Title

Structure-function relationships in the phosphagen kinase and ribulose-phosphate binding barrel superfamilies

Permalink

https://escholarship.org/uc/item/2kw5p0px

Author

Novak, Walter Ray Pendola

Publication Date

2004

Structure-Function Relationships in the Phosphagen Kinase and Ribulose-Phosphate Binding Barrel Superfamilies

by

Walter Ray Pendola Novak

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry and Chemical Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Date University Librarian

Degree Conferred:..

for my wife, Kathleen

and in memory of my brother, Paul

PREFACE

This thesis does not reflect the work of a single individual, but rather reflects the guidance, support and friendships I have cultivated with a variety of people. My thesis advisor, Patsy Babbitt, deserves my utmost gratitude and respect. She has given me experience in developing and pursuing research projects, mentoring students, collaborations with other professors, and in reviewing and writing manuscripts. Because of her guidance, I feel I am well prepared for whatever tasks the future may present.

I wish to thank Tom Ferrin and Teri Klein for their guidance as members of my thesis committee. I enjoyed working closely with the Computer Graphics Laboratory, which Tom was always quick to support. Teri has always impressed me with her sincere concern for my interests, and I genuinely appreciate having such an advocate.

The members (or past members) of the Computer Graphics Laboratory have been great to work with. Tom Ferrin, Conrad Huang, Eric Pettersen, Greg Couch, Al Conde, Tom Goddard, Dan Greenblatt, Andrew Jewett and Sean Mooney (as a graduate student, he spent hours hiding out in the Treasure room), thank you.

I cannot thank John Cantwell enough for taking the time to teach a chemist molecular biology. If you know John, you know his attention to detail. It was to my benefit to learn from such a skilled individual. In addition, I will always remember the mornings in lab diverting ourselves with musical discussions and listening to KFOG.

Scott Pegg has become a great coworker and friend. He was always available to help me with any programming questions that I had. He is very straightforward and has a knack for identifying potential project problems before too much time has been wasted. I think I wondered for two years whether Scott liked me or not. In the end, I decided I liked him. He is a member of team Advil.

The entire, now very large, Babbitt lab deserves thanks for providing hours of conversation and other diversion. Ray Ray, thanks for being such a good target, and a good cook.

A final, deep thanks go out to all of my family for all of their support during this long journey. I am especially thankful to have had such a wonderful partner as Kathleen to share in this experience. We often alternated leaning on the other's shoulder and being that shoulder. I consider myself very blessed to know so many wonderful people.

ABSTRACT

Understanding the sequence-structure-function paradigm remains an important biological question. It is a question whose answer will change the way in which drugs are produced, the environment is cleaned and new products are synthesized. Great steps are being made towards understanding how a protein's sequence determines its structure, which, in turn, determines its function; however, there is still much to learn. In an effort to understand this paradigm, two enzyme superfamilies have been investigated with computational and experimental techniques.

Studies of the first superfamily, the phosphagen kinases, in Chapters 1-4 represent an attempt to extract detailed information from available sequences and structures in order to understand the specificity principles used by these enzymes. Human muscle creatine kinase (CK) and its homologs were examined using a variety of computational techniques. These experiments led to the investigation of the effects of two residues on the delivery of substrate specificity, Ile 69 and Val 325. Mutations at these positions were evaluated experimentally. Val 325 was identified as a "specificity switch," and two mutations at this position, Val 325 to Ala and Val 325 to Glu, altered CK to prefer cyclocreatine or glycocyamine, respectively.

The ribulose-phosphate binding barrel proteins (RPBB), the second superfamily investigated here in Chapter 5, in contrast to the phosphagen kinases, are a very diverse superfamily, and members often share less than 10% sequence identity. RPBB members possess the $(\beta/\alpha)_8$, (TIM-barrel) fold, and include the tryptophan and histidine biosynthesis enzymes, D-ribulose-5-phosphate 3-epimerases, and the orotidine 5'-monophosphate and 3-keto-L-gulonate 6-phosphate decarboxylases. It has been hypothesized that these (and many other TIM-barrels) have evolved from a common ancestor, yet, because of extremely low sequence identities, the foundation for this assertion remains unclear. Our studies indicate that many of the sequence links are achieved solely through similarities in the β7-8, or phosphate binding, region and lack similarity in the β1-6 region. Our data also suggest that these phosphate binding sites have evolved independently of the remainder of the barrel. Further, we demonstrate that other small two-β strand units within the β1-6 region appear to be circularly permuted within this superfamily.

Walter R. P. Novak
Doctoral Candidate

Patricia C. Babbitt

Patricia C. Babbitt, Ph.D.
Thesis Advisor

TABLE OF CONTENTS

PREFACE

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

APPENDIX 1

LIST OF FIGURES AND SCHEMES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

APPENDIX 1

APPENDIX 2

INTRODUCTION TO CHAPTER 1

When I began my rotation in Patsy's lab, I did not know much more than I wanted to work on a project that would involve both experimental and computational techniques. On the experimental side, I studied under John Cantwell, a postdoc in the Babbitt lab. I began learning the protein purification process and how to perform biochemical assays in order to investigate the catalytic mechanisms of phosphagen kinases. Study of the phosphagen kinases, particularly how substrate specificity is delivered in this system, makes up the bulk of my thesis and is discussed in other chapters. However; in addition to learning protein biochemistry, I was also encouraged to pursue my computational interests. Thus, I began my long relationship with the UCSF Computer Graphics Laboratory (CGL).

I was introduced to Andrew Jewett, who was then developing *MinRMS* (Jewett, et al., 2003), a program to produce a set of structural superpositions between a pair of enzymes. I became an alpha tester for this program, and the software under development by the CGL to view these superpositions, *Chimera.*

My greatest contributions to Chapter 1 are most likely those not explicitly expressed, but rather are hidden in the text and took place in the form of constant conferences with the developers. I learned a great deal about the development of software in this rotation, and I hope that my suggestions and complaints to the many developers were as helpful to them. As well as gaining exposure to the arena of software development, I learned a little about the structural similarities between creatine kinase and glutamine synthetase. Creatine kinase is a member of the phosphagen kinase enzyme superfamily, and, in contrast to the majority of protein folds, this fold has been currently found to perform only one function, the reversible transfer of the γ-phosphate of ATP to a guanidino substrate. A distant structural similarity between glutamine synthetase and creatine kinase was first noted by Kabsch and Fritz-Wolf (1997), and provided the rationale for my investigations into the basis of this similarity. A portion of these results were used as an example of the unique ability of the *Chimera* software package (Pettersen, et al., 2004) to facilitate comparisons between sequence and structure.

Jewett, A. I., C. C. Huang and T. E. Ferrin (2003). "MINRMS: an efficient algorithm for determining protein structure similarity using root-mean-squared-distance." *Bioinformatics* **19**(5): 625-34.

Kabsch, W. and K. Fritz-Wolf (1997). "Mitochondrial creatine kinase-a square protein." *Curr. Opin. Struct. Biol.* **7**(6): 811-818.

Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin (2004). "UCSF Chimera-A visualization system for exploratory research and analysis." *J. Comput. Chem.* **25**(13): 1605-12.

CHAPTER 1

Integrated Tools for Structural and Sequence Alignment and Analysis

Conrad C. Huang, Walter R. Novak, Patricia C. Babbitt, Andrew I. Jewett, Thomas E. Ferrin, and Teri E. Klein*

Published in the *Pacific Symposium on Biocomputing*, 2000, pp. 230-241.

Departments of Pharmaceutical Chemistry and Biopharmaceutical Sciences
University of California, San Francisco
San Francisco, California 94143-0446

*Corresponding Author: klein@cgl.ucsf.edu

ABSTRACT

We have developed new computational methods for displaying and analyzing members of protein superfamilies. These methods (*MinRMS*, *AlignPlot* and *MSFviewer*) integrate sequence and structural information and are implemented as separate but cooperating programs to our *Chimera* molecular modeling system. Integration of multiple sequence alignment information and three-dimensional structural representations enable researchers to generate hypotheses about the sequence-structure relationship. Structural superpositions can be generated and easily tuned to identify similarities around important characteristics such as active sites or ligand binding sites. Information related to the release of *Chimera, MinRMS, AlignPlot* and *MSFviewer* can be obtained at http://www.cgl.ucsf.edu/chimera.

INTRODUCTION

By July of 1999, the number of non-redundant protein sequences in the Genbank database had reached ~400,000 and included completed genome sequences for 23 organisms. These data provide an opportunity to explore the evolution of functional diversity by probing the entire repertoire of many organisms. One powerful approach to this study is the comparative analyses of large numbers of protein structures and their associated functions through primary sequence analysis and computer-assisted modeling of three-dimensional structures.

For example, discovery of a large enzyme superfamily whose members represent a surprising range of different chemical functions extended the insight that the evolution of new functions is linked to chemical capabilities associated with a given tertiary fold. (*1*, *2*) Because it illuminates the constraints built into the evolution of protein structure, this focus on chemistry is a crucial element for learning how new enzyme functions evolve from pre-existing structural scaffolds. This observation provides the conceptual framework for developing computational tools that integrate sequence and structure, and provides the basis for formulating hypotheses about function. The function of an unknown reading frame is rather easily deduced from sequence similarity when the function is the same as that of its homologs. For more divergent proteins, the predictions can be much more difficult because the function of each unknown enzyme may have no apparent relationship to that of its homologs. In each case, the crucial clues are provided by hidden similarities in chemical mechanisms that can be inferred from the structural comparisons. Because the most interesting insights come from relationships among such highly dissimilar proteins, we have developed methods to identify these distant sequence relationships (*3*) and to interpret them using tools designed to integrate sequence and structural information.

Aspects of this problem have been solved by a number of investigators. There are several examples of homology modeling packages such as the Swiss-Model (*4*) web server, Molecular Applications Group's LOOK (*5*) and Molecular Simulations Inc. (*6*) Homology and Insight II. There are also tools such as DINAMO (*7*), CINEMA (*8*), and PROTALIGN (*9*) and PROMUSE (*10*) which are useful in analyzing structure-sequence alignments. However, these tools have limitations such as extensibility, interactive real-time three-dimensional graphics display and analysis, and/or cost.

NEW COMPUTATIONAL AND ANALYSIS TOOLS

The set of tools, *MinRMS* (*11*), *AlignPlot* and *MSFviewer* were developed for sequence and structural alignment and analysis. These methods were easily integrated with *Chimera* (*12*) using Python (*13*), *Wrappy* (*14*), the *Object Technology Framework* (*15*), C and C ++. *MinRMS,* written in C ++, generates a family of structural alignments, allowing the user to explore the similarities between two proteins, including highly divergent structures (Figure 1). The unique ability to examine the optimum RMSD (Root Mean Square Distance) superpositions generated from the α-carbons of the structures being compared provides a much richer environment for exploring structural similarities than methods that produce a single pairwise alignment (*16*, *17*). Details of *MinRMS* and *Chimera* are published elsewhere (*11*, *12*).

The focus of this paper is on new tools for structural and sequence analysis and visualization. *AlignPlot*, written in C ++ and Python, provides a graphical representation of the RMSD values for each alignment in the set, allowing the user to quickly identify the regions of two structures that are most similar. Particularly important, it provides a user-friendly way to display specific alignments on the screen and navigate among them. *MSFviewer*, written in Python, provides an integrated link to sequence space, displaying multiple alignments of related sequences on the screen and providing for interactive highlighting of a selected structural alignment and the associated multiple sequence alignment.

MinRMS. Holm and Sander (*16*), Godzik (*17*), Fenz and Sippl (*18*), and Orengo et al. (*19*) have suggested that determining the single best structural alignment may not be possible. Given two proteins with experimentally-determined or modeled three-dimensional coordinates, *MinRMS* (*11*) solves this issue by generating multiple structural alignments and their corresponding sequence alignments. The *MinRMS* algorithm performs an exhaustive analysis of all plausible shape similarities between two proteins using RMSD between aligned α-carbon atoms. This method generates alignments containing all possible amino acid residues in a single pass without the need of parameters.

MinRMS uses intermolecular RMSD as the metric for comparing protein structures. The appropriateness of RMSD as a metric for comparing protein structures has been discussed elsewhere (*20-22*). The main advantage of the RMSD in that it is easy to interpret. The *MinRMS* algorithm is a two-step process: (1) Two proteins are rotated and translated to bring similarly shaped regions into close proximity; and, (2) With the two proteins fixed at a particular relative position, corresponding residues are chosen between the two proteins which minimize RMSD. Candidate superpositions are generated by selecting every fragment of 4 consecutive residues for each of the proteins and superimposing them by least-squared distance between α-carbons using the method described by Diamond (*23*). Given the relative positions of the two structures, we developed a new dynamic programming algorithm to choose the matching residues between the proteins that minimizes RMSD. Similar to the Needleman and Wunsch (*24*) algorithm, our algorithm is recursive and blind to "nontopological" similarities (*25*). For each candidate superposition, the algorithm generates multiple alignments containing different numbers of corresponding residues which minimize the intermolecular RMSD (*11*). The dynamic programming algorithm is applied to all candidate superpositions between the proteins with small local regions well matched. Typical execution time for aligning two proteins with an average length of 300 residues is less than 1 hour on an SGI Onyx 2.

The output of *MinRMS* is a large table of data that contains, for each structural alignment, the number of matched residues for the two proteins, the RMSD for the alignment, and the longest distance between any pair of matched residues. For comparison purposes, the - *log(P)* is calculated where *P* is the probability that a better alignment is found between two unrelated proteins occurring in the SCOP 26 database as described by Levitt and Gerstein (*22*). Structural alignment is presented in sequence alignment form as MSF (Multiple Sequence Format) files (Table 1). Relative positions are stored as comments in the MSF file. The program *Chimera*, in cooperation with *AlignPlot* and *MSFViewer*, is used to view the volumes of data produced from *MinRMS* .

Chimera. Chimera is a molecular visualization graphics package developed at the UCSF Computer Graphics Laboratory. *Chimera* is written in the Python programming language with C ++ extensions and uses standard multi-platform libraries such as the Tk toolkit for it's graphical user interface and OpenGL for three-dimensional graphics primitives. *Chimera* also uses the *Object Technology Framework* object class library for manipulating molecular data.

A major design goal for *Chimera* is program extensibility. By choosing Python as the *Chimera* command language, users can create complex command "scripts" (*e.g.*, with iterative loop and conditional execution) which in turn allow for sophisticated operations to be performed on multiple molecular models. Python has an extensive library (*13*) that include interfaces to Tk. This means that users are easily able to create their own custom graphical user interface (GUI) elements such as menus and dialog boxes as part of their extensions. *Chimera* itself is implemented with a small set of core functionalities, including graphical display, Protein Data Bank (PDB) input, and basic user interface elements (menu bars, tool bar, command line, reply window and status line). More advanced features are constructed on top of the core using Python extension modules. This results in a program architecture in which new functionality is easily added when needed. The separate applications *AlignPlot* and *MSFviewer* are example extensions of *Chimera*.

AlignPlot. AlignPlot displays three different representations that summarize the data from *MinRMS*. The bottom graph (Figure 1: RMSD *vs*. *N*) displays three numerical quantities as a function of matched residue pairs (*N*): RMSD, - *log(P)* of Levitt and Gerstein (*22*) and the longest pairwise distance between matched residues. *MinRMS* and Levitt and Gerstein scores are displayed to provide multiple evaluation criteria. Levitt and Gerstein favor matching more residues over better geometric fit. Thus, their method is less distance sensitive than *MinRMS*. The user can easily select a particular alignment by point and click with the mouse in the graph. The corresponding three-dimensional superposition is visualized in *Chimera*. Matched residues closer than one angstrom are denoted by a small sphere. Matched residues with a distance greater than one angstrom have a line drawn between them. This plot allows the user to discern patterns over the entire set of solutions.

The middle representation (Figure 1: Orientation Clusters) in *AlignPlot* uses a genetic algorithm (GA) to condense the data from *MinRMS* by selecting a small set of orientations to represent the entire data set. For any given set of representative orientations, a structure in a non-representative orientation contributes a penalty proportional to the RMSD from the most similar representative orientation. The GA "fitness" metric is the sum of penalties of all non-representative orientations. The GA uses the fitness metric to find a good representative set, which is then used to divide the data set into clusters. The clustering results are displayed as a table where the columns represent alignments and the rows represent clusters. The cells of the table are color-coded and the brightness of each cell is proportional to RMSD from the representative of that cluster. The cluster plot classifies the solutions into a small number of groups which reduce the amount of information that the user needs to process.

The top representation (Figure 1: Sequence vs. Sequence) is a two-dimensional histogram of residue pairs. Each cell of the histogram represents a pair of residues, one from each structure. The value of the cell is the number of *MinRMS* alignments that match the two residues. The value is converted to color. The color scale is blue to red representing values that range from 1 to the maximum cell value. If there is no match, the cell is colored like the background. Information displayed in this manner provides easy identification of matching patterns (*e.g.*, secondary structure matches appear as diagonal runs).

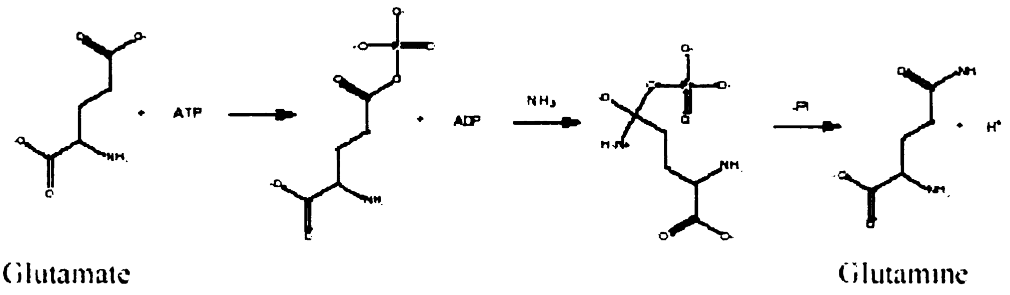
Using these three tools together, one can identify structural alignments of interest. The orientation cluster plot reduces hundreds of alignments into tens of alignments. The RMSD vs. N plot illustrates the trade-off between the number of matched residues and closeness of global superpositioning. Lastly, the Sequence vs. Sequence plot typically identifies secondary structural elements important to the alignment. These tools used in combination facilitates the analysis of a large data set.

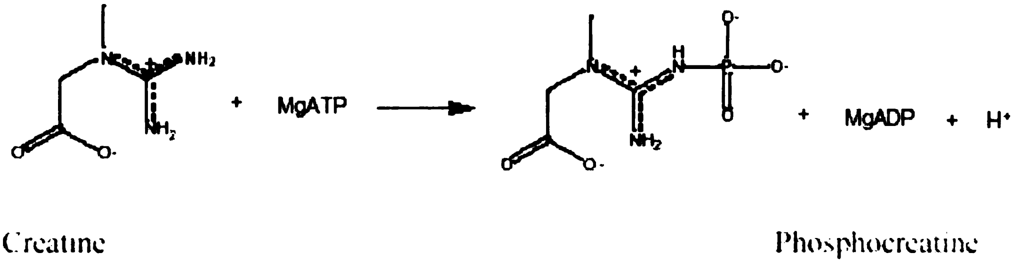
MSFviewer. *MSFviewer* was developed independently of *AlignPlot* to view multiple sequence alignment in MSF format (*e.g.*, an output option of commonly used multiple alignment programs). Fragments of the sequence can be selected and highlighted on the structure, allowing the user to focus on secondary structure elements, active site residues, monitoring of residues of interest and filtering of data (Figure 1). The alignment can be edited interactively, saved in MSF format or printed for presentation purposes (Figure 2).

MSFviewer cooperates with *AlignPlot* through *Chimera* for the selection and mapping of sequences to their structures. Selecting a specific alignment in *AlignPlot* will highlight the matched residues in *MSFviewer*. Upon selecting specific residues in *MSFviewer*, *AlignPlot* displays the matching statistics of these residues for each alignment. *Chimera* serves as the data repository and communication channel between *AlignPlot* and *MSFviewer*.

EXAMPLE: STRUCTURAL COMPARISONS OF GLUTAMINE SYNTHETASE WITH CREATINE KINASE AND OTHER GUANIDINO KINASES

The study of structure-function relationships is important to the understanding of proteins and provides guidance for protein engineering. We have attempted to better understand structure-function relationships through the structural comparison of glutamine synthetase (GS) with creatine kinase (CK) and other guanidino kinases. While GS and CK have no significant sequence similarity, they both have multimeric forms, have been proposed to have similar tertiary structures (Figure 3) (*27*), and catalyze similar reactions. GS catalyzes the reaction of glutamate and ammonia to form glutamine through a phosphorylated intermediate, while CK catalyzes the transfer of a phosphate group from ATP to creatine to yield phosphocreatine.





Liaw and Eisenberg (*28*) solved a series of crystal structures of GS to elucidate the mechanism of glutamine synthesis and to identify residues possibly involved in ATP binding and in transfer of the γ-phosphate. This structure of GS was superimposed with an available CK structure (*29*) using *MinRMS*. A family of several hundred structural superpositions resulted reflecting many possible orientations of GS and CK (Figure 1). Simultaneous viewing of the three-dimensional and sequence alignments and interactive editing of the sequence alignments allowed for comparison of catalytic residues and binding domains using all of the sequence and structural information available (Figures 4 and 5). These tools allowed us to examine the ATP-binding residues of GS and CK using sequence alignments informed by the structural superpositions. While crystal structures of CK bound with MgATP or substrate are not available, our studies indicate that many of the ATP binding residues in GS have potential homologs in CK.

The information gained from this analysis supports the previous suggestion that a similar scaffold is used in both GS and CK (*27*). The analyses of this work suggest that this scaffold also utilizes potentially homologous residues to bind ATP and assist in the transfer of the γ-phosphate group. Use of the tools described here have provided a useful model to continue the study of structure-function relationships in the guanidino kinases. Prior to using these tools, it was difficult to obtain a useful structural alignment.

CONCLUDING REMARKS

Superfamily analysis frequently involves proteins whose sequence similarities may fall below the level of *statistical significance* but whose relationships are nonetheless *biologically significant*. *MinRMS*, *AlignPlot*, *MSFviewer* along with Shotgun (*3*), in cooperation with *Chimera*, provide a set of tools for generating and testing hypotheses about sequence, structure and functional relationships of such proteins.

Initial testing of this software has suggested additional functionalities to allow users to choose the subsets of alignments that provide the best overlap over specific residues such as active site residues. More extensive editing capabilities will be added to facilitate correcting the registration between (1) sub-group multiple alignments of very distantly related sequences based on the structural alignments; and (2) very distantly related sequences based on the structural alignments of representative sub-group members. Lastly, we are exploring non-distance methods for comparing more than two proteins at one time.

Information on the availability of the software tools described here can be found at http://www.cgl.ucsf.edu/chimera.

ACKNOWLEDGMENTS

This work is supported by the Department of Energy (DE-FG03-96ER62269), NIH National Center for Research Resources (P41-RR01081) and NIH (AR17323).

REFERENCES

1. P.C. Babbitt, G.T. Mrachko, M.S. Hasson, G.W. Huisman, R. Kolter, D. Ringe, G.A. Petsko, G.L. Kenyon. and J.A. Gerlt, "A Functionally Diverse Enzyme Superfamily that Abstracts the α-protons of Carboxylic Acids." *Science* **267**: 1159-1161, 1995.

2. P.C. Babbitt, M. Hasson, J.E. Wedekind, D.J. Palmer, M.A. Lies, G.H. Reed, I. Rayment, D. Ringe, G.L. Kenyon, and J.A. Gerlt., "The Enolase Superfamily: A General Strategy for Enzyme-Catalyzed Abstraction of the α-protons of Carboxylic Acids." *Biochem.* **35**: 16489-16501, 1996.

3. S.C.-H. Pegg and B.C. Babbitt, "Shotgun: Getting More from Sequence Similarity Searches." *Bioinformatics* **15**(9):729-470, 1999.

4. N. Guex and M.C. Peitsch, "SWISS-MODEL and the Swiss-PdbViewer: An Environment for Comparative Protein Modeling," *Electrophoresis* **18**:2714-2723, 1997.

5. Molecular Applications Group, 607 Hansen Way, Building One, Palo Alto, California 94304. See http://www.mag.com/.

6. Molecular Simulations Inc., 9685 Scranton Road, San Diego, California 92121. See http://www.msi.com/.

7. M. Hansen, J. Bentz, A. Baucom and L. Gregoret, "DINAMO: A Coupled Sequence Alignment Editor/Molecular Graphics Tool for Interactive Homology Modeling of Proteins", *PSB* 106-117, 1998.

8. T.K. Attwood, A.W.R. Payne, A.D. Michie and D.J. Parry-Smith, "A Colour INteractive Editor for Multiple Alignments - CINEMA," *EMBnet.news* **3**, 1997.

9. D. Meads, M.D. Hansen and A. Pang, "PROTALIGN: A 3-Dimensional Protein Alignment Assessment Tool," *Pacific Symposium on Biocomputing* 354-367, 1999.

10. M.D. Hansen, E. Charp, S. Lodha, D. Meads and A. Pang, "PROMUSE: A System for Multi-Media Data Presentation of Protein Structural Alignments," *Pacific Symposium on Biocomputing* 368-379, 1999.

11. A.I Jewett, C. C. Huang, C. and T.E. Ferrin, “MinRMS: An Efficient Algorithm for Determining Protein Structure Similarity.” *Bioinformatics* **19**(5):625-634, 2003.

12. C.C. Huang, G.S. Couch, E.F. Pettersen and T.E. Ferrin, “Chimera: An Extensible Molecular Modeling Application Constructed using Standard Components”, *Pacific Symposium on Biocomputing*, 724, 1996.

13. See http://www.python.org/.

14. G.S. Couch, “Wrappy -- A Python Wrapper Generator for C++ Classes,” in O'Reilly Open Source Convention Python Conference Proceedings, 1999, http://conferences.oreilly.com/.

15. C.C. Huang, E.F. Pettersen, G.S. Couch, T.E. Ferrin, A.E. Howard and T.E. Klein, “The Object Technology Framework (OTF): An Object-Oriented Interface to Molecular Data and Its Application to Collagen.” *Pacific Symposium on Biocomputing*, 349-361, 1998.

16. L. Holm and C. Sander, “Protein Structure Comparison by Alignment of Distance Matrices”, *J. Mol. Biol.* **233**:123-138, 1993.

17. A. Godzik, “The Structural Alignment Between Two Proteins: Is there a Unique Answer?”, *Protein Science* **5**:1325-1338, 1996.

18. Z.K. Feng and M.J. Sippl, “Optimal Superimposition of Protein Structures: Ambiguities and Implications,” *Folding & Design* **1**:123-132, 1996.

19. C.A. Orengo, M.B. Swindells, A.D. Michie, M.J. Zvelebil, P.C. Driscoll, M.D. Waterfield and J.M. Thornton, “Structural Similarity Between the Pleckstrin Homology Domain and Verotoxin: The Problem of Measuring and Evaluating Structural Similarity,” *Protein Science* **4**:1977-1983, 1995.

20. F.E. Cohen and M.J.E. Sternberg, "On the Prediction of Protein Structure: The Significance of the Room Mean Squared Deviation," *J. Mol. Biol.* **138**:321-333, 1980a.

21. A. Falicov and F.E. Cohen, "A Surface of Minimum Area Metric for the Structural Comparison of Proteins," *J. Mol. Biol.* **258**:871-892, 1996.

22. M. Levitt and M. Gernstein, "A Unified Statistical Framework for Sequence Comparison and Structure Comparison," *PNAS* **95**:5913-5920, 1998.

23. R. Diamond, "A Note on the Rotational Superposition Problem," *Acta Cryst.* **A44**:211-216, 1988.

24 S.B. Needleman and C.D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," *J. Mol. Biol.* **48**:443-453, 1970.

25. I.N. Shindyalov and P.E. Bourne, "Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path," *Protein Engineering* **11**:739-747, 1998.

26. T.J. Hubbard, B. Ailey, S.E. Brenner, A.G. Murzin and C. Chothia, "SCOP, Structural Classification of Proteins Database: Applications to Evaluation of the Effectiveness of Sequence Alignment Methods and Statistics of Protein Structural Data," *Acta Crysta.* **D54**:1147-1154, 1998.

27. W. Kabsch and K. Fritz-Wolf, "Mitochondrial Creatine Kinase--A Square Protein," *Curr. Op. in Struct. Bio.*, **7**:811-818, 1997.

28. S-H Liaw and D. Eisenberg, "Structural Model for the Reaction Mechanism of Glutamine Synthetase, Based on Five Crystal Structures of Enzyme-Substrate Complexes," *Biochemistry*, **33**:675-681, 1994.

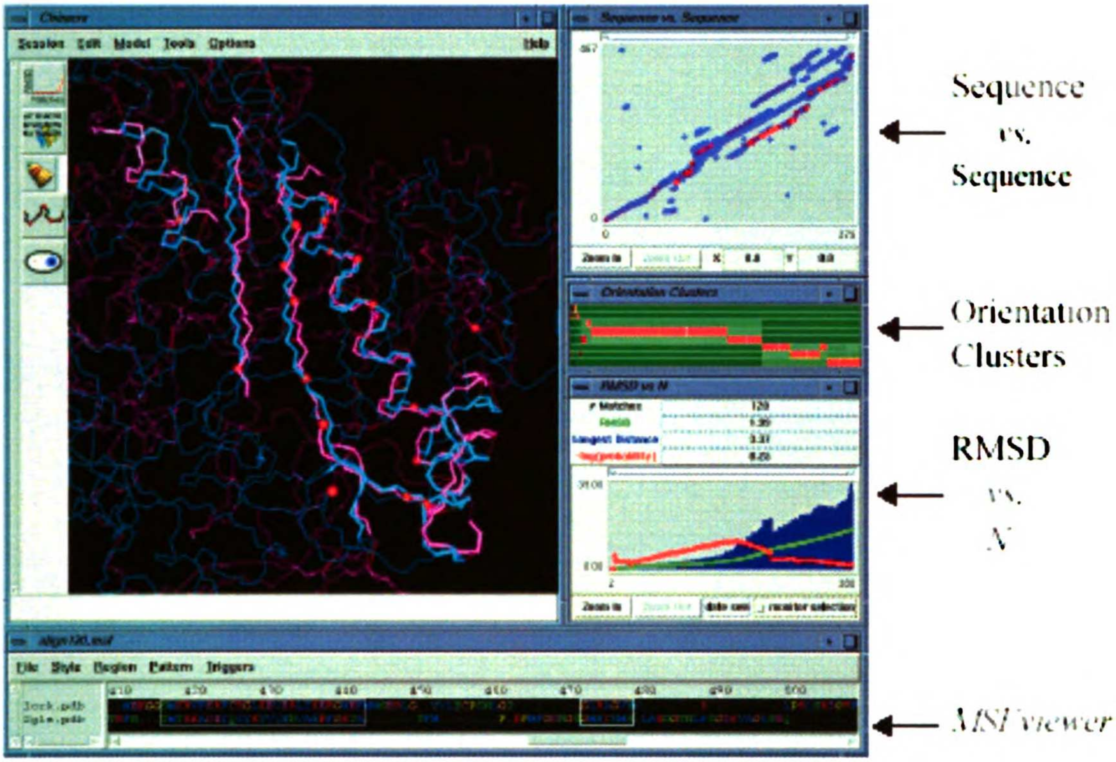
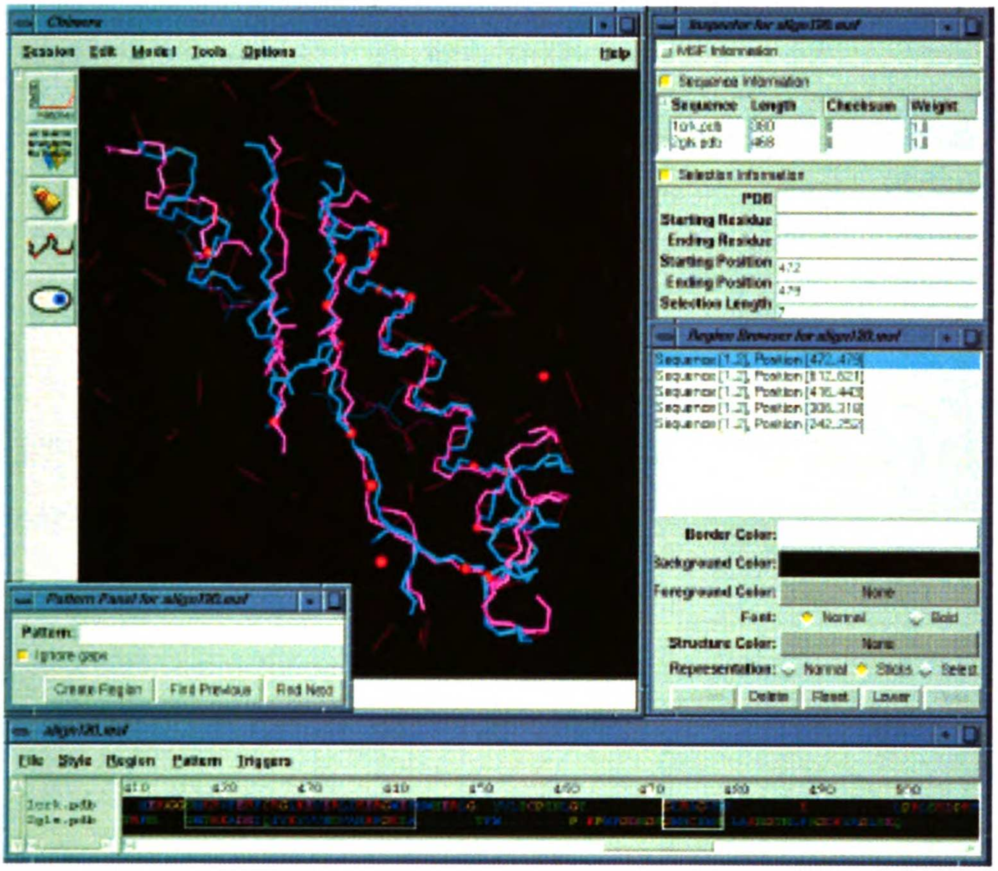


Figure 1. Screen display of *AlignPlot, MSFviewer* and *Chimera.* Glutamine synthetase is in magenta and creatine kinase is in cyan. Matched residue pairs are highlighted by red spheres and lines. See *sections 2.3 & 2.4* for detail descriptions.



Sequence	Length	Checksum	Weight
1ock.pdb	380	0	1.0
2gle.pdb	468	0	1.0

PDB
Starting Residue
Ending Residue
Starting Position 472
Ending Position 479
Selection Length 7

Sequence [1..2], Position [472..479]
Sequence [1..2], Position [612..621]
Sequence [1..2], Position [416..443]
Sequence [1..2], Position [308..318]
Sequence [1..2], Position [242..252]

Border Color:
Background Color:
Foreground Color: None
Font: ● Normal ○ Bold
Structure Color: None
Representation: ○ Normal ● Sticks ○ Select

Figure 2. Graphical user interface elements of *MSFviewer* are displayed.

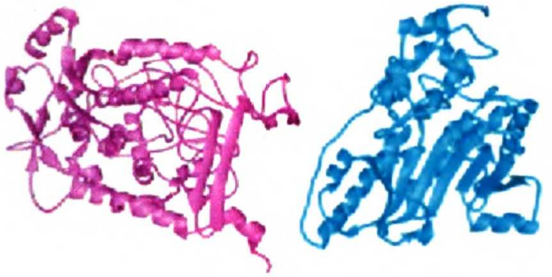


Figure 3. Ribbon representations of glutamine synthetase (magenta) and creatine kinase (cyan) prior to alignment with *MinRMS*.

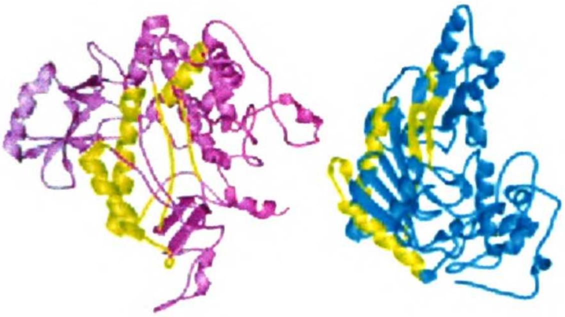


Figure 4. Ribbon representations of glutamine synthetase (magenta) and creatine kinase (cyan) post alignment with *MinRMS*. Matched regions are highlighted (yellow). The associated sequence alignment is seen in Figure 5.

```
Chimera minimal RMSD structural alignment with 120 equivalences.
RMSD = 1.988821
-----
Transform Matrix to apply to molecule: 2gls.pdb
0.580381 -0.537281 -0.611953 -9.842606
-0.744292 -0.654900 -0.130905 57.874092
-0.330435 0.531447 -0.779985 15.317198

 Name: 1crk.pdb          Len:   380  Check:     0  Weight:  1.00
 Name: 2gls.pdb          Len:   468  Check:     0  Weight:  1.00

1crk.pdb TVHEKRKLFP PSADYPDLRK HNNCMAECLT PAIYAKLRDK LTPNGYSLDQ CIQTGVDNPG
2gls.pdb .......... .......... .......... .......... .......... ..........

1crk.pdb HPFIKTVGMV AGDEESYEVF AEIFDPVIKA RHNGYDPRTM KHHTDL.... ..........
2gls.pdb .......... .......... .......... .......... ......SAEH VLTMLNEHEV

1crk.pdb .......... ..DAS..... .......... .......... .......... ..........
2gls.pdb KFVDLRFTDT KGK..EQHVT IPAHQVNAEF FEEGKMFDGS SIGGWKGINE SDMVLMPDAS

1crk.pdb .......... .......... .......... .KI...T..H GQF....... ..DERYVLS.
2gls.pdb TAVIDPFFAD STLIIRCDIL EPGTLQGYDR DP.RSIAKRA .E.DYLRATG IADT.....V

1crk.pdb .SRVRTGRSI R......... ........G. LSL....... .......... ....PPACSR
2gls.pdb LFGPEPEFFL FDDIRFGASI SGSHVAIDDI EG.AWNSSTK YEGGNKGHRP GVKGG.....

1crk.pdb .....AERRE VENVVVTAL. AGL..KG.DL SGKYYSLTNM SERDQQQLID DHFLFDKPVS
2gls.pdb YFPVPPVD.S AQDIRSE.MC L.VMEQ.MGL .......... .......... ..........

1crk.pdb PLLTCAGMAR DWPDARGIW. HNNDKTFLV. WINEED.... ..HTRVIS.. MEKGGNMKRV
2gls.pdb .......... .........V V......E.A HHH..EVATA GQNE.VA.TR FN...TMTKK

1crk.pdb FERFCRGLKE VERLIKERGW EFMWNERLG. .YVLTCPSNL GT........ .GLRAGVHV.
2gls.pdb ADEIQIYKYV VHNVAHRFGK TA.......T FM........ P.KPMFGDNG SGMHCHMS.L

1crk.pdb .......K.. ........LP RLSKDPRFPK I.....L..E NLRL...... ..........
2gls.pdb AKNGTNLFSG DKYAGLSEQ. .......... .ALYYIGGVI KHA.KAINAL ANPTTNSYKR

1crk.pdb .......... .......... .QKRGTGGVD .TAAVADVY. .....DI.SN LD.RMGRS..
2gls.pdb LVPGYEAPVM LAYSARNRSA SI.RIPV... VA.......S PKARRI.EV. ..RF....PD

1crk.pdb ..EVEL...V .QIVIDGVNY .LVDCEKKLE KGQDIKVPPP LP........ ..........
2gls.pdb PAAN..PYLC FAALLMAGLD GI..K..... ....N..... ..KIHPGEPM DKNLYDLPPE

1crk.pdb ........Q. ....FGR... .......... .......... .....K.... ..........
2gls.pdb EAKEIPQVAG SLEEA..LNA LDLDREFLKA GGVFTDEAID AYIALRREED DRVRMTPHPV

1crk.pdb ........
2gls.pdb EFELYYSV
```

Figure 5. MSF output from MinRMS of the sequence alignment for glutamine synthetase and creatine kinase. This structure alignment corresponding to this sequence alignment is displayed in Figure 4.

INTRODUCTION TO CHAPTER 2

Chapter 2 describes work spearheaded by John Cantwell, a former postdoctoral researcher in the Babbitt lab, and published in *Biochemistry*. This work sought to investigate the roles of two acidic residues in the catalytic mechanism of creatine kinase (CK). Dissecting the important catalytic residues in CK has remained a difficult task to this day. It largely appears that the precise positioning of the substrates is as important to catalysis as proton abstraction though there is much debate on the relative contributions. While several unliganded CK structures were available, at the time of this work, a thorough understanding of the catalytic mechanism was hindered by the lack of a CK crystal structure with a bound transition state analog complex. However, a crystal structure of the CK homolog, arginine kinase (AK), had recently been solved, and we set out to extract as much data as we could through sequence and structure analyses in order to guide the mutagenesis studies.

Thus, my contribution to this chapter of my thesis centered on the computational comparisons of the sequences and structures of CK and AK. From the AK structure, two glutamic acid residues (E225 and E314) were implicated in the catalytic mechanism. The first of these, E225, is strictly conserved in all phosphagen kinases, while the second, E314 is strictly conserved in all phosphagen kinases except CKs. In CKs the homologous residue to E225 is E232, and E314 appeared to be homologous to D326, one position to the C-terminus in sequence alignments. E225 is positioned in towards the center of the enzyme, while E314 is located on a C-terminal flexible loop region.

Although the mutagenesis work focused on the human muscle isoform of CK, for the structural comparisons we chose to use the structure of the ubiquitous mitochondrial isoform (uMtCK) over the available human muscle form. The motivation for this selection was dependent on the completeness of the available crystal structures. The human muscle CK structure was missing the C-terminal flexible loop region, and this was a primary region of focus in this work. Structural superpositions were performed between uMtCK (PDB 1CRK) and AK (PDB 1BG0) using *MinRMS*. The superposition we selected superimposed 236 of the ~380 α-carbons with an RMSD of 1.25 Å, and revealed significant differences in the locations of the C-terminal flexible loop. From the computational studies we further hypothesized that differences in the substrate specificities of these enzymes would be present on the C-terminal loop region.

This study found that mutations at E232 affected catalysis more than mutations at D326 in CK, implicating E232 as the catalytic base for proton abstraction. It was later discovered, with the advent of a CK structure with a transition state analog complex bound, that D326 is not homologous to AK E314. A hydrophobic binding pocket is formed in CK by the interaction of V325 (from the C-terminal loop) and I69 (from the N-terminal loop), and this hydrophobic pocket appears to serve the same role as E314 in AK. The roles of V325 and I69 in specificity and catalysis are the focus of Chapter 4. The work leading to the first crystal structure of CK bound with a transition state analog complex is the focus of Chapter 3.

CHAPTER 2

Mutagenesis of Two Acidic Active Site Residues in Human Muscle Creatine Kinase: Implications for the Catalytic Mechanism

John S. Cantwell[†], Walter R. Novak[†], Pan-Fen Wang[§], Michael J. McLeish[§], George L. Kenyon[§], and Patricia C. Babbitt[†*]

Reproduced with permission from J.S. Cantwell, W.R. Novak, P.F. Wang, M.J. McLeish, G.L. Kenyon and P.C. Babbitt, Mutagenesis of two acidic active site residues in human muscle creatine kinase: implications for the catalytic mechanism. *Biochemistry* **40** (2001), pp. 3056–3061. Copyright 2001 American Chemical Society.

[†] *Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, California 94143-0446,*

[§] *College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109-1065*

Revised Manuscript Received January 17, 2001

[*] To whom correspondence should be addressed.

Supported by USPHS Grant AR17323 to GLK and PCB.

ABSTRACT

Creatine kinase (CK) catalyzes the reversible phosphorylation of the guanidine substrate, creatine, by MgATP. Although several X-ray crystal structures of various isoforms of creatine kinase have been published, the detailed catalytic mechanism remains unresolved. A crystal structure of the CK homologue, arginine kinase (AK), complexed with the transition-state analogue (arginine-nitrate-ADP), has revealed two carboxylate amino acid residues (Glu225 and Glu314) within 2.8 Å of the proposed transphosphorylation site. These two residues are the putative catalytic groups that may promote nucleophilic attack by the guanidine amino group on the -phosphate of ATP. From primary sequence alignments of arginine kinases and creatine kinases, we have identified two homologous creatine kinase acidic amino acid residues (Glu232 and Asp326), and these were targeted for examination of their potential roles in the CK mechanism. Using site-directed mutagenesis, we have made several substitutions at these two positions. The results indicate that of these two residues the Glu232 is the likely catalytic residue while Asp326 likely performs a role in properly aligning substrates for catalysis.

Abbreviations: CK, creatine kinase; HMCK, human muscle creatine kinase; AK, arginine kinase; TSAC, transition-state analogue complex; RMCK, rabbit muscle creatine kinase; MtCK, mitochondrial creatine kinase; sMtCK, sarcomeric mitochondrial creatine kinase; uMtCK, ubiquitous mitochondrial creatine kinase; LB medium, Luria-Bertani medium; CyCr, cyclocreatine; ORF, open reading frame; EDTA, ethylenediaminetetraacetic acid; PMSF, phenylmethanesulfonyl fluoride; SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis.

INTRODCUTION

Creatine kinase (CK (*1*); EC 2.7.3.2) is responsible for the production and maintenance of a "high-energy" phosphate intermediate, phosphocreatine (PCr), that can be converted to ATP in cells requiring short, rapid bursts of energy. The isoforms of CK are named for the tissues and organelles with high-energy demands in which they were originally and predominantly found, namely the muscle [M-type (*1*)], brain [B-type (*2*)], and mitochondrial [Mt-type (*3*)] CK isozymes. Because of its central importance in cellular metabolism, cellular.CK concentrations and/or activity are perturbed in a wide range of human disease states, including cardiac infarct (*4*), cancer (*5-7*), muscular dystrophies (*8*), and neurodegenerative diseases (*9*, *10*).

CK has been a topic of active research efforts for over 50 years and serious investigations of its mechanism have spanned at least the last 30. The advances in understanding of the mechanism have coincided with the advent of new techniques. Cross-linking and affinity labeling studies have implicated a reactive cysteine [Cys283, HMCK numbering (*11*, *12*)], as well as lysine (*13*, *14*) and arginine residues (*15*) that interact with negatively charged phosphates of the substrates. In addition to cross-linking studies (*16*, *17*), NMR investigations (*18*) have suggested a reactive histidine, while pH profile data have suggested the presence of a catalytic histidine residue together with an assisting carboxylate (*19*).

Nearly two decades later, site-directed mutagenesis has been employed to confirm, modify, or reject the suspected mechanistic importance of some of these residues. For example, mutagenesis studies of the so-called "essential" cysteine (Cys283) show that significant activity can be retained even after amino acid substitution (*20*, *21*). In other cases, little or no activity is observed when this residue is replaced (*11*). However, despite these many investigations, the precise catalytic role of the cysteine remains to be determined. More readily interpreted results confirm that both arginine (*21*) and tryptophan (*22-24*) residues function as nucleotide-binding groups. Finally, mutagenesis of all five highly conserved histidine residues have excluded a centrally important catalytic function for any of them (*25*). This result strongly suggests that the long-held model of the chemical mechanism, which involves a histidine residue acting as a potential general base, must now be reevaluated.

The recent solution of several crystal structures of CK, including ubiquitous mitochondrial CK from chicken heart [uMtCK (*26*)], rabbit muscle CK [RMCK (*27*)], chicken brain b-form CK (*28*), and human ubiquitous mitochondrial CK (*3*), provides evidence confirming the roles of many residues previously implicated in substrate-binding or dimer and tetramer formation and/or stabilization. Unfortunately, however, no crystal structure of CK bound to either substrate or product analogues is yet available. Thus, the primary players in the CK mechanism, the catalytic residues (and therefore the details of the CK mechanism) remain largely unidentified. To some extent, this problem has been ameliorated by the recent publication of an arginine kinase structure (AK-TSAC) complexed with a transition-state analogue, arginine-nitrate-ADP (*29*). Arginine kinase (AK; EC 2.7.3.3) and CK are structurally similar members of the guanidino-kinase family, exhibit ~40% identity in their amino acid sequences, and show high levels of sequence conservation for all residues/motifs known to be functionally important. The structure of this liganded AK is superimposable upon the unliganded structure of RMCK at an RMSD of 1.25 Å over 236 of 381 total -carbons. Given these similarities, the AK-TSAC structure can be used as an approximate model of the CK active site. As described by Zhou et al. (*29*), two arginine kinase carboxylate residues, Glu 225 and Glu 314, are located in an ideal position to facilitate proton donation and/or abstraction from the N1 and N2 amine groups of arginine. In human muscle CK, the analogous carboxylates are Glu232 and Asp326, respectively. To provide insight concerning how the cytosolic CKs function at the molecular level, we have chosen to examine the human muscle variant of CK, both because it is important to human health and because it is highly similar (96% identity) to the enzyme from rabbit muscle (RMCK), the form on which most of the previous mechanistic work has been done. In this paper, we describe the replacement of each of these carboxylates with residues that are either functionally or structurally homologous. The results of these substitutions are compared with those from a nonconservative substitution, also generated for each of these carboxylates. Using this system, we can evaluate the roles of the carboxylates in both substrate binding and catalysis.

MATERIALS AND METHODS

Materials. Unless otherwise stated, chemicals and reagents were purchased from either Fisher Scientific or Aldrich-Sigma Chemicals. Restriction enzymes were purchased from New England Biolabs.

Bacterial Strains and Vectors. HMCK wild-type construct was created as previously described (*30*). HMCK mutant plasmid DNA was transformed into DH5α cells (Gibco BRL) for mutant construction, propagation, and sequencing.

Site-Directed Mutagenesis. Amino acid substitutions were created using the QuikChange kit (Stratagene), with mutagenic primers containing unique restriction site identifiers. Plasmids were purified using the NucleoSpin Plus Miniprep kit (Clontech), and mutant clones were identified by restriction site analysis. Sequence confirmation and oligonucleotide primer synthesis were performed by the Biomolecular Resource Center (BRC, University of California, San Francisco).

Growth and Overexpression. Plasmids containing the HMCK gene were transformed into BL21(DE3)pLysS cells (Stratagene) and spread onto LB agar plates containing carbenicillin (50 g/mL) and chloramphenicol (34 g/mL). Single colonies were used to inoculate 1-2 L of LB medium containing the appropriate antibiotics. Cells were grown in an incubator/shaker at 37 °C to an absorbance reading (A_{600}) of 0.8-1.0 at which point IPTG (0.4 mM) was added and the flasks were cooled to 20 °C for overnight growth (~15 h). Cells were harvested by centrifugation (15 min at 5000*g*), the cell pellet (~2.5-5 g/L) was washed with 25 mM Tris buffer, pH 7.5, collected and stored at -20 °C.

Purification. Cells were lysed in 25 mM Tris, 50 mM NaCl, 5 mM EDTA, and 0.1 mM PMSF by lysozyme treatment (0.3 mg/mL), three freeze-thaw cycles, and sonication. The cell lysate was centrifuged (30 min at 100000*g*), and the supernatant passed over a Q-Sepharose (Amersham-Pharmacia) anion-exchange column equilibrated with 25 mM Tris buffer, pH 7.0. CK eluted with the flow-through fraction. This fraction was then applied to a Blue Sepharose (Amersham-Pharmacia) dye affinity column which had been equilibrated in 20 mM MOPS buffer, pH 7.0. A 50 to 500 mM NaCl gradient was applied, and CK eluted between 200 and 300 mM NaCl. Fractions containing CK were combined and concentrated using an Amicon concentrator fitted with 30K molecular weight cutoff membrane.

Protein Concentration and Determination of Purity. Protein concentration was determined by the Bradford assay (*31*) with bovine serum albumin as the protein standard. Purity was determined by SDS-PAGE. NuPAGE 10% Bis-Tris precast gels (Invitrogen-Novex) were used with MOPS, pH 7.0, running buffer per manufacturer's instructions. CK-bands at 43 kDa represented 95% of the total protein per lane as determined by densitometry. Native gels were used to verify dimerization and charge shift of mutant CKs. NuPAGE 7% Tris-acetate precast gels were run with Tris-glycine native running and sample buffer per manufacturer's instructions.

Enzyme Assays. Creatine kinase activity was determined at 30 °C in the forward direction (PCr formation) by the NADH-linked assay method (*32*). The reverse direction (ATP production) was measured by the NADP-linked assay method (*33*). V_{max} and K_m values were calculated using Hyper.exe, a hyperbolic regression analysis program (version 1.0, copyright J. S. Easterby, 1992). Cyclocreatine, synthesized as previously described (*34*), was a kind gift of Dr Liangren Zhang.

Structural Superpositions. Structural superpositions were generated using MinRMS (*35*). The superpositions of creatine kinase (1CRK, A chain only) and arginine kinase (1BG0) were visualized using Chimera and the Chimera extension, AlignPlot (*35*). The superposition aligns 236 α-carbons with an average RMSD of 1.25 Å. The graphic shown in Figure 2 was created using MidasPlus (*36*).

RESULTS

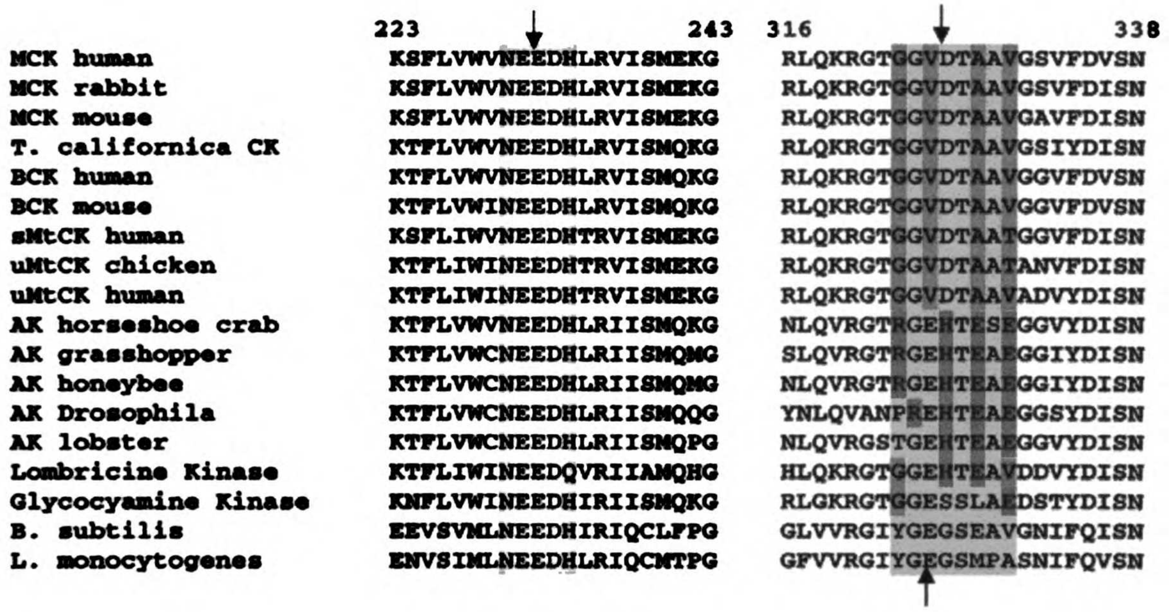
Sequence Homology. Figure 1 presents the primary sequence alignments of several species and isoforms of CK and AK in the regions of interest. The first region, residues 223-243 (numbered using the HMCK sequence), shows high sequence similarity among all known guanidino kinases, and includes a region we have designated the "NEED-box". This region is highlighted in Figure 1. A second region, HMCK residues 316-338 in Figure 1, has been identified in several of the guanidino kinase superfamily structure studies as containing an active-site flexible loop. This flexible loop, also highlighted in Figures 1 and 2 (HMCK residues 323-330), has been associated with a conformational change that occurs upon substrate binding (*37*). Two putative bacterial guanidino-kinase ORFs (*Bacillus subtilis* and *Listeria monocytogenes*) have been included in the alignments in order to contrast the relative homology between and within species and isozymes for AKs and CKs. The NEED-box represents a highly conserved motif across all of these sequences with ~80% identity for each pairwise comparison in this region. The highlighted NEED-box region has only one amino acid difference among all of the sequences shown, including the very distantly related bacterial ORFs. By contrast, the flexible loop exhibits much more variability across the alignment, with all CKs highly similar to each other, all AKs highly similar to each other, but with only 38% identity between CKs and AKs. Within a core of eight residues that surround the AK Glu314 or CK Asp326 position there is a distinct variation of side-chain properties for four of the eight residues. In CK there is a predominance of small nonpolar groups (Gly323, Val325, Ala328, and Val330), whereas the corresponding AK residues are charged (Arg312, His315, Glu317, and Glu319).

Purification. Wild-type and mutant HMCKs display nearly identical elution profiles from Blue-Sepharose. SDS-PAGE analysis reveals similar levels of yield, purity and subunit migration distances (data not shown). In a native gel, the E232D and D326E mutants comigrate with wild-type HMCK, while the E232Q and E232A also comigrate with each other, but at a slower rate than wild-type HMCK. Finally, the D326N and D326A variants have identical (to each other) but even slower migration behavior. Conspicuous differences in these native gel migration patterns of the purified mutant and wild-type HMCKs presumably are attributable to charge differences.

Kinetics. Functionally-, structurally-, and nonconservative amino acid substitutions, generated at each of the two residues, E232 and D326, led to large differences in their respective V_{max} values (Table 1). A functionally conservative substitution at the Glu232 position (E232D) results in a 500-fold loss of activity, while the comparable mutation at Asp326 (D326E) results in only a 3-fold loss in activity. Structurally conservative substitution results in a nearly 100000-fold decrease in activity for E232Q, while the D326N construct produces only a 20-fold decrease in catalytic rate. Last, the nonconservative E232A substitution completely eliminates detectable enzymatic activity, while the analogous variant, D326A, retains 0.1% of wild-type activity. In agreement with previously published values (*38*), the reverse reaction activity rates are 2-3 times higher than the forward reaction rates. The sole exception is the D326A mutant, which has 10-fold greater activity in the reverse direction than in the forward direction.

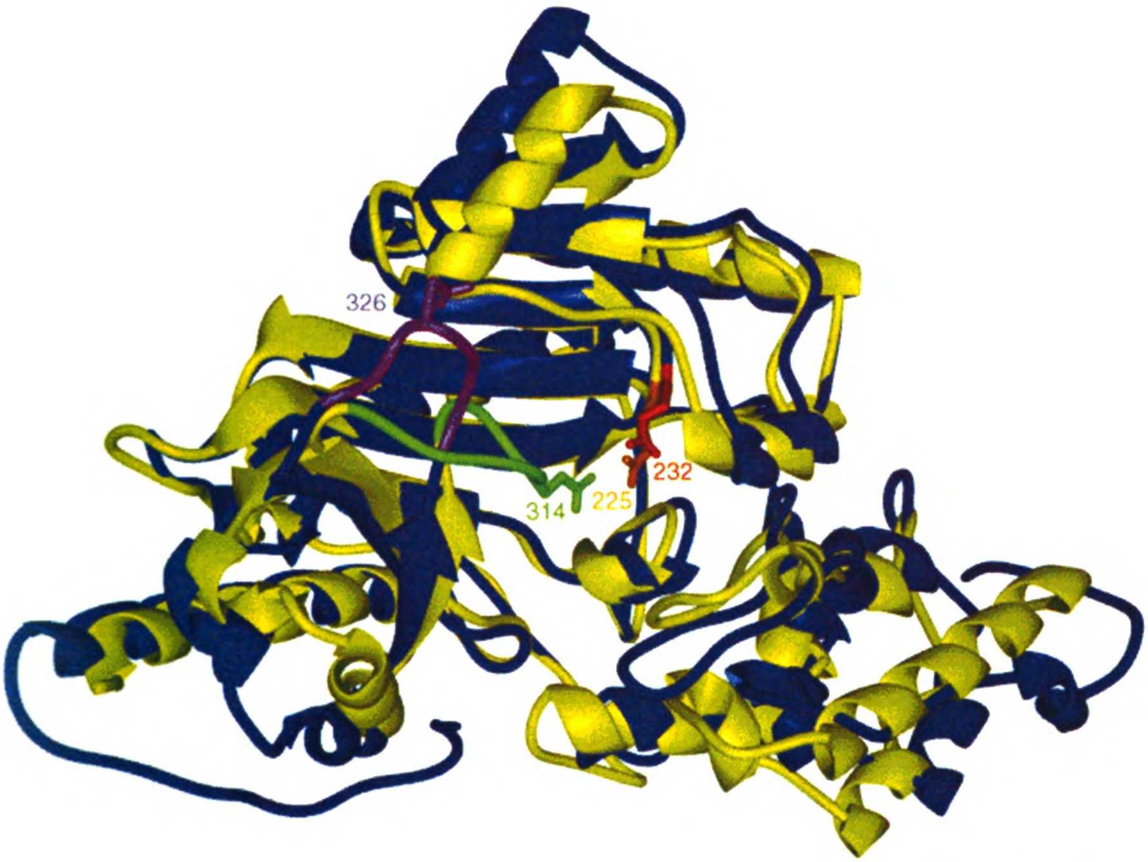
The Michaelis constants (K_m values), shown in Table 1, show only relatively minor changes from wild-type. In the forward reaction, no significant differences in K_m were observed for E232D (Table 1). Only D326A had any noticeable increase in both K_m(ATP) and K_m(Cr); about 3- and 4-fold, respectively. For the reverse reaction, D326A again possessed the most conspicuous increase in K_m(PCr), about 6-fold, while the E232D mutant had a 2.5-fold increase in K_m(PCr). From these values, it would seem that neither of these residues has a major impact on substrate interactions.

Reaction with Cyclocreatine. Of the available creatine analogues, cyclocreatine (CyCr) exhibits the most reactivity with CK (*34*). In cyclocreatine, the positions of the atoms and the bond angles fixed by the ring structure are expected to be very close to those adopted by creatine in the enzyme-substrate complex (*39*). However, the addition of a methylene bridge to form cyclocreatine fixes the stereochemistry of the two possible phosphorylation sites, which makes it possible to orient this site on the substrate relative to residues in the active-site cleft of CK (Figure 3). We studied the reaction of wild type and variant CKs with cyclocreatine with the expectation that CyCr might provide additional discrimination between the functions of the two carboxyl residues. Because phosphocyclocreatine (1-carboxymethyl-2-imino-3-phospho-4-imidazoline) is a much poorer substrate for CK than phosphocreatine (*40*), the reaction with cyclocreatine was studied only in the forward reaction. Table 2 shows that for wild-type CK the *K*m for cyclocreatine is about 20 mM, i.e., about 2-fold higher than that for creatine. Substitutions with the two noncharged amino acid substitutions, D326N, and D326A, result in the phosphorylation of CyCr at equal or higher rates than those measured for the phosphorylation of Cr. In contrast, the wild-type and negatively charged CK variants, E232D and D326E, phosphorylate CyCr at only 30% of their respective Cr phosphorylation rates.



	223 243	316 338
MCK human	KSFLVWVNEEDHLRVISMEKG	RLQKRGTGGVDTAAVGSVFDVSN
MCK rabbit	KSFLVWVNEEDHLRVISMEKG	RLQKRGTGGVDTAAVGSVFDISN
MCK mouse	KSFLVWVNEEDHLRVISMEKG	RLQKRGTGGVDTAAVGAVFDISN
T. californica CK	KTFLVWVNEEDHLRVISMQKG	RLQKRGTGGVDTAAVGSIYDISN
BCK human	KTFLVWVNEEDHLRVISMQKG	RLQKRGTGGVDTAAVGGVFDVSN
BCK mouse	KTFLVWINEEDHLRVISMQKG	RLQKRGTGGVDTAAVGGVFDVSN
sMtCK human	KSFLIWVNEEDHTRVISMEKG	RLQKRGTGGVDTAATGGVFDISN
uMtCK chicken	KTFLIWINEEDHTRVISMEKG	RLQKRGTGGVDTAATANVFDISN
uMtCK human	KTFLIWINEEDHTRVISMEKG	RLQKRGTGGVDTAAVADVYDISN
AK horseshoe crab	KTFLVWVNEEDHLRIISMQKG	NLQVRGTRGEHTESEGGVYDISN
AK grasshopper	KTFLVWCNEEDHLRIISMQMG	SLQVRGTRGEHTEAEGGIYDISN
AK honeybee	KTFLVWCNEEDHLRIISMQMG	NLQVRGTRGEHTEAEGGIYDISN
AK Drosophila	KTFLVWCNEEDHLRIISMQQG	YNLQVANPREHTEAEGGSYDISN
AK lobster	KTFLVWCNEEDHLRIISMQPG	NLQVRGSTGEHTEAEGGVYDISN
Lombricine Kinase	KTFLIWINEEDQVRIIAMQHG	HLQKRGTGGEHTEAVDDVYDISN
Glycocyamine Kinase	KNFLVWINEEDHIRIISMQKG	RLGKRGTGGESSLAEDSTYDISN
B. subtilis	EEVSVMLNEEDHIRIQCLFPG	GLVVRGIYGEGSEAVGNIFQISN
L. monocytogenes	ENVSIMLNEEDHLRIQCMTPG	GFVVRGIYGEGSMPASNIFQVSN

Figure 1. Multiple sequence alignment of important regions of several guanidino-kinases proposed to contain active-site residues. Alignments were generated using ClustalW (*46*). Regions of interest as described in the text are highlighted in yellow, and conserved carboxyl residues, based on structural alignments, are marked with arrows. Residue numbering is based on the HMCK sequence. For the flexible loop region, residues specific to CK specific are in blue and residues specific to AK are in magenta. MCK human, gi125305; MCK rabbit, gi125307; MCK mouse, gi125306; T. californica CK, gi125309; BCK human, gi125294; BCK mouse, gi417208; sMtCK human, gi125312; uMtCK chicken, gi2497494; uMtCK human, gi125315; AK horseshoe crab, gi1708613; AK grasshopper, gi1688218; AK honeybee, gi7434587; AK Drosophila, gi1346366; AK lobster, gi585342; Lombricine Kinase, gi3183058; Glycocyamine Kinase, gi1730042; B. subtilis, gi2127054; L. monocytogenes, gi1314296.



UCSF MidasPlus

Figure 2. Structural superposition of uMtCK and AK. uMtCK α-carbon trace is in blue, while AK is in yellow. The CK flexible loop is in purple, the AK loop is in green. Glu225 of the AK structure is colored orange while the CK Glu 232 is in red.

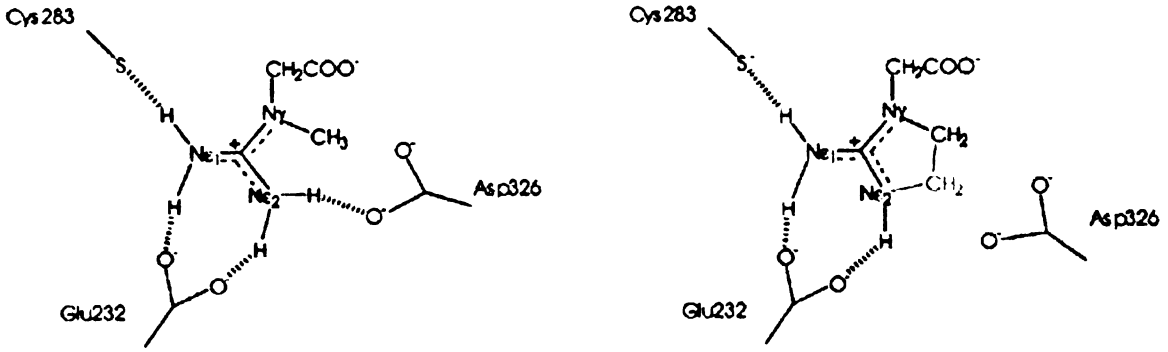


Figure 3. Schematic drawing of proposed interaction between HMCK and creatine or cyclocreatine. Hashed bonds represent potential hydrogen bonds. Red bonds represent the additional methylene bridge of cyclocreatine.

Table 1: Maximum Specific Activity[a] and Michaelis Constants[b] for Wild-Type and Recombinant HMCKs

Protein	V_{max} forward (units/mg)	V_{max} reverse (units/mg)	K_m(Cr)	K_m(ATP)	K_m(PCr)	K_m(ADP)
Wild-type	114 ± 7.4	304 ± 6.0	10.3 ± 2.4	0.504 ± 0.11	0.714 ± 0.061	0.0484 ± 0.0027
E232D	0.204 ± 0.012	0.591 ± 0.016	16.9 ± 2.4	0.492 ± 0.045	1.80 ± 0.15	0.0757 ± 0.0051
E232Q	0.0015 ± 0.0005	0.0035 ± 0.001	*n.d.*[c]	*n.d.*[c]	*n.d.*[c]	*n.d.*[c]
E232A	*n.d.*[c]	*n.d.*[c]	*n.d.*[c]	*n.d.*[c]	*n.d.*[c]	*n.d.*[c]
D326E	32.2 ± 1.6	116 ± 2.8	18.0 ± 5.3	1.03 ± 0.12	0.932 ± 0.088	0.0387 ± 0.0031
D326N	5.72 ± 0.16	16.0 ± 0.31	14.0 ± 5.5	0.838 ± 0.062	1.48 ± 0.051	0.0210 ± 0.0015
D326A	0.0776 ± 0.0020	0.834 ± 0.014	47.4 ± 11	1.37 ± 0.082	4.27 ± 0.16	0.0458 ± 0.0048

[a] Reaction conditions: Forward, pH 9.1, saturating ATP and creatine. Reverse, pH 7.0, saturating ADP and phosphocreatine. [b] Units are in millimolar. [c] Not detected.

Table 2: Kinetic Parameters for Wild-Type and Recombinant HMCKs Using Cyclocreatine as Substrate

Protein	k_{cat}[a] (s^{-1})	K_m (mM)	k_{cat}/K_m (s^{-1} M^{-1})	k_{cat}(Cycr)/k_{cat}(Cr)
Wild-type	38.4 ± 2.7	22.8 ± 2.8	1680	0.31
E232D	0.0462 ± 0.0061	25.0 ± 2.6	1.85	0.23
D326E	10.4 ± 0.75	19.4 ± 2.7	536	0.32
D326N	6.23 ± 0.84	27.9 ± 6.4	223	1.1
D326A	0.308 ± 0.055	39.4 ± 11	7.82	2.6

[a] k_{cat} for forward reaction.

DISCUSSION

Evidence for a difference in function for the two carboxylates under investigation is provided in part by primary sequence alignments. The NEED-box region (residues 223-243 in Figure 1), in which E232 is located, is highly conserved across all members of the guanidino-kinase family for the 20 residue stretch shown in Figure 1. Conversely, the flexible loop (HMCK residues 316-338 in Figure 1), which contains D326, is only ~40% identical for all pairwise comparisons within the CK family. Moreover, this region is clearly and consistently different between the CK and AK families. The importance of this flexible loop to the overall function of all superfamily members is suggested by measurements using small-angle X-ray scattering (*37*). These show that, notwithstanding their primary structure differences, CK and AK seem to behave similarly with respect to the movement of the flexible loops upon substrate binding.

The strikingly different levels of activity loss associated with substitutions generated at these two residues also suggest different catalytic roles for each. Glu232 is intolerant of any change; even a conservative substitution (E232D), which retains the functional group but shortens the carbon chain by a single methylene unit, results in 500-fold loss of activity. Replacement of the carboxyl group by substitution with the structurally similar but uncharged glutamine nearly eliminates activity. Consistent with these results, the nonconservative replacement of Glu232 with Ala results in complete loss of activity, within the limits of detection (>10-6 units/mg). This trend of loss in activity is characteristic of the removal of a critical catalytic residue. Further, the log $V_{\max}$ vs pH and log $V_{\max}/K_{\mathrm{m}}$ vs pH profiles for wild-type and E232D variants are very similar (data not shown). Consequently, the possibility that the activity observed for E232D is due to contamination by revertants to wild-type cannot be ruled out. In contrast, Asp326 is much more tolerant of amino acid replacement than Glu232. The conservative mutation (D326E) retains a significant proportion of the wild-type enzymatic activity. Exchange of a charged for an uncharged residue causes a significant drop (20-fold) in catalysis (D326N), while complete removal of the carboxylate group (D326A), produces an approximately 3 orders of magnitude loss of activity. Taken together these data suggest that, although Asp326 is required for optimal activity, CK can still function at a reasonable rate without it.

Assuming that the TSAC-AK structure is a suitable model for the analogous substrate-bound CK, then the flexible loop in CK should contribute to catalysis by folding from an open to a closed position, thus completing the active-site pocket and bringing into proximity residues that might contribute to catalysis (Figure 2). In the AK structure, the flexible loop residue Glu314, the putative homologue of CK Asp326, comes into proximity with the transphosphorylation site. Here it could coordinate either the guanidine or phosphoguanidine groups, thereby providing optimal alignment for in-line attack (*41*). Data from our kinetic experiments with CyCr support this concept; a decrease in k_{cat} resulting from a charged-to-nonpolar residue substitution (D326A) may be partially alleviated by a compensatory change in the substrate, i.e., the addition of the methylene bridge in CyCr (Figure 3). On the basis of these observations, we propose that in the CK mechanism, Asp326 may have a role in properly aligning the substrate by virtue of its residence on the flexible loop and the subsequent positioning of this loop proximal to the creatine or phosphocreatine molecule prior to transphosphorylation. We suggest that the differences in loop design, i.e., small nonpolar amino acids in the CK loop (HMCK residues 323-330) versus the largely negatively charged residues of the homologous AK loop (residues 312-319 in Figure 1), are likely to account for differences in substrate specificity between CK and AK.

Although the relative importance of both Glu232 and Asp326 in the HMCK mechanism have been confirmed, the details of their contributions remain unresolved. However, it is informative to view the current data in terms of earlier models of CK action. Nearly 20 years ago Cook et al. (*19*) proposed that, for RMCK, a group with a p*K*a of 7 must be unprotonated in the forward reaction and protonated in the reverse reaction, thereby acting as an acid-base catalyst. On the basis of a p*K*a value of 7 and an observed decrease in that p*K*a during solvent perturbation experiments, this acid-base catalyst was originally proposed to be a histidine. This putative active-site histidine was proposed to facilitate catalysis by accepting a proton from or donating a proton to the N2 of creatine and phosphocreatine, either helping to create a good nucleophile or making the phosphate on PCr a better leaving group. However, in a later study, Chen et al. (*25*) carried out a mutagenic analysis of all the conserved histidine residues in RMCK, concluding that none was essential for catalysis.

On the basis of the results described here, we now propose that Glu232 may perform the role of the acid-base catalyst that was assigned to a histidine in the earlier model (*19*). This residue was first suggested by Zhou et al. (*29*) as being a contributing feature in the AK reaction. One major concern with this hypothesis is that the p*K*a of Glu232 would need to be shifted from 4 to a value of nearly 7 in order to facilitate reversible proton transfer. However, studies of other enzymes provide a precedent for the elevation of the p*K*a of glutamate or aspartate residues, generally as a result of acidic or hydrophobic residues in the vicinity of the ionizable group. These include human lysozyme (*42*, *43*) and the phospholipase C reaction (*44*). On the basis of the structure of RMCK, which has 96% sequence identity with HMCK, the residues Glu231, Asp233, Pro143, Thr281, and Leu202, are potential contributors to such an elevation in p*K*a for Glu232. Further, some of the substrates also carry acidic groups (ADP, ATP, or PCr) that could, upon binding, contribute to an elevated p*K*a.

A complex picture of the CK reaction is now emerging from recent mutagenic and structural research. The mechanism must explain the participation of reactive glutamic acid and cysteine residues, the involvement of the flexible loop, as well as nucleotide and guanidinium binding residues, in both the forward and reverse reactions. Clearly, there must be several contributors to the overall mechanism (*45*). The AK structure provides a detailed view of a guanidino-kinase active site at the midpoint in the phosphoryl transfer reaction, which can be used to infer some mechanistic properties of other superfamily members, e.g., CKs, and guided our selection of two conserved carboxylates, Glu232 and Asp326, for amino acid replacement. Finally, between Glu232 and Asp326, our results have identified Glu232 as the acidic residue more vital for catalysis.

REFERENCES

1. Kuby, S. A., Noda, L., and Lardy, H. A. (1954) *J. Biol. Chem. 209*, 191-201.
2. Dawson, D. S., Eppenberger, H. M., and Kaplan, N. O. (1965) *Biochem. Biophys. Res. Commun. 21*, 346-353.
3. Eder, M., Fritz-Wolf, K., Kabsch, W., Wallimann, T., and Schlattner, U. (2000) *Proteins 39*, 216-225.

4. Apple, F. S. (1999) *Coronary Artery Dis. 10*, 75-79.

5. Hoosein, N. M., Martin, K. J., Abdul, M., Logothetis, C. J., and Kaddurah-Daouk, R. (1995) *Anticancer Res. 15*, 1339-1342.

6. Schiffenbauer, Y. S., Meir, G., Cohn, M., and Neeman, M. (1996) *Am. J. Physiol. 270*, C160-169.

7. Zarghami, N., Giai, M., Yu, H., Roagna, R., Ponzone, R., Katsaros, D., Sismondi, P., and Diamandis, E. P. (1996) *Br. J. Cancer 73*, 386-390.

8. Ozawa, E., Hagiwara, Y., Yoshida, M. (1999) *Mol. Cell. Biochem. 190*, 143-151.

9. David, S., Shoemaker, M., and Haley, B. E. (1998) *Brain Res. Mol. Brain Res. 54*, 276-287.

10. Aksenova, M. V., Aksenov, M. Y., Payne, R. M., Trojanowski, J. Q., Schmidt, M. L., Carney, J. M., Butterfield, D. A., and Markesbery, W. R. (1999) *Dementia Geriatr. Cognit. Disord. 10*, 158-165.

11. Zhou, H.-M., and Tsou, C.-L. (1987) *Biochim. Biophys. Acta 911*, 136-143.

12. Buechter, D. D., Medzihradszky, K. F., Burlingame, A. L., and Kenyon, G. L. (1992) *J. Biol. Chem. 267*, 2173-2178.

13. Kassab, R., Roustan, C., and Pradel, L. A. (1968*) Biochim. Biophys. Acta 167*, 308-316.

14. James, T. L., and Cohn, M. (1974) *J. Biol. Chem. 249*, 2599-2604.

15. Wood, T. D., Guan, Z., Borders, C. L., Jr., Chen, L. H., Kenyon, G. L., and McLafferty, F. W. (1998) *Proc. Natl. Acad. Sci. U.S.A. 95*, 3362-3365.

16. Pradel, L. A., and Kassab, R. (1968) *Biochim. Biophys. Acta 167*, 317-325.

17. Clarke, D. E., and Price, N. C. (1979) *Biochem. J. 181*, 467-475.

18. Rosevear, P. R., Desmeules, P., Kenyon, G. L., and Mildvan, A. S. (1981) *Biochemistry 20*, 6155-6164.

19. Cook, P. F., Kenyon, G. L., and Cleland, W. W. (1981) *Biochemistry 20*, 1204-1210.

20. Furter, R., Furter-Graves, E. M., and Wallimann, T. (1993) *Biochemistry 32*, 7022-7029.

21. Lin, L., Perryman, M. B., Friedman, D., Roberts, R., and Ma, T. S. (1994) *Biochim. Biophys. Acta 1206*, 97-104.

22. Zhou, H.-M., and Tsou, C.-L. (1985) *Biochim. Biophys. Acta 830*, 59-63.

23. Gross, M., Furter-Graves, E. M., Wallimann, T., Eppenberger, H. M., and Furter, R. (1994) *Protein Sci. 7*, 1058-1068.

24. Hagemann, H., Marcillat, O., Buchet, R., and Vial, C. (2000) *Biochemistry 39*, 9251-9256.[Full text - ACS]

25. Chen, L. H., Borders, C. L., Vasquez, J. R., and Kenyon, G. L. (1996) *Biochemistry 35*, 7895-7902.[Full text - ACS]

26. Fritz-Wolf, K., Schnyder, T., Wallimann, T., and Kabsch, W. (1996) *Nature 381*, 341-345.

27. Rao, J. K., Bujacz, G., and Wlodawer, A. (1998) *FEBS Lett. 439*, 133-137.

28. Eder, M., Schlattner, U., Becker, A., Wallimann, T., Kabsch, W., and Fritz-Wolf, K. (1999) *Protein Sci. 8*, 2258-2269.

29. Zhou, G., Somasundaram, T., Blanc, E., Parthasarathy, G., Ellington, W. R., and Chapman, M. S. (1998) *Proc. Natl. Acad. Sci. U.S.A. 95*, 8449-8454.

30. Chen, L. H., White, C. B., Babbitt, P. C., McLeish, M. J., and Kenyon, G. L. (2000) *J. Protein Chem. 19*, 59-66.

31. Bradford, M. M. (1976) *Anal. Chem. 72*, 248-254.

32. Tanzer, M., and Gilvarg, C. (1959) *J. Biol. Chem. 234*, 3201-3204.

33. Rosalki, S. B. (1967) *J. Lab. Clin. Med. 69*, 696-705.

34. Rowley, G. L., Greenleaf, A. L., and Kenyon, G. L. (1971) *J. Am. Chem. Soc. 93*, 5542-5551.

35. Huang, C. C., Jewett, A. I., Novak, W. R., Ferrin, T. E., Babbitt, P. C., and Klein, T. E. (2000) in *Pacific Symp. Biocomput. 2000* (Altman, R. B., Dunker, A. K., Hunter, L., and Klein, T. E., Eds.) pp 230-241, World Scientific, Singapore.

36. Ferrin, T. E., Huang, C., C., Jarvis, L. E. and Langridge, R. (1988) *J. Mol. Graphics 6*, 13-27, 36-37.

37. Forstner, M., Kriechbaum, M., Laggner, P., and Wallimann, T. (1998) *Biophys. J. 75*, 1016-1023.

38. Jacobs, H. K., and Kuby, S. A. (1980) *J. Biol. Chem. 255*, 8477-8482.

39. Phillips, G. N., Thomas Jr., J. W., Annesley, T. M., and Quiocho, F. A. (1979) *J. Am. Chem. Soc. 101*, 7120-7121.

40. Annesley, T. M., and Walker, J. B. (1977) *Biochem. Biophys. Res. Commun. 74*, 185-190.

41. Hansen, D. E., and Knowles, J. R. (1981) *J. Biol. Chem. 256*, 5967-5969.

42. Inoue, M., Yamada, H., Yasukochi, T., Kuroki, R., Miki, T., Horiuchi, T., and Imoto, T. (1992) *Biochemistry 31*, 5545-5553.

43. Muraki, M., Goda, S., Nagahora, H., and Harata, K. (1997) *Protein Sci. 6*, 473-476.

44. Martin, S. F., and Hergenrother, P. J. (1998) *Biochemistry 37*, 5755-5760.

45. Stroud, R. M. (1996) *Nat. Struct. Biol. 3*, 567-569.

46. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res. 22*, 4673-4680.

INTRODUCTION TO CHAPTER 3

Chapter 3 takes the form of a manuscript published in *Protein Expression and Purification*. The focus of this work is on the marked improvement of the expression and purification of the creatine kinase isoform from *Torpedo californica* (TcCK). My advisor, Patsy Babbitt, initially subcloned, expressed and purified TcCK as a portion of her thesis work (Babbitt, et al., 1988). Unfortunately, the gene product expressed at low levels, never expressed solubly and retained only very low activity after refolding. Initially, we wanted to perform mutagenesis studies on the "essential" cysteine in TcCK, and therefore we set out to increase the solubility and activity of this enzyme.

I made significant contributions to this manuscript. The TcCK gene was first subcloned into the pET17b expression vector. Although this vector had been successfully used in the lab to solubly express the human muscle and brain isoforms of creatine kinase, TcCK still expressed in inclusion bodies. However, there was a large improvement in the protein expression level. I tested several refolding and purification protocols, and assayed the refolded enzyme. I initially had nucleation problems causing the refolded protein to slowly precipitate. John Cantwell suggested running the refolded protein over a Blue Sepharose column, which he tested, eliminating the precipitation problem. The TcCK project ended at this point in our lab. There was some difficulty obtaining the mutants we wished to examine, and my thesis project had become well focused on substrate specificity.

Michael McLeish and my advisor, Patsy Babbitt, hypothesized that this CK isoform may be able to yield an X-ray crystal structure with a bound transition state analog. This was reasoned from the fact that the only phosphagen kinase crystallized in this manner, arginine kinase, was purified from inclusion bodies. Thus we sent the pETTcCK clone and our purification protocols to George Kenyon's lab at the University of Michigan, where Michael McLeish and Pan-Fen Wang fine-tuned the expression and purification procedures.

This work resulted in large increases in both the amount of purified protein and TcCK activity. Purified TcCK was increased from less than 1 mg/L in the original expression system to 54 mg/L in the new system. The specific activity of TcCK was increased from 5.6 U/mg to 76.5 U/mg. Finally, this work resulted in crystals with a bound transition state analog complex. This structure was ultimately solved by Karen Allen's lab at Boston University (Lahiri, et al., 2002).

Babbitt, P. C., B. L. West, I. D. Kuntz and G. L. Kenyon (1988). "Purification of the Insoluble Aggregate Obtained from the Expression of Creatine Kinase in E. coli Yields Greatly Improved Specific Activity in the Refolded Protein." *Biochemistry* **27**: 3093.

Lahiri, S. D., P. F. Wang, P. C. Babbitt, M. J. McLeish, G. L. Kenyon and K. N. Allen (2002). "The 2.1 A structure of Torpedo californica creatine kinase complexed with the ADP-Mg(2+)-NO(3)(-)-creatine transition-state analogue complex." *Biochemistry* **41**(47): 13861-7.

CHAPTER 3

Expression of Torpedo californica creatine kinase in Escherichia coli and purification from inclusion bodies

Pan-Fen Wang[†], Walter R.P. Novak[§], John S. Cantwell[§1], Patricia C. Babbitt[§], Michael J. McLeish[†*] and George L. Kenyon[†]

Reproduced with permission from *Protein Expression and Purification*, 2002, Vol. 26, pp. 89-95. Copyright 2002 Elsevier.

[†] *College of Pharmacy, University of Michigan, 428 Church St., Ann Arbor, MI 48109-1065, USA*

[§] *Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, CA 94143-0446, USA*

Received 7 March 2002; revised 30 April 2002. Available online 18 September 2002.

[*] Corresponding author. Fax: 1-734-615-3079; email: mcleish@umich.edu

[1] Present address: Arriva Pharmaceuticals Inc., 2430-B Mariner Square Loop Alameda, CA 94501.

ABSTRACT

The pET17 expression vector was used to express creatine kinase from the electric organ of *Torpedo californica* as inclusion bodies in *Escherichia coli* BL21(DE3) cells. The insoluble aggregate was dissolved in 8 M urea and, following extraction with Triton X-100, the enzyme was refolded by dialysis against Tris buffer (pH 8.0) containing 0.2 M NaCl. After two buffer changes, chromatography on Blue Sepharose was used as a final step in the purification procedure. Approximately 54 mg active protein was recovered from a 1 L culture and the refolded enzyme had a specific activity of 75 U/mg. The molecular mass of the purified protein was consistent with that predicted from the amino acid sequence and the CD spectrum of the refolded enzyme was essentially identical to that of creatine kinase from human muscle (HMCK). The K_m values of ATP and ADP were also similar to those of HMCK, while the K_m values for both phosphocreatine and creatine were approximately 5–10-fold higher. The purification described here is in marked contrast with earlier attempts at purification of this isozyme where, in a process yielding less than 1 mg/L culture, enzyme with a specific activity of ca. 5 U/mg was obtained.

INTRODUCTION

Creatine kinase (CK, EC 2.7.3.2) is found in all vertebrates and catalyzes the reversible phosphorylation of creatine (Cr). The product, phosphocreatine (PCr), is considered to be a reservoir of "high-energy phosphate," which is able to supply ATP, the primary energy source in bioenergetics, on demand. As a consequence, creatine kinase plays a significant role in energy homeostasis of cells, particularly those requiring short, rapid burst of energy, such as neurons and both skeletal and cardiac muscles (*1*, *2*). The major isozymes of CK are named for the tissues and organelles from which they were originally isolated. These include the M-type, isolated from muscle (*3*), the B-type, isolated from brain (*4*), and the mitochondrial (Mt) isozymes (*5*). There is, in addition, a heterodimeric MB isozyme and elevated levels of this isozyme are used as a marker for myocardial infarction (*6*). In fact, changes in cellular CK concentrations or activities have been linked to a wide range of disease states including cancer (*7*, *8*), Alzheimer's disease (*9*), and muscular dystrophies (*10*).

Commensurate with the cellular importance of creatine kinase, considerable effort has been expended in trying to understand the structural and functional aspects of the CK mechanism. In addition to the early mechanistic studies generally carried out on the rabbit muscle isozyme (*11*), cDNAs encoding creatine kinase from a wide range of species have been cloned and sequenced. Considerable homology has been observed, especially within the individual isozyme classes (*12*, *13*). Each cDNA encodes a protein of about 40 kDa and X-ray structures are now available for all three major isozymes (*5*, *14*, *15*). Unfortunately, there are no structures of CK bound to either substrates, products or analog and, consequently, much of the detail of the CK molecular mechanism is yet to be elucidated. The structure of arginine kinase (AK), a guanidino kinase with ~40% sequence identity similar to those of most creatine kinases, complexed with transition-state analog, arginine–nitrate–ADP, has recently been solved (*16*). This structure shows that there is considerable movement of residues following substrate binding, notably those of a flexible active-site loop (*16*, *17*). In CK, too, this flexible loop has been associated with a conformational change upon substrate binding (*18*) and it appears that this loop may account for some of the different substrate specificities between CK and AK (*19*). These structural differences emphasize the prominent part a liganded structure of CK will play in ultimately understanding the role of individual residues in the mechanism of creatine kinase.

We have recently overexpressed, in soluble form, both the human muscle and human brain CK isozymes (*20*). Attempts have been made to crystallize the two proteins under a variety of conditions, both with and without ligand, again, with little or no success. This is possibly due to microheterogeneity problems similar to those observed previously for recombinant rabbit muscle CK (*21*). As with the rabbit muscle enzyme, both human muscle and human brain CK show significant heterogeneity on isoelectric focusing gels (*20*). It has been suggested that the problem lies in the purification (*21*), although deamidation cannot be ruled out (*22*). Accordingly, we have looked for an alternative isoform of CK, one that requires a significantly different purification scheme, that may provide a more homogeneous product, and, ultimately, a crystalline CK-transition state analog complex.

It was noted that the enzyme used to obtain the X-ray structure of the transition state analog complex (TSAC) of arginine kinase was purified from inclusion bodies (*26*). Several years ago, the DNA sequence (GenBank M36427) of creatine kinase from the electric organ of *Torpedo californica* (TcCK) was reported (*23*). CK is prevalent in this tissue, presumably as a result of the high-energy requirements of the electric discharge. The sequence showed considerable similarity (nearly 90%) to the sequences of both rabbit and human muscles. Unfortunately, attempts to express TcCK in *Escherichia coli* led to an inactive, insoluble aggregate and, while it was possible to refold the protein and restore activity, the yield of active TcCK was considerably less than 1 mg/L culture (*24*, *25*). However, the possibility that TcCK, purified from inclusion bodies, may also provide a crystalline TSAC suggested that another attempt be made to obtain this isozyme. Here, we describe an improved procedure for both expressing and purifying *T. californica* creatine kinase, one that permits the isolation of tens of milligrams of active enzyme per liter of culture, and which has allowed us to carry out large-scale crystallization studies.

MATERIALS AND METHODS

Creatine, phosphocreatine, ATP, ADP, NADH, NADP, and phosphoenolpyruvate were purchased from Sigma Chemical. The coupling enzymes, pyruvate kinase, lactic dehydrogenase, hexokinase, and glucose-6-phosphate dehydrogenase, were also purchased from Sigma. 3-(1-Pyridino)-1-propanesulfonate (NDSB-201) was purchased from Calbiochem. Human muscle creatine kinase (HMCK) was available from a previous study (*27*). Buffer salts and other reagents were of the highest quality available. The following buffer solutions were used in the protein preparation.

Buffer A: 25 mM Tris, 0.1 mM PMSF, 5 mM EDTA, and 50 mM NaCl, pH 7.5.

Buffer B: Buffer A containing 8 M urea and 50 mM DTT.

Buffer C: 25 mM Tris, 20 mM NaCl, 1 mM DTT, and pH 7.0.

Buffer D: 25 mM Tris, 0.2 M NaCl, 1 mM DTT, 1 M NDSB-201, and pH 7.0.

Construction of a TcCK expression vector. Plasmid pKTCK3F, containing the TcCK gene, was available from a previous study (*25*). To obtain better yields of protein, we placed the TcCK gene into pET17b (Novagen), which uses the strong T7 promoter for high-level expression. This was achieved by engineering a *Sal*I restriction site into pKTCK3F, downstream of the C-terminus of TcCK. The mutagenesis was carried out using the QuikChange mutagenesis kit (Stratagene) employing the following primers (the *Sal*I site is underlined):

5'-GCCTTGAAATATCACAGAACTTtGAACTTTCCC-3';

5'-GGGAAAGTTCAAAGTTCTGTGATATTTCAAGGC-3'.

Following mutagenesis, the template DNA was removed by treatment with *Dpn*I and the PCR products were transformed into *E. coli* strain DH5 (Gibco-BRL). Single colonies were picked and their isolated DNA was screened for the presence of the new *Sal*I site. The presence of the *Sal*I site and the fidelity of the PCR amplification of the cloned TcCK gene were confirmed by sequencing. The mutated pKTCK3F vector, now containing both *Nde*I and *Sal*I sites, was digested with these enzymes and the fragment containing the TcCK gene was purified from an agarose gel using the Geneclean Spin kit (Bio101). The purified fragment was ligated into the vector pET17b, which had been digested with both *Nde*I and *Sal*I. The resulting construct, designated pETTcCK, was then transformed into *E. coli* BL21(DE3)pLysS (Stratagene) for protein expression.

Expression of TcCK and preparation of inclusion bodies. A 50 mL culture of BL21(DE3)pLysS, freshly transformed with pETTcCK, was used to inoculate 1 L LB media containing ampicillin and chloramphenicol. The cells were grown at 37 °C until OD_{600}

reached 1.0. Protein expression was then induced with 0.5 mM IPTG. After growth for an additional 4 h at 30 °C, the cells were harvested by centrifugation at 6300*g* for 8 min at 4 °C. The cell pellet was resuspended in 25 mL Buffer A and the cells were lysed by treatment of the cell suspension with lysozyme (0.3 mg/mL) for 30 min at room temperature. Triton X-100 was added to 5% (v/v) and the suspension was sonicated, on ice, four times for 1 min each. The suspension was then treated with DnaseI (110 U/mL) and RnaseA (3 U/mL) for 15 min at room temperature. The inclusion bodies were sedimented by centrifugation at 30,000*g* for 20 min at 4 °C. To remove additional lipids and potential proteolytic activity, the inclusion bodies were washed with 25 mL buffer A containing 5% Triton X-100, followed by centrifugation at 30,000*g* for 20 min at 4 °C. The excess Triton X-100 was removed by washing twice with 25 mL buffer A, followed by centrifugation at 30,000*g* for 20 min at 4 °C.

Denaturation, refolding, and purification of TcCK. The pellet containing inclusion bodies from 0.5 L cell culture was resuspended in 25 mL of Buffer A and solid urea and DTT were added to final concentrations of 8 M and 50 mM, respectively. The solution was incubated at 4 °C for 1 h. The insoluble material was removed by centrifugation at 100,000*g* for 15 min at 4 °C. At this point, the protein concentration in the supernatant was estimated using the Bradford assay (*28*). Buffer B was used to dilute the supernatant to a protein concentration of ~2 mg/mL. The solution of denatured TcCK was dialyzed at 4 °C overnight against 25 mM Tris buffer, pH 8.0, containing 0.2 M NaCl, 5 mM DTT, and 1 mM EDTA. The precipitate that formed during the dialysis was removed by centrifugation at 3500*g* for 30 min at 4 °C. The supernatant was again dialyzed overnight at 4 °C, this time against 25 mM Tris buffer containing 50 mM NaCl, 1 mM DTT and 1 mM EDTA at pH 8.0. The solution was once more clarified by centrifugation at 3500*g* for 30 min and the refolded protein was subjected to a final overnight dialysis step against 25 mM Tris, pH 7.0, containing 20 mM NaCl, 1 mM DTT, and 1 mM EDTA. The dialyzed protein was then loaded onto a 2.6 × 14.5 cm Blue Sepharose CL-6B column (Pharmacia Biotech), which had been preequilibrated in Buffer C. Unbound material was removed by washing the column with two column volumes of Buffer C and TcCK was eluted using a linear gradient of 20–500 mM NaCl over two column volumes. The peak of TcCK was centered at around 140 mM NaCl. The fractions containing TcCK were pooled and dialyzed against Buffer C containing 1 mM EDTA at 4 °C overnight, before being concentrated using an Amicon Centricon concentrator unit. The purified enzyme was stored in aliquots at -80 °C. The enzyme concentration was determined using an A280 of 0.88 for a 1 mg/mL sample (*3*). The final yield of active TcCK was ca. 54 mg per liter of cell culture.

Refolding of denatured protein using nondetergent sulfobetain (NDSB). A sample (3 mg/mL) of denatured TcCK, prepared as described above, was diluted with vigorous stirring into cold folding buffer (Buffer D). Following the addition of the denatured protein, stirring was continued for an additional 2 min at 4 °C and the solution was left undisturbed at 4 °C for 1 h. The solution was filtered to remove the protein that precipitated during the refolding process and purified by chromatography on Blue Sepharose as described above.

Isoelectric focusing electrophoresis. Isoelectric focusing (IEF) of TcCK was performed on PhastGel IEF 3–9 using a PhastSystem (Pharmacia Biotech). The gel, which covers a p*I* range of 3–9, was prefocused at 2000 V for 10 min. TcCK was applied to the gel at 200 V for 5 min. After the sample had been applied, focusing was continued at 2000 V for

30 min, with the temperature maintained at 15 °C. The gel was fixed in 20% trichloroacetic acid and stained with PhastGel Blue R (Coomassie R 350).

Mass spectrometry. Electrospray mass spectrometry was performed on a Finnigan LCQ mass spectrometer interfaced with an analytical HPLC (Hitachi) equipped with a Vydac C18 reversed phase column (4.6 × 250 mm^2). The mobile phase was composed of 0.1% of acetic acid, 0.02% of TFA, and a linear gradient of 5–75% acetonitrile. TcCK eluted at about 60% MeCN.

Circular dichroism spectropolarimetry. CD spectra were recorded on a Jasco J-810 spectropolarimeter calibrated with *d*-10-camphorsulfonic acid. The TcCK and HMCK samples were both at a concentration of 0.3 mg/mL, in 10 mM potassium phosphate, pH 7.0. The spectra were an average of 5 scans recorded at 20 °C with a bandwidth of 1 nm, a 0.1 nm step size, and a 1 s time constant, and were baseline corrected.

Determination of creatine kinase activity. Standard assays of creatine kinase activity were carried out in the direction of creatine phosphorylation using a pH-stat method (*29*), with ATP and creatine concentrations of 5 and 100 mM, respectively. Steady-state kinetic analyses of TcCK were carried out in both forward and reverse directions at 30 °C, using coupled assays, as described previously (*27*, *30*). In the forward direction, creatine phosphorylation is coupled to the reactions of pyruvate kinase, and lactate dehydrogenase, and followed by monitoring the decrease in absorbance at 340 nm due to bleaching of NADH. The assay mixture contained 75 mM TAPS buffer (pH 9.0), 0.36 mM NADH, 0.36 mM phosphoenolpyruvate, 1 mM magnesium acetate, 13 mM potassium acetate, variable MgATP (0.3–5 mM), variable creatine (6–100 mM), and 9.3 nM TcCK. The concentrations of the coupling enzymes, pyruvate kinase and lactate dehydrogenase are 28 and 54 U/mL, respectively. In the reverse direction, ADP phosphorylation is coupled to the reactions of hexokinase, and glucose-6-phosphate dehydrogenase, and followed by monitoring the increase in absorbance at 340 nm due to reduction of NADP. The assay mixture contained 75 mM HEPES buffer (pH 7.0), 5 mM glucose, 1 mM NADP, 5 mM magnesium acetate, variable MgADP (10–200 M), variable phosphocreatine (1–15 mM), and 2.4 nM TcCK. The concentrations of the coupling enzymes, hexokinase and glucose-6-phosphate dehydrogenase, are 4.1 and 8.2 U/mL, respectively.

RESULTS AND DISCUSSION

As shown in Fig. 1 (lanes 2 and 3), the pETTcCK vector gave excellent inducible expression of TcCK. It provided a considerably greater expression of TcCK than the earlier pKTC3F (data not shown) but, again, the enzyme was recovered as an insoluble aggregate. Little or no CK activity was detected in the soluble extracts nor was any found in the inclusion bodies. It had been reported that proteolysis could be a problem during solubilization and refolding of TcCK, but the proteolytic activity could be removed by extraction with a buffer containing Triton X-100 (*24*). Accordingly Triton X-100 was added to the cell lysis buffer and, in addition, the pelleted inclusion bodies were also extracted with Triton X-100.

Solubilization of the aggregates was achieved under strongly denaturing conditions (8 M urea) in the presence of reducing agent (50 mM DTT). Earlier attempts to refold TcCK used rapid dilution into Tris buffer (*24*). In this study, attempts were made to refold TcCK in two ways: by stepwise dialysis and by dilution into a folding buffer containing the mild solubilizing/stabilizing agent NDSB-201 (*31*). While, in milligrams of protein, the yield of active enzyme was broadly comparable for both methods, the specific activity of the enzyme refolded using the direct dialysis method (76.5 U/mg) was considerably higher than of the enzyme refolded by the dilution technique (10.3 U/mg).

The final purification step of the refolded TcCK is a modification of the method used for HMCK (*20*). As can be seen from Fig. 1 (lanes 4 and 5), neither the solubilized inclusion bodies nor the refolded enzyme contained many contaminants and it proved possible to complete the purification of the refolded TcCK in a single step using chromatography over Blue Sepharose column (Fig. 1, lane 6). Blue Sepharose is commonly used to purify CK isozymes (*20*, *32*) and, an additional anion exchange chromatography step, essential for the purification of HMCK, did not lead to any improvement in specific activity of TcCK. Further, it was noted that the TcCK refolded by dilution into NDSB-201 did not bind well to the Blue Sepharose column. This observation, combined with its relatively low specific activity, suggested that TcCK refolded by rapid dilution may not fully attain its native conformation. The dialysis method is the more tedious of the two, requiring three overnight steps before the chromatography step. However, given the significantly higher specific activity of the refolded enzyme, and the fact that NDSB-201 is not inexpensive, particularly for large-scale preparations, the dialysis method of refolding emerged as the method of choice.

Table 1 shows that this purification results in an excellent yield of highly active TcCK. Once refolded, neither dialysis nor chromatography results in much loss of CK activity. The overall purification, 1.7-fold, emphasizes that expression in inclusion bodies can, under some circumstances, in itself provide an excellent initial purification. The specific

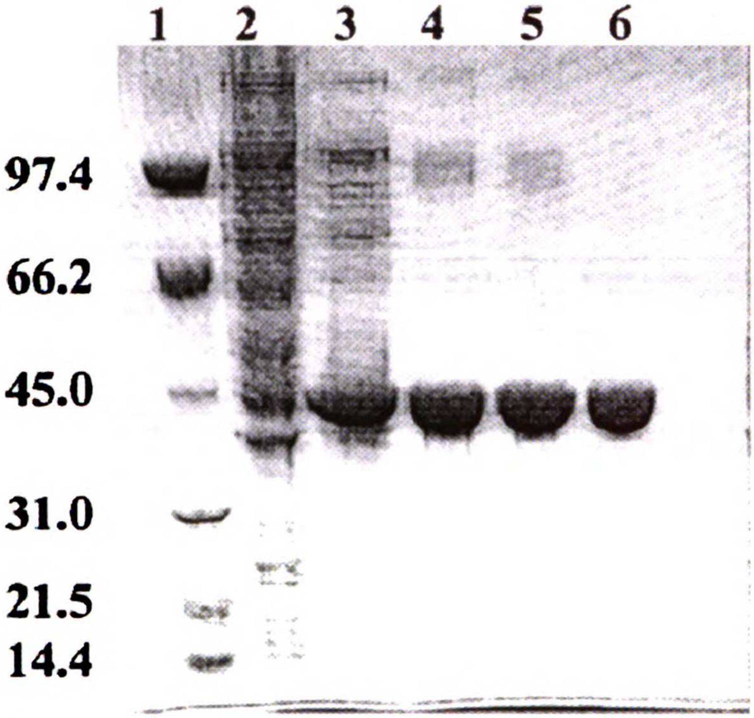


Figure 1. SDS–PAGE analysis of a typical expression and purification of TcCK. Lane 1, molecular weight markers; lane 2, uninduced Bl21(DE3)pLysS cells containing pETTcCK/; lane 3, whole cell lysate 3 h post-induction with 0.5 mM IPTG; lane 4, solubilized inclusion bodies; lane 5, TcCK after dialysis; lane 6, TcCK following Blue Sepharose chromatography.

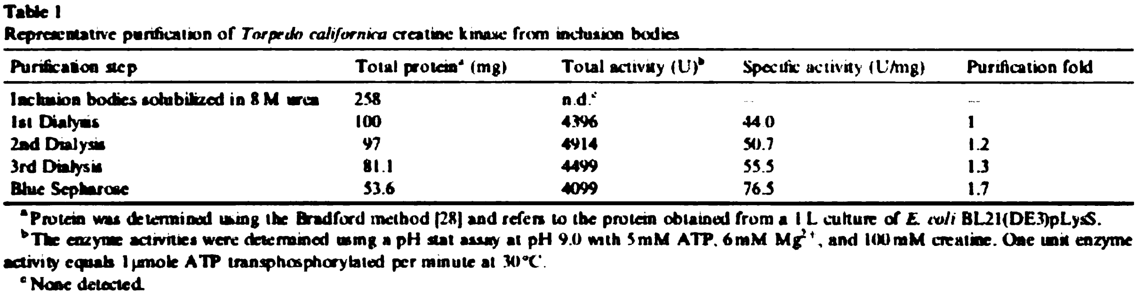


Table 1

Representative purification of *Torpedo californica* creatine kinase from inclusion bodies

Purification step	Total protein[a] (mg)	Total activity (U)[b]	Specific activity (U/mg)	Purification fold
Inclusion bodies solubilized in 8 M urea	258	n.d.[c]	--	--
1st Dialysis	100	4396	44.0	1
2nd Dialysis	97	4914	50.7	1.2
3rd Dialysis	81.1	4499	55.5	1.3
Blue Sepharose	53.6	4099	76.5	1.7

[a] Protein was determined using the Bradford method [28] and refers to the protein obtained from a 1 L culture of *E. coli* BL21(DE3)pLysS.

[b] The enzyme activities were determined using a pH stat assay at pH 9.0 with 5 mM ATP, 6 mM Mg^{2+}, and 100 mM creatine. One unit enzyme activity equals 1 µmole ATP transphosphorylated per minute at 30°C.

[c] None detected.

Table 1. Representative purification of *Torpedo californica* creatine kinase from inclusion bodies

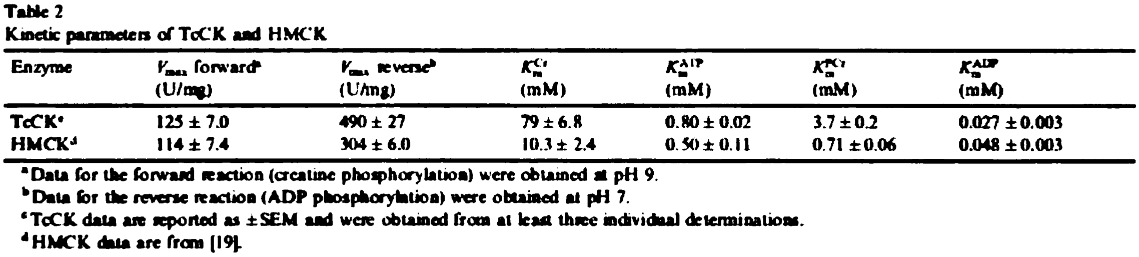


Table 2

Kinetic parameters of TcCK and HMCK

Enzyme	V_{max} forward[a] (U/mg)	V_{max} reverse[b] (U/mg)	K_m^{Cr} (mM)	K_m^{ATP} (mM)	K_m^{PCr} (mM)	K_m^{ADP} (mM)
TcCK[c]	125 ± 7.0	490 ± 27	79 ± 6.8	0.80 ± 0.02	3.7 ± 0.2	0.027 ± 0.003
HMCK[d]	114 ± 7.4	304 ± 6.0	10.3 ± 2.4	0.50 ± 0.11	0.71 ± 0.06	0.048 ± 0.003

[a] Data for the forward reaction (creatine phosphorylation) were obtained at pH 9.

[b] Data for the reverse reaction (ADP phosphorylation) were obtained at pH 7.

[c] TcCK data are reported as ±SEM and were obtained from at least three individual determinations.

[d] HMCK data are from [19].

Table 2. Kinetic parameters of TcCK and HMCK

activity of the purified TcCK, 76.5 U/mg, lies between that of human muscle CK (210 U/mg) (*27*) and rabbit muscle (33.5 U/mg) (*21*), and is comparable to that of chicken sarcomeric CK (73 U/mg) (*33*) and human brain CK (88 U/mg) (*34*). In toto, both the yield, 54 mg/L, and the specific activity of the TcCK prepared here are a considerable improvement over those obtained in previous attempts to express TcCK in *E. coli*, <1 mg/L and 5.6 U/mg, respectively (*24*).

The amino acid sequence of TcCK (*23*) translates to a protein with a predicted subunit MW of 42,927 Da. Electrospray mass spectrometry indicated the molecular mass of purified TcCK to be 42797±3 Da, which is in agreement with the recombinant enzyme having its N-terminal methionine residue removed. The removal of the N-terminal methionine residue has been observed previously for both the human muscle and human brain isozymes when expressed in *E. coli* (*20*).

The circular dichroic spectrum of TcCK is essentially identical to that of human muscle CK (Fig. 2). This is not surprising given the high level of sequence similarity, but it does serve to confirm that the refolded enzyme has structural as well as kinetic integrity.

The steady-state kinetics of recombinant TcCK have been studied in both forward and reverse directions. The kinetic parameters are listed in Table 2 and contrasted with comparable data for recombinant HMCK. In the forward direction, at pH 9.0, and in the reverse direction, at pH 7.0, both isozymes exhibit a random binding mechanism. Both isozymes have similar K_m values for ADP and ATP, but TcCK has much higher K_m values for both phosphocreatine and creatine. The high K_m values for the latter two substrates require that the V_{max} data be obtained by extrapolation to saturating substrate levels. The V_{max} data show minor differences between the two isozymes, but these are not dissimilar to

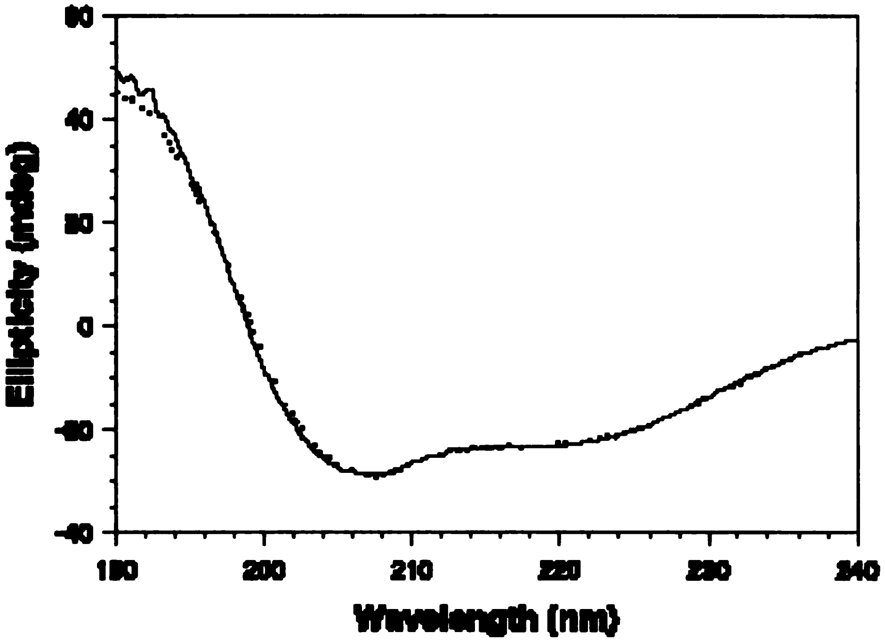


Figure 2. CD spectra of TcCK and HMCK The spectra of TcCK (—) and human muscle CK (···) were obtained at 20 °C, at a protein concentration of 0.3 mg/mL in KPO4, pH 7.0.

differences observed between, for example, the human muscle and brain isozymes (*20*). However, one observation of note is that TcCK shows a dramatic decrease in activity when it is subjected to repeated freezing/thawing. The activity can be recovered by treatment with DTT (50 mM DTT in 50 mM HEPES buffer, pH 8) but it is not an immediate recovery and overnight incubation is required. This loss of activity is not observed with HMCK. Alignment of the sequences of HMCK and TcCK shows that the latter has three additional cysteine residues and it appears reasonable to suggest that the reversible loss of activity may be due to oxidation of the additional cysteine(s).

In addition to obtaining improved yields of TcCK, one of the initial aims of this work was to obtain a homogeneous preparation of a creatine kinase isozyme that would be amenable to crystallization. Highly purified rabbit muscle CK (*21*), as well as human muscle and human brain CK, display multiple isoforms on an IEF gel (*20*). It was suggested that the heterogeneity may arise as an artifact of the purification (*21*) or more likely may be due to post-translational amidation/deamidation (*22*). In this study, we found that TcCK showed only one major band on an IEF gel (pH 3–9). However, in contrast to sharply focused bands of HMCK (p*I* 6.9–7.8), the TcCK band appeared to be quite diffuse with a p*I* between 4.5 and 5.5 (not shown). A closer examination showed that the diffuse band was really a collection of several closely spaced finer bands, indicating that TcCK also displays the microheterogeneity common to many CK isozymes. Happily, this problem was not sufficient to prevent crystal formation in the crystallization trials that will be described elsewhere.

ACKNOWLEDGEMENTS

This work was supported by USPHS Grant AR17323 (to G.L.K.). Thanks to Brian DeSmet for carrying out some of the kinetic assays.

REFERENCES

1. T. Wallimann, M. Wyss, D. Brdiczka, K. Nicolay and H.M. Eppenberger, Intracellular compartmentation, structure and function of creatine kinase isoenzymes in tissues with high and fluctuating energy demands: the `phosphocreatine circuit' for cellular energy homeostasis. *Biochem. J.* **281** (1992), pp. 21–40.
2. M. Wyss and R. Kaddurah-Daouk, Creatine and creatinine metabolism. *Physiol. Rev.* **80** (2000), pp. 1107–1213.
3. S.A. Kuby, L. Noda and H.A. Lardy, Adenosinetriphosphate-creatine transphosphorylase: I. Isolation of the crystalline enzyme from rabbit muscle. *J. Biol. Chem.* **209** (1954), pp. 191–201.
4. D.M. Dawson, H.M. Eppenberger and N.O. Kaplan, The comparative enzymology of creatine kinases: II. Physical and chemical properties. *J. Biol. Chem.* **242** (1967), pp. 210–217.
5. M. Eder, K. Fritz-Wolf, W. Kabsch, T. Wallimann and U. Schlattner, Crystal structure of human ubiquitous mitochondrial creatine kinase. *Proteins* **39** (2000), pp. 216–225.
6. A.H. Wu, Creatine kinase isoforms in ischemic heart disease. *Clin. Chem.* **35** (1989), pp. 7–13.

7. N.M. Hoosein, K.J. Martin, M. Abdul, C.J. Logothetis and R. Kaddurah-Daouk, Antiproliferative effects of cyclocreatine on human prostatic carcinoma cells. *Anticancer Res.* **15** (1995), pp. 1339–1342.

8. N. Zarghami, M. Giai, H. Yu, R. Roagna, R. Ponzone, D. Katsaros, P. Sismondi and E.P. Diamandis, Creatine kinase BB isoenzyme levels in tumour cytosols and survival of breast cancer patients. *Br. J. Cancer* **73** (1996), pp. 386–390.

9. S. David, M. Shoemaker and B.E. Haley, Abnormal properties of creatine kinase in Alzheimer's disease brain: correlation of reduced enzyme activity and active site photolabeling with aberrant cytosol-membrane partitioning. *Brain Res. Mol. Brain Res.* **54** (1998), pp. 276–287.

10. E. Ozawa, Y. Hagiwara and M. Yoshida, Creatine kinase, cell membrane and duchenne muscular dystrophy. *Mol. Cell. Biochem.* **190** (1999), pp. 143–151.

11. G.L. Kenyon and G.H. Reed, Creatine kinase: structure–activity relationships. *Adv. Enzymol. Relat. Areas Mol. Biol.* **54** (1983), pp. 367–426.

12. P.C. Babbitt, G.L. Kenyon, I.D. Kuntz, F.E. Cohen, J.D. Baxter, P.A. Benfield, J.D. Buskin, W.A. Gilbert, S.D. Hauschka, J.P. Hossle, C.P. Ordahl, M.L. Pearson, J.-C. Perriard, L.A. Pickering, S.D. Putney, B.L. West and R.A. Ziven, Comparisons of creatine kinase primary structures. *J. Prot. Chem.* **5** (1986), pp. 1–14.

13. S.M. Muhlebach, M. Gross, T. Wirz, T. Wallimann, J.C. Perriard and M. Wyss, Sequence homology and structure predictions of the creatine kinase isoenzymes. *Mol. Cell. Biochem.* **133/134** (1994), pp. 245–262.

14. J.K. Rao, G. Bujacz and A. Wlodawer, Crystal structure of rabbit muscle creatine kinase. *FEBS Lett.* **439** (1998), pp. 133–137.

15. M. Eder, U. Schlattner, A. Becker, T. Wallimann, W. Kabsch and K. Fritz-Wolf, Crystal structure of brain-type creatine kinase at 1.41 Å resolution. *Protein Sci.* **8** (1999), pp. 2258–2269.

16. G. Zhou, T. Somasundaram, E. Blanc, G. Parthasarathy, W.R. Ellington and M.S. Chapman, Transition state structure of arginine kinase: implications for catalysis of bimolecular reactions. *Proc. Natl. Acad. Sci. USA* **95** (1998), pp. 8449–8454.

17. G. Zhou, W.R. Ellington and M.S. Chapman, Induced fit in arginine kinase. *Biophys. J.* **78** (2000), pp. 1541–1550.

18. M. Forstner, M. Kriechbaum, P. Laggner and T. Wallimann, Structural changes of creatine kinase upon substrate binding. *Biophys. J.* **75** (1998), pp. 1016–1023.

19. J.S. Cantwell, W.R. Novak, P.F. Wang, M.J. McLeish, G.L. Kenyon and P.C. Babbitt, Mutagenesis of two acidic active site residues in human muscle creatine kinase: implications for the catalytic mechanism. *Biochemistry* **40** (2001), pp. 3056–3061.

20. L.H. Chen, C.B. White, P.C. Babbitt, M.J. McLeish and G.L. Kenyon, A comparative study of human muscle and brain creatine kinases expressed in *Escherichia coli*. *J. Protein Chem.* **19** (2000), pp. 59–66.

21. L.H. Chen, P.C. Babbitt, J.R. Vasquez, B.L. West and G.L. Kenyon, Cloning and expression of functional rabbit muscle creatine kinase in *Escherichia coli*. Addressing the problem of microheterogeneity. *J. Biol. Chem.* **266** (1991), pp. 12053–12057.

22. T.D. Wood, L.H. Chen, N.L. Kelleher, D.P. Little, G.L. Kenyon and F.W. McLafferty, Direct sequence data from heterogeneous creatine kinase (43 kDa) by high-resolution tandem mass spectrometry. *Biochemistry* **34** (1995), pp. 16251–16254.

23. B.L. West, P.C. Babbitt, B. Mendez and J.D. Baxter, Creatine kinase protein sequence encoded by a cdna made from *Torpedo californica* electric organ mRNA. *Proc. Natl. Acad. Sci. USA* **81** (1984), pp. 7007–7011.

24. P.C. Babbitt, B.L. West, D.D. Buechter, I.D. Kuntz and G.L. Kenyon, Removal of a proteolytic activity associated with aggregates formed from expression of creatine kinase in *Escherichia coli* leads to improved recovery of active enzyme. *Biotechnology (New York)* **8** (1990), pp. 945–949.

25. P.C. Babbitt, B.L. West, D.D. Buechter, L.H. Chen, I.D. Kuntz and G.L. Kenyon, Active creatine kinase refolded from inclusion bodies in *Escherichia coli*. In: E. DeBernadez and G. Georgiou, Editors, *Protein Refolding*, American Chemical Society, New York (1991), pp. 153–168.

26. G. Zhou, G. Parthasarathy, T. Somasundaram, A. Ables, L. Roy, S.J. Strong, W.R. Ellington and M.S. Chapman, Expression, purification from inclusion bodies, and crystal characterization of a transition state analog complex of arginine kinase: a model for studying phosphagen kinases. *Protein Sci.* **6** (1997), pp. 444–449.

27. P.F. Wang, M.J. McLeish, M.M. Kneen, G. Lee and G.L. Kenyon, An unusually low p*K*a for cys282 in the active site of human muscle creatine kinase. *Biochemistry* **40** (2001), pp. 11698–11705.

28. M.M. Bradford, A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Chem.* **72** (1976), pp. 248–254.

29. T.A. Mahowald, E.A. Noltmann and S.A. Kuby, Studies on adenosine triphosphate transphosphorylases: III. Inhibition reactions. *J. Biol. Chem.* **237** (1962), pp. 1535–1548.

30. P.F. Cook, G.L. Kenyon and W.W. Cleland, Use of pH studies to elucidate the catalytic mechanism of rabbit muscle creatine kinase. *Biochemistry* **20** (1981), pp. 1204–1210.

31. L. Vuillard, C. Braun-Breton and T. Rabilloud, Non-detergent sulphobetaines: a new class of mild solubilization agents for protein purification. *Biochem. J.* **305** Pt 1 (1995), pp. 337–343.

32. O. Marcillat, C. Perraut, T. Granjon, C. Vial and M.J. Vacheron, Cloning, *Escherichia coli* expression, and phase-transition chromatography-based purification of recombinant rabbit heart mitochondrial creatine kinase. *Protein Expr. Purif.* **17** (1999), pp. 163–168 doi:10.1006/prep.2001.1511 .

33. R. Furter, P. Kaldis, E. Furter-Graves, T. Schnyder, H.M. Eppenberger and T. Wallimann, Expression of active octameric chicken cardiac mitochondrial creatine kinase in *Escherichia coli. Biochem. J.* **288** (1992), pp. 771–775.

34. C.B. White, Human brain creatine kinase, Ph.D. Dissertation. University of California, San Francisco, 1996.

INTRODUCTION TO CHAPTER 4

Chapter 4 describes investigations into the substrate specificity of creatine kinase (CK). This manuscript was submitted for publication in *Biochemistry*. The phosphagen kinases catalyze the reversible phosphorylation of a guanidino substrate by ATP. Unlike the majority of protein folds, the phosphagen kinase fold is unique and only performs this one function. CK shares ~40% sequence identity with other phosphagen kinases such as arginine kinase (AK) and glycocyamine kinase (GK), yet these enzymes are highly specific for their individual substrates.

When I began this project, only AK was crystallized with a transition state analog complex (Zhou, et al., 1998). Therefore computational comparisons between AK and CK were necessary to understand how active site differences relate to the delivery of substrate specificity. These studies identified class-specific differences in two flexible loop regions, one N-terminal (N-loop) and one C-terminal (C-loop). The N-loop primarily contacts the carboxy terminus of the substrate, while the C-loop makes substrate contacts near the guanidino terminus. My initial studies were concerned with complete swapping of the N-terminal and C-terminal loops of CK and GK (whose substrates differ only by a single *N*-methyl group) in a stepwise manner. After much progress was made towards the development of these mutants it became apparent that simply swapping these loops was not altering the specificity of the CK mutants in such a manner as to have high activity with glycocyamine.

Almost immediately after Patsy and I decided to focus on V325 (homologous to AK E314) as a specificity determinant, we heard from Michael McLeish that a CK transition state analog complexed structure had been solved of TcCK (see Chapter 3) (Lahiri et al., 2002). Examination of this structure verified that V325 made contacts with the *N*-methyl group, and also implicated I69. Computational studies failed to identify I69 as being in proximity to the ligand due to loop length differences between the AK and CK structures used in our comparative studies.

Investigations of the roles of I69 and V325 revealed that mutations at I69 were unable to direct the specificity of the enzyme away from its natural substrate, creatine. While specificity of the enzyme was affected, several mutations at this position affected the synergy of the enzyme. In contrast, both mutations we created at V325, V325A and V325E, resulted in the preference of CK for either cyclocreatine or glycocyamine, respectively. Thus, we have described a "specificity switch" at position V325, where mutations alter the substrate preference of the enzyme.

Zhou, G., T. Somasundaram, E. Blanc, G. Parthasarathy, W. R. Ellington and M. S. Chapman (1998). "Transition state structure of arginine kinase: implications for catalysis of bimolecular reactions." *Proc. Natl. Acad. Sci. U.S.A.* **95**(15): 8449-54.

Lahiri, S. D., P. F. Wang, P. C. Babbitt, M. J. McLeish, G. L. Kenyon and K. N. Allen (2002). "The 2.1 A structure of Torpedo californica creatine kinase complexed with the ADP-Mg(2+)-NO(3)(-)-creatine transition-state analogue complex." *Biochemistry* **41**(47): 13861-7.

CHAPTER 4

Isoleucine 69 and Valine 325 Form a Specificity Pocket in Human Muscle Creatine Kinase[†]

Walter R.P. Novak[‡], Pan-Fen Wang[§], Michael J. McLeish[§], George L. Kenyon[§] and Patricia C. Babbitt[*,‡]

Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry
University of California, San Francisco, 600 16th St., San Francisco, CA 94143
College of Pharmacy, University of Michigan, 428 Church St., Ann Arbor, MI 48109

[*] To whom correspondence should be addressed. Telephone: (415) 476-3784. Fax: (415) 514-4260. E-mail: babbitt@cgl.ucsf.edu.

[†] Supported by NIH grants GM60595 to PCB and AR17323 to GLK. WRPN is a Burroughs Wellcome predoctoral fellow.

[‡] University of California, [§] University of Michigan

Running Title: The Specificity Pocket of Creatine Kinase

Abbreviations: CK, creatine kinase; AK, arginine kinase; TES, 2-[(2-hydroxy-1,1-bis[hydroxymethyl]ethyl)amino]-ethanesulfonic acid; TAPS, 2-[(2-hydroxy-1,1-bis[hydroxymethyl]ethyl)amino]-1-propanesulfonic acid; DTT, dithiothreitol; PMSF, phenylmethanesulfonyl fluoride; IPTG, isopropyl-β-D-thiogalactopyranoside; SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis; CD, circular dichroism.

Footnotes:

[1] In several AKs, the SGV (residues 63-65) motif and D62 are not conserved. This sub-class of AKs presumably has a unique substrate recognition system (*31*).

[2] For clarity, *N*- refers to the N_γ position of creatine and the creatine analogs and the analogous N_ε of arginine, as applicable.

[3] An α value equal to unity indicates that the binding of the first substrate has no effect on the binding of the second. A value less than unity signifies an increased affinity for the second substrate, and a value greater than unity indicates a decrease in the affinity of the enzyme for the second substrate.

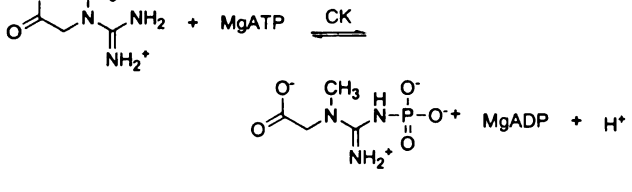
ABSTRACT

Creatine kinase (CK) catalyzes the reversible phosphorylation of creatine by ATP. From a structural perspective, the enzyme utilizes two flexible loop regions to sequester and position the substrates for catalysis. There has been debate over the specific roles of the flexible loops in substrate specificity and catalysis in CK and other related phosphagen kinases. In CK, two hydrophobic loop residues, I69 and V325, make contacts with the *N*-methyl group of creatine. In this study, we report the alteration of the substrate specificity of CK through the mutagenesis of V325. The V325 to glutamate mutation results in a more than 100-fold preference for glycocyamine while mutation of V325 to alanine results in the slight preference of the enzyme for cyclocreatine (1-carboxymethyl-2-iminoimidazolidine). This study enhances our understanding of how the active sites of phosphagen kinases have evolved to recognize their respective substrates and catalyze their reactions.

INTRODUCTION

Koshland originally suggested that conformational changes upon substrate binding could serve as a basis for substrate specificity, e.g., a good substrate induces an active conformation of an enzyme, and a poor substrate induces an inactive conformation (*1*). Frequently, these changes involve the movement of flexible loops which assist in substrate recognition and catalysis (*2-4*). While the contribution of induced fit to the substrate specificity of an enzyme has been greatly debated (*5-8*), it is clear that determinants of specificity can often be found on flexible loop regions (*4, 9, 10*). Here, we investigate the contributions of two flexible loop regions of creatine kinase (CK) to the substrate specificity of the enzyme.

CK (E.C. 2.7.3.2) is a member of the phosphagen kinase superfamily of enzymes whose members catalyze the reversible phosphorylation of guanidino substrates by ATP. The CK reaction is shown in Scheme 1. The products of these reactions, phosphagens, act as reservoirs of high-energy phosphate which are utilized to rapidly regenerate ATP in cells with variable energy needs (i.e., muscle and nerve cells). In addition to their central role in energy homeostasis, the phosphagen kinases have been used as a paradigm for understanding the kinetics and mechanisms of bimolecular reactions (*11-13*).



Scheme 1.

The phosphagen kinases possess a unique structural fold with only distant similarity to glutamine synthetase (*14, 15*). High resolution crystal structures of two phosphagen kinases, arginine kinase (AK) and CK, have been determined in the unbound form (*16-19*) and bound with a transition state analog complex (TSAC) (*20, 21*). These structures indicate that large movements of two flexible loop regions (termed here N-terminal and C-terminal loops) occur upon substrate binding. The N-terminal loop region (N-loop, CK residues 60-73) is variable in length across the phosphagen kinase superfamily. The length of this loop shows an inverse relationship to the size of the cognate substrate, suggesting that it may be involved in determining specificity. In some studies, the N-loop has been referred to as the guanidino specificity region (*20, 22*). The length of the C-terminal loop (C-loop, CK residues 323-332) is invariant across the phosphagen kinases, although the sequence motifs found in the loop are class (family) specific (e.g., there are differences between AKs and CKs, etc.). The gross movements of the N- and C-loops are similar in AK and CK but, consistent with the sequence differences among families, the specific interactions that occur between these two loops in the TSAC enzymes differ significantly. In the AK-TSAC structure, the N-loop is short and cannot make contacts with the C-loop. In contrast, the CK-TSAC structure reveals significant interactions between the N- and C-loops (*21*).

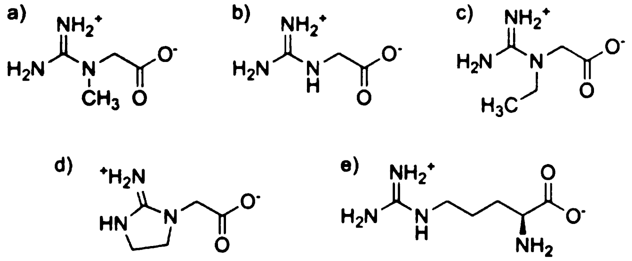


Figure 1. Natural (a, b, e) and synthetic (c, d) guanidino substrates. a) creatine b) glycocyamine c) *N*-ethylglycocyamine d) cyclocreatine e) arginine

Our examination of the structures of AK and CK suggests that CK residues I69 and V325 may act as a specificity determining system in that enzyme. Accordingly, the roles of each of these residues in substrate recognition and catalysis have been investigated using site-directed mutagenesis. Several mutants at I69 and V325 were constructed and assayed with creatine and creatine analogs (Figure 1a-d). Although the structural evidence suggests that both I69 and V325 make hydrophobic contacts with the substrate, our study found that single site mutants at I69 are unable to direct the specificity of CK away from its natural substrate. In contrast, results from single site mutations at V325 suggest that this residue can function as a "specificity switch." Changing the identity of the residue at this position alters the specificity of the enzyme among three guanidino substrates. In the context of these mutation studies, the roles of both residues in interactions with the creatine substrate for specificity and catalysis are discussed.

MATERIALS AND METHODS

Materials. Unless otherwise noted, all chemicals and enzymes were purchased from Sigma-Aldrich Chemical Company (St. Louis, MO). Microplate readings were taken on a SPECTRAmax 340 spectrophotometer from Molecular Devices (Sunnyvale, CA) using UV-transparent Costar 96 well plates from Corning (Corning, NY).

Synthesis of cyclocreatine and N-ethylglycocyamine. Cyclocreatine (1-carboxymethyl-2-iminoimidazolidine) and *N*-ethylglycocyamine (*N*-ethyl-*N*-amidinoglycine) were synthesized according to previously published methods (*23*).

Site-Directed Mutagenesis. Amino acid substitutions in the human muscle CK isozyme were carried out using the QuikChange Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA) with the vector pETHMCK (*24*) acting as the DNA template. Double mutants were created either by performing a second mutagenesis step on the appropriate mutant template, or by digesting the appropriate mutant plasmids with NcoI and PvuI and ligating the fragments. The forward primers are shown below with the lowercase letters indicating the base change from the wild type; the codon encoding the mutation is underlined.

I69A 5'-CAGGTCACCCCTTCgcCATGACCGTGGGC-3'

I69V 5'-CCCAGGCCACCCCTTCgTCATGACCGTCGGCTGCG-3'

I69L 5'-CCCAGGCCACCCCTTCcTCATGACCGTCGGCTGCG -3'

V325A 5'-GGGGTACAGGTGGCGcGGACACAGCTGCCGTGGG-3'

V325E 5'-GAGGGGTACCGGTGGCGaGGACACAGCTGCAGTGGGC-3'

After treatment with DpnI to remove template DNA, the PCR product was transformed into *E. coli* DH5α (Invitrogen, Carlsbad, CA). The presence of the mutation and fidelity of the mutagenesis were confirmed by sequencing the entire gene.

Expression and Purification. Proteins were expressed and purified using previously described methods (*25*) with some minor modifications. The plasmids were transformed into *E. coli* strain BL21(DE3)pLysS (Stratagene), and the transformed cells were grown in LB medium at 37 °C to an A_{600} = 0.6 – 1.0. The cells were cooled to 25 °C, and protein expression was induced with 0.4 mM IPTG. After 6 hours of growth, the cells were harvested by centrifugation and were resuspended in MES buffer (10 mM MES, 20 mM KCl, 1 mM DTT, pH 6.0) containing 0.1 mM PMSF. The cells were lysed by freeze-thawing, followed by the addition of DNase to a final concentration of ~60 units/mL. The suspension was centrifuged at 22000×*g* for 30 minutes, and the supernatant was loaded onto a Blue Sepharose CL-6B column (Amersham Biosciences, Piscataway, NJ) as described previously (*25*). Mutant proteins were eluted with TES buffer (10 mM TES, 1 mM DTT, pH 8.0) containing 20 mM KCl. The CK mutants were further purified on a HiPrep Q (Amersham Biosciences) column (*25*). Each purified mutant enzyme appeared as a single band on SDS PAGE.

Protein Stability Measurements. Thermal denaturation profiles were performed on a Jasco J-715 spectropolarimeter (Easton, MD) equipped with a Jasco PTC-348WI Peltier-effect temperature control device and in-cell stirring. Native CK and mutant concentrations were 0.02 mg/mL in 50 mM KPi, 200 mM KCl and 38% ethylene glycol at pH 6.8. The CD spectrum at 223 nm was monitored in 0.2 °C increments from 20-70 °C. Melting temperatures for the gross unfolding of creatine kinase were determined using the EXAM software program (*26*).

Kinetic Characterization. Steady-state kinetic analyses of CK were determined using a previously described coupled assay (*27, 28*) modified for a microplate format. Activity measurements were performed at 30 °C in 96-well UV-transparent microplates. Assay controls using the native enzyme found the results from the microplate procedure to be similar to previous studies (*24, 25, 29*). The final assay mixture contained 75 mM TAPS buffer, pH 9.0, 0.36 mM NADH, 0.36 mM phosphoenolpyruvate, 1 mM $Mg(OAc)_2$, 13 mM KOAc, and variable concentrations of MgATP, guanidino substrate and CK in a final volume of 300 μL. Concentrations of the coupling enzymes, pyruvate kinase and lactic acid dehydrogenase, were 28 U/mL and 54 U/mL, respectively. The reaction was initiated by the addition of the guanidino substrate.

Activity was determined by monitoring the oxidation of NADH at 340 nm using a molar extinction coefficient of 6.22 $\text{mM}^{-1}\ \text{cm}^{-1}$. The microplate assay uses a mean pathlength of 0.8 cm. Data points were collected every 12 seconds for 10 minutes and the maximum rate was determined using at least 5 points sampled over 1 minute. At pH 9.0, CK operates by a rapid equilibrium, random bi-bi mechanism (*11*). Data were fitted to Eq. (1), using SigmaPlot 8.0 with the Enzyme Kinetics module from SPSS (Chicago, IL),

$$v = V_{\text{max}}[\text{A}][\text{B}] \,/\, (\alpha(K_{\text{ia}}K_{\text{ib}} + K_{\text{ib}}[\text{A}] + K_{\text{ia}}[\text{B}]) + [\text{A}][\text{B}]) \qquad (1)$$

where [A] and [B] are the substrate concentrations of guanidino substrate (GS) and MgATP respectively, and K_{ia} and K_{ib} are the dissociation constants for the E-GS and E-MgATP complexes respectively, and αK_{ia} and αK_{ib} are the dissociation constants of GS and MgATP, respectively, from the E-GS-MgATP complex. Thus the term α quantifies how the binding of one substrate affects the binding of the other.

Where mutant activity was too low to perform detailed kinetic analysis or when substrate solubility was poor, *V*/*K* conditions were used to estimate $k_{\text{cat}}/K_{\text{m}}$. The assay mixture was prepared as above with the following changes. Guanidino substrate was varied at non-saturating levels while the concentration of MgATP was maintained at 5 mM (saturating). Velocity values for at least three independent trials were plotted against substrate concentrations. Under these conditions, the slopes of the plots yield $V_{\text{max}}/K_{\text{m}}$.

RESULTS

Sequence and Structural Similarities and Differences Among Guanidino Kinases. In many aspects, the functions of the flexible loop regions in the characterized phosphagen kinases, AK and CK, are essentially identical. Each loop recognizes the appropriate substrate and assists in positioning the substrate for catalysis. In addition, the loops undergo substantial movement upon substrate binding leading to stabilization of the active conformation. A multiple sequence alignment of the flexible loop regions of a representative set of phosphagen kinases is shown in Figure 2.

The N-loops exhibit significant differences in the length and amino acid composition across the various phosphagen kinases (Figure 2) (*20, 22*). In most[1] AKs, the amino groups of the N-loop residues 63-65 (SGV) make contacts with the carboxylate of arginine and D62 stabilizes the N-loop through a hydrogen bond with R193 (Figure 3a) (*20*). Mutagenesis studies have shown that the mutation of either D62 or R193 to glycine reduces activity ~100-fold (*30*). The N-loop of the CKs is longer than that found in AKs and possesses a conserved residue, H66, which appears to interact with D326 from the C-loop, possibly acting to stabilize the active conformation (Figure 3b) (*21*). Recent studies (P.-F. Wang, A.J. Flynn, M.J. McLeish and G.L. Kenyon, unpublished results) indicate that mutations at H66 or D326 have a significant effect on catalysis, with mutation of either residue resulting in greatly reduced activity. In CK, none of the conserved N-loop residues except I69 make specific contacts with the creatine substrate (*21*). In the CK-TSAC structure (Figure 3b), residue I69 makes hydrophobic contacts with both the *N*-methyl[2] of creatine and V325. The interaction of I69 with V325 form a "specificity pocket" for the *N*-methyl group in CK (*21*).

```
                    59                78   318              338
                              ↓                   ↓
LK-Eisenia          PSVDNTG-----RIIGLVAGD  QKRGTGGEHTEAVDDVYDISN
GK-Neanthes         TGVDNPGNKFYGKKTGCVFGD  GKRGTGGESSLAEDSTYDISN
CK-Rat.M            TGVDNPGHPFIM-TVGCVAGD  QKRGTGGVDTAAVGAVFDISN
CK-Human.M          TGVDNPGHPFIM-TVGCVAGD  QKRGTGGVDTAAVGSVFDVSN
CK-Chicken.M        TGVDNPGHPFIM-TVGCVAGD  QKRGTGGVDTAAVGAVFDISN
CK-Torpedo          TGVDNPGHPFIM-TVGCVAGD  QKRGTGGVDTEAVGSIYDISN
CK-Human.B          TGVDNPGHPYIM-TVGCVAGD  QKRGTGGVDTAAVGGVFDVSN
CK-Rat.B            TGVDNPGHPYIM-TVGAVAGD  QKRGTGGVDTAAVGGVFDVSN
CK-Chicken.B        TGVDNPGHPFIM-TVGCVAGD  QKRGTGGVDTAAVGGVFDVSN
CK-Rat.Mi           TGVDNPGHPFIK-TVGMVAGD  QKRGTGGVDTPATADVFDISN
CK-Human.Mi         TGVDNPGHPFIK-TVGMVAGD  QKRGTGGVDTAATGGVFDISN
CK-Chicken.Mi       TGVDNPGHPFIK-TVGMVAGD  QKRGTGGVDTAATANVFDISN
AK-Lobster          SGVENLD-----SGVGIYAPD  QVRGTRGEHTEAEGGIYDISN
AK-Honeybee         SGIENLD-----SGVGIYAPD  QVRGTRGEHTEAEGGIYDISN
AK-Shrimp           SGVENLD-----SGVGIYAPD  QVRGTRGEHTEAEGGIYDISN
AK-Limulus          SGVENLD-----SGVGIYAPD  QVRGTRGEHTESEGGVYDISN
AK-Battilus         SGCLNLD-----SGVGIYACD  QARGIHGEHTESEGGVYDLSN
```

Figure 2. Alignment of the N- and C-loop regions of guanidino kinases. The sequence alignment was generated using ClustalW (*41*) and edited based on structural alignments of AK (1BG0) and CK (1N16) created with MinRMS (*42*). Numbering is in reference to human muscle CK. Loop regions are boxed. The hydrophobic specificity pocket residues I69 and V325 in CK are indicated with arrows on the multiple sequence alignment. Swissprot accession numbers are as follows: lombricine kinase (LK) *Eisenia* (O15991); glycocyamine kinase (GK) *Neanthes* (P51546); CK Rat.M (P00564); CK Human.M (P06732); CK Chicken.M (P00565); CK *Torpedo* (P00566); CK Human.B (P12277); CK Rat.B (P07335); CK Chicken.B (P05122); CK Rat.Mi (P25809); CK Human.Mi (P12532); CK Chicken.Mi (P70079); AK Lobster (P14208); AK Honeybee (O61367); AK Shrimp (Q95V58); AK *Limulus* (P51541); AK *Battilus* (O15989).

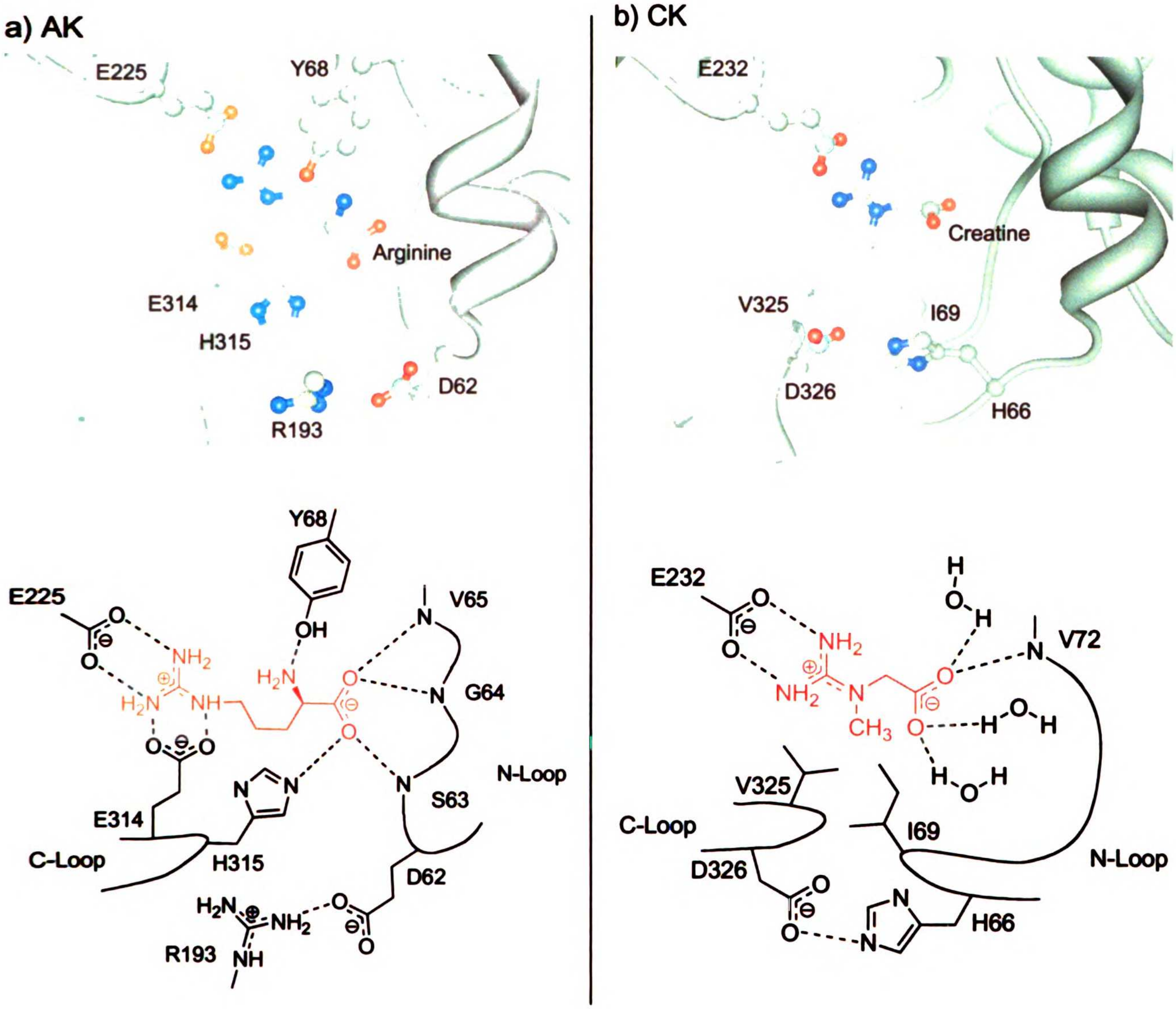


Figure 3. Active site comparisons of AK and CK. Ribbon diagrams with ball and stick ligands and side chains generated with Chimera (*15*) and Chemdraw representations of the active sites of AK (1BG0) and CK (1N16). Requirements of the flexible loop regions for both enzymes include precise positioning of the guanidino substrate and stabilization of the loops in the active conformation. Only the glutamate at 225 in AK and 232 in CK is strictly conserved between these active sites. a) In the AK-TSAC structure the guanidino terminus of arginine is positioned through C-loop residues E314 and H315, while the amino terminus is

positioned through interactions with the backbone nitrogens of N-loop residues 63-65. The N-loop of AK is stabilized by the interaction of D62 with R193. b) In the CK-TSAC structure the guanidino terminus of creatine is positioned through the hydrophobic interaction of I69 and V325 with the methyl group of the creatine substrate. The carboxy-terminus of CK is stabilized primarily through interactions with conserved, buried waters. The CK loops are stabilized by the interaction of H66 and D326, from the N- and C-loops, respectively.

Comparisons of phosphagen kinase C-loops reveal class-specific sequence differences although the lengths of the loops are identical across all superfamily members (Figure 2). All CKs have a conserved VD motif (CK residues 325-326) which aligns structurally with the EH motif (AK residues 314-315) conserved across all AKs and lombricine kinases and the ES motif of glycocyamine kinase (Figure 2). The AK-TSAC structure shows the C-loop EH motif contacts only the bound arginine substrate (*20*). E314 forms two critical hydrogen bonds with the guanidino terminus of arginine (Figure 3a) (*20*). The conservative E314Q mutation reduces the k_{cat} of the enzyme ~300-fold (*29, 31*). H315 forms a single hydrogen bond with the carboxylate of arginine, and is too distant to interact with the AK N-loop (*20*). This is in contrast to CK, where C-loop residues V325 and D326 contact N-loop residues I69 and H66, respectively (Figure 3b) (*21*). V325 and I69 also make contacts with the *N*-methyl of the bound creatine substrate, whereas D326 and H66 do not make any substrate contacts, but may stabilize the active, closed loop conformation (*21*). It is important to note that AK E314 and CK V325 are aligned in structural superpositions (Figure 3) although, prior to the solution of the CK-TSAC structure, it was often assumed that AK E314 was homologous to, and would align with, CK D326 (*29, 32*).

Protein Expression and Stability. CK and all variants were routinely expressed and purified essentially as described previously (*25*). In order to ensure that the overall stability of the protein is unaffected by the mutations, thermal denaturation profiles were obtained. The mutant enzymes exhibit profiles identical to native CK (data not shown).

Enzyme Activity and Kinetic Parameters. Kinetic parameters for the reaction of native and mutant CK enzymes in the forward direction (phosphagen formation) were determined in a microplate format (see Methods). Creatine, glycocyamine, *N*-ethylglycocyamine and cyclocreatine were employed as substrates (Figure 1a-d). Binding constants and α values are listed in Table 1, except for glycocyamine, which, due to its poor solubility, could not be fully characterized kinetically. These substrates were selected based on their similarity to creatine, and their ability to probe specific enzyme-substrate interactions. For example, glycocyamine lacks the *N*-methyl group of creatine, while cyclocreatine and *N*-ethylglycocyamine have bulkier rigid and flexible *N*-substituents, respectively. The k_{cat} and specificity constants (k_{cat}/K_m) are listed in Table 2. None of the mutations we made significantly perturb the K_d value for MgATP. Although the I69A/V325A and I69V/V325A double mutants were also prepared and tested, the activities of these mutants were too low to allow accurate interpretation (data not shown).

Mutations at I69. The effects of mutagenesis at I69 vary significantly depending on the mutation and substrate utilized. The I69A mutant possesses the highest activity with creatine as the substrate; however, the k_{cat} is reduced 20-fold in comparison to the native enzyme. The K_d for creatine is increased approximately two-fold over native enzyme, while the K_m for creatine is increased over 25-fold, indicative of a change in the synergy (α value[3]) of the enzyme. Binding synergy has been previously observed for some CK isozymes, although the degree of synergy varies and may be affected by many factors (*11, 33*). Under our conditions, native CK exhibits synergy with creatine and has an α value of 0.24. In contrast, the I69A mutation results in a greater than 15-fold increase in α with creatine (α=3.7), indicating negative synergy with creatine. The increase in this α value also dramatically affects the K_m value, resulting in an enzyme with a 500-fold reduction in its specificity for creatine.

Table 1. α Values and Dissociation and Michaelis Constants for CK Mutants with Several Guanidino Substrates[ab]

			MgATP, mM		Guanidino Substrate, mM	
Enzyme	Substrate[c]	α	K_d^{MgATP}	K_m^{MgATP}	K_d^{GS}	K_m^{GS}
WT	Cr	0.24	0.90 ± 0.12	0.22 ± 0.05	38.6 ± 6.3	9.3 ± 2.5
	CyCr	0.37	0.79 ± 0.09	0.29 ± 0.07	67 ± 12	24.5 ± 7.1
	EG	0.68	0.91 ± 0.07	0.62 ± 0.12	103 ± 14	70 ± 16
I69A	Cr	3.7	0.86 ± 0.09	3.17 ± 1.02	69 ± 11	250 ± 90
	CyCr	*n.d.*	*n.d.*	*n.d.*	*n.d.*	*n.d.*
	EG	*n.d.*	*n.d.*	*n.d.*	*n.d.*	*n.d.*
I69V	Cr	0.49	1.11 ± 0.10	0.54 ± 0.11	75 ± 10	36.4 ± 8.5
	CyCr	0.51	0.95 ± 0.10	0.47 ± 0.14	130 ± 28	66 ± 22
	EG	1.7	1.00 ± 0.09	1.66 ± 0.38	81 ± 11	135 ± 34
I69L	Cr	0.52	0.84 ± 0.06	0.44 ± 0.09	137 ± 20	72 ± 17
	CyCr	0.71	0.89 ± 0.11	0.64 ± 0.17	75 ± 13	53 ± 16
	EG	1.3	0.85 ± 0.10	1.1 ± 0.3	58.2 ± 9.1	76 ± 21
V325A	Cr	0.63	1.01 ± 0.07	0.64 ± 0.12	113 ± 15	71 ± 15
	CyCr	0.46	0.95 ± 0.07	0.44 ± 0.08	109 ± 14	50 ± 10
	EG	0.83	0.85 ± 0.09	0.70 ± 0.20	100 ± 19	83 ± 27

[a] Values are shown ± SE. K_d and K_m values are discussed in the Methods section. Each data point is an average of at least 3 individual measurements. GS, guanidino substrate. [b] *n.d.*, activities are too low for full kinetic characterization. [c] Cr, creatine; CyCr, cyclocreatine; EG, *N*-ethylglycocyamine.

Table 2. Velocity and Specificity Constants of CK Mutants[ab]

Enzyme	Substrate[c]	k_{cat}, s^{-1} (% WT)	k_{cat}/K_m, s^{-1} M^{-1}	(k_{cat}^{GS}/K_m^{GS})
			(% WT)	$/ (k_{cat}^{Cr}/K_m^{Cr})$
	Cr	85.3 ± 1.5 (100)	9190 (100)	1
WT	CyCr	35.2 ± 0.8 (100)	1440 (100)	0.16
	EG	14.5 ± 0.5 (100)	207 (100)	0.02
	Gly	*n.d.*	21[d] (100)	0.002
	Cr	4.2 ± 0.5 (5)	17 (0.2)	1
I69A	CyCr	*n.d.*	0.5[d] (0.03)	0.03
	EG	*n.d.*	0.5[d] (0.2)	0.03
	Gly	*n.d.*	*n.d.*	*n.d.*
	Cr	82.7 ± 2.2 (97)	2270 (25)	1
I69V	CyCr	4.4 ± 0.2 (13)	67 (5)	0.03
	EG	12.0 ± 0.7 (83)	89 (43)	0.04
	Gly	*n.d.*	1.4[d] (7)	0.001
	Cr	45.0 ± 1.3 (53)	627 (7)	1
I69L	CyCr	1.5 ± 0.1 (4)	28 (2)	0.04
	EG	0.55 ± 0.02 (4)	7.2 (3)	0.01
	Gly	*n.d*	0.3[d] (1)	0.0005
	Cr	15.4 ± 0.5 (18)	216 (2)	1
V325A	CyCr	18.0 ± 0.4 (51)	359 (25)	1.66
	EG	4.3 ± 0.2 (30)	52 (25)	0.24
	Gly	*n.d.*	1.7[d] (8)	0.008
	Cr	*n.d.*	0.14[d] (0.002)	1
V325E	CyCr	*n.d.*	*n.d.*	*n.d.*
	EG	*n.d.*	*n.d.*	*n.d.*
	Gly	*n.d.*	15[d] (71)	104

[a] Each data point is an average of at least 3 individual measurements. Unless otherwise noted, k_{cat} and k_{cat}/K_m values are determined from Table 1. Percent of WT values are shown in parentheses. Values are shown ± SE. GS, guanidino substrate. [b] *n.d.*, activities are too low for determination. [c] Cr, creatine; CyCr, cyclocreatine; EG, ethylglycocyamine; Gly, glycocyamine. [d] Values were determined under pseudo-first order conditions discussed in the Methods section. SE are estimated to be ± 10%.

The I69V mutant exhibits a k_{cat} similar to that of the native enzyme with both creatine and *N*-ethylglycocyamine as substrates, while the k_{cat} for cyclocreatine is reduced 8-fold. While the k_{cat} values for creatine and *N*-ethylglycocyamine are near WT values, the K_m values for these substrates are significantly higher, thereby also affecting the specificity of the enzyme. The K_m for cyclocreatine is lower than that for *N*-ethylglycocyamine; however, this tighter binding does not result in an increase in the k_{cat} value with the cyclocreatine substrate. Instead, the decrease in K_m is accompanied by a decrease in k_{cat} for cyclocreatine. This phenomenon may indicate nonproductive binding with cyclocreatine as the substrate (*34*). Despite the slight changes in activity and binding for *N*-ethylglycocyamine, the I69V mutation results in an increase in α from 0.7 in the native enzyme to 1.7 in the mutant. This negative synergy is similar to that observed for I69A with creatine, albeit less pronounced.

In addition to mutations which generate a larger substrate binding pocket at I69, we also investigated the I69L mutant. This substitution reduces the size of the binding pocket and would be expected to decrease the enzyme's affinity for creatine and bulkier substrates. However, it is conceivable that the extra methyl group at this position could alter the enzyme to prefer glycocyamine by mimicking the *N*-methyl group of the creatine substrate. The k_{cat}/K_m for I69L with glycocyamine is reduced ~100-fold in comparison to the WT enzyme with this substrate and indicates that such a compensation does not occur. The k_{cat} value with this mutant is also reduced 2-fold with the creatine substrate, and 25-fold with cyclocreatine and *N*-ethylglycocyamine in comparison to the WT values with these substrates. The K_d for cyclocreatine is similar to that for WT, and the K_d for *N*-ethylglycocyamine is almost 2-fold lower than WT, despite the added bulk nearer the substrate binding pocket. The specificity constants (k_{cat}/K_m) for each of the substrates are significantly reduced, with values for all substrates ranging from 1-7% of WT. The α values for I69L are similar to I69V for all substrates, including the negative synergy observed with *N*-ethylglycocyamine.

Mutations at V325. While mutations at I69 affect both the synergy and specificity of the enzyme, mutations at this position were unable to switch the overall preference of CK away from its natural substrate, creatine. In contrast, both mutations at V325 alter the specificity of the enzyme away from creatine and towards one of its analogs. Mutation of V325 to alanine creates a slightly larger substrate binding pocket, which we speculated would direct the enzyme to prefer bulkier substrates such as cyclocreatine and *N*-ethylglycocyamine. However, the V325A mutant fails to improve the K_d or K_m values (Table 1) for either of the bulkier substrates. In fact, the mutation of V325 to alanine results in increased K_m values, accompanied by decreased activities, for all substrates. The ability of V325A CK to bind and catalyze the reaction with creatine as the substrate is reduced more than with the cyclocreatine substrate. This results in a modest 1.7-fold preference for the cyclocreatine substrate over creatine.

Mutation of V325 to glutamate results in a striking preference for glycocyamine over creatine. As shown in Table 2, comparisons of the specificity constants of V325E for creatine and glycocyamine reveal that the enzyme prefers glycocyamine by two orders of magnitude. The k_{cat}/K_m value with creatine as substrate is reduced nearly 70,000-fold relative to wild type, while the specificity constant with glycocyamine remains essentially unchanged in comparison to the native enzyme. Examination of the sequences and structures suggests an explanation for this preference. Both the larger size and negative charge of the glutamate at this position should interfere with the binding of the hydrophobic *N*-methyl group of creatine (see Discussion). In addition, the glutamate may be able to accept a hydrogen bond from the glycocyamine substrate, a potentially stabilizing interaction which the native valine at that position cannot provide.

DISCUSSION

The use of flexible loops in the binding of and discrimination between substrates as well as catalysis is common among many enzyme families. In these cases, a conformational change, such as the closure of flexible loops, occurs upon ligand binding and is required to properly orient the catalytic machinery and substrates for catalysis. There is substantial precedent in several enzymes that the alteration of specificity may be achieved through variations in flexible loops (*4, 9, 10*). Inspection of the unliganded (*16-19*) and TSAC (*20, 21*) structures for both AK and CK indicates that these phosphagen kinases also utilize flexible loop regions to orient active site elements and substrates for catalysis. These structures also suggest that some determinants of substrate specificity are present on these loop regions (*30, 35, 36*). However, the precise roles of many of the residues in the loops remain unclear, and, prior to this study, researchers have been unable to alter the specificity of either of these enzymes by making changes in these loops (*36*).

Literature reports of the alteration of substrate specificity have generally been limited to sites distant from or only indirectly associated with the residues involved in catalysis (*4, 10, 37*). In contrast, the specificity determinants in the phosphagen kinases are close to the site of chemistry, and may even assist in catalysis through the optimal positioning of the substrate. Thus, the phosphagen kinases present a unique system in which to examine the delivery of substrate specificity through flexible loop regions.

The reasons for a dual role of the flexible loop residues both in determining specificity and mediating catalysis can be rationalized by examination of the TSAC structures of AK and CK. First, the loop residues interacting with the substrates are important for catalysis (*20, 31*). The sp^2 hybridization of the guanidino nitrogens in all guanidino kinase substrates results in a rigid, planar substrate at the site where chemistry occurs. This planarity allows the precise alignment of the guanidino group necessary for productive catalysis to be controlled through the positioning of any rigid *N*-substitutions at the guanidino group, or by the positioning of the guanidino group itself. Therefore, creatine may be optimally aligned for catalysis through interactions with its rigid *N*-methyl substitution at the guanidinium ion (Figure 1a), while the other naturally occurring guanidino substrates (Figure 1b, e) may be positioned through direct interactions with the guanidino group. The importance of precise substrate positioning has been demonstrated in AK where structures of AK-TSAC mutants reveal that subtle perturbations in the substrate alignment lead to a loss of activity (*31*). Second, these same residues are important in the discrimination of substrates. In AK, the arginine substrate appears to be precisely positioned for catalysis, at least in part, through the interaction of E314 with the guanidino group (Figure 3a) (*20*). In CK, precise alignment of the substrate is partially achieved through the interaction of the *N*-methyl group of creatine with the hydrophobic residues I69 and V325 (Figure 3b) (*21*). AK E314 and CK V325 align structurally, and our results show that the substrate specificity of CK may be altered through the mutagenesis of V325. Thus, in CK, and possibly in other phosphagen kinases, we see that specific flexible loop residues may play a role in both catalysis and substrate recognition.

The results reported here are consistent with the interpretation that precise positioning of the substrates is important in catalysis and help to explain results from previous studies on

AK and CK. These studies have generated much debate over the contributions of acid-base chemistry (*25, 29, 38, 39*), strain and optimal substrate alignment to catalysis (*16, 20, 29, 31, 32*). Prior to the solution of the CK-TSAC structure, the candidates for the putative catalytic base were identified as either the strictly conserved E232 (E225 in AK) or the C-loop residue D326 (E314 in AK) (Figure 3a, b) (*20, 29, 32*). A study by Cantwell et al. (*29*) found that in CK the functionally conservative mutation of E232D results in a 500-fold decrease in activity compared with a modest 3-fold decrease in activity for the D326E mutation. This suggests that E232 is more likely to act as a catalytic base; in AK, a parallel result was found for the mutation of E225 (*31*).

Pruett et al. (*31*) have suggested that in AK, the flexible loop residue which interacts with the guanidino terminus, E314, is not acting as a catalytic base. Rather, its role lies in the precise positioning of the substrate. Our results support this interpretation. Examination of the CK-TSAC structure shows that there is no candidate on the C-loop to act as a catalytic base. CK residue V325 aligns both structurally (Figure 3a, b) and in the multiple sequence alignment (Figure 2) with E314 in AK. Clearly, valine is unable to act as a catalytic base. Instead, V325 is important in determining the specificity of CK. The V325E mutation severely restricts the ability of the enzyme to turn over creatine, as determined from the k_{cat}/K_m, while the k_{cat}/K_m value with glycocyamine as a substrate is essentially unaffected. This results in a ~100-fold preference of V325E CK for glycocyamine over creatine. Our results also show that by increasing the size of the binding cavity at this position (V325A) we reduce the k_{cat}/K_m value for creatine to 2% of WT, whereas the k_{cat}/K_m value with cyclocreatine is only reduced to 25% of WT. Thus, we obtain a modest (1.7-fold) preference for the rigid substrate, cyclocreatine.

In comparison to V325, the role of I69 is more subtle with respect to its contribution to either catalysis or specificity. While mutations at I69 do affect substrate specificity, mutations at this position alone fail to alter the preference of the enzyme for a substrate other than creatine. Even the I69L mutation, which might be expected to increase activity with glycocyamine by compensating for the loss of the *N*-methyl group, fails to direct the enzyme to prefer this substrate. Thus, the low activities of I69L and I69A may largely be associated with the fact that specificity in CK is conferred through the interaction of V325 and I69 with the *N*-methyl group in creatine, in a composite system whose precise interactions do not easily tolerate changes. These interactions appear to be unnecessary in homologous enzymes such as AK and GK, whose substrates lack the *N*-methyl group. We speculate that, as in AK and the V325E CK mutant, glycocyamine may prefer a hydrogen bond acceptor at the V325 position for optimal positioning and/or catalysis (note that an H-bond acceptor is not strictly required as WT CK has activity with glycocyamine).

While mutations at I69 alone fail to direct CK alter substrate preference, single site mutations at I69 do affect the binding synergy. The loss of synergy resulting from the I69A, I69V and I69L mutations is similar to that recently reported in an I69G mutant of the *Danio* CK isozyme (*35*). We propose the role of I69 and V325 in CK is to both act as a specificity filter and to aid in the optimal positioning of the substrate for catalysis.

Although much has been learned about phosphagen kinase catalysis in general from studies on CK and AK, transfer of insight about catalysis from either AK to CK or *vice versa* should be approached with some caution due to important differences in the nature and behavior of their flexible loops. The specific roles of the flexible loop residues in AK appear to differ significantly from those in CK. AK is considered a more primitive phosphagen kinase than CK (*40*) and its flexible loops do not interact. In the study by Pruett et al. (*31*), the C-loop in AK was substituted with a CK-like C-loop (R312G/E314V/H315D/E317A/E319V). This mutant has near wild-type activity with arginine, leading those investigators to question the role of the C-loop interactions with substrate in the AK reaction. By contrast, in the more evolved CK, the loop residues have significant interactions. These interactions may place additional functional constraints on the CK enzyme not present in AK and may help in rationalizing the differences between the results of the Pruett study (*31*) and those reported here. A more informative understanding of these results may require a TSAC structure of the AK mutant and more detailed studies of the effects of the individual mutations in its C-loop.

Very recently, Azzi et al. (*36*) investigated the roles of the N-loop residues as specificity determinants in AK and showed that the specificity of AK by is not altered by substituting a longer CK-like N-loop for the short AK N-loop. In our study of CK, the results for the V325E mutant provide a possible explanation for the lack of change in specificity in the AK study. Our results show that the bulk and/or negative charge of the glutamate at position V325 has a critically negative impact on the catalytic efficiency of the enzyme with creatine as the substrate (a ~70,000-fold decrease). Pertinent to these observations, Azzi et al. (*36*) also obtained a structure of AK with creatine and ADP bound. In this structure, creatine binds similarly to its native conformation in CK, except, as expected, there is a distortion in the guanidino plane of the substrate, presumably caused by a steric clash of the methyl group of creatine with E314. We note that the loop mutants of Azzi et al. (*36*) all contain E314. Based on the results reported here, it is possible these AK loop mutants have no detectable activity with creatine simply because E314 interferes with the optimal positioning of creatine. Based on observations from our study, and the Pruett (*31*) and Azzi (*36*) studies, we suggest that the conversion of AK to CK will not be achieved solely through N-loop mutations and, minimally, will require the additional mutation of E314 to a small hydrophobic residue such as valine or alanine.

While it is now clear there are significant differences between the flexible loops of AK and CK, early interpretation of the roles of these loops in recognition and catalysis in CK was compromised by the incorrect assumption that the homologous residue to AK E314 was CK D326 (*20, 29, 32*). The CK-TSAC structure allows the correction of this mistake and suggests that the E314 of AK has been replaced in CK with the hydrophobic pocket formed by I69 and V325. Our mutagenesis results support this suggestion. Moreover, we now postulate that the specific role of position V325 in CK is to recognize and align the substrate for catalysis. We therefore have termed position V325 a "specificity switch," where the identity of the amino acid at this position determines the enzyme's preference for creatine, cyclocreatine, or glycocyamine.

Given the results of this and earlier studies, we may now suggest a new idea for how the phosphagen kinase superfamily of enzymes has evolved to recognize a variety of substrates and catalyze their reactions. Phosphagen kinase substrates may be described in relative terms as long (arginine, lombricine), short (creatine, glycocyamine), having an *N*-methyl group (creatine) and lacking an *N*-substitution (arginine, glycocyamine). Similarly, the cognate enzymes of these substrates have evolved accordingly. Although the study by Azzi et al. (*36*) raises questions about the importance of loop length relative to substrate size, at least in nature, long substrates are catalyzed by enzymes with short N-loops, while short substrates are catalyzed exclusively by enzymes with long N-loops. Substrates with an *N*-

methyl group such as creatine may only be catalyzed by enzymes possessing a small hydrophobic group on the C-loop, while substrates lacking this *N*-substitution such as arginine and glycocyamine require an acidic group, e.g., a glutamate, at the analogous position. Only one of the four possible substrate-enzyme combinations (a long, *N*-methylated substrate) does not occur naturally. It may be that the interactions between the N- and C-loops are required for recognition of an *N*-methylated substrate (note the reduced activity of the I69A mutant) and therefore, limitations in the size of the associated binding pocket may preclude the use of a substrate such as *N*-methylarginine.

Finally, in the context of induced fit theories, we have identified an enzyme superfamily where the specificity factors presented on flexible loops are also important in catalysis. Thus, in addition to proton abstraction, precise positioning of the substrates is also required for optimal catalysis. Here, we have described a system where one of these residues responsible for correct positioning of the substrate may be altered to direct the enzyme to prefer an alternative substrate.

ACKNOWLEDGMENTS

We would like to thank Dr. Brian Shoichet and Ray Nagatani, Jr. for assistance with the CD experiments.

REFERENCES

1. Koshland, D. E. Jr. (1958) Application of a theory of enzyme specificity to protein synthesis, *Proc. Natl. Acad. Sci. 44*, 98-104.

2. Joseph, D., Petsko, G. A., and Karplus, M. (1990) Anatomy of a conformational change: hinged "lid" motion of the triosephosphate isomerase loop, *Science 249*, 1425-8.

3. Gerstein, M., and Chothia, C. (1991) Analysis of protein loop closure. Two types of hinges produce one motion in lactate dehydrogenase, *J. Mol. Biol. 220*, 133-149.

4. Hedstrom, L., Szilagyi, L., and Rutter, W. J. (1992) Converting trypsin to chymotrypsin: the role of surface loops, *Science 255*, 1249-53.

5. Fersht, A. R. (1974) Catalysis, binding and enzyme-substrate complementarity, *Proc. R. Soc. Lond. B. Biol. Sci. 187*, 397-407.

6. Fersht, A. R. (1985) in *Enzyme Structure and Mechanism* pp 331-333, W.H. Freeman and Co., New York.

7. Herschlag, D. (1988) The role of induced fit and conformational changes of enzymes in specificity and catalysis, *Bioorg. Chem. 16*, 62-96.

8. Post, C. B., and Ray, W. J., Jr. (1995) Reexamination of induced fit as a determinant of substrate specificity in enzymatic reactions, *Biochemistry 34*, 15881-5.

9. Graf, L., Jancso, A., Szilagyi, L., Hegyi, G., Pinter, K., Naray-Szabo, G., Hepp, J., Medzihradszky, K., and Rutter, W. J. (1988) Electrostatic complementarity within the substrate-binding pocket of trypsin, *Proc. Natl. Acad. Sci. U.S.A. 85*, 4961-4965.

10. Wilks, H. M., Hart, K. W., Feeney, R., Dunn, C. R., Muirhead, H., Chia, W. N., Barstow, D. A., Atkinson, T., Clarke, A. R., and Holbrook, J. J. (1988) A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework, *Science 242*, 1541-4.

11. Morrison, J. F., and James, E. (1965) The mechanism of the reaction catalyzed by adenosine triphosphate creatine phosphotransferase, *Biochem. J. 97*, 37-52.

12. Morrison, J. F., and Cleland, W. W. (1966) Isotope exchange studies of the mechanism of the reaction catalyzed by adenosine triphosphate: creatine phosphotransferase, *J. Biol. Chem. 241*, 673.

13. Cleland, W. W. (1967) Enzyme Kinetics, *Ann Rev Biochem 36*, 96-99.

14. Kabsch, W., and Fritz-Wolf, K. (1997) Mitochondrial creatine kinase-A square protein, *Curr. Opin. Struct. Biol. 7*, 811-818.

15. Huang, C. C., Novak, W. R., Babbitt, P. C., Jewett, A. I., Ferrin, T. E., and Klein, T. E. (2000) Integrated tools for structural and sequence alignment and analysis, *Pac. Symp. Biocomput.*, 230-241.

16. Yousef, M. S., Clark, S. A., Pruett, P. K., Somasundaram, T., Ellington, W. R., and Chapman, M. S. (2003) Induced fit in guanidino kinases-comparison of substrate-free and transition state analog structures of arginine kinase, *Protein Sci. 12*, 103-111.

17. Fritz-Wolf, K., Schnyder, T., Wallimann, T., and Kabsch, W. (1996) Structure of mitochondrial creatine kinase, *Nature 381*, 341-345.

18. Rao, J. K., Bujacz, G., and Wlodawer, A. (1998) Crystal structure of rabbit muscle creatine kinase, *FEBS Lett. 439*, 133-7.

19. Eder, M., Schlattner, U., Becker, A., Wallimann, T., Kabsch, W., and Fritz-Wolf, K. (1999) Crystal structure of brain-type creatine kinase at 1.41 A resolution, *Protein Sci. 8*, 2258-69.

20. Zhou, G., Somasundaram, T., Blanc, E., Parthasarathy, G., Ellington, W. R., and Chapman, M. S. (1998) Transition state structure of arginine kinase: implications for catalysis of bimolecular reactions, *Proc. Natl. Acad. Sci. U.S.A. 95*, 8449-54.

21. Lahiri, S. D., Wang, P. F., Babbitt, P. C., McLeish, M. J., Kenyon, G. L., and Allen, K. N. (2002) The 2.1 A structure of Torpedo californica creatine kinase complexed with the ADP-Mg(2+)-NO(3)(-)-creatine transition-state analogue complex, *Biochemistry 41*, 13861-7.

22. Suzuki, T., Kawasaki, Y., Furukohri, T., and Ellington, W. R. (1997) Evolution of phosphagen kinase. VI. Isolation, characterization and cDNA-derived amino acid sequence of lombricine kinase from the earthworm Eisenia foetida, and identification of a possible candidate for the guanidine substrate recognition site, *Biochim. Biophys. Acta 1343*, 152-159.

23. Rowley, G. L., Greenleaf, A. L., and Kenyon, G. L. (1971) On the Specificity of Creatine Kinase. New Glycocyamines and Glycocyamine Analogs Related to Creatine, *J. Am. Chem. Soc. 93*, 5542-5551.

24. Chen, L. H., White, C. B., Babbitt, P. C., McLeish, M. J., and Kenyon, G. L. (2000) A comparative study of human muscle and brain creatine kinases expressed in Escherichia coli, *J. Protein Chem. 19*, 59-66.

25. Wang, P. F., McLeish, M. J., Kneen, M. M., Lee, G., and Kenyon, G. L. (2001) An unusually low pK(a) for Cys282 in the active site of human muscle creatine kinase, *Biochemistry 40*, 11698-705.

26. Kirchhoff, W. (1993), National Institute of Standards and Technology, Gaithersburg, MD.

27. Tanzer, M. L., and Gilvarg, C. J. (1959) Creatine and Creatine Kinase Measurement, *J. Biol. Chem. 234*, 3201-3204.

28. Wang, P. F., Novak, W. R., Cantwell, J. S., Babbitt, P. C., McLeish, M. J., and Kenyon, G. L. (2002) Expression of Torpedo californica creatine kinase in Escherichia coli and purification from inclusion bodies, *Protein Expr. Purif. 26*, 89-95.

29. Cantwell, J. S., Novak, W. R., Wang, P. F., McLeish, M. J., Kenyon, G. L., and Babbitt, P. C. (2001) Mutagenesis of two acidic active site residues in human muscle creatine kinase: implications for the catalytic mechanism, *Biochemistry 40*, 3056-61.

30. Suzuki, T., Fukuta, H., Nagato, H., and Umekawa, M. (2000) Arginine kinase from Nautilus pompilius, a living fossil. Site-directed mutagenesis studies on the role of amino acid residues in the Guanidino specificity region, *J. Biol. Chem. 275*, 23884-23890.

31. Pruett, P. S., Azzi, A., Clark, S. A., Yousef, M. S., Gattis, J. L., Somasundaram, T., Ellington, W. R., and Chapman, M. S. (2003) The putative catalytic bases have, at most, an accessory role in the mechanism of arginine kinase, *J. Biol. Chem. 278*, 26952-26957.

32. Zhou, G., Ellington, W. R., and Chapman, M. S. (2000) Induced fit in arginine kinase, *Biophys. J. 78*, 1541-1550.

33. Maggio, E. T., and Kenyon, G. L. (1977) Properties of a CH3-blocked creatine kinase with altered catalytic activity. Kinetic consequences of the presence of the blocking group, *J. Biol. Chem. 252*, 1202-7.

34. Fersht, A. R. (1998) *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*, W.H. Freeman and Co., New York, NY.

35. Uda, K., and Suzuki, T. (2004) Role of amino acid residues on the GS region of Stichopus arginine kinase and Danio creatine kinase, *The Protein Journal 23*, 53-64.

36. Azzi, A., Clark, S. A., Ellington, W. R., and Chapman, M. S. (2004) The role of phosphagen specificity loops in arginine kinase, *Protein Sci. 13*, 575-85.

37. Venekei, I., Szilagyi, L., Graf, L., and Rutter, W. J. (1996) Attempts to convert chymotrypsin to trypsin, *FEBS Lett. 383*, 143-7.

38. Cook, P. F., Kenyon, G. L., and Cleland, W. W. (1981) Use of pH studies to elucidate the catalytic mechanism of rabbit muscle creatine kinase, *Biochemistry 20*, 1204-10.

39. Eder, M., Stolz, M., Wallimann, T., and Schlattner, U. (2000) A conserved negatively charged cluster in the active site of creatine kinase is critical for enzymatic activity, *J. Biol. Chem. 275*, 27094-9.

40. Suzuki, T., Kawasaki, Y., and Furukohri, T. (1997) Evolution of phosphagen kinase. Isolation, characterization and cDNA-derived amino acid sequence of two-domain arginine kinase from the sea anemone Anthopleura japonicus, *Biochem. J. 328 (Pt 1)*, 301-6.

41. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res. 22*, 4673-80.

42. Jewett, A. I., Huang, C. C., and Ferrin, T. E. (2003) MINRMS: an efficient algorithm for determining protein structure similarity using root-mean-squared-distance, *Bioinformatics 19*, 625-34.

INTRODUCTION TO CHAPTER 5

The work described in Chapter 5 began a couple years ago when Patsy came to me describing a new paradigm in enzyme evolution that she and John Gerlt were working on (Gerlt and Babbitt, 2001). In this paradigm the active site of a protein is conserved, but the new protein uses the active site residues to perform a different catalytic mechanism, i.e., there is no conservation of a common catalytic step. Proteins exhibiting this type of similarity are said to belong to an enzyme *suprafamily*. The original suprafamily was comprised of two proteins, orotidine 5'-monophosphate (OMPDC) and hexulose-6-phosphate synthase (HPS). Both of these enzymes possess a conserved Asp-X-Lys-X-X-Asp motif, yet the catalytic mechanisms are very different. In addition, it was suggested that D-ribulose-5-phosphate 3-epimerase (RPE) may be another distant relative of this suprafamily.

I began examining the sequences and available structures using evolutionary trace techniques. The more I examined these proteins the less clear their evolutionary connection became. Eventually, we decided to shelve this project and focus on other research.

Fast forward to more recent times, when we decided to revisit the links between these proteins. The first seven $(\beta/\alpha)_8$, or TIM-barrel, superfamilies in SCOP have a similar phosphate binding site (SPB). I decided to begin with the examination of the entire SCOP ribulose phosphate binding barrel (RPBB) superfamily. Since all members reportedly have this conserved domain, I decided to test the effects of omitting this region of the barrel from sequence similarity searches. This was the piece of data we were missing before. I found that removal of this sequence portion eliminates many of the sequence links found with the full length sequences. Further, searching only with the SPB region generates many sequence links. Thus, we have found that sequence searches using full length proteins may often lead to unclear relationships between proteins, and that by dividing these sequences into functional sections can help to clarify these relationships. In addition, I also came across what appears to be another circular permutation in TIM-barrels, this time between OMPDC and the histidine biosynthesis enzymes. It remains to be seen whether randomly generating fragments for Blast or hmmsearches will lead to a better understanding of the evolutionary relationships in other TIM-barrels and other protein folds, but this is a collaboration Patsy and I intend on pursuing in the future.

Gerlt, J. A. and P. C. Babbitt (2001). "Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies." *Annu. Rev. Biochem.* **70**: 209-46.

CHAPTER 5

Dissecting Distant Evolutionary Relationships in the Ribulose-Phosphate Binding Barrel Proteins

Walter R.P. Novak and Patricia C. Babbitt

Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry

University of California, San Francisco, 600 16th St., San Francisco, CA 94143

ABSTRACT

Proteins within the ribulose-phosphate binding barrel superfamily (RPBB) as defined by SCOP are able to catalyze a wide variety of reactions on an equally wide variety of substrates. RPBB members include the tryptophan and histidine biosynthesis enzymes, D-ribulose-5-phosphate 3-epimerases, and the orotidine 5'-monophosphate and 3-keto-L-gulonate 6-phosphate decarboxylases, and all share the $(\beta/\alpha)_8$, or TIM-barrel, fold. These enzymes display distant similarities between their protein structures and their substrates. Each substrate has a ribulose-like backbone structure and each substrate is phosphorylated. This phosphate group is bound in a conserved location, the β7-8 loop region, across all superfamily members. In addition, sequence searches such as PSI-Blast and hmmsearch with specific RPBB members are often able to find other members of this superfamily. Thus, it has been hypothesized that these (and many other TIM-barrels) have evolved from a common ancestor, yet, because of extremely low sequence identities, the foundation for this assertion remains unclear. Here we investigate the sequence links in the RPBB superfamily in an attempt to better understand these distant relationships. Hidden Markov Models (HMMs) were generated from the full length, β1-6 and β7-8 regions for each family. The HMMs were searched against the NR database and subjected to congruence analysis and tree generation. Our results indicate that many of the sequence links are achieved solely through similarities in the β7-8, or phosphate binding, region, and lack similarity in the β1-6 region. Our results indicate that these phosphate binding sites have evolved independently of the remainder of the barrel. Further, we demonstrate that other small two-β strand units within the β1-6 region appear to be circularly permuted within this superfamily.

INTRODUCTION

The $(\beta/\alpha)_8$, or TIM-barrel fold, is one of the most common protein folds (*1, 2*), comprising approximately 10% of currently solved protein structures (*3*). TIM-barrel proteins are capable of performing a wide range of catalytic functions, with fold members represented in five of the six primary EC categories. Consistent with the wide diversity of functions performed by this fold, the TIM-barrel proteins as a whole fail to show any sequence similarity across the different superfamilies. Yet despite this sequence diversity, TIM-barrels are thought to have evolved from a single common ancestor (*1, 4-6*). Perhaps the best evidence for the divergent evolution of this fold is the conservation of a similar phosphate-binding site in a subset of TIM-barrels (*2, 7*). Here, we investigate the contribution of the sequence signal associated with phosphate binding in the identification, classification and evolution of TIM-barrel superfamilies, particularly the ribulose-phosphate binding barrel proteins.

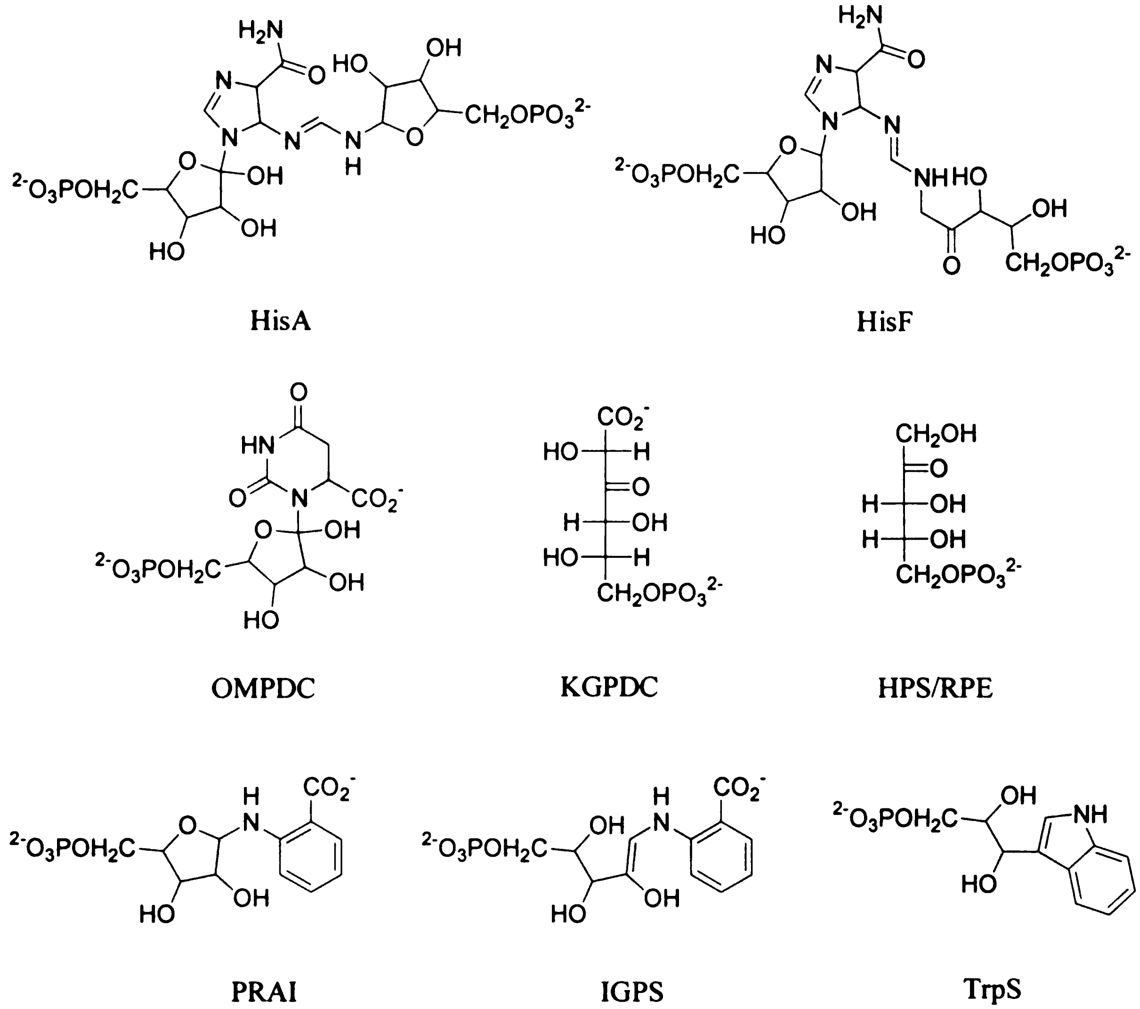
Two hierarchical databases, SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) (*8*) and CATH (http://www.biochem.ucl.ac.uk/bsm/cath/) (*9*), reveal a subset of TIM-barrels which possess a similar phosphate binding (SPB) domain (*2, 7*). The phosphate binding site of the SPB is formed by the loop regions between β7-α7 and β8-α8. In SCOP, the first seven superfamilies of the TIM β/α barrel fold possess an SPB, while in CATH, the Homologous superfamily 3.20.20.90 contains all TIM-barrels with the SPB domain.

A subset of SPBs has further been grouped by SCOP into the ribulose-phosphate binding barrel superfamily (RPBBs) (Table 1). In addition to the similar phosphate-binding site, the substrates utilized by the RPBBs have structural elements built on the sugar ribulose

Table 1. SCOP Ribulose-Phosphate Binding Barrel Superfamily.[a]

SCOP family (Abbreviation)	Pfam protein families (Abbreviation)	E.C. Number(s)
Histidine biosynthesis enzymes (His)	Histidine biosynthesis protein (His)	5.3.1.16 4.1.3.-
D-ribulose-5-phosphate 3-epimerase (RPE)	D-ribulose-5-phosphate 3-epimerase (RPE)	5.1.3.1 5.1.3.-
Decarboxylase (DCase)	Orotidine 5'-monophosphate decarboxylase / hexulose-6-phosphate synthase (DCase)	4.1.1.23 4.1.2.-
Tryptophan biosynthesis enzymes (Trp)	N-(5'phosphoribosyl)anthranilate isomerase (PRAI)	5.3.1.24
	Indole-3-glycerolphosphate synthase (IGPS)	4.1.1.48
	Trp synthase alpha-subunit (TrpS)	4.2.1.20

[a] Where the SCOP families are defined by a single Pfam family, the same abbreviation is used.



HisA

HisF

OMPDC

KGPDC

HPS/RPE

PRAI

IGPS

TrpS

Figure 1. Substrates utilized in the RPBB superfamily.

(Figure 1). The RPBBs are diverse in sequence and function, with proteins in this superfamily often sharing less than 10% identity. RPBB proteins perform reactions in two primary EC categories (lyase and isomerase) and include the tryptophan (Trp) and histidine biosynthesis enzymes (His), D-ribulose-5-phosphate 3-epimerases (RPEs), and the orotidine 5'-monophosphate and 3-keto-L-gulonate 6-phosphate decarboxylases (DCases) (Table 1) (*10*). Despite the differences in these enzymes, sensitive sequence similarity search tools such as PSI-BLAST (*11*) and hmmsearch (*12*) are able to find sequence links between many of these proteins. For example, Copley and Bork (*13*) used PSI-BLAST to identify homologous TIM-barrel families, and provided evidence that many of the TIM-barrel superfamilies share a common ancestor. However, the basis for these sequence connections remains unelucidated.

Lang et al. proposed an approach by which extant proteins could be deconvoluted by subdivision into small ancestral precursor domains, and used this approach experimentally to show that the histidine biosynthesis enzymes from the HisA and HisF genes are likely to have evolved from a common half-barrel ancestor (*14*). Here, in an attempt to understand the basis for the sequence connections in the RPBB superfamily, we utilize a computational approach in which the sequences of superfamily members are subdivided into the β1-6 (active site) and β7-8 (phosphate binding) regions. These fragments along with the full length sequences are used to perform sequence similarity searches and to construct SATCHMO (http://www.drive5.com/lobster) (*15*) trees. We show that only searches using the β7-8 region are able to link all RPBB superfamily members and that the clustering of these proteins varies dramatically depending on the seed sequence segment used. With these results we further develop the evolutionary model of TIM-barrel proteins, and suggest that the phosphate binding regions of these proteins evolved along a different path than the domains associated with the remainder of the barrel (β1-6 region). Additionally, we demonstrate that identification of subsequence signals allows the rationalization of sequence hits found in the noise region of sequence similarity searches.

RESULTS

Congruence Analyses. Hidden Markov models (HMMs) were constructed for the full length, β1-6 and β7-8 sequence regions of each Pfam (http://www.sanger.ac.uk/Software/Pfam/) (*16*) family represented in the RPBB superfamily (Table 1). In SCOP, four families make up the RPBBs. Three families can each be described by a single Pfam family, while the tryptophan biosynthesis enzymes require multiple Pfam families (TrpS, PRAI and IGPS) to fully describe the family. Hmmsearch was used to perform similarity searches on each HMM (see Methods). Overlaps in the database searches were analyzed with the program Intersect. Sequence links identified between families based on the various sequence regions are illustrated in Figure 2.

Intersect (http://www.babbittlab.ucsf.edu/software/Intersect) (*17*) analysis reveals that similarity searches using full length HMMs only find four links (out of 15 possible) between the six RPBB Pfam families (Figure 2a). Although the three tryptophan biosynthesis enzyme Pfam families comprise a single SCOP family, searches were unable to find sequence links between each of the members. Only a link between the IGPS and PRAI families was found; however, this link is misleading as PRAI and IGPS are merged in a bifunctional enzyme, so that each search identifies only its homologous domain within the fused protein.

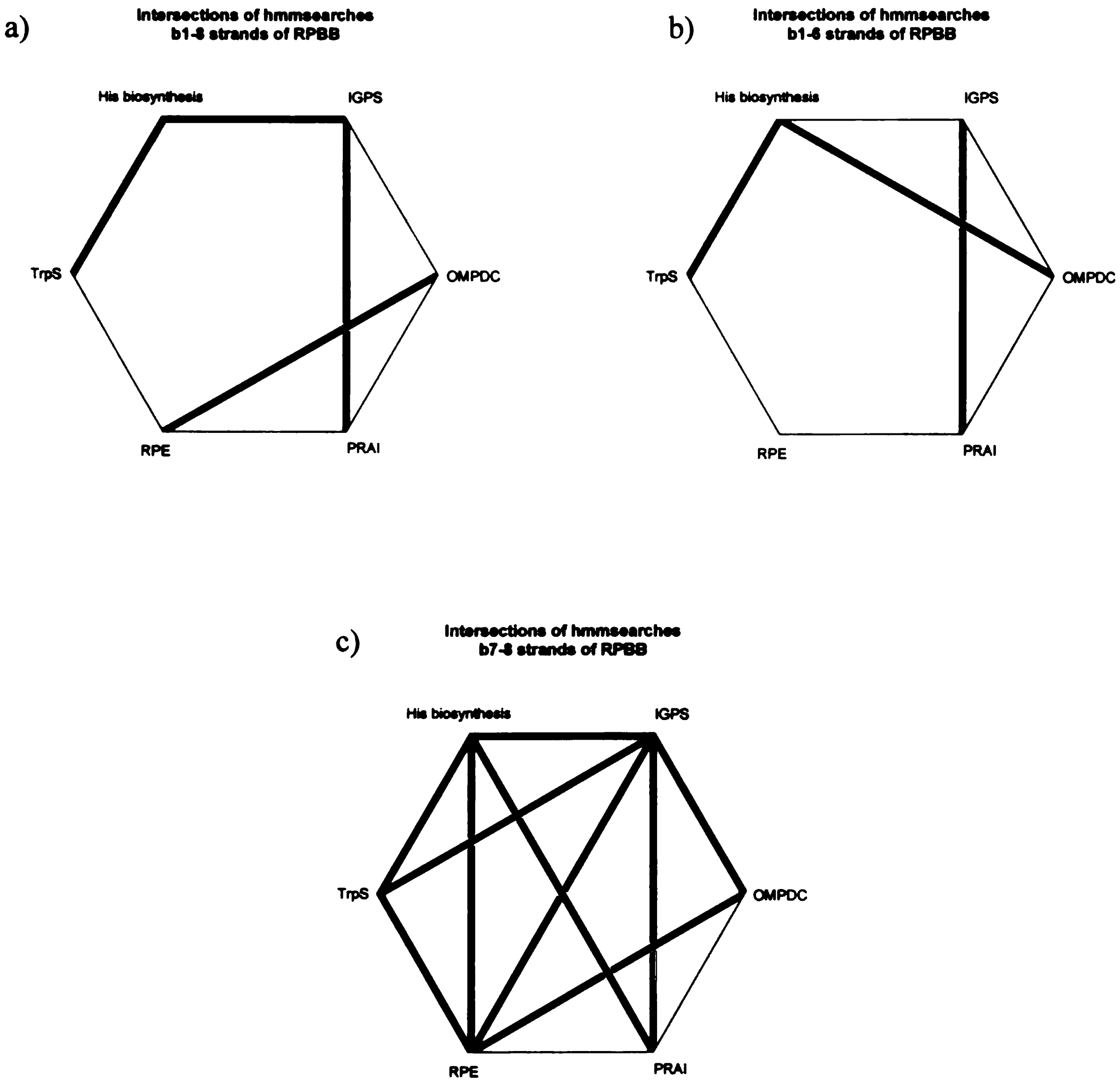


Figure 2. Overlaps between database searches.

Investigations of the sequence links are not clearly identified as bifunctional enzymes show that they indeed are merged in a single protein and that the HMM for each of these families is still identifying a different domain.

The His proteins, though diverse with respect to sequence identity, are defined by a single Pfam family. Searches with the full length His HMM find links between both the IGPS and TrpS families, but the links to each of these families are tenuous. Intersect identifies only a single sequence link between His and TrpS, a tryptophan synthase from *Methanothermobacter*, and it is identified by the His HMM with an E-value of 5.7. The link between His and IGPS is consists of three sequences, an IGPS from *Pyrococcus* and two putative N-acetylmannosamine-6-phosphate 2-epimerases from *Thermoanaerobacter*. Each of these links has at least one E-value of 2 or greater.

The remaining sequence link found using the full length HMMs is between RPE and DCase. In contrast to the previously described results, where no or only a few valid sequence links between families are found, searches using full length HMMs for RPE and DCase find 60 overlapping sequences. The majority of the sequences found by both searches are identified as RPEs, and three of these are identified with both E-values less than 1. These families are also linked by 20 unknown environmental sequences and two D-arabino 3-hexulose 6-phosphate synthases (HPSs). HPSs along with the 3-keto-L-gulonate 6-phosphate decarboxylases (KPGDCs) have been shown to be related to the orotidine 5'-monophosphate decarboxylases (OMPDCs) in a *supra*family context (*18*), though no HPS structure is currently available. RPEs have also been hypothesized to be distantly related to this suprafamily; however, the nature of this relationship has not been elucidated (*18*).

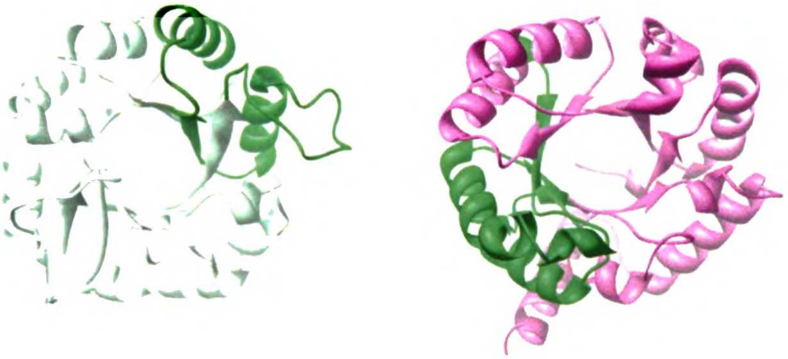
For the groups linked, the sequence region identified by each family was also examined to ensure that each linker is recognized based on the same sequence motif. In each of the above cases, the full length sequence HMMs for each family identify the same sequence region.

Sequence links between families based on the β1-6 regions are shown in Figure 2b. Here, only three sequence links are found, and one of these is the invalid link described above between IGPS and PRAI. The remaining links are between TrpS and His and between His and OMPDC, and in each case the sequence links are found with few sequences and relatively high E-values.

Searches with HMMs constructed from the β1-6 regions of the His and TrpS families result in the identification of 10 linking sequences. Eight of these sequences are TrpSs, one is a putative IGPS and one is an unknown environmental sequence. In each case, the sequence is more distant to the His family with E-values in the 2.3-9.3 range.

The His and OMPDC families also find few overlapping sequences. Two of these are HPSs from *Archaeoglobus* and *Methanosarcina* and one is a hypothetical protein from *Methanosarcina*. Again, the His sequences are more distant with E-values ranging from 1.8-6.9.

Examination of the sequence and structural regions identified by the β1-6 region HMMs reveal that, although six strands were used to build the HMM, only a region corresponding to approximately two βα units (in some cases the first β-strand is not part of this motif) is identified by both HMMs (Figure 3). Interestingly, this sequence region is not co-localized on each of the barrels. Figure 3a shows that the region of overlap between the a)

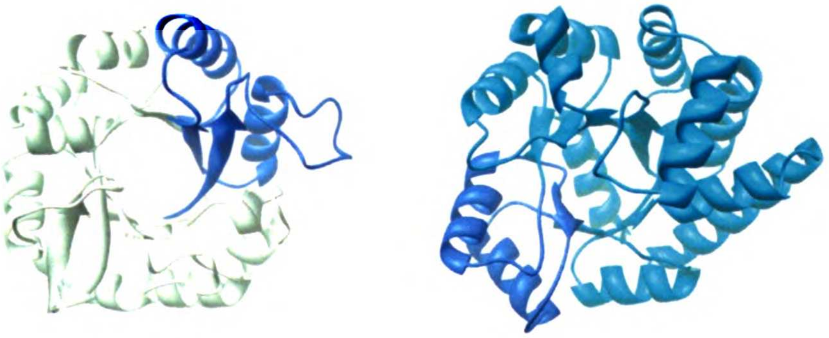


```
His
Consensus   k G  V       +        l+ + DP+  Ak  ee G+d+L ++v +d  + G +  l+v++ +
11498467    KYGCEV------M-----ADLINVPDPASRAKEVEELGVDYLNVhVGIDQQMKGLD-PLEVLKDV

OMPDC
Consensus   +yg             ++++ +++++++++ +++e++++g  +++  vh ++d++ +
11498467    KYG-------------CEVMADLINVPDPASRAKEVeELG--VDYLNVHVGIDQQMK---------

OMPDC
Consensus        G+ + + +                d+
11498467    -----GLDPLEVLK---------------DV
```

b)



```
His         rIIPalDlkdGrvVRLykGdynYPvfknlvyagDPvelAkryeeeGAdeLHfvD.LdAAk.eGrpvnld
Consensus   ++ P             d+     + l+ +           ++ G+ +L+ v + ++A++  ++  l+
136261      IVAP------------TTDN-----QRLKMISE--------VSSGFHYLVSVMgVTGARsRVEESTLE

His         vieriaeevfiP
Consensus   +ier+  + ++P
136261      LIERVKGAGSLP

TrpS        LvAPtTsdeRlktiaeaasGFiYlVSraGVTGararavneqldelvarlKkytnvP
Consensus   +vAPtT+++Rlk+i+e++sGF+YlVS++GVTGar+ +v+e++ el++r+K + ++P
136261      IVAPTTDNQRLKMISEVSSGFHYLVSVMGVTGARS-RVEESTLELIERVKGAGSLP
```

Figure 3. Sequence regions identified by the β1-6 HMMs for a) His (PDB 1THF, white) and OMPDC (PDB 1DVJ, magenta) (green motif) and b) His and TrpS (PDB 1TTQ, cyan) (blue motif). Sequence alignments are from the hmmsearch output. Linking sequences are identified by their GI number. E-values for the His-OMPDC intersection (GI 11498467) are 6.9 and 4.2e-39, respectively. E-values for the His-TrpS intersection (GI 136261) are 2.3 and 2.6e-119, respectively. Structural alignments were obtained from the Structural Superposition Database (SSD) (*19*) extension of Chimera (*20*).

HMMs of His and OMPDC corresponds roughly to the β1-2 region in His and the β5-6 region in OMPDC. A similar result is shown for the His-TrpS intersection (Figure 3b).

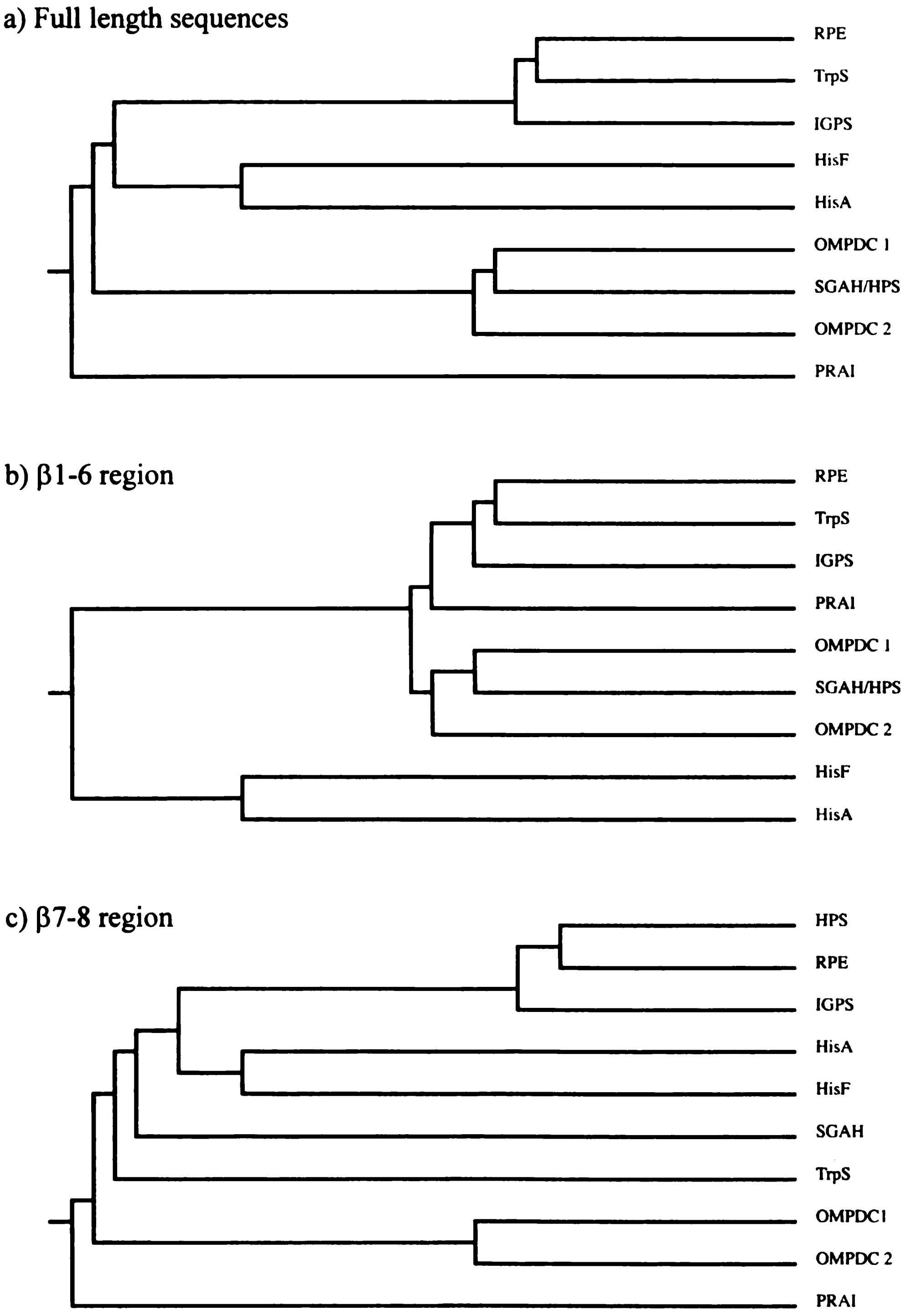
Figure 2c illustrates the sequence links found using HMMs constructed from the β7-8 region of the RPBBs. Many more links are found based on HMMs built from this region as compared to searches with either the full length or β1-6 region HMMs. The His, RPE and IGPS families each show links to four other families (not including the IGPS-PRAI link). The number of linking sequences and best linking sequences between all RPBB families (except the IGPS-PRAI intersection) are shown in Table 2.

Using only the β7-8 region HMMs, the strongest link between any two families within the RPBB superfamily is between RPE and DCase. More than 100 linking sequences are identified by each of these families, with 62 identified as RPEs. Other linking sequences include 13 HPSs, one thiamin phosphate synthase (TPS), one OMPDC and a variety of unknown and hypothetical sequences. The best linking sequence (an RPE) is found by both families with E-values less than 10^{-4}. RPEs are also linked to IGPSs primarily via other RPEs, although 4 TPSs also link these families. It is interesting to note that the best linking sequence for the IGPSs and RPEs is also the best linking sequence between the IGPSs and OMPDC families (Table 2).

The TrpS family shows a strong link to both the His and IGPS proteins, with 90 and 76 linking sequences identified, respectively. Of the linking sequences between TrpS and His, 23 are identified as TrpSs, while 48 sequences of unknown function make up over one-half of the linkers. The sequences of known function which link TrpS and IGPS are split between TrpSs and a group of pyridoxine biosynthesis proteins for which there is no available structure.

Table 2. Linking Sequence Information for the β7-8 region HMM searches.

Families Linked	# Linking Sequences	Best Linking Sequence (GI #, Name)	E-values	
His-TrpS	90	136261, TrpS	TrpS	1.5e-29
			His	0.27
His-IGPS	29	21227606, PRAI	His	1.3e-14
			IGPS	0.75
His-PRAI	1	136343, PRAI	PRAI	2e-07
			His	8
His-RPE	23	37525513, PRAI	His	6.9e-13
			RPE	0.4
RPE-OMPDC	113	15606282, RPE	RPE	2.6e-11
			OMPDC	2.1e-05
RPE-IGPS	25	15608546, RPE	RPE	2.5e-14
			IGPS	0.73
RPE-TrpS	10	20094080, Predicted phosphate-binding enzyme	RPE	0.35
			TrpS	1.4
TrpS-IGPS	76	34897216, Put. IGPS	TrpS	4.8e-20
			IGPS	0.76
IGPS-OMPDC	7	15608546, RPE	OMPDC	0.0042
			IGPS	0.73



a) Full length sequences

b) β1-6 region

c) β7-8 region

Figure 4. SATCHMO trees for the a) full length, b) β1-6 region and c) β7-8 region.

Because the β7-8 region HMMs are small, essentially the same sequence region of the linking sequence is identified by each HMM.

SATCHMO Tree Construction. The sequence sets used to develop HMMs were also used for alignment and tree construction by SATCHMO. SATCHMO was chosen based on its ability to model structural similarities between groups of related proteins (*15*). However, SATCHMO trees should not be interpreted as a model of the evolutionary history of a set of proteins, although the trees may be consistent with such a model (*15*). Reduced set representations of the SATCHMO trees constructed from the full length, β1-6 and β7-8 sequence regions are shown in Figure 4.

The SATCHMO tree constructed from the full length sequences closely resembles the SCOP families with a few differences. The His proteins are clustered in a single, diverse branch and are split into two subgroups which reflect the HisA and HisF gene products. The SCOP DCases are also clustered into a single branch. There are four subgroups present in the DCase tree: OMPDC 1, OMPDC 2, HPS and KGPDC. The differences between the SCOP families and the full length tree are apparent in the clustering patterns of Trps and RPEs. The PRAI branch is split off very early from the rest of the tryptophan biosynthesis proteins, while the IGPS and TrpS subgroups cluster closely. Further, the RPEs cluster very closely with the TrpS proteins. This close clustering of RPEs to the TrpSs contrasts with the Intersect results where no link was found between these proteins using full length sequences.

The SATCHMO tree constructed from the β1-6 region is essentially identical to the tree constructed from the full length sequences with one exception. Now, using only the β1-6 region, all tryptophan biosynthesis enzymes cluster more closely. This tree most closely represents the SCOP RPBB superfamily, but again the RPEs cluster closely with the TrpS subgroup.

Another clustering variation is apparent when only the β7-8 sequence region is used to construct the SATCHMO tree. In this version of the tree, only the His cluster remains intact. The OMPDC 1 and 2 subgroups cluster together, but the KGPDC and HPS branches are now removed from this cluster. Interestingly, the HPSs now most closely cluster with the RPEs. The tryptophan biosynthesis enzymes are also split, though now the IGPSs are more closely clustered to the RPEs. This contrasts with the full length and β1-6 trees where the TrpS proteins cluster most closely with RPEs. The PRAIs branch off distantly as they do in the full length SATCHMO tree.

DISCUSSION

Understanding the design principles utilized by nature to develop new functions from existing protein folds remains an important question in molecular biology. Such studies have implications for genome annotation as well as protein engineering efforts. Though the question of divergent evolution has been studied extensively in the TIM-barrel proteins (*1, 2, 5, 14, 21-23*), many questions remain unanswered.

Relationships between similar enzymes can be described as occurring through one or more of the following evolutionary strategies: specificity dominant, chemistry dominant and active site architecture dominant. Specificity dominant evolution was originally proposed by Horowitz in 1945 (*24*). Horowitz suggested that biochemical pathways may have evolved backwards, thus the substrate of last enzyme of a pathway would be the product of the enzyme just before it. The histidine (*25, 26*) and tryptophan (*7*) biosynthetic pathways are examples of this type of evolution. The chemistry dominant evolutionary model relies on the retention of a common catalytic step in the overall mechanism of an enzyme (*27*). Along with the retention of this step, the subset of active site residues responsible for performing the step is also retained. The chemistry dominant model is similar to the active site architecture model (*18*) in that a subset of active site residues are retained; however, where active site architecture alone is dominant, there is no retention of a common chemical step. The DCase superfamily is the first example of this evolutionary strategy (*18*).

Members of the RPBB superfamily are mechanistically diverse and are clustered together by SCOP based upon 1) the conservation of the SPB domain and 2) the similarity of their substrates to the sugar ribulose (Figure 1). Thus, it appears that the RPBB superfamily may exist as a result of divergent, substrate dominant evolution. Copley and Bork provided statistical evidence for the clustering of 12 TIM-barrel superfamilies and suggested that HisA most closely resembles the ancestral protein (*13*).

Considering the conservation of the SPB domain throughout the RPBB superfamily, it is valid to question the contribution of this sequence region in the identification of homologous sequences. In addition, recent evidence indicates that TIM-barrels may not have evolved as a single unit (*14, 28*). Structural comparisons between the HisA N- and C-terminal and HisF N- and C-terminal half-barrel ($((\beta/\alpha)_4)$) structures reveal a number of residues invariant in both sequence and structural space (*14*). These and other studies (*25, 26*) suggest the evolution of HisA and HisF from a common half-barrel ancestor. Höcker et al. showed that the HisF-N and -C half-barrels formed stable, inactive structures when expressed separately (*28*). Furthermore, they demonstrated that co-expression or joint refolding of the half-barrels results in a fully active HisF-NC complex (*28*). It has also been suggested that TIM-barrels may have evolved through the duplication of even smaller units (*29*), and several studies support this suggestion. Luger et al. transposed the β7-8 region of yeast PRAI to the amino terminus of the protein resulting in full activity (*30*). Similarly, the β1-6 fragment of this protein alone has a high degree of secondary structure. When the β1-6 fragment is complemented with the β7-8 fragment, the result is an active protein (*31*). Finally, Matthews and co-workers have demonstrated that the β1-6 region has a high degree of stability in comparison to the β1-α5 structure (*32*). Taken together, these studies support the fragmentation of the RPBB proteins into the β1-6 and β7-8 segments in our study.

Examination of the Intersect results for the full length RPBBs shows that two continuous chains are created: one from TrpS to His to IGPS to PRAI, and one from DCase to RPE. However, we fail to find a valid link between IGPS and PRAI, even at poor expectation values. Since IGPS and PRAI are fused in a bifunctional protein, each HMM recognizes only its homologous domain, and does not recognize a common sequence signature among these two proteins. The sequence links between the DCases and RPEs are extremely strong in comparison to the other sequence links, with 60 sequences found by both HMMs. Taken alone, the links found with the full length HMM sequences suggest that the His and Trp proteins should not be clustered into the same superfamily as DCases and RPEs.

However, such a model is not entirely consistent with the SATCHMO tree data. The HMM and the tree agree that the PRAI subgroup is distant from the remainder of the sequences. However, the tree suggests that the RPEs are not closely related to the DCases, and instead, cluster closely with two of the Trp proteins, TrpS and IGPS. Copley and Bork found a similar (but not the same) pattern using PSI-BLAST. In their scheme, RPEs find both

DCases and TrpSs, and similar to our Intersect results, the TrpSs utilize the His proteins as a link to IGPS.

Unfortunately, all these data agree on only one thing, and that is that they do not agree. Thus while there is statistical evidence of similarities between proteins in the RPBB superfamily, it is difficult to understand much about the underlying basis for these links. It is the understanding of this basis that is perhaps the most relevant to the correct annotation of the genome, understanding the evolution of function and successful protein engineering efforts. Thus we have fragmented these sequences in an effort to understand the nature of the sequence links discovered in similarity searches.

Upon fragmentation of the TIM-barrel sequence into the β1-6 and β7-8 regions, we are left with two subdomains. These are the β1-6 subdomain which, in theory, possesses much of the substrate specificity and catalytic residues, and the β7-8 subdomain, which is only able to bind the phosphate moiety of the ligand. We can now ask the question, "What are the sequence similarities in the β1-6 (or β7-8) region of the RPBB members?"

Similarity searches using HMMS constructed from only the β1-6 regions show fewer sequence links than using the full length sequence. Of particular interest is the loss of the connection between the DCases and RPEs. The RPEs have been hypothesized to be distant members of the DCase suprafamily (*18*). In one case, an HPS has been identified with 28% identity to an RPE (J.A. Gerlt, personal communication); however, there are no conserved catalytic residues between these proteins. Our HMM results do not support the clustering of the RPE proteins with the DCase suprafamily. The SATCHMO results also provide evidence against RPE membership in the DCase suprafamily, where RPEs remain closely clustered with the Trp proteins.

The suggestion that the TrpS proteins are related to the His proteins is evidenced by the Intersect connections between both the full length and the β1-6 regions of these proteins. Examination of the sequence regions identified by both HMMs shows that these regions are not co-localized on the structures. Approximately the first two β-strands of the His proteins are similar to the β5-6 region of TrpS. A similar pattern is observed in the comparison of the His β1-6 HMM with the DCase β1-6 HMM. Again, it is the β1-2 region of the His proteins that shows similarities to the β5-6 region of the DCase OMPDC. These findings may indicate a half-barrel or even quarter-barrel evolution for proteins other than the His proteins, which has not been previously described. In addition, the SATCHMO trees advocate the suggestion that the His proteins may most resemble the ancestral protein. This tree also shows that, despite the HMMs not linking the Trp proteins, the PRAI branch now clusters with the rest of the Trp family.

The β7-8 region results bring a final piece of information to the understanding of the organization and evolution of the RPBB proteins. The Intersect results using the β7-8 HMMs clearly show this common thread between the RPBBs. The SATCHMO trees add detail to this picture and show that subtle differences between the SPB domains may help to explain many previous results. The DCase tree cluster formed using the full length and β1-6 regions has now been scattered along subfamily lines. The OMPDCs remain together, but the KPGDC and HPS branches have been split from the OMPDC cluster. The HPS branch clusters closely with the RPEs, indicating the HPS SPB domain is more similar to the RPE SPB than to the SPB of other DCases. This finding may account for the link many have suggested between the RPE and DCase families (*18, 33*).

In conclusion, it is often easy to cluster proteins based on overall sequence similarity, and quite often, clustering proteins in this manner is sufficient to enhance our understanding of evolutionary relationships between protein families. However, among very distantly related proteins, especially TIM-barrels, where circular permutations are often allowed, global sequence comparisons may lead to unclear or even contradictory results. Fragmentation of the sequence into subdomains for the RPBB proteins may enable us to not only understand the basis of the sequence links found, but also to understand the functional significance of these links as well.

Thus, while our study supports a divergent evolutionary model for the RPBB proteins, it does not necessarily favor a model of general enzyme recruitment and modification. Rather, it appears that, at least in part, evolution has proceeded via quarter barrel shuffling. The SATCHMO tree constructed using the SPB quarter barrel (which differs from the trees constructed using either full length or β1-6 regions), and the prominent non-localization of the sequence signals linking the His-Trp and His-DCase families support this finding. Such a view questions whether or not ancient enzymes need have possessed broad specificities, but also suggests that studies based on the subdivision of sequence into smaller subdomains may yield complex, yet robust evolutionary links between distantly related protein sequences.

METHODS

Sequence Database Searching. The RPBB superfamily is defined by the SCOP database (1.65 release) (*10*). Sequences for each RPBB family were obtained from Pfam (version 14.0) (*34*). Sequences were subdivided into the β1-6 and β7-8 regions by mapping secondary structure elements to the Pfam alignments and HMMs were generated on the full length and fragment alignments using hmmer 2.3.2 (http://hmmer.wustl.edu/) (*12*). HMMs were searched against the NR protein database from the NCBI.

Data Analyses. Congruence analyses between hmmsearch results were performed using Intersect v. 1.2 (*17*). Trees were generated using SATCHMO simultaneous alignment and tree building tool (*15*). Structure comparisons were obtained from the SSD database extension (*19*) of Chimera (*20*).

REFERENCES

1. Farber, G. K., and Petsko, G. A. (1990) The evolution of alpha/beta barrel enzymes, *Trends Biochem Sci 15*, 228-34.

2. Branden, C. I. (1991) The TIM Barrel-the most frequently occurring folding motif in proteins, *Curr. Opin. Struct. Biol. 1*, 978-983.

3. Gerlt, J. A. (2000) New wine from old barrels, *Nat Struct Biol 7*, 171-3.

4. Janecek, S. (1995) Similarity of different beta-strands flanked in loops by glycines and prolines from distinct (alpha/beta)8-barrel enzymes: chance or a homology?, *Protein Sci 4*, 1239-42.

5. Janecek, S., and Balaz, S. (1995) Functionally essential, invariant glutamate near the C-terminus of strand beta 5 in various (alpha/beta)8-barrel enzymes as a possible indicator of their evolutionary relatedness, *Protein Eng 8*, 809-13.

6. Janecek, S. (1996) Invariant glycines and prolines flanking in loops the strand beta 2 of various (alpha/beta)8-barrel enzymes: a hidden homology?, *Protein Sci 5*, 1136-43.

7. Wilmanns, M., Hyde, C. C., Davies, D. R., Kirschner, K., and Jansonius, J. N. (1991) Structural conservation in parallel beta/alpha-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis, *Biochemistry 30*, 9161-9.

8. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol 247*, 536-40.

9. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997) CATH--a hierarchic classification of protein domain structures, *Structure 5*, 1093-108.

10. Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2002) SCOP database in 2002: refinements accommodate structural genomics, *Nucleic Acids Res 30*, 264-7.

11. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res 25*, 3389-402.

12. Eddy, S. R. (1996) Hidden Markov models, *Curr Opin Struct Biol 6*, 361-5.

13. Copley, R. R., and Bork, P. (2000) Homology among (beta/alpha)(8) barrels: implications for the evolution of metabolic pathways, *J Mol Biol 303*, 627-41.

14. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R., and Wilmanns, M. (2000) Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion, *Science 289*, 1546-50.

15. Edgar, R. C., and Sjolander, K. (2003) SATCHMO: sequence alignment and tree construction using hidden Markov models, *Bioinformatics 19*, 1404-11.

16. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004) The Pfam protein families database, *Nucleic Acids Res 32 Database issue*, D138-41.

17. Pegg, S. C., Novak, W. R., and Babbitt, P. C. (2003) Intersect: identification and visualization of overlaps in database search results, *Bioinformatics 19*, 1997-9.

18. Gerlt, J. A., and Babbitt, P. C. (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies, *Annu Rev Biochem 70*, 209-46.

19. Chiang, R. A., Meng, E. C., Huang, C. C., Ferrin, T. E., and Babbitt, P. C. (2003) The Structure Superposition Database, *Nucleic Acids Res 31*, 505-10.

20. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera-A visualization system for exploratory research and analysis, *J Comput Chem 25*, 1605-12.

21. Gerlt, J. A., and Raushel, F. M. (2003) Evolution of function in (beta/alpha)8-barrel enzymes, *Curr Opin Chem Biol 7*, 252-64.

22. Henn-Sax, M., Hocker, B., Wilmanns, M., and Sterner, R. (2001) Divergent evolution of (beta/alpha)8-barrel enzymes, *Biol Chem 382*, 1315-20.

23. Nagano, N., Orengo, C. A., and Thornton, J. M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions, *J Mol Biol 321*, 741-65.

24. Horowitz, N. H. (1945) On the evolution of biochemical syntheses, *Proc. Natl. Acad. Sci. U.S.A. 31*, 153-157.

25. Fani, R., Lio, P., Chiarelli, I., and Bazzicalupo, M. (1994) The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the hisA and hisF genes, *J Mol Evol 38*, 489-95.

26. Thoma, R., Schwander, M., Liebl, W., Kirschner, K., and Sterner, R. (1998) A histidine gene cluster of the hyperthermophile Thermotoga maritima: sequence analysis and evolutionary significance, *Extremophiles 2*, 379-89.

27. Babbitt, P. C., and Gerlt, J. A. (1997) Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities, *J Biol Chem 272*, 30591-4.

28. Hocker, B., Beismann-Driemeyer, S., Hettwer, S., Lustig, A., and Sterner, R. (2001) Dissection of a (beta/alpha)8-barrel enzyme into two folded halves, *Nat Struct Biol 8*, 32-6.

29. Thornton, J. M., Orengo, C. A., Todd, A. E., and Pearl, F. M. (1999) Protein folds, functions and evolution, *J Mol Biol 293*, 333-42.

30. Luger, K., Hommel, U., Herold, M., Hofsteenge, J., and Kirschner, K. (1989) Correct folding of circularly permuted variants of a beta alpha barrel enzyme in vivo, *Science 243*, 206-10.

31. Eder, J., and Kirschner, K. (1992) Stable substructures of eightfold beta alpha-barrel proteins: fragment complementation of phosphoribosylanthranilate isomerase, *Biochemistry 31*, 3617-25.

32. Zitzewitz, J. A., Gualfetti, P. J., Perkons, I. A., Wasta, S. A., and Matthews, C. R. (1999) Identifying the structural boundaries of independent folding domains in the alpha subunit of tryptophan synthase, a beta/alpha barrel protein, *Protein Sci 8*, 1200-9.

33. Wise, E., Yew, W. S., Babbitt, P. C., Gerlt, J. A., and Rayment, I. (2002) Homologous (beta/alpha)8-barrel enzymes that catalyze unrelated reactions: orotidine 5'-monophosphate decarboxylase and 3-keto-L-gulonate 6-phosphate decarboxylase, *Biochemistry 41*, 3861-9.

34. Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains, *Nucleic Acids Res 26*, 320-2.

CONCLUSION

This thesis represents attempts to understand the sequence-structure-function paradigm in two enzyme superfamilies using computational and experimental techniques. The computational methods focus on understanding the evolutionary relationships between enzymes in order to not only comprehend how sequence relates to the function of a protein, but also how we can use this understanding to experimentally engineer enzymes to perform alternate functions.

Chapters 1-4 detail investigations into the substrate specificity of creatine kinase (CK). These studies began with sequence information for CK and several homologs, unliganded CK structures, and one transition-state analog complexed structure for the CK homolog arginine kinase. These early studies demonstrate the ability of computational methods to identify only one (Val 325) of the two residues eventually found to be important in binding the *N*-methyl of the creatine substrate. It took the solution of a CK structure bound with a transition state analog complex to implicate Ile 69 on the N-terminal loop as being important. Finally, in chapter 4, our studies identified the roles of these residues, specifically, that while Val 325 and Ile 69 both contact the *N*-methyl of creatine, only mutations at Val 325 were able to alter CK to prefer alternative substrates. However, it is the interaction of both Ile 69 and Val 325 with the *N*-methyl of creatine that proves to be important in the optimal positioning of the substrate for catalysis.

In chapter 5, the evolutionary relationships among the ribulose-phosphate barrel (RPBB) enzymes, a very diverse superfamily, are examined. It has been suggested that these enzymes have all evolved from a common ancestor; however, even statistical evidence of this link using the full length sequences can be questioned. These questions led us to split the RPBB enzyme sequences into two functional domains: catalytic and phosphate binding. Evolutionary trees were generated using SATCHMO. Hidden Markov Models (HMMs) for each of these domains were used to search the NR database at NCBI. The search results were subjected to congruence analysis using the Intersect program, developed in the Babbitt lab. These results show that many of the sequence links between RPBB members are solely based on similarity of the phosphate binding region. Further, the fact that RPBB members whose active sites are clearly related often have less similar phosphate binding sites suggests that the phosphate binding region has evolved independently of the catalytic domain.

The first four chapters of this thesis demonstrate both the effectiveness and the limitations of computational biology in deciphering specificity determinants. These chapters also validate the notion that computational investigations into the natural evolution of function, through methods such as evolutionary trace, may yield insight directly applicable to protein engineering efforts. In fact, computational methodologies are rapidly improving and, recently, Dwyer, et al. presented the rational design of a functional enzyme from ribose-binding protein.

While the results presented in chapter 5 represent an advancement in the dissection of distant relationships, the applicability of these finding to protein design efforts has yet to be proven. It is likely that studies investigating the nature of distant relationships will not yield much to rational design efforts; however, we speculate that the identification of relationships between minimal functional protein subdomains may enhance DNA shuffling experiments. Patsy and I have begun a collaboration to investigate the applicability of the methods outlined in chapter 5 in understanding evolutionary relationships in other enzyme superfamilies.

Dwyer, M. A., L. L. Looger, and H. W. Hellinga (2004). "Computational Design of a Biologically Active Enzyme." *Science* **304**(5679): 1967-1971.

APPENDIX 1

Comparison of microplate and cuvette-based methods for the kinetic characterization of rabbit muscle creatine kinase

Walter R. P. Novak[a] and Patricia C. Babbitt[ba]

[a]Department of Pharmaceutical Chemistry and

[b]Department of Biopharmaceutical Sciences

University of California, San Francisco, Box 2240, 600 16th St., San Francisco, CA 94143-2240, USA

ABSTRACT

Creatine kinase catalyzes the reversible transfer of the γ-phosphorous of ATP to creatine to form phosphocreatine. Because it plays a key role in cellular energetics and in several disease states, creatine kinase has remained an active area of research. The widely used assay method of Tanzer and Gilvarg has remained relatively unchanged for over 50 years. However, in order to rapidly assay multiple mutant enzymes with a variety of substrates, a microplate based assay is needed. Here we describe the development and validation of such an assay, and compare it to the standard cuvette based assay. The development of this assay increases our assay efficiency greater than 10 fold, and even exhibited lower standard error than the cuvette assay.

Keywords: creatine kinase; NADH-linked; microplate assay

INTRODUCTION

Creatine kinase (CK; EC 2.7.3.2) catalyzes the reversible phosphorylation of creatine by ATP. A widely used assay for CK is an NADH-linked assay developed by Tanzer and Gilvarg [1]. This assay utilizes two enzymes, pyruvate kinase and lactic acid dehydrogenase to link the consumption of ATP to the oxidation of NADH, which may be followed spectroscopically at 340 nm (Fig. 1a). Because the CK reaction requires two substrates, the determination of the kinetic constants for each substrate requires the independent variation of ATP and creatine over at least 5 different concentrations. Including controls, performing such an analysis can require up to 80 assays and can be tedious for the evaluation of several mutants and/or substrates.

James Florini [2] described a microplate assay for the determination of CK activity in the reverse direction (ATP formation) using a modified NAD-linked assay. This protocol utilized glucose-6-phosphate dehydrogenase from *Leuconostoc meseteroides*, which can reduce thio-NAD. This activity can be monitored spectroscopically at 405 nm. More recently, Schulte *et al.* [3] used a microplate-based NADH-linked assay to perform a rapid kinetic characterization of mevalonate kinase, but failed to compare the results to the standard cuvette assay. In this paper, we describe an NADH-linked microplate assay for the rapid kinetic analysis of CK. Further, we compare the microplate assay data with the cuvette based assay.

MATERIALS AND METHODS

Enzymatic assay preparation. Unless otherwise noted, all chemicals and enzymes were purchased from Sigma-Aldrich Chemical Company (St. Louis, MO). Tween-20 was purchased from Bio-Rad (Hercules, CA). BSA and Superblock® (a nonspecific site blocking agent) are from Pierce (Rockford, IL).

A 100 mM stock of MgATP was prepared by mixing equimolar amounts of $Mg(OAc)_2$ and ATP in 75mM TAPS buffer. For the microplate assay, 45 µL of assay buffer containing TAPS, NADH, phosphoenolpyruvate, $Mg(OAc)_2$, KOAc, pyruvate kinase, lactic acid dehydrogenase and CK and 15 µL of MgATP were added to each well. The reaction was initiated by the addition of 240 µL of creatine. For the cuvette assay all volumes were doubled. In both the microplate and cuvette experiments, the final assay mixture contained 75 mM TAPS buffer, pH 9.0, 0.36 mM NADH, 0.36 mM phosphoenolpyruvate, 1 mM $Mg(OAc)_2$, 13 mM KOAc, variable MgATP (0.2-5 mM), variable creatine (6-96 mM) and 9.3 nM rabbit muscle creatine kinase. Concentrations of the linking enzymes pyruvate kinase and lactic acid dehydrogenase are 28 U/mL and 54 U/mL respectively.

Data collection and fitting. Cuvette and microplate readings were taken on the SPECTRAmax 384 spectrophotometer from Molecular Devices (Sunnyvale, CA). UV-transparent microplates were Costar 96 well plates from Corning (Corning, NY).

Activity was determined by monitoring the oxidation of NADH at 340 nm at 30 °C. The molar extinction coefficient for NADH at 340 nm is 6.22 mM^{-1} cm^{-1}. For the microplate assay a mean pathlength of 0.8 cm was determined using the spectrophotometer's Pathcheck® feature. Data points were collected every 12 seconds and the maximum rate was determined using at least 5 points sampled over 1 minute. At pH 9.0, the kinetic mechanism of rabbit muscle CK can be described by a rapid equilibrium, random bi-bi mechanism [4]. Data was fitted to Eq. (1), using SigmaPlot 8.0 with the Enzyme Kinetics module from SPSS (Chicago, IL):

$$v = V_{max}[A][B] / (\alpha(K_aK_b + K_b[A] + K_a[B]) + [A][B]). \qquad (1)$$

Where [A] and [B] are the substrate concentrations of creatine and MgATP respectively, and K_a and K_b are the dissociation constants for the $ES_{creatine}$ and ES_{MgATP} complexes respectively. The term α quantifies how the binding of one substrate affects the binding of the other [5]. For example, $\alpha K_{creatine}$ is a measure of the affinity of creatine for the ES_{MgATP} complex and is referred to as K_m (Cr).

RESULTS AND DISCUSSION

Comparison of microplate and cuvette data. As shown in Figure 1b, CK activity varies linearly over a range of 0.5-5.0 μg of CK per assay. CK activity data from the microplate assay fits well to Eq. 1 with $R^2 = 0.994$, compared with the cuvette assay where $R^2 = 0.965$. The assay is relatively insensitive to inter-well variability and inter-plate differences (data not shown). Data collection for the entire set of 80 assays requires only minutes, while the time required for the preparation of stock solutions is similar to that required for the cuvette based assay. Since the assay volume of the microplate assay is halved in comparison to the cuvette assay, less CK is utilized in kinetic determinations. This is an important advantage when assaying low activity mutants where CK concentrations that exceed 100 µg/mL may be required for detection of activity.

Table 1 shows the kinetic parameters of CK determined by both the microplate and cuvette assay methods. For both methods all kinetic parameters are similar, although consistently lower for the microplate assay. Activity values at the lowest substrate concentrations are approximately equal, indicating a similar limit of detection for the assays. The standard error measurements for the microplate assay were approximately 2-fold lower than the cuvette assay error.

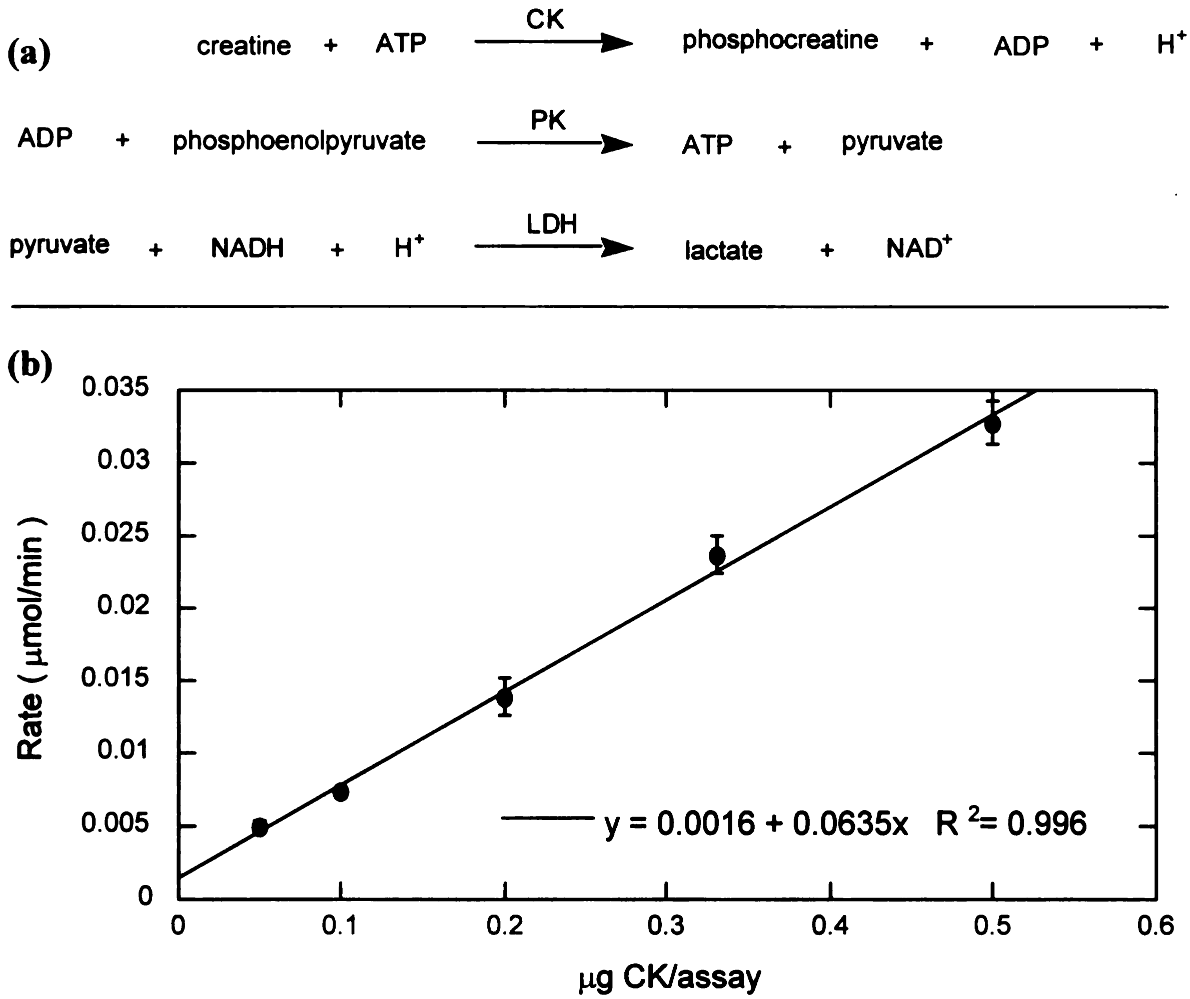
To assess whether the differences between the cuvette and microplate assays were due to unfavorable interactions of assay mix components with the microplate, a variety of detergents and blocking agents were analyzed. The addition of nonionic detergents, Tween-20, Triton X-100, and a zwitterionic detergent, CHAPSO, at 1%, 0.1% and 0.01% had no effect on microplate activity. Similarly, incubating the microplate overnight with BSA 2.0 mg/mL and Superblock® (1x) had no effect (data not shown).

Although the kinetic parameters from the microplate assay differ marginally from the cuvette assay, in our experience these variations are no greater than normal inter-laboratory determinations [6, 7]. The reproducibility, low standard deviations and rapidity of the microplate assay make it a viable method for the determination of kinetic parameters. With the microplate assay, we will be able to rapidly characterize and compare kinetic parameters with a variety of substrates on a many CK mutants in future studies. This assay should be applicable to other kinases with similar results, provided the enzyme does not interact unfavorably with the microplate.

ACKNOWLEDGMENTS

We would like to thank Dr. James McKerrow and Elizabeth "Joey" Hansell for the use of the SPECTRAmax 384 plate reader. We would also like to thank Drs. Paul Cook, Michael McLeish, John Cantwell, Susan Miller and Rich Ledwidge for helpful discussions. This work was supported by NIH RO1 GM60595.

Figure 1.



(a) Schematic indicating the coupling of CK activity to the oxidation of NADH through pyruvate kinase (PK) and lactic acid dehydrogenase (LDH). The oxidation of NADH may be followed spectroscopically at 340 nm. **(b)** The relationship between enzyme concentration and reaction rate. The amount of CK was varied between 0.05-0.5 µg/assay. Creatine and ATP concentrations were maintained at 96 mM and 5 mM respectively. Data points presented are mean ± SD (n = 3).

Table 1: Kinetic constants for CK determined using the microplate and cuvette based assays[a]

	MgATP, mM		creatine, mM		
Method	K_d (MgATP)	K_m (MgATP)	K_d (Cr)	K_m (Cr)	V_{max}, U/mg
Microplate	0.76 ± 0.06	0.13 ± 0.03	44.75 ± 4.86	7.84 ± 1.95	74.94 ± 0.70
Cuvette	0.97 ± 0.18	0.19 ± 0.10	58.61 ± 15.21	11.39 ± 6.83	116.13 ± 3.09

[a] Kinetic constants are shown ± Standard Error. K_d and K_m values are discussed in the Methods section. Each data point is an average of at least 3 individual measurements.

REFERENCES

[1] M.L. Tanzer and C.J. Gilvarg, Creatine and creatine kinase measurement, J. Biol. Chem. 234 (1959) 3201-3204.

[2] J.R. Florini, Assay of creatine kinase in microtiter plates using thio-NAD to allow monitoring at 405 nM, Anal. Biochem. 182 (1989) 399-404.

[3] A.E. Schulte, R. van der Heijden, and R. Verpoorte, Microplate enzyme-coupled assays of mevalonate and phosphomevalonate kinase from Catharanthus roseus suspension cultured cells, Anal. Biochem. 269 (1999) 245-54.

[4] J.F. Morrison and E. James, The mechanism of the reaction catalyzed by adenosine triphosphate creatine phosphotransferase, Biochem. J. 97 (1965) 37-52.

[5] I.H. Segel, Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems, Wiley-Interscience, New York, 1975.

[6] C. Perraut, E. Clottes, C. Leydier, C. Vial, and O. Marcillat, Role of quaternary structure in muscle creatine kinase stability: tryptophan 210 is important for dimer cohesion, Proteins. 32 (1998) 43-51.

[7] J.M. Cox, C.A. Davis, C. Chan, M.J. Jourden, A.D. Jorjorian, M.J. Brym, M.J. Snider, C.L. Borders, Jr., and P.L. Edmiston, Generation of an active monomer of rabbit muscle creatine kinase by site-directed mutagenesis: the effect of quaternary structure on catalysis and stability, Biochemistry. 42 (2003) 1863-71.

APPENDIX 2

Intersect: identification and visualization of overlaps in database search results

Scott C-H Pegg[*], Walter R P Novak[$], Patricia C Babbitt[*$¶]

[*] *Department of Biopharmaceutical Sciences*

[$] *Department of Pharmaceutical Chemistry*

University of California, San Francisco, San Francisco, CA 94143

ABSTRACT

Summary: The determination of distant evolutionary relationships remains an important biological problem, and distant homologs often appear in statistically insignificant regions of sequence similarity searches. Intersect is a computer program designed to identify and visualize the overlaps between sets of sequences reported by multiple database searches. This capability aids researchers in identifying the individual sequences that best bridge sequence families and superfamilies.

Availability: The Intersect program is available from the Babbitt laboratory website at http://www.babbittlab.ucsf.edu/software/intersect

Contact: babbitt@cgl.ucsf.edu

Identifying the sequence links between groups of homologous sequences aides in the determination of function and the analysis of evolutionary conservation. For analysis of divergent families and superfamilies, such linking sequences often have very low sequence similarity to most or all of the sequences contained within the families/superfamilies that they bridge, rendering them very difficult to detect, especially when searching with one sequence at a time. Moreover, while many sequences may appear congruently in search outputs using distantly related sequences as queries, it is not trivial to determine which of these would be most useful for use in subsequent iterations of a complex search strategy. We have developed the Intersect program in an effort to help researchers identify and visualize the sequence links among sets of homologous sequences, including those representing very divergent relationships.

Several useful programs exist to aid in the identification of potential homologs within sequence databases. These include FASTA (Pearson and Lipman, 1988), BLAST (Altschul et al., 1990), PSI-BLAST (Altschul et al, 1997), and several approaches using hidden Markov models (Eddy, 1996, Sjolander, 1996). The output files of these searches often contain distant homologs hidden in regions of low statistical significance. Examination of the overlap between and among multiple searches, each performed with a different but related query sequence, can often distinguish these true homologs out of the noise (Pegg and Babbitt, 1999). In addition, breaking multiple database searches into sets associated with the original query sequences allows the user to search for sequences that bridge the sets of search results. For example, a user may perform 5 database searches using query sequences from family A, and 5 from family B (a total of 10 database searches). Sequences bridging families A and B can be found by looking for sequences reported in both the output files for families A and those for family B. Intersect is also useful for examining new hypotheses about relationships among 2 or more families/superfamilies. Sequences from families hypothesized to be related can be used as queries for congruent database searches followed by Intersect analysis. The linking sequences found may provide evidence of potential relationships that can be examined further by independent methods. A non-automated version of this approach was successfully used to deduce that the *pcpA* gene, found in a pathway associated with the metabolism of pentachlorophenol in *S. chlorophenolica*, belongs to a new and unusual class of extradiol dioxygenases (Xu, 1999).

The Intersect program allows users to identify the sequences reported across multiple sets of output files by providing congruence analysis functionality via a graphical user interface. Each user-defined set of output files is designated by a color (chosen automatically, but also configurable by the user). This color-coding allows for the rapid identification of the sets contributing to each overlapping region. The current version of Intersect supports FASTA, BLAST (both NCBI and Wustl versions), and PSI-BLAST output files. Sets of output files may contain any mixture of these formats. As a result, Intersect can by used to rapidly estimate the specificity and selectivity of these search methods on a particular query or database.

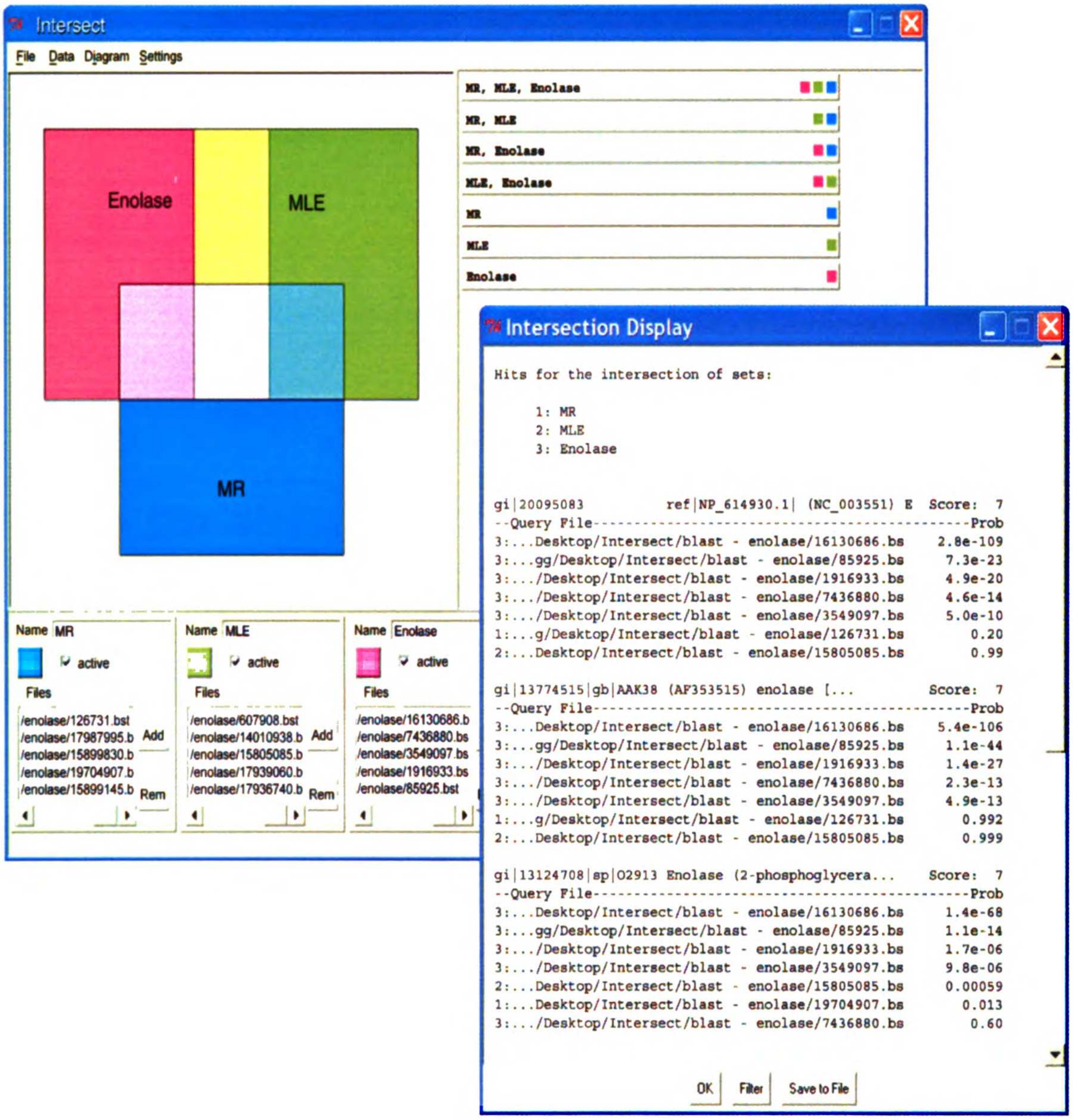
Each non-empty region of the set space is displayed as a color-coded tab in the upper right panel of the interface. Three sets of output files (A, B, and C, for example) could have at most seven non-empty regions (A, B, C, A+B, A+C, B+C, and A+B+C). Clicking on a tab displays the sequences contained within the region, including information regarding which individual search file reported each sequence and it's associated significance score (as reported by the database search program, e.g. BLAST E-value). Intersect allows the user to filter the sequences reported according the their significance scores with both high and low cutoff values.

When the total number of sets is under five, a Venn diagram of the set overlap can be displayed as overlapping rectangles. (Note that a five set Venn diagram has at most 31 non-empty subsets, making it too busy to be practically useful in this application.) The area of the overlapping regions is not scaled by the number of sequences contained within them, and empty regions are colored black. The program allows for the diagram to be easily labeled and printed. Clicking on any of the colored regions brings up the text-based window displaying the details of the sequences contained within it.

Intersect is written in Python to allow for platform independence. It is available (along with a comprehensive user's manual) free of charge to investigators from academic or other non-profit institutions at http://www.babbittlab.ucsf.edu/software/intersect.

ACKNOWLEDGMENTS

This work was supported by NIH RO1 GM60595.



Name MR
☑ active
Files
/enolase/126731.bst
/enolase/17987995.b
/enolase/15899830.b
/enolase/19704907.b
/enolase/15899145.b

Name MLE
☑ active
Files
/enolase/607908.bst
/enolase/14010938.b
/enolase/15805085.b
/enolase/17939060.b
/enolase/17936740.b

Name Enolase
☑ active
Files
/enolase/16130686.b
/enolase/7436880.bs
/enolase/3549097.bs
/enolase/1916933.bs
/enolase/85925.bst

```
Hits for the intersection of sets:

     1: MR
     2: MLE
     3: Enolase


gi|20095083          ref|NP_614930.1| (NC_003551) E  Score:  7
--Query File---------------------------------------------Prob
3:...Desktop/Intersect/blast - enolase/16130686.bs   2.8e-109
3:...gg/Desktop/Intersect/blast - enolase/85925.bs    7.3e-23
3:.../Desktop/Intersect/blast - enolase/1916933.bs    4.9e-20
3:.../Desktop/Intersect/blast - enolase/7436880.bs    4.6e-14
3:.../Desktop/Intersect/blast - enolase/3549097.bs    5.0e-10
1:...g/Desktop/Intersect/blast - enolase/126731.bs       0.20
2:...Desktop/Intersect/blast - enolase/15805085.bs       0.99

gi|13774515|gb|AAK38 (AF353515) enolase [...         Score:  7
--Query File---------------------------------------------Prob
3:...Desktop/Intersect/blast - enolase/16130686.bs   5.4e-106
3:...gg/Desktop/Intersect/blast - enolase/85925.bs    1.1e-44
3:.../Desktop/Intersect/blast - enolase/1916933.bs    1.4e-27
3:.../Desktop/Intersect/blast - enolase/7436880.bs    2.3e-13
3:.../Desktop/Intersect/blast - enolase/3549097.bs    4.9e-13
1:...g/Desktop/Intersect/blast - enolase/126731.bs      0.992
2:...Desktop/Intersect/blast - enolase/15805085.bs      0.999

gi|13124708|sp|O2913 Enolase (2-phosphoglycera...    Score:  7
--Query File---------------------------------------------Prob
3:...Desktop/Intersect/blast - enolase/16130686.bs    1.4e-68
3:...gg/Desktop/Intersect/blast - enolase/85925.bs    1.1e-14
3:.../Desktop/Intersect/blast - enolase/1916933.bs    1.7e-06
3:.../Desktop/Intersect/blast - enolase/3549097.bs    9.8e-06
2:...Desktop/Intersect/blast - enolase/15805085.bs    0.00059
1:...Desktop/Intersect/blast - enolase/19704907.bs      0.013
3:.../Desktop/Intersect/blast - enolase/7436880.bs       0.60
```

Figure 1. The Intersect interface. At the bottom left are three sets of BLAST output files grouped by the family membership of the query sequences. All query sequences were members of the enolase superfamily (Babbitt et al., 1996), and were members of either the mandelate racemase (blue), muconate lactonizing enzyme (green), or enolase (pink) subfamily. The upper right panel displays a color-coded clickable tab for each non-empty overlap between the search results or for each set of search results for a single subfamily. The

upper left panel shows the Venn diagram for the same data. Clicking on either the tab or the corresponding region on the Venn diagram brings up the inset window in which information about the sequences populating the region is displayed.

REFERENCES

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

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.

Babbitt, P.C., Hasson, M.S., Wedekind, J.E., Palmer, D.R., Barrett, W.C., Reed, G.H., Rayment, I., Ringe, D., Kenyon, G.L., and Gerlt, J.A. (1996) The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry*, **35**, 16489-501.

Eddy, S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361-5.

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.*, **85**, 2444-8.

Pegg, S.C. and Babbitt, P.C. (1999) Shotgun: getting more from sequence similarity searches. *Bioinformatics*, **15**, 729-40.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S., and Haussler, D. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, **12**, 327-45.

Xu, L., K. Resing, et al. (1999). "Evidence that pcpA encodes 2,6-dichlorohydroquinone dioxygenase, the ring cleavage enzyme required for pentachlorophenol degradation in Sphingomonas chlorophenolica Strain ATCC 39723." Biochem. **38**: 7659-7669.

APPENDIX 3

Useful python programs and scripts

INTRODUCTION

This section of the appendix includes python scripts which may be especially useful to others. The first three programs are generally used to manipulate large numbers of sequences: unique.py, correctTfas.py and Find_and_Align.py. The next three programs deal primarily with building HMMs. Not that building HMMs is difficult, it isn't. What these programs allow you to do is to generate HMMs on protein subsequences. For example, truncateFasta.py will truncate a concatenated fasta file given start and stop values. Given this new truncated fasta file, one can run makehmms.py, which will build and calibrate the hmm file automatically. Finally, hmmsearch.py performs multiple hmmsearches so you don't have to wait to start the next round. The next two programs are designed to help in populating the structure function linkage database (SFLD). Readme files are also included for both makeEFD.py and makeEFD-TinyXML.py. The final two utilities are pdb_remove_h.py, which does just that (removes hydrogens from pdb files), and SAB.py, which stands for Split and Blast. Split and Blast has not been thoroughly tested. I wrote this to try to find regions of high homology in malarial proteins which have low sequence identity to other organisms.

```
#!/sw/bin/python


##########################################################################
#
# unique.py
#
# Generates a comma-delimited nonredundant list given a file full of
# redundants!
# so there.
# fine.
# go ahead.
# use me.
# check your python path.
# WRPN
# 08/19/03
#
##########################################################################

import sys, string

##########################################################################
# Retrieves the name of the input file from the command line, or if the
# file is
# not specified, prints an error message and exits the program
##########################################################################
def getInFileNames(argv):
   """hey this is from shoshana!"""

   errorString = "Usage: unique.py <inputFile> [-o<outputFile>]"

   outFileName = "unique.gi"

   if len(argv) < 2:
      print errorString
      sys.exit(0)
   else:
      inFileName = argv[1]
      if len(argv) > 2 and argv[2][:2] == "-o":
         outFileName = argv[2][2:]

   return inFileName, outFileName


##########################################################################
# Opens a file and creates a dictionary.
# Places in dict.
##########################################################################
def makeDict(dictFile):

   dict = {}

   inFile = open(dictFile, "r")

   while 1:
      line = inFile.readline()
      if not line:
         break
      items = string.split(string.strip(line), ',')
      for item in items:
         dict[item] = 1

   return dict


###########################################################################
# Main program
###########################################################################
def unique(argv):

   inFileName, outFileName = getInFileNames(argv)

   dictionary = makeDict(inFileName)

   out = open(outFileName, "w")
   for key in dictionary.keys():
      out.write(key)
      out.write("\n")

unique(sys.argv)
```

```
#!/sw/bin/python

##########################################################################
#
# correctTfas.py
#
# I need to classify the seed sequences into families
# This program opens a concatenated tfa file and returns
# the first 80 characters of the description line for
# sequences of appropriate length. Appropriate length is
# defined as .8 to 1.2 times the searching sequence length.
#
# Author: Wally Novak
# Last Modified: 10/17/01
#
##########################################################################

import os
import sys
import string

##########################################################################
# Gets name of the input file
##########################################################################
def getInFileName(argv):

   errorString = "Usage: correctTfas.py <inputFile>"

   if len(argv) < 2:
      print errorString
      sys.exit(0)
   else:
      inFileName = argv[1]

   return inFileName


##########################################################################
# Get the first 80 chars of the desc line and write to descript.out
##########################################################################
def makeDescrip(inFileName, outFileName):

   inFile = open(inFileName, "r")
   outFile = open('temp.fix', "w")

   giList = []
   one = 0
   length = 0

   # this section populates giList with the GI numbers and the lengths of
   # the protein seqs
   while 1:
      line = inFile.readline()
      if not line:
         giList.append(length)
         break
      if line[:1] == ">":
```

```
      if length > 0:
         giList.append(length)
         if one == 0:
            goodLen = length
            one = 1
      line = string.strip(line[:80]) + '\n'
      outFile.write(line)
      line = line[4:]
      index = string.find(line, "|")
      if (index == (-1)) or (index > 10):
         index = string.find(line, " ")
      giNum = line[:index]
      giList.append(giNum)
      length = 0
   else:
      length += (len(line) -1)
      outFile.write(line)
      continue

outFile.close ()
inFile.close()

low = (.8 * goodLen)
high = (1.2 * goodLen)

badList = []
i = 0

# this section populates badList with sequences of incorrect length
while i < (len(giList) - 1):
   if giList[i+1] > high or giList[i+1] < low:
      badList.append(giList[i])
   i = i + 2

inFile = open('temp.fix', "r")
outFile = open(outFileName, "w")

good = 1
line = inFile.readline()

# this section writes only the proteins of correct length to the
# .fix file
while 1:
   if not line:
      break
   if line[:1] == ">":
      linetemp = line[4:]
      index = string.find(linetemp, "|")
      if (index == (-1)) or (index > 10):
         index = string.find(linetemp, " ")
      giNum = linetemp[:index]
      if not (giNum in badList):
         good = 1
         outFile.write(line)
         line = inFile.readline()
      else:
         good = 0
```

```
            line = inFile.readline()
            continue


      elif good == 1:

         # check for bad characters
         GOODCHARS = ['Q', 'q', 'W', 'w', 'E', 'e', 'R', 'r', 'T', 't',
'Y', 'y', 'I', 'i', 'P', 'p', 'A', 'a', 'S', 's', 'D', 'd', 'F', 'f', 'G',
'g', 'H', 'h', 'K', 'k', 'L', 'l', 'C', 'c', 'V', 'v', 'N', 'n', 'M', 'm',
'\n']
         for i in line:
            if i not in GOODCHARS:
               line = string.join(string.split(line, i), "-")

         outFile.write(line)
         line = inFile.readline()

      else:
         line = inFile.readline()
         continue


##########################################################################
# Main program
##########################################################################
def getDesc(argv):

   inFileName = getInFileName(argv)

   outFileName = inFileName + '.fix'

   makeDescrip(inFileName, outFileName)

getDesc(sys.argv)
```

```
#!/sw/bin/python

##########################################################################
#
# Find_and_Align.py
# 1/09/02
# Wally Novak
#
# This script runs a set of programs commonly used to generate an
# initial set of related sequences for super/suprafamily or
# evolutionary trace analysis.
#
##########################################################################

from string import *
from time import *
import sys, os, glob
from types import *

ROUNDS = 15 # the number of PSI-BLAST rounds

##########################################################################
# Get the filename and strip the last 3 chars
##########################################################################
def getInFileName(argv):

   errorString = "Usage: Find_and_Align.py <inputFile.tfa>\nNote the input
file must have the .tfa extension and be in fasta format.\n"

   if len(argv) < 2:
      print errorString
      sys.exit(0)
   else:
      inFileName = argv[1][:-3]

   return inFileName

##########################################################################
# Runs a blast search on the sequence for the approriate number of rounds
##########################################################################
def blast(file):

   infile = file + "tfa"
   outfile = file + "blastp"

   print "\nRunning blast search on %s." % (infile)

   cmd = "blastpgp -i %s -a2 -I -j%d -b0 -o %s" % (infile, ROUNDS,
outfile)
   os.system(cmd)

   print "Blast search completed. Output saved as %s.\n" % (outfile)

   return outfile
```

```
##########################################################################
# Gets the gi numbers from the blast file. Utilizes getAllGIs.py by
# Shoshana Brown.
##########################################################################
def get_GIs(file):

   print "Retrieving GI numbers."

   cmd = "getAllGIs.py %s" % (file)
   os.system(cmd)

   print "GI numbers stored in psi10.ggi.\n"

##########################################################################
# Gets a concatenated tfa file containing all sequences. Uses the fastacmd
# program on socrates.cgl.ucsf.edu
##########################################################################
def get_fastas(file):

   outfile = file + "tfas"

   print "Retrieving sequence files."

   cmd = "fastacmd -i psi10.ggi -o %s" % (outfile)
   os.system(cmd)

   print "Sequences stored in %s.\n" % (outfile)

   return outfile

##########################################################################
# Corrects the long lines in the tfa file and removes long and short seqs.
# Uses the correctTfas.py program (written by me).
##########################################################################
def correct_tfas(file):

   outfile = file + ".fix"

   print "Correcting %s." % (file)

   cmd = "correctTfas.py %s" % (file)
   os.system(cmd)

   print "Corrected format stored in %s.\n" % (outfile)

   return outfile

##########################################################################
# Performs a clustalw alignment on the corrected tfa file.
##########################################################################
def clustal(file):

   align_out = file[:-3] + "aln"
   dnd_out = file[:-3] + "dnd"

   print "Aligning sequences in %s with clustalw." % (file)
   cmd = "clustalw %s > temp.output" % (file)
   os.system(cmd)

   print "Clustal alignment stored in %s.\nClustal dendogram stored in
%s.\n" % (align_out, dnd_out)

##########################################################################
# Removes temporary files
##########################################################################
def cleanup():

   cmd = "rm temp*"
   os.system(cmd)

##########################################################################
# Main program
##########################################################################
def Find_and_Align(argv):

   inFileName = getInFileName(argv)

   blastFile = blast(inFileName)

   get_GIs(blastFile)

   tfa_file = get_fastas(inFileName)

   tfa_corrected = correct_tfas(tfa_file)

   clustal(tfa_corrected)

   cleanup()

   print "\nFind_and_Align completed successfully.\n"

Find_and_Align(sys.argv)
```

```
#!/sw/bin/python

###########################################################################
#
# truncateFasta.py
#
# I want to truncate fasta alignment files to generate different hmm
# profiles.
#
# This program will truncate fasta alignment files given start and stop
# values.
#
############
#
# Why use it?
#
# #3: It allows you to generate hmm files on segments of your alignment.
# #2: Comes in handy for multiple domains
# and the #1 reason to use truncateFasta.py...Because everyone else is
# doing it
#
############
#
# Usage:
#
# truncateFasta.py <inputFile> <start> <stop>
#
############
#
# Author: Wally Novak
# Last Modified: 08/11/03
#
###########################################################################

import os
import sys
import string

###########################################################################
# Gets name of the input file
###########################################################################
def getInFileName(argv):

   errorString = "Usage: truncateFasta.py <inputFile> <start> <stop>"

   if len(argv) < 4:
      print errorString
      sys.exit(0)
   else:
      inFileName = argv[1]
      start = argv[2]
      stop = argv[3]

   return inFileName, start, stop
```

```
##########################################################################
# Truncate each sequence in the fasta alignment to the start and end
# values
##########################################################################
def makeDescrip(inFileName, outFileName, start, end):

   inFile = open(inFileName, "r")
   outFile = open(outFileName, "w")

   seqname = ""
   seq = ""
   stop = 0

   line = inFile.readline()

   while 1:
      if line[:1] == ">":
         seqname = line
         line = inFile.readline()
         if not line:
            break
         while line[:1] != ">":
            seq = seq + string.strip(line)
            line = inFile.readline()
            if not line:
               stop = 1
               break
         seq = seq[start:end]
         newseq = formatString(seq, 69)

         outFile.write(seqname)
         outFile.write(newseq)
         seq = ""
         seqname = ""

         if stop == 1:
            break

##########################################################################
# Inserts newlines
##########################################################################
def formatString(string, charsPerLine):

   index = 0
   newstring = ""

   while index < len(string):
      newstring = newstring + string[index:(index + charsPerLine)]
      newstring = newstring + "\n"
      index = index + charsPerLine

   return newstring

##########################################################################
# Main program
##########################################################################
def getDesc(argv):
   inFileName, start, end = getInFileName(argv)

   outFileName = inFileName + '.trun2'

   start = string.atoi(start)
   end = string.atoi(end)

   makeDescrip(inFileName, outFileName, start, end)

getDesc(sys.argv)
```

```
#!/sw/bin/python

##########################################################################
#
# makehmms.py
#
# Takes a list of aligned fasta files and builds and calibrates hmms
# for each file in the list.
#
# Author: Wally Novak
# Last Modified: 08/11/03
#
##########################################################################

import os
import sys
import string

##########################################################################
# Gets name of the input file
##########################################################################
def getInFileName(argv):

   errorString = "Usage: makehmms.py <inputFile>"

   if len(argv) < 2:
      print errorString
      sys.exit(0)
   else:
      inFileName = argv[1]

   return inFileName


##########################################################################
# Gets the filenames and creates and calibrates the hmm
##########################################################################
def makeDescrip(inFileName):

   inFile = open(inFileName, "r")

   while 1:
      line = inFile.readline()
      if not line:
         break
      line = string.strip(line)
      index = string.find(line, ".")
      if index != -1:
         out = line[:index] + ".hmm"
      else:
         out = line + ".hmm"
         print "error parsing filename"
      cmd = "hmmbuild %s %s" % (out, line)
      os.system(cmd)
      cmd2 = "hmmcalibrate %s" % out
      os.system(cmd2)
   inFile.close()

###########################################################################
# Main program
###########################################################################
def getDesc(argv):

   inFileName = getInFileName(argv)

   makeDescrip(inFileName)

getDesc(sys.argv)
```

```
#!/sw/bin/python

##########################################################################
#
# hmmsearch.py
#
# Takes a list file of hmms and performs hmmsearches with the hmm against
# the nrdb.
#
# Author: Wally Novak
# Last Modified: 08/11/03
#
##########################################################################

import os
import sys
import string

##########################################################################
# Gets name of the input file
##########################################################################
def getInFileName(argv):

   errorString = "Usage: hmmsearch.py <inputFile>"

   if len(argv) < 2:
      print errorString
      sys.exit(0)
   else:
      inFileName = argv[1]

   return inFileName


##########################################################################
# Takes a list file of hmms and performs hmmsearches with the hmm against
# the nrdb.
##########################################################################
def makeDescrip(inFileName):

   inFile = open(inFileName, "r")

   while 1:
      line = inFile.readline()
      if not line:
         break
      line = string.strip(line)

      index = string.find(line, ".")
      if index != -1:
         out = line[:index] + ".out"
      else:
         out = line + ".out"
         print "error parsing filename %s" % out

      cmd = "hmmsearch %s nr > %s" % (line, out)
      #print cmd + "\n"
      os.system(cmd)

   inFile.close()


##########################################################################
# Main program
##########################################################################
def getDesc(argv):

   inFileName = getInFileName(argv)

   makeDescrip(inFileName)

getDesc(sys.argv)
```

README.makeEFD_and_makeEFD-TinyXML

Here's the deal.

makeEFD.py and makeEFD-TinyXML.py both require 1 additional argument, the xml file.

Check the python path as it is different on my mac than on socrates or your computer.

I suggest using makeEFD-TinyXML, which requires that you download the TinySeqXML file from NCBI. The tinyseqxml is not available on the first menu on ncbi's site, so first display fasta and then tinyseqxml. Download the file and use this with makeEFD-TinyXML.py.

If you want to use makeEFD.py you must download the standard xml file. BEWARE! this file may be extremely large due to sequencing projects. ie a tinyseqxml file may be 100KB, while the standard xml file may be over 1GB!!!!

#############
How to get the XML file:

Go to: http://www.ncbi.nlm.nih.gov/Entrez/

Enter all the gi #'s, comma separated in the search window and make sure protein is selected.

When the output is displayed...select display 1000 hits and type=fasta and push display.

When the output is displayed the second time...select TinySeqXML and push display.

When the output is displayed the third time...save it to a file.

Use as below.

#############
Usage:

```
makeEFD-TinyXML.py <xml file> > output
```

Output:

a tab-delimited file suitable for import into the EFD excel table for the SFLD

```
#!/sw/bin/python

##########################################################################
#
# makeEFD.py
#
# Generates and EFD file that may be imported into excel. It takes an ncbi
# xml file as input
# It also outputs the fasta sequence file
#
# WRPN
# 08/18/03
#
##########################################################################

import sys, string

from xml.sax import saxutils, handler, make_parser

# --- The ContentHandler

class ContentGenerator(handler.ContentHandler):

    def __init__(self, out = sys.stdout):
        handler.ContentHandler.__init__(self)

        self.gi = ""
        self.protname = ""
        self.species = ""
        self.length = ""
        self.seq = ""

        self._out = out

        self.isGI = 0
        self.isName = 0
        self.isSpecies = 0
        self.isLength = 0
        self.isSeq = 0
        self.inNameContent = 0

    # ContentHandler methods


    def startElement(self, name, attrs):
        if name == "Seq-id_gi": self.isGI = 1
        if name == "Seqdesc_title":
            self.isName = 1
            self.inNameContent = 1
        if name == "Org-ref_taxname": self.isSpecies = 1
        if name == "Seq-inst_length": self.isLength = 1
        if name == "NCBIeaa": self.isSeq = 1

    def endElement(self, name):
        if name == "TSeq_defline":
            self.inNameContent = 0
            self.isName = 2
        if self.isGI == 2 and self.isName == 2 and self.isSpecies == 2 and
self.isLength == 2 and self.isSeq == 2:
            self._out.write("%s\t%s\t%s\t%s\n" % (self.gi, self.protname,
self.species, self.length))

            fname = "%s.fa" % self.gi

            self.seq = formatString(self.seq, 69)

            outfile = open(fname, "w")
            outfile.write(">gi|%s| %s\n%s" % (self.gi, self.protname,
self.seq))
            outfile.close()


            self.isGI = 0
            self.isName = 0
            self.isSpecies = 0
            self.isLength = 0
            self.isSeq = 0

            self.gi = ""
            self.protname = ""
            self.species = ""
            self.length = ""
            self.seq = ""

    def characters(self, content):
        if self.isGI == 1:
            self.gi = (saxutils.escape(content))
            self.isGI = 2
        if self.isName ==1:
            self.protname = self.protname + (saxutils.escape(content))
            #self.isName = 2
        if self.isSpecies == 1:
            self.species = (saxutils.escape(content))
            self.isSpecies = 2
        if self.isLength == 1:
            self.length = (saxutils.escape(content))
            self.isLength = 2
        if self.isSeq == 1:
            self.seq = (saxutils.escape(content))
            self.isSeq = 2

    def ignorableWhitespace(self, content):
        self._out.write(content)

    def processingInstruction(self, target, data):
        self._out.write('<?%s %s?>' % (target, data))

##########################################################################
# Inserts newlines
##########################################################################
def formatString(string, charsPerLine):

   index = 0
   newstring = ""

   while index < len(string):
      newstring = newstring + string[index:(index + charsPerLine)]
      newstring = newstring + "\n"
      index = index + charsPerLine

   return newstring


# --- The main program

parser = make_parser()
parser.setContentHandler(ContentGenerator())
parser.parse(sys.argv[1])
```

```
#!/sw/bin/python

##########################################################################
#
# makeEFD-TinyXML.py
#
# Generates and EFD file that may be imported into excel. It takes an ncbi
# tinySeqxml file as input
# It also outputs the fasta sequence file
# WRPN
# 08/18/03
#
##########################################################################

import sys, string

from xml.sax import saxutils, handler, make_parser

# --- The ContentHandler

class ContentGenerator(handler.ContentHandler):

    def __init__(self, out = sys.stdout):
        handler.ContentHandler.__init__(self)

        self.gi = ""
        self.protname = ""
        self.species = ""
        self.length = ""
        self.seq = ""

        self._out = out

        self.isGI = 0
        self.isName = 0
        self.isSpecies = 0
        self.isLength = 0
        self.isSeq = 0
        self.inNameContent = 0

    # ContentHandler methods


    def startElement(self, name, attrs):
        if name == "TSeq_gi": self.isGI = 1
        if name == "TSeq_defline":
            self.isName = 1
            self.inNameContent = 1
        if name == "TSeq_orgname": self.isSpecies = 1
        if name == "TSeq_length": self.isLength = 1
        if name == "TSeq_sequence": self.isSeq = 1

    def endElement(self, name):
        if name == "TSeq_defline":
            self.inNameContent = 0
            self.isName = 2
        if self.isGI == 2 and self.isName == 2 and self.isSpecies == 2 and
self.isLength == 2 and self.isSeq == 2:
            self._out.write("%s\t%s\t%s\t%s\n" % (self.gi, self.protname,
self.species, self.length))

            fname = "%s.fa" % self.gi

            self.seq = formatString(self.seq, 69)

            outfile = open(fname, "w")
            outfile.write(">gi|%s| %s\n%s" % (self.gi, self.protname,
self.seq))
            outfile.close()


            self.isGI = 0
            self.isName = 0
            self.isSpecies = 0
            self.isLength = 0
            self.isSeq = 0

            self.gi = ""
            self.protname = ""
            self.species = ""
            self.length = ""
            self.seq = ""

    def characters(self, content):
        if self.isGI == 1:
            self.gi = (saxutils.escape(content))
            self.isGI = 2
        if self.isName == 1:
            if self.inNameContent == 1:
                self.protname = self.protname + (saxutils.escape(content))
            #else: self.isName = 2
        if self.isSpecies == 1:
            self.species = (saxutils.escape(content))
            self.isSpecies = 2
        if self.isLength == 1:
            self.length = (saxutils.escape(content))
            self.isLength = 2
        if self.isSeq == 1:
            self.seq = (saxutils.escape(content))
            self.isSeq = 2


#########################################################################
# Inserts newlines
#########################################################################
def formatString(string, charsPerLine):

   index = 0
   newstring = ""

   while index < len(string):
      newstring = newstring + string[index:(index + charsPerLine)]
      newstring = newstring + "\n"
```

```
        index = index + charsPerLine

    return newstring


# --- The main program

parser = make_parser()
parser.setContentHandler(ContentGenerator())
parser.parse(sys.argv[1])
```

```
#!/sw/bin/python

###########################################################################
#
# pdb_remove_h.py
#
# remove lines with Hydrogens from a PDB
#
# Wally
# 1-13-04
#
###########################################################################

import os
import sys
import string

###########################################################################
# Retrieves the name of the input file from the command line, or if the
# file is not specified, prints an error message and exits the program
###########################################################################
def getInFileNames(argv):

   errorString = "Usage: pdb_remove_h.py <inputFile>"

   if len(argv) < 2:
      print errorString
      sys.exit(0)
   else:
      inFileName = argv[1]

   return inFileName


###########################################################################
# Retrieves each unique GI number from the file specified by inFileName,
# and adds it to the list giList.  Returns giList
###########################################################################
def findH(inFileName):

   inFile = open(inFileName, "r")
   outfile = inFileName[:-4] + "-h.pdb"
   out = open(outfile, "w")

   while 1:
      line = inFile.readline()
      if not line:
         break
      if line[:4] == "ATOM" and line[13] == "H":
         print "Removing line: %s" % line
      else:
         out.write(line)

   inFile.close()
   out.close()
```

```
##########################################################################
# Main program
##########################################################################
def pdb_remove_h(argv):

   inFileName = getInFileNames(argv)

   findH(inFileName)

pdb_remove_h(sys.argv)
```

```
#!/usr/bin/python

##########################################################################
#
# SAB.py Split And Blast
# 06/06/03
# Wally Novak
#
# Here we want to split a sequence into smaller chunks and blast them to
# try and determine regions of higher homology. We also want to determine
# secondary structure using something like FSSP.
#
##########################################################################

from string import *
from time import *
import sys, os, glob
from types import *

SIZE = 30
ROUNDS = 10
DATABASE = "nr" #"Pfa3D7"
MATRIX = "BLOSUM62"
hist = ""

seqList = []

class SEQ:
   def __init__(self):
      seq = ""
      start = 0
      end = 0
      outFile = ""

##########################################################################
# getInFileName - Get the filename and strip the last 3 chars
##########################################################################
def getInFileName(argv):

   errorString = "Usage: SAB.py <inputFile.tfa>\nNote the input file must
have the .tfa extension and be in fasta format.\n"

   if len(argv) < 2:
      print errorString
      sys.exit(0)
   else:
      inFileName = argv[1]

   return inFileName

##########################################################################
# splitSeq - Split the sequence into blastable chunks
##########################################################################
def splitSeq(inFileName):

   inFile = open(inFileName, "r")
   fullSeq = ""

   while 1:
      line = inFile.readline()
      if not line:
         break
      if line[:1] == ">":
         seqName = line
      else:
         fullSeq = fullSeq + strip(line)


   inFile.close()

   #print fullSeq

   #print len(fullSeq)

   bean = 0

   while (bean + SIZE) < len(fullSeq):

      newSeq = SEQ()

      newSeq.seq = fullSeq[bean:(bean + SIZE)]
      newSeq.start = bean
      newSeq.end = (bean + SIZE)
      newSeq.outFile = 'out%d.tfa' % newSeq.start

      outFile = open(newSeq.outFile, "w")
      descriptor = '>%s\n' % newSeq.outFile
      outFile.write(descriptor)
      outFile.write(newSeq.seq)
      outFile.close()

      seqList.append(newSeq)

      bean = bean + (SIZE/3)

##########################################################################
# blast - Run a blast search
##########################################################################
def blast(file):

   infile = file
   outfile = file[:-3] + "blastp"

   print "\nRunning blast search on %s." % (infile)

   cmd = "blastpgp -d %s -M %s -i %s -I -j%d -b0 -o %s" % (DATABASE,
MATRIX, infile, ROUNDS, outfile)
   os.system(cmd)

   print "Blast search completed. Output saved as %s.\n" % (outfile)

   parseOutput(outfile)
```

```
##########################################################################
# parseOutput - gets the number of hits per blast search
##########################################################################
def parseOutput(file):

   inFile = open(file, "r")
   global hist
   count = 0

   while 1:
      line = inFile.readline()
      if not line:
         break
      if line[:3] == "Pfa":
         count = count + 1

   inFile.close()

   index = find(file, ".")
   number = file[3:index]
   output = "%s\t%d\n" % (number, count)
   hist = hist + output

##########################################################################
# Main program
##########################################################################
def SAB(argv):

   inFileName = getInFileName(argv)
   global hist
   splitSeq(inFileName)

   for seq in seqList:
      blast(seq.outFile)

   outFile = open("histogram.txt", "w")
   outFile.write(hist)
   outFile.close()

   print "\nSAB completed successfully.\n"

SAB(sys.argv)
```