Bioinformatics-Driven Enzyme Engineering: Work On Adenylate Kinase

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By
Nicholas Callahan, B.A.
Biophysics Graduate Program

The Ohio State University
2015

Dissertation Committee:
Dr. Thomas Magliery, Advisor
Dr. William Ray, Co-Advisor
Dr. Ralf Bundschuh
Abstract

Despite the large databases of protein sequence and structure data presently available, predicting the effect of mutations on an enzyme is still a major challenge. Isolating and interpreting evidence in sequence alignments of major evolutionary events is one way to probe the chemical logic of enzymes' functional plasticity. In this thesis, I outline work on predicting mutations of structural significance in proteins using bioinformatic techniques and the characterization of said mutations in functional enzymes. For this work, I use the well-studied enzyme adenylate kinase (ADK) as my model system.

In Chapter 1, I review the history and current research in bioinformatics-driven mutation prediction and chemical, biological, and biophysical aspects of adenylate kinase. Chapter 2 presents a series of bacterial adenylate kinase constructs which demonstrate key transitional variants, predicted from an alignment of ADKs, bridging functional ADK from B. subtilis and E. coli. Significantly, characterization of these variants has found necessary pre-conditions for the evolutionary elimination of a zinc binding site from B. subtilis ADK that is otherwise required for function. Chapter 3 presents the algorithmic implementation of an improved protocol for the prediction of consensus mutations, previously discussed by Dr. Brandon Sullivan and Dr. Thomas Magliery, and several consensus constructs of adenylate kinase, triosephosphate isomerase, and the DNA binding domain of fructose repressor FruR. Many of these
constructs show thermal stabilization and native-like function, revealing the robustness of the new protocol. Additionally, I analyze of the common features across both our successful and unsuccessful constructs. Chapter 4 outlines preliminary work towards the further biophysical characterization of the ADK transitional variants discussed in chapter 2. Chapter 5 presents our attempts to produce constructs from key features of the human dp71 gene product, for the purpose of characterizing mutations implicated in Duchenne muscular dystrophy.
Acknowledgements

I would like to acknowledge my advisors and primary planners on my projects - Dr. Thomas Magliery and Dr. William Ray; my co-workers - David Bowles, Dr. Adeline Vulin-Chaffiol, Dr. Venuka Durani-Goyal, Deepti Mathur, Sidharth Mohan, Dr. Deepa Perera, Dr. Eva Petrik, Kimberly Stephany, Dr. Brandon Sullivan, and Jendy Weppler; and those who’s advice helped me – Dr. Kiran Doddapaneni, Tran Nguyen, Dr. Lihua Nie, Dr. Brinda Ramasubramanian, and Dr. Shiladitya Sen. I would also like to acknowledge the technicians of the Magliery Lab – Sarah Johnston, Srividya Murali, and Kimberly Stephany – whose day-to-day assistance was of immeasurable value to my research. Additionally, I would like to acknowledge Dr. Chunhua Yuan of the OSU Campus Chemical Instrument Center for the collection of NMR data.
Vita

2003 .................................................................St. Ignatius High School

2007 .................................................................B.A., Case Western Reserve University

2005 – 2006 ....................Technician, Shoham Lab, CWRU Dept. of Biochemistry

2007 .................................................................Technician, Bzik Lab, Dartmouth Medical Center

2009 – Present ...................Graduate student, Magliery Lab, OSU Dept. of Chemistry

2010-2011 ..........................................................TA, OSU Dept. of Chemistry

2007 .................................................................Cum Laude, CWRU

2009-2010 .............................University Fellowship, OSU Graduate Program

2012 .........................Travel Award, Computational Biophysics to Systems Biology conference

Fields of Study

Major work: Biophysics
Table of Contents

Abstract .......................................................................................................................... ii
Acknowledgements ....................................................................................................... iv
Vita ................................................................................................................................. v
Table of Contents .......................................................................................................... vi
List of Tables .................................................................................................................. vii
List of Figures ............................................................................................................... viii
Chapter 1: Introduction ................................................................................................. 1
Chapter 2: Analysis of transitional variants in the ADK Lid domain ......................... 51
Chapter 3: Improvements in Consensus Mutant Prediction ........................................ 95
Chapter 4: Biophysical studies of ADK transition variants ...................................... 131
Chapter 5: Purification of Dystrophin gene products .................................................. 141
Appendix A: Selection scores for consensus constructs ............................................ 124
Appendix B: Construct sequences with secondary structure .................................... 128
Appendix C: Thermal unfolding curves ....................................................................... 130
Appendix D: Kinetic measurements ............................................................................ 133
Appendix E: Comparison of selection scores ............................................................... 134
References .................................................................................................................... 153
List of Tables

Table 1. Popular Scoring Methods for Coevolution Detection……………………………7

Table 2.1 Explanation of variant ADK constructs…………………………………….…76

Table 2.2 Melting temperatures and activity of ADK constructs……………………..…85

Table 3.1 Sources and effects of consensus mutants…………………………………...104
List of Figures

Figure 1.1 Evolutionary transitions in sequence space ..............................................2
Figure 1.2 Example of coevolution detection ..............................................................4
Figure 1.3 Grouping of correlation residues from Halabi et. al (reprint with permission) ........................................................................................................9
Figure 1.4 ATP/AMP phosphotransfer mechanism ....................................................13
Figure 1.5 Reaction mechanism for ADK .................................................................13
Figure 1.6 Domains of ADK .....................................................................................15
Figure 1.7 Binding of ADK and AP₅A ..................................................................19
Figure 1.8 Hinge regions of ADK .........................................................................19
Figure 1.9 Reaction pathway of ADK from Åden et al. (reprint with permission) ......21
Figure 1.10 Substates of ADK from Haran et al. (reprint with permission) ..........27
Figure 1.11 Folding pathway of ADK from Takada et al. (reprint with permission) ....28
Figure 1.12 Variations of the ADK Lid domain .......................................................38
Figure 1.13 Phylogeny of ADK from Leipe et al. (reprint with permission) ..........40
Figure 2.1 Alignment of bsADK and ecADK structures .........................................50
Figure 2.2 Residues of interest in the Lid domain ..................................................50
Figure 2.3 Coupled reactions for kinetic measurements .......................................57
Figure 2.4 Occupancy of ADK alignments ..............................................................60
Figure 2.5 Alignment of bsADK and ecADK sequences ...........................................62
Figure 2.6 Conservation at ADK positions .............................................................63
Figure 2.7 Correlation between ADK positions .....................................................65
Figure 2.8 Unique residue combinations found in Lid domain ..............................67
Figure 2.9 Sequence logo of Lid variants ...............................................................68
Figure 2.10 Comparison of Lid and AMPbd interface for bsADK and ecADK
Figure 2.11 MRE for bsADK, ecADK, bsDI, bsHexa, bsTetra, ecCys-4
Figure 2.12 Unfolding curves for bsADK, ecADK, bsDI, bsHexa, bsTetra, ecCys-4
Figure 2.13 $^1$H-$^{15}$N HSQC spectra for bsADK and bsTetra
Figure 2.14 Kinetic curves for bsADK, ecADK, bsDI, bsHexa, bsTetra, ecCys-4
Figure 2.15 MRE and unfolding curves for bsPenta-135 and bsPenta-155
Figure 2.16 Kinetic curves for bsPenta-135 and bsPenta-155
Figure 2.17 Kinetic curves for interface variants
Figure 3.1 Conservation and correlation scores of "Algo" TIM construct
Figure 3.2 GAP-dependent activity of ecTIM-bac
Figure 3.3 Illustration of consensus mutants made at ecTIM dimer interface and ecADK hinge regions
Figure 3.4 $^1$H-$^{15}$N HSQC and unfolding/refolding curves of FruR-bac
Figure 3.5 Types of mutations seen in consensus constructs
Figure 3.6 Phylogenetic trees of TIM and ADK alignments
Figure 4.1 Thermal unfolding curves of ADK variants in urea
Figure 4.2 Unfolding of ADK variants at 25 °C
Figure 4.3 Effects of urea on ADK melting temperature
Figure 4.4 Crystallization of ADK variant
Figure 5.1 Purification of dp71 fusion proteins
Figure 5.2 Circular dichroism of dp71 over temperature range
Figure 5.3 ANS binding of dp71 fusion proteins
Figure 5.4 Refolding dp71-Zf fragment
Figure 5.5 Circular dichroism of dp71-Zf fragment
Chapter 1: Introduction

The biochemical properties of a protein stem from the atomic topology of its folded state. As Anfinsen demonstrated, a protein will assume a fold which minimizes its free energy and the identity of this fold will depend on the unique amino acid sequence of the protein. However, comparison of sequence homology between proteins with similar folds and similar function shows that homologous proteins often diverge in sequence. For a given protein family, can we say these diverse sequences are “equally fit”? When Bornberg-Bauer et al. overlaid predicted stabilities on a neural network of related sequences, they found that the more stable variants clustered near the center of the network, where mutation-tolerant variants were found, while destabilized, mutation-intolerant variants appeared at the fringe. This suggests that the space of possible protein sequences could be described as a rugged “terrain” with valleys that represent optimal fitness and span diverse coordinates. Proteins would traverse this sequence space through chemically conservative mutations, as per Dayhoff’s model, with a bias towards “low terrain,” until a shift in fitness requirements changes the terrain. Povolotskaya and Kondrashov recently demonstrated through the statistical analysis of protein phylogeny that the rate of protein evolution appears to be consistently linear, rather than reaching a plateau, a behavior suggesting unsaturated sequence space. In this case, there could be an unknown number of novel sequences that fall within the fitness minima for a given
Fig. 1.1 Evolutionary transitions in sequence space

The starting sequence comes under selective stress.

The sequence relieves stress through a random walk of mutations, until fitness is achieved.

Analysis of available sequences allows us to hypothesize about the collection of fit sequences and the likeliest intermediate forms.
fold and so could serve to inform us about fold mechanics and fitness requirements beyond what is apparent in natural sequences.

Section 1: Correlation approaches to protein engineering

In the absence of a method to predict folding with absolute certainty and computational economy, it has been profitable to design novel proteins based on accumulated sequencing, structural, and functional data. Sequence-based design holds the appeal of large volumes of available data and higher relative speed compared to other methods. Consensus methods have demonstrated success in this area, but in numerous cases, a significant number of consensus-guided mutations can be destabilizing. These destabilizing mutations can arise from repacking issues that require more nuanced compensatory mutations to resolve, something consensus can be blind to.

Altschuch et al. showed in a 1987 study of the coat proteins of Tobacco Mosaic Virus and its homologs that positions with coordinated substitution patterns in multiple sequence alignments were found to be close in space while not adjacent in the primary structure, a result successfully repeated the following year in similar studies of serine proteases, cysteine proteases, and hemoglobin. An example of such a sequence pattern is provided in Fig. 1.1.2. In 1990, the Westhof lab modeled the tertiary structure of the Group I catalytic intron using similar methods. The growth of available structural and sequence databases allowed for this phenomenon to be tested computationally. Korber and co-workers used information theory correlation metrics to predict functional linkages
Fig. 1.2 A short example alignment demonstrates what we mean by “co-evolving” residues. By translating the amino acid identities into numbers based on the appearance of the mutation, we can render the sequence data in an abstract manner amenable to pattern detection. We can hypothesize a link between positions 1-5 and 3-7, because they share the same pattern of mutations.
between residues of the V3 loop of HIV-1 envelope proteins.\textsuperscript{19} A later study on myoglobin by Neher showed an overlap between crystallographic contacts and the high correlation coefficients for pairs of alignment positions.\textsuperscript{20} A similar study by Pazos and coworkers showed a link between correlated mutations and protein-protein docking contacts.\textsuperscript{21} It would be a logical suggestion, then, that residues observed as statistically linked in an alignment could serve as the targets for functionally-informative mutagenesis.

The Ranganathan group made a ground-breaking contribution to this field when they demonstrated a linear correlation between statistical coupling of aligned residues in the PDZ domain and the change in free energy from mutations.\textsuperscript{22} In a follow up paper, the Ranganathan group demonstrated that networks described by these statistical couplings corresponded to the known allosteric pathways of G-protein-coupled receptors, chemotrypsin, and hemoglobin.\textsuperscript{23} Fodor & Aldrich tested the reliability of using statistical coupling by comparing their own sequence analysis to published double mutant studies. They found that coupled residues were reliably thermodynamically coupled, but that not all thermodynamically coupled residues pairs were statistically detectable.\textsuperscript{24} Magliery & Regan found that statistically coupled residues worked well as descriptors of the binding preferences of TPR motifs, which use hyper-variable regions to modulate the binding activity of an otherwise constant fold.\textsuperscript{15} The Ranganathan group fleshed out this observation by showing that, using correlated residue pairs in WW domains and their cognate binding partners, they could produce novel binding folds with native-like properties but unique activities.\textsuperscript{25,26} Sullivan and coworkers were able to use comparisons
of conservation and correlation to detect such variable regions in the core of triosphosphate isomerase and showed that accounting for them greatly improved the success of consensus-guided stabilization.\textsuperscript{27} Using high-throughput selection system that coupled PDZ binding to GFP expression and quantified fitness by FACS sorting, the Ranganathan group recently showed that clusters of correlated residues represented fitness-altering hotspots, both for loss of function and switching of function.\textsuperscript{28} Overall, the body of work over the last decade suggests that these correlated positions in proteins represent useful selective determinants and must be considered during engineering efforts.

Given our incomplete understanding of the evolutionary record and the total range of present-day sequences, there is no definitive mathematical convention to detect co-evolution between protein residues. Numerous modes of analysis have been proposed, and we will only describe what we feel are significant milestones in theoretical development (Table 1). Lockless et al. initially treated probability distributions in sequences as energetic states and described coupling as pseudo free energy changes.\textsuperscript{22} Dekkar et al. simplified this approach by using perturbation of sequence probabilities to calculate the maximum likelihood of coupled states.\textsuperscript{29} The Dunn and Wahl groups, in collaboration, use the mutual information formalism, the test between two hypotheses (dependence and independence), to test for co-evolution by equating co-evolution and statistical dependence.\textsuperscript{30,31} A test of available methodologies in 2006 by the Nussinov group, using known contacts in the Cohesin-Dockerin families as a benchmark, found these approaches to be roughly equivalent in success and rigor, with no significant
Table 1. Popular scoring methods for coevolution detection (Described by Halperin et al., Proteins 63, pp. 832-45, 2006)

**Pearson Correlation Coefficient**

\[
CM_{ij} = r_{ij} = \frac{1}{N^2} \sum_{kl} W_{kl} \left( \frac{s_{kl} - \langle s_k \rangle \langle s_l \rangle}{\sigma_k \sigma_l} \right)
\]

The correlation coefficient \( r \) between two positions \( i \) and \( j \), for sequences pairs \( k \) and \( l \), for total sequences \( N \). \( W \) is the fraction of non-positions, \( s \) is a similarity score determined from a reference matrix, and \( \sigma \) is the standard deviation.

**Observed minus expected squared, OMES**

\[
N_{ex} = \frac{N_{xi} N_{yj}}{N_{valid}} \quad r_{ij} = \sum_{l} \frac{(N_{obs} - N_{ex})^2}{N_{valid}}
\]

The correlation coefficient \( r \) between two positions \( i \) and \( j \), for residues \( x \) and \( y \), is a function of the number \( N \) of observed pairs and expected pairs. In this case, 'valid' stands for pairs in which positions are not a gap.

**Mutual Information, MI**

\[
MI_{ij} = \sum_{xy} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}
\]

The mutual information bit score \( MI \) for positions \( i \) and \( j \) is a function of the probabilities of residues \( x \) and \( y \) and the probability of their co-occurrence, for all possible residues.

**Statistical coupling analysis**

\[
\Delta \Delta G_{ij} = \sqrt{\sum_x \left( \ln P_{ij|\sigma_j}^x - P_i^x \right)^2}
\]

The statistical energetic coupling between positions \( i \) and \( j \) is a function of the probability of residue \( x \) at position \( i \), for the whole alignment and for a subset conditioned on the residue at position \( j \).

**Explicit likelihood of subset covariation (Perturbation)**

\[
\prod_{x=1}^{20} \left( \frac{N_{xj}}{n_{xj}} \right)^{n_{xj}} \left( \frac{N_{xj}}{m_{xj}} \right)^{m_{xj}}
\]

For a user-defined subset of the alignment, \( N \) is the number of residue \( x \) at position \( j \) for all sequences, \( n \) is the number in the subset, and \( m \) is the number outside the subset.

**Two-State Maximum Likelihood**

\[
LR = -2 \ln \left( \frac{L_I}{L_D} \right)
\]

The statistic \( LR \) is the comparison of the likelihoods of two models, determined by parsimony, stating from the assumption of either independence \( I \) or dependence \( D \) for two positions.
difference between the thermodynamic and informatic assumptions. Rigor of the mutual information approaches could be improved by subtraction of the average result, to limit background arising from finite sequence availability. Additionally, results can be further improved by weighting sequences by redundancy and representation. Recent efforts have focused on expanding analysis from pairwise to multi-residue descriptions. Li et al. have used conditional random fields, comparing probability matrices of whole sequences, to predict protein-protein interactions, which has obvious precedent in the analysis of correlated core residues. The Ranganathan group has moved towards categorizing groups of pairwise interactions by eigenvalue decomposition of pairwise probability matrices. (Fig. 1.1.3)

Although the computational methods are still under debate, these correlation-guided approaches have shown clear success in protein engineering projects, beyond even proof-of-principle experiments. Co-evolving residues pairs have been used to stabilize KDO8P synthase to remove metal dependence and describe evolutionary constraints in the antibody evolution. Identification of co-evolving residues has also proven useful in redesign of protein-protein interaction networks, including finding determinates of specificity for RXR heterodimers and histidine kinase and their response regulators. Comparisons of co-evolution in alignments can contain homologs with differential activity have aided in the identification of allosteric pathways in chaperones and pyruvate kinases, drug binding sites in dihydrofolate reductase, and distinct natural interaction surfaces in receptors. Most impressively, comparison of co-evolution and structural data helps in locating surface sites for novel allosteric regulation.
Fig. 1.3 Groups of highly correlated positions mapped to the crystal structures of PDZ (A), PAS (B), SH2 (C), and SH3 (D) domains, originally published by Halabi and coworkers  and reproduced with permission. Correlation scores are displayed by heat maps that show clustering of high values. In these cases, mapped regions are either contiguous or represent points of contact with substrates.
Section 2: Adenylate kinase – Kinetics and Binding

Adenylate kinase (ADK, EC 2.7.4.3.) is an enzyme with a rich literature of sequence identification and comparison, biophysical and biochemical characterization which makes it an ideal platform for studying the intersection of evolutionary events and protein biophysics. Accordingly, we’ve used it as a model system for testing the significance of detected correlations. ADK catalyzes the magnesium-dependent transfer of the γ-phosphate from adenosine triphosphate (ATP) to adenosine monophosphate (AMP), subsequently generating two adenosine diphosphate (ADP). This chemistry opens up a range of possible cellular functions for ADK. In mammalian genetics, isoforms of ADK have been implicated in maintaining energy production during cell division, nucleotide biosynthesis during DNA replication, sensing low ATP levels, and acting as a position-dependent link in signal transduction.

Adenylate kinase belongs to the P-loop kinase superfamily, whose two major features are the phosphate-binding Walker-motif and the core α-β-α sandwich. Along with closely related family members guanidine kinase and tyrosine kinase, ADK is noteworthy for the way substrate binding leads to global conformation rearrangement of its two nucleotide binding domains and burial of the binding site. This large scale rearrangement has been the subject of both extensive molecular dynamics and structural experiments, aimed at explaining its mechanism. Although this phenomenon was first observed in porcine ADK isoform 1, the focus of many of these efforts has been E. coli ADK, due to an evolutionarily-common 27 residue insert that forms a third domain and similarly participates in the binding-driven rearrangement as an active site lid. Additional
experimental focus on ADK stems from the wide range of homologs available from different environments – thermophiles, mesophiles, psychrophiles, and halophiles. These diverse but similar folds provide the opportunity to examine how a protein can be adapted to changing environments. They also allow for researchers to observe how environmental adaptation can alter the function of an enzyme and how detrimental evolutionary effects are avoided. Here, we seek to usefully summarize the current body of work for ADK. We will focus on allosteric biophysics and evolutionary adaptation, as well as current understanding of ADK’s fundamental enzymatic activity.

Mid-century work on the catalytic activity of ADK has been collected in a chapter by Lafayette Noda. The transfer of the $\gamma$-phosphate from Mg·ATP to AMP is commonly referred to as the forward reaction, while the reverse reaction transfers the $\beta$-phosphate of ADP to Mg·ADP. (Fig. 1.2.1) The bivalent cation complex is essential to catalysis. Reaction rates, from greatest to least, follow the sequence Mg$^{2+}$, Ca$^{2+}$, Mn$^{2+}$, and Ba$^{2+}$ or Co$^{2+}$. Early isotope exchange experiments showed the mechanism follows random Bi-Bi kinetics for two binding sites. The binding site for the nucleotide-cation complex appeared to interact primarily via the phosphates and cation and could accept a wide variety of purine bases, albeit with less binding energy. The AMP/ADP binding site, by contrast, is highly specific for adenosine. Given its high $K_i$, diadenosine pentaphosphate (AP5A) is generally accepted as a transition state analog.

$^{31}$P NMR showed symmetrical peaks for P1 and P5 and for P2 and P4 of free AP5A, but that these peaks would lose symmetry in the presence of ADK, demonstrating significant variation between the ATP and AMP binding states. Structural
studies of the bound state\textsuperscript{5,52} found that substrate binding induced large conformational transitions in ADK, which close the active off from solution. I will discuss these conformational effects in more detail below and will for now focus on their kinetic implications. Two transitions were isolated by FRET\textsuperscript{53} – a partial 9 Å transition for ATP or AMP binding, then a 12 to 13 Å transition for AP5A, corresponding to the fully closed active site. This suggests a two-step closure mechanism. Kinetic data showed reaction inhibition at high concentrations of AMP. These combined facts suggest that the first step of active site closure depends on Mg·ATP binding and that concentrated AMP can compete with Mg·ATP, with the enzyme trapped in the first step and unable to continue to the second.

An extensive kinetic study of the forward and reverse reactions by Sheng et al.\textsuperscript{54} produced a revised iso-random Bi Bi mechanism. (Fig. 1.2.2) Not only does AMP inhibit Mg·ATP binding, but ADP inhibits Mg·ADP, as determined by varying MgCl\textsubscript{2} concentration to modulate the ratio of ADP to Mg·ADP. This confirms that not only does substrate inhibition occur by blocking the nucleotide-cation binding site, but it also suggests that the nucleotide-cation complex must form before enzyme binding. However, despite similar binding mechanisms, these studies demonstrate a pronounced difference between the forward and reverse reactions. For the forward reaction, AP5A acts as a competitive inhibitor. But, it acts as a mixed noncompetitive inhibitor of the reverse reaction. The iso-random Bi Bi mechanism, then, postulates that ADK has two distinct isoforms for binding substrates or products.
**Fig. 1.4** Substrates and products of the ADK-catalyzed phosphotransfer reaction. The oxygen of the AMP phosphate makes an in-line nucleophilic attack on the phosphorous of the γ-phosphate.

**Fig. 1.5** Iso-random bi-bi reaction mechanism proposed by Sheng and co-workers. $E_1$ and $E_2$ are conformations of the enzyme. M is AMP, T is Mg·ATP, $D_1$ is ADP, and $D_2$ is Mg·ADP.
Tan et al. recently suggested a revision of this mechanism, which incorporates allosteric activation by Mg$^{2+}$. The basis of this hypothesis is an observed rise in activity despite being at already saturating concentrations of ATP. Although these results are believable, it is not clear that ADK binds Mg$^{2+}$. Tan et al. claim that binding is demonstrated by titrating enzyme into a solution containing MgCl$_2$ and the fluorescent magnesium indicator mag-fluo-4, from Invitrogen. As enzyme is added, fluorescence is quenched, interpreted as sequestration of Mg$^{2+}$ from the dye by specific enzyme binding. However, this assay, by design, produces linear data and cannot demonstrate a point of saturation. This reviewer feels it cannot be reliably interpreted as binding data, if saturation cannot be positively determined.

ADK binds its substrates in a cleft composed of three independently-folding domains. I will refer to these domains as Core, AMPbd, and Lid. (Fig. 1.2.3) Historically, the Lid domain may also be referred to as Insert. A highly conserved feature of the binding site, one shared among several related classes of kinases, is a flexible glycine-rich “P-loop” in the Core domain which binds anions with both main chain and side chain contacts. In ADKs, two prominent features of this loop are a conserved lysine and a spatially close arginine from the Lid domain. Mutations made to prolines and glycines in this loop lower binding affinity, but do not affect $V_{\text{max}}$. However, mutation of the conserved lysine to a glutamate disables catalytic activity. NMR of chicken muscle ADK K21A shows that the P-loop lysine dictates the orientation of the ATP triphosphate and stabilizes the entire active site, an effect which cannot be rescued by K21R.
Schulz and co-workers solved the structure of *E. coli* ADK (ADKec) bound to $\text{AP}_5\text{A}$. Although $\text{AP}_5\text{A}$ is symmetrical, the ATP and AMP binding sites could be inferred by their features and previously mentioned work. Many of the crystal structure

![Lid domain](image1)

**Fig. 1.6** Structure of unbound ADK ec (PDB: 4AKE). The Lid domain is highlighted in purple, the AMPbd domain in red, and the P-loop of the Core domain in yellow.
observations are backed-up by NMR of yeast and muscle ADK by the Tsai group.\textsuperscript{60,61} In the ATP binding site, the adenine and ribose features are both loosely bound. The adenine forms a single hydrogen bond with K200 of the Core domain and stacks its ring against the guanidine of the R119 side chain, providing its low binding specificity. The ribose forms a hydrogen bond to Y133 of the Lid domain. The ATP $\alpha$- and $\beta$-phosphates are more strongly bound. The $\alpha$-phosphate forms two hydrogen bonds with the P-loop main chain. The $\beta$-phosphate forms five hydrogen bounds with the P-loop chain and interacts with the P-loop's K13 and R123. The $\gamma$-phosphate, which is transferred during the reaction, weakly interacts with two additional arginines, R156 and R167, as well as K13. The high flexibility of analogous K21 of rabbit muscle ADK – predicted by chemical modification from different phosphopyridoxal species\textsuperscript{62} – leads to the hypothesis that the P-loop lysine 'shuttles' the transferred phosphate from the substrate to product state. Mutation of R123, R156, and R167 to lysine lowers $k_{\text{cat}}$, implying that specific guanidine geometry is necessary to interact with the transition state.\textsuperscript{60} The orientation of the active site arginines is further restrained by D158 and D159 of the Lid domain. Alanine scanning of the active site shows that all of these charged side chain interactions are essential for catalysis.\textsuperscript{63}

The AMP binding site has a collection of specific main chain and side chain contacts with the adenine base, from both the Core and AMPbd domains, resulting in greater specificity for AMP binding.\textsuperscript{5} Mutation of the residue L66 of chicken muscle ADK to other hydrophobic residues greatly lowers AMP affinity.\textsuperscript{64} The AMP $\alpha$-phosphate interacts with two AMPbd arginines, R36 and R88.\textsuperscript{5,60} R88 is especially key
for activity and is protected from chemical modification by AMP binding. These interactions restrain AMP phosphate geometry. A subsequent structure using AMP and AMPPNP validates this interpretation and shows AP₅A to be an effective approximation of ATP/AMP binding, despite its additional phosphate-4. Early reaction studies by the Frey group using chiral substrates predict inversion of the ATP γ-phosphate by in-line nucleophilic attack from the AMP phosphate. Geometry of substrates in both the Schulz and Phillips's structures support this prediction. Molecular dynamics simulations based on these structures show the substrate orientations to be stable over the timescale of the reaction.

A later structure of yeast ADK with AP₅A successfully incorporates a bound Mg²⁺. This magnesium does not perturb the ligand geometry from what was observed in ADKec and forms a distorted octahedral coordination sphere incorporating the three ATP phosphates and two waters bound to nearby D89. (Fig. 1.2.4) Using ³¹P and ²⁵Mg NMR, the Tsai group probed the role of this aspartate by making an aspartate-to-alanine mutation and recording chemical shift differences for the two constructs. (Due to chain numbering differences, the Tsai group identifies this mutation as D93A.) They found that D93A resulted in a drop in magnesium affinity, but that activity could not be restored by only increasing magnesium concentration. Differences in ³¹P ensembles for the ADK/AP₅A complex point to a change in the energy of the transition state resulting from the D93A mutation. Shifts in proton peaks also showed that D93 interacts with the structurally-stabilizing H36, but the D93A construct showed no difference in stability.
Comparison of ADK structures in the bound and unbound states\textsuperscript{1,52} show that the Lid and AMPbd domains undergo large conformational shifts upon substrate binding. These shifts are best described a rigid-body motions around structural “joints”, moving from an extended structure to a more globular one. (Fig. 1.2.5) NMR studies demonstrate that while the secondary structure content of these domains remains constant with binding, the rate of hydrogen exchange with the solution significantly slows.\textsuperscript{72} The Lid and AMPbd have nanosecond-scale backbone motions which are quenched by ligand binding.\textsuperscript{73} Although this trend dominates for most of the structure, a section of the Core domain – at the junction of α4 and α5 and the β-sheet, as well as nearby α1 – show a higher rate of hydrogen exchange and higher crystallographic B-factors in the bound structure.\textsuperscript{1,72} The hypothesis has been put forward\textsuperscript{1} that these regions act a “counterweight” to the ligand-induced stabilization of Lid and AMPbd.

It is possible to crystallize ADKs in partial states of the Open-to-Close transition. As a follow-up to their bound ADKec structure, the Schulz group crystallized mitochondrial matrix ADK bound solely to AMP, to confirm the predicted AMP binding site. This structure not only agreed with their predictions, but showed the AMPbd domain in the same conformation as in AP\textsubscript{5}A-bound ADKec, without the corresponding closure of the Lid domain.\textsuperscript{74} A similar phenomena was seen in a fortuitous structure of AP5A-bound \textit{B. stearothermophilus} ADK obtained by the Phillips group.\textsuperscript{75} In this case, despite the AP\textsubscript{5}A occupying the ATP binding site on the P-loop and the AMPbd domain assuming the bound position, the Lid domain remained in a partially open conformation. The authors hypothesized this was caused by crystal contacts overpowering the normal
**Fig. 1.7** Structure of *B. subtilis* ADK bound to AP₅A and magnesium (II).
(PDB: 1P3J)

**Fig. 1.8** Flexible “hinge” regions of ADKec (PDB: 4AKE) highlighted in red, as identified by Muller and coworkers. These regions undergo repacking of secondary structure while the rest of the protein moves as a rigid body.
ligand interactions. Noteworthy features of this structure are the unique conformations of Arg127 and Arg160, whose analogs are predicted from the AP₅A-ADKec to interact with the ligand phosphates and drive Lid closure. The Schulz group also obtained structures of inactive mutant yeast ADK (D89V, R165I) unbound, bound to AP₅A, and bound to ATP-analog AMPPCF₂P. The AP₅A-bound structure showed that the full-substrate bound state was unperturbed and that the mutations, at residues that interact with the catalytic magnesium and the transferred phosphate respectively, only perturbed the reaction mechanism. The AMPPCF₂P-bound structure showed a closed Lid domain, but a fully extended AMPbd domain.

The Kern group attempted to observe the rate-limiting action of the active site using CMPG relaxation dispersion NMR methodologies to fit the kinetic parameters for the micro-to-millisecond exchange of chemical shifts. However, they found the dominant shift to be a global transition of the protein structure, on the time scale as the overall reaction. This chemical exchange was identical for true substrate and substrate analogs, thus implicating the closure step as overall rate limiting.\textsuperscript{76} Determination of the rates of opening and closing in both ADKec and thermophilic \textit{A. aeolicus} ADK showed that both enzymes had similar closure rates, but different, temperature-dependent opening rates, which could account for the different temperature of maximum activity.

NMR experiments by the Wolf-watz group measured residual dipolar coupling of ADKec in an anisotropic media. This allowed for them to model the in-solution closure of Lid and AMPbd in the presence of saturating ATP, AMP, or AP₅A.\textsuperscript{77} Interestingly, ATP and AMP produced intermediate peaks for their associated domains, which were
Fig. 1.9 The scheme proposed by Wolf-watz and co-workers for the conformation mediated reaction pathway of ADK. Originally published in Åden et al\textsuperscript{75} and reproduced with permission.
linear combinations of the unbound enzyme and the AP₅A-bound state. The individual substrates, then, induce an ensemble of open and closed states for the Lid and AMPbd domains. A noteworthy result is that AMP only induced closure for the AMPbd domain, but ATP affected the Lid domain and then the AMPbd domain to a lesser extent. This led the authors to suppose one-way energetic communication between the Lid and AMPbd domains. This experiment provides a schema for in-solution binding of substrate to ADK. (Fig. 1.2.6)

In a second experiment, the Wolf-watz group created a series of ADKec constructs in which ATP binding at key sites was sequentially knocked out, to determine specificity of interaction and individual contributions to active site components. Strength of binding was determined by SPR and site of binding was determined by ¹H-¹⁵N HSQC NMR.₇⁸ With the mutations R123A, R156A, and R167A, they successfully erased interactions between ATP and the Lid domain. In this construct, ATP bound to the P-loop and the AMPbd domain. Additional mutation of known adenine-interacting residues in AMPbd restricted binding to the P-loop. A reverse construct retained the AMPbd mutations and eliminated the P-loop by circular permutation, but kept the key arginines for Lid interaction. However, this construct completely lost ATP binding activity. In this way, the Wolf-watz group established that the Lid domain only interacts with ATP after P-loop binding. Interestingly, the R123A-R156A-R167A constructs were still able to carry out the phosphotransfer reaction, as observed by ³¹P NMR. However, two reactions were observed – phosphorylation of the α-phosphate of AMP and hydrolysis of Mg•ATP to ADP. Hydrolysis was also observed for only enzyme and ATP. However, it does not
occur when ATP in mixed with wild-type enzyme. This provides solid evidence that the mechanistic purpose of Lid domain interactions is to prevent hydrolysis, as an unproductive side reaction.
Section 3: Adenylate Kinase – Folding and Dynamics

Given the complex relationship between the dynamics of its three domains, ADK unsurprisingly has non-trivial convolutions in its folding pathway. Denaturation experiments by urea and guanidine hydrochloride both show two inter-converting intermediates between the native and denatured states, as observed by CD, fluorescence, and SEC. Thermal denaturation fits the two-state model, but is irreversible unless ADK is treated with 3 M GuHCl after cooling. Thermal inactivation precedes observable perturbations in the secondary or tertiary structure or aggregation. Interestingly, at low concentrations of denaturant, there will be a spike in activity before inactivation, which corresponds to a spike in binding of 8-anilino-1-naphthalenesulfonic acid (ANS). ANS is a dye which fluoresces in hydrophobic environments. It has been shown crystallographically to bind to the AMP binding pocket and produces fluorescent signal due to the binding pocket's hydrophobic nature. This has lead the Zhou group to propose that increased flexibility of the Lid and AMPbd domains leads to more frequent binding site exposure, before critical unfolding events destroy activity. This corresponds well with observations by NMR that the Lid and AMPbd have faster intrinsic motions than the Core domain and that transitions between the open and closed state are rate limiting. NMR of ADKs under thermal denaturing conditions and with perturbing mutations shows that the three domains fold independently of each other and in a non-cooperative fashion.

Ruan, Qiao et al. demonstrated that ADKec could be reversibly unfolded by an increase in hydrostatic pressure and used tryptophan mutants at six well-distributed
sites to probe local folding rates.\textsuperscript{85} They found that five of the positions refolded at comparable rates, but that Y193W folded at a slower rate than the rest of the protein. A similar experiment by the Haas group used a FRET donor-acceptor pair at positions 169 and 188 to measure the rate of folding in this helical region during stop-flow denaturation and refolding.\textsuperscript{86} They saw that the emission decay of the donor and acceptor change rapidly, but that the distribution of distances did not narrow to that typical of \(\alpha\)-helices until a second, slow folding step. This implies the labeled sites changed chemical environment suddenly, but did not form secondary structure until after the environment shift. A second experiment by the Haas group repeated this labeling study, now with the addition of probes at 188/203 (a \(\beta\)-strand region) and 28/203, which would report on global diameter.\textsuperscript{87} This showed that the chain as a whole collapsed to a sphere by 5 ms (roughly the dead time of the experiment), but the end-to-end distances indicative of secondary structure for 169/188 and 188/203 appeared at 2-3 seconds.

The Haas group then performed a number of FRET experiments on ADKec aimed at describing the role of loop closure during folding. Early experiments used labels at 28/71 to measure the collapse of the AMPbd domain as a loop,\textsuperscript{88,89} while using labels at 28/86 and 58/86 monitored the overall foldedness of the AMPbd domain and it's interaction with the Core domain.\textsuperscript{90} The residues at 28/71 reach native-like distance before the spherical collapse of the entire chain, while 28/86 and 58/86 do not show any packing of the AMPbd and Core domains. Using ensembles of emission decay from position 58, the Haas group demonstrated that the 28/71 loop collapse occurred before native-like folding of the entire AMPbd domain was achieved.\textsuperscript{90} The 18/203 pair was
used to monitor the overall, cooperative two-state transition to the folded protein.\textsuperscript{89} The 28/71 pair closes together as a loop three orders of magnitude before the cooperative transition.\textsuperscript{89} A recent follow-up experiment uses nine donor-acceptor pairs to compare folding times of three loops (I, III, IV) and three core β-strands.\textsuperscript{88} Two of the loops showed native-like contacts within 1 ms. One loop and one β-strand showed partial closure in a fast burst, then finish forming native contacts on the slow time scale. Finally, the remaining two β-strand fold on the slow time scale. Interestingly, emission decay ensembles show that the loops are tightly packed early in the folding transition, then become perturbed as β-strand structure resolves.

The Haran group used a high-throughput FRET methodology to characterize the number of ADKec substates at varying concentrations of guanidinium chloride.\textsuperscript{91} This was accomplished by trapping individual proteins in lipid vesicles by emulsion and affixing the vesicles to a glass-supported bilayer. Using labels at 73/203 to span the Core domain, FRET trajectories were collected at equilibrium for ADKec. These trajectories were then fit to a Hidden Markov model to describe their connectivity and calculate the likeliest number of substates. This identified six distinct substates. At lower concentrations of denaturant, multiple distributed pathways were evident. However, as concentration increased, the number of pathways narrowed, until a single sequential pathway dominated. (Fig. 1.3.1) The Takada group used these results to qualify their atomic-interaction course-grained model, which reduced residues to spheres while applying their known atomic interactions.\textsuperscript{92} Their methodology predicts seven distinct substates along the folding pathways, five of which are characterized by variations within
Fig. 1.10 The FRET substates of ADKec and the transition rates between them, at varying concentrations of denaturant. Originally published by Haran and coworkers\textsuperscript{89} and reproduced with permission.
Fig. 1.11 Proposed folding pathways for ADKec, originally published by Takada and co-workers\textsuperscript{93} and reproduced with permission. Blue highlighting indicates well-ordered regions, and red indicates disorder. Arrow colors indicate observed folding trajectories. Arrow size indicates propensity.
the structure of the Core domain. These simulations first predict folding of the Lid domain, followed by a distribution of states in which folding is driven by collapse of the AMPbd domain or the core β-sheet. As these substates approach the native state, the Lid domain becomes increasingly disordered again. Raising simulation temperature narrows the distribution of substates to favor full folding of the Core domain before folding of the AMPbd domain, which is notably counter to the sequence of events observed by the Haas group under refolding conditions. (Fig. 1.3.2)

ADK makes an attractive target for the study of protein dynamics, both computationally and by techniques such as NMR and FRET. The picture of ADK’s closure mechanism is constantly evolving and points to a combination of traditional models. In 2005, Magarakis and Karplus created a plastic network model of active site closure, using then 45 published crystal structures of various homologs and bound conformers to constrain the pathway. Their model agreed with the hypothesis that local unfolding of “hinge” regions contributed to closure and predicted that Lid closure must precede AMPbd closure. Using distance replica exchange methodology (which increased the range of conformation sampling), Lou and Cukier published a similar model in 2006, with the observation that the apo state could sample conformers similar to the ligand bound state. A later simulation by Kubitzki and Groot affirms these results without adding directional restraints. Shapiro and Meirovitch used a combination of inversion recovery, CPMG, and steady-state NOE NMR measurements on 15N-labeled ADKec to calculate its tumbling rate and the rate of domain motion. They found that not only was the rate of domain motions roughly half of tumbling, but that the activation energy of
domain motion (31 kJ/mol) was well below the expected activation energy of phosphotransfer, making it unlikely that these motions were exclusively coupled to catalysis. In 2007, the Yang group published a single-molecule FRET experiment which demonstrated ADKec equilibrated between closed and open states even in the absence of substrate and that substrate shifted the equilibrium to the closed state. A simulation by Arora and Brooks published at the same time predicted that rather than substrate binding the open state, ADK could first assume the near closed conformation and then substrate would bind with virtually no energetic barrier. Overall, this work points to a population selection model of ADK binding, in which the protein finds the substrate binding conformation, then binds its target. A principal component analysis of MD trajectories for ADK published by Cukier agrees with this model.

However, a Go-model simulation of ADKec by Daily, Phillips, and Cui disagrees with the above conclusion. In this model, although both the apo and holo enzymes sample open and closed conformations, holo-ADK leads with closure of the Lid domain, while apo-ADK leads with the AMPbd domain. A similar study comparing go-models of mesophilic and thermophilic ADKs repeats this result despite varying entropy of the open Lid domain at different temperatures and across computational mutants. A molecular dynamic trajectory study of ADK opening with explicit waters by the Bagchi group, published in 2011, comes to a number of interesting conclusions about the relative motions of the Lid and AMPbd domains. They first note that the Lid domain has a much more diverse and rugged set of trajectories than AMPbd, indicating a greater freedom of motion. Second, the opening of AMPbd has a much steeper transition
energy than Lid. Third, the trajectories of the Lid and AMPbd show strong correlation. Taken together, these results suggest a model into opening and closure of AMPbd is facilitated by Lid, which is in turn restricted by temperature and NTP binding.

This controversy persists even for very recent simulation studies. The Kidera group has produced by simulation using “on-the-fly” string methodology which reaffirms the AMPbd-then-Lid sequence of closure,\textsuperscript{103} due to a requirement that that Lid requires the energy from contacts with the AMPbd domain to successfully dehydrate the P-Loop. The Kidera model also proposes a partially closed state of the Lid, seen prior in crystal structures, may exist in solution and be responsible for experimental results that show the presence of the closed state for apo-ADK. A 2013 simulation of trajectory evolution from either the closed or open state by Song and Zhu\textsuperscript{104} agrees with this interpretation, bycontending that the closed apo state is far less stable than the open state. The Papoian group published free energy calculations for ADK substates using updated umbrella sampling techniques\textsuperscript{105} and came to the conclusion that many substates of ADK occupy a wide, swallow energy basin in which many conformers maintain Lid-AMPbd contacts. An extensively-sampled model by the Wang group succeeds in bringing together previously disparate computational results.\textsuperscript{106} They identified two pathways, one with AMPbd closure leading and one leading with Lid closure, which are preferences by temperature. The Lid-first pathway dominates at room temperature, while the AMPbd-first pathway dominates at the optimal reaction temperature. This would make the AMPbd pathway entropically drive and the Lid pathway enthalpically driven.
I have previously mentioned the connections made between the closure of the ADK binding site and the flexibility of the Lid and AMPbd domains. I will here elaborate on work done on the packing of closure related elements and side chain motions which contribute to structural transition. Bae and Phillips generated a series of chimeras exchanging the Lid, AMPbd, the hinge regions, and “thermodynamic counterweight” regions of mesophilic and thermophilic ADKs. Each chimera was tested for melting temperature and temperature of optimal kinetic activity. None of these mutations greatly perturbed the melting temperature; melting temperature appeared to be dictated by the origin of the Core domain. Each chimera had perturbed temperature-dependence of kinetics, but by far, the largest perturbations resulted from the exchange of the Lid and AMPbd domains. Work by the Kern group compares the ps-to-ns fluctuation of amide bonds between a meso/thermo ADK pair using the internal correlation time and order parameters of NOE NMR peaks. They found that the thermophile had slower backbone flux than the mesophile at 303 K. However, using experimental order parameters to first qualify room temperature simulations, they demonstrated by simulation that the thermophile's order parameters at its optimal kinetic temperature were similar to the mesophile's at 303 K. These two results suggest that activity-dictating movements in ADK are caused by the thermal motions of the mobile domains. A simulation by Pontiggia et al. found that steps along the trajectory of the Lid from open to closed were punctuated by increases and then release of local strain in several positions along the Lid sequence. This poses the possibility that the closure of mobile domains may be caused by folding and unfolding of the domains, which is addressed experimentally by
By mutating Lid domain residues to glycines, they produced a series of ADKec mutants which lost kinetic activity but not thermal stability. Both crystal structures and HSQC NMR confirm that when bound to AP$_5$A, these variants have the same structure as wild-type. However, ITC experiments determine a difference in binding energies highly similar to the difference in unfolding energies for the Lid domain variants, as predicted by the COREX algorithm. A combination of MD and statistical coupling analysis was used by Armenta-Medina et al.\textsuperscript{111} and found a strong relationship between co-evolution and coupled motions between residues in Lid and AMPbd, indicating strong selective pressures worked to preserve these small-scale dynamics.

The Kern group noted\textsuperscript{108} that the hinge regions with kinetically-correlated chain flexibility were helix regions with residues of low helical propensity. Brokaw and Chu used simulations of apo-ADKec in the open and closed states to probe the role of one such residue, Pro177, in stabilizing each conformation.\textsuperscript{112} A 100 ns simulation showed that the apo closed state transitioned towards the open state, while the open state was stable. When Pro177 was restrained in the unfolded position for the first 11 ns, the closed state fluctuated around closed and partially closed, with several unique hydrogen bonds forming between the Lid and substrate binding site and the Lid and AMPbd and around the Pro177 hinge. Olsson and Wolf-watz characterized two hinge region mutations, I116G and L168G, and demonstrated that ATP binding affinity is inversely related to thermal stability.\textsuperscript{113} Both mutations lower melting temperature and cause local unfolding of the hinge region, shown by HSQC. ITC shows higher ATP and AP$_5$A affinities for both mutants. However, $V_{\text{max}}$ for the mutants is decreased relative to wild-type. As noted
in earlier sections, the opening of the Lid is rate limiting. Since a lack of HSQC peak perturbations shows that the contacts between ADK and substrate are undisturbed, Olsson and Wolf-watz interpret the high substrate affinity to show that a lack of folding energy for the hinge regions favors the closed state and inhibits the opening motion. This agrees with a simulation by Whitford et al.\(^\text{114}\) which showed that local unfolding in the hinge regions was preceded by a rise in main chain strain.

Lu and Wang provide a detailed analysis of inter-residue contacts during the open-to-closed transition, for both Lid-leading and AMPbd-leading scenarios, in ADKec.\(^\text{115}\) A striking feature they noted was the distinct orientations of the solvent-sensitive contact pair K97-N190, of the Core domain, in either closure pathway. Initial steps in Lid closure were driven by hydrogen-bond contacts with the P-loop, aided by hydrophobic contacts F137-N138-V202. Initial contact at this stage with AMPbd is mediated by A127-S129-G130 (Lid) and D33 (AMPbd). AMPbd closure is driven by hydrophobic contacts with the Core domain. At the second stage of closure for either scenario, a series of clustered charged interactions are made between Lid and AMPbd. Four of these pairs (D54-K157, R36-D158, D33-R156, K57-E170) were computationally identified as salt bridges by Beckstein and coworkers.\(^\text{116}\) The Wolf-watz group produced a mutant of one of these residues, E170A, and demonstrated by NMR that apo-ADK equilibrium had been shifted towards the open state and consequently catalytic activity was decreased.\(^\text{117}\)
Section 4: Adenylate Kinase – Evolution

Adenylate kinases are diverse, ubiquitous proteins necessary for signaling and energy homeostasis in eukaryotes and bacteria, being critical for cellular development and drug interactions and being implicated in oncogenesis. These proteins have been studied for nearly sixty years and enjoy extensive structural and kinetic characterization. ADKs have the twin virtues that they are amenable for large-scale purification and yet possess a range of complex biophysics relevant to questions of multi-domain folding and allostery. ADKs are globular proteins with three domains (Core, Lid, and AMPbd) which coordinate to close over substrates and form an active site to catalyze the transfer of phosphate from a Mg·ATP complex to AMP. This fortuitous intersection of wide diversity, compelling biophysics, and well-established methodology make ADKs ideal for addressing questions of the sequence-to-structure relationship in highly dynamic proteins. We present here a short review of known evolutionary relationships for ADKs and evolution of ADKs in a laboratory setting.

Kinases, defined by their ability to use ATP as a substrate for phosphorylation, can have a wide range of folds related by divergent and convergent evolution and are considered an ancient class of enzymes. ADKs belong to a group of kinases with a Rossmann-like fold. Rossmann folds have a core β-sheet with anti-parallel α-helices packed against both sides. ADKs specifically have five β-α pairs making up the Core domain, with the β-strands packing in the order of 23145, and the loops of pairs 2 and 4 extending to form the AMPbd and Lid domains, respectively. ADKs further fall into the sub-group of P-loop kinases, where β1-loop-α1 and β3 form two sides of the NTP
binding site. P-loops can contain the Walker A (GXXXXGKT/S) and Walker B (hhhhD/E) motifs – Walker A is a triphosphate binding site and Walker B forms part of the coordination sphere for a divalent cation with the substrate phosphates. The lysine of the Walker A motif is considered integral for phosphotransfer. The P-loop is characteristic shared by rNMP and nucleoside kinases and GTPases. The defining divergence of these two classes is the Lid domain, whose conserved arginine interacts with the adenine base of ATP. rNMP kinases are further defined by arginines after the Walker B motif and on helix-2 (NMP binding domain) which interact with the NMP phosphate. Among rNMP kinases, ADKs share a glycine insertion after the lysine of the Walker A motif. Archaeal ADKs have been identified with no lysine in the P-loop, with the catalytic role instead performed by an arginine in the Lid domain.

A major feature of ADK divergence is the length of the Lid domain. In certain ADKs, the Lid domain is shortened by 27 residues. Based on phylogenetic distribution, this represents a deletion event that took place in an ancestral Lid-containing kinase, closely related to guanylate kinase, before the differentiation of bacteria and eukaryotes. The ADK from *H. halobium* has a long-length Lid, but closer sequence similarity to short-length ADKs, raising the possibility that it is close to the common ancestor. Although short-Lid ADKs primarily appear in eukaryotes, notable bacterial examples exist, raising the possibility of several independent deletion events. Characterization of a short-Lid ADK from *M. tuberculosis* found that, relative to *E. coli* ADK, it gained thermostability but lost activity and substrate specificity. Human cells can express nine different isoforms of ADK, with varying localizations and tissue
distributions. Phylogenetic models suggest that pre-eukaryotic ancestors had short-Lid isoforms and gained long-Lid isoforms via mitochondrial absorption. Two short-Lid isoforms, AK1 and AK5, are closely related and cytosolic. A third, AK6, appears to be a much more ancient divergence and is localized to the nucleus.

Another major division among ADKs is the ability to bind zinc in the Lid domain. For *B. stearothermophilis* and *B. subtilis*, Lid domains containing cysteines, similar to the zinc finger motif, were found to be protected against modification and able to bind cobalt. Removal of metal resulted in decrease of thermal stability and loss of activity, although not specificity. Zinc-binding would be later confirmed by electrospray mass spectroscopy. Mutation of the zinc binding site in *B. subtilis* to *E. coli* equivalent residues results in a phenotype similar to the wild-type enzyme with the metal removed. However, the metal-binding motif can be added to the *E. coli* Lid domain and produces an active, stabilized ADK. The Barzu group developed a protocol for the high-throughput testing of zinc-binding by ADKs. Bacteria were grown with $^{65}$ZnCl$_2$, lysed, and poured over a blue sepharose column, then eluted with AP$_5$A. Detection of radioactivity in the final sample confirmed zinc incorporation. Using a broad range of species, they saw the trend that ADKs of gram-positive species bound zinc and those of gram-negative species did not, with a small number of exceptions.

The binding motif is Cys-X$_2$-Cys-X$_{16}$-Cys-X$_2$-Cys/Asp. The first cysteine may be replaced by a histidine, but results in a much higher $K_D$. Metal-binding homologs from the gram-negative species, *D. gigas* and *D. sulfuricans*, substitute a histidine for the second cysteine. These homologs appear to use both zinc, cobalt, and iron in vivo.
Fig. 1.12 A comparison of *B. subtilis* ADK (zinc binding Lid), *E. coli* ADK, and human ADK 1A (short Lid), bound to AP$_3$A.

(PDB: 1P3J, 3HPQ, and 1Z83.)
When bound with iron, *D. gigas* ADK has notably lower activity and melting temperature and has shifts in UV absorption that suggests tertiary structure perturbation.\(^{132}\) Crystal structures of *D. gigas* ADK with all three metals were obtained and have similar folds,\(^{133}\) indicating that iron incorporation perturbs the equilibrium between the open and closed conformations.

UMP-CMP kinase (UK) is considered a member of the ADK family, descended from a common ancestor of short-Lid ADKs.\(^ {119,122}\) UK has been shown to complement ADK knock-out and temperature-sensitive strains of yeast and *E. coli*.\(^ {134}\) Crystal structures of UK with ADK substrates show an identical transition state.\(^ {135}\) Although the NMP binding site is large enough to facilitate an adenine base, contacts between uracil and the enzyme are bridged by a water molecule.\(^ {136}\) Recently, an archaeal variation of ADK was found with high sequence similarity to UK.\(^ {137}\) This homolog, dubbed ADKb, is expressed along with canonical ADK in *Methanocaldococcus*. It is believed that ADKb may be the evolutionary link between ADK and UK.

Trimeric ADKs have been crystallized from various archaeal extremophiles, of the genii *Methanococcus* and *Sulfolobus*.\(^ {138-142}\) The homotrimer arrangement is achieved by the packing of helix 7 around the axis of symmetry.\(^ {138,139}\) This trimeric packing arrangement does not correlate to high thermostability, as these homologs come from a mix of mesophiles and thermophiles and show optimum activity as each of their parents strains optimum growth temperatures.\(^ {138}\) As mentioned earlier, *Methanococcus* ADKs lack the “invariant” lysine of the Walker A motif. This lysine is still present in *Sulfolobus* ADKs, but the crystal structure of *S. acidocaldarius* shows the lysine to be oriented away
Fig. 1.13 Evolutionary classifications of ADK originally published by Leipe and coworkers\textsuperscript{117} and reproduced with permission. The minimal secondary structures of the P-loop kinase superfamily (A) and the ADK family (B) show AMPbd is an insert which confers AMP specificity. The phylogenetic tree of ADKs and UKs (C) demonstrates that deletion within the Lid domain appears at multiple, independent points.
from the substrate. This suggests archaeal ADKs may stabilize the transition state by an alternate set of contacts.\textsuperscript{139}

Multiple ADKs from the archaeal genus \textit{Methanococcus} have been characterized and crystallized. Although these ADKs share a short Lid domain and a homotrimer tertiary structure, both thermophilic and mesophilic homologs have been found.\textsuperscript{139,140} Haney et al. produced chimeras of thermophilic \textit{M. jannaschii} ADK and mesophilic \textit{M. voltae} ADK, which share high sequence similarity, by exchanging their N-terminal, C-terminal, or both.\textsuperscript{143} While exchange of either terminal produced a modest change in melting temperature, exchange of both would shift the melting temperature a full 20\textdegree C.

Examination of the differing residues in these chimeras found that almost all mutated sites were hydrophobic residues, packed in the $\beta$-sheet core. Similar conclusions were drawn from the comparison of crystal structures for \textit{M. voltae} and \textit{M. thermolithotrophicus},\textsuperscript{140} although another noteworthy feature was that while both homologs had an equal number of potential salt bridges, those in \textit{M. thermolithotrophicus} ADK formed a more extension network of contacts with each other. \textit{M. maripaludis} ADK has been shown to have the melting temperature and activity range comparable to other thermophilic ADKs, but comes from a mesophilic organism, suggesting it may be in the early stages of mesophilic adaptation.\textsuperscript{139}

A range of psychrophilic, mesophilic, and thermophilic ADKs have also been characterized for the genus \textit{Bacillus}.\textsuperscript{6} All of these variants have a zinc binding motif in the Lid domain. Despite not being a distinct feature of thermophilic ADKs, the zinc binding motif can thermally and chemically stabilize E. coli ADK while maintaining
Comparison of psychrophilic *B. globisporus* ADK, mesophilic *B. subtilis* ADK, and thermophilic *B. stearothermophilus* ADK show the same dependence on hydrophobic burial to stabilize the protein at the ideal temperature for its species of origin. ADKste has greater buried surface area than ADKsub, while ADKglo has less. Interestingly, salt bridges also seem to account for the stability difference between ADKste and ADKsub, despite playing no role in the stabilization of ADKsub relative to ADKglo. Three of four predicted salt bridges from ADKste are predicted by simulation to stabilize ADKsub, which is born out by ADKsub mutants. Chimeras of ADKsub and ADKste show that the key determinant of the global stability is the Core domain, whereas the Lid and AMPbd domains are determinants of optimal kinetic temperature. A large scale comparison of stabilization mutants found by bioinformatic and library methods recapitulated this trend. In addition to ADKbg, the ADK of another psychophile, *M. marinus*, has been characterized and the crystal structure has been solved. Despite a low RMSD with ADKbg, ADKmm shows a kinetic temperature dependence similar to mesophilic ADKbs.

A comparison by NMR of fast time scale (ps-ns) motions in *E. coli* and *A. aeolicus* ADKs showed that rate of fast motions corresponding to optimum kinetic temperature. NMR demonstrated that these motions are faster in ADKec than ADKaa at room temperature. Simulation predicts that at 80 °C, ideal growth temperature for *A. aeolicus*, ADKaa has fast motions comparable to ADKec at room temperature. An interesting homolog, *T. neapolitana* ADK, has similar kinetic parameters from 30 °C to 80 °C, making it the exception to this rule. H/D exchange NMR shows that while the
Lid and AMPbd domains in ADKtn are more rigid than ADKec, hinge regions which selectively unfold for closure over the active site have the same flexibility. Combined with the chimera results mentioned above, it may be reasonable to hypothesize that the Lid and AMPbd domains determine the upper bound of the activity range, while the hinge regions determine the lower bound.

Expression of split fragments of both ADKbs and ADKtn in a thermally-sensitive *E. coli* ADK knockout strain showed that thermophilic ADKtn was more successful in reassembling and complementing the knock-out strain. ADKtn fragments showed greater retention of secondary structure and zinc binding properties. Circularly permutated libraries of ADKbs and ADKtn in the same knock-out strain demonstrated a greater robustness of ADKtn. Both homologs were able to accommodate circular permutation in the Core and AMPbd domains, but only ADKtn could allow the N and C termini to be placed in the Lid domain.

Thermal adaptation of an ADK can be most directly observed in the work of the Shamoo group. They successfully produced a strain of thermophilic *Geobacillus stearothermophilus* with its natural ADK knocked-out and replaced with *B. subtilis* ADK. This range of available growth temperatures for this strain was limited by the melting temperature of ADKbs. Growth of bacteria over 1500 generations under conditions of slowly rising temperature yielded a small library of ADKbs variants with stabilizing mutations. These allowed bacteria to flourish at higher temperatures. At the thermal limit of ADKbs, a single stabilizing mutant appeared and dominated, followed by a burst of double-mutants and finally a single double-mutant dominating the culture.
“gateway mutant,” Q199R, created a new surface-exposed charge-charge interaction and resulted in the repacking of charged residues to create a network of further interactions. Both melting temperature and optimum kinetic temperature shifted upwards and $k_{\text{cat}}$ increased in a temperature dependent fashion. Calorimetric analysis of substrate binding showed that enthalpy now dominated over entropy. Characterization of Q199R double-mutants showed a variety of stabilization strategies at work – optimization of hydrophobic packing, creation of new salt bridges and hydrogen bonds, and an increase in ground-state dynamics. Most of the mutants lowered the rate of unfolding, while modestly raising the rate of folding. A193V, however, raised both rates by several orders of magnitude, with the ratio overwhelmingly favoring folding. Nearly all of the secondary mutants were destabilizing or neutral without Q199R present. One exception, Q16L, was independently stabilizing, but Q199R/Q16L proved greater than additive. Q199R/Q16L would reproducibly dominate culture growth.

The Wilson group used the Shamoo group’s set of mutations as both the scaffolds and training set for a multistate rational design algorithm. The approach could successful predict the stability and activity of all Shamoo variants and was used to generate 100 ADK variants. On the whole, the computational method was successful in predicted hydrophobic repacking, but failed in predicting the behavior of charge-charge interactions, despite recapturing key variants from the Shamoo set. Q199R/Q16L behaved poorly as a scaffold, producing only destabilizing mutations, leading the authors to hypothesize that Q199R/Q16L represents a fully optimized packing arrangement and
that further stabilizing could only be accomplished with adding charge-charge interactions.
Chapter 2: Analysis of transitional variants in the ADK Lid domain

**Contributions:** Initial bsADK and ecADK constructs were obtained and cloned by Deepti Mathur, an undergraduate of the Ohio State University. Dr. Deepa Perera, of Muskingum University, cloned the initial series of ADK variants (bsDi, bsTetra, bsHexa) and developed the protocol for kinetic measurements. Jendy Weppler, an undergraduate of Muskingum University, performed additional cloning work. Initial computational analysis was performed with the help of Venuka Durani, a graduate student of the OSU chemistry department. StickWRLD analysis of ADK was provided by Dr. William Ray of OSU. Project planning was the combined work of Dr. William Ray and Dr. Thomas Magliery. I personally contributed to the majority of the computational analysis, the cloning of several constructs, biophysical and kinetic measurements, NMR sample preparation and analysis, planning of further mutations, and later stage project planning.

Although we understand a great deal about the forces responsible for protein folding, it is difficult to use that information to predict the effects of mutations on protein structure and stability. It is a great deal harder still to predict mutational effects on function or dynamics. This limits our understanding of epistasis – that is, a protein’s retention of selective advantage while exploring sequence space to optimize for changing conditions and discover novel functions. One potential way to gain added
insight into the sequence determinants of protein structure and function is to statistically examine a multiple sequence alignment (MSA) of proteins with similar structures or functions. Some positions in proteins are found to be highly variable, and others have highly biased distributions or are even restricted to a single amino acid (i.e., are conserved). Likewise, the distributions in different positions may vary independently of other positions, or they may co-vary with other positions (i.e., be correlated or anticorrelated), even if they show little or no conservation independently.

A number of experiments have indicated that correlation information may capture additional structural or function information than consensus analysis alone. However, little is understood about how information is distributed between positional and correlated information, and the exact meaning of most detectable sequence correlations is unknown (if indeed it means anything at all in some contexts). For example, Ranganathan and colleagues showed that inclusion of correlation data in re-designs of WW domains resulted in a far greater fraction of folded proteins than when consensus alone was considered. Magliery and coworkers showed that correlation information was required to capture native-like surface electrostatics in tetratricopeptide repeats engineered from consensus. More recent work has suggested specific roles for sequence correlations in dynamics, allosteric regulation, or protein interactions. It may even be possible to segregate correlations into groups that act together (‘sectors’) in different aspects of protein structure or function. Ray et al recently suggested that networks of co-evolving residues function together in a biological analog of a Conditional Random Field to cooperatively determine function. (Ref needed) Our group
recently engineered from consensus information two variants of triosephosphate isomerase, which differ only in a small number of unconserved but highly correlated positions, one of which was well-folded and had wild-type like kinetics and one of which was molten and down 105-fold in activity.\textsuperscript{160} We went on to show that accurate prediction of the stabilizing effect of consensus mutations could be improved by removing sites that are correlated to other positions, suggesting a distributed role for sequence correlations in protein stability.\textsuperscript{161}

Recently, one of us developed a visualization modality for sequence correlation data called StickWRLD/MAVL. The tool was initially applied to nucleic acid sequences, but in its first description in application to proteins, an interesting sequence correlation was noted in the lid domain of the adenylate kinase (ADK) family.\textsuperscript{162,163} ADK (ATP:AMP phosphotransferase, EC 2.7.4.3) is a monomeric enzyme that catalyzes the disproportionation of two molecules of ADP into one AMP and one ATP. This enzyme is seen in many species and is essential to energy economy.\textsuperscript{49} Two consistent features of ADK are ATP-binding and AMP-binding domains, which together close to form the active site.\textsuperscript{1,5} Additionally, an important phylogenetic feature in ADK is a 27-residue loop insert found in bacterial ADK as well as ADK2 and ADK3 of vertebrates; this insert is identified as the Lid domain.\textsuperscript{122} Following substrate binding, the three domains close over the active site in a coordinated fashion, making bacterial ADK of great interest in dynamic studies. The lid domain has no direct catalytic role, but has been suggested to aid catalysis by excluding water from the active site.\textsuperscript{164}
Furthermore, important phylogenetic distinctions in bacteria can be identified on the packing of the lid domain. The lid domain of ADK from Gram-positive bacteria (ADKgp), like B. subtilis, has a zinc, cobalt, or iron chelating site, a feature absent from the ADK of Gram-negative (ADKgn) species such as E. coli (with certain thermophilic exceptions).\textsuperscript{126,130,165}

The dynamics and catalytic mechanism of ADKgp are not constitutively different from ADKgn, although the optimal catalytic temperatures differ based on the ‘flexibility’ of each enzyme.\textsuperscript{166,167} Dynamic characterization of ADK structures by FRET and NMR have shown that ADK accesses its catalytically active conformation in the absence of substrate, transitioning through multiple native states.\textsuperscript{168} ADK activity, therefore, depends not only on global stability, but also on maintaining the energetic balance of its multiple states, thus requiring that certain folded regions are not so stable that they cannot unfold easily at room temperature and that relatively unfolded regions are not too far below the threshold of stability.

In the StickWRLD/MAVL analysis of the ADK lid domain, two additional positions were observed to have significant correlations to the positions occupied by the chelating amino acids in ADKgp. A close inspection showed that those two residues, while variable across all lid domains, were largely unconserved in the ADKgp variants but were highly biased in the ADKgn variants. Based on this, we hypothesized that the non-chelating lid of ADKgn variants might be swapped into ADKgp variants, but only if the additional correlated positions were also interconverted. Of course, we did not know
**Fig. 2.1** A.) Structure of ecADK in the open conformation. Highlighted are the phosphate binding loop (orange), the AMP binding domain (red), and the Lid domain (purple). B.) Aligned structures of ecADK and bsADK in the closed conformation, with the substrate AP₅A.⁵,⁶

**Fig. 2.2** Comparison of lid domains from *B. subtilis* (a) and *E. coli* (b) with residues of interest highlighted in blue.
how these correlated positions might be important (that is, to stability, function, dynamics, etc.). To further explore the interrelatedness of positions in the ADK bacterial family, we carried out a mutual information analysis. Mutual information (MI) uses patterns of co-evolution to score the amount of information each position in a MSA contains about all other positions. \(^{30,158}\) Information content can be defined as the extent to which the selection of a particular identity at one position lowers or raises the informatics entropy at another – the greater the change in entropy, the more information that is present. \(^{169}\) MI, however, does not reveal the underlying changes in sequence responsible for the positional correlation. To do that, we use a method we call perturbation, which is similar to the method used by Ranganthan and colleagues in the original implantation of statistical coupling analysis (SCA) \(^{22,158}\) except that it measures the change in distribution using relative entropy (RE), an information theoretic measure of the “distance” between two distributions. Here, we present that analysis, which confirms the sequence correlations observed in StickWRLD/MAVL and suggests further correlations, as well.

To examine the effects of these correlated mutations on ADK and the lid domain, we swapped the amino acids in the four positions that ligate a metal ion in ADKgp from \textit{B. subtilis} for the ones found in ADKgn from \textit{E. coli}. We additionally mutated the correlated positions in the \textit{B. subtilis} background, and we examined the effect of those mutations one at a time. It was previously shown that incorporation of a zinc chelating motif into \textit{E. coli} ADK produces an enzyme with an elevated melting point and lowered activity. \(^{170}\) Here we show that the reverse conversion results in an inactive enzyme, but
that activity is partially restored by the correlated mutations. However, the correlated mutations have essentially no effect on the overall stability of the protein.

During out statistical analysis of ADK sequences, we also found two highly correlated residues with no apparent correlation to other parts of the structure. These residues sit on the interface of the LID domain and the AMP binding domain, which come into contact during the closed conformation of the enzyme. Using these residues as a starting point, we discovered that the inclusion of an additional positive charge at this interface greatly enhances the activity of bsADK. However, the removal of an analogous charged residue from ecADK does not alter activity. This raises questions about the role of charge-charges in ADK and their relative evolutionary flexibility.

MATERIALS & METHODS

Sequence Alignment

Sequences and alignment were downloaded from the ADK_LID category of the Pfam 23.0 database (http://pfam.sanger.ac.uk/) on 6/17/09, specifically limiting the species to bacteria.\textsuperscript{171} Given that our species of interest (\textit{E. coli} and \textit{B. subtilis}) both had LID domains 36 residues in size, we only kept aligned sequences between 20 and 40 residues. We then removed any positions with less than 50% occupancy. The final alignment had 999 sequences and 36 positions.

The full length alignment was made with sequences taken from the UNIPROT database,\textsuperscript{172} as downloaded on 07/29/09. The initial sequences for the seed alignment (21 bsADK homologs and 12 ecADK homologs) were found through NCBI Blast\textsuperscript{173} searches.
for bsADK and ecADK. HMMER 3.0\textsuperscript{174} was installed locally from the binary made available through the Sanger Institute and used to generate the seed HMM and the alignment.

\textit{Mutual Information Analysis}

In information theory, Mutual information (MI) measures the data a given output provides about the possible identities of the input.\textsuperscript{169} Directly applied to the analysis of multiple sequence alignments, MI measures the constraints the identity at one position places on the identity at another. MI is logarithmically related to the multinomial probability of observing the actual joint distribution of two positions if the distribution frequencies at the two positions are actually independent. The higher the MI, the further the actual joint distribution is from independence, and thus the more complex the relationship between positions. MI can be computed from the relative entropy between the joint distribution and the distribution predicted from the product of the frequencies in the individual distributions:

\begin{equation}
\text{MI}(x,y) = \sum_{y=0}^{n} \sum_{x=0}^{n} P_{(x,y)} \ln \left( \frac{P_{(x,y)}}{P_x P_y} \right)
\end{equation}

Calculations were performed using Microsoft Excel 2003. To account for background biases in the alignments, ten versions of the alignment with randomized identities at each position were generated and the MI calculations were applied for each.
The average maximum MI value for these ten alignments was taken to be the “noise” cut-off for the true alignment MI.

**Perturbation Analysis**

The frequency of co-occurrence for different combinations of identities at the positions of interest was observed by using an Excel spreadsheet to filter sequences for subsets of interest. Filtering was accomplished by selecting sequences that met a given criteria at positions of interest. The amino acid distributions at these positions were counted and compared to the values from the full set. The total number of unique combinations was found by concatenating alignments positions and filtering out repeats.

**Primer Construction and Cloning**

The gene for *B. subtilis* wild-type *adk* was obtained from a genomic prep prepared by the Taylor Lab at OSU. *E. coli* wild-type *adk* was obtained by colony PCR from DH10B cells. Both genes were incorporated into a pHLIC plasmid with ampicillin resistance, an N-terminal 6×His tag with TEV-site linker, and a T7 transcription site. The Di, Tetra, Hexa, Penta-135, and Penta-155 mutants were created by mutation of the *B. subtilis adk* construct by overlap PCR. Similarly, the Cys4 mutation was derived from the *E. coli adk* construct by overlap PCR. All constructs were cloned and stocked in the DH10B *E. coli* cell line, and were confirmed by sequencing (Genewiz).
ADK Expression and Purification

Electrocompetent BL21(DE3) E. coli cells were transformed with the pHLIC plasmids bearing the adk genes, and grown in 2YT media at 37 °C with 1 mL of 100 mg/mL Ampicillin. Upon reaching an appropriate cell density, expression of 6×His-tagged protein was induced by addition of 0.1 mM Isopropyl β-D-1thiogalactopyranoside (IPTG). After three hours of induction, cells were harvested by centrifugation; resuspended in 50 mM Tris-HCl (pH 7.8), 300 mM NaCl, 10 mM imidazole, 6 mM MgCl₂, 6 mM CaCl₂, and 1.2% Triton-X; and lysed by sonication. Soluble material was separated by centrifugation at 15K rpm and 4 °C for 30 minutes. Lysis supernatant was mixed with 1 mL of Ni-NTA resin from Thermo Scientific and incubated overnight at 4 °C. The resin was separated from the supernatant by decanting into a Bio-rad 20 mL pre-fretted chromatography column and washed with 50 mM Tris pH 7.8, 300 mM NaCl, and 50 mM imidazole; then protein was eluted with 50 mM Tris pH 7.8, 300 mM NaCl, and 200 mM imidazole. Eluted protein was incubated at 30 °C with 5 mM DTT and TEV protease overnight, to remove the 6×His tag and linker region. A second Ni-NTA purification was used to separate TEV-cleaved tag and any uncleaved protein from protein with the tag removed. Protein samples intended for kinetic characterization were dialyzed into 50 mM Tris-HCl (pH 7.8), 50% glycerol and stored at -80 °C. Samples intended for CD characterization were dialyzed into 100 mM sodium phosphate buffer (pH 7.8) and 300 mM NaCl for short-term storage at 4 °C. Sample concentrations were determined by 1.) absorbance at 280 nm after the sample was confirmed to be pure on
SDS-PAGE gel (stained with Coomassie Brilliant Blue dye) and 2.) optical density comparison of the sample on SDS-PAGE gel to serial dilutions of 2 mg/mL of bovine serum albumin.

**CD Spectroscopy and Thermal Denaturation**

CD measurements were taken on a Jasco J-318 Circular Dichroism spectrometer (Ohio State Department of Chemistry & Biochemistry). 20 μM samples were prepared in 100 mM phosphate buffer (pH 7.1) with 300 mM NaCl, and full-spectrum measurements were taken at 25 °C with a path length of 1 mm. Thermal denaturation was tracked at 222 nm at a rate of 1 °C/min over a range of 25 °C to 95 °C. Data was fit as described by Matouschek and Fersht.176

**NMR Analysis**

Protein samples for NMR were N\textsuperscript{15} labeled during expression in minimal media.177 1H-\textsuperscript{15}N heteronuclear single quantum coherence (HSQC)\textsuperscript{178} spectra were collected on a Bruker DRX 600 MHz spectrometer and analyzed using the NMRpipe software package.179
**Kinetic Characterization**

Variants were characterized by the rate of catalyzing the conversion of AMP and ATP to ADP, at saturating ATP and as a function of AMP concentration. Rate of the forward reaction was monitored through a coupled reaction consuming NADH, which absorbs at 340 nm.\(^5\)

![Full reaction schematic of kinetic experiments.](image)

Scatter plots of kinetic data were fit to curves using Kaleidagraph v. 4.03. ADK is known to display substrate inhibition to AMP. Accordingly, when substrate inhibition was observed, we fit the curve to a Bi-Bi iso-random mechanism.
\[ V = \frac{V_{\text{max}[AMP]}}{K_m + [AMP]\left(1 + \frac{1}{K_i}\right)} \]

**RESULTS**

*Alignments*

Previous computational work using StickWRLD/MAVL showed a strong network of correlations between the four positions involved in the zinc-chelating motif of the ADK lid domain, as well as several intersecting networks of weaker correlations branching off of each chelating position.\(^{162}\) We speculated that these correlations were the result of hydrogen bonding interactions in the lid domains of ADK from Gram-negative bacteria. Using correlation analysis, we sought to determine the minimal mutations required to restore a stable, functional hydrogen bonding network to the lid. Aligned sequences for Adenylate Kinase homologs were obtained from the Pfam online database, under the ADK and ADK LID family entries. Both families have entries for *E. coli* and *B. Subtilis* ADK. The usefulness of the alignments were decided by a.) successful alignment of bsADK and ecADK within the larger alignment and b.) high degree of occupancy at positions corresponding to bsADK and ecADK residues. For the ADK lid domain alignment, we found that the bs and ec sequences aligned perfectly to each other (36 residues matched to 36 residues) and that among the aligned positions, all had an occupancy above 95% (with a maximum of 99.23% for the entire alignment). The full sequence ADK alignment available from Pfam showed the bs and ec sequences to be
well-aligned, except for a four residue insert in bsADK at the c-terminal of Helix 5 (resi 98 to 102). Positions varied in occupancy from 37% to 98%, with a mean of 82%.

However, a major issue with the Pfam ADK alignment was the absence of the c-terminal portion of the Core domain. This fragment is conserved between bsADK and ecADK, but is poorly aligned by HMMer due to truncated sequences used in the seed alignment.

We chose to generate a new full sequence ADK alignment using the program HMMer\textsuperscript{174} and a locally saved copy of the UNIPROT database. We generated the seed alignment in ClustalW\textsuperscript{180} using 12 homologs of ecADK and 21 homologs of bsADK, owing to a higher level of similarity among ecADK homologs. The resulting alignment not only covered the whole sequences of bsADK and ecADK, but also showed higher relative occupancy overall. Occupancy varies from 14% to 99%, but the mean is 92.7% and visual inspection of the occupancy-per-position shows that the low occupancy positions in the c-terminal region are excluded from the Pfam seed alignment. The insert in bsADK (resi 98 to 102) was recapitulated; additionally, an insert was seen in ecADK from positions 189 to 194, a similar loop following the c-terminal of a helix. (Fig. 2.4) Noteworthy is that the ecADK 189-194 insert occurs significantly less frequently than the bsADK 98-102 insert. We chose to pursue correlation study using this alignment and the Pfam ADK LID alignment.

Initially, the ADK lid alignment contained 1039 sequences aligned over 67 positions and the full sequence alignment contained 717 sequences aligned over 481
positions. Two steps were taken to lessen the influence of outliers. First, the lid alignment was filtered for sequences between 20 and 40 residues and the full alignment for sequences between 200 and 256 residues. Second, all positions below 50% occupancy were deleted, bringing the lid alignment down to 36 positions and the full alignment to 214 positions. Doing so eliminates the ecADK(189-194) insert and the bsADK(215-217) c-terminal from consideration in the full alignment. Relative entropy calculation, which captures the relative level of conservation at each position, shows an even distribution of

Fig. 2.4 Occupancy of each position obtained from both the Pfam and our custom alignment of full length ADK.
conservation and variable positions across the full alignment, with a slightly lower average at the c-terminal positions. The highest conservation in the LID alignment is at the chelating positions; the n- and c-terminals of the domain where electrostatic contacts are made with the binding site; and at two prolines in positions 17 and 18.

Mutual Information

Satisfied with our curation, we used Excel to calculate the JDRE (joint distribution relative entropy) for both alignments. (See methods.) This serves as an estimation of the mutual information between the alignment positions. For both experiments, we generated ten “broken correlation” alignments in which the identities at each position had been randomized by column. The average maximum JDRE between the ten randomized alignments was taken to be the potential noise level of the alignment, generated by phylogenetic artifacts and relative rate of mutation. The noise level served as the cut off of possible relevance of a pair of positions. We found that for the lid domain alignment, the noise lay below the average MI (Noise: 0.07, Average: 0.17), while for the full length alignment, the noise greatly exceeded the average (Noise: 0.60, Average: 0.17).

When the lid domain correlations are plotted to a heat map (Fig. 2.7), it is immediately obvious that positions 4, 7, 24, and 27 strongly correlate to each other – these correspond to the chelating motif in bsADK. Positions 4, 7, and 27 also strongly correlate to position 29, and position 27 makes an additional strong correlation to position 9. The 4-9, 7-9, and 24-9 correlations all appear at a mid-tier level, as does the 24-29
**Fig. 2.5** Alignment of *E. coli* and *B. subtilis* sequences as appears in our full length ADK alignment. Green indicates β-sheets and red indicates α-helices. Asterisks denotes positions implicated in zinc chelation. Brackets delineate the lid domain.
Fig. 2.6 Conservation as scored by relative entropy for the full length ADK and ADK lid alignments.
correlation. The strongest correlation in the lid alignment is between positions 5 and 32, but these do not show strong correlations to any other positions.

The MI plot for the full length alignment is dominated by high correlations with the 98-102 and 189-194 regions – these correspond to the loop length differences between bsADK and ecADK. However, in the region corresponding to the lid domain, we find that the pattern of correlations from the lid alignment is recapitulated by mid-tier correlations. The lid residues also show correlations to 98-102 and 189-194 regions, but given that these loops vary widely in length and show correlations with the entire sequence, it is hard to interpret what specific interactions they might have. The strongest correlation between those regions and the lid is through position 131, which is synonymous with position 5 in the lid alignment. Although we will not pursue it further in this work, this correlation between loops of differing length may be related to observations by the Haas group, that loop closure precedes secondary structure formation for the ADK core domain.\textsuperscript{90,181} Curiously, the 18-21 region, which corresponds to part of the ATP-binding Walker A motif, shows several strong correlations but little with the lid domain – despite forming opposite sides of the ATP binding pocket.\textsuperscript{56}

\textit{Perturbation and Mutant Selection}

We sought to understand the cause of the correlations in the lid domain by looking at all apparent combinations of the residues in the alignment. Although this does not strictly tell us what is structurally permissible, we believe it can elucidate what is structurally
Fig. 2.7 A.) Heat map of the mutual information between positions in the ADK lid alignment. B.) Correlations from the entire protein alignment. C.) Location of major correlated positions in relation to the bsADK Lid domain. (PDB: 1P3J)
favored. We found in the lid alignment 984 out of 999 sequences in which positions 4, 7, 9, 24, 27, and 29 were all occupied. In those sequences, there were 66 unique combinations with the chelating motif and 7 unique combinations without. However, there were 440 total sequences with the chelating motif, but 540 sequences lacking the motif. Moreover, 515 of the non-chelating sequences had the combination H(X2)S(X)R(X14)D(X2)T(X)E, while only 207 of the chelating sequences, less than half, had the combination, C(X2)C(X)A(X14)C(X2)C(X)G. This implies that the chelating motif places fewer constraints of the identities of positions 9 and 29. Examination of the top 20 most prevalent combinations demonstrates that sequences with the chelating motif allows a wide variety at positions 9 and 29, with little common chemical properties, and that the non-chelating sequences restrictively show R at position 9 and EDQ at position 29. Both types allow some small variation at position 27 – C/D is allowed in chelating sequences and T/S is allowed otherwise. The same exploration of the prevalent combinations in the full length alignment, using the corresponding positions 130-133-135-150-153-155, shows a similar pattern, with differences most likely due to the lower number of sequences (551 sequences containing 59 unique combinations, 50 sequences with the chelating motif and 9 without). It is notable that R and EDQ are permissible identities for positions 9/135 and positions 27/155 in the context of the chelating motif and do in fact appear among the top twenty most occupied combinations. We used these correlations as a guide when planning our mutations in bsADK and ecADK. ecADK belongs to the H(X2)S(X)R(X14)D(X2)T(X)E subset, which, mentioned above, dominates the distribution of non-chelating sequences. bsADK belongs to a subset substituting the
fourth cysteine in the chelating motif with an aspartate, which is composed of 56 sequences in the LID alignment, with 16 combinations present. bsADK belongs to a further subset C(X₂)C(X)T(X₁₄)C(X₂)D(X)G that contains a plurality (21 of 56) of sequences. Within the D27 subset, R9 and E/Q29 are still permissible identities. The correspondence between the alignment positions and the bsADK and ecADK residues and the planned constructs are shown in Table 2.1.

**Fig. 2.8** Top twenty populated perturbations of positions 4-7-9-24-27-29 (ADK lid alignment) and 130-133-135-150-153-155 (full length ADK alignment). The ADK lid sequences are taken from a total population of 984 sequences and the full length ADK sequences are taken from a total population of 551.
Fig. 2.9

Top: Sequence logo of the curated alignment downloaded from the Pfam database for the ADK lid domain in bacterial species.  Middle, Bottom: Sequence logos for the 4H (Gram-negative) and 4C (Gram-positive, chelating) subsets.  Note that 9 and 29 are only weakly conserved among Gram-positive species but are entirely conserved among Gram-negative.  Sequence logos were created using the Web Logo online service.³
The ecADK Cys4 construct is identical to one already produced by Perrier et al. The bsADK constructs still show the greatest homology to bsADK itself, while a NCBI Blast search identifies identical-length ADK of *B. amyloliquefaciens* and *B. mojavensis* to be the closest homologs, as is also true for the wild-type sequence. Moreover, the wild-type shows a great percent identity with these homologs than the mutants, assuring us that we are producing unique constructs.

Using overlap PCR, a series of mutated constructs was generated for both *B. subtilis* and *E. coli* ADK. The *E. coli* Cys4 construct (H126C S129C D146C T149C), which embeds the consensus chelating motif in *E. coli* ADK, is identical to one previously described. We initially engineered three variants of the *B. subtilis* ADK, with mutations only at the chelating residues (bsTetra, C130H C133S C150D D153T), only at the correlated positions (bsDi, T135R G155E), or at all six of the chelating and correlated positions (bsHexa).

<table>
<thead>
<tr>
<th>ADK lid</th>
<th>full length ADK</th>
<th>bsADK</th>
<th>ecADK</th>
<th>Construct</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>130</td>
<td>C130</td>
<td>H126</td>
<td>T135R G155E</td>
</tr>
<tr>
<td>7</td>
<td>133</td>
<td>C133</td>
<td>S129</td>
<td>bsTetra C130H C133S C150D D153T</td>
</tr>
<tr>
<td>9</td>
<td>135</td>
<td>T135</td>
<td>R131</td>
<td>bsHexa C130H C133S T135R C150D D153T G155E</td>
</tr>
<tr>
<td>24</td>
<td>150</td>
<td>C150</td>
<td>D146</td>
<td>bsPenta-135 C130H C133S T135R C150D D153T</td>
</tr>
<tr>
<td>27</td>
<td>153</td>
<td>D153</td>
<td>T149</td>
<td>bsPenta-155 C130H C133S C150D D153T G155E</td>
</tr>
<tr>
<td>29</td>
<td>155</td>
<td>G155</td>
<td>E151</td>
<td>ecCys4 H126C S129C D146C T149C</td>
</tr>
<tr>
<td>5</td>
<td>131</td>
<td>S131</td>
<td>A127</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>158</td>
<td>Y158</td>
<td>T154</td>
<td></td>
</tr>
<tr>
<td>--</td>
<td>161</td>
<td>A161</td>
<td>K157</td>
<td></td>
</tr>
</tbody>
</table>
We were also interested in the correlation between LID alignment positions 5 and 32, due to the high mutual information score and seeming isolation from any other network for correlations. These positions correspond to S131 and Y158 in bsADK and A127 and T154 in ecADK. When initial kinetic results suggested no particular interaction between these residues, we expanded our investigation to nearby residues using the available crystal structures of substrate-bound bsADK and ecADK. Residue K40 is conserved in both homologs and while it appears to interact with bsADK Y158, its ecADK structure is perturbed due to nearby K157. Since K40 remains constant, we chose to repeat the 5/32 mutations in the context of A or K at position 161.

**Fig. 2.10** Comparison of the LID domain and AMP binding domain interface for bsADK bound to AP3A (left) and ecADK bound to ADP and AMP (right).
In the full ADK alignment, position 40 is a highly conserved position with a positively charged side chain in 257 sequences. Of this subset, 45% have a negatively charged side chain at position 161, with a strong preference for aliphatic or hydrophobic amino acids at positions 131 and 158. This immediately suggests that the electrostatic interaction between 40 and 161 and the burial of 131 and 158 help drive interface closure. However, polar and even charged side chains still appear at positions 131 and 158. When position 161 is neutral or positively charged, these polar residues are enriched at 131 and 158. The interface between the LID domain and the AMP binding domain, then, must contain the potential for multiple modes of interaction which allow for such a wide walk through sequence space.

Structure and Stability

All variants were expressed in *E.coli* at 37 °C and purified via cleavable affinity tags. All purification steps took place at room temperature. Samples were then transferred to buffer for storage and characterization by dialysis. The structure of the wild type proteins and variants was probed by circular dichroism spectroscopy. Mutations in the lid domain had little overall effect on the shapes of the far-UV CD spectra, suggesting the same overall folds. The spectra from both the bsADK and ecADK variants are consistent with a mixed α/β protein. ecADK is 16% β-sheet and 50% α-helix, while bsADK is 16% and 51%, respectively. The presence or absence of the metal-binding motif had little effect on the spectral shapes. Although there is some variation in the mean-residue ellipticity,
these spectra likely indicate that the variants fold very similarly to the wild-type scaffolds.

We also monitored the thermal denaturation of the variants by observing the ellipticity at 222 nm as a function of temperature. ecADK and bsADK both melt at 60 °C. Addition of the ecCys4 chelating motif stabilized the *E. coli* homolog 12 °C, consistent with the results of Perrier et al. Replacement of the chelating residues in bs ADK with the polar residues in ec ADK destabilized the protein about 10 °C. However, in the bsADK scaffold, the correlated mutations T135R and G155E had almost no effect on the stability (i.e., wild-type versus Di, and Tetra vs. Hexa). These results show that the chelating motif greatly stabilizes the protein, while the T135R and G155E mutations do little to affect stability of the protein in the context of either the chelating motif or the polar residue network. Surprisingly, the positions correlated to the polar network did not offset the loss of stability from the loss of the chelating motif as we had expected. All of the ADK variants in this study melt irreversibly.

It is important to note – in previous literature, the melting temperature of bsADK is reported for 50 mM phosphate and zero NaCl. Under these conditions, the melting temperature reported is 50 °C. In these experiments, buffer conditions included 300 mM NaCl. Salt concentration serves to increase the melting temperature of bsADK, with a linear increase as NaCl concentration is increased from 0 to 400 mM, over a range of 50 to 60 °C. ecADK does not demonstrate a similar property, with identical melting temperatures at 0 and 500 mM NaCl. When BS Tetra is tested under similar conditions,
Fig. 2.11 Mean residue ellipticities (MRE) from far-UV CD scans of (a) *B. subtilis* wild type ADK and variants and (b) *E. coli* wild type and variant.

Fig. 2.12 Thermal denaturaturation of the ADK variants monitored by ellipticity at 222 nm.
we find a linear dependence similar to bsADK, but over a range of 40 to 50 °C. We chose to maintain these conditions for ease of handling the destabilized variants.

To further characterize the structural differences between variants, we collected $^1$H-$^{15}$N HSQC spectra for bsADK, BS Di, BS Tetra, and BS Hexa. As expected from previous NMR studies of ADKs, bsADK has a well-dispersed spectrum, while BS Di produces an overlapping spectrum with wild-type. BS Tetra and BS Hexa have identical spectra with peaks clustered in 8.0 to 8.5 ppm region, indicative of rapid conformational interconversion.

Fig. 2.13 $^1$H-$^{15}$N HSQC spectra of bsADK (red) and BS Tetra (blue).
Kinetic Characterization

We characterized the activity of the ADK variants by monitoring the rate of consumption of adenosine 5’ monophosphate (AMP) coupled to the oxidation of nicotinamide adenine dinucleotide (NADH), monitored by absorption at 340 nm. The reactions were carried out at 37 °C, in accordance with kinetic measurements carried out for *B. stereothermophilus* ADK, the archetypical Gram-positive ADK variant. The optimal reaction temperature for ecADK is 30 °C, but we chose to compare all reactions at the same temperature. To validate our experimental procedure, we compared the reaction rate obtained for ecADK at 300 μM AMP and 30 °C to an expected result calculated using constants presented by Glaser et al. The result was 57000 min⁻¹ versus 43000 min⁻¹, a difference of only 1.3 fold and acceptable for our purposes.

bsADK activity appears to be below ecADK activity, but bsADK is known to exhibit substrate inhibition by AMP as we observe here. The $k_{cat}$ and $k_{cat}/K_{m}$ values are within two-fold for the wild-type enzymes, although the $K_i$ for bs ADK is 420 μM, compared to the $K_m$ of 91 μM. ecADK does not exhibit significant substrate inhibition by AMP until the low millimolar range, although the reduction in rate at the highest AMP concentration for E. coli wild-type and Cys4 may be due to substrate inhibition. ecCys4 was down only 10% in activity compared to wild-type, although the $k_{cat}$ is reduced somewhat more (2.5-fold) with a 2-fold improvement in the $K_m$. 

75
In contrast, the Tetra mutant of B. subtilis ADK is virtually devoid of activity (down 400-fold to 80 M$^{-1}$ s$^{-1}$ from $3 \times 10^5$ M$^{-1}$ s$^{-1}$). As was observed for the structure and stability, mutation to the correlated residues alone in the context of the chelating motif (Di) had little effect on the activity, especially taking into the account the relative uncertainty of the Ki and its effect on the uncertainty in $k_{cat}/K_m$. In contrast, the two correlated mutations rescue a significant amount of the activity relative to the Tetra mutant, resulting in a 30-fold increase for Hexa over Tetra and a $k_{cat}/K_m$ within 15-fold of

![Fig. 2.14 Rate/[ADK](μM) versus [AMP] (μM). For the complete coupled reaction scheme, see Methods. The initial reaction velocity at 37 °C, normalized for concentration of ADK, is plotted as a function of concentration of AMP. B. subtilis wild type and Di show substrate inhibition characteristic of B. subtilis adenylate kinase.](image)
the wild-type. Surprisingly, the addition of the correlated mutants restored a considerable amount of activity to the apparently-incomplete polar motif. Moreover, the correlated mutations restore that activity without restoring any of the stability lost upon removal of the metal binding site, demonstrating that the loss in Tetra activity is not simply due to a loss in global stability. The Tetra and Hexa mutants show no indication of substrate inhibition in the concentration range we tested. This is somewhat surprising, since the presumed mechanism of substrate inhibition is binding of AMP to the ATP site. The increase in $K_m$ with respect to AMP would be expected to make that inhibition more pronounced; the lack of it suggests that the mutations to the lid domain are affecting the interaction of the AMP and ATP binding sites. Kinetic parameters are summarized in Table II.

<table>
<thead>
<tr>
<th></th>
<th>$T_m$ (°C)</th>
<th>$K_m$ (μM)</th>
<th>$K_{cat}$ (Min$^{-1}$)</th>
<th>$K_I$ (μM)</th>
<th>$K_{cat}/K_m$ (M$^{-1}$ S$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC wt</td>
<td>60.5</td>
<td>140 ± 30</td>
<td>3300 ± 200</td>
<td>--</td>
<td>30000 ± 10000</td>
</tr>
<tr>
<td>EC Cys4</td>
<td>72.2</td>
<td>70 ± 30</td>
<td>1300 ± 100</td>
<td>--</td>
<td>30000 ± 10000</td>
</tr>
<tr>
<td>BS wt</td>
<td>60.9</td>
<td>90 ± 7</td>
<td>1900 ± 80</td>
<td>415 ± 34</td>
<td>30000 ± 3000</td>
</tr>
<tr>
<td>BS Di</td>
<td>59.6</td>
<td>40 ± 20</td>
<td>2200 ± 500</td>
<td>317 ± 127</td>
<td>80000 ± 50000</td>
</tr>
<tr>
<td>BS Tetra</td>
<td>48.5</td>
<td>900 ± 200</td>
<td>46 ± 7</td>
<td>--</td>
<td>80 ± 30</td>
</tr>
<tr>
<td>BS Hexa</td>
<td>47.6</td>
<td>200 ± 40</td>
<td>300 ± 20</td>
<td>--</td>
<td>2500 ± 500</td>
</tr>
<tr>
<td>BS Penta-135</td>
<td>47.7</td>
<td>400 ± 40</td>
<td>600 ± 30</td>
<td>--</td>
<td>2500 ± 300</td>
</tr>
<tr>
<td>BS Penta-155</td>
<td>48.4</td>
<td>2000 ± 1000</td>
<td>80 ± 50</td>
<td>--</td>
<td>60 ± 60</td>
</tr>
</tbody>
</table>

Table 2.2
**Penta Mutants**

To understand better the roles of the correlated positions, we also constructed two mutants combining the Tetra mutant (in the *B. subtilis* background) with the individual correlated mutations: Penta-135 (Tetra T135R) and Penta-155 (Tetra G155E). Both of these variants had very similar far-UV CD spectra to Tetra and Hexa, and their thermal stabilities were approximately the same as Tetra and Hexa. Surprisingly, Penta-135 has catalytic parameters that are essentially identical to Hexa, while Penta-155 has catalytic parameters essentially identical to Tetra. This suggests that the T135R mutation is alone responsible for the 30-fold rescue in activity of the Tetra mutant, and that the G155E mutation has virtually no effect. This is especially surprising since the correlation analysis allowed us to identify T135R as being important to the non-chelating lid motif, and G155E is correlated to both the T135R as well as to the mutations to the metal binding residues. Of course it is possible that the G155E mutation co-evolves with T135R for a reason other than what we assaying here (AMP + ATP → 2ADP activity and thermal stability). For example, the reverse reaction, solubility, expression level, or some kind of regulation might require G155E. As with the Tetra and Hexa mutants, there is no sign of significant substrate inhibition for either of the Penta variants at the concentrations tested here.
Fig. 2.15 MRE and thermal denaturation of bsPenta-135 and bsPenta-155 compared to bsTetra and bsHexa.

Fig. 2.16 Kinetic curves of bsPenta-135 and bsPenta-155 compared to bsTetra and bsHexa.
LID-AMP binding interface

We generated a second series of mutants in bsADK and ecADK to investigate the role of the correlations found on the LID side of the interface between the LID and AMP-binding domains. Single and multi-mutants were made for S131A/Y158T/A161K in bsADK and A127S/T154Y/K157A in ecADK. Initial melts made in 50 mM phosphate buffer and no NaCl shows little shift in melting temperature for most of the single mutants. The cumulative effect of combined ecADK mutants lowers the melting temperature in a consistent fashion. The Y158T single mutant of bsADK shows a greater decrease in melting temperature relative to other single mutants, with a similar result for the S131A-Y158T multi-mutant. The bsADK A161K mutant does not appear to stabilize bsADK relative to wild-type, but seems to recover some stability lost from Y158T.

Kinetic assays demonstrate that for ecADK, the K157A mutation appears very close to wild-type in activity, while T154Y and A127S are debilitating to activity as both single and multi-mutants. S131A and Y158T show similar debilitation for bsADK. Significant is that in all cases, A161K greatly improves activity of bsADK and appears to give the greatest improvement for the S131A-Y158T-A161K triple mutant. This suggests that an evolutionary mechanism exists for bsADK to recover binding strength in this interface, which has diminished in importance for ecADK.
DISCUSSION

The loss of ADK activity from making major structural perturbations in the Lid domain is easily predicted. It has long been known that the Lid domain folds independently and that denaturating conditions can lead to inactivation of ADK before global unfolding transitions occur. A recent model describes the open-to-closed transition of ADK as a stepwise build-up of strain in the Lid domain, followed by relief through conformational transition. This model can be corroborated by experiments which show that the introduction of conformationally-relaxing mutations into the Lid lowers ADK activity, while keeping identical contacts to AP5A in the bound state. The loss of zinc chelation, then, considerably lowers the amount of conformational restriction in the Lid and would

**Fig. 2.17** Kinetic activity curves for interface variants
be expected to decrease the rate of this transition. Normally, ADKs exist in solution as an
ensemble of open and closed states, favoring the closed, and addition of substrate pushes
this ensemble to nearly 100% closed molecules. Perturbation of contacts which
stabilize the closed state have been shown to shift this equilibrium towards the open
state. Our HSQC results show that the no-zinc variants, although folded, exist as a
much more flexible ensemble, i.e. predominantly open. Millisecond NMR of ADK
provides a linear combination of peaks between the open and closed states, which is
sensitive to the open/closed equilibrium. If the T135R mutation restored activity by
restoring conformational restraints and shifting this equilibrium back towards the closed
state, then bs ADK, bs Tetra, and bs Hexa should all have unique HSQC spectra. The
similar spectra of the inactive and restored variant, then, imply they have a similar
open/closed equilibrium.

Examining the kinetic parameters for bs Tetra, bs Hexa, and bs Penta-T135R, we
can see that the $K_m$, within error, differ by a factor of 3 or less. Rather, $k_{cat}$ makes the
greatest contribution to the increase in activity. The difference between $k_{cat}$ for bs Tetra
and bs Penta-T135R is greater than an order of magnitude. The work of the Kern group
measuring ADK dynamics by $\mu$s and ns-ps NMR found the open-to-closed and closed-to-
open transitions to be on the same time scale as the overall reaction rate, with the closed-
to-open transition being the overall rate determining step. In normal apo ADK, the
closed-to-open transition in slower than the open-to-close rate. Since the no-zinc variants
are predominantly in the flexible, open conformation, it is logical to infer that in these the
open-to-closed transition is rate limiting. This implies that T135R does indeed provide
interactions that favor the closed state, but that closed species exist fleetingly in solution, on a time scale far below the wild-type. Despite this, they are able to catalyze phosphotransfer within an order of magnitude of the wild-type.

A similar hypothesis may apply to the activating qualities of the A161K mutation. As we noted in the comparison of the *E. coli* and *B. subtilis* ADK crystal structures, K40 is conserved, but must be reoriented to account for a second nearby lysine. A computational model by Beckstein and coworkers demonstrates that a collection of charge-charge contacts stabilize the interface between the Lid and AMPbd domains during the closure transition. Disruption of one or more of these contacts should destabilize the closed state and speed up the rate-limiting opening transition, as A161K demonstrates. Noteworthy is that the correlated mutation pair, S131A and Y158T, lower the reaction rate in the context of the domain interface’s normal charge-charge interactions – presumably by aiding the opening transition or impairing the closing transition until it becomes rate-limiting – but then increase the reaction rate even more when the normal charge-charge interactions have been disrupted. Given that S131A and Y158T are both size-decreasing mutations, the simplest hypothesis is that increasing space in the interface between Lid and AMPbd disrupts the packing arrangements which optimizes activity -- let us stress, we cannot say if this makes the interface more or less stable – but that this abandonment of optimized packing is preferred when a fortuitous rearrangement is stumbled upon. However, we cannot discount the possibility that changes in the side chains points to new interactions along the trajectory of open/closed
transition. All we can be certain of presently is that these residues are very sensitive to the change in context brought about by a single nearby mutation.

Epistasis is a pressing problem in protein evolution. Logically, homologs with similar folds should be linked in sequence space. But the significant differences in sequence identity between homologs raises the question of what sort of mutations allow the fold to stay functional, while forcing non-trivial rearrangements in packing. The chemical and biophysical properties of such ‘gateway’ mutations are issues of great importance, especially if one wishes to expand the sequence space that can be explored while engineering desired properties. Such a mutation has been discovered in *B. subtilis* ADK before, during selection for thermophilic properties.\textsuperscript{155,186} Here, we present two related scenarios – a mutation which can restore functional motions after catastrophic loss of dynamic constraints and a mutation which dramatically alters the optimized packing at a domain-domain interface.

Our statistical investigation of ADK sequence alignments shows that the T135R mutation logically precedes the loss of zinc-binding activity in the Lid domain. Although the natural loss of the zinc-binding motif may not be so dramatic as in our laboratory, C130H has been shown to preserve zinc binding at lower affinity, making it a possible intermediate state.\textsuperscript{130} Obviously, at some point, bacteria were pressed to stop relying on zinc to stabilize the Lid structure. G155E is conditioned on the appearance of T135R but as no effect on activity itself. G155E may represent the first step in the evolution of the *E. coli* ADK Lid domain, exploiting the new T135R interactions to begin re-optimizing packing. The situation of S131A-Y158T is less clear, because their beneficial properties
rely on A161K, but they show no stand-out correlations to it. Their correlation to each other seems to rely on exploring sequence space at the same time due to A161K, not a dependence on each other. The key difference may be that loss of zinc-binding can inactivate the enzyme entirely, impacting survival of the organism, whereas A161K is a gain of function that allows for exploration of new sequence space. Major challenges to organism survival, such as the depletion of environmental zinc, may leave a more tractable statistical pattern on a protein’s evolutionary history than random events that allow for dominance of the new variants.

Regardless, our methods are mathematically crude and plausible alternatives exist. Armenta-Medina and coworkers used a comparison of correlation information from sequences and coupled motions found in molecular dynamics simulations to predict an evolutionary network for ADK similar to our own initial work.\textsuperscript{187} Additionally, other groups have begun using conditional random field (CRF) approaches to model protein-protein interaction sites and predicting fold function.\textsuperscript{188-190} CRFs also use probabilistic landscapes to describe pathways and dependencies through sequence space, as we have done, but are more tractable for datasets where pairways assumptions limit the scope of information accessible and have been shown to be more robust against the introduction of bias.\textsuperscript{191}
Chapter 3: Improvements in Consensus Mutant Prediction

Contributions: Execution of the consensus mutant selection algorithm was assisted by Dr. Brandon Sullivan and Sidharth Mohan, of the Ohio State University. Design of gene sequences for consensus constructs was the combined effort of Sidharth Mohan, David Bowles, Dr. Eva Petrik, and myself. All characterization FruR constructs is the contribution of Dr. Eva Petrik, also of OSU. All characterization of TIM constructs is by Sidharth Mohan and Kimberly Stephanie. All characterization of ADK constructs and analysis of mutation patterns and phylogenies is done by myself.

Consensus design of proteins is a technique met with great success, but still demands improvement for reliable industrial use. Previously, the Magliery group reported a consensus variant of yeast triosephosphate isomerase created by combining conservation and co-variation statistics, improving on the results produced by conservation statistics alone. We have now produced four additional consensus enzymes, two TIMs and two ADKs, which show this method’s utility over a range of species. We have also produced consensus constructs of the small DNA binding domain FruR and tumor suppressor p53. Our consensus constructs are in good agreement with previous studies of stabilized and thermophilic homologs of these proteins, pointing to the general applicability of our methods. We also provide detailed sequence analysis of our failed constructs generated in
the course of this study and hypothesize on indicators that may be used to avoid problems in consensus design.

Protein chemistry is a topic of great industrial interest, and the demand for molecules tailored for function is now greater than ever, thanks to recent advancements in protein pharmaceuticals and biosynthesis. Current efforts seek to both redesign known molecules and identify novel molecules to meet this demand. Chief among present research concerns is finding reliable methods to structurally stabilize protein folds without sacrificing other useful properties. Advancements in this area would speed implementation of industrial-grade chemistry, improve shelf-life and cost of protein pharmaceuticals, and aid novel protein characterization by creating homologs more amenable to laboratory handling.

Consensus design is a methodology well suited to meet these demands. This approach uses the comparison of multiple aligned protein sequences to predict the “best fit” amino acid for a given position in a common fold. In the past, consensus design has famously produced stabilized antibody fragments, robust homologs of enzymes such as phytases, and model molecules for functional studies of tumor suppressor p53. Given the increased availability of sequence data over recent years thanks to metagenomic efforts and the creation of alignment databases such as Pfam, consensus design is an attractive option for non-expert users. However, among initial mutations made by consensus, 25 to 60 % are often neutral or destabilizing. Final constructs are usually arrived at by iterative screening of individual, additive single mutants, which can be costly and laborious. To date, only three major comparisons of
multiple consensus mutations in a single construct have taken place – the design of hyperthermophilic consensus phytases by Lehmann et al.; the design of consensus triosephosphate isomerases by the Magliery group; and the design of a consensus sucrose phosphorylase by Aerts et al. Each of these studies showed a strong dependence on the similarity of the starting wild-type sequence to the aligned homologs, suggesting that phyla-specific co-mutations differentiate otherwise similar folds and therefore not every consensus mutation was a true “best fit”.

Co-varying residues have been shown to be energetically coupled, and statistical co-variation in multiple sequence alignments can be used in the search for protein-protein interactions and intramolecular allosteric networks. In the survey of consensus TIMs made by our group, mentioned above, we observed that the most successful combinations of consensus mutations were those which avoided residues implicated in networks of co-varying amino acids. Similar work by Dietrich et al. found that alanine scans of co-varying residues could destabilize molecules or disrupt catalytic activity. Additional work by the Magliery group showed that a consensus TIM construct produced by picking all conserved mutations and then excluding those high degrees of co-variation showed a higher increase in melting temperature than a similar construct made by combining all consensus mutations proven to be individually stabilizing. That is, we showed that statistically screening possible consensus mutants could be more effective than experimentally screening individual mutants and assuming additivity. This present work expands on this methodology and tests its applicability to other well-studied folds.
We selected a wide range of proteins to make consensus constructs in: triosephosphate isomerase from *E. coli* and human, adenylate kinase from *E. coli* and *B. subtilis*, and the DNA binding domain of *E. coli* metabolic-repressor FruR. Additionally, multiple consensus constructs for target TIMs and ADKs were made using alignments of varying character, allowing us to examine what factors in MSAs might influence consensus design. We were able to generate stabilized consensus constructs of each of our target sequences, some with a melting temperature increase of nearly 20 degrees, while preserving the activity of enzymes. Happily, we found that mutations made in successful constructs paralleled observations from studies of thermophilic homologs, lending confidence to our hope in the general applicability of this work.

**Materials & Methods**

*Alignment and consensus screening*

Sequences for alignment were obtained through the Uniprot protein database, either independently or through the Pfam meta-database service. All alignments were made using HMMer. Hand-curated triosephosphate isomerase alignments were prepared as described previously. Automated curation was accomplished by binning sequences by order of blank and occupied positions and discarding distinct minorities. The purpose of curation is to exclude fragments and poorly-aligned sequences that may produce anomalous results.
The degree of conservation for each residue was calculated as relative entropy (RE), a bit score describing perturbation from a ‘neutral drift’ probability distribution. (Eq. 1) The codon usage probability in *S. cerevisiae* was treated as the neutral drift state, in accordance with the previous work of ourselves and others. The magnitude of RE scores depends on the size and alphabet of the sample set, so it is only useful for comparing positions in the same alignment. With this in mind, we use the average RE score as the amount of conservation expected from purely statistical effects. Residues with RE scores higher than average were considered conserved by evolutionary pressure and marked as potential consensus mutation sites.

\[
\text{(Eq. 3.1) } RE(x) = \sum_{AA} p(x) \ln \frac{p(x)}{q(x)}
\]

If a residue shows higher-than-average conservation and yet is not the consensus AA in the native protein, the possibility exists that interactions another AA side chain prohibit the consensus identity. As a check against this, we calculated the mutual information (MI) for each pair of alignment positions. MI describes interaction between two sets of data by comparing the real informatics entropy of the combined set to the predicted entropy if the sets were truly independent. Although structural contacts are not the only possible cause of high set dependencies, a positive correlation between the two has been shown. Based on previous empirical work, residues with a mutual information score above the 99th percentile were excluded from the sites of potential consensus mutations.

\[
\text{(Eq 3.2) } MI(x, y) = H(x) + H(y) - H(x, y)
\]
(Eq 3.3) \[ MI(x, y) = \sum_x \sum_y p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \]

These two scores define a window for each residue to be considered for consensus mutation – RE must be above the average, MI must be below the 99th percentile. For an example, Fig. 3.1 shows the distribution of scores for the scTIM-uni-2 construct. Statistical results for all alignments can be found in Appendix A.

Constructs and cloning.

Genes for consensus constructs were synthesized by Genewiz and shipped in pUC57-kan plasmids. Genes were cloned into DH10B cells on pHLIC plasmids and transformed in BL21-DE3 cells for expression. Expression and purification protocols followed previous work for wild-type proteins. Consensus constructs aligned with their wild-type counterparts can be found in Appendix B.

Kinetic Measurements

ADK and TIM kinetics were measured using reactions coupled to the oxidation of NADH at 25°C. Reactions coupling NADH oxidation to ADP or dihydroxyacetone phosphate (DHAP) production have been described previously.\textsuperscript{54,160} To summarize – ADK phosphorylates AMP in excess ATP, producing two ADP, which are used by pyruvate kinase to de-phosphorylate phosho-enol-pyruvate and produce pyruvate; this is used by lactate dehydrogenase to produce lactate, using NADH as the reducing agent.
Fig. 3.1 Relative entropy and Mutual Information scoring used in the design of the “Algo” TIM construct.
TIM catalyzes isomerization of D-glyceraldehyde 3-phosphate (GAP) to DHAP, which then converted to glycerol-3-phosphate by glycerol-3-phosphate dehydrogenase, using NADH as the reducing agent.

**CD Scans and unfolding**

CD measurements were taken on a Jasco J-318 Circular Dichroism spectrometer (Ohio State Department of Chemistry & Biochemistry). 20 μM samples were prepared in 100 mM phosphate buffer (pH 7.1), and 200 to 260 nm measurements were taken at 25 °C with a path length of 1 mm. Thermal denaturation was tracked at 222 nm at a rate of 1 °C/min over a range of 25 °C to 95 °C. Data was fit as described by Matouschek and Fersht.

**NMR spectra**

Protein samples for NMR were N¹⁵ labeled by expression in minimal media. H¹⁻¹⁵N heteronuclear single quantum coherence (HSQC) spectra were collected on a Bruker DRX 600 MHz spectrometer (OSU Campus Chemical Instrument Center) and analyzed using the NMRpipe software package.
## Table 3.1

<table>
<thead>
<tr>
<th>Construct</th>
<th># Mutations/Chain Length</th>
<th>% positions mutated</th>
<th>Phylogeny</th>
<th>Families</th>
<th># of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Algo&quot; TIM</td>
<td>15/248</td>
<td>6.00%</td>
<td>All</td>
<td>TIM</td>
<td>781</td>
</tr>
<tr>
<td>scTIM-uni-2</td>
<td>33/248</td>
<td>13.30%</td>
<td>All</td>
<td>TIM</td>
<td>900</td>
</tr>
<tr>
<td>ecTIM-uni</td>
<td>19/255</td>
<td>7.50%</td>
<td>All</td>
<td>TIM</td>
<td>781</td>
</tr>
<tr>
<td>ecTIM-bac</td>
<td>17/255</td>
<td>6.70%</td>
<td>Bacteria</td>
<td>TIM</td>
<td>516</td>
</tr>
<tr>
<td>huTIM-uni</td>
<td>13/249</td>
<td>5.20%</td>
<td>All</td>
<td>TIM</td>
<td>781</td>
</tr>
<tr>
<td>huTIM-euk</td>
<td>5/249</td>
<td>2.00%</td>
<td>Eukaryotes</td>
<td>TIM</td>
<td>224</td>
</tr>
<tr>
<td>bsADK-bac</td>
<td>21/217</td>
<td>9.70%</td>
<td>All</td>
<td>long-chain ADK</td>
<td>574</td>
</tr>
<tr>
<td>eeADK-lc</td>
<td>12/214</td>
<td>5.60%</td>
<td>All</td>
<td>long-chain ADK</td>
<td>574</td>
</tr>
<tr>
<td>ecADK-gn</td>
<td>6/214</td>
<td>2.80%</td>
<td>Gram-negative</td>
<td>long-chain ADK</td>
<td>335</td>
</tr>
<tr>
<td>FruR-bac</td>
<td>9/49</td>
<td>18.40%</td>
<td>Bacteria</td>
<td>LacI, FruR, PurR</td>
<td>8330</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Construct</th>
<th>$T_m$(°C)</th>
<th>$\Delta T_m$</th>
<th>% Max Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Algo&quot; TIM</td>
<td>68.4</td>
<td>+9.2</td>
<td>100%</td>
</tr>
<tr>
<td>scTIM-uni-2</td>
<td>66.4</td>
<td>+7.2</td>
<td>150%</td>
</tr>
<tr>
<td>ecTIM-uni</td>
<td>66.9</td>
<td>+6.2</td>
<td>92%</td>
</tr>
<tr>
<td>ecTIM-bac</td>
<td>71.0</td>
<td>+10.3</td>
<td>1%</td>
</tr>
<tr>
<td>huTIM-uni</td>
<td></td>
<td></td>
<td>Does Not Express</td>
</tr>
<tr>
<td>huTIM-euk</td>
<td>68.4</td>
<td>+18.2</td>
<td>52%</td>
</tr>
<tr>
<td>bsADK-bac</td>
<td>69.0</td>
<td>+17.8</td>
<td>160%</td>
</tr>
<tr>
<td>eeADK-lc</td>
<td>61.8</td>
<td>+1.8</td>
<td>0%</td>
</tr>
<tr>
<td>ecADK-gn</td>
<td>64.1</td>
<td>+4.1</td>
<td>69%</td>
</tr>
<tr>
<td>FruR-bac</td>
<td>68.7</td>
<td>+18.7</td>
<td>Not characterized</td>
</tr>
</tbody>
</table>
Results

Previous work by members of the Magliery lab engineered triosephosphate isomerases (TIM) for stability using consensus mutations in *S. cerevisiae* TIM and novel consensus sequences from alignments of bacterial, eukaryotic, and archaeal TIMs. Among the consensus constructs of scTIM originally reported, we generated one (dubbed “Algo TIM” at the time) using the screening algorithm discussed here. We repeated the design of a stabilized scTIM to account for two factors – our move from manual to partially automated curation of the sequence alignment and the addition of new TIM sequences to the Uniprot database. ("scTIM-uni-2") Both scTIM constructs display similar stabilities and activities. This gives us confidence in the general utility of our methods and that expert intervention in alignment generation can be minimized.

Melting temperatures and kinetic activities of our constructs are summarized on the above table. Melt curves can be found in Appendix C, and kinetic measurements can be found in Appendix D. We targeted TIMs from two new species for stabilization – *E. coli* and human. For both, we made two constructs, with consensus mutations screened from either a universal or a domain-specific alignment. ecTIM-uni, ecTIM-bac, and huTIM-euk demonstrate a 5-20 °C increase in melting temperature. huTIM-euk also demonstrates more consistent two-state unfolding behavior than its wild-type counterpart. However, huTIM-uni could not be expressed and characterized.

Because we cannot express huTIM-uni, we cannot be certain what prevents proper folding. However, alignment of the screening-generated huTIM-uni sequence to
wild-type shows two deletions, L29 and V196. These deletions appear in the consensus sequence because L29 and V196 are poorly occupied in the universal TIM alignment.

The significant stabilization of huTIM-euk by a mere five mutants appears to be the result of optimizing helicity and hydrophobic packing, similar to results obtained by previous studies by the Goraj group. GAP kinetics of huTIM-euk show it to be active, but with much more pronounced substrate inhibition. Whether this is due to the stabilization of an inactive conformation or a disruption of the substrate binding site is not easy to decide.

Fig. 3.2 GAP-dependent rate measurements of ecTIM-bac activity.
The ecTIM-uni and ecTIM-bac constructs provide a more useful comparison. These two constructs show close amounts of stabilization, but differ greatly by kinetics. Both have a cluster of mutations which repack the hydrophobic core between β-strands 4 and 5 and α-helix 6. By specific activity at 4 mM GAP, ecTIM-bac is two orders of magnitude less active than wild-type, whereas ecTIM-uni has wild-type like activity. Interestingly, ecTIM-bac shows a sharp increase in activity at 2 mM GAP, consistent with allosteric activation. (Fig. 3.2) It may be reasonable to suppose that dimerization, which is required for enzymatic activation, has been perturbed in ecTIM-bac. Both constructs showed a progressive loss of activity over time, with ecTIM-bac activity dropping much more quickly. Both constructs have the mutant T78I on loop 3, a key part of the dimer interface, as well as the adjacent I47L mutant. (Fig. 3.3a) But only ecTIM-bac has the nearby mutant L12M, which packs in the center of loop 3, while ecTIM-uni has V66C neighboring T78I. In both alignments, methionine is the consensus for position 12, but in the universal alignment, its high correlation to other positions gets it excluded by our screening process. (See Appendix E.)

ADKs present an interesting target for stabilization because they consist of three independently folding domains, whose coordinated dynamics are required for activity. Moreover, numerous previous studies have been done on ADK stabilization, providing a useful benchmark for our results. An alignment of ADKs was produced from Uniprot using the program HMMer. Such an alignment is available through Pfam, but the Pfam entry does not differentiate between short and long chain isomers. Our own HMMer-generated alignment only used ADKs with the “long chain” loop insert, a feature which
**Fig. 3.3** A.) Interface of the ecTIM dimer, showing residues L12, I47, V66, and T78. B.) Complete ecADK structure displaying ecADK-ic mutations (blue), ecADK-gn mutations (magenta), and shared (yellow). PDB: 4IOT, 4AKE
dominates in bacterial homologs but is not well-represented in archaea and eukaryotes. Our targets for stabilization were ADKs from B. subtilis and E.coli. bsADK-lc gained 18 °C to its melting temperature, as well as retaining activity comparable to wild-type. ecADK-lc, by contrast, showed little stabilization, while losing kinetic activity. We addressed this by manually curating our long-chain ADK alignment for sequences from only gram-negative bacteria and screening this smaller alignment for consensus mutations. Subsequently, ecADK-gn retained wild-type activity, while its melting temperature increased by 4.1 °C.

Examination of the ecADK-lc mutants on the ecADK crystal structure suggests activity loss due to mutations which interfered with ATP binding, especially with a mutation made in the NTP-binding Walker motif (A8P) and the mobile regions of the Lid (E170V, Q173K, G180D) and AMPbd domain (I72L).1 (Fig. 3.3b) However, we cannot be certain, because identical mutations appear in bsADK-bac without any loss of activity. What is striking about ecADK-lc and ecADK-gn is that many of the same positions passed the screen for mutation, but were already the consensus amino acid. Meanwhile, while E170V was excluded by high correlation, Q173K and G180D were excluded by low overall conservation. (See Appendix E.) This means that by going to the gram-negative specific alignment, we excluded problematic positions not by better detection of correlations, but because their conservation went below the cut-off to be considered a viable consensus position. Equally interesting is the distribution of mutations. Work by both the Shamoo152 and Phillips’s167 labs comparing thermophiles and mesophiles implies that only positions in the Core domain affect the over-all temperature of ADK’s thermal
unfolding, giving ADK a pseudo-two-state unfolding curve despite its independently folding domains. 82% of all screened mutations were in the Core domain, with 18% in the NMPbd domain and none in the Lid domain. NMPbd’s tolerance of neutral mutation is reasonable, considering that most substrate contacts are made with the backbone. Among the 16 core mutations in bsADK-bac, 5 correspond to known stabilizing mutations reported previously in the literature. 4 of the 5 Core mutations in ecADK-gn also appear in ecADK-bac, suggesting that stabilizing mutations made in ecADK-bac were balanced by equally deleterious mutations.

Our next target sequence was the short FruR DNA binding domain. This afforded us the chance to study the stabilization of small peptides without well-defined hydrophobic cores, backed up by well-annotated NMR literature. Due to the low number of unique FruR sequences (<200), we choose to screen mutations from the Pfam alignment of the LacI superfamily, of which FruR is a member. FruR-bac differs from its wild-type sequence by nine mutations. This constitutes up to 18.4% of the sequence. $^1$H-$^{15}$N HSQC NMR confirms that the FruR-bac construct is well-folded, although the change in side-chain content means it is not directly comparable to the wild-type. (Fig. 3.4a) Five of the mutated positions have closely packed side-chains and arguably optimize contacts between themselves. The other four are located on the third helix; two optimize helix electro-statics and one substitutes an alanine for a helix-disrupting valine. Overall, FruR-bac has a melting temperature 13.7 °C greater than FruR wild-type. Unlike wild-type FruR, FruR-bac remains soluble after denaturation and folds reversibly upon cooling. However, after refolding, the stabilization is lost. It can be unfolded and refolded
Fig. 3.4 A.) $^1$H-$^{15}$N HSQC NMR of FruR-Bac. B.) Melting and refolding curves of FruR and FruR-bac.
Discussion

Because we've chosen to make our consensus mutations in aggregate, instead of individual, there is little we can infer about the biophysics of single mutations. Even though there is a net gain in the melting temperature for each construct, it is not unlikely for a single mutation to have a neutral or deleterious effect that is masked by the overall collection. However, because we have five globular enzymes which enjoy the same transformation – an increase in melting temperature with a retention of activity – we find it reasonable to think overall trends in our screened mutations are informative.

Summation of the mutations shows a sizeable preference for branched hydrophobic amino acids – valine, leucine, and isoleucine – to alternate between themselves. We should emphasize, this does not correspond any enrichment or depletion in branched hydrophobic amino acids, although there does seem to be a preference for β-branched isoleucine and valine. (Fig. 3.5) Overall, this suggests simultaneous optimization of core side-chain packing and core main-chain secondary structure. A curious trend among our successful variants is that although they show varying ΔT_m, their final melting temperatures all approach 70 ºC. Since our successful constructs all seem to rely on stabilized core packing, our results may be considered in light of ADK structural comparison by the Philips lab,⁶ who found that electrostatic stabilization did not...
Fig 3.5 Total mutations seen across active consensus constructs.
differentiate psychrophiles and mesophiles but did mark the difference between mesophiles and thermophiles, in accordance with the temperature range where the dielectric constant makes electrostatic contributions significant. If our consensus methods favor core packing mutations, then the limit may represent the threshold at which natural evolution relies on electrostatic interactions rather than hydrophobic packing.

We are fortunate to have two cases in which we have two consensus variants of the same target sequence, using different alignments, and both constructs show net stabilization while only one shows compromised enzymatic activity. These are ecADK-lc/ecADK-gn and ecTIM-bac/ecTIM-cc. These provide an opportunity to compare the sequence content of their source alignments, with an eye towards possible indicators that the consensus sequence does not preserve kinetics.

We find that the full bacterial ADK alignment predicts beneficial mutations for bsADK and detrimental ones for ecADK. Interestingly, if we compare the pure, unscreened consensus of the alignment to both wild-type sequences by alignment, we find it has 61% identity with bsADK and 72% identity with ecADK. However, by generating a phylogenetic tree reflecting the alignment, we see that ecADK lays on a subtree which diverges very early, whereas bsADK lies on the main tree and shares internal nodes with a wide majority of aligned sequences. When we look at the alignment of gram-negative ADKs, which predicts beneficial mutations for ecADK, we find that the phylogeny is a single tree, with ecADK at the root. Going by this example, we may posit that the tree predicted by an alignment is the good indicator of its utility for informing mutations in a target sequence – one should prefer a tree that positions the branch of the
target sequence along as many common nodes as possible within the main tree. The availability of simple tree creation by ClustalW\textsuperscript{202} and viewing software such as EvolView\textsuperscript{203} makes this an easy criteria for non-experts to work with.

The ecTIM-bac sequence uses consensus derived from a bacteria-only alignment of TIMs and while it's overall melting temperature has a greater increase, it demonstrates badly lowered kinetics due to destabilization of the dimer interface. The ecTIM-uni sequence, which uses consensus from a universal TIM alignment, shows weaker dimer binding than wild-type but retains comparable activity for a longer period of time. The main difference between the alignments is that the universal one contains a minority percentage of eukaryotic and mesozoic sequences. Strong evidence does exist that eukaryotic TIMs are the result of a horizontal gene transfer from gram negative bacteria.\textsuperscript{204}

Counter-intuitively, both alignments are very similar. Comparison between wild-type ecTIM and the pure alignment consensus of both alignments shows a similarity of 60\% for the universal consensus and 61\% for bacterial consensus. Comparing our post-selection consensus constructs to the consensus of their source alignments shows 68\% for both. Using the alignments to inform a phylogenetic tree for TIM sequences places ecTIM in a similar subtree for both – a group of 62-65 sequences branching off at a very early internal node. We can see by looking at the normalized MI scores that the N and C-terminals of ecTIM co-vary more relative to other positions in the universal alignment. This causes the exclusion of several positions at the dimer interface in ecTIM-uni, which passed the statistical test for ecTIM-bac. Thus, we see that our algorithm is still sensitive
Fig. 3.6 Phylogenetic trees from the TIM universal (A), TIM bacterial (B), ADK long chain (C), and ADK gram negative (D) alignments.
to differences that may not be readily apparent by cursory indicators and demand careful alignment building.

Conclusion

We are pleased to report that our consensus screening algorithm works consistently for two enzymes, over a range of species, provided that starting alignments meet reasonable criteria which can easily be assessed using widely-available software. MSAs must reflect the common nature of sequences, a facet captured poorly by similarity scores but well by phylogenetic common ancestors. Screening conserved mutations by co-variation can protect enzymatic activity, but requires that the alignment present as close to a contiguous evolutionary history as possible, to correctly capture statistical effects. This makes enzymes which have undergone horizontal gene transfer difficult targets. However, this also means such enzymes may make for good benchmarks for the robustness of statistical analysis.

Among the successful constructs we report, the trend is towards rearrangement of the hydrophobic core by alternating between branched hydrophobic amino acids. A variation of this theme, though, is the small peptide FruR, an irreversible folder whose consensus construct enjoys increased solubility but which still unfolds irreversibly. Comparison of our huTIM and bsADK constructs to past studies of stabilizing and thermophilic mutations shows good agreement with their conclusions. This gives us faith in the generality of our approach.
Chapter 4: Biophysical studies of ADK transition variants

As discussed in chapter 2, we designed a series of *B. subtilis* ADK variants that probed how enzymatic activity could be retained in the event of mutations that ablated zinc binding by the Lid domain, by using aligned sequence homologs. What we discovered is that loss of structural zinc dependence is preconditioned on the mutation T135R. A variant possessing this mutation will retain activity after the zinc chelating site has been completely mutated away. When the zinc binding site is removed, the temperature of thermal unfolding drops severely, but the T135R mutation does not restore stability, only activity. Interestingly, a strong correlation in the alignment can be found between the loss of zinc chelating residues, the T135R mutation, and the mutation G155E. This second mutation has no effect on an inactive variant, but does alter kinetic activity of a variant rescued by T135R, to show a greater degree of substrate inhibition.

Neither T135R nor G155E alters the melting temperature of bsADK. However, as mentioned in our introduction, chimeras of *B. subtilis* and *B. stearothermophilus* demonstrate the principal determinate of ADK’s melting temperature is the Core domain, assuming zinc is bound to the Lid domain in both homologs. This suggests that the structural effects of mutations in the Lid domain not related to zinc chelation may be invisible during thermal denaturation. To further probe the effects of these mutations, then, we analyzed our transitional variants using a gradient combination of thermal and chemical denaturation. Normally, such an analysis would allow us to fit the unfolding behavior to thermodynamic parameters. But this fitting requires the assumption that...
folding occurs as a two-state equilibrium. We cannot make this assumption due to the complicated nature of ADK’s folding pathway. Regardless, these results are enlightening as to the possible interaction between T135R and G155E that leads to their evolutionary correlation.

We have also tried to generate crystals of key variants for X-ray defraction. The intention is to specifically model rearrangement of the Lid domain upon the inclusion of the T135R mutant. Although this work is not complete, it is recorded here as it stands at the time of writing.

**Materials and Methods**

*Protein Purification*

Electrocompetent BL21(DE3) E. coli cells were transformed with the pHLIC plasmids bearing the adk genes, and grown in 2YT media at 37 °C with 1 mL of 100 mg/mL Ampicillin. Upon reaching an appropriate cell density, expression of 6×His-tagged protein was induced by addition of 0.1 mM Isopropyl β-D-1thiogalactopyranoside (IPTG). After three hours of induction, cells were harvested by centrifugation; resuspended in 50 mM Tris-HCl (pH 7.8), 300 mM NaCl, 10 mM imidazole, 6 mM MgCl2, 6 mM CaCl2, and 1.2% Triton-X; and lysed by sonication. Soluble material was separated by centrifugation at 15K rpm and 4 °C for 30 minutes. Lysis supernatant was mixed with 1 mL of Ni-NTA resin from Thermo Scientific and incubated overnight at 4 °C. The resin was separated from the supernatant by decanting into a Bio-rad 20 mL pre-
fretted chromatography column and washed with 50 mM Tris pH 7.8, 300 mM NaCl, and 50 mM imidazole; then protein was eluted with 50 mM Tris pH 7.8, 300 mM NaCl, and 200 mM imidazole. Eluted protein was incubated at 30 °C with 5 mM DTT and TEV protease overnight, to remove the 6×His tag and linker region. A second Ni-NTA purification was used to separate TEV-cleaved tag and any uncleaved protein from protein with the tag removed. Samples intended for CD characterization were dialyzed into 50 mM sodium phosphate buffer (pH 7.5) for short-term storage at 4 °C. Sample concentrations were determined by 1.) absorbance at 280 nm after the sample was confirmed to be pure on SDS-PAGE gel (stained with Coomassie Brilliant Blue dye) and 2.) optical density comparison of the sample on SDS-PAGE gel to serial dilutions of 2 mg/mL of bovine serum albumin.

Chemical and Thermal Denaturation

Samples were prepared for CD at a final protein concentration of 20 μM in 400 μL with 25 mM sodium phosphate buffer (pH 7.5). 200 μL of a sample was prepared with 40 μM protein in 50 mM buffer, then diluted with 200 μL of urea at variable concentration. Samples were equilibrated overnight at 4 °C. When transferred to 0.1 cm cuvettes and placed in the instrument (Jasco J-318 Circular Dichroism spectrometer, Ohio State Department of Chemistry & Biochemistry), samples were allowed to equilibrate 10 °C for ten minutes. Samples were heated from 10 to 55 °C at a ramp rate of 1 degree every two minutes. At five degree intervals, samples were allowed to equilibrate for five
minutes and then the spectrum from 200 to 260 nm was measured. The minimum at 222 nm was used to measure the fraction of the sample folded. Melting curves were fit to a two state model as described by Matouschek and Fersht.\textsuperscript{199}

\textit{Crystallization Screening}

The ADK variants bsDi, bsHexa, and bsTetra with the ligand AP\textsubscript{5}P (see Chapter 2) were screened using the NeXtal Tubes Classics Suite from Qiagen. Further screening of bsDI and bsADK-A161K was done by incrementally altering the crystallization conditions for \textit{B. subtilis ADK}, described by Bae et al.\textsuperscript{6}

\textbf{Results and discussion}

\textit{Chemical denaturation}

The melting temperatures of bsADK, bsDi, bsTetra, bsHexa, bsPenta-135, and bsPenta-155 were tested at varying concentrations of urea. In all cases, we observed a rise in the CD signal for the baseline of folded material. Fig. 4.1 shows the unfolding curves for bsADK treated with urea. At 1.6 M urea, bsADK still denatures under heat as though it is fully folded material, but displays less ellipticity than material which has not been treated with denaturant. This indicates separate unfolding pathways for observable thermal and chemical denaturation.
We can compare the chemically-induced unfolding of these variants at 25 °C (Fig. 4.2) and see that variants lacking a zinc binding site unfold at a lower concentration of urea than bsADK, similar to the trend seen with thermal unfolding. Interestingly, though, the mutation G155E causes a very pronounced drop in chemical stability, which is repaired by the addition of T135R (bsHexa). Another notable feature is

![Graph showing thermal unfolding curves of ADK variants in urea](image)

**Fig. 4.1** Thermal unfolding curves of ADK variants in urea
that when T135R and G155E are combined in a zinc binding variant (bsDi), they still confer a resistance to chemical denaturation greater than bsADK. Alternatively, we can compare the melting temperatures of our chemically perturbed samples. We see another interesting phenomenon if compare the trend in melting temperatures between bsHexa, bsPenta-135, and bsTetra – for this instance, the melting temperatures of bsPenta-155 were so perturbed that melting curves could not be reliably fit. (Fig. 4.3) Here we see that although overall melting points for bsPenta-135 are lower, they remain parallel with the trend of melting points for bsTetra. For bsHexa, however, the trend remains linear, but is less shallow. A similar difference is seen in the chemically-perturbed melting points of bsADK and bsDi. Considering the accumulated data, this suggests that T135R and G155E individually make bsADK more susceptible to chemical denaturation, but interact in a fashion that affects the pathway of chemical unfolding but not the pathway of thermal unfolding.
Fig. 4.2 Unfolding of ADK variants at 25 °C
Fig. 4.3 Effects of urea on ADK melting temperature
Crystallography

We found two conditions which allowed ADK variants to be crystallized for X-ray defraction. Crystals of 10 mg/mL bsHexa can be precipitated in 100 mM Tris (pH 8); 200 mM MgCl$_2$; 2.5 M 1,6-hexanediol; and 3.5 mM AP$_5$A. (Fig. 4.4a) Crystals of 10 mg/mL bsADK-A161K can be precipitated in 50 mM HEPES (pH 8), 50 mM CaCl$_2$, 35% poly-ethylene glycol 1500, and 3.5 mM AP$_5$A. (Fig. 4.4b) Crystallization conditions have been found for 18 mg/ml bsDi, in 50 mM HEPES (pH 7), 50 mM CaCl$_2$, 35% poly-ethylene glycol 1500, and 3.5 mM AP$_5$A. However, these crystals in branched needles, making them unideal for X-ray defraction. (Fig. 4.4c)
Fig. 4.4 Crystallization of ADK variants
Chapter 5: Purification of Dystrophin gene products

Contributions: The full length dp71 constructs were prepared by Dr. Adeline Vulin-Chaffiol, from the lab of Dr. Kevin Flanigan at Nationwide Children’s Hospital. All other work is done by myself.

Duchenne muscular dystrophy (DMD) is a developmental disorder, mapped to the X-chromosome, characterized by degeneration of skeletal muscle and cognitive impairment. This disorder has been linked to mutations in the gene product dystrophin, a 427-kDa protein translated from 79 exons. Dystrophin is comprised of four regions: the N-terminal fragment with an actin binding motif, a linker domain comprised of 24 spectrin repeats, a cysteine-rich fragment which binds to the sarcolemmal glycoprotein complex (CR), and a c-terminal region of unknown function (CT). The combined CR-CT fragment is of additional interest because it is expressed as the gene product dp71. Whereas full-length dystrophin is only expressed in skeletal muscle, dp71 is one of the several alternative-splicing products expressed in non-muscular tissue and is the sole gene product found in neuronal tissue. This implies that dp71 may be key in understanding cognitive symptoms found in DMD patients.

The CR domain can be further divided into four distinct domains: a WW domain, a ZZ-type zinc finger domain, and two EF-hand domains. Previously, the individual WW
domain and the EF-hand domains have both been demonstrated to bind with beta-dystroglycan, a component of the sarcolemmal glycoprotein complex. The ZZ fragment, however, enhances the binding of the WW domain, but does not bind itself. The role of the ZZ fragment is significant because two disease-linked mutations, Cys3313Phe and Cys3340Tyr, potentially disrupt the zinc binding site of the ZZ fragment.

It is, therefore, important to biophysically characterize the ZZ fragment, its interaction with the WW domain, and the exact effect of these disease-linked mutations. Here, we attempted to express and purify dp71 fusion proteins that have been used in previous genetic studies and the ZZ fragment from dp71. Unfortunately, despite their use in binding studies and antibody growth, we have found both of these constructs to be poorly folded and that the dp71 fusion proteins show evidence of being molten globules.

Materials and Methods

Cloning

A gene was produced encoding the cysteine-rich (CR) and c-terminal (CT) regions of human dystrophin, covering residues 3051 to 3678. A nickel-binding six-histidine region (His-Tag) and then either maltose binding protein (MBP) or glutathione S-transferase (GST) were added to the n-terminal with a TEV protease site. The final fusion protein is His-MBP/GST-TEV-CR-CT. This total gene was prepared and cloned into the pHLIC expression vector by Dr. Adeline Vulin-Chaffiol.
The Dp71 ZF fragment gene was produced by overlap PCR of fragments produced by GeneWiz and then amplified. Dp71 ZF fragment was then cloned into the pHLIC expression vector with an n-terminal His tag.

**Expression and Purification**

Constructs in the pHLIC expression vector were expressed in the BL21-DE3 *E. coli* strain by induction with 100 μM IPTG, after cultures had been allowed to reach linear growth phase. Expression took place overnight, at 16°C, in two liter volumes of 2YT media. Cells were collected by pelleting at a force of 4424 g for 10 minutes.

For CR-CT fusion constructs, cells were resuspended in 50 mM Tris (pH 7.5), 100 mM NaCl, and 10% glycerol. Cells were lysed by pressure, in the presence of Roche cOmplete, EDTA-free protease inhibitor tablets. Lysate was treated with DNase 1, RNase A, lysozyme, 10% Triton-X 100, and 10 mM PMSF. Insoluble material was cleared from the lysate by pelleting at 18,500 g for thirty minutes. MBP fusion proteins were collected using amylose resin columns and eluted with amylose. GST fusion proteins were collected using Ni-NTA columns and eluted using imidazole. Fusion proteins were treated with TEV protease in 5 mM DTT at 10 °C overnight. The CR-CT construct was isolated using anion exchange for characterization.

For the dp71 ZF fragment, cells were resuspended in 20 mM Tris (pH 7.2) and lysed by sonication, then pelleted at 18,500 g for thirty minutes. Soluble fraction was disposed of and pellet was resuspended in 20 mM Tris (pH 7.2), 100 mM NaCl, 10 mM
ZnCl, 6M Urea, and 2mM β-ME. We incubated this suspension at 37°C for three hours, then passed it over a fretted column to remove remaining solid particles. ZF fragment was isolated by binding to a Ni-NTA column and elution with 250 mM imidazole.

*Circular Dichroism*

CD measurements were taken on a Jasco J-318 Circular Dichroism spectrometer (Ohio State Department of Chemistry & Biochemistry). 20 μM samples were prepared in 100 mM phosphate buffer (pH 7.1) with 300 mM NaCl, and full-spectrum measurements were taken at 25 °C with a path length of 1 mm. Thermal denaturation was tracked at 222 nm at a rate of 1 °C/min over a range of 15 °C to 95 °C.

*ANS dye binding*

4 μM of 1,8 Anilinonapthalene-8-Sulfonic Acid (ANS) was added to 2 μM of protein under room temperature conditions. Samples were excited at 372 nm and fluorescence was collected over a range of 400 to 600 nm. Fluorescence was collected on an Olis DM-45P Fluorimeter with a 3.16 mm slit width and a 13 mm band pass.
Fig. 5.1 Gels showing the purification of MBP (A) and GST (B) fusion constructs.

A.) TEV cleavage of MBP fusion construct (1), flow through and wash of 8 mL anion exchange column (2, 3), and elution of column with sodium chloride (4 to 9). Ladder is 10-225 kDa USB protein marker.

B.) Cell lysis from GST expression (1), flow through of Ni-NTA column (2), wash (3) and elution from Ni-NTA column (4), and TEV cleavage of GST fusion protein (5). Ladder is Full Range Rainbow.
Results and discussion

Lack of tertiary structure by CD

The His-MBP-Tev-CR-CT fusion protein was expressed and TEV cleaved (Fig. 5.1a), and the CR-C fragment was isolated, as described in our methods. The CR-C fragment’s secondary and tertiary structure features were assessed by progressive scans for circular dichroism at rising temperature. CD shows a sharp peak at 210 nm, indicative of high β-sheet content. Following this peak over rising temperatures shows a linear slope (Fig. 5.2). This suggests a lack of consistent tertiary structure, from the lack of cooperative unfolding behavior.
**ANS binding**

Given the lack of cooperative unfolding behavior seen by CD, we pursued the possibility that the CR-CT fragment had been purified as a molten globule. We used 1,8-ANS to test the CR-CT fragment for exposed hydrophobic surface area. ANS fluorescence is quenched by water, but fluoresces strongly in hydrophobic environments, similar to the core of a collapsed protein. ANS fluorescence is a good test for exposed, accessible core regions and thus unstructured protein, i.e. a molten globule.

Strong fluorescence was detected for both His-MBP-Tev-CR-CT and for isolated CR-CT, with no fluorescence detected for isolated His-MBP-Tev or empty buffer. We changed to the His-GST-Tev-CR-CT construct, on the basis of its frequent use in literature as a model. However, we found similar results. We could not isolate His-GST-Tev and CR-CT by anion exchange without significant product loss, but we could individually express and purify His-GST for comparison, using an identical protocol (Fig. 5.1b). The His-GST-Tev/CR-CT mixture showed strong ANS fluorescence, whereas His-GST alone did not. Our overall conclusion from these studies is that, as purified, the CR-CT fragment is not a folded protein.

**Dp71 ZF fragment**

We decided that the CR-CT fragment would not make a functional model for the biophysical characterization we sought to accomplish. Instead, we sought to characterize the smallest possible fragment still containing the implicated cysteine mutations. This is the zinc-finger fragment comprised of residues 3307 to 3348.
Fig. 5.3 ANS binding of dp71 fusion proteins
Fig. 5.4 Dialysis of dp71 to 6 M (1), 3 M (2), 2 M (3), and 1 M (4) urea. Ladder is Full Range Rainbow.
The dp71 ZF fragment was purified from inclusion bodies in good yield. Refolding was attempted by dialyzing out urea in two hour steps, in four steps – 6 M to 3 M, 3 M to 2 M, 2 M to 1 M, 1 M to no urea. In-between steps, precipitation was pelleted out of the sample and fresh TCEP was added as a reducing agent. After successful dialysis to urea-free buffer, the remaining sample was examined by CD. Although a broad, shallow peak was seen at 210-213, temperature scans found it invariant to rising temperature. This may be, then, indicative of unfolded, random coil.

As our results demonstrate, both the dp71 MBP and GST fusion proteins show characteristics of molten globules – proteins with folded secondary structure but poorly defined tertiary structure. This has two immediate implications. Firstly, since these constructs have been previously used to quantify binding, the CR domain must be able to fold in response to binding partners. Second, it raises the question of whether these constructs represent the true physiological state of these proteins, as are relevant to their disease states. Given the physiological importance of dystrophin and dp71, these are important questions that must be addressed to fully understand DMD and related diseases.
Fig. 5.5 Circular dichroism of dp71-Zf fragment
Appendix A. Selection scores for consensus constructs
Appendix B. Construct sequences with secondary structure

Alignments produced by ClustalW 2.1
Secondary structure taken from Uniprot database entries.

α-Helix β-Sheet

scTIM  
MARTFFVGNFKLNGSQAISEIIVERLNTASIPENVVCPPATYDYSVSLVKKQVIT 60  
scTIM-uni-2  
MARTFFVGNFKLNGTQSIKEIIVERLNTASIPENVENVCCPFYLDYSVSLVKKQVIT 60

scTIM  
VGACQNYLASKAQTGENSVQIKDDVWLVILGHSERRSFYHDFSDKFTAKTVFLQGG 120  
scTIM-uni-2  
VGACQNYLASKAQTGENSVQIKDDVWLVILGHSERRSFYHDFSDKFTAKTVFLQGG 120

scTIM  
VGVLIC1GETLEERKAGKTLDDLQVNLNAEVKQWDTVVAVAYEPWAOITGLAATFFED 180  
scTIM-uni-2  
LGVLIC1GETLEERAKGKTLDDLQVNLNAEVKQWDTVVAVAYEPWAOITQKATFFEQ 180

scTIM  
DGDIHASIRKFLASKLGDKAASELRILYGGSANGSNAVTFKDKADVDGFLVGGASLKEF 240  
scTIM-uni-2  
AQEVNASIRKFLASKGAAEVRILYGGSVKASNAVELKQPDIDGFLVGGASLKEF 240

huTIM  
MAPSRKFVGGNWKMNGRKQQLGELIGTLNAAKVPADTEVCAAPPAYIFDARQKLDPKI 60  
huTIM-euk  
MAPSRKFVGGNWKMNGRKQQLGELIGTLNAAKVPADTEVCAAPPAYIFDARQKLDPKI 60  
huTIM-uni  
MAPSRKFVGGNWKMNGRKQQLGELIGTNAAKVPADTEVCAAPPAYIFDARQKLDPKI 59

huTIM  
AVAAQNCYKVTNGAFTGEISPGMIKDCGATWVILGHSERRHVFGESDELQKVAHALAE 120  
huTIM-euk  
AVAAQNCYKVTNGAFTGEISPGMIKDCGATWVILGHSERRHVFGESDELQKVAHALAE 120  
huTIM-uni  
AVAAQNCYKVTNGAFTGEISPGMIKDCGATWVILGHSERRHVFGESDELQKVAHALAE 119

huTIM  
GLGVIACIGEKLDEREAGITEKVVFEQTKVIADNVKDWSKVVLAYEPWAOITQKATFFEQ 180  
huTIM-euk  
GLGVIACIGEKLDEREAGITEKVVFEQTKAIADNVKDSKVVLAYEPWAOITQKATFFEQ 180  
huTIM-uni  
GLGVIACIGEKLDEREAGITEKVVFEQTKVIADNVKDSKVVLAYEPWAOITQKATFFEQ 179

ecTIM  
MGHPLVMGNWKLNGSRBHVMHELSSNLRELLAGYAGCAYADVAFPEMYIMKREAEQGSHM 60  
ecTIM-bac  
MRHPLVMGNWKLNGSRBHVMHELSSNLRELLAGYAGCAYADVAFPEMYIMKREAEQGSHM 60  
ecTIM-uni  
MRHPLVMGNWKLNGSRBHVMHELSSNLRELLAGYAGCAYADVAFPEMYIMKREAEQGSHM 60

ecTIM  
LGQYNVLDSLNGAFGETSAAMLKDSIGAQYIIIGHSERRYTHKESDELIKAFVAKLKEQG 120  
ecTIM-bac  
LGQYNVLDSLNGAFGETSAAMLKDSIGAQYIIIGHSERRYTHKESDELIKAFVAKLKEQG 120  
ecTIM-uni  
LGQYNVLDSLNGAFGETSAAMLKDSIGAQYIIIGHSERRYTHKESDELIKAFVAKLKEQG 120

133
ecTIM  LTPVLCIGETEAEAGKTEEVCAQRQDAVLKGTQGAADFEGAVIAEPVWAIGTGKSATP  180
ecTIM-bac LTPILCVGETEAEAGKTEEVCAQRQDAVLKGTQGAADFEGAVIAEPVWAIGTGKSATP  180
ecTIM-uni LTPILCVGETEAEAGKTEEVCAQRQDAVLKGTQGAADFEGAVIAEPVWAIGTGKSATP  180

ecTIM  AQAQAVHKFIRDHIAKVDANIAEQVIIQYGGSVNASNAAELEFAQPIDDAVLGGASLKA  240
ecTIM-bac AQAQAVHKFIRDHIAKVDANIAEQVIIQYGGSVNASNAAELEFAQPIDDAVLGGASLKA  240
ecTIM-uni AQAQAVHKFIRDHIAKVDANIAEQVIIQYGGSVNASNAAELEFAQPIDDAVLGGASLKA  240

ecTIM  AFAVIVKAAEAAKQA  255
ecTIM-bac AFAVIVKAAEAAKQA  255
ecTIM-uni AFAVIVKAAEAAKQA  255

ecADK  MRLILLGAPGAGKTQAGIMEKYGIPQISTGDMLRAAVKSGTELQGKADIMDAGKLVT  60
ecADK-gn MRLILLGAPGAGKTQAGIMEKYGIPQISTGDMLRAAVKSGTELQGKADIMDAGKLVT  60
ecADK-bac MRLILLGAPGAGKTQAGIMEKYGIPQISTGDMLRAAVKSGTELQGKADIMDAGKLVT  60

ecADK  DELVIALVKERIAQEDCRNFLLDGFPRTIPQADAKMEGQNDVYIEFDVPLDELVRDI 120
ecADK-gn DELVIALVKERIAQEDCRNFLLDGFPRTIPQADAKMEGQNDVYIEFDVPLDELVRDI 120
ecADK-bac DELVIALVKERIAQEDCRNFLLDGFPRTIPQADAKMEGQNDVYIEFDVPLDELVRDI 120

ecADK  VGRRVHAPSRGVRHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRYLTVKMTAPLIG 180
ecADK-gn VGRRVHAPSRGVRHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRYLTVKMTAPLIG 180
ecADK-bac VGRRVHAPSRGVRHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRYLTVKMTAPLIG 180

ecADK  YYSKEAEAGNTKRYAKVDTKPAEVRADEKILG  214
ecADK-gn YYSKEAEAGNTKRYAKVDTKPAEVRADEKILG  214
ecADK-bac YYSKEAEAGNTKRYAKVDTKPAEVRADEKILG  214

bsADK  MNLVMLGAPGAGKTQGTERIVEDYGIPISTGDMLRAAVKSGTELQGKADIMADKELVP  60
bsADK-bac MNLVMLGAPGAGKTQGTERIVEDYGIPISTGDMLRAAVKSGTELQGKADIMADKELVP  60
bsADK-bac MNLVMLGAPGAGKTQGTERIVEDYGIPISTGDMLRAAVKSGTELQGKADIMADKELVP  60

bsADK  DEVTIGVKERLKEGCERFLLDGFPRTVQAASEALEELEYGKPIDYVINVKDVVL 120
bsADK-bac DEVTIGVKERLKEGCERFLLDGFPRTVQAASEALEELEYGKPIDYVINVKDVVL 120
bsADK-bac DEVTIGVKERLKEGCERFLLDGFPRTVQAASEALEELEYGKPIDYVINVKDVVL 120

bsADK  MERLTGRRISVCGTTTVLFNPKPFTPIGDKEVYQRRADDNEETVSKREVKMYQQTQ  180
bsADK-bac MERLTGRRISVCGTTTVLFNPKPFTPIGDKEVYQRRADDNEETVSKREVKMYQQTQ  180
bsADK-bac MERLTGRRISVCGTTTVLFNPKPFTPIGDKEVYQRRADDNEETVSKREVKMYQQTQ  180

bsADK  PLLDYSEKGYLANVQIQDIQVYADVKDLGGGKLK  217
bsADK-bac PLLDYSEKGYLANVQIQDIQVYADVKDLGGGKLK  217
bsADK-bac PLLDYSEKGYLANVQIQDIQVYADVKDLGGGKLK  217

FruR  KLDEIARLAGVRSRRSASYVINGKAKCQVKSDDKTVEVKVEVVMVREHNYHPN  49
FruR-LacI  TIDDIAARLAGVRSRRSASYVINGKAKCQVKSDDKTVEVKVEVVMVREHNYHPN  49
Appendix C. Thermal unfolding curves
Appendix D. Kinetic measurements

![Graphs showing kinetic measurements with various conditions and concentrations.](image-url)
Appendix E. Comparison of selection scores

Blue indicates positions where the consensus between alignments differ. Red indicates scores which fail to meet the cut-off.

### E. coli ADK

<table>
<thead>
<tr>
<th>Conservation</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Universal Gram-N</td>
<td>Universal Gram-N</td>
</tr>
<tr>
<td>A8</td>
<td>1.7786137</td>
</tr>
<tr>
<td>S43</td>
<td>1.9632841</td>
</tr>
<tr>
<td>L63</td>
<td>1.3740268</td>
</tr>
<tr>
<td>V64</td>
<td>1.4894215</td>
</tr>
<tr>
<td>I72</td>
<td>1.564856</td>
</tr>
<tr>
<td>V103</td>
<td>1.3371373</td>
</tr>
<tr>
<td>L107</td>
<td>1.4251333</td>
</tr>
<tr>
<td>L120</td>
<td>1.5828176</td>
</tr>
<tr>
<td>E170</td>
<td>1.206965</td>
</tr>
<tr>
<td>Q173</td>
<td>0.6691744</td>
</tr>
<tr>
<td>G180</td>
<td>0.9242171</td>
</tr>
<tr>
<td>V196</td>
<td>1.636131</td>
</tr>
<tr>
<td>L209</td>
<td>1.3493348</td>
</tr>
</tbody>
</table>

LC alignment | GN alignment
RE avg | 0.6383535 | 0.6383535 | RE avg | 1.14073 |
MI cut off | 0.4432283 | 0.60139 | MI cut off | 0.507384 |

### E. coli TIM

<table>
<thead>
<tr>
<th>Conservation</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Universal Bacterial</td>
<td>Universal Bacterial</td>
</tr>
<tr>
<td>V6</td>
<td>1.391178</td>
</tr>
<tr>
<td>L12</td>
<td>2.289369</td>
</tr>
<tr>
<td>A37</td>
<td>1.115698</td>
</tr>
<tr>
<td>I40</td>
<td>1.625213</td>
</tr>
<tr>
<td>A41</td>
<td>1.490216</td>
</tr>
<tr>
<td>I47</td>
<td>1.549381</td>
</tr>
<tr>
<td>A54</td>
<td>0.891512</td>
</tr>
<tr>
<td>V66</td>
<td>1.66494</td>
</tr>
<tr>
<td>T78</td>
<td>1.493663</td>
</tr>
<tr>
<td>A88</td>
<td>1.578137</td>
</tr>
<tr>
<td>I91</td>
<td>1.797509</td>
</tr>
<tr>
<td>I93</td>
<td>1.45592</td>
</tr>
<tr>
<td>S105</td>
<td>1.091038</td>
</tr>
<tr>
<td>I109</td>
<td>1.666515</td>
</tr>
<tr>
<td>L116</td>
<td>1.823261</td>
</tr>
<tr>
<td>V124</td>
<td>1.795841</td>
</tr>
<tr>
<td>J127</td>
<td>1.68361</td>
</tr>
<tr>
<td>N124</td>
<td>1.912942</td>
</tr>
<tr>
<td>C143</td>
<td>1.412108</td>
</tr>
<tr>
<td>A181</td>
<td>1.057631</td>
</tr>
<tr>
<td>K188</td>
<td>0.802183</td>
</tr>
<tr>
<td>J194</td>
<td>1.119343</td>
</tr>
<tr>
<td>Q208</td>
<td>1.492223</td>
</tr>
<tr>
<td>N214</td>
<td>1.305785</td>
</tr>
<tr>
<td>A215</td>
<td>1.19664</td>
</tr>
<tr>
<td>S216</td>
<td>0.733443</td>
</tr>
<tr>
<td>D240</td>
<td>0.764174</td>
</tr>
</tbody>
</table>

### Universal alignment | Bacterial alignment
RE avg | 0.693239 | 0.693239 | RE avg | 1.434491 |
MI cut off | 0.433444 | 0.433444 | MI cut off | 0.507384 |
References


34. Buslje CM, Santos J, Delfino JM, Nielsen M. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics 2009;25(9):1125-1131.


198. Dietrich S, Borst N, Schlee S, Schneider D, Janda J-O, Sterner R, Merkl R. Experimental assessment of the importance of amino acid positions identified by


