely very dilute with respect to both nucleotides and amino acids, recent work has shown that convection currents driven by thermal gradients in narrow, pore-like structures can lead to accumulation of mononucleotides to concentrations of over 10^8 over that of the outside environment [23]. The remaining sections of this chapter will focus on this scenario.

The other scenario builds upon the environment suggested by the PR experiments themselves: a substrate of some sort, perhaps a solid two-dimensional mineral surface and perhaps coated with organic molecules of some sort, presents an interface with the aqueous environment and the amino acid PR exploits this interface in a way such that certain amino acids are more likely to be found spatially separated from others. One possibility that connects the world of short nucleotide sequences with amino acids in such a way that the genetic code may have arisen will be discussed in the next chapter, wherein a mineral surface is saturated with nucleotides and amino acids are introduced into the solution.

2.7 Arginine-Adenine Associations

From Figure 2.4 it is clear that the DMP molecules in solution are generally disordered and exhibit little oriented interactions with the amino acids. In contrast, results from our MD simulations of amino acid solutions with the four standard nucleobases indicate persistent and directed intermolecular interactions between the nucleobases and amino acids. Solutions of adenine show a high degree of base stacking which leads to strong stereochemical interactions with some of the amino acids. Arginine, in particular, exhibited a strong interaction with nucleobases. The first peak in the angular correlation function (Fig. 2.13) corresponds to the time-averaged alignment of arginine's side chain to a single adenine molecule. A second Ade molecule stacks in an alternately parallel and anti-parallel fashion to the first (corresponding to the second and third peaks, respectively, in the figure). The orientation of the side chain relative to the adenine electrostatic dipole displays a much more pronounced orientational correlation to Ade than it does in the DMP simulations. There, the orientation is essentially random with $|\langle \cos \theta \rangle| \leq 0.3$ for the entire range of radial distance observed. Like a standard radial pair distribution function, the ergodic principle applies. Over an equilibrated portion of the trajectory, the average captures the overall behavior of the system. It is interesting to note, in light of the affinity of Arg and Ade in solution, that arginine is found near AMP in the active sites of class II aminoacyl tRNA synthetases, such as AsnRS. The separation distance ($\sim 4\text{\AA}$) between the arginine and the first adenine found in our simulation is typical of those found between AMP and Arg in the active sites of class II aminoacyl-tRNA synthetases [see for Asn, 24]. It is possible that the intrinsic affinity for Arg and Ade may have been "frozen into" the structures of the class II synthetases.

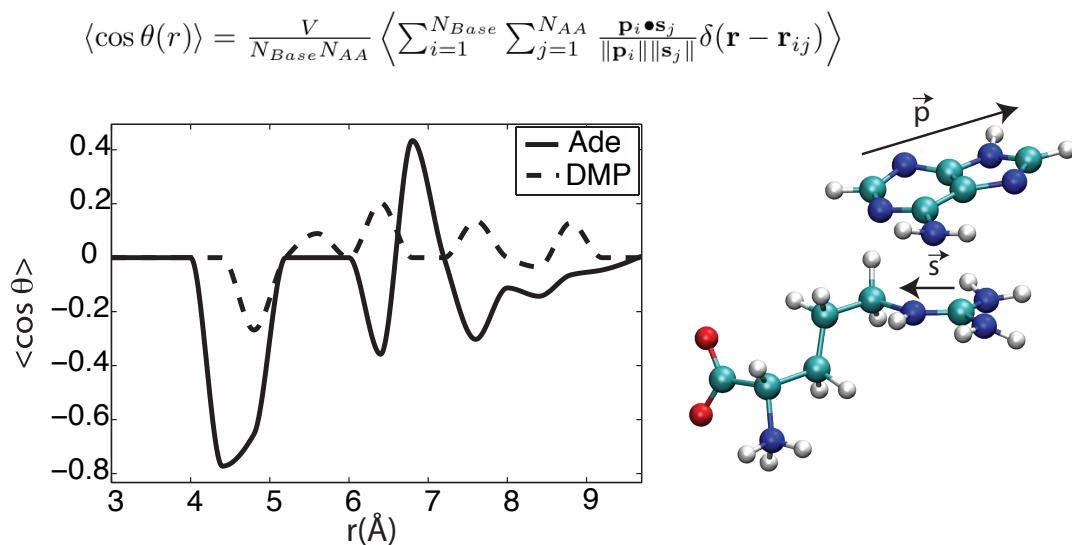


Figure 2.13: **Left** Angular distribution function for arginine relative to the plane of adenine, depicting strong association between the two molecules. Equation for the calculation of the radial angular distribution function displayed above. **Right** Snapshot from MD simulation used to calculate $\langle \cos \theta(r) \rangle$ with arrows indicating the base dipole moment, \mathbf{p} , and Arg sidechain orientation, \mathbf{s} .

2.8 Arginine Associations with Guanine

In addition to the specific interaction between arginine and adenine, arginine is also observed to have an persistent interaction with the nucleobase guanine in solution. This interaction cannot be observed using the angular correlation analysis, however. Figure 2.14 is a snapshot from the simulation, showing that the guanine aligns with its plane perpendicular to the vector along the arginine sidechain, which leads to an average cosine of zero for the angle between the vectors. The correlation between these two molecules can be captured, however, by a standard radial distribution, presented in figure 2.15.

Guanine is the middle position base in arginine's codon, and thus it is interesting that we observe such strong interaction between the two in solution. The middle codon position is also highly important with respect to conservation of PR in the genetic code. However, the first position base for arginine's codons is cytosine, with which arginine was not observed to have any significant interaction. Arginine, in fact, has an adenine in only one of its four codons, which is essentially meaningless since the third position for arginine's codon is perfectly degenerate (it can be any of the four bases).

It is difficult to conclude, based on these observations, that a direct association between nucleotides and amino acids may have resulted in the establishment of the genetic code. It is possible that amino acid-specific interactions with nucleotides or short sequences of nucleotides may be particularly weak and

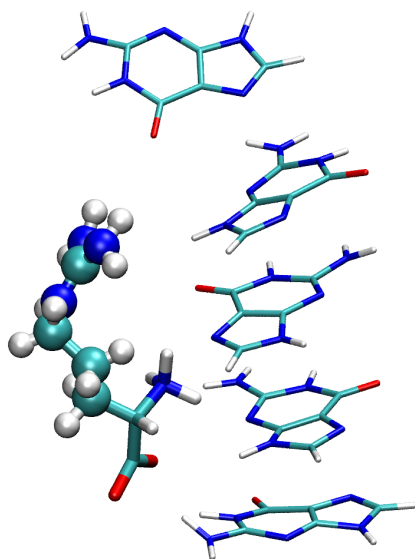


Figure 2.14: Interaction between arginine and guanine in solution.

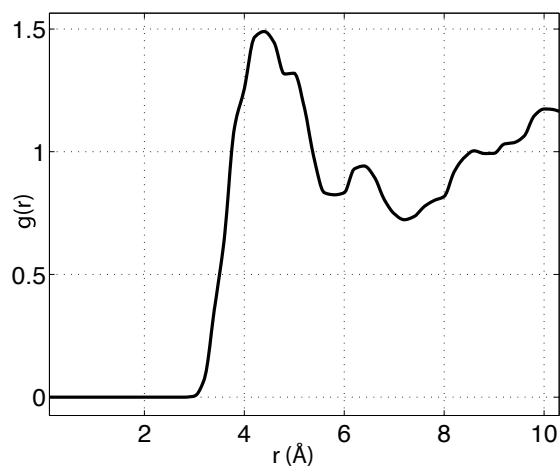


Figure 2.15: Radial distribution function between arginine and guanine sidechain.

difficult to observe in the short times covered by our simulations. A weak association may have still been able to bias an early amino acid selection process that was correlated with the genetic code, however it is more likely that the genetic code's origins come from an as yet unknown third party. Perhaps this third party was a mineral surface that presented an interface to the aqueous environment of its surroundings. Amino acids with low PR would preferentially adhere to the surface or some organic mixture which had already adhered to the surface previously. Nucleotides may have also associated with this surface in ways peculiar

to their own structures. Purines, for instance, may have partitioned to this organic phase more readily than pyrimidines, which might have led to an association of low PR amino acids with purines and established a correlation between amino acids and their anticodons, which would agree with the correlation between amino acid PR and the middle position of an amino acid’s anticodon (see Figure 2.10 and replace the middle position bases with their anticodon counterparts).

2.9 Conclusions

Using statistical mechanical treatments we have investigated the physical basis of the PR values for amino acids. The patterns seen in the radial distribution functions of the amino acid side chains in binary solutions of water and the nucleobase are sensitive to the local solvent environment. The rate of change in the peak RDF values with concentration is strongly correlated ($R^2 = 0.92$) with the experimental PR values obtained from water-DMP solutions (Fig. 2.9) indicating that our proposed microscopic model captures the physical phenomenon behind the PR measurements. It is important to note that PR is a function of the binary solution and will vary with the choice of the nucleobase and its functional groups. Experiments performed using 2,6-diisopropylpyridine, for example, resulted in a wider spread in the PR values of amino acids with U in the middle position of their codons [25], while unsubstituted pyridine results in a compressed range of observed PR values [22].

The only amino acid showing a large discrepancy with experimental values is tyrosine, which can be attributed to a systematic experimental error stemming from direct interaction of tyrosine molecules with cellulose of the chromatogram. In addition, intermolecular interactions between amino acid molecules in experimental conditions could obfuscate the behavior of the isolated, individual molecule. Our computed values represent a property of individual solvated amino acids, and has no such external influences from cellulose amino acid intermolecular interactions. It is interesting to note that our calculated PR value for tyrosine leads to a more highly ordered codon table. Using the calculated PR for tyrosine, there is less of a spread in values for amino acids with an Ade in their middle codon position. By using infinite dilute systems, we cannot include the effects of ionic strength on the degree of phase separation nor can we compensate for possible intermolecular interactions between multiple amino acids. These effects would be present in partition chromatogram experiments. By avoiding these effects, the computed PR values capture an essential amino acid property to a greater extent than possible from experiments.

From our volumetric occupancy probabilities studies, we were only partially successful in capturing the power law dependence of R_m on the water content. While the calculated partition coefficient displayed the linear behavior with concentration without requiring any scaling as in the infinite dilute treatment, the PR values obtained from the calculated slopes were consistently smaller than those observed experimentally. The disparity could be due in part to the fact that our analysis did not take into account the reorganization rates of the two phases of the binary solution and simply used a time average of these components’ volumetric occupancies. A more detailed analysis would also have to take into account the volume of the amino acid clusters in assigning the simultaneous occupancy of a cell for the solute and each of the two solution phases. A better method of distinguishing the “mobile” and “stationary” phases would also need to be employed.

By observing the microscopic environments of the amino acids in binary solution, it is apparent that

the PR is related to how an amino acid partitions across a polar-nonpolar interface. The fact that earlier theoretical work has found a high degree of error tolerance in the genetic code when PR is used as a measure of amino acid similarity suggests that polar-nonpolar interfaces may have played a role in the establishment or development of the early genetic code.

Our work in elucidating the polar requirement invites further inquiry and speculation. While much can still be gleaned through efforts along the lines of those taken by us and our collaborators [4] as well as earlier work by other [13] in exploring the genetic code’s optimality, the answers to many of our questions will lie behind “the Veil through which we might not see” to borrow a phrase from Omar Khayyam. The genetic code was almost certainly established long before life’s Darwinian transition to a traceable phylogenetic hierarchy. The Darwinian transition marks the point in life’s history beyond which we cannot glean a complete understanding; it marks the threshold between what we can hope to know about the evolution of life and the murky world of rampant gene transfer and rapid innovation development and sharing. The ordering of amino acids by polar requirement based on the middle base of their codons, for instance, supports the possibility of a single-base, early version of the genetic code. Perhaps the ancient translational mechanism of the pre-Darwinian stage of evolution manufactured “statistical proteins” whose monomers were selected by a single nucleotide “codon” and needed only to have polar requirement values within a certain range.

An inorganic surface coated by nucleotides would present an ordered aqueous-organic interface that could be utilized for both selecting and orienting amino acids from solution based on PR. Such a system would be helpful in establishing early associations between amino acids and their codons. Investigations into nucleotide interactions with inorganic surfaces have produced interesting observations along these lines. Ferris and coworkers have found that activated nucleotides adhere to and polymerize on montmorillonite clay surfaces [26; 27]. Others have observed large scale accumulation of nucleotides in simulated hydrothermal vent environments [28]. Interestingly, RNA-binding sites for several amino acids are unusually rich in their cognate codon/anticodon triplets, suggesting amino acid/nucleotide associations may have played a role in setting the genetic code [29]. By selecting amino acids based on PR, the early code would evolve to minimize PR differences due to single point mutations in the codons, just as the modern code does.

2.10 Methods

Model of Partition Chromatography The PR experiments employed partition chromatography techniques, whose theoretical interpretation was originally derived from the model of Martin and Synge [30]. The solvent used to develop the chromatogram consists of water and a partially miscible organic phase. In their model, the hydrophilic nature of the cellulose substrate creates a water-rich stationary phase close to the surface of the paper. The resulting water-depleted mobile phase moves on the surface of the stationary phase. The upward velocity of the sample, v_{sample} (Eq. 2.1), is determined by its frequency of transfer between, and lifetime within each of the two phases. From a macroscopic perspective, the mobility is related to the concentration of the solute present in the mobile phase. The mobile phase moves vertically up the support at the speed of the advancing solvent front, $v_{solvent}$ in Eq. 2.1.

For the PR experiments, a binary solution of water-DMP was used to develop the chromatogram. For partially miscible liquids such as these, the two-phase picture no longer strictly applies in the microscopic

scale. In reality, the concentration profile for either of the two species is more likely a continuous function of distance from the cellulose support. In this case, the water concentration is at a maximum adjacent to the cellulose and decays with distance to a value reflecting the miscibility limit of water in DMP. The average location of the solute sample within the liquid layer is determined by its equilibrium properties in the binary mixture. That is to say, the equilibrium solvent environment surrounding a solute sample in the binary mixture will map onto its average position along the water concentration gradient within the liquid film of the chromatogram. If we assume laminar flow of the liquid, we can infer a linear velocity profile in the flow rate. Fluid flow rate in our model is zero directly adjacent to the support, and increases linearly to the rate of the solvent front advance. This maximum velocity occurs in the region where the water is at its miscibility limit in the organic solvent. Studies of the viscosity and thermomechanical properties of water-DMP mixtures in the near-critical regime [31; 32] as well as of the interfacial region of sheared elastic liquids [33] support this approximation. Given this model, the observed relative velocity of a sample is a linear function of its relative distance from the support which in turn is directly related to its equilibrium solvent environment when immersed in a mixture of water and DMP of the same ratios used in the chromatographic experiment. Should an amino acid have a local equilibrium solvent environment that is more water-rich, it will likely occupy a spacial regime closer to the cellulose in the chromatogram and vice versa. Given the assumption of a linear velocity profile, this position is directly related to its relative velocity relative to the solvent front, and therefore directly related to its chromatographic retardation factor. With sufficient sampling, it could be shown that a linear trend in local water concentration surrounding an amino acid with respect to water:DMP molar ratios would correspond to the observed power-law trends in retardation as a function of mole fraction water.

To investigate the solvent environment as a function of distance from the support, we have simulated mixtures of water-DMP in contact with an aqueous monolayer representing a generic hydrophilic surface such as cellulose. Results indicate that rather than two distinct phases (organic-aqueous, or even mobile-stationary) there is in fact a continuous blending from a water-rich regime into an organic-rich one, and that the water concentration decays in the regime close to this surface at a rate inversely proportional to the bulk concentration (data not shown). For a given water-DMP molar ratio, the local density observed around the amino acid will correspond to a region on this exponential curve. If this distance has a power law dependence on the bulk mole fraction of water, the microscopic laminar flow model predicts the power law trends leading to the polar requirement.

MD Simulations The molecular dynamics software package, NAMD2, was used with an NPT ensemble [34] and the Charmm 27 forcefield [35; 36]. A pressure of 1 atmosphere and temperature of 300K were maintained for each simulation. Periodic boundary conditions were enforced, with PME full electrostatics. Simulations for RDF analysis were run for 4ns for Ala, Asp, Gly, His, Met, Pro, Tyr, Trp, and Val and the results compared from the first and last 2ns intervals. For amino acids with medium and high experimental PR, it was determined that the statistics obtained for 2ns was sufficient to obtain the same behavior observed over 4ns of simulation. Longer simulation times were needed for some amino acids with lower PR to collect sufficient statistical sampling for analysis due to limited water contacts. For Ala and Gly, the longer simulation times were needed to separate the weak signal of the side chain or α -hydrogen radial water

distributions from the backbone environment. Asp was run for 4ns to verify that longer simulation times did not effect results for polar and charged amino acids.

VMD was used to create the systems for simulation and for analysis after the MD simulations were complete [37]. The systems simulated consisted of a single amino acid in the center of a $6.4 \times 10^4 \text{ \AA}^3$ cubic box containing water and DMP molecules. For each amino acid five different water to DMP mole fractions were simulated, spanning a concentration range of 50-80 mole percent water.

Radial Distribution Function As a Measure of Local Solvent Environment A radial distribution function, $g(r)$, (Eq. 2.5) of solvent molecules relative to the amino acid is a measure of the solvent density as a function of radial distance from the amino acid [16; 38]. The intensity of the first RDF peak is therefore proportional to the local water density surrounding the reference atom. As described above, the local solvent density from the RDF can be used to infer the average location of the amino acid within a water-DMP film in contact with the cellulose support for a given water:DMP molar ratio. We calculate $g(r)$ from our MD trajectories using the script written by Axel Kohlmeyer and included in the VMD graphical analysis software package.

$$g(r) = \frac{V}{N_{sel}N_{ref}} \left\langle \sum_{j,ref} \sum_{i,sel} \delta(\mathbf{r} - \mathbf{r}_{ij}) \right\rangle \quad (2.5)$$

To probe the differences in local solvent environments around each amino acid, we need to limit the influence of the backbone groups as these are similar for all amino acids. To do this we used the most distant atoms of the amino acid side chain as reference atoms with water oxygen or hydrogen atoms (as appropriate) for the selections and evaluate the maximum $g(r)$ value instead of integrated values. For glycine, which has no side-chain, we chose the α -hydrogen atoms as reference atoms for the RDFs. Special care was taken for amino acids with multiple, non-equivalent hydrogen-bond centers. For Asn, Gln, and His the multiple centers for water hydrogen bonding were used as reference atoms individually for RDFs and the average peak height determined from each reference was used.

Averages were made over the entire trajectories, excluding the first 400 ps for equilibration. The change in maximum values of the peak RDFs from our analysis were used to extrapolate a calculated polar requirement. The peak values of each RDF were plotted as a function of water mole fraction for each amino acid, and the slopes of the resulting lines used to calculate the amino acid polar requirements.

Probabilistic Mapping Depicts Amino Acid Partitioning Martin and Syngue obtained an expression relating the ratio of the solute concentration in the stationary phase to that in the mobile phase, α , to the R_m factor measured in partition chromatography (Eq. 2.6, where A_S and A_M are the cross section areas of the stationary and mobile phases respectively). It was shown that the partition coefficient derived in this manner agrees well with the bulk partition coefficient that has been determined from extraction experiments [39; 40].

$$R_m = \alpha (A_S/A_M) \quad (2.6)$$

The above expression indicates that the experimentally observed R_m in the PR experiments is a measure of the partitioning of amino acids between the aqueous and organic phases of a binary solution. Using

probabilistic mapping, we attempt to relate the macroscopically observed chromatographic results with the partitioning of the individual amino acids in solution.

Volumetric probability maps have been employed by others for studies of gas migration pathways in myoglobin [41] and for mapping ion densities in and near membrane pores [42]. MD simulations for our studies involved 32 amino acid molecules rather than the infinite dilute simulations used for RDF analysis. MD simulation parameters for these systems were the same as for those used for our RDF analysis, except that all were run for 4ns. The simulated box was divided into $\sim 1\text{\AA}^3$ cells, and the probability of occupancy within each cell was calculated over the length of the equilibrated MD trajectory for every atom of the amino acids, DMP, and water molecules. Of the three components of a water-amino acid-DMP mixture, the fastest reorganization times are expected to be in the amino acid and water components. For our study, we ignore drift and currents of the two phases by averaging over time. We attempt to recover an approximate power law dependence of R_m on water content as observed experimentally. The partition coefficient can be approximated from the probability overlap of the amino acid and water or base summed over all cells. Equation 2.7 was used to determine a bulk partition coefficient. Here, the first term provides an overall normalization (total occupancies for DMP and water), and the second term gives the probability overlap of the amino acid with the two phases. P_i^{AA} , P_i^{DMP} , and P_i^{Water} are the average occupancy probabilities of the i^{th} cell for atoms of amino acid, DMP, and water molecules respectively.

$$\alpha = \left(\frac{P^{DMP}}{P^{Water}} \right) \sum_{i=1}^{N_{cells}} \left(\frac{P_i^{AA}(x, y, z) \cdot P_i^{Water}(x, y, z)}{P_i^{AA}(x, y, z) \cdot P_i^{DMP}(x, y, z)} \right) \quad (2.7)$$

Analysis of Amino Acid/Nucleobase Interactions in Solution The electrostatic dipole of the base, \mathbf{p} , was used to describe the base orientation. A vector along the amino acid side chain, \mathbf{s} , was used to describe its orientation. The average of the dot products of these vectors was collected as a function of radial separation between the amino acid and base. Dividing out the magnitudes of the two vectors gives the average of the cosine of the relative angle between the two molecular species (Eq. 2.8). A perfect orientational correlation over time would correspond to an average value of ± 1 . The sign of the correlation indicates parallel (positive) or anti-parallel (negative) orientations between the two vectors. No orientational correlation, or a perpendicular orientation, would correspond to an average value of zero.

$$\langle \cos \theta(r) \rangle = \frac{V}{N_{Base} N_{AA}} \left\langle \sum_{i=1}^{N_{Base}} \sum_{j=1}^{N_{AA}} \frac{\mathbf{p}_i \cdot \mathbf{s}_j}{\|\mathbf{p}_i\| \|\mathbf{s}_j\|} \delta(\mathbf{r} - \mathbf{r}_{ij}) \right\rangle \quad (2.8)$$

Chapter 3

The Catalytic Role of Mineral Surfaces in the Formation of Oligonucleotides from Activated Mononucleotides

3.1 Background

The “RNA World” view of the origins of life holds that chains of RNA became self-replicating, information-bearing molecules that may have also performed rudimentary catalysis of simple, biologically-relevant reactions. Some suggest “metabolism first” scenarios, where a simple (possibly autocatalytic) chemical reaction cycle was the first innovation on the way to the development of life. In either case, the question remains: How did geochemistry turn into biochemistry? In other-words, given the geophysical/chemical environment of the early earth, what features could have played a potential role in the origins of life? What chemicals played the role that modern protein catalysts play today? The starting point to answer these questions should logically begin with the determination of the circumstances of origins and the building blocks available in the initial phase. Did the archaic reactions leading to life begin in our oceans possibly near hydrothermal vents, on land, or in water droplets surrounding grains in the dense clouds that may have formed the early atmosphere? Astrophysicists and astrochemists have identified a wide range of molecular species in the interstellar medium and clouds surrounding nascent stars. The astronomers are aware of small molecules like H_2 , N_2 , SO_2 , H_2S , HCN , CO , and CS , small molecular radicals, and a host of larger species that give rise to unidentified infrared bands [43]. These larger species are a class of compounds known as polycyclic aromatic nitrogen heterocycles (PANHs), polycyclic aromatic hydrocarbons with one or more nitrogen atoms substituted into their carbon skeleton [43; 44]. Geochemists have identified the clays and minerals forming the hydrothermal vent fields and the gases and fluids discharged from the vents. Contemporary hydrothermal vents, both black and white smokers, are surrounded by highly porous mineral and clay precipitates. Large pores exist within these structures, varying from millimeter to micrometers in size. There is also a steep temperature gradient across the vent from several hundreds of degrees in the chimney, to the surface in contact with the ocean water at $2^\circ C$. Typical clays and minerals found in or near these vents include various

The contents of this chapter are based in part on work previously published as Damien Mathew and Zaida Luthey-Schulten. “Influence of Montmorillonite on Nucleotide Oligomerization Reactions: A Molecular Dynamics Study,” *Ori Life Evol Biosp*, 40:303-317 (2010).

forms of smectite, chlorite, and metal sulfide minerals like pyrite, greigite, and other iron-nickel sulfides (for a review see [45]).

There have also been a number of suggestions about the specific reaction chemistry associated with each origins of life scenario. The discoveries demonstrating the catalytic and genetic roles of RNAs gave rise to the RNA world theory for the origin of life (for a review see [46]). Taken to the extreme, the RNA world would have little or no polypeptides or proteins. Attempts to explain the formation of nucleobases, sugars, sugar-phosphate backbones, and reactions leading to the formation of polynucleotides, have also led to the debate of whether a metabolic cycle akin to an archaic TCA cycle has a higher probability to emerge from a geochemical environment than the formation of information storing, self-replicating catalytic oligomers[47–53]. This debate may be more one of emphasis than of substance. In his theory of surface metabolism, Wächtershäuser (1988) laid out a scenario for the formation of metabolites, amino acids, and nucleotides including their polymerization on pyrite and the transition to a cytoplasmic metabolism following formation of early lipid membranes. Both Cody and Wächtershäuser have gone on to experimentally test several of the steps in their proposed reaction schemes[47; 54]. They show that under extreme pressures and temperatures similar to those found in the hydrothermal vents that trace amounts of amino acids and pyruvic acid are formed.

Leaving for the moment questions about the formation of the building blocks, in this chapter we focus on the potential roles that inorganic clay could have played in catalyzing reactions relevant to the formation of the first functional primitive polymers.

An important step in any origins of life scenario is the buildup of complexity from simple molecules of nucleic acids and amino acids into longer chains. The chains of polynucleotides and polypeptides may have been statistically formed initially, but at a later stage cellular machinery arose such that a genetic link between polypeptides and polynucleotides was established and the polynucleotides acquired their information storing property. It has been suggested by many groups that inorganic surfaces were the sites of the reactions in the statistical phase (For recent reviews, see [46; 55–57]). In particular iron sulfide minerals have been proposed as a possible catalyst for prebiotic reactions, and have been shown to catalyze synthesis of amino acids in solution [47; 48]. Considerable experimental work lead by James Ferris and his coworkers has been done to show that the clay montmorillonite can concentrate and oligomerize nucleotides up to 50 in length, especially when the nucleotides are activated with a 1-Methyladenine (1-Mead) or imidazole group at their 5' end [26; 58–64]. Here, we present the results of molecular dynamics (MD) simulations that shed light on the influence of the clay surface on the polymerization reactions.

Montmorillonite is a phyllosilicate clay, and an erosion product of volcanic ash. It is widely distributed, geographically (see Figure 3.1, and can be found in deep sea sediments including near the “Lost City” hydrothermal vent range, which has been an area of significant study for marine scientists and geochemists interested in origins of life, since hydrothermal vent ranges offer chemical and thermal gradients which could have been exploited by early, biologically relevant reaction cycles and can act as a source for chemical raw materials. It has a structure consisting of two layers of tetrahedral silicon oxide sandwiching a layer of octahedral aluminum oxide forming a sheet. The sheets stack like a deck of cards with $\sim 10\text{\AA}$ spaces separating them in which water and ions are present. With no defects or substitutions, its chemical formula

would be $\text{Al}_4\text{Si}_8\text{O}_{20}(\text{OH})_4$. However, montmorillonite is characterized by defects in the silicon oxide (silicon replaced by aluminum) and aluminum oxide (aluminum replaced by magnesium) layers, leading to a net negative charge of $\sim -10^{-4} \text{Coul} \cdot \text{cm}^{-2}$. The negative charge results in the accumulation of positive counter ions in the “gallery” regions between the layers. When dried, the gallery region is occupied by partially hydrated cations which hold the layers together. When wetted, the gallery regions expand and fill with water leading to a large increase in the interlayer distance of more than 20\AA . Cations in a wetted clay become completely hydrated and are able to diffuse freely within the gallery. The expanded interlayer in montmorillonite and related clays is wide enough to accommodate polymers [65; 66] and polynucleotide chains [67], and interlayer cations are known to mediate the reversible binding of RNA oligonucleotides [68; 69]. In addition, montmorillonite catalyzes the formation of vesicles [70] and RNA bound to the clay can be encapsulated within these vesicles [69].

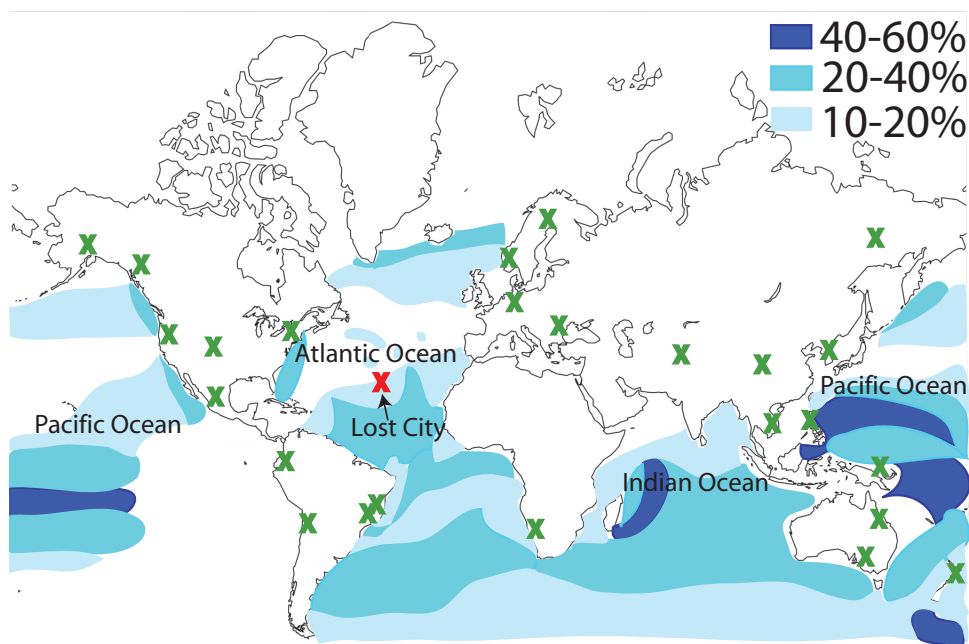


Figure 3.1: Oceanic distributions of the clay mineral montmorillonite measured from deep sea sediments [2] shown as a fraction of the total sediment by color coding. Green X's mark dry-land locations of relatively high purity montmorillonite deposits (taken from the MinDat online mineral data archive: <http://www.mindat.org/min-2821.html>). The red X marks the location of one of the mid-Atlantic hydrothermal vent fields known as “Lost City.”

Montmorillonite’s ability to catalyze nucleotide polymerization of up to 50 mers in length is remarkable considering that only slightly longer synthetic RNA chains such as the Bartel ribozyme have been shown to exhibit RNA polymerase activity [71]. It has also been shown that polymerization reactions on montmorillonite are regioselective, preferring elongation through the 5’ phosphate of the mononucleotide with the 3’ site of the elongating chain’s ribose as opposed to its 2’ site [59]. Experiments have not been able to determine the specific property of montmorillonite that is responsible for the catalytic effect, nor is it known with certainty where in the material the catalysis is taking place: surface, gallery, or step-defect, though evidence has pointed to the interlayer region as being the site for catalysis [72; 73]. The molecular dynamics

simulations presented in this paper provide the microscopic details to address these questions and provide insight into additional observations.

In our simulations, we apply a model of montmorillonite with characteristic charge defects that was developed by Heinz and coworkers [66] and shown to accurately recover spectroscopic and mechanical properties such as compressibility. Using this model we compare two possible scenarios for oligonucleotide synthesis. Simulations are performed with nucleotides in contact with either an exposed surface of montmorillonite or within a confined hydrated interlayer environment to determine the location of catalysis. The confined environment in the interlayer region of the clay is simulated by applying periodic boundary conditions and randomly placing reactants in the aqueous layer between the two images of a single clay sheet.

The step-wise reaction rates for montmorillonite-catalyzed nucleotide polymerization have been experimentally determined, and the dimer to trimer step found to have the fastest rate [74]. While classical MD simulations cannot determine reaction barriers, information about kinetic reaction rates can be inferred from both orientation and encounter distances of the reactants. In our study we compare the differences in interaction of monomeric reactants with each other and perform radial distribution analysis to infer relative reaction rates for $2' - 5'$ and $3' - 5'$ linkages of pairs of monomers. We also analyze interactions with monomers and activated dimers to determine why the formation of a trimer is faster than that of a dimer. Encounter distances are determined from a calculation of the radial distribution function $g(r)$ (see Methods). The steady state reaction rate constant for a bimolecular reaction, k^+ , can be shown to be related to the radial pair distribution function, $g(r)$ (3.1) [75].

$$k^+ = 4\pi \int_0^\infty k^\circ(r)g(r)r^2dr \quad (3.1)$$

where $k^\circ(r)$ is an intrinsic reactivity function related to the probability of a reaction going forward given a reactant separation distance of r . $g(r)$ reflects the likelihood of the reactant pair to be at a distance r relative to that of the bulk density. This formula assumes that both $k^\circ(r)$ and $g(r)$ are isotropic. If a reaction takes place only within a certain encounter radius, R , k^+ can be shown to be directly proportional to the magnitude of a radial distribution function evaluated at R (3.2) where we substitute $k^\circ(r) = k^\circ\delta(r - R)/4\pi r^2$.

$$k^+ = k^\circ g(R) \quad (3.2)$$

Experiments have also shown that there is a sequence dependence to the reaction rates for dimerization. Our simulations verify the observed sequence selectivity and suggest that the variation in rates observed when activated pyrimidines (purines) are added to purines (pyrimidines) are due to differences in adhesion of the two forms of nucleotides to the surface.

Experimentally oligonucleotide synthesis on montmorillonite has been found to prefer homochiral products of either all D- or all L- nucleotides incorporated in the polynucleotide chain when mixtures of D- and L- mononucleotides are present. Our simulations indicate that the homochiral pairs are the dominant configurations which is important for questions of chirality selection in origins of life scenarios.

3.2 Role of Confinement in Catalysis

Montmorillonite as a catalyst for oligonucleotide synthesis is primarily of interest due to the regioselective nature of the activity. While biologically synthesized nucleotide chains are always connected by 3' – 5' linkages, this is not the most favorable product based on reactivity. Without a regioselective catalyst, aqueous reactions resulting in 5' – 5' pyrophosphate and 2' – 5' phosphodiester linkages are more favorable [46]. Here we study the physical feature of montmorillonite responsible for the regioselective catalysis, paying particular attention to orientations of monomers and dimers adhered to the surface and quantifying the favorability of reaction by analysis of $g(r)$ measured between the phosphorus and 2' or 3' oxygens of two separate reactants (see Methods).

Mononucleotides adhere to the montmorillonite surface in our simulations in different ways. Unactivated monomers tend to adhere to the clay surface through interaction of their bases which allows hydrogen bonding with their ribose hydroxy groups (Fig 3.2 left). Monomers activated with 1-Mead adhere through interactions of both the activating group and the exocyclic functional groups of the base, with the backbone facing away from the surface (Fig 3.2 right).

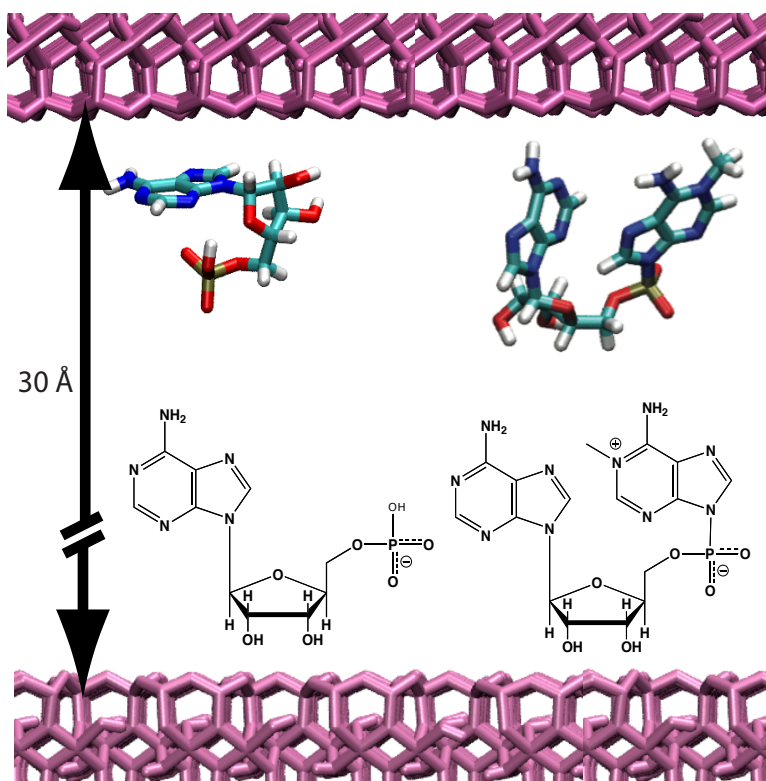


Figure 3.2: **Left** Unactivated AMP adheres to a montmorillonite surface through stacking of the base and interaction with the 2' hydroxy **Right** 1-MeadpA adhesion orientation. Colored molecules represent structures from the simulations. Structures for both molecules are included below the snapshots. Lower clay sheet added to indicate interactions take place in the interlayer, however the separation distance has been compressed in the figure. (Other monomers, ions and water omitted for clarity.)

Dimer adhesion to montmorillonite differs from the adhesion of monomers. When the activating group

1-MeAd is added to the 5' end of the dimer, dinucleotides are observed to anchor to an exposed surface through either the activating group or a base, allowing the molecule to extend into the solvent above the surface (Fig. 3.3 right). In the interlayer region the entire dimer is seen to remain close to the montmorillonite surface with little drift into solvent (Fig. 3.3 left). The environment of the clay interlayer region has a significantly higher ionic strength than that above an exposed surface, which may assist in the adhesion of the nucleotide.

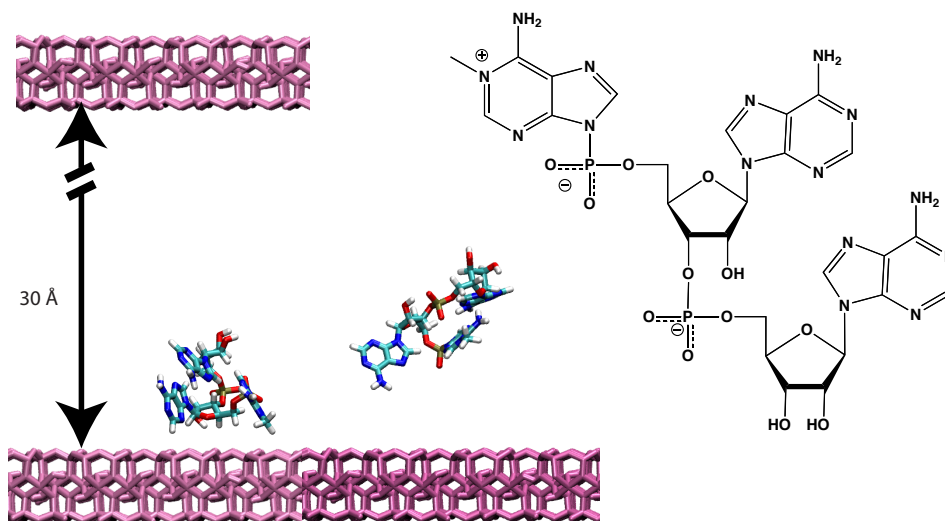


Figure 3.3: Adhesion of activated adenosine dimers (1-MeAdpApA) with (**Left**) and without (**Center**) confinement in the interlayer. **Right** Molecular structure. (Other dimers, ions, and water omitted for clarity.)

Simulations with both an exposed surface and an interlayer environment show that the elongation reactions are more favored in the interlayer regions. Analysis of the trajectories suggests that the relatively higher ionic strength of the solvent as well as the conformational differences in dimer adhesion in the interlayer are responsible for the catalytic effect. Complexes of monomers with activated dimers on an exposed surface show a preference for 5' – 5' linkage as measured by the peak position and magnitude of the $g(r)$ in Fig. 3.4 (left). It is only when the activated dimers are placed in the confinement of an interlayer environment that the 3' – 5' regioselectivity is observed. This observation is in agreement with experiments in which a chemical blocking of the interlayers resulted in a dramatic decrease in the reaction yield [72].

In the interlayer, the $g(r)$ s show a strong preference for orientation of the 5' phosphate of the monomer to the 3' hydroxy terminal group of the activated dimer, which is the linkage predominately observed experimentally. These groups remain in close proximity for several nanoseconds once they have come in contact (See Fig. 3.4 right). According to Eq. 3.2 both k° and g_{max} contribute to the reaction rate. We make a direct comparison of g_{max} to estimate relative reaction rates using Eq. 3.2, assuming that differences in k° are relatively small for the narrow range of distances examined. However, peaks occurring at larger radial distances reflect a smaller reaction rate since the intrinsic reactivity function, k° , is assumed to be a rapidly decreasing function of the separation distance between reactants. Without confinement (Fig. 3.4 left) g_{max} for 5' linkages dominate, and little signal for 2' – 5' or 3' – 5' linkages is observed. In the interlayer g_{max}

for the $5' - 5'$ linkage is only 25 and occurs 0.9\AA farther than the much stronger peak for the $3' - 5'$ linkage at 3.8\AA and a value of 48. The peak for $2' - 5'$ linkage occurs at 9.8\AA and is a result of the strong $3' - 5'$ affinity. Comparing g_{max} for the exposed and interlayer environments suggests that an enhanced catalytic effect occurs in the interlayer, and that reactions on the exposed surface should result almost exclusively in $5' - 5'$ linkages.

Calcium ions play an important role in bringing reactants together in the dimer to trimer elongation reaction. As discussed in the Methods section, we screened the $g(r)$ s for the $2' - 5'$ and $3' - 5'$ calculations for the presence of calcium in close proximity. Interactions between the $3'$ hydroxyl oxygens are nearly always mediated by a calcium ion. Calcium strongly mediated the $3' - 5'$ linkage, but was only rarely observed in the $2' - 5'$ linkage.

The elongation occurs primarily according to two scenarios, both mediated by calcium positioned as Ca_A^{2+} or Ca_B^{2+} respectively in Fig. 3.5a. Once the calcium coordinates with the phosphate, the monomer undergoes two dimensional diffusion along the surface until it comes in close proximity to a dimer. At this point, the calcium simultaneously coordinates with both the AMP's phosphate and $3'$ terminal hydroxy of the dimer, which puts the two groups in close contact for a sufficiently long duration to allow a reaction to take place as shown in the mechanism in Fig. 3.5a. Fig. 3.5b is a snapshot of such a simulation. During the 7 ns animation, the $3'$ terminal hydroxyl remains on average within 3\AA of the surface. Divalent metal ions have been shown to catalyze nucleotide polymerization without the presence of a surface [76], however the catalysis had been shown to form $2' - 5'$ linkages. The cation may modify the pKa of the hydroxy groups to make the reaction more favorable. In our simulations, the ion interacts with the $3'$ hydroxy, which may force the regioselectivity of catalysis to form $3' - 5'$ links as suggested in Fig. 3.4 (Right).

Fig. 3.5c shows an alternative configuration that still brings the reactant groups in close proximity. In this orientation, the calcium ion (Ca_B^{2+} in Fig. 3.5a) coordinates between the $5'$ phosphates on the activated monomer and AMP which positions the AMP close to the terminal hydroxyl group on the dimer as shown in Fig 3.5a. This orientation leads to the $5' - 5'$ linkage as suggested by the $g(r)$ in Fig. 3.4(Right).

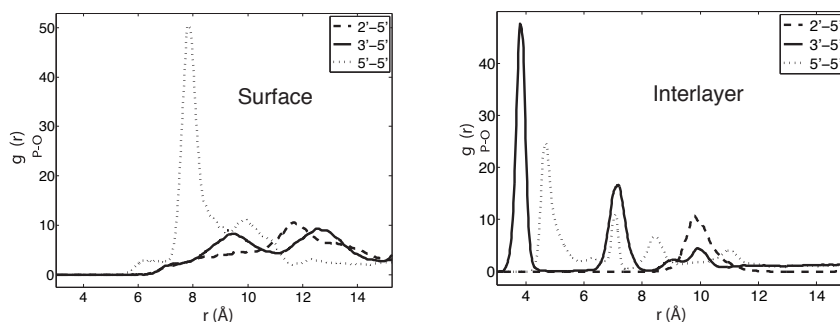


Figure 3.4: Trimer formation from 1-MeadpApA + AMP . Radial distribution functions for **Left** exposed surface and **Right** Interlayer.

Simulations were performed on a system consisting of the same reactants as described above but without the presence of the montmorillonite surface. The $g(r)$ analysis is presented in Figure 3.6. Clearly, the peak intensities for the $3' - 5'$ linkage is significantly lower than observed in the presence of montmoril-

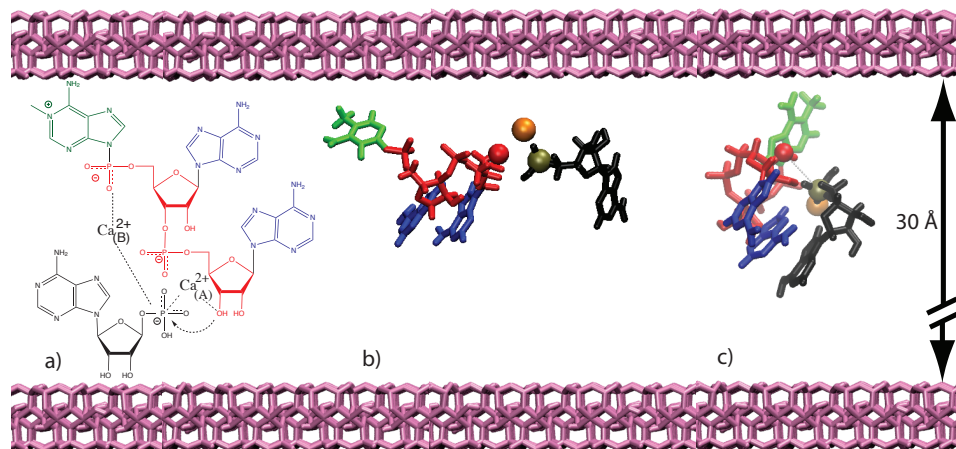


Figure 3.5: 1-MeadpApA + AMP elongation reaction simulated in montmorillonite. **(a)** Two mechanisms for elongation of a dimer to a trimer that depend on positions Ca^{2+} are shown. **(b)** Snapshot from simulation of the dominant mechanism. Calcium ion (orange) ($\text{Ca}^{2+}_{(A)}$) mediates interaction of a terminal 3' oxygen (red sphere) with the phosphate (gray sphere) of the incoming AMP. **(c)** Alternate configuration in which $\text{Ca}^{2+}_{(B)}$ brings the reactants in close proximity through 5' – 5' interactions. Both configurations contribute to the g_{max} in Fig. 3.4(Right). The 3' terminal oxygen and 5' AMP phosphorus are within 3.8 Å in both cases. Water molecules surrounding the system have been hidden for clarity, but are explicitly present in the simulations.

lonite, and a strong preference for 5' – 5' linkages is suggested. Without the surface, 2' – 5' linkages are a possibility as they have a significant peak in these simulations as opposed to those with the surface and in the interlayer. Additionally, calcium ions were rarely involved in assisting the 2' – 5' or 3' – 5' linkages in the solution simulations. These differences in are an indication of the strong catalytic properties of the montmorillonite surface.

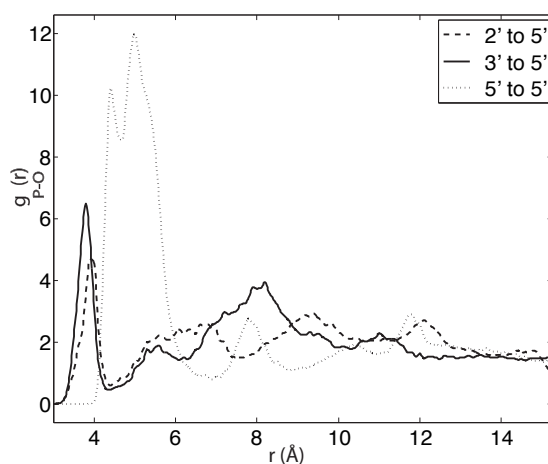


Figure 3.6: Radial distribution functions for trimer formation from 1-MeadpApA + AMP in bulk solution without any surface present.

3.3 Sequence Selectivity in Dimer Formation

In experiments that added nucleotide monophosphates (NMPs) to activated nucleotides, it was reported that the addition of the purine monophosphate (AMP) to activated pyrimidines resulted in the highest yields and also exhibited the strongest regioselectivity for 3' – 5'[63]. Conversely, adding the pyrimidine monophosphate (UMP) to activated purines resulted in the lowest yields, while adding pyrimidine monophosphates to activated pyrimidines actually showed slight preference for 2' – 5' linkages.

Our simulations on mixtures of adenosine and uridine in the clay interlayer confirm the strong sequence selectivity of montmorillonite oligonucleotide catalysis. In one set of simulations the adenosine is activated and uridine is in its UMP form (UMP + 1-MeadpA). A second simulation switches the activation to the uridine, and leaves adenosine in its monophosphate form (AMP + 1-MeadpU). Simulations of both mixtures were carried out for 50 ns, and the pair distribution function $g(r)$ for the approach of the 5' phosphate on the monophosphate to the terminal hydroxyl groups and the 5' phosphate on the activated nucleotide analyzed. The $g(r)$ s in Fig. 3.7 clearly show that the reactants are preferentially arranged in the AMP + 1-MeadpU solution for dimerization utilizing 3' – 5' linkages. The UMP + 1-MeadpA scenario shows minimal 3' – 5' alignment, with the g_{max} significantly less intense than for the AMP + 1-MeadpU scenario (Fig. 3.7), however a strong 5' – 5' signal is observed.

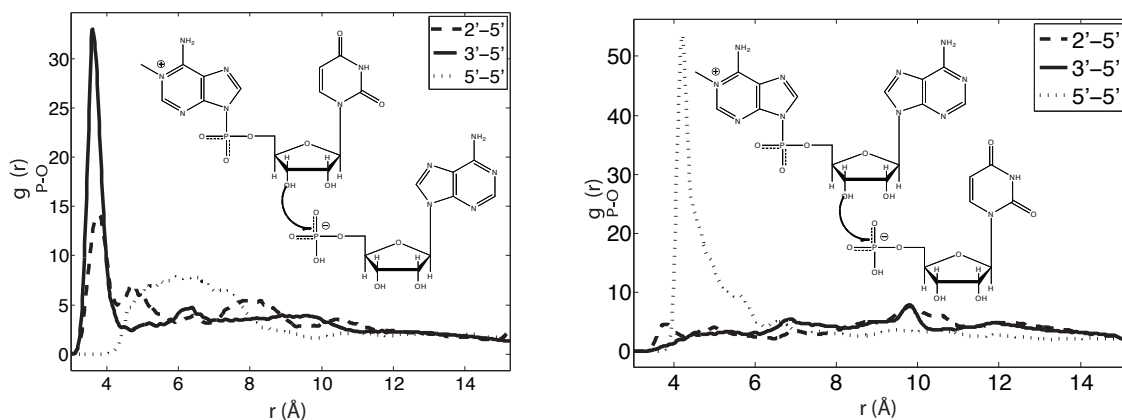


Figure 3.7: Sequence selectivity for U-A and A-U dimer formation. $g(r)$ s for (Left) AMP to 1-MeadpU and (Right) UMP + 1-MeadpA).

Analysis of the trajectories reveal that UMP appears to have a weaker adhesion to the clay surface than AMP which likely prevents the 2D surface diffusion with a bound Ca^{2+} ion observed in the AMP + 1-MeadpU as well as the AMP + 1-MeadpApA simulations of trimer formation discussed earlier. The lack of dimensional constraint to the surface results in interactions between UMP and 1-MeadpA that are not favorable for 3' – 5' linkage dimerization. Namely the base on UMP stacks with the 1-Mead group and base pairs with the adenine which contribute to the second and third peaks in Fig. 3.7. In the case of AMP + 1-MeadpU, 1-MeadpU adheres more strongly to the surface than UMP while AMP is able to diffuse along

the surface and easily attack the 3' oxygen in a similar manner as observed in trimer formation. The lower value of g_{max} for AMP + 1-MeadpU dimer formation relative to the value for trimer formation with only adenosine suggests that this step in elongation would have a lower reaction probability. This is in agreement with experimental studies of the stepwise reaction rates, which have shown the formation of trimers from dimers and monomers to be the fastest step in the montmorillonite-catalyzed formation of oligonucleotides [74].

3.4 Influence of Chirality on Catalytic Efficiency

Chirality is an important consideration when studying the origins of life [77]. Examination of the Murray and Murchison meteorites, two highly regarded extraterrestrial sources of non-biological amino acids, revealed a slight enantiomeric bias towards L-form amino acids [78]. However, this bias is very slight (only 1.0 to 9.2%). To account for the exclusive use of L-amino acids and D-sugars (i.e. ribose in nucleotide backbones) in modern biology, further enrichment is required. This can be accomplished through chiral-selective adhesion to a surface. Though montmorillonite does not present a chiral crystal surface, recent work has shown that adhesion of molecules to an achiral surface can induce chiral recognition of other molecules which preferentially adhere in response to a change in the surface's electronic structure [79]. Here, we present results from MD simulations that demonstrate a selective property for montmorillonite to prefer linkages between monomers of similar chirality. This behavior, coupled with the possibility for a slight deviation from racemic abundance of naturally-occurring nucleotides could account for the homochiral polynucleotides utilized by modern life.

Experimental work has indicated montmorillonite catalysis favors the synthesis of homochiral oligonucleotides [80; 81]. In the presence of activated D- and L- AMP, the ratio of homochiral oligomers (characterized by the sum of all D- and all L- members) to mixed oligos was reported to be 60:40. To gain a deeper understanding of this effect and to determine the contributions of the four different pairings, we modeled an L-1-MeadpA (Fig. 3.8) and L-AMP and performed 4 sets of simulations using various mixtures summarized in Table 3.1 . The simulations consisted of a system containing four D-AMP molecules and four L-1-MeadpA molecules, a system of four L-AMP and four D-1-MeadpA molecules, and two homochiral systems each consisting of four activated and four monophosphate L- and D- form adenosine molecules respectively. For comparison, we also performed simulations without the montmorillonite present. While the experimental findings for homochirality used an imidazolid activating group, we retained the 1-Mead group as used in the other simulations reported in this paper to allow for consistent comparison. Each system contained identical montmorillonite surfaces and interlayer separation used in the systems discussed above. Results show that the L-AMP and L-1-MeadpA adhere to the surface in a manner similar to the D-forms shown in Fig. 3.2.

The maximum in $g(r)$ was used to measure the approach of the 5' phosphate group of the AMP to the 3' hydroxy group of the 1-MeadpA . Based on the comparison of the g_{max} , the results in Table 3.1 show a preference for homochiral association of monomers in orientations conducive for dimerization. To relate the findings in the table to a ratio of homochiral to heterochiral products, we sum the g_{max} values for L+L and D+D. This value is then divided by the sum of the g_{max} for all four combinations (L+D, D+L, L+L,

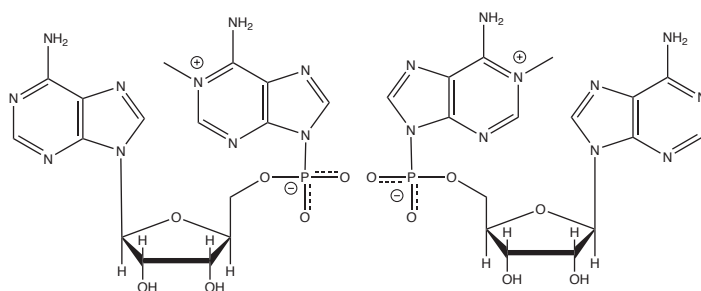


Figure 3.8: L- (**Left**) and D- (**Right**) Stereoisomers of 5' 1-Mead activated Adenosine

and D+D) resulting in a value proportional to the reaction rate for homochiral product as a fraction of the rate for total product. In the presence of the montmorillonite catalyst, we see both stronger peaks in the pair distribution of reactants relative to the solution simulations as well as a 65:35 ratio in the sums of the homochiral versus heterochiral $g(r)$ peaks. For simulations carried out in solution with no montmorillonite present, the homochiral to heterochiral ratio was found to be 43:57

Table 3.1: Peak $g(r)$ values for hetero- and homo- chiral monomer interaction

1-MeadpA Chirality	AMP Chirality	g_{max} catalyzed	g_{max} in solution
L	D	5.3	4.9
D	L	5.5	1.5
L	L	12.2	3.4
D	D	7.7	1.5

3.5 Generalized Forms of Positive Surfaces and Their Potential Role in Polynucleotide Formation

So far in this chapter we have dealt with a specific mineral surface that, through random substitutions of magnesium for aluminum in its aluminum oxide layers, possessed a net negative charge. The potential role of positively charged surfaces, particularly iron-sulfides such as greigite ($\text{Fe}^{2+}\text{Fe}_2^3\text{S}_4$) and pyrite (FeS_2), in the origins of life has also been discussed in the literature [48]. It is also important to note that, while we have observed that surface adhesion to the montmorillonite surface played a role in promoting the elongation of a nucleotide chain, a solvated cation played an equally important role in stabilizing configurations of the reactants that would allow the reaction to go forward. In this section, we address the possibility of a positively-charged surface playing a similar role to that of montmorillonite in promoting nucleotide oligomerization.

Greigite is an especially compelling mineral, since it is associated with modern biological systems. Magnetotactic bacteria, for instance, have greigite crystals in their magnetosomes and greigite is also formed by sulphate reducing bacteria. So, greigite is a biologically relevant mineral in life today, and while greigite

is not readily used in essential metabolic cycles in life nor is it ubiquitous in life, its presence in life today suggests that it may have been utilized for other purposes in early life. Additionally, greigite is associated with clay sediments, and is thus often found in deposits containing montmorillonite, and can also be found in deep-ocean hydrothermal vent stacks which have been proposed as potential sites for the origins of life [82]. Greigite's positive charge is derived from substitutions of Ni^{3+} or Fe^{3+} for Fe^{2+} , and results in a nominal charge of 2.5 per unit cell [83].

Generally, the proposed scenarios involving iron sulfides focus on a “metabolism first” view of the origins of life, in which reaction cycles forming biologically relevant molecules are catalyzed or otherwise assisted by greigite [84].

Here, we present results from a generalized positive surface modeled, for the purposes of our study, by an array of sodium atoms, constrained to two-dimensional surface. We used MD simulations to investigate the microscopic interactions between this surface and uracil nucleotides in aqueous solution with free calcium ions as counter ions. The nucleotides are observed to adhere directly to the surface through electrostatic interactions between the phosphate groups and the sodium ions.

In addition, we have simulated a mixture of amino acids above this nucleotide-impregnated surface to determine if the surface promotes the formation of oligopeptides. This mixture consists of the glutamic acid, leucine, 0.7M NaCl, and water. We find that the system can selectively promote a 5' to 3' linkage between nucleotides, which is an essential structure for biological activity. We also find a possible promotion of polymerization of amino acids from this system. We observe a slight preference for Glu pairing over Leu pairing, which would be consistent with association of Glu with its anticodon, which has a U in its middle (most critical) position.

3.6 Promotion of Nucleotide Oligomerization and Associations with Amino Acids

3.6.1 Selectivity in Oligonucleotide Formation

We observed that the nucleotides placed above the surface adhered to the sodium ions by electrostatic interaction with the negatively charged oxygens on the 5' terminal phosphate group. This behavior fixes the positions of the 5' end of each nucleotide, but allows for pivoting of the molecule about this point. For this reason, we do not expect to see 5'-5' linkages being made. Biological RNA polymers are always linked from the 5' phosphate group to the 3' group of the next member of the chain, however, there is no reason for this specific linkage to exclusively occur in solution. We examine the radial distribution functions of the 5' terminal oxygens on the phosphate group to the 2' oxygens and 3' oxygens separately to determine if there is a preferential association of these groups which would lead to promotion of one bond formation over the other. There is no need to examine 5' to 5' radial distributions, as these groups are fixed to lattice positions by electrostatics.

Figure 3.9 depicts the radial distribution of 2' oxygens relative to the 5' phosphate oxygens of the nucleotides. We see that there is a peak value of $g(r) = 5.8$ which occurs at 4 Å.

Figure 3.10 gives the radial distribution function of 3' oxygens relative to the 5' phosphate oxygens of the nucleotides. In this case we see a larger peak value, $g(r) = 6.8$ at the same radial distance of 4 Å. The

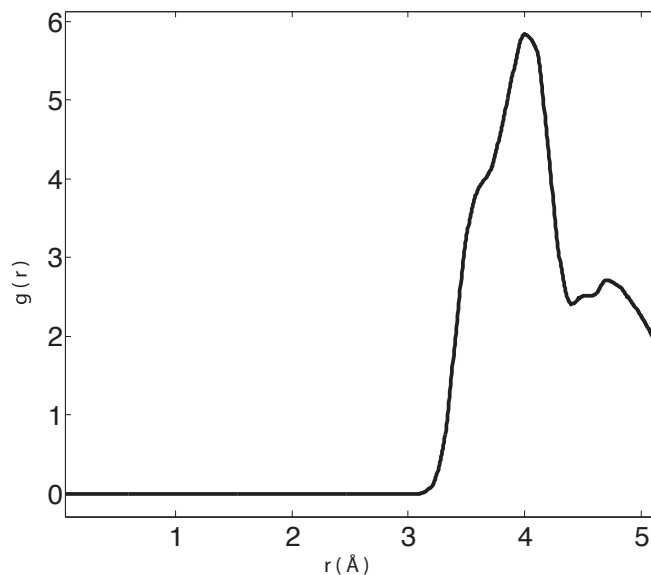


Figure 3.9: Radial distribution function of 5' phosphate oxygens to the 2' oxygens

larger $g(r)$ for 5' to 3' implies that this association is stronger than for 5' to 2', and that should a reaction occur, the 5' to 3' linkage would be preferred over 5' to 2'. We should note that the distance of 4 Å may be too large for a reaction to take place. However, while we terminated our nucleotides with the standard 5' terminal phosphate and 3' terminal hydroxy groups, experimentally a 1-methyladenine 5' terminal group has been shown to promote oligomerization on montmorillonite [59]. It is possible that this larger terminal group may help reduce the separation distance and promote the reaction.

3.6.2 Promotion and Selectivity of Oligopeptide Formation

We will now focus on the promotion of oligopeptide formation by the nucleotide coated surface. We chose a mixture of leu and glu as a test of specificity. Leu, with U in its middle codon position and with a polar requirement of 4.9 is taken to be much different than glu with an A in its middle codon position and a polar requirement of 12.5. It was believed that if an association between amino acids and their codons exist, that a Uracil impregnated surface should preferentially promote leu oligomerization. Figure 3.11 gives the radial distribution function of leu carboxy oxygens to amine hydrogens. We see that the peak is located at ~ 1.8 Å with a value of $g(r) = 4.1$. This short distance is sufficient to allow for a condensation reaction to at least dimerize leu.

To determine if Leu oligomerization is preferred, we examined the radial distribution function of Glu backbone carboxy oxygens to Glu amine hydrogens. The results of this calculation are presented in Figure 3.12. We again see a peak located at ~ 1.8 Å, but with a peak value of $g(r) = 5.5$, larger than the $g(r)$ peak for leu-leu dimerization.

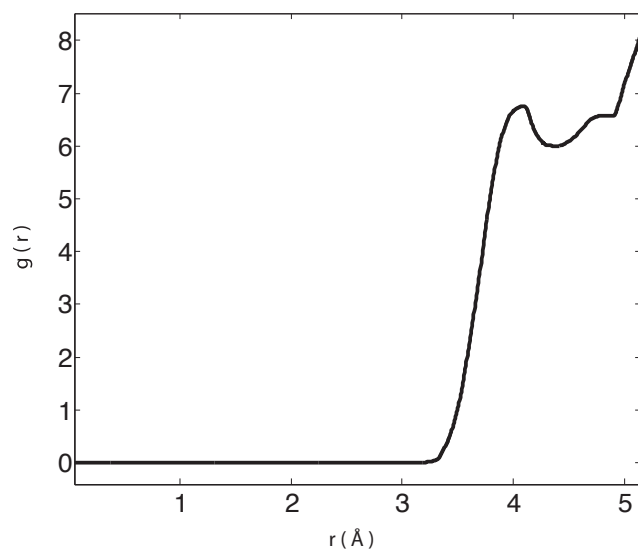


Figure 3.10: Radial distribution function of 5' phosphate oxygens to the 3' oxygens

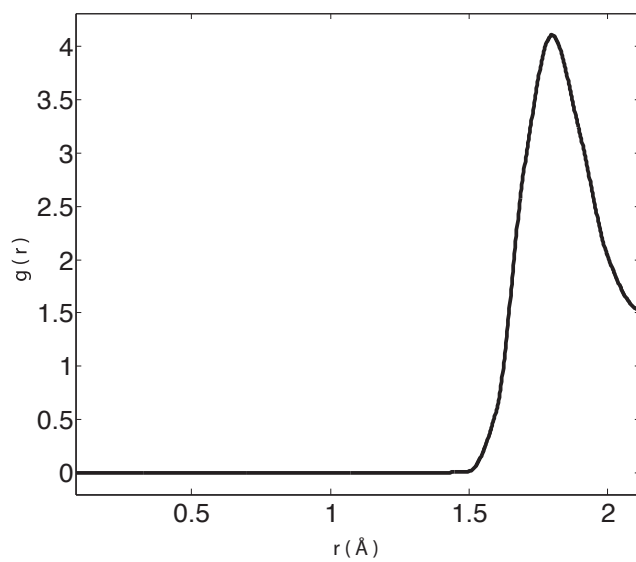


Figure 3.11: Radial distribution function of backbone amine hydrogens to carboxy oxygens of leu

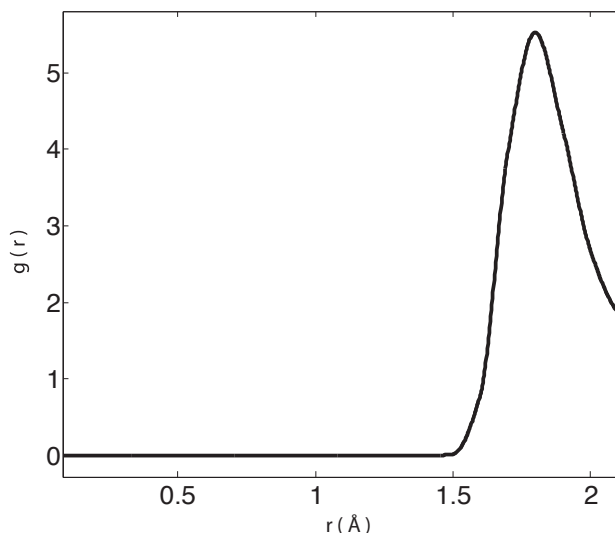


Figure 3.12: Radial distribution function of backbone amine hydrogens to backbone carboxy oxygens of glu

To determine the rate of Glu dimerization with Leu, we calculated the radial distribution function for Glu backbone carboxy oxygens relative to leu amine hydrogens. The results are presented in Figure 3.13. Once again, we see a peak located at $\sim 1.8\text{\AA}$, this time with a peak value of $g(r) = 5.7$.

3.7 Conclusions

We have used classical molecular dynamics techniques to study the interactions of nucleotides with montmorillonite. The simulated environment is similar to the experimental conditions used to study the catalytic effect of montmorillonite on nucleotide oligomerization in that we consider interactions of both activated and phosphate-terminated nucleotides at the exposed surface and within the clay interlayer along with the catalytic effects of divalent ions (Ca^{2+}). We have demonstrated the ability of the Heinz montmorillonite model to recover the trends in product propensities observed experimentally. Although the equilibrium simulations are only 50 ns long, they give insight into the reaction mechanisms. Our simulations confirm that oligonucleotide synthesis takes place primarily in the clay's interlayer region, and that the synthesis should proceed in the $3' - 5'$ direction, just as in template-directed synthesis in the RNA polymerase. Trimer formation on an exposed surface gives the non-biological $5' - 5'$ linkage over the biological $3' - 5'$ as shown by a comparison of the g_{max} values in Fig. 3.4. Simulations with the Heinz model also reveal the preference for forming $3' - 5'$ linkages as well as the sequence selectivity reported by experiments, showing that a preference for catalyzing the AMP + 1-MeadpU over the UMP + 1-MeadpA reaction is due to a relatively weak adhesion of UMP to the clay.

Analysis and comparison of the peak radial distribution values, g_{max} , for both $3' - 5'$ and $2' - 5'$ ap-

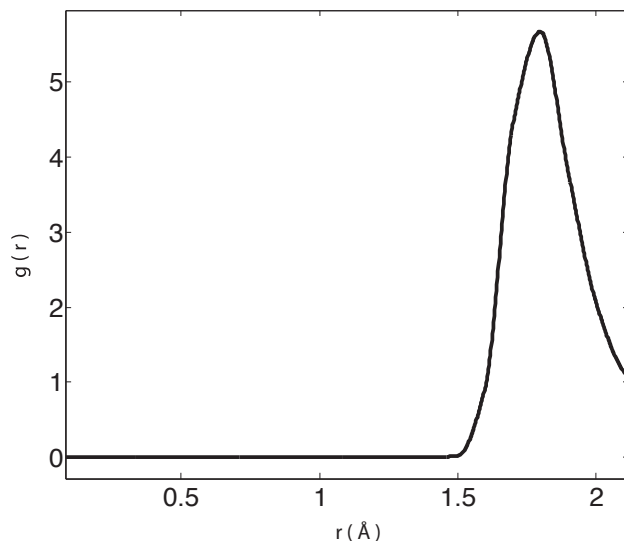


Figure 3.13: Radial distribution function of backbone amine hydrogens of Glu to carboxy oxygens of Leu

proaches of reactants in both the interlayer and exposed surface environments of montmorillonite, suggests both an increased regioselectivity for $3' - 5'$ linkage formation as well as an overall increased catalytic effect when the reactions take place in the interlayer. In addition to g_{max} the intrinsic rate constant k^o is also a factor in determining the overall rate constant for a reaction. In our interpretation, we assume that k^o is similar for $5'$ phosphate attack on both the $2'$ and $3'$ hydroxy. We have also not taken into account that k^o is also a function of separation distance between and relative orientation of the reactants. However, it is interesting to note that in the interlayer, the g_{max} for the $2' - 5'$ is smaller and occurs at a larger separation distance than for the $3' - 5'$ linkage, which suggests that the rate for $2' - 5'$ linkage formation may be less. Without knowing the exact distance dependence of k^o it is difficult to quantify how significant of an effect this would have on the reaction rate constant, though it suggests that the rate constant for the $2' - 5'$ linkage may be lower than what is inferred from a simple comparison of g_{max} . The reason for the increased catalytic effect in the interlayer is likely due to the relatively higher ionic strength in the interlayer which changes the orientation of adhesion of the reactants and also facilitates a constrained diffusion of NMPs coordinated with divalent ions along a surface, thereby improving the likelihood of reactant encounters.

The influence of reactant chirality on the effectiveness of montmorillonite's catalysis in our simulations is in agreement with the results obtained in the experiments. A preference for catalysis of homochiral products is observed, however our simulations indicate a preference for L- over D- homochiral products in dimer formation.

The analysis method is not without some limitations. A radial distribution function neglects a consideration of the relative orientation of the reactants, which may be of some importance for a nucleophilic

reaction. Additionally, the analysis yields only information about the kinetics of a reaction. To study thermodynamic barriers a method such as quantum mechanical / molecular mechanical (QM/MM) simulations or free energy perturbations should be employed. The degree of agreement between the simulations and the experiments suggest that classical MD techniques could be used to also investigate the link between polynucleotide formation and the establishment of the genetic code through associations between amino acids and nucleotide chains on montmorillonite surfaces.

3.8 Methods

Molecular Dynamics Simulations The molecular dynamics software package, NAMD2, was used with an NPT ensemble [34] and the CHARMM27 forcefield [35; 36]. A pressure of 1 atmosphere and temperature of 300K were maintained for each simulation. Periodic boundary conditions (PBC) were enforced, with PME full electrostatics. Harmonic constraints of $2 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^2$ were applied to a silicon atom at each of the four corners of the montmorillonite sheet to maintain orientation. To ensure adequate wetting of the clay surface two minimization and equilibration cycles were performed prior to a longer simulation for analysis. Water was added to the system and minimization was carried out for 5000 iterations, followed by a 1 ns equilibration. Water was again added to the system to fill voids formed during the wetting process and the same minimization and equilibration procedure was followed. The process of adding water in stages results in a swelling of the clay interlayer region to an equilibrated separation distance of 30\AA , except for those simulations which were intentionally made larger to simulate an exposed clay surface. Once the surface was fully wetted, 50 ns of equilibrated MD simulation were calculated.

VMD [37] was used to create the systems for simulation and for analysis after the MD simulations were complete. The interlayer systems simulated consisted of a random distribution of nucleotides of interest near a $52 \text{\AA} \times 52 \text{\AA}$ square of montmorillonite surface, and a 15\AA layer of water on each side of the surface. With PBC, the result is an infinite montmorillonite crystal with an interlayer region size of 30\AA when fully wetted. For simulations of reaction on a single exposed surface the separation was increased to 70\AA so that the effect of the image of the surface was negligible. We have modeled the 1-methyladenine (1-Mead) activating group by analogy from the topology of the unmodified adenine and corrected the partial charges to be consistent with those found from Hartree Fock calculations. This activating group was used in experiments resulting in the 50-mer oligonucleotides [59]. Some studies of montmorillonite catalysis of oligonucleotide synthesis used a phospho-imidazolid (Imp) 5' terminal group instead of 1-Mead. In order to allow for comparison between our simulations, we use the 1-Mead activating group in all our simulations. A model for L-adenosine used in chirality studies was made by analogy to the D-form. For systems in which we study dimerization, four molecules of each species are included in the simulation, with one set of reactants activated by 1-Mead (uridine (1-MeadpU), or adenosine (1-MeadpA)) and the other terminated with a monophosphate group (AMP or UMP). Studies involving the formation of trimers consist of four adenosine dimers activated with 1-Mead (1-MeadpApA) and seven adenosine monomers terminated with a single phosphate (AMP). For this system, an excess of unreacted monomer was chosen to reflect the likely relative concentrations under experimental conditions. Calcium ions were added in sufficient number to neutralize the system. For all systems, nucleotides and ions were placed randomly. The total number of

atoms for each simulation varies depending on the specific reactants included, but is generally $\sim 10,000$ atoms for each simulated system. The choice of pairing activated with monophosphate-terminated reactants was made based on a proposed reaction scheme for stepwise oligonucleotide synthesis on montmorillonite [74].

Montmorillonite has a monoclinic crystal structure[85; 86] with basis vectors given in equation 3.3. The modeled montmorillonite surface from the work of Heinz et al. [66] was used in our simulations (see Figure 3.14). This model has been shown to reproduce experimentally observed bond vibrations from IR spectra as well as mechanical properties such as swelling and compression in the presence of intercalated polymers. Force field parameters compatible with CHARMM for montmorillonite were also obtained from this work. The chemical formula for the simulated montmorillonite, including charge defects, is $(\text{Al}_{3.34}\text{Mg}_{0.66})\text{Si}_8\text{O}_{20}(\text{OH})_4\text{X}_{0.33}$. In our model, X is divalent calcium in the interlayer region. A number of other exchangeable cations have been shown to be effective in catalytic montmorillonite [87]. Calcium was chosen for its strong electrostatic attraction for the negatively charged terminal phosphate group on the nucleotides as well as for the negatively charged surface. Calcium also exhibits a relatively short-lived binding to water molecules as compared to magnesium which rarely dissociates from water and thus may not contribute to direct associations with the reactants as frequently. The Heinz model has substitutions of Mg for Al, and includes fewer defects than those present in the clay used by Ferris. This montmorillonite model has 17% of the octahedral aluminum sites replaced by magnesium, resulting in a surface charge density, $\sigma = -2.4 \times 10^{-5} \text{Coul}/\text{cm}^2$. The clay used in the experiments contained some iron substitutions and roughly 46% magnesium, resulting in a higher surface charge density. We find in this work that the abundance of defects present in the model adequately recovers the trends from experimental results and there was no need to re-parameterize a new montmorillonite model. Other published work on MD simulations of montmorillonite have used the CLAYFF forcefield parameters adapted for the Amber MD simulation package [88]. This forcefield uses only non-bonded forces to maintain the clay structure. Due to its compatibility with CHARMM and its proven accuracy in recovering properties in simulating clay with intercalated polymers, the Heinz model was chosen for the simulations in this study.

$$a = 5.200\text{\AA}, \quad b = 9.200\text{\AA}, \quad c = 10.130\text{\AA}, \quad \alpha = \gamma = 90^\circ, \quad \beta = 99^\circ \quad (3.3)$$

Radial Distribution Function Chemical reactions cannot be simulated using the classical MD methods employed in this study. However, a reaction rate can be inferred from an analysis of the simulation using radial distribution functions. A radial distribution function ($g(r)$) of atoms involved in the reactions of interest yields the relative abundance of reactant A from reactant B as a function of their separation distance [38; 89]. The $g(r)$ is calculated using the Radial Pair Distribution Function tool built into the VMD software package, using the default value of 0.1\AA for histogram bin size, Δr . In our analysis, the reactants are the terminal (2' or 3') hydroxy groups of a 5' activated mono- or di-nucleotide and the phosphate a 5' monophosphate-terminated nucleotide. We use the hydroxy oxygens and the phosphorus atoms in the calculation of Eq. 3.4, where V is the volume of the simulation cell, N_s are the numbers of the selected atoms (either phosphorus or oxygens), M is the total number of trajectory frames sampled, and r is the radial separation distance. Using $g(r)$ s calculated from our classical MD simulations, we are able to quantify the likelihood of a reaction

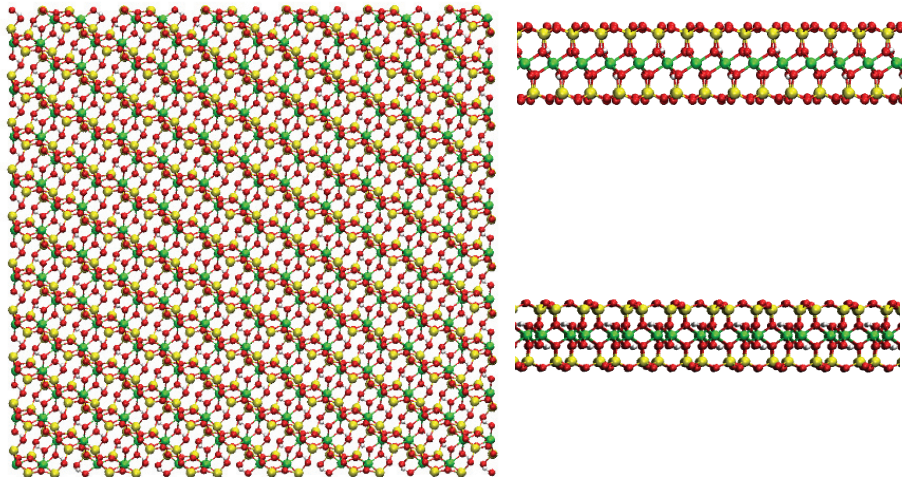


Figure 3.14: Right: Two side views of the montmorillonite structure used. Left: Top view of structure. Silicon atoms represented in yellow, Oxygens in red, Aluminum and Magnesium in green, and Hydrogens in white

taking place, as this is directly related to the contact time of the reactants and contact time is directly related to $g(r)$ peak intensity, g_{max} . To attempt to capture as closely as possible the relationship between the $g(r)$ and the reaction rate constant,

$$g_{PO}(r) = \frac{V}{4\pi r^2 \Delta r M N_P N_O} \sum_{m=1}^M \sum_{i=1}^{N_P} \sum_{j=1}^{N_O} Q_m(r; r_{P_i}, r_{O_j}) \quad (3.4)$$

Where the counting function, Q , is defined by:

$$Q(r; r_{P_i}, r_{O_j}) = \begin{cases} 1 & \text{if } (r - \Delta r/2) \leq (r_{P_i} - r_{O_j}) < (r + \Delta r/2) \\ 0 & \text{otherwise} \end{cases} \quad (3.5)$$

Averages were made over the entire 50ns trajectories, after excluding the first nanosecond for equilibration.

Some additional considerations must be made in order to use radial distribution functions as a measure of relative kinetic rate constants. First, if the 5' end of the monophosphate is close to both the 2' and 3'

groups of the other reactant the reaction is equally likely to take place linking either $3' - 5'$ or $2' - 5'$. To take into account this shared likelihood, when the $5'$ phosphorus atom is simultaneously within 4 \AA of the $2'$ and $3'$ groups of the other reactant, we multiply the $g(r)$ calculated at that time step by a factor of 0.5 to reflect this shared probability. Second, calcium ions, when present, may enhance the reaction rate. For simulations in which calcium appears to remain close to the terminal hydroxyl groups, we calculate $g(r)$ s only for those time steps in which a calcium ion is within 4 \AA of these groups.

A Generalized Form of a Positively Charged Surface These simulations were carried out under an NPT ensemble with 1 atmosphere of pressure and a temperature of 300K. Periodic boundary conditions were enforced with PME full electrostatics. The system was minimized for 4000 iterations followed by 4 ns of molecular dynamics using a 1 fs timestep.

The “generalized” positive mineral surface was constructed using an array of sodium ions placed at the exchangeable cation sites of the montmorillonite surface. While montmorillonite has a net negative surface charge, it’s structure was used as an analogy for a mineral surface consisting of regularly repeating positive charge sites for the purposes of this study.

Sodium ions were placed at the exchangeable cation positions of the clay located at 000 and repeated by lattice translation vectors appropriate for monoclinic crystals to construct a 2D surface of sodium ions representing the saturated cationic structure of the clay given in equation 3.6. These sodium ions are constrained to their initial positions by a harmonic restoring force with a force constant of $20 \text{ kcal/mol \AA}^2$.

$$\vec{A}_1 = \frac{1}{2}a\hat{x} - \frac{1}{2}b\hat{y}, \quad \vec{A}_2 = \frac{1}{2}a\hat{x} + \frac{1}{2}b\hat{y} \quad (3.6)$$

Uracil nucleotides were placed above this array in a ratio of one nucleotide for every two cation sites. Above this, a charge-neutralizing 0.7M NaCl solution containing equal numbers of Glu and Leu residues was placed. The size of the simulation was $60\text{\AA} \times 40\text{\AA} \times 50\text{\AA}$. The last 3 ns of simulation were used for analysis. Radial distribution functions were calculated to determine selectivity of oligonucleotide formation catalyzed by the surface, as well as promotion of oligopeptide formation. Figure 3.15 shows the arrangement of sodium ions with the equilibrated mononucleotides adhering to the surface.

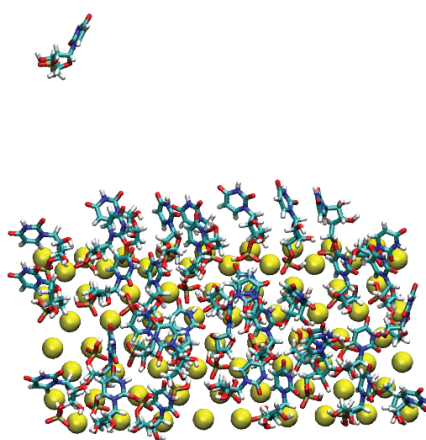


Figure 3.15: Sodium ions (yellow) in an array with equilibrated uracil mononucleotides electrostatically bound. The surface is completely saturated by uracil nucleotides, indicated by the lone uracil mononucleotide adrift in the solution. Waters, ions, and amino acids not shown.

Chapter 4

Reaction Pathway for Charging tRNA in an Aminoacyl tRNA Synthetase

4.1 Background

George Whitesides, in a lecture at the Santa Fe Institute in 2007, drew an analogy to a bridge as he discussed research into the origins of life. On one side of the Whitesides bridge are the geochemists who strive to glean insight into early life and/or proto-metabolism from a purely physical-chemical inspection of what the early earth would have provided. On the other side, we have those who look to the past through the lens of modern biology. They begin with biochemistry as their axioms and attempt to reverse engineer them to gain an understanding of what must have been early proto-biology. The preceding chapters were devoted to studies of interactions and reactions directly related to putative origins of life scenarios. Essentially, the work presented in this dissertation up to this point has been from the first side of the “Whitesides Bridge.”

In this chapter, we take an approach from the other side of the bridge, and begin by examining a reaction that is essential to maintaining the genetic code in modern biology. Here, we report on our investigations of the tRNA charging reaction within amino-acyl tRNA synthetases. In particular, we have examined the charging of tRNA^{Glu} from an adenylate of glutamate in GluRS, based on the crystal structure of the tRNA^{Glu}-GluRS-adenylate analog complex from *Thermus Thermophilus* [90]. A reaction mechanism has been proposed by others [91] for the homologous system of GlnRS-tRNA^{Gln} which consists of an acyl substitution reaction with the 2' oxygen of the tRNA's A76 acting as the nucleophile and attacking the carboxy carbon of the Glu-adenylate. The acyl substitution reaction's initial nucleophilic attack is believed to occur simultaneously with a proton transfer from the 2' hydroxy of the tRNA's A76 to the phosphate of the adenylate, according to the proposed scenario. The reaction scheme is depicted by the black arrows in figure 4.1. We have chosen two reaction coordinates, denoted as **X** and **Y** in figure 4.1. **X** represents the distance between the 2' hydrogen and the nearest oxygen of the phosphate of the Glu-adenylate and **Y** is the distance between the 2' oxygen and the carbonyl carbon of the adenylate.

We have performed an initial survey of the uncatalyzed reaction (i.e. with the reactant's only, and not including the reaction center of the synthetase). Using the knowledge gained from this initial assessment, we apply a multi-level geometry optimization algorithm using both classical force fields and quantum chemical methods to delve deeper into the catalyzed reaction.

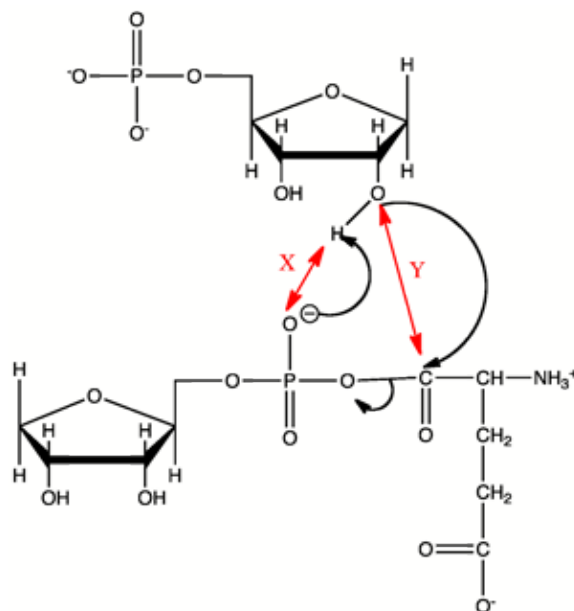


Figure 4.1: Simplified model and reaction coordinates for the GluRS-tRNA^{Glu} charging reaction.

4.2 Reaction Pathway: Uncatalyzed

In an initial attempt to survey the reaction path for transfer of glutamate to tRNA^{Glu}, we have a potential energy surface for a simplified model of the reactants in “gas phase” (no surrounding protein from the active site of the GluRS included). Energetic barrier encountered in this system are to be considered those of the uncatalyzed reaction, since none of the synthetase is included in the system being calculated. We calculate the potential energy surface by scanning each collective variable independently, and optimizing the geometry at each point, constraining **X** and **Y**. If the reaction is concerted, given the optimization at each point, we expect to observe a minimal path along the potential energy surface formed as a function of these collective variables. The bases of the nucleotides, for the purposes of this initial study, have been replaced with hydrogens to simplify our calculations. For the uncatalyzed reaction, we have determined a potential energy barrier slightly over 33 kcal/mol. However, the reaction path seems to be in agreement with the proton transfer occurring simultaneously with the nucleophilic attack, in agreement with the proposed reaction mechanism.

A significant peak is observed in the potential energy surface at $\mathbf{X} \simeq 1.3\text{\AA}$ and $\mathbf{Y} \simeq 1.5\text{\AA}$. An alternate reaction path is observed crossing between two maxima, corresponds to a nucleophilic attack on the adenylate carboxy carbon prior to proton transfer, however the barrier height for this path is slightly larger than the one for the concerted reaction (simultaneous proton transfer with the nucleophilic attack).

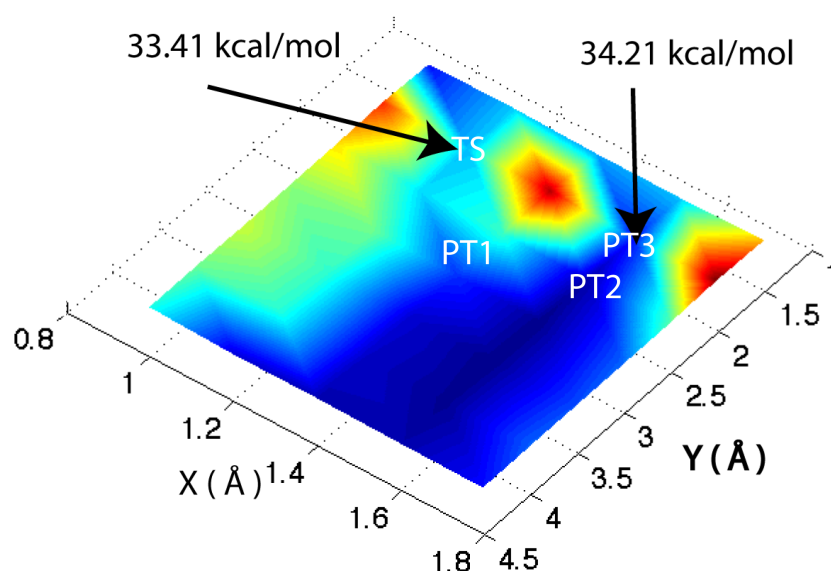


Figure 4.2: Potential energy surface along the two reaction coordinates. Two plausible reaction paths are located. Both exhibit high potential energy barriers. Points labeled on plot (TS, PT1, PT2, and PT3) are used for calculations of the catalyzed reaction.

4.3 Reaction Pathway: Catalyzed

Building from our work on the uncatalyzed reaction scheme described above, we have begun to examine the reaction including the influence of the enzyme. In this work, we use the ONIOM method in Gaussian 09, in which molecular mechanics and/or lower level QM theory may be used to treat a portion of the system being studied while higher level methods (e.g. Hartree-Fock) may be used for the reactants atoms. This approach would allow for a more complete representation of the environment within the catalytic site. Interactions based on both electrostatic effects as well as van der Waals forces would be taken into account in this approach, allowing for hydrogen bond formation, polarization of reactant electron orbitals, and the formation of salt bridges.

In order to include the effects of the environment within the catalytic site of the GluRS synthetase, we have selected the amino acid residues from of the GluRS known to make contacts with tRNA A76 or the Glu-AMP molecules [92] and included additional residues which were found to be within 5 Å of the reactants, even if they did not make direct contact. The decision to include near-by residues that did not make obvious interactions with the reactants was made to ensure that the electrostatic environment surrounding the reactants was as realistic as possible for this simplified model. Unlike the earlier survey study, the nucleobases for both the Glu-AMP and the A76 are included in the calculations described in this section in order to more realistically represent the conditions in which the reaction takes place. The calculations in this section also involve a two-level treatment of the atoms using the ONIOM method included in the Gaussian quantum chemistry software package. A low level classical force field (UFF [93]) was used for the atoms of

the GluRS catalytic site listed in Table 4.1 and an unrestricted Hartree-Fock (UHF) method with the 6-31G* basis set was applied to all the atoms in A76 and Glu-AMP. The electrostatic charge from the atoms given the UFF treatment were embedded in the electronic structure integrals for the atoms receiving the UHF treatment.

Table 4.1: Residues from the GluRS catalytic site included at the classical level for calculations of the catalyzed tRNA^{Glu} charging reaction

Residue number	Residue type	Interaction with reactants
5	Arg	H-bonds with Glu-AMP sidechain
6	Ile	Near Glu-AMP backbone amine
7	Ala	H-bond with Glu-AMP backbone amine
8	Pro	Near Glu-AMP
9	Ser	H-bonds with Glu-AMP backbone amine and phosphate
15	His	Interacts with Glu-AMP base
17	Gly	Near Glu-AMP
18	Thr	Near Glu-AMP
20	Tyr	Near Glu-AMP
21	Ile	Near Glu-AMP
41	Glu	Electrostatic interaction with Glu-AMP backbone amine
43	Thr	H-bond through water with A76 3' hydroxy
44	Asp	Near A76
47	Arg	Near A76
180	Lys	Near Glu-AMP sidechain
185	Pro	Near Glu-AMP sidechain
186	Thr	Near Glu-AMP sidechain
187	Tyr	Near Glu-AMP sidechain
188	Hse	Near Glu-AMP sidechain
191	Asn	Backbone interacts with Glu-AMP sidechain
203	Val	Near Glu-AMP
204	Ile	Near Glu-AMP
205	Arg	Electrostatic interaction with Glu-AMP sidechain
206	Ala	Backbone interacts with Glu-AMP 3' hydroxy
208	Glu	Near Glu-AMP ribose
209	Trp	Interacts with Glu-AMP 3' hydroxy
212	Ser	Near Glu-AMP
234	Pro	Near Glu-AMP
235	Leu	Near Glu-AMP
236	Leu	Interacts with Glu-AMP
243	Lys	Near Glu-AMP base
244	Ile	H-bond to Glu-AMP exocyclic amine

Glu41 is a particularly influential residue in the reaction center of GluRS. It has been noted that the pKa for the Glu41 sidechain in its local environment can range from 8.06 to 8.90 and that of the backbone, α amine, group of the Glu-AMP is 7.23 [92]. Given these pKas, it is possible that the Glu41 sidechain may

be in a neutral, protonated, form, having received a proton from the Glu-AMP α amine. We have performed geometry optimization calculations for each of the two possible protonation states of these moieties for both the pre- and post-transfer states of the system (A76 with Glu-AMP and Glu-A76 with H-AMP respectively). Initial configurations for each of the two stages of the reaction (pre- and post-transfer) were taken from equilibrated MD trajectories of the entire GluRS and tRNA^{Glu} system. The residues listed in table 4.1 were selected from a representative frame from each trajectory and treated with UFF while the reactants were calculated with UHF. The energies from the optimized geometries are listed in table 4.2. All energies are scaled relative to the energy of the anionic Glu41 and cationic α amine Glu-AMP.

Table 4.2: Relative configurational energies taken from optimized geometries for reactant and product states of the tRNA-charging reaction

Protonation state	Reaction step	Relative E (kcal/mol)
Negative Glu41	pre-transfer	0
Negative Glu41	TS	-14.56
Negative Glu41	PT1	0.87
Negative Glu41	PT2	No convergence
Negative Glu41	PT3	No convergence
Negative Glu41	post-transfer	-27.66
Neutral Glu41	pre-transfer	0
Neutral Glu41	TS	5.65
Neutral Glu41	PT1	No convergence
Neutral Glu41	PT2	-13.53
Neutral Glu41	PT3	17.00
Neutral Glu41	post-transfer	-8.34

For the calculation including an anionic Glu41 sidechain and cationic Glu-AMP α amine, we see a change in potential energy, ΔE , as the reaction goes forward of -27.66 kcal/mol. This compares to a ΔE for the reaction when both the Glu41 and Glu-AMP's backbone amine are neutralized of -8.34 kcal/mol. Dealing only with the change in potential energy, the reaction seems favorable for either set of protonation states. However, the absolute energies when both Glu41 and the Glu-AMP backbone amine groups are neutral prove to be significantly higher than for those when these moieties are charged. This is partially attributable to the electrostatic interaction energy between the two charged groups when they are charged ($V_{electrostatic} = e^2/3\text{\AA} = 111 \text{ kcal/mol}$).

For the case in which the Glu41 side chain and the α amine group are left charged, there is a strongly negative potential energy at the TS point. This point was believed to be the transition state, based on Figure 4.2, however, given the slight positive value for PT1 with this charge state, it would appear that the transition state has shifted from where it was in the gas phase calculations.

When the α amine and Glu41 are both neutral, the TS point has a positive potential energy relative to the reactant state. While PT2 shows a strongly negative potential energy given this charge configuration, PT3 shows a highly unfavorable potential energy barrier, suggesting that the alternate reaction path suggested by Figure 4.2 remains unavailable for the catalyzed reaction. In order to make more definitive statements

about the favorability of the reaction, a free energy calculation of the pre- and post-transfer states must be obtained.

To gain a clearer understanding of what may be taking place in the catalyzed reaction, we begin by considering the changes in the collective variables described in figure 4.1 (table 4.3). For TS, PT1, PT2, and PT3, the **X** and **Y** coordinates were fixed, but are included in the table to Here, we observe that when Glu41's sidechain and Glu-AMP's α amine are neutral, the key collective variables being monitored drift significantly far from what would be considered a distance for a reaction to occur. However, it has been suggested that deprotonation of the α -amine on Glu-tRNA^{Glu} takes place prior to dissociation of the charged tRNA from the GluRS. Based on the unfavorable distances in **X** and **Y** observed for the neutral Glu41 and Glu α -amine groups, it seems clear that this proton transfer either takes place immediately following or at the same time as the charging reaction.

Table 4.3: Reaction coordinate values for calculations of pre- and post-transfer states as well as points indicated in Figure 4.2 with and without proton transfer between Glu α -amino and Glu41 (Neutral/Charged, respectively)

Stage of reaction	Neutral or Charged Glu41	X (Å)	Y(Å)
Pre-transfer	Charged	1.90	3.55
TS	Charged	1.1	1.5
PT1	Charged	1.3	2.7
PT2	Charged	1.55	2.25
PT3	Charged	1.55	1.75
Post-transfer	Charged	0.95	1.34
Pre-transfer	Neutral	7.58	7.97
TS	Neutral	1.1	1.5
PT1	Neutral	1.3	2.7
PT2	Neutral	1.55	2.25
PT3	Neutral	1.55	1.75
Post-transfer	Neutral	0.95	1.34

Given the relatively small change in **X**, and the wide potential energy valley in Figure 4.2, we have also explored the possible reaction scenario of having the 2' hydrogen from A76 being transferred first to the Glu-AMP phosphate prior to the acyl substitution. We consider the α -amine group to be in its cationic form, and the Glu41 sidechain to be in its deprotonated form. In an MD simulation of this state (see Figure 4.3), we observe a **Y** distance of 3.8 Å, but also observe that the sidechain of Glu41 has flipped to point away from the α -amine group of the Glu-AMP, and a strong electrostatic interaction between the cationic α -amine and the anionic (deprotonated) 2' hydroxy of A76. This strong interaction could interfere with the nucleophilic attack of the 2' oxygen on the carbonyl carbon.

4.4 Conclusions

Given both the MD and QM calculations pertaining to the tRNA charging reaction, we can conclude that deprotonation of the α amino group of Glu, with the proton being transferred to Glu41, likely takes place

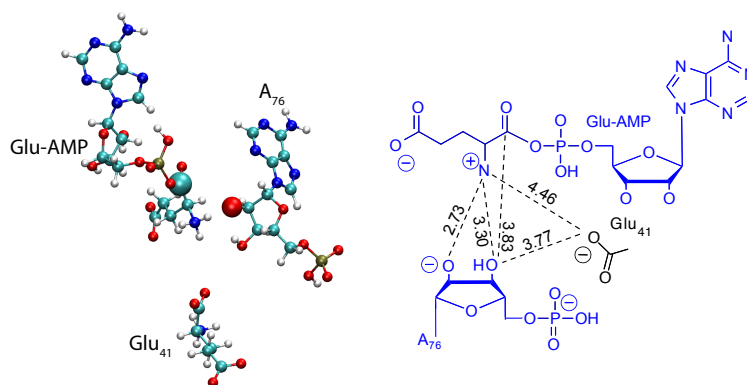


Figure 4.3: **Left** Snapshot from MD simulation including active site residues (all but Glu41 not shown). A76 O2' represented as larger red sphere, and Glu-AMP carboxyl carbon represented as large cyan sphere. **Right** Schematic of interactions in MD snapshot, with distances labeled in Angstroms.

shortly after or at the same time as the acyl substitution reaction that loads Glu onto tRNA^{Glu}. From the results of the MD simulation, a proton relay may also be taking place. The 2' hydrogen of A76 may transfer to the Glu-AMP phosphate first. For the acyl substitution reaction to proceed, an additional proton transfer from the Glu-AMP α -amine to Glu41 reduces the electrostatic interaction with A76, and allows the 2' oxygen the freedom to perform the nucleophilic attack for the substitution reaction. Further calculations are being performed to examine this scenario in greater detail.

4.5 Methods

We have used Gaussian03 and Gaussian09 in all the calculations performed in this chapter. Systems were set up using the free molecular visualization and editing software package, Avogadro. A preliminary calculation was performed to obtain a potential energy surface for a reduced system consisting of the two reactants, A76 from the tRNA and and Glu-AMP adenylate, with the adenine bases replaced with hydrogens to simplify the system. Other calculations used the multi-level ONIOM method in Gaussian to allow a classical forcefield treatment of atoms not directly involved in the reaction of interest, and a high level quantum chemical treatment of the reactant molecules themselves. This method provides a relatively computationally cheap means to include effects of the surrounding environment on the reaction.

For the MD simulations described in this section, the same residues included in the ONIOM calculations were used with neutral terminating groups for each peptide chain. Harmonic constraints were applied to the terminal nitrogen and carbon atoms to preserve the relative positions of the peptides. An NPT ensemble was used at a temperature of 300K and pressure of 1 atmosphere. Simulations were performed for 10 ns, and the equilibrated portions of the resulting trajectories used for analysis.

Bibliography

- [1] Woese, CR, Dugre, DH, Dugre, SA, Kondo, M, Saxinger, WC (1966) On the fundamental nature and evolution of the genetic code. *Cold Spring Harb Symp Quant Biol* 31:723–736.
- [2] Rateev, MA, Gorbunova, ZN, Lisitzyn, AP, Nosov, GL (1969) The distribution of clay minerals in the oceans. *Sedimentology* 13:21–43.
- [3] Peeters, Z, Botta, O, Charnley, SB, Ruiterkamp, R, Ehrenfreund, P (2003) The astrobiology of nucleobases. *The Astrophysical Journal Letters* 593:L129.
- [4] Butler, T, Goldenfeld, N, Mathew, D, Luthey-Schulten, Z (2009) Extreme genetic code optimality from a molecular dynamics calculation of amino acid polar requirement. *Phys Rev E* 79:060901.
- [5] Mathew, DC, Luthey-Schulten, ZA (2008) On the physical basis of the amino acid polar requirement. *J. Mol. Evol.* 66:519–528.
- [6] Mathew, D, Luthey-Schulten, Z (2010) Influence of montmorillonite on nucleotide oligomerization reactions: A molecular dynamics study. *Orig. Life Evol. Biosph.* 40:303–317.
- [7] Woese, CR (1973) Evolution of the genetic code. *Naturwissenschaften* 60:447–459.
- [8] Woese, CR, Dugre, DH, Saxinger, WC, Dugre, SA (1966) The molecular basis for the genetic code. *Proc Natl Acad Sci U S A* 55:966–974.
- [9] Alff-Steinberger, C (1969) The genetic code and error transmission. *Proc. Natl. Acad. Sci. USA* 64:584–591.
- [10] Woese, CR, Olsen, GJ, Ibba, M, Söll, D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* 64:202–236.
- [11] Caporaso, JG, Yarus, M, Knight, R (2005) Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code. *J. Mol. Evol.* 61:597–607.
- [12] Haig, D, Hurst, LD (1991) A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* 33:412–417.
- [13] Freeland, SJ, Hurst, LD (1998) The genetic code is one in a million. *J Mol Evol* 47:238–248.

- [14] Freeland, SJ, Knight, RD, Landwebber, LF, Hurst, LD (2000) Early fixation of an optimal genetic code. *Mol Biol Evol* 17:511–518.
- [15] Di Giulio, M (1989) The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *J. Mol. Evol.* 29:288–293.
- [16] McQuarrie, DA (2000) *Statistical Mechanics* (University Science Books, Sausalito, California), 2nd edition, pp 257–274.
- [17] Campo, MG (2006) Molecular dynamics simulation of glycine zwitterion in aqueous solution. *J. Chem. Phys* 125:114511.
- [18] Nigmatullin, R, Lovitt, R, Wright, C, Linder, M, Nakari-Setälä, T, Gama, M (2004) Atomic force microscopy study of cellulose surface interaction controlled by cellulose binding domains. *Colloids Surf. B, Biointerfaces* 35:125–135.
- [19] Nimlos, MR, Matthews, JF, Crowley, MF, Walker, RC, Chukkapalli, G, Brady, JW, Adney, WS, Cleary, JM, Zhong, L, Himmel, ME (2007) Molecular modeling suggests induced fit of Family I carbohydrate-binding modules with a broken-chain cellulose surface. *Protein Eng. Des. Sel* 20:1–9.
- [20] Crick, FHC (1968) The origin of the genetic code. *J. Mol. Biol.* 38:367 – 379.
- [21] Grantham, R (1974) Amino acid difference formula to help explain protein evolution. *Science* 185:862–864.
- [22] Knight, RD (2001) Ph.D. thesis (Princeton University).
- [23] Baaske, P, Weinert, FM, Duhr, S, Lemke, KH, Russell, MJ, Braun, D (2007) Extreme accumulation of nucleotides in simulated hydrothermal pore systems. *Proc. Natl. Acad. Sci. USA* 104:9346–9351.
- [24] Nakatsu, T, Kato, H, Oda, J (1998) Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class II aminoacyl-tRNA synthetase. *Nat. Struct. Biol* 5:15–19.
- [25] Woese, CR (2007) (Private communication).
- [26] Huang, W, Ferris, JP (2003) Synthesis of 35-40 mers of RNA oligomers from unblocked monomers. A simple approach to the RNA world. *Chem Comm* 12:1458–1459.
- [27] Ferris, J, Joshi, P, Wang, KJ, Miyakawa, S, Huang, W (2004) Catalysis in prebiotic chemistry: application to the synthesis of RNA oligomers. *Adv Space Res* 33:100–105.
- [28] Baaske, P, Weinert, FM, Duhr, S, Lemke, KH, Russell, MJ, Braun, D (2007) Extreme accumulation of nucleotides in simulated hydrothermal pore systems. *Proc. Natl. Acad. Sci. USA* 104:9346–9351.
- [29] Yarus, M, Caporaso, JG, Knight, R (2005) Origins of the genetic code: The escaped triplet theory. *Annu. Rev. Biochem.* 74:179–198.

- [30] Martin, AJP, Synge, RLM (1941) A new form of chromatogram employing two liquid phases 1. A theory of chromatography 2. Application to the micro-determination of the higher monoamino-acids in proteins. *Biochem. J.* 35:1358–1368.
- [31] Mirzaev, SZ, Behrends, R. and Heimburg, T, Haller, J, Kaatze, U (2006) Critical behavior of 2,6-dimethylpyridine-water: Measurements of specific heat, dynamic light scattering, and shear viscosity. *J. Chem. Phys.* 124:144517.
- [32] Mirzaev, SZ, Iwanowski, I, Zaitdinov, M, Kaatze, U (2006) Critical dynamics and kinetics of elementary reactions of 2,6-dimethylpyridine-water. *Chem. Phys. Lett.* 431:308–312.
- [33] Miller, JC, Rallison, JM (2007) Interfacial instability between sheared elastic liquids in a channel. *J. Non-Newtonian Fluid Mech.* 143:71–78.
- [34] Phillips, JC, Braun, R, Wang, W, Gumbart, J, Tajkhorshid, E, Villa, E, Chipot, C, Skeel, RD, Kale, L, Schulten, K (2005) Scalable molecular dynamics with NAMD. *J Comp Chem* 26:1781–1802.
- [35] MacKerell, Jr., AD, Banavali, N (2000) All-atom empirical force field for nucleic acids: 2) Application to molecular dynamics simulations of DNA and RNA in solution. *J Comp Chem* 21:105–120.
- [36] MacKerell, Jr., AD, Bashford, D, Bellott, M, Dunbrack Jr., R, Evanseck, J, Field, M, Fischer, S, Gao, J, Guo, H, Ha, S, Joseph-McCarthy, D, Kuchnir, L, Kuczera, K, Lau, F, Mattos, C, Michnick, S, Ngo, T, Nguyen, D, Prodhom, B, Reiher, III, W, Roux, B, Schlenkrich, M, Smith, J, Stote, R, Straub, J, Watanabe, M, Wiorkiewicz-Kuczera, J, Yin, D, Karplus, M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102:3586–3616.
- [37] Humphrey, W, Dalke, A, Schulten, K (1996) VMD - Visual Molecular Dynamics. *J Molec Graphics* 14:33–38.
- [38] Allen, M, Tildesley, D (1987) *Computer Simulation of Liquids* (Clarendon Press, Oxford), p 55.
- [39] Consden, R, Gordon, AH, Martin, AJP (1944) Qualitative analysis of proteins: a partition chromatographic method using paper. *BCHJ* 38:224–232.
- [40] England, A, Cohn, EJ (1935) Studies in the physical chemistry of amino acids, peptides and related substances. iv. the distribution coefficients of amino acids between water and butyl alcohol. *JACS* 57:634–637.
- [41] Cohen, J, Arkhipov, A, Braun, R, Schulten, K (2006) Imaging the migration pathways for O₂, CO, NO, and Xe inside myoglobin. *Biophys. J.* 91:1844–1857.
- [42] Sotomayor, M, Vásquez, V, Perozo, E, Schulten, K (2007) Ion conduction through MscS as determined by electrophysiology and simulation. *Biophys. J.* 92:886–902.
- [43] Snow, TP, McCall, BJ (2006) Diffuse atomic and molecular clouds. *Annual Review of Astronomy and Astrophysics* 44:367–414.

- [44] Hudgins, DM, Bauschlicher, CW, Allamandola, LJ (2005) Variations in the peak position of the 6.2 μ m interstellar emission feature: a tracer of N in the interstellar polycyclic aromatic hydrocarbon population. *Astrophysical Journal* 632:316–332.
- [45] Lackschewitz, KS, Botz, R, Garbe-Schönberg, D, Scholten, J, Stoffers, P (2006) Mineralogy and geochemistry of clay samples from active hydrothermal vents off the north coast of Iceland. *Mar. Geol.* 225:177–190.
- [46] Gesteland, RF, Cech, TR, Atkins, JF, eds (2006) *The RNA world* (Cold Spring Harbor Laboratory Press), 3rd edition.
- [47] Huber, C, Wächtershäuser, G (2006) α -hydroxy and α -amino acids under possible hadean, volcanic origin-of-life conditions. *Science* 314:630–632.
- [48] Wächtershäuser, G (1992) Groundworks for an evolutionary biochemistry: the iron-sulfur world. *Prog. Biophys. Molec. Biol.* 58:85–201.
- [49] Orgel, LE (2008) The implausibility of metabolic cycles on the prebiotic earth. *PLoS Biol.* 6:e18.
- [50] Eschenmoser, A (2007) On a hypothetical generational relationship between HCN and constituents of the reductive citric acid cycle. *Chem Biodivers* 4:554–573.
- [51] Morowitz, HJ, Kostelnik, JD, Yang, J, Cody, GD (2000) The origin of intermediary metabolism. *Proc. Natl. Acad. Sci. USA* 97:7704–4408.
- [52] Smith, E, Morowitz, HJ (2004) Universality in intermediary metabolism. *Proc. Natl. Acad. Sci. USA* 101:13168–13173.
- [53] Wächtershäuser, G (1988) Before enzymes and templates: theory of surface metabolism. *Microbiol Rev* 52:452–484.
- [54] Cody, GD, Boctor, NZ, Filley, TR, Hazen, RM, Scott, JH (2000) Primordial carbonylated iron-sulfur compounds and the synthesis of pyruvate. *Science* 289:1337–1340.
- [55] Russell, MJ (2008) in *Energy from the Sun*, 14th International Congress on Photosynthesis, eds Allen, JF, Gantt, E, Golbeck, JH, Osmond, B (Springer).
- [56] Monnard, PA (2007) Question 5: Does the RNA-world still retain its appeal after 40 years of research? *Orig Life Evol Biosph* 37:387–390.
- [57] Monnard, PA (2005) Catalysis in abiotic structured media: an approach to selective synthesis of biopolymers. *Cell Mol Life Sci* 62:520–534.
- [58] Ferris, JP (2006) Montmorillonite-catalysed formation of RNA oligomers: the possible role of catalysis in the origins of life. *Phil. Trans. R. Soc. B* 361:1777–1786.

- [59] Huang, W, Ferris, JP (2006) One-step, regioselective synthesis of up to 50-mers of RNA oligomers by montmorillonite catalysis. *J. Am. Chem. Soc.* 128:8914–8919.
- [60] Zagorevskii, DV, Aldersley, MF, Ferris, JP (2006) MALDI analysis of oligonucleotides directly from montmorillonite. *J. Am. Soc. Mass Spectrom.* 17:1265–1270.
- [61] Ferris, JP (2005) Mineral catalysis and prebiotic synthesis: montmorillonite-catalyzed formation of RNA. *Elements* 1:145–149.
- [62] Ferris, JP, Joshi, PC, Wang, KJ, Miyakawa, S, Huang, W (2004) Catalysis in prebiotic chemistry: application to the synthesis of RNA oligomers. *Adv Space Res* 33:100–105.
- [63] Miyakawa, S, Ferris, JP (2003) Sequence- and regioselectivity in the montmorillonite-catalyzed synthesis of RNA. *J. Am. Chem. Soc.* 125:8202–8208.
- [64] Ertem, G, Snellinger-O'Brien, A, Ertem, M, Rogoff, D, Dworkin, J, Murray, Johnston, M, Hazen, R (2008) Abiotic formation of RNA-like oligomers by montmorillonite catalysis: part II. *Astrobiology* 7:1–7.
- [65] Mazo, MA, Manevitch, LI, Gusarova, EB, Shamaev, MY, Berlin, AA, Balabaev, NK, Rutledge, GC (2008) Molecular dynamics simulation of thermomechanical properties of montmorillonite crystal 2. montmorillonite crystals with PEO oligomer intercalates. *J. Phys. Chem B* 112:3597–3604.
- [66] Heinz, H, Koerner, H, Anderson, KL, Vaia, RA, Farmer, B (2005) Force field for mica-type silicates and dynamics of octadecylammonium chains grafted to montmorillonite. *Chem. Mater.* 17:5658–5669.
- [67] Thyveetil, M, Coveney, PV, Greenwell, HC, Suter, JL (2008) Computer simulation study of the structural stability and materials properties of DNA-intercalated layered double hydroxides. *J. Am. Chem. Soc.* 130:4742–4756.
- [68] Franchi, M, Ferris, JP, Gallori, E (2003) Cations as mediators of the adsorption of nucleic acids on clay surfaces in prebiotic environments. *Orig. Life. Evol. Biosphere* 33:1–16.
- [69] Hanczyc, MM, Fujikawa, SM, Szostak, JW (2003) Experimental models of primitive cellular compartments: encapsulation, growth, and division. *Science* 302:618–622.
- [70] Hanczyc, MM, Mansy, SS, Szostak, JW (2007) Mineral surface directed membrane assembly. *Orig Life Evol Biosph* 37:67–82.
- [71] Lawrence, MS, Bartel, DP (2005) New ligase-derived RNA polymerase ribozymes. *RNA* 11:1173–1180.
- [72] Ertem, G, Ferris, JP (1998) Formation of RNA oligomers on montmorillonite: site of catalysis. *Orig Life Evol Biosph* 28:485–499.
- [73] Ertem, G (2004) Montmorillonite, oligonucleotides, RNA and origin of life. *Orig Life Evol Biosph* 34:549–570.

- [74] Kawamura, K, Ferris, JP (1999) Clay catalysis of oligonucleotide formation: Kinetics of the reaction of the 5'-phosphorimidazolides of nucleotides with the non-basic heterocycles uracil and hypoxanthine. *Orig Life Evol Biosph* 29:563–591.
- [75] Keizer, J (1987) *Statistical Thermodynamics of Nonequilibrium Processes* (Springer), pp 437–438.
- [76] Sleeper, H, Orgel, L (1979) The catalysis of nucleotide polymerization by compounds of divalent lead. *J. Mol. Evol.* 12:357–264.
- [77] Bailey, J (2000) Chirality and the origin of life. *Acta Astronautica* 46:627 – 631.
- [78] Pizzarello, S, Cronin, JR (2000) Non-racemic amino acids in the murray and murchison meteorites. *Geochimica et Cosmochimica Acta* 64:329 – 338.
- [79] Blankenburg, S, Schmidt, WG (2007) Long-range recognition due to substrate locking and substrate-adsorbate charge transfer. *Phys. Rev. Lett.* 99.
- [80] Joshi, PC, Pitsch, S, Ferris, JP (2000) Homochiral selection in the montmorillonite-catalyzed and uncatalyzed prebiotic synthesis of RNA. *Chem Comm* 24:2497–2498.
- [81] Joshi, PC, Pitsch, S, Ferris, JP (2007) Selectivity of montmorillonite catalyzed prebiotic reactions of D,L-nucleotides. *Orig Life Evol Biosph* 37:3–26.
- [82] Russell, MJ, Arndt, NT (2005) Geodynamic and metabolic cycles in the Hadean. *Biogeosciences* 2:97–111.
- [83] Russell, MJ, Hall, AJ (2006) in *Evolution of early earth's atmosphere, hydrosphere, and biosphere: constraints from ore deposits*, eds Kesler, SE, Ohmoto, H (Geological Society of America).
- [84] Russell, MJ, Martin, W (2004) The rocky roots of the acetyl-coa pathway. *Trends in Biochemical Sciences* 29:358 – 363.
- [85] Bailey, SW, ed (1988) , *Reviews in Mineralogy and Geochemistry* (Mineralogical Society of America) Vol. 19.
- [86] Tsipursky, SI, Drits, VA (1984) The distribution of octahedral cations in the 2:1 layers of dioctahedral smectites studied by oblique-texture electron diffraction. *Clay Minerals* 19:177–193.
- [87] Ferris, J, Ertem, G (1993) Montmorillonite catalysis of RNA oligomer formation in aqueous solution. A model for the prebiotic formation of RNA. *J. Am. Chem. Soc.* 115:12270–12275.
- [88] Cygan, RT, Liang, JJ, Kalinichev, AG (2004) Molecular models of hydroxide, oxyhydroxide, and clay phases and the development of a general force field. *The Journal of Physical Chemistry B* 108:1255–1266.
- [89] Cramer, C (2004) *Essentials of Computational Chemistry* (John Wiley and Sons), 2nd edition.

- [90] Sekine, S, Nureki, O, Dubois, D, Bernier, S, Chenevert, R, Lapointe, J, Vassilyev, DG, Yokoyama, S (2003) ATP binding by glutamyl-tRNA synthetase is switched to the productive mode by tRNA binding. *EMBO J* 22:676–688.
- [91] Perona, JJ, Rould, MA, Steitz, TA (1993) Structural basis for transfer RNA aminoacylation by *escherichia coli* glutamyl-tRNA synthetase. *Biochemistry* 32:8758–8771.
- [92] Pyrkosz, AB, Eargle, J, Sethi, A, Luthey-Schulten, Z (2010) Exit strategies for charged tRNA from GluRS. *Journal of Molecular Biology* 397:1350 – 1371.
- [93] Rappe, A, Casewit, C, Colwell, K, Goddard III, W, Skiff, W (1992) UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* 114:10024–10035.

Appendix A

Supporting Information

Copies of all supporting information for this dissertation including, but not limited to, source code, data files, and laboratory notes have been deposited with Zan Luthey-Schulten, UIUC. Written laboratory notebooks were physically transferred and electronic materials archived to tape and then deposited.

Curriculum Vitae

DAMIEN MATHEW
Chemical Life Science Laboratory, A540, Mailbox 68-6
600 S Mathews Dr, Urbana, IL 61801
P:217-244-5983, F:217-244-3186
dmathew2@illinois.edu

ACADEMIC TRAINING

December 2010	University of Illinois at Urbana-Champaign, Urbana, IL Ph.D., Biophysics and Computational Biology Thesis: <i>Computational Studies of Origins of Life Scenarios</i> Advisor: Zaida (Zan) Luthey-Schulten
2003	University of Wisconsin, Madison, WI M.A., Physics
2000	University of North Dakota, Grand Forks, ND B.S. (Summa cum Laude), Major: Physics, Minor: Mathematics

RESEARCH EXPERIENCE

May 2006 – present	Graduate Research Assistant, University of Illinois Zan Luthey-Schulten Group <ul style="list-style-type: none">◦ Investigate physical properties of amino acids relating to their codon assignments, and relate this information to the origin and structure of the genetic code◦ Investigate catalytic properties of inorganic minerals relating to reactions leading to biochemistry. In particular, simulate the interaction of activated nucleotides with montmorillonite to ascertain atomic details behind the clay's catalysis of nucleotide polymerization.◦ Use quantum chemical methods to study the reaction path for charging tRNA inside an aminoacyl tRNA synthetase
June – August 2000	Summer Graduate Research Assistant, University of Wisconsin

PEER-REVIEWED PUBLICATIONS

- [3] **Damien Mathew** and Zaida Luthey-Schulten. “Influence of Montmorillonite on Nucleotide Oligomerization Reactions: A Molecular Dynamics Study,” *Orig Life Evol Biosph*, 40:303-317 (2010).
- [2] Tom Butler, Nigel Goldenfeld, **Damien Mathew**, and Zaida Luthey-Schulten. “Extreme genetic code optimality from a molecular dynamics calculation of amino acid polar requirement,” *Phys Rev E*, 79:06091 (2009).
- [1] **Damien Mathew**, and Zaida Luthey-Schulten. “On the physical basis of the amino acid polar requirement,” *J Mol Evol*, 66(5):519-528 (2008).

OTHER PUBLICATIONS

Andrew Magis, Ke Chen, **Damien Mathew**, John Eargle, Elijah Roberts, and Zan Luthey-Schulten. “Evolution of the ribosome at the molecular level,” *NSF Center for the Physics of Living Cells Summer School* (2009).

<http://www.scs.uiuc.edu/schulten/tutorials/ribosome>.

HONORS, AWARDS AND FELLOWSHIPS

2005 Graduate College Fellow, University of Illinois

2000 Phi Beta Kappa, Alpha of North Dakota

1999 Golden Key National Honor Society

1998 Bjerkaas Physics Scholarship, University of North Dakota

1996-2000 Anton and Anna M. Thompson Memorial Scholarship, University of North Dakota

1996 Presidential Freshman Scholar

TEACHING EXPERIENCE

Spring 2008 Teaching Assistant, CHEM470 Computational Chemical Biology, UIUC

Fall 2000 Teaching Assistant, Physics laboratory and discussion sections, University of Wisconsin

PROFESSIONAL EXPERIENCE

2003 – 2005 North Dakota State University, Fargo, ND

Laboratory Manager, Department of Coatings and Polymeric Materials