AN ANALYSIS OF ANCESTRAL SEQUENCE RESURRECTION IN THE CONTEXT OF GUANYLATE KINASE EVOLUTION

by

WILLIAM CAMPODONICO-BURNETT

A THESIS

Presented to the Department of Biochemistry
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

July 2014

An Abstract of the Thesis of

William Campodonico-Burnett for the degree of Bachelor of Arts
in the Department of Biochemistry to be taken July 2014

Title: An Analysis of Ancestral Sequence Resurrection in the Context of Guanylate Kinase Evolution

Approved: 

Kenneth E. Prehoda

Ancestral sequence resurrection (ASR) is an important tool for studying evolution on a molecular scale. The process takes a broad range of extant samples and, using sequence alignment and evolutionary prediction algorithms, determines the most likely sequence to have evolved into modern-day proteins. While ever-improving technologies allow for increasingly reliable predictions, it is impossible to prove whether a reconstruction is in fact the true ancestor. This project will analyze the fidelity of the ASR process in the context of the divergence of enzymatically inactive guanylate kinase-like binding domains and enzymatically active guanylate kinases from a common ancestor. A maximum likelihood ancestor has already been predicted, so by comparing relative enzymatic activity of this ancestor, a variety of mutants, Bayesian predictions, and extant enzymes, we will be able to assess the validity of ASR for this billion-year-old evolutionary event.

Acknowledgements

I would like to thank Professors Prehoda and Southworth and Dr. Hetrick for helping me to fully examine the intricacies of protein evolution on a molecular scale, and for providing a diverse set of perspectives for framing my research. I would also like to thank all the professors, faculty, and staff that have given me the foundation of knowledge and skills required to complete an undergraduate thesis, as well as my fellow researchers within the Prehoda Lab for their guidance, wisdom, and encouragement through my research. Finally, I would like to thank my family for their constant support and encouragement, and for providing the ability to attend this wonderful institution.

Table of Contents

Introduction 1

Background 2

General Information 2

Ancestral Sequence Resurrection (ASR) 6

Membrane Associated Guanylate Kinases (MAGUK) 9

Enzyme Kinetics 14

Results 20

Catalytic Turnover (k_{cat}) of Ancestral Guanylate Kinases 20

"Substrate Affinity" (K_M) of Ancestral Guanylate Kinases 21

Catalytic efficiency (k_{cat}/K_M) of Ancestral Guanylate Kinases 22

Discussion & Conclusion 24

Mean and Variance data show that k_{cat} and K_M are statistically comparable 24

Analysis of individual point mutations can provide insight into their negligible impact on overall catalytic efficiency 24

Kinetics 26

Conclusion 26

Materials and Methods 28

Plasmid construction, expression, and purification 28

Enzymatic activity coupled assay 29

Future Directions 31

Appendix A – Protein Concentrations 32

Appendix B: Compiled Enzyme Data 33

Appendix C: Reconstruction Facts and Values 34

Individual Mutations 34

Additional Protein Characteristics: Bayesian 35

Appendix D: Phylogenetic Tree 37

Works Cited 38

List of Figures

Figure 1: The (simplified) Central Dogma of Molecular Biology 2
Figure 2: Transcription and Translation 4
Figure 3: A General Overview of Ancestral Sequence Resurrection 8
Figure 4: Reaction Catalyzed by Guanylate Kinase Enzymes 10
Figure 5: MAGUK Domain Architecture 11
Figure 6: Characterization of the Likelihood of AncGK0 12
Figure 7: A Characteristic Michaelis-Menten Kinetics Plot 16
Figure 8: A Characteristic Sigmoidal Plot for an Allosteric Enzyme 18
Figure 9: Compiled Enzyme Turnover Data 21
Figure 10: Compiled Substrate Affinity Data 22
Figure 11: Compiled Catalytic Efficiency Data 23

Introduction

Approximately a billion years ago, a chromosomal duplication event occurred in a common ancestor of Choanoflagellates and Metazoa[1], creating two copies of the gene coding the guanylate kinase (GK) enzyme. Over the course of the next billion years, these initially identical genes diverged to develop unique functions which still exist today, one branch maintaining its enzymatic ability, the other losing kinase activity but simultaneously gaining a protein binding function. Through the process of ancestral sequence resurrection (ASR), the amino acid sequence of this ancient protein was statistically predicted to help elucidate the process by which this divergence occurred, and this paper follows a rigorous mutagenic analysis of the resurrection to determine its relevance to the guanylate kinase system.

[1] de Mendoza, 2010

Background

General Information[2]

This paper will be working with the expectation of a rudimentary knowledge of biological processes. While far from exhaustive, the following explanation will hopefully provide enough information to follow the study. At the core of all cellular processes lies the central dogma of molecular biology, which generally describes the flow of genetic information and for the purpose of this study can be simplified to say 'DNA leads to RNA leads to proteins'.

Figure 1: The (simplified) Central Dogma of Molecular Biology

In a simplistic view, this figure describes the central dogma of molecular biology, which generally states, "DNA produces RNA produces proteins". The figure also indicates that DNA can remake itself. It does not show reverse transcription, another process whereby DNA can be synthesized using and RNA template, nor does it consider a large variety of other RNA function.

While in actuality the theory is far more nuanced, this hits the following important points. Deoxyribonucleic acid (or DNA) serves as the information carrier of the cell, coding all a cell needs to survive and reproduce in countless different environments. It is the genetic material that is passed from generation to generation and ensures that life will continue beyond the span of a single organism's time. During the course of an organism's life, this genetic information is copied to ribonucleic acid (RNA) through the process of transcription, and this RNA then acts as the messenger, carrying

[2] Background information adapted from Voet & Voet

information to the ribosome. Once at the ribosome RNA is translated into proteins. As the words imply, transcription (conversion of DNA into RNA) is a process of copying – both DNA and RNA are composed of nucleic acids. DNA cannot leave the nucleus and RNA can, so the genetic information is converted to RNA to reach the cytoplasmic ribosomes. On the other hand, translation (as the name implies) takes the nucleotide sequence and changes it into amino acids, which are structurally significantly different from nucleotides.

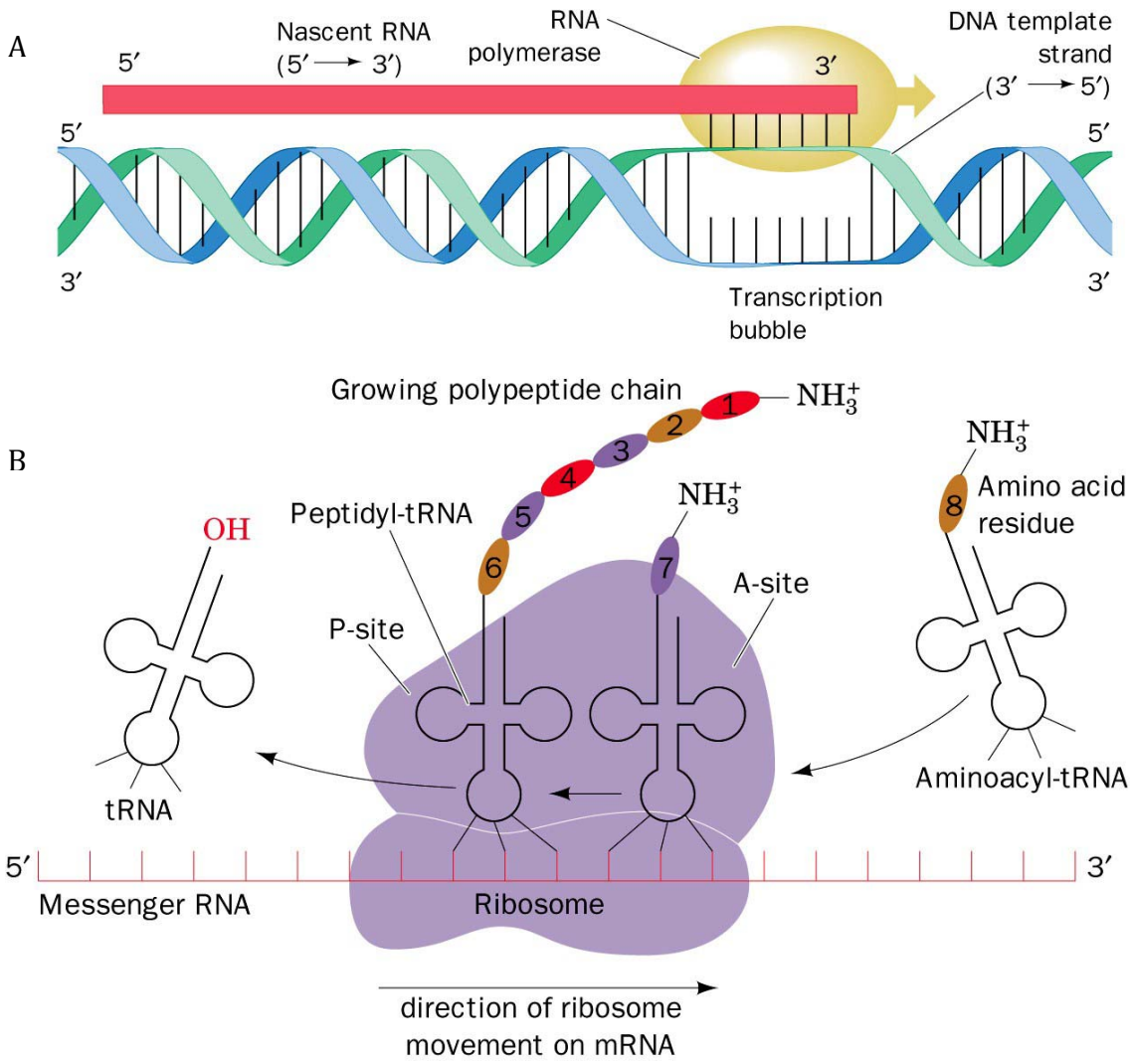


Figure 2: Transcription and Translation[3]

A. The flow of genetic information begins in the nucleus where DNA is transcribed into mRNA. This mRNA then leaves the nucleus and is transported to the ribosome, where B. it is translated into a protein chain. Three-nucleotide codons indicate which amino acid to add based on the mRNA sequence.

These proteins are the functional end of this genetic flow, and will carry out a nearly unimaginable range of functions depending on their makeup. To use a business analogy, DNA is high management, making decisions and passing them along to RNA,

[3]Voet & Voet

the middle management, which then carries orders and oversees the proteins performing the work itself.

As mentioned above, DNA and RNA are both chains of nucleic acids made up of sequences of five subtly different nucleotides – adenine, guanine (purines), thymine, cytosine, and uracil (pyrimidines). Proteins are made up of chains of 20 different natural amino acids, which allows these molecules much greater functional diversity than strings of nucleotides. During translation, unique groups of three nucleotides, known as codons, are read by the ribosome to represent individual amino acids – for example, a sequence of the three nucleotides guanine, adenosine, cytosine (GAC for short) would be translated into an aspartic acid amino acid while the subtly different ACC codon translates to a dramatically different alanine. A protein's function relies heavily on both the linear sequence of amino acids, known as primary structure, and the 3D shape into which it folds, or secondary and tertiary structures, which are largely dictated by the primary structure. At the core, this folding process is all about minimizing energy to create the most energetically stable structure, and this stability is largely determined by the chemical properties of substituent amino acids, including size, rigidity, pKa, and hydrophobicity. Folding also gives rise to functional domains within a protein, such as hydrophobic cores (regions of largely organic, non-polar amino acids that clump together to keep away from water) and enzymatic active sites (groups of amino acids that create chemical microenvironments which help catalyze a chemical reaction), to name a few. Different combinations of these amino acids, and the resulting shapes into which they fold, lead to the nearly unimaginable range of protein functions observed within even the least complex forms of life.

Ancestral Sequence Resurrection (ASR)

Ancestral Sequence Resurrection is a method of statistically predicting and creating the ancestors of extant proteins. ASR as a process was first hypothesized in the 1960's, but for several decades technology could not make it a viable reality. In the 1990's, the process was shown to be possible, but due to technological inadequacies was not truly practical until the turn of the 21st century[4]. Since then, the field has only improved, with more complicated statistical models, larger extant datasets (along with the computing power to process these data), and improved DNA synthesis techniques conspiring to make ASR a viable strategy for studying evolution.

How ASR works[5]

Ancestral sequence resurrection involves 5 main steps (Figure 3). First, a wide range of modern sequences for proteins descended from the ancestral construct are obtained and aligned. It is common throughout life for different organisms to have proteins descended from a common ancestor that perform identical functions but have subtle differences in primary sequence. By comparing a large set of these subtly different proteins, it is possible to develop a phylogenetic tree, which groups the modern sequences based on similarity (Appendix D). Next, a statistical model is used to predict the most likely ancestral sequence that could have evolved into the observed proteins. Every amino acid position within the protein is assigned a posterior probability, and a compilation of all individual predictions determines the overall likelihood of the total protein. Once an acceptable ancestor has been predicted, the gene

[4] Hanson-Smith, Kolaczowski, and Thornton, 2010.
[5] Adapted from Thornton, 2004 and Hanson-Smith, Kolaczowski, and Thornton, 2010.

must be synthesized by one of several methods – PCR, cloning, or mutagenesis could all produce the desired results. At this point, the DNA is often codon-optimized – while an amino acid can be coded by several different codons, organisms often translate different codons with different efficiencies, so building the most favorable sequence can increase the efficiency of later protein production. Fourth, this newly created gene is inserted into a DNA vector, transformed into a cell culture, and expressed at high levels. Finally, a protein that has not existed for over a billion years can be purified and subjected to a range of analyses. This basic framework was used to create a 188-residue ancestor for all extant GK enzymes and domains.

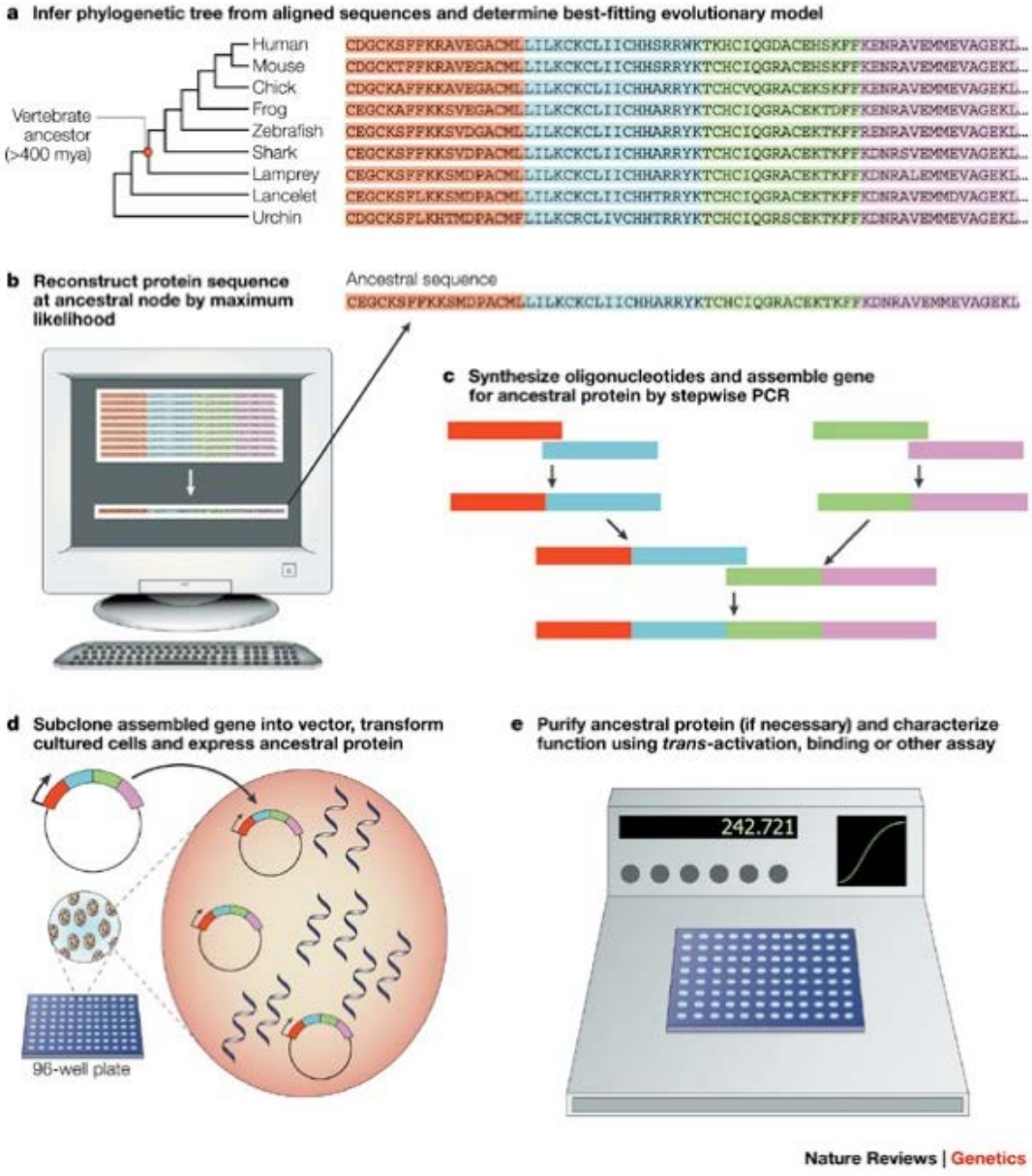


Figure 3: A General Overview of Ancestral Sequence Resurrection[6]

[6] Thornton, 2004

Ancestral Sequence Resurrection (ASR) generally follows the following process. A. A large diversity of extant sequences are gathered and aligned to produce a phylogenetic tree, showing the most likely evolutional divergence of the various sequences. B. Statistical models are used to predict the maximum likelihood ancestral sequence. C. This maximal likelihood ancestor is synthesized by PCR (or other methods). D. Newly created gene is transformed into competent cells, expressed, and purified. E. The newly synthesized protein can now be characterized by a variety of biochemical methods.

As mentioned above, ASR relies on a statistical method to recreate extinct proteins. Over the years, several methods have been used, beginning with the consensus method[7]. Under this framework, the most conserved sequence across examined extant species was assumed to be the ancestor. This system was heavily biased based on the modern-day proteins selected, and was replaced in the early 1980's by maximum parsimony (MP)[7]. By taking into consideration the phylogenetic tree, a more accurate sequence can be determined, eliminating the unconscious selection bias. In spite of its benefits over consensus, MP also had inherent weaknesses, and eventually maximum likelihood (ML) came to be the accepted norm for ASR[7]. The largest improvement is the ability of ML to consider a known evolutionary model in recreating nodes in a phylogenetic tree. This progress to ML prediction, along with modern advances in computing and gene synthesis, is primarily responsible for allowing ASR to excel.

Membrane Associated Guanylate Kinases (MAGUK)

The Membrane-Associated Guanylate Kinase (MAGUK) lineage is a superfamily of scaffolding proteins that rely heavily on the GK binding domain for

[7] Thornton, 2004

function[8]. This GK binding domain, along with extant guanylate kinase enzymes, evolved from a common ancestor that existed approximately a billion years ago. In contrast to the protein binding functions exhibited by the GK binding domain, both the ancestral protein and extant guanylate kinase enzymes are catalytically active, assisting with the transfer of an inorganic phosphate from ATP to GMP (Figure 4).

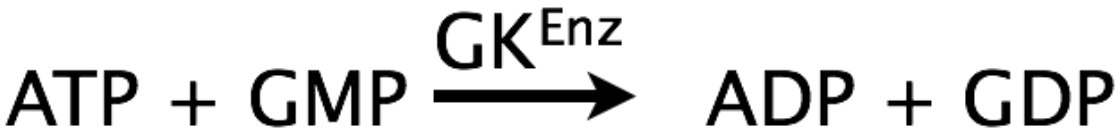


Figure 4: Reaction Catalyzed by Guanylate Kinase Enzymes

Both extant guanylate kinase enzymes and the ancestral resurrection AncGK0 are catalytically active. They are responsible for transferring an inorganic phosphate, PO_4^{-3} from ATP to GMP, producing ADP and GDP. Both compounds are important energy and metabolic control molecules in the cell.

Members of the MAGUK superfamily, characterized by the presence of PDZ, SH3, and GK binding domains (Figure 5), play important biological roles including mitotic spindle orientation, cell-cell interaction, synaptogenesis, and postsynaptic organization[9]. This project is positioned in a larger context of spindle orientation exploration; as improper orientation is a characteristic of cancer, MAGUKs are relevant to the study of the disease.[10]

[8] Funke, Dakoji, and Bredt, 2005
[9] Funke et al 2005, de Mendoza et al 2010, Olivia et al 2011
[10] Hoover et al 1998

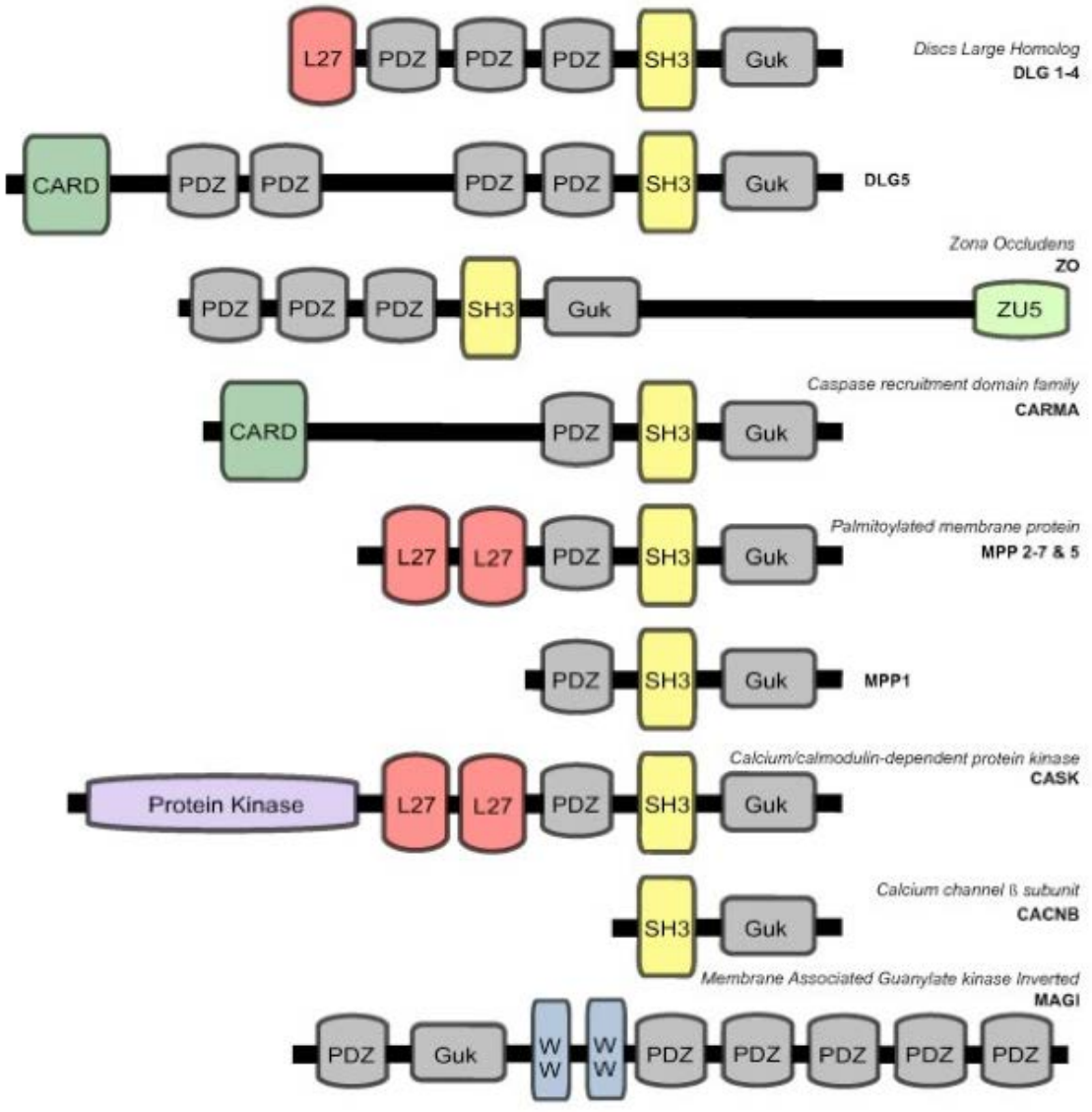


Figure 5: MAGUK Domain Architecture[11]

Membrane-Associated Guanylate Kinases are characterized by their domain architecture, including a series of PDZ repeats, and SH3 domain, and a guanylate kinase-like binding domain (Guk). Members of this family will also commonly contain L27 domains, another protein-protein interaction motif, as well as other functional domains.

[11] de Mendoza, 2010

As discussed above, it is impossible to predict an ancestral sequence with complete confidence. Though the most statistically likely guanylate kinase ancestor (in this study referred to as AncGK0) has a 93.73% chance of being the actual sequence - very well characterized for such an ancient gene – it is important to take into consideration the possibility that a different protein was actually the ancestor.

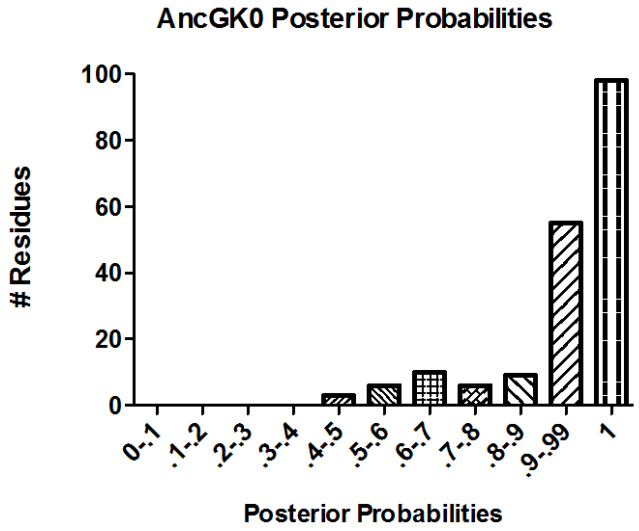


Figure 6: Characterization of the Likelihood of AncGK0

As can be seen, the majority of the amino acid positions within AncGK0 were predicted with 100% confidence, and a great excess over majority were predicted with greater than 90% confidence. However, there still exist a series of less well-predicted residues that represent ideal candidates for experimentation.

To demonstrate the fidelity of this resurrection, the enzymatic properties of a range of other possible ancestors were tested. Three increasingly stringent levels of variation were tested – single amino acid mutants of AncGK0 (hereafter referred to as the "point mutants"), compiled alternates (AltAll20%), and Bayesian constructs, a statistical method for generating a random sample of possible sequences (for a compilation of all mutations, see Appendix B).

The point mutants were designed to test sequences most similar to AncGK0. As mentioned above, each position within the protein has a statistical likelihood ranging from 0-100% confidence. While the maximum likelihood prediction had an overall likelihood of 93.73% of being the correct sequence, it contained 20 amino acids with a statistically significant possible alternate (where statistically significant is considered greater than 20% likelihood). Of these, 10 were randomly selected for further screening. These mutants, each with a single variation from AncGK0, were expressed and purified, then enzymatic activity compared to that of AncGK0. Demonstrating that these activities were within a statistically relevant range of AncGK0 activity would help show that, even if the ASR process failed to determine the exact sequence, a minor variation would not impact the overall relevance of the ancestral construct.

Point mutants provide compelling evidence for the strength of ASR, but it is valuable to extrapolate even further, since mutating one amino acid out of the 188 total gives an almost identical protein. Rather than single mutations, the next level of stringency was accomplished by the AltAll20% construct. This construct is the compilation of every alternate with at least 20% likelihood, therefore containing a total of 20 variations from AncGK0. Because of the compounding effect of decreased likelihood alternates, AltAll20% is several orders of magnitude less likely than AncGK0. Showing that this construct still has comparable enzymatic activity, despite its relative unlikelihood, again strengthens the argument that a slight inaccuracy in the AncGK0 sequence could still lead to an ancestor capable of evolving into both the GK enzyme and binding domain observed in extant organisms.

The final level of variation from AncGK0 consists of the Bayesian constructs. Bayesian inference was used to calculate the posterior probabilities of every possible sequence allowed by the phylogenetic tree used for ASR, and 5 sequences were randomly selected to represent extremely unlikely options (see Appendix B). To put this more simply, five proteins were selected from the compilation of every possible sequence consisting of any combination of any alternate, no matter how poorly predicted. Due to the extremely low probability that these compilations represent, all five Bayesian samples are extraordinarily unlikely, in spite of only having 5-13 mutations from AncGK0. However, this merely takes the trend to an extreme, showing that even a sequence that is between 10^{11} and 10^{26} times less likely than AncGK0 was a viable ancestor, capable of enzymatic activity and therefore evolution into extant species.

Enzyme Kinetics[12]

Enzymes represent an extraordinarily vital class of molecules for life. Of the countless chemical reactions occurring in the body at any given second, nearly all are mediated by enzymes. These molecules are biological catalysts and, like inorganic catalysts, primarily serve to speed chemical reactions within living organisms. Enzymes are far superior to their man-made inorganic counterparts, though, with higher reaction rates and greater specificity possible under less harmful reaction conditions. Without this added reaction speed, basic and necessary cellular processes would be impossible, making life as a whole impossible; needless to say, understanding enzyme function is important.

[12] Adapted from Voet & Voet

In studying enzyme kinetics, two important factors to take into consideration are turnover rate and binding affinity. Turnover rate is the more intuitive of the two; simply put, it is a measure of how many times an enzyme can catalyze a reaction in a given time and is generally reported as k_{cat}, with units of s^{-1}. Increasing turnover rate would increase catalytic efficiency, and this is one way enzymes can evolve to be superior.

In addition to turnover rate, binding affinity plays an important role in determining enzymatic properties. While an enzyme's ability to catalyze a reaction is important, enzyme turnover would be entirely irrelevant if the molecule could not bind its substrate in the first place. Binding affinity is typically reported as the dissociation constant K_d, which is an equilibrium constant for the dynamic binding and dissociation of an enzyme and its substrate. This measure of how tightly an enzyme can bind its substrate and its importance is twofold; an enzyme must be able to both effectively bind a substrate to allow catalysis and also release the product after the chemical reaction has occurred (if product release were not possible, enzymes would essentially be inactivated after one round of catalysis). Several different models of binding exist to help explain experimentally observed kinetic trends. The most basic model is described by the Michaelis-Menten equation, which describes initial velocity V_0 as a function of substrate concentration. This model assumes that the enzyme binds its substrate with a constant affinity to form the enzyme-substrate complex, which can then catalyze the reaction and release the product, regenerating the enzyme. Also, V_0 increases according to a predictable parabolic function, as seen in a representative Michaelis-Menten plot (Figure 7).

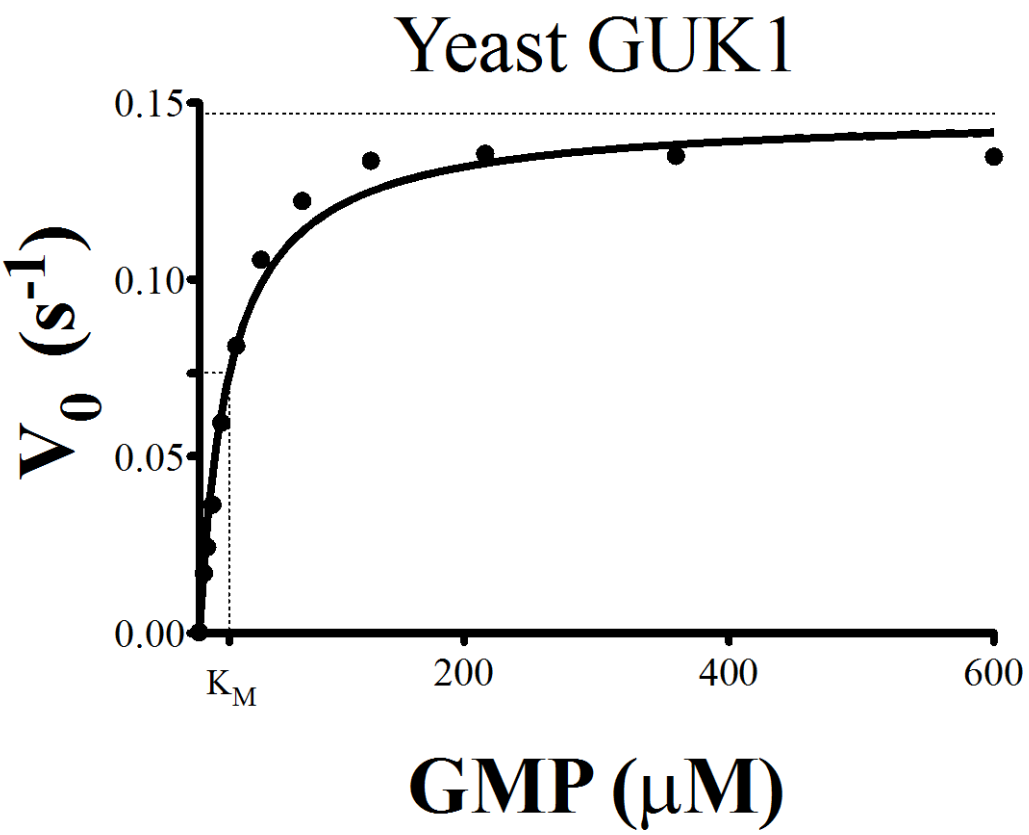


Figure 7: A Characteristic Michaelis-Menten Kinetics Plot

Standard Michaelis-Menten kinetics follow a rectangular parabolic function, steadily approaching a maximum velocity under saturated conditions. K_M, the substrate concentration required for half-maximal enzyme activity, is an important value for describing the enzyme's function.

An important value in this model is K_m, the Michaelis-Menten constant, which is defined as the concentration of substrate necessary for the enzyme to function at half its maximal efficiency. This constant is unique to an enzyme and is a common and effective way of reporting the binding of an enzyme to its substrate. It should be noted that K_M is *not* a true measurement of binding affinity (which needs an exclusive equilibrium between the bound and unbound enzyme states), but rather "is a dynamic or pseudo-equilibrium constant expressing the relationship between the actual steady-state

concentrations, rather than the equilibrium concentrations… Nevertheless, K_M represents a valuable constant that relates the velocity of an enzyme-catalyzed reaction to the substrate concentration"[13]. Thus, since all reactions in this work represent a steady-state condition rather than true enzyme-substrate association-dissociation equilibrium, K_M will be used as the metric to represent substrate affinity. It should also be noted that, unlike turnover rate, decreased values of K_M imply a "better" enzyme. Since K_M measures the substrate concentration necessary for half-maximal activity, having a lower value implies concentrations necessary for saturation and therefore higher affinity binding. Combined together, k_{cat} and K_M describe the efficiency of a given enzyme.

While the Michaelis-Menten model is a valuable and effective tool for simple enzyme systems, in practice many enzymes display a greater degree of complexity, meaning more complicated kinetics models are necessary. A primary source of this complexity is cooperativity. Michaelis-Menten operates under the assumption of a constant substrate affinity, leading to a generally linear increase in initial rate until the enzyme approaches saturation. With a cooperative system, however, an enzyme can bind a substrate at a non-catalytic site, leading to an allosteric change (change in shape) which impacts affinity at the catalytic site. Because of this cooperativity the graph deviates from the usual rectangular parabolic function seen in standard Michaelis-Menten kinetics, adopting instead a sigmoidal character (Figure 8). While the allosteric-sigmoidal curve does not give a true Michaelis-Menten constant as a measure of binding affinity, it is possible to calculate the substrate concentration necessary for

[13] Segel, 1976

half-maximal enzymatic activity, which still serves as an effective metric to describe the enzyme.

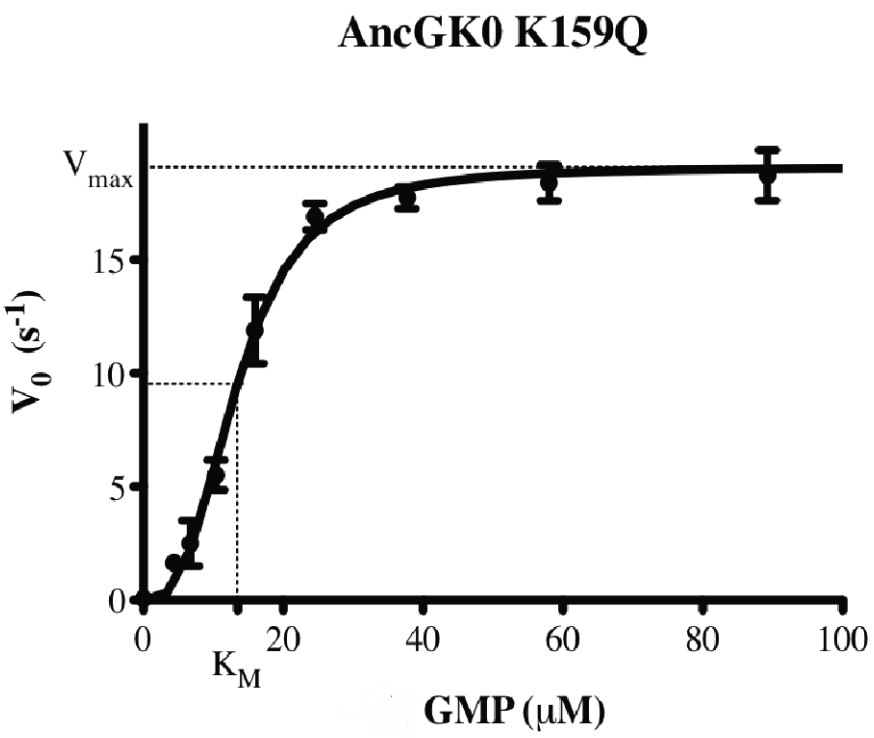


Figure 8: A Characteristic Sigmoidal Plot for an Allosteric Enzyme

While the sigmoidal plot has generally the same shape as a Michaelis-Menten plot, rising before plateauing under saturating conditions, there is an important difference at low substrate concentrations. The gradual acceleration phase before a more linear ascent is the defining characteristic of a sigmoidal regression for an allosteric enzyme. While the substrate concentration at half-maximal velocity is not truly a Michaelis-Menten constant, it is nonetheless a valuable metric for assessing an enzyme's substrate affinity.

It is important to recognize that neither K_M nor k_{cat} alone provides an accurate description of an enzyme's function; an enzyme with an extremely fast turnover rate and extremely poor substrate binding could easily be less efficient than an enzyme with average K_M and k_{cat} values. To overcome this discrepancy, it is common to report

enzyme characteristics in terms of catalytic efficiency, which is defined as k_{cat}/K_M and provides a normalized value to compare different enzymes. This value will be reported as the final comparison between various constructs.

Results

The modern-day human guanylate kinase enzyme had a catalytic turnover rate of $k_{cat} = 32.85s^{-1}$, $K_M = 19.14\mu M$, and a catalytic efficiency of $1.72\mu M^{-1}s^{-1}$. The maximum likelihood AncGK0 had a catalytic turnover rate of $k_{cat} = 7.24s^{-1}$, $K_M = 10.26\mu M$, and a catalytic efficiency of $0.71\mu M^{-1}s^{-1}$. This implies that AncGK0 is just above two-fold less efficient than extant enzymes, which is not surprising.

Catalytic Turnover (k_{cat}) of Ancestral Guanylate Kinases

Compared to the extant control (human GK), all ancestral resurrections have a lower catalysis rate, with k_{cat} ranging from $2.77s^{-1}$ to $25.73s^{-1}$, though all initial rates are well within an order of magnitude, making them comparable to modern day enzymes. The maximum likelihood construct AncGK0 falls almost exactly in the middle of all predictions with a k_{cat} of $7.24\ s^{-1}$. The point mutants generally cover a wider range than Bayesian constructs, and on average have a higher rate. AltAll20% falls towards the top of the ancestral resurrections with approximately a 3-fold increase over AncGK0 enzyme turnover. k_{cat} values are compiled in Figure 9.

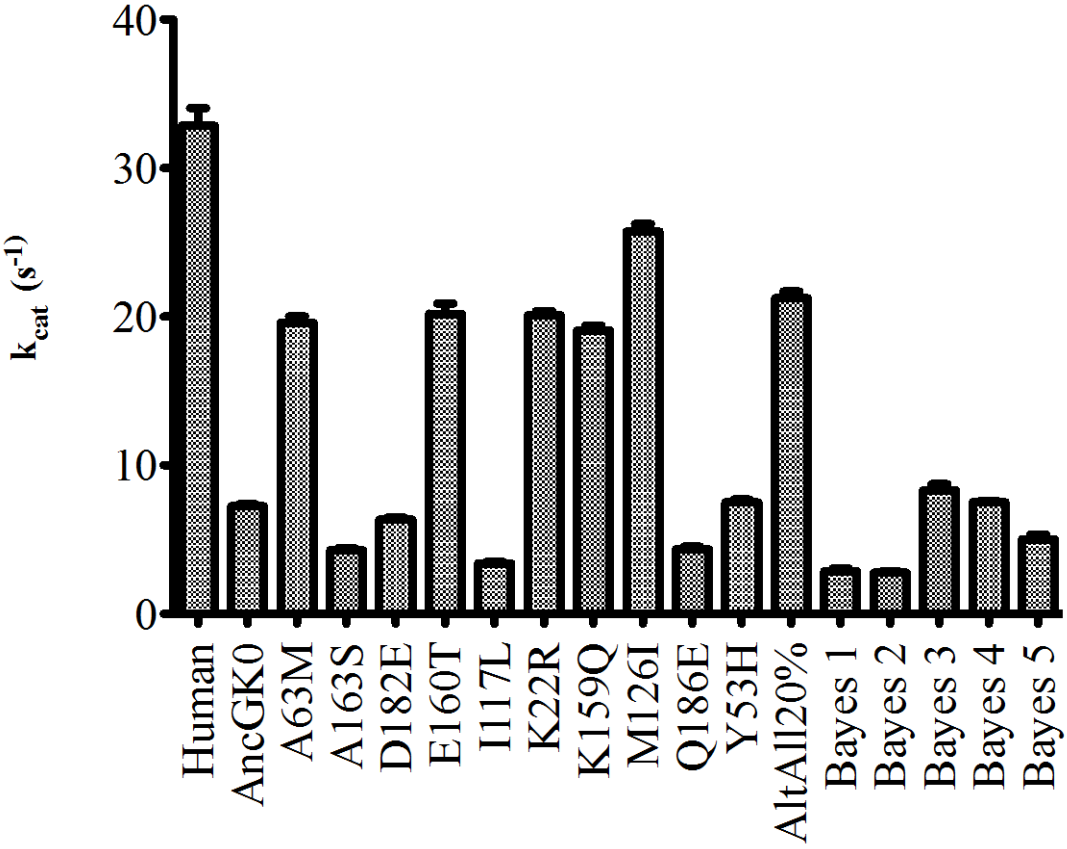


Figure 9: Compiled Enzyme Turnover Data

All k_{cat} values for the ancestral resurrections and extant control. Values ranged from $2.77s^{-1}$ to $25.73s^{-1}$.

"Substrate Affinity" (K_M) of Ancestral Guanylate Kinases

Looking at K_M values, AncGK0 had an increased but comparable affinity to human GK. Interestingly, all but three of the ancestral predictions have a higher substrate affinity than extant human GK. As with catalytic turnover, AncGK0 falls generally in the middle of the predictions, while AltAll20% has an affinity almost identical to human GK. Bayesian constructs have the largest range of values, representing both the best and worst binding. Substrate affinity data are compiled in Figure 10.

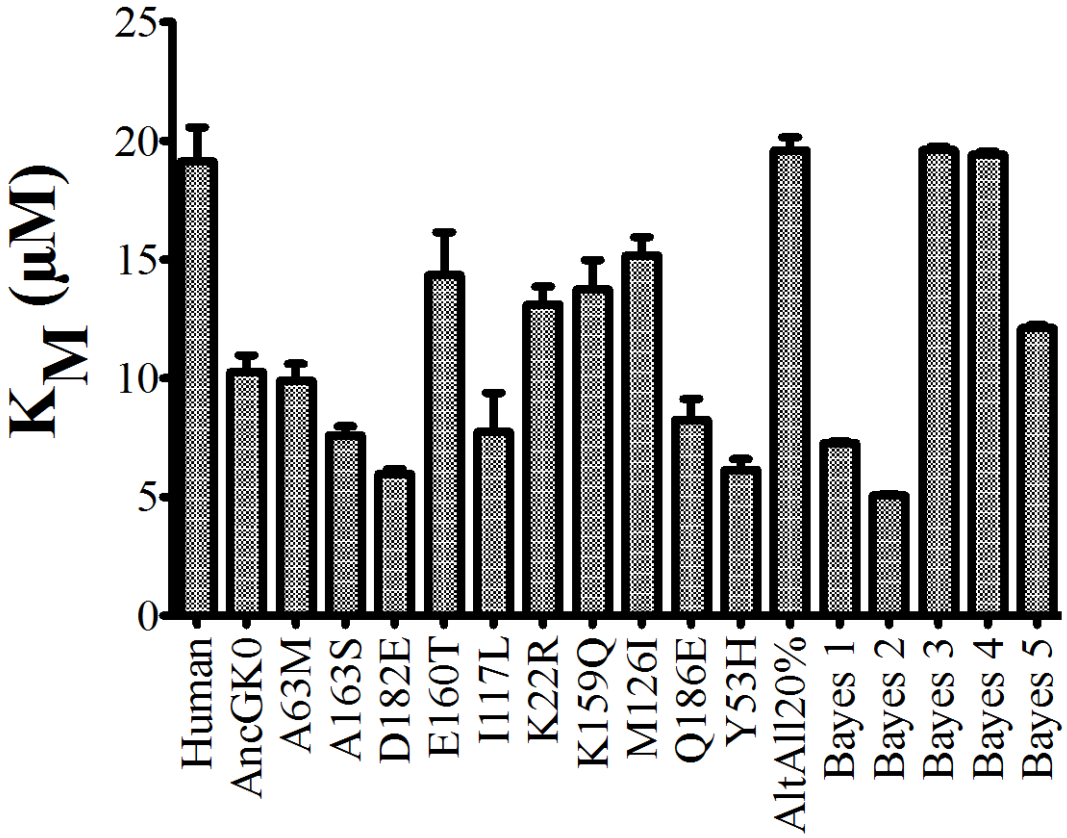


Figure 10: Compiled Substrate Affinity Data

All K_M values for the ancestral resurrections and extant control. Values ranged from 5.06µM to 19.60µM.

Catalytic efficiency (k_{cat}/K_M) of Ancestral Guanylate Kinases

As mentioned above, neither turnover rate nor substrate affinity alone provides an accurate picture of enzyme kinetics. By combining the two to find a ratio of k_{cat}/K_M, a more reliable representation of catalytic efficiency is established. Catalytic efficiencies ranged from $0.39\mu M^{-1}s^{-1}$ to $1.98\mu M^{-1}s^{-1}$, with AncGK0 again falling in the middle with a value of $0.71\ s^{-1}\mu M^{-1}$. Catalytic efficiency data are compiled in Figure 11.

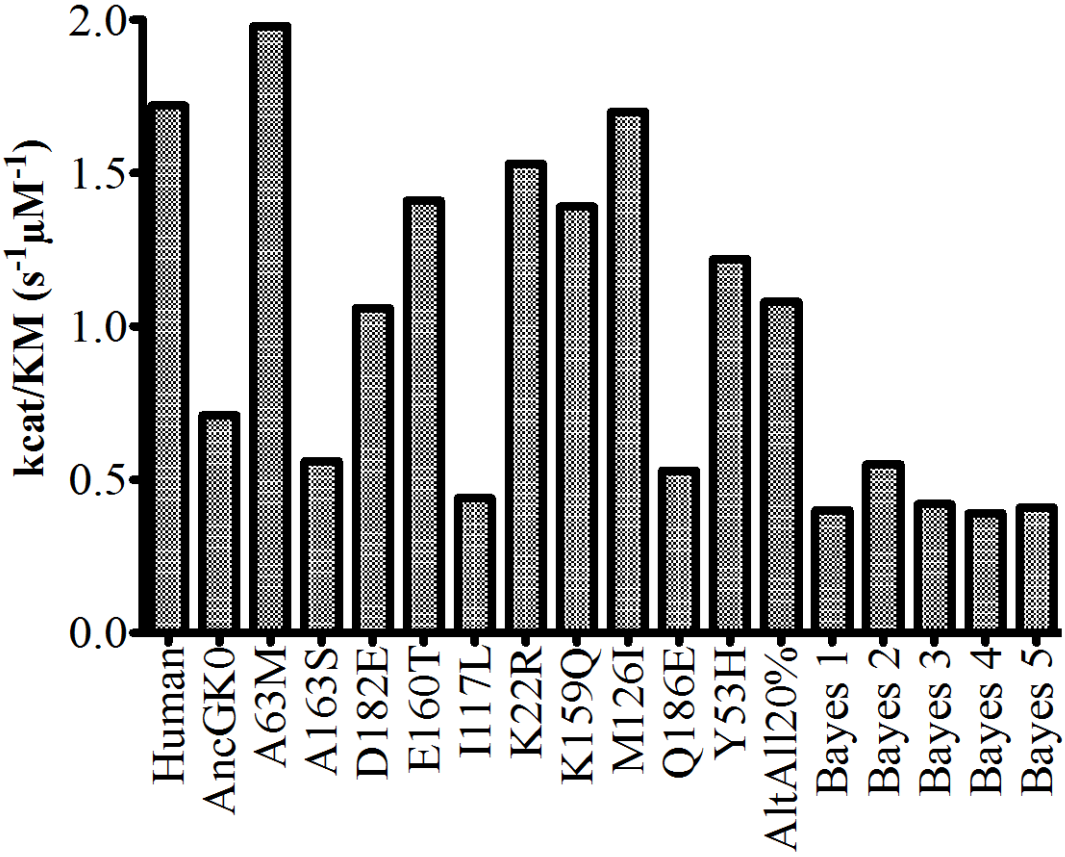


Figure 11: Compiled Catalytic Efficiency Data

Catalytic efficiency, k_{cat}/K_M, provides a more accurate way of comparing enzymes than does either k_{cat} or K_M individually. Values ranged from $0.39\mu M^{-1}s^{-1}$ to $1.98\mu M^{-1}s^{-1}$ for the ancestral resurrections and extant control.

Discussion & Conclusion

Mean and Variance data show that k_{cat} and K_M are statistically comparable

Upon first examination of the results, it appears as though there is significant variation both within the ancestral resurrections and when compared to the extant human GK. However, by subjecting these data to a t-test, it is possible to determine whether statistical importance can be assigned to the results. By creating three groups, "extant", "AncGK0, point mutants, & AltAll20%", and "Bayesian", it can be shown that the difference in the mean and variance (of both k_{cat} and K_M) of any given group is not statistically significantly different from any other. In other words, a t-test definitively demonstrates that all ancestral constructs are statistically comparable both to each other and to extant guanylate kinase activity.

In addition to statistical tests showing no significant difference between enzymatic activities of ancestral predictions, an allegorical example helps illustrate the insignificance of the variation within this group. Previous studies have shown that a single point mutation can take an enzyme with a k_{cat} of approximately $6s^{-1}$ and entirely destroy enzyme ability[14]. Compared to this dramatic change of function, the observed changes in catalytic efficiency are insignificant.

Analysis of individual point mutations can provide insight into their negligible impact on overall catalytic efficiency

A closer examination of the individual mutations will hopefully provide insight into the observed changes to affinity and turnover. To begin with the point mutants, of

[14] Johnston et al, 2011

the 10 mutations, five (A63M, I117L, M126I, D182E, K22R) represent changes within the same classification of amino acids (for example, hydrophobic to hydrophobic or negatively charged to negatively charged). Thus, it follows that such minor changes to protein primary structure would result in comparably minor changes in protein activity. The remaining five point mutants represent changes within classes (such as a hydrophobic residue being replaced by a positively charged one). However, all of these mutations are on the surface of the enzyme, and none represent changes *to* a hydrophobic residue, which on the surface could have more significantly changed stability.

While single point mutations can either increase or decrease catalytic efficiency, when these mutations are combined as in the case of AltAll20%, the compounded effect results in an enzyme with poor binding and above average turnover, culminating in a slightly above average catalytic efficiency when compared to all other constructs examined. Since the individual mutations that make up AltAll20% have both positive and negative effects on catalytic efficiency, it is unsurprising that the effects cancelled and this construct is within a reasonable range of AncGK0.

Finally, the Bayesian constructs contain mutations in all three domains of AncGK0, both on the surface and in the interior, and both changing and maintaining amino acid characteristics. Somehow a combination of all these mutations led to constructs with both increased and decreased affinity and velocity. Interestingly, Bayes 1&2 both had significantly lower k_{cat} and K_M values than the rest of their group, implying that, in this case, superior velocity was correlated to inferior binding ability.

To summarize, as expected mutations that would be expected to only slightly change protein structure led to correspondingly small changes in kinetic characteristics, leading to several groupings of ancestral constructs that are significantly similar to one another as well as extant guanylate kinase enzymes.

Kinetics

Given that extant guanylate kinase enzymes obey standard Michaelis-Menten kinetics, it was surprising to see cooperative binding in the ancestral reconstructions. While the ancestors could be forcibly fit with a rectangular parabola characteristic of Michaelis-Menten kinetics, there was clear cooperativity and the sigmoidal graph much better represented the data. However, this observed cooperativity could make sense when considering the age of AncGK0 and its mutants. Over the last billion years, extant guanylate kinase enzymes have had time to refine their binding to accept only select substrates. On the other hand, it would make sense for the ancestors to practice more promiscuous binding, as selectivity had not yet developed.

Conclusion

As ancestral sequence resurrection is nothing more than a method of targeted statistical prediction, it is impossible to say definitively that a given sequence prediction was in fact the protein that existed a billion years ago. Rather, at best we can say that we have predicted the most likely sequence based on currently existing proteins, but as extensively discussed, alternates are possible. The purpose of this study, then, is perhaps not to argue that the maximum likelihood prediction was in fact the one that existed, but rather that it does not matter whether our most likely prediction, or another

similar construct, was the true ancestor. By showing that the next 10 most likely sequences were all viable enzymatic ancestors, then stepping further back and showing that the AltAll20%, with a total of 20 mutations from maximum likelihood was still viable, and again taking things further and saying that the extremely poorly predicted Bayesian constructs were also viable, we have shown that it does not matter if the ML sequence was incorrect. We have successfully used ASR to predict viable ancestors with the potential to evolve into extant proteins.

Materials and Methods

Plasmid construction, expression, and purification

Once the ancestral amino acid sequence was obtained, the corresponding DNA was synthesized by Integrated DNA Technologies. The codons were optimized for E. Coli, the organisms used for expression. Proteins were ligated into pET-T7 vector (containing a poly-His tag) cut with BamHI/XhoI restriction enzymes. Plasmids were transformed into BL21(DE3) competent E. Coli cells then plated on LB agarose spiked with 400µg/mL ampicillin and grown overnight at 37°C. Bacteria were transferred to a 50 mL LB-Amp starter culture for one hour at 37°C. This culture was added to 2L LB-Amp and grown to an optical density (OD^{600}) of 0.6-0.8 at 37°C, then shifted to 18°C, induced with 500µL IPTG, and incubated overnight.

Protein samples were purified using Ni-NTA agarose affinity (Qiagen) followed by anion exchange chromatography with a Source 30 Q column (GE Life Sciences) after it was observed that the protein of interest was copurifying with an inhibitor when purified by Ni-NTA alone. After collecting and concentrating the protein, typical concentrations were in the range of 1-40mM (for individual concentrations see Appendix A). These proteins were stored at -80°C in buffer (20mM Tris pH 7.5, 10mM NaCl, 5mM DTT) to limit degradation. Protein purity was assessed by SDS-PAGE, and accuracy of protein concentrations (acquired by Bradford Assay) was determined by comparing Coomassie staining intensity to a known standard of BSA.

Enzymatic activity coupled assay

Direct measurement of guanosine diphosphate (GDP, the product of GK enzymatic activity) is difficult, time consuming, and requires radioactivity, so a creative way to measure GK catalysis more quickly, efficiently, and safely was desired. A coupled assay was developed to create a measurable system. When initiated by addition of GMP, guanylate kinase catalyzes the transfer of inorganic phosphate from ATP to GMP, producing GDP and ADP:

$$ATP + GMP \rightarrow ADP + GDP$$

Pyruvate kinase then transfers a phosphate from excess phospho(enol)pyruvate to ADP, regenerating the reaction starting material ATP:

$$PEP + ADP \rightarrow ATP + Pyruvate$$

Finally, lactate dehydrogenase catalyzes the conversion of this newly generated pyruvate to lactate, oxidizing NADH to NAD+ along the way:

$$Pyruvate + NADH \rightarrow Lactate + NAD+$$

Depletion of NADH can be spectroscopically monitored by measuring absorbance at λ=340nm. Since all other reagents are in excess and all other enzymes are saturated, the rate of NADH disappearance is directly correlated to the rate of GDP production, which gives GK catalytic activity. By running the assay with a titration of GMP, it is possible to produce a concentration dependence curve showing both catalytic rate (V_{max}) and substrate binding efficiency (K_M). By normalizing V_{max} according to enzyme concentration, k_{cat} and K_M are finally obtained and can be reported. All enzyme assays were performed in triplicate to determine standard deviation. Background enzymes

(pyruvate kinase and lactate dehydrogenase) were purchased from Sigma Aldrich. Enzyme concentrations were assayed between 10-500nM, with GMP concentrations spanning 0-500μM. Experiments were performed on a Tecan Safire^2 microplate reader at 30°C and a pH of 7.5. Measurement began upon addition of GMP and absorbance was measured every 15 seconds for a total of 30 cycles.

Future Directions

As repeatedly mentioned, the ancestral resurrections are merely statistical predictions. We see comparable activity amongst all mutants surveyed, the *in vitro* assays thus offering proof that AncGK0 can functionally catalyze GDP production as expected in the ancestral enzyme, and any errors in the ancestral resurrection are insignificant relative to the protein's function. Transitioning to a live cell assay would prove the small differences observed in *in vitro* analysis are not functionally significant, and that the constructs are functionally viable as enzymes. Additionally, while this study extensively characterized AncGK0, there are more recent evolutionary nodes within the phylogenetic tree that could be characterized in a similar manner. Finally, the cooperativity in the ancestral resurrections was well established but is not understood. Exploring the mechanism of this cooperativity, and the loss of cooperativity in extant enzymes, could provide further insight into the evolution of the guanylate kinase binding domain.

Appendix A – Protein Concentrations

Full Name	Concentration (mg/mL)	Concentration (mM)
Human	**2.37 mg/mL**	98.3 μM
AncGK0	**2.961 mg/mL**	**0.0505 mM**
AncGK0 A63M	**45.28 mg/mL**	**2.12 mM**
AncGK0 A163S	**9.06 mg/mL**	**0.424 mM**
AncGK0 D182E	**4.16 mg/mL**	**0.173 mM**
AncGK0 E160T	**34.39 mg/mL**	**1.43 mM**
AncGK0 I117L	**69.86 mg/mL**	**3.27 mM**
AncGK0 K159Q	**32.20 mg/mL**	**1.34 mM**
AncGK0 K22R	**34.70 mg/mL**	**1.44 mM**
AncGK0 M126I	**33.05 mg/mL**	**1.37 mM**
AncGK0 Q186E	**25.31 mg/mL**	**1.18 mM**
AncGK0 Y53H	**3.86 mg/mL**	**0.0161 mM**
AncGK0 AltAll20%	**428.3 mg/mL**	**19.98 mM**
AncGK0 Bayesian 1	**131.18 mg/mL**	**5.49 mM**
AncGK0 Bayesian 2	**362.4 mg/mL**	**15.06 mM**
AncGK0 Bayesian 3	**195.5 mg/mL**	**8.07 mM**
AncGK0 Bayesian 4	**241.5 mg/mL**	**10.04 mM**
AncGK0 Bayesian 5	**171.23 mg/mL**	**21.03 mM**

Appendix B: Compiled Enzyme Data

Full Name	Turnover Rate (k_{cat})	Substrate Affinity (K_M)	Catalytic Efficiency (k_{cat}/K_M)
Human	32.85 s^{-1}	19.14 µM	1.72 s^{-1} μM^{-1}
AncGK0	7.24 s^{-1}	10.26 µM	0.71 s^{-1} μM^{-1}
AncGK0 A63M	19.59 s^{-1}	9.90 µM	1.98 s^{-1} μM^{-1}
AncGK0 A163S	4.292 s^{-1}	7.60 µM	0.56 s^{-1} μM^{-1}
AncGK0 D182E	6.343 s^{-1}	5.97 µM	1.06 s^{-1} μM^{-1}
AncGK0 E160T	20.19 s^{-1}	14.34 µM	1.41 s^{-1} μM^{-1}
AncGK0 I117L	3.401 s^{-1}	7.74 µM	0.44 s^{-1} μM^{-1}
AncGK0 K159Q	20.09 s^{-1}	13.11 µM	1.53 s^{-1} μM^{-1}
AncGK0 K22R	19.08 s^{-1}	13.74 µM	1.39 s^{-1} μM^{-1}
AncGK0 M126I	25.73 s^{-1}	15.17 µM	1.70 s^{-1} μM^{-1}
AncGK0 Q186E	4.364 s^{-1}	8.25 µM	0.53 s^{-1} μM^{-1}
AncGK0 Y53H	7.507 s^{-1}	6.15 µM	1.22 s^{-1} μM^{-1}
AncGK0 AltAll20%	21.25 s^{-1}	19.60 µM	1.08 s^{-1} μM^{-1}
AncGK0 Bayesian 1	2.88 s^{-1}	7.28 µM	0.40 s^{-1} μM^{-1}
AncGK0 Bayesian 2	2.77 s^{-1}	5.06 µM	0.55 s^{-1} μM^{-1}
AncGK0 Bayesian 3	8.31 s^{-1}	19.61 µM	0.42 s^{-1} μM^{-1}
AncGK0 Bayesian 4	7.52 s^{-1}	19.42 µM	0.39 s^{-1} μM^{-1}
AncGK0 Bayesian 5	5.02 s^{-1}	12.11 µM	0.41 s^{-1} μM^{-1}

Appendix C: Reconstruction Facts and Values

Individual Mutations

AncGK0 primary sequence:

MAPRPVVLSGPSGSGKSTLLKKLLKEFPDEFGFSVSHTTRKPRPGEVNGK
DYYFVTREEMEQAIEKGEFIEHAEFSGNLYGTSKKAVQDVQSQGKICILD
IDMQGVKNIKKTDLNPIYIFIQPPSMEELEKRLRGRGTETEESLQKRLATA
KEEMEYGKEPGAFDHIIVNDDLEKAYEELKDFIIQEK

AltAll20% primary sequence (mutations highlighted):

MAPRPVVLSGPSGSGKSTLLKRLFKEFPDEFGFSVSHTTRKPRPGEVNGK
DYHFVTREEMEQMIEKGEFIEHAEFSGNLYGTSKKAVQDVQSQGKICIL
DIDMQGVKQIKKTDLNPLYIFIQPPSIEELEKRLRGRGTETEESLQKRLAA
AREEMEYAQTPGSFDHVIVNDDLDKAYEKLKEFIMEEI

Bayes 1 primary sequence (mutations highlighted):

MASRPVVLSGPSGSGKSTLLKKLLKEFPDEFGFSVSHTTRKPRPGEVNGK
DYYFVTREEMEEAIEKGEFIEHAEFSGNLYGTSKKAVRDVQAQGKICILD
IDMQGVKNIKKTDLNPIYIFIQPPSMEELEKRLRGRGTETEESLQKRLAA
AREEMEYGKEPGSFDHIIVNDDLEKAYEELKDFIIQEK

Bayes 2 primary sequence (mutations highlighted):

MAPRPVVLSGPSGSGKSTLLKKLLKEFPDEFGFSVSHTTRKPRPGEVNGK
DYYFVTREEMERAIKKGEFIEHAEFSGNLYGTSKKAVQDVQSQGKICIL
DIDMQGVKNIKKTDLNPIYIFIQPPSMEELEKRLRGRGTETEESLQKRLAT
AKEEMEYGKKPGAFDHIIINDDLEKAYEELKDFIVQEK

Bayes 3 primary sequence (mutations highlighted):

MAPRPVVLSGPSGSGKSTLLKKLFQEFPDKFGFSVSHTTRKPRPGEVNG

KDYYFVTREEMEQAIKKGEFIEYAEFSGNLYGTSKKAVQDVQAQGKICI

LDIDMQGVKNIKKTDLNPIYIFIQPPSMEELEKRLRGRGTETEESLRKRLA

TAREEMEYGKTPGAFDHIIVNDDLEKAYEKLKDFIIEEK

Bayes 4 primary sequence (mutations highlighted):

MAPRPVVLSGPSGSGKSTLLKRLFKEYPDEFGFSVSHTTRKPRPGEVNG

KDYHFVTREEMEQAIEKGEFIEHAEFSGNLYGTSKKAVQDVQSQGKICIL

DIDMQGVKQIKKTDLNPLYIFIQPPSIEELEKRLRGRGTETEESLQKRLAA

AKEEMEYSKTPGAFDHVIVNDDLEAAYDELKDFIIQEK

Bayes 5 primary sequence (mutations highlighted):

MAPRPVVLSGPSGSGKSTLLKRLLKEFPDEFGFSVSHTTRKPRPGEVNGK

DYHFVTREEMEQAIEKGEFIEWAEFSGNLYGTSKKAVQDVQSNGKICIL

DIDMQGVKNIKKTDLNPLYVFIQPPSMEELEKRLRGRGTETEESLQKRLA

TAKDEMEYGKEPGAFDHIIVNDDLDKAYEELKEFIIDEK

Additional Protein Characteristics: Bayesian

AncGK0 Bayes1

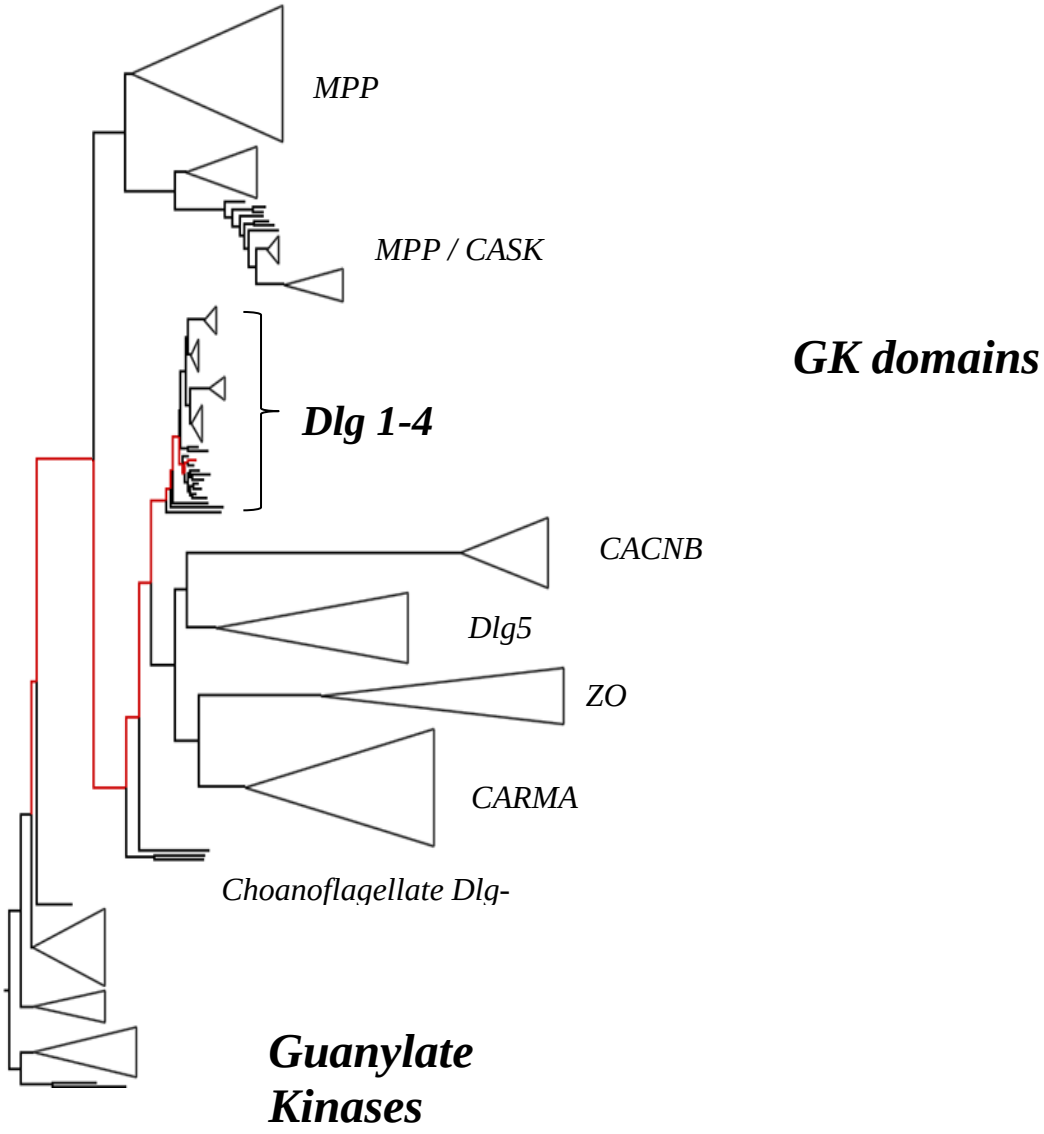
Log Likelyhood: -25.328
ML: -14.530
Avg PP: 0.918
changes: 7

AncGK0 Bayes4

Log Likelyhood: -34.914
ML: -14.530
Avg PP: 0.909
changes: 13

AncGK0 Bayes2 Log Likelyhood: -28.087 ML: -14.530 Avg PP: 0.921 # changes: 5	AncGK0 Bayes5 Log Likelyhood: -36.696 ML: -14.530 Avg PP: 0.910 # changes: 10
AncGK0 Bayes3 Log Likelyhood: -29.897 ML: -14.530 Avg PP: 0.910 # changes: 11	

Appendix D: Phylogenetic Tree



Works Cited

Funke L., Dakoji S., Bredt D. (2005). Membrane-Associated Guanylate Kinases Regulate Adhesion and Plasticity at Cell Junctions. Annu. Rev. Biochem. 74, 219-245.

Hanson-Smith, V., Kolaczkowski, B., Thornton, J. (2010). Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. Mol. Bio. Evol. 27(9), 1988-1999.

Hoover, KB., Liao SY., Bryant PJ. (1998). Loss f the night junction of MAGUK ZO-1 in Breast Cancer: The Relationship to Gladular Differentiation and Loss of Heterzygosity. Am J Pathol. 153(6), 1767-1773.

Johnston, C., Whitney, D., Volkman, B., Doe, C., Prehoda, K. (2011). Conversion of the Enzyme Guanylate Kinase into a Mitotic-spindle Orienting Protein by a Single Mutation that Inhibits GMP-Induced Closing. PNAS, 108(44), 973-8.

de Mendoza, A., Suga, H., Ruiz-Trillo, I. (2010). Evolution of the MAGUK Protein Gene Family in Premetazoan Lineages. BMC Evol. Biol. 10:93.

Olivia, C., Escobedo, P., Astorga, C., Molina, C., Sierralta, J. (2011). Role of the MAGUK Family in Synapse Formation and Function. Dev. Neurobiol. 72(1), 57-72.

Segel, I. *Biochemical Calculations* (Wiley, New York, 1976)

Thornton, J. (2004). Resurrecting Ancient Genes: Experimental Analysis of Extinct Molecules. Nature Reviews Genetics 5, 366-375.

Voet & Voet, *Biochemistry* (Wiley, New York, ed. 4, 2011)