J|A|C|S
ARTICLES

Published on Web 11/04/2006

FAST-NMR: Functional Annotation Screening Technology Using NMR Spectroscopy

Kelly A. Mercier,[†] Michael Baran,[§] Viswanathan Ramanathan,[‡] Peter Revesz,[‡]
Rong Xiao,[§] Gaetano T. Montelione,[§] and Robert Powers*[,†]

Contribution from the Department of Chemistry and Department of Computer Science and Engineering, University of Nebraska−Lincoln, Lincoln, Nebraska 68588, and Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey 08854

Received July 27, 2006; E-mail: rpowers3@unl.edu

Abstract: An abundance of protein structures emerging from structural genomics and the Protein Structure Initiative (PSI) are not amenable to ready functional assignment because of a lack of sequence and structural homology to proteins of known function. We describe a high-throughput NMR methodology (FAST-NMR) to annotate the biological function of novel proteins through the structural and sequence analysis of protein−ligand interactions. This is based on basic tenets of biochemistry where proteins with similar functions will have similar active sites and exhibit similar ligand binding interactions, despite global differences in sequence and structure. Protein−ligand interactions are determined through a tiered NMR screen using a library composed of compounds with known biological activity. A rapid co-structure is determined by combining the experimental identification of the ligand binding site from NMR chemical shift perturbations with the protein−ligand docking program AutoDock. Our CPASS (Comparison of Protein Active Site Structures) software and database are then used to compare this active site with proteins of known function. The methodology is demonstrated using unannotated protein SAV1430 from *Staphylococcus aureus*.

Introduction

The availability of the human genome sequence has just begun to provide a wealth of information on cell biology, development, evolution, and physiology that is expanding our understanding of disease and making beneficial contributions to medicine and human health.[1] Contributing to this expanding knowledge base is the extensive amount of protein structures emerging from structural genomics and the Protein Structure Initiative (PSI).[2]

To date, 261 genomes have been completely sequenced with ∼1158 ongoing genome projects[3] where 30 000−90 000 proteins are predicted to be from the human genome alone.[1] Unfortunately, ∼60% of recent structures emerging from structural genomics correspond to folds that provide little insight into function (Figure 1). The prospect of obtaining functional information for this extensive collection of hypothetical proteins by traditional biochemical approaches presents an extremely overwhelming and daunting task.

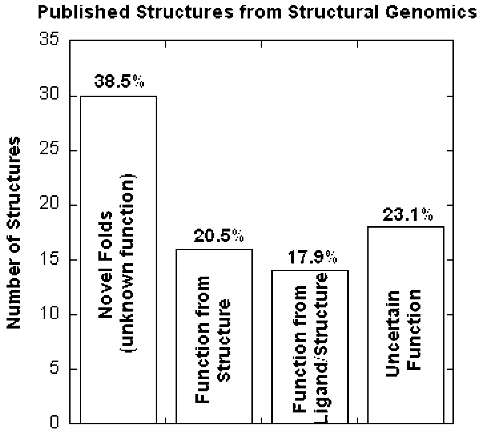


Figure 1. Functional information for hypothetical proteins accumulated from the literature can be classified into four general categories. Functional identification was obtained for ∼40% of the proteins where ligand binding played a key role.

A fundamental component to our understanding of the biological role of a protein is through the identification of its functional ligand(s), the identification of its active site or binding site, and the corresponding determination of a structure for this complex. The active sites of proteins interact with unique and specific targets, as is well-documented in numerous metabolic and signaling pathways.[4] Promiscuity of binding is inherently detrimental to the overall biological process. The comparison and prediction of putative ligand binding sites from both

[†] Department of Chemistry, University of Nebraska−Lincoln.

[‡] Department of Computer Science and Engineering, University of Nebraska−Lincoln.

[§] Rutgers University.

(1) Venter, C.; Adams, M. D.; et al. *Science* **2001**, *297*, 1304.
(2) Burley, S. K. *Nat. Struct. Biol.* **2000**, *7*, 932.
(3) Frishman, D.; Mokrejs, M.; Kosykh, D.; Kastenmuller, G.; Kolesov, G.; Zubrzycki, I.; Gruber, C.; Geier, B.; Kaps, A.; Albermann, K.; Volz, A.; Wagner, C.; Fellenberg, M.; Heumann, K.; Mewes, H.-W. *Nucleic Acids Res.* **2003**, *31*, 207.
(4) Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. *Nucleic Acids Res.* **2004**, *32*, D277.

10.1021/ja0651759 CCC: $33.50 © 2006 American Chemical Society

structural and sequence information has proven to be a valuable approach for functional analysis of proteins.[5] Distant evolutionary relationships between enzymes with no sequence similarity have been identified on the basis of the comparison of conserved active site structures.[6] Similarly, a number of hypothetical proteins determined from structural genomics have had a function attributed to them by the serendipitous observation of a bound ligand in the resulting structure.[7,8]

We have established the functional annotation screening technology using NMR spectroscopy (FAST-NMR) to provide initial functional information for proteins originating from structural genomics that lack functional information. By experimentally identifying ligands that have an associated biological function that bind to hypothetical proteins, identifying the protein's active site from this interaction, and determining a corresponding co-structure of this protein−ligand complex, it is possible to readily infer a potential function for these novel proteins. Functional annotation will emerge from a combined bioinformatics approach that incorporates the identity of the functional ligands that bind the protein, the comparison of the ligand-defined active site identified from FAST-NMR with structures of protein−ligand complexes for proteins of known function using our CPASS software and database,[9] and other readily available methodologies and software.

Analysis of genome sequences for structural genomics target selection has indicated that the typical protein domain length is ∼100 residues, where the majority of larger domains fall in the range of >160−300 residues.[10] This indicates that a significant percentage of the proteome is amenable to analysis by our FAST-NMR methodology,[11] where further progress in NMR technology may continue to extend the range of viable protein targets.[12] Recent statistical analysis indicates that ∼20−40% of protein structures determined from structural genomics may be amenable to analysis by NMR,[13,14] further supporting the potential general utility of FAST-NMR. Additional proteins may also become available for analysis when secondary efforts to optimize conditions are applied outside the practical limits imposed by the high-throughput structure determination process.[15] This paper describes our development of the FAST-NMR screen and the functional assignment of unannotated protein SAV1430 from *Staphylococcus aureus*, a typical target of the Northeast Structural Genomics (NESG) consortium.

Experimental Section

Functional Chemical Screening Library. To date, our screening library is composed of 414 compounds with known biological activity in a total of 113 mixtures comprising 3−4 compounds.[16] Reference NMR spectra are collected for both the individual compounds and the mixtures. These spectra provide NMR assignments for the compounds, while verifying compound solubility, compatibility, and the presence of distinct NMR resonances attributed to each compound in the mixture to avoid deconvolution. Identical chemical shifts, coupling patterns, and line-widths between the mixture and individual NMR spectra verify that no interaction is occurring between the compounds and that the compounds are equally soluble and stable in the mixture. All of the compounds have been individually weighed, dissolved to a concentration of 20 mM in D_2O or d_6-DMSO (Sigma Aldrich, St. Louis, MO), and stored in standard 2 mL 96-well plates at −80 °C.

1D NMR Line-Broadening Experiments. The 1D NMR line-broadening samples consist of 100 μM of each compound and 25 μM protein in a 20 mM D-bis-Tris buffer with 11.1 μM TMSP as a NMR reference in 100% D_2O at pH 7.0 (uncorrected). The NMR spectra were collected on a Bruker 500 MHz Avance spectrometer equipped with a triple-resonance, Z-axis gradient cryoprobe, a BACS-120 sample changer, and Icon NMR software for automated data collection. ^{1}H NMR spectra were collected with 128 transients at 298 K with solvent presaturation of the residual HDO, a sweep-width of 6009 Hz and 32K data points, and a total acquisition time of 6 min. The NMR spectra were processed automatically using a macro in the ACD/1D NMR manager. The NMR data were Fourier transformed, zero-filled, phased, and baseline corrected. Each spectrum was referenced with the TMSP peak set to 0.0 ppm and peak-picked. Compounds were identified as binding to *S. aureus* protein SAV1430 by a visual comparison of the free ligand and protein−ligand NMR spectra. The entire functional chemical library composed of 113 mixtures containing 414 compounds was used in the 1D NMR experiments. Software for the automated analysis of 1D NMR line broadening data is currently being developed.[17]

2D ^{1}H−^{15}N HSQC NMR Experiments. Twenty-two 2D ^{1}H−^{15}N HSQC were collected with 16 transients at 298 K with a sweep-width of 6009 Hz and 1K data points in the direct dimension and 1612 Hz and 256 data points in the indirect dimension for a total acquisition time of 2.5 h. Each NMR sample consists of 100 μM protein and 400 μM compound in a bis-Tris buffer in 95% H_2O, 5% D_2O at pH 7.0 (uncorrected). The spectra were processed using NMRPipe on a Linux Workstation.[18] Chemical shift differences were identified by comparing a reference 2D ^{1}H−^{15}N HSQC spectrum of the free protein to the spectrum of the compound−protein samples. Chemical shift perturbations were then mapped onto the surface of the SAV1430 structure (PDB ID: 1PQX) and visualized using VMD-XPLOR.[19]

Rapid Determination of Protein−Ligand Co-structures. AutoDock was used to obtain co-structures, binding, and docking energies for all of the 21 compounds identified to bind SAV1430 from the NMR experiments.[20] Chemical shift changes identified a consensus binding site comprising residues I6−P10, T14−K16, and I61−V63. A grid was manually defined within AutoDock to encompass this experimentally determined ligand binding site by adjusting the *x*, *y*, and *z* coordinates for the center of the grid box to position the grid in the binding pocket. The grid size is determined by the number of points in the *x*, *y*, and *z* dimensions and is just visibly large enough to fit the entire ligand.

[5] Campbell, S. J.; Gold, N. D.; Jackson, R. M.; Westhead, D. R. *Curr. Opin. Struct. Biol.* **2003**, *13*, 389.
[6] Hasson, M. S.; Schlichting, I.; Moulai, J.; Taylor, K.; Barrett, W.; Kenyon, G. L.; Babbitt, P. C.; Gerlt, J. A.; Petsko, G. A.; Ringe, D. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 10396.
[7] Parsons, J. F.; Lim, K.; Tempczyk, A.; Krajewski, W.; Eisenstein, E.; Herzberg, O. *Proteins* **2002**, *46*, 393.
[8] Eswaramoorthy, S.; Gerchman, S.; Graziano, V.; Kycia, H.; Studier, F. W.; Swaminathan, S. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2003**, *59*, 127.
[9] Powers, R.; Copeland, J.; Germer, K.; Mercier, K. A.; Ramanathan, V.; Revesz, P. *Proteins* **2006**, *65*, 124.
[10] Liu, J.; Hegyi, H.; Acton, T. B.; Montelione, G. T.; Rost, B. *Proteins* **2004**, *56*, 188.
[11] Kanelis, V.; Forman-Kay, J. D.; Kay, L. E. *IUBMB Life* **2001**, *52*, 291.
[12] Tugarinov, V.; Kay, L. E. *ChemBioChem* **2005**, *6*, 1567.
[13] Snyder, D. A.; Chen, Y.; Denissova, N. G.; Acton, T.; Aramini, J. M.; Ciano, M.; Karlin, R.; Liu, J.; Manor, P.; Rajan, P. A.; Rossi, P.; Swapna, G. V. T.; Xiao, R.; Rost, B.; Hunt, J.; Montelione, G. T. *J. Am. Chem. Soc.* **2005**, *127*, 16505.
[14] Yee, A. A.; Savchenko, A.; Ignachenko, A.; Lukin, J.; Xu, X.; Skarina, T.; Evdokimova, E.; Liu, C. S.; Semesi, A.; Guido, V.; Edwards, A. M.; Arrowsmith, C. H. *J. Am. Chem. Soc.* **2005**, *127*, 16512.
[15] Liu, Z.-J.; Shah, A. K.; Habel, J. E.; Ng, J. D.; Kataeva, I.; Xu, H.; Horanyi, P.; Yang, H.; Chang, J.; Zhao, M.; Huang, L.; Chang, S.; Tempel, W.; Chen, L.; Zhou, W.; Lee, D.; Lin, D.; Zhang, H.; Newton, M. G.; Rose, J.; Wang, B.-C. *J. Struct. Funct. Genomics* **2005**, *6*, 121.
[16] Mercier, K. A.; Germer, K.; Powers, R. *Comb. Chem. High Throughput Screening* **2006**, *9*, 515.
[17] Ramanathan, V.; Mercier, K.; Powers, R.; Revesz, P. *Using Databases and Computational Techniques to Infer the Function of Novel Proteins*; IEEE International Conference on Electro Information Technology: Lincoln, NE, 2005.
[18] Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277.
[19] Schwieters, C. D.; Clore, G. M. *J. Magn. Reson.* **2001**, *149*, 239.
[20] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639.

CPASS Database and Software for Functional Annotation. Comparison of Protein Active Site Structures (CPASS) database and software were developed to aid in the functional annotation of hypothetical proteins by using a protein−ligand co-structure.[9] The CPASS database contains ∼26 000 ligand-defined active sites identified from structures deposited in the PDB. CPASS returns a similarity score (0−100%) and an interactive 3D graphically display of the structural alignment for each active site comparison. The CPASS software runs on 16-node Beowulf Linux cluster with a simple web-based interface. Each comparison averages ∼40 s requiring ∼18 h to complete a comparison against the entire database.

Bioinformatics Analysis of SAV1430. The SAV1430 PDB structure (http://www.ebi.ac.uk/dali/Interactive.html) was uploaded to the DALI web server to identify homologous 3D structures.[21] The Dali output contains a list of aligned structures and corresponding Z-scores, where a higher Z-score indicates a better 3D fit. Identical structures yield a Z-score of ∼30, similar structures a score of ∼15, where scores below 2−3 are generally insignificant and mean the structures are dissimilar. Dali analysis of SAV1430 only identified structures with Z-scores < 3, which included a number of hypothetical proteins of unknown functions or ferredoxin-like folds.

A web server (http://consurf.tau.ac.il/) interface for the ConSurf software enables a rapid identification of conserved residues by comparison with structural homologues of known functions and generating a phylogenic tree.[22] ConSurf outputs an amino-acid conservation score for each residue in the SAV1430 structure, where the highest score represents the highest evolutionary conserved residue. The scores are normalized such that the average score is zero. The ConSurf amino acids with high scores were color-coded onto the SAV1430 surface using VMD-XPLOR.

Rosetta Stone analysis (http://www.igs.cnrs-mrs.fr/FusionDB/main.html) of SAV1430 was done using FusionDB, a database of bacterial and archael gene fusion events.[23] Submitting the protein sequence for SAV1430 indicated a gene fusion for proteins SAV1430 and SAV0936 based on hypothetical proteins Q92GV4 and the Q9ZCQ2 in *Rickettsia conorii* and *Rickettsia proazeki*, respectively.[24] The quality of the predicted fusion event is measured by the separation index where a value >0.6 is indicative of a real gene fusion event. A separation index score of 0.79 was determined for the predicted fusion of genes SAV1430 and SAV0936.

BLAST,[25] FASTA,[26] and ClustalW[27] were used with the SAV1430 and SAV0936 sequences to identify bacterial proteins with homology to both SAV1430 and SAV0936. A prevalence of hypothetical proteins was identified from the sequence alignments, but a mixture of small (∼80aa) and large (∼190aa) NifU-like proteins corresponding to the C-terminal domain of the multidomain NifU protein was also observed. Specifically, the sequence alignment of SAV1430 and SAV0936 against the C-terminal NifU domain from *Brucella melitensis* yielded 30% and 47% sequence identity, respectively.

Results and Discussion

Description of the FAST-NMR Methodology. The overall protocol for the FAST-NMR assay is illustrated in Figure 2. There are three major components to the process: (i) identify functional ligands that bind the protein, (ii) use the ligands to determine protein−ligand co-structures, and (iii) use the co-structures with bioinformatics to infer function. An NMR-based screen is used to determine which ligands from a chemical library bind a hypothetical protein that has been identified from structural genomics. The structure and NMR assignments for the hypothetical protein are available prior to initiating the screen. The chemical library that is screened in the FAST-NMR assay is composed of small molecules with defined functional and biological activity.[16] This library is composed of amino acids, carbohydrates, cofactors, fatty acids, hormones, inhibitors, known drugs, metabolites, neurotransmitters, nucleotides, substrates, and vitamins.

The tiered NMR screening assay minimizes valuable resources, such as protein samples and NMR instrument time, while increasing throughput. To further improve efficiency, ligands are screened as mixtures of 3−4 compounds based on a statistical analysis to optimize NMR-based screens.[28] The mixtures were designed to avoid deconvolution and maximize structural and functional diversity.

The first one-dimensional (1D) ^{1}H line-broadening NMR experiment is most suitable for screening a larger number of compounds using minimal resources while providing preliminary binding information. A protein−ligand binding interaction is identified by a change in line-width or complete disappearance of the ligands ^{1}H NMR resonances in the presence of the protein (Figure 3a,b). Effectively, a bound ligand acquires the broad NMR line-widths of the large MW protein.

The second 2D ^{1}H−^{15}N HSQC NMR experiment is only conducted on positive “hits” from the first NMR experiment, because it is more resource intensive. This experiment monitors NMR chemical shifts for the backbone amide resonances for each amino-acid residue in the protein, which has been previously assigned during the determination of the protein’s NMR structure.[29] These backbone amide chemical shifts are sensitive to the local environment and change proportionally to the residues’ proximity to a bound ligand. The residues that incur a chemical shift change in the presence of the bound ligand can be mapped onto the protein surface to identify the ligand binding site (Figure 3d). Nonspecific binders will exhibit a random distribution or complete lack of chemical shift changes.

A protein−ligand co-structure is rapidly generated by combining this experimentally determined ligand binding site with AutoDock. A grid (Figure 3d) is used to direct AutoDock to dock the ligand in this NMR binding site and determine the lowest energy conformation for the complex. Thus, AutoDock permits an initial structure of the protein−ligand complex to be calculated in minutes. The AutoDock-derived co-structure provides a complete and detailed description of the ligand binding site. The chemical-shift perturbation mapping of the ligand binding site is generally incomplete because of chemical shift overlap. Moreover, a lack of a perturbation in the NH NMR resonances may occur, because the primary interaction between the ligand and the amino-acid residue is with the side-chain instead of the backbone. Also, amino-acid residues not in direct contact with the ligand may incur a chemical shift perturbation because of local structural changes. Additionally, the NMR

[21] Dietmann, S.; Park, J.; Notredame, C.; Heger, A.; Lappe, M.; Holm, L. *Nucleic Acids Res.* **2001**, *29*, 55.

[22] Glaser, F.; Pupko, T.; Paz, I.; Bell, R. E.; Bechor-Shental, D.; Martz, E.; Ben-Tal, N. *Bioinformatics* **2003**, *19*, 163.

[23] Suhre, K.; Claverie, J.-M. *Nucleic Acids Res.* **2004**, *32*, D273.

[24] Renesto, P.; Ogata, H.; Audic, S.; Claverie, J.-Ml; Raoult, D. *FEMS Microbiol. Rev.* **2005**, *29*, 99.

[25] Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J. Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389.

[26] Pearson, W. R. *Methods in Molecular Biology (Totowa, New Jersey)* **2000**, *132*, 185.

[27] Thompson, J. D.; Higgins, D. G.; Gibson, T. J. *Nucleic Acids. Res.* **1994**, *22*, 4673.

[28] Mercier, K. A.; Powers, R. *J. Biomol. NMR* **2005**, *31*, 243.

[29] Ferentz, A. E.; Wagner, G. *Q. Rev. Biophys.* **2000**, *33*, 29.

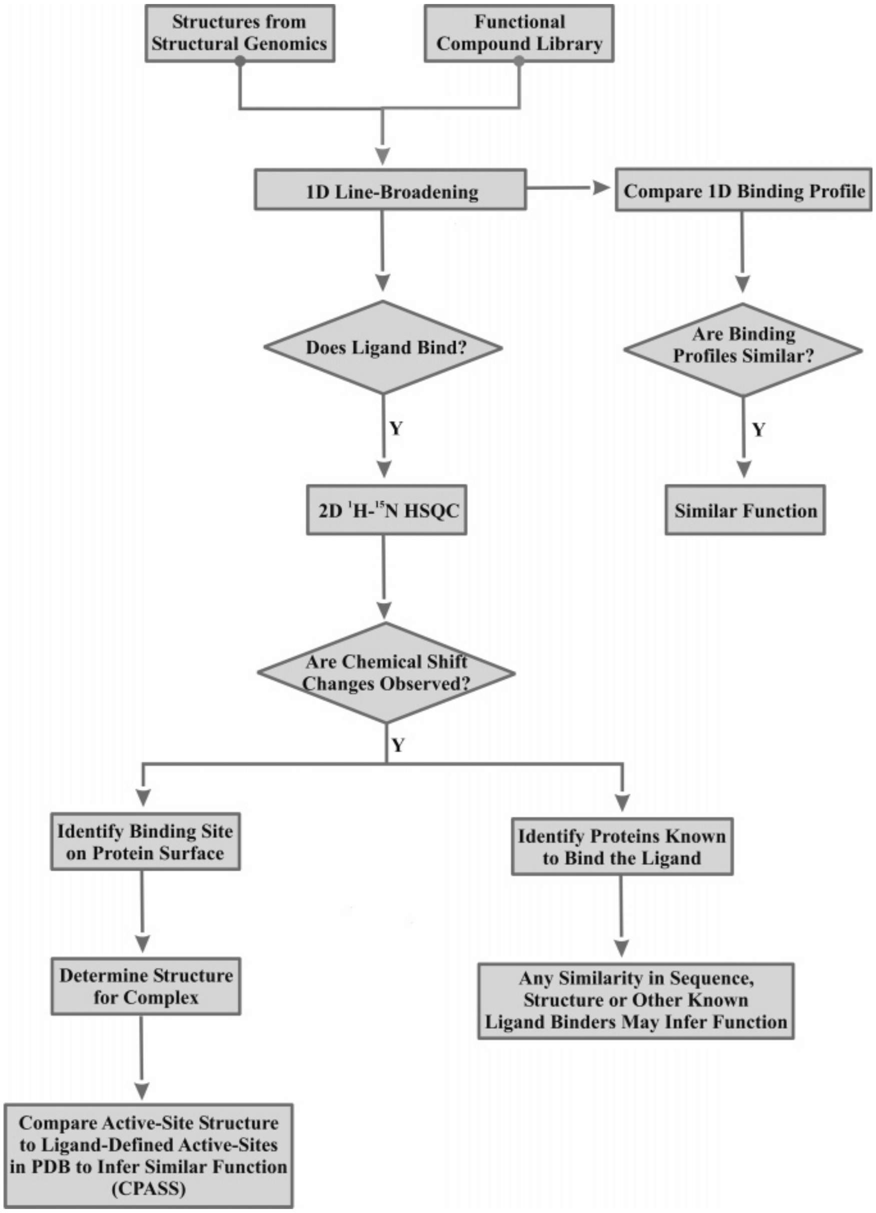

Structures from Structural Genomics
Functional Compound Library
1D Line-Broadening
Compare 1D Binding Profile
Does Ligand Bind?
Are Binding Profiles Similar?
Y
Y
2D 1H-15N HSQC
Similar Function
Are Chemical Shift Changes Observed?
Y
Identify Binding Site on Protein Surface
Identify Proteins Known to Bind the Ligand
Determine Structure for Complex
Any Similarity in Sequence, Structure or Other Known Ligand Binders May Infer Function
Compare Active-Site Structure to Ligand-Defined Active-Sites in PDB to Infer Similar Function (CPASS)

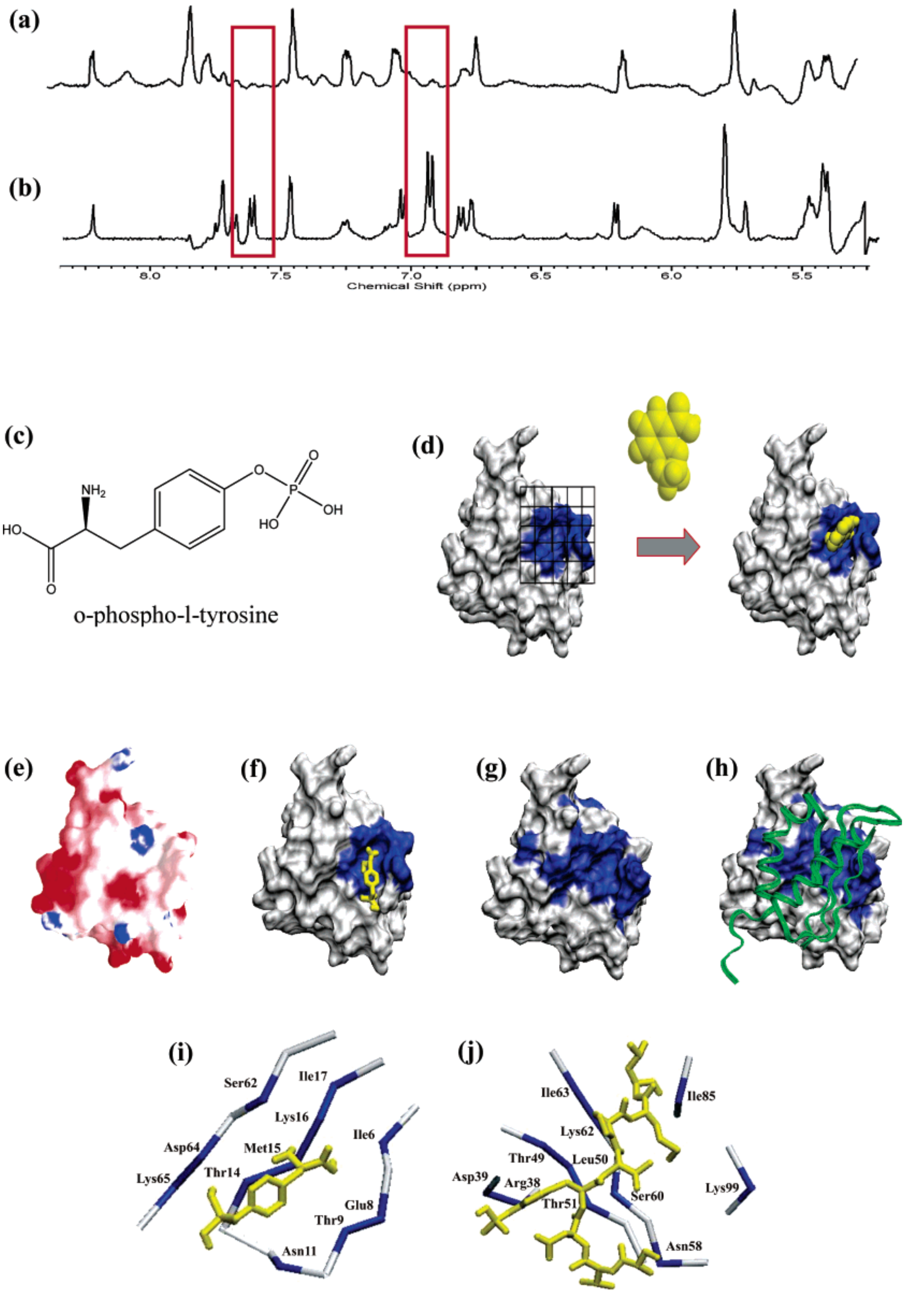
Figure 2. Flow chart for the FAST-NMR screen.

chemical shift perturbations only identify residues involved in an interaction with the ligand. The AutoDock co-structure also identifies relative distances between the ligand and the amino-acid residues.

The Comparison of Protein Active Site Structures (CPASS) database and software enable the comparison of experimentally identified ligand binding sites to infer biological function.[9] Proteins that exhibit a high similarity in the structural characteristics of their active sites would suggest a corresponding similarity in function. This is simply a further refinement of generally accepted paradigms where global sequence and structure homology are routinely used to annotate function.[30] The AutoDock structure of the protein−ligand complex is used to generate a ligand-defined active site, which comprises any residue in the protein that contains at least one atom that is ≤6 Å from any atom in the ligand. The entire PDB database has been similarly analyzed where each protein that contains a bound ligand had its corresponding ligand-defined active site extracted as part of the CPASS database. The ligand-defined active site identified from the FAST-NMR assay using AutoDock is then compared with the CPASS database using the CPASS software. CPASS iteratively compares the ligand-defined active site with the database to maximize both the sequence and the structural alignments and reports a similarity score ranging from 0 to 100%. Protein(s) of known function that exhibit a high similarity with the unannotated protein's active site would suggest potential functions for the unknown protein. It is important to note that the alignment of the ligand-defined active sites does not include the ligands, but relative distances between the ligands and the active site residues are used in the CPASS scoring function. The identity of the ligand(s) that bind the unannotated protein in conjunction with other available bioinformatics tools provide additional insight into possible functions for the unannotated protein.

Validation of FAST-NMR Using *Staphylococcus aureus* Protein SAV1430. *Staphylococcus aureus* (*S. aureus*) protein

[30] Whisstock, j. C.; Lesk, A. M. *Q. Rev. Biophys.* **2003**, *36*, 307.



SAV1430	I6	E8	T9	N11	T14	M15	K16	I17	S62	D64	K65
PDB ID:1OO4	I85	K99	S60	N58	T51	L50	K62	I63	T49	D39	R38

Figure 3. FAST-NMR analysis of SAV1430. Example of 1D NMR line-broadening spectra obtained in the SAV1430 screen. Expanded NMR region corresponding to the aromatic residues (a) in the presence of SAV1430 and (b) the free mixture, illustrating the binding of SAV1430 by changes in observed NMR line-widths. (c) Chemical structure of *O*-phospho-L-tyrosine shown to bind SAV1430. (d) Illustration of the use of the NMR-defined ligand binding site to direct a protein−ligand co-structure determination with AutoDock.[20] SAV1430 surface where residues that incur a chemical shift change are colored blue. (e) GRASP[33] electropotential surface where positive and negative regions are blue and red, respectively. (f) NMR-based ligand-defined SAV1430 binding site where residues that incur a chemical shift change upon binding pTyr are colored blue. (g) Mapping of functionally conserved residues (colored blue) identified by ConSurf[22] on the SAV1430 surface. (h) Molecular model of SAV1430−SE0630 (structural homologue of SAV0936) complex determined by Hex.[42] SE0630 is shown as a ribbon. Sequence and structural alignment of (i) SAV1430−pTyr binding site to the (j) Src SH2 domain's pTyr-peptide binding site (PDB ID: 1OO4). The sequence alignment is shown below the structure where the aligned residues are colored blue. The pTyr and pTyr-peptide ligands are colored yellow.

SAV1430 is an example of a protein of unknown biological function that is a typical target of NESG. Dali analysis suggests SAV1430 has a topology similar to ferredoxin-like folds; however, the Z-score of <3 is not significant.[21] The only significant sequence homology was to hypothetical proteins. Taken together, bioinformatic analysis alone failed to assign a reliable function to SAV1430. The NMR resonance assignments and corresponding solution structure for SAV1430 were determined by the effort at NESG.[31,32]

The 1D NMR line-broadening experiments using our entire functional chemical library identified 21 compounds that bind SAV430 (Figure 3a,b). The compounds exhibit similar structural features that are consistent with binding in an optimized active site for SAV1430 and suggestive of the natural ligand. The compounds are generally similar to *O*-phospho-L-tyrosine (pTyr). The structures tend to contain one or more aromatic rings that overlap with the phenyl ring in pTyr with one or more hydrophilic substituents that overlaps with either the amino acid or the phosphate group in pTyr (e.g., 2-amino-4-methylphenol, nicotinic acid, and acetylsalicylic acid).

These 21 compounds were further evaluated for SAV1430 binding in the second tiered 2D $^1H-^{15}N$ HSQC NMR experiments. Specifically, pTyr was identified as one of the compounds that exhibited higher chemical shift perturbations through a visual inspection of histogram plots of the chemical shift perturbations for the 21 compounds (Figure 3c). Analysis of the 2D $^1H-^{15}N$ HSQC NMR spectra implies a distinct, consensus binding site that comprises residues I6−P10, T14−K16, and I61−V63. Based on these NMR results, protein−ligand co-structures have been generated using AutoDock targeting the identified binding pocket (Figure 3d−f). This binding site corresponds to a shallow cleft on the SAV1430 surface surrounded by relatively flat structural features. A GRASP representation of the electropotential surface indicates that most of the ligand binding site is hydrophobic in nature (Figure 3e).[33] A small positive patch (Lys 16) and negative patch (Asp 64) are part of or proximal to the ligand binding site. The weak ligand binding interactions with SAV1430 in conjunction with the indistinct and generally flat hydrophobic structural characteristics of the consensus ligand binding site are strongly suggestive of a protein−protein interaction site.

ConSurf was used to identify residues in the SAV1430 structure that are conserved with homologous proteins with a known structure. These conserved residues are predicted to be functionally important to the biological activity of SAV1430 and were mapped onto the protein's surface (Figure 3g). The experimentally determined SAV1430 ligand binding site is consistent with the functionally conserved residues identified by ConSurf supporting the biological relevance of this region of SAV1430. Furthermore, the FAST-NMR defined ligand binding site corresponds to the most prominent structural feature (binding cleft) within the larger surface area described by ConSurf.

CPASS analysis of the SAV1430−pTyr binding site identified a Src SH2 domain complexed with a phosphotyrosine containing peptide, along with other SH2 domains, as a significant hit (37% similarity). Visual comparison of the two ligand binding sites clearly indicates a similarity in local structure, primarily a flat β-sheet, and sequence composition (Figure 3i,j). SH2 domains are typically part of multidomain proteins involved in cell signaling and form a protein−protein complex with a kinase after phosphorylation of a tyrosine.[25] Phosphorylations of Ser, Thr, and Tyr are also common mechanisms for regulating protein activity in bacteria.[35,36] The similarity in the characteristics of the SAV1430 and Src SH2 ligand binding sites and the fact that SAV1430 binds pTyr further supports the general proposal that SAV1430 functions by forming a protein−protein complex.

Bioinformatics Analysis of FAST-NMR Results of SAV1430. The NIF (nitrogen fixation), ISC (iron−sulfur cluster), and SUF (sulfur) [Fe−S] cluster assembly networks are essential for the viability of bacteria.[37,38] The assembly of [Fe−S] clusters is a complex, poorly understood process involving multiple proteins. [Fe−S] clusters are utilized in >100 proteins for a variety of functions. NifU is a multi-protein complex within the NIF [Fe−S] cluster assembly network that is involved in the overall biosynthesis of FeMo-cofactor. The current viewpoint is that NifU provides a scaffold for the assembly of transient [2Fe−2S] units for the formation of a [Fe−S] cluster in nitrogenase, an important component of biological nitrogen fixation. The [Fe−S] assembly scaffold is located in the N-terminus of the NifU homodimer structure, the central region of the protein contains a permanent [2Fe-2S] cluster, and the C-terminus appears to contain a second transient [Fe−S] assembly scaffold. NifU functions as a homodimer, where a number of proteins have been identified as components of the FeMo-cofactor biosynthesis pathway. NifU has been shown to form a macromolecular complex with NifS.[39] Also, the identification of two transient [Fe−S] cluster assembly scaffolds within NifU suggests target specificity between NifU and the other components of the FeMo-cofactor biosynthesis pathway. The assembly of macromolecular complexes appears to be an important component of the FeMo-cofactor biosynthesis pathway.

SAV1430 is homologous to the N-terminal domain of the *S. cerivisae* NIFU protein. Rosetta Stone analysis reveals that the C-terminal domain of *S. cervisae* NIFU is also found in a distant part of the *S. aureus* genome as gene SAV0936. This analysis suggests that SAV0936 may be a binding partner of SAV1430. To validate the FAST-NMR active site prediction, we were able to use a homologue of SAV0936, the *Staphylococcus epidermidis* protein SE0630[40,41] (95% identity), to perform in-silico docking experiments. An unbiased model of a SAV1430−

[31] Baran, M. C.; Aramini, J. M.; Huang, Y. J.; Xiao, R.; Acton, T. B.; Shih, L.-y.; Montelione, G. T. *Solution Structure Determination of the Staphylococcus Aureus Hypothetical Protein SAV1430. Northeast Structure Consortium target ZR18*; Department of Biochemistry, University of Wisconsin-Madison, 2003; BMRB accession number 5844.

[32] Baran, M. C.; Aramini, J. M.; Xiao, R.; Huang, Y. J.; Acton, T. B.; Shih, L.; Montelione, G. T. *Solution Structure of the Hypothetical Staphylococcus Aureus protein SAV1430. Northeast Structural Genomics Consortium target ZR18;* 2003; PDB ID: 1PQX.

[33] Petrey, D.; Honig, B. *Methods Enzymol.* **2003**, *374*, 492.

[34] Marengere, L. E. M.; Pawson, T. *J. Cell Sci. Suppl.* **1994**, *18*, 97.

[35] Kennelly, P. J.; Potts, M. *J. Bacteriol.* **1006**, *178*, 4759.

[36] Alzari, P. M. *Structure* **2004**, *12*, 1923.

[37] Schilke, B.; Voisin, C.; Beinert, H.; Craig, E. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 10206.

[38] Olson, J. W.; Agar, J. N.; Johnson, M. K.; Maier, R. J. *Biochemistry* **2000**, *39*, 16213.

[39] Yuvaniyama, P.; Agar, J. N.; Cash, V. L.; Johnson, M. K.; Dean, D. R. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 599.

[40] Baran, M. C.; Huang, Y. P.; Acton, T.; Xiao, R.; Montelione, G. T. *Solution Structure Of The Staphylococcus Epidermis Protein SE0936. Northeast Structural Genomics Consortium Target SeR8;* Department of Biochemistry, University of Wisconsin-Madison, 2004; BMRB accession number 6355.

[41] Baran, M. C.; Huang, Y. P.; Acton, T.; Xiao, R.; Montelione, G. T. *Solution Structure Of The Staphylococcus Epidermidis Protein SE0630. Northeast Structural Genomics Consortium Target SeR8;* PDB ID: 1CHJ.

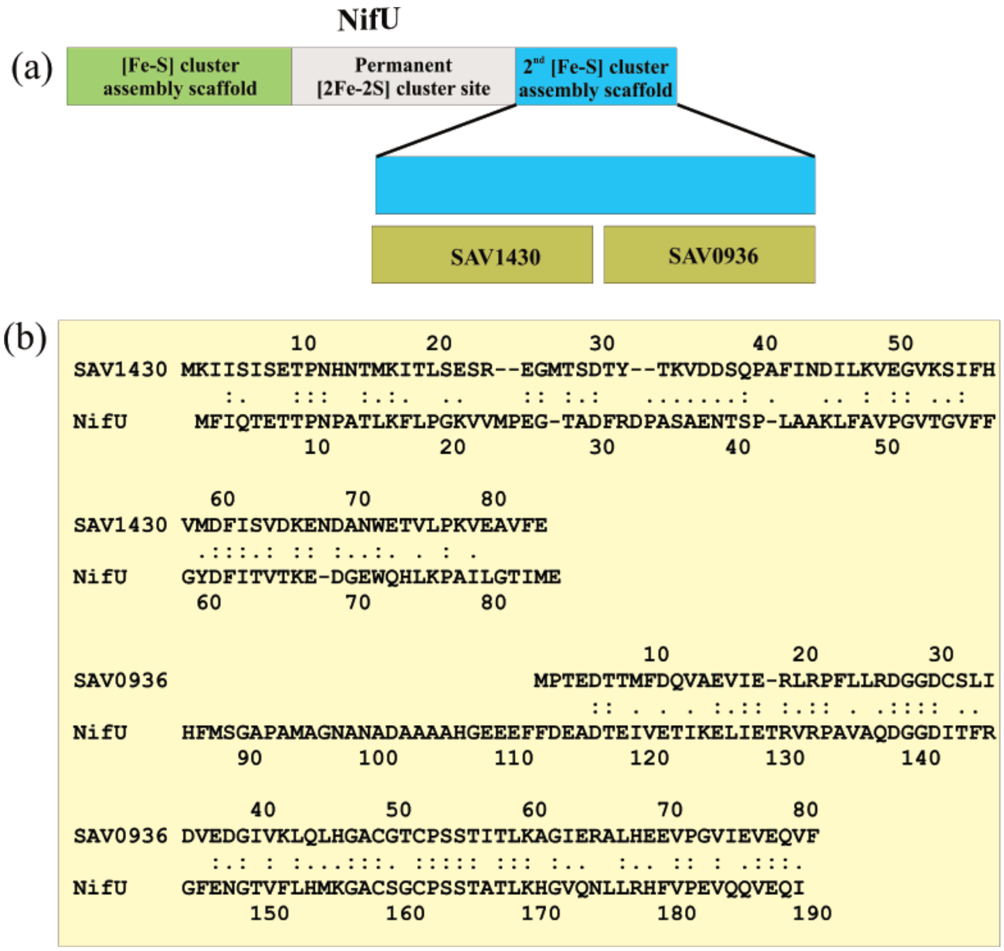

NifU
(a)
[Fe-S] cluster assembly scaffold
Permanent [2Fe-2S] cluster site
2nd [Fe-S] cluster assembly scaffold
SAV1430
SAV0936
(b)
10 20 30 40 50
SAV1430 MKIISISETPNHNTMKITLSESR--EGMTSDTY--TKVDDSQPAFINDILKVEGVKSIFH
NifU MFIQTETTPNPATLKFLPGKVVMPEG-TADFRDPASAENTSP-LAAKLFAVPGVTGVFF
10 20 30 40 50
60 70 80
SAV1430 VMDFISVDKENDANWETVLPKVEAVFE
NifU GYDFITVTKE-DGEWQHLKPAILGTIME
60 70 80
10 20 30
SAV0936 MPTEDTTMFDQVAEVIE-RLRPFLLRDGGDCSLI
NifU HFMSGAPAMAGNANADAAAAHGEEEFFDEADTEIVETIKELIETRVRPAVAQDGGDITFR
90 100 110 120 130 140
40 50 60 70 80
SAV0936 DVEDGIVKLQLHGACGTCPSSTITLKAGIERALHEEVPGVIEVEQVF
NifU GFENGTVFLHMKGACSGCPSSTATLKHGVQNLLRHFVPEVQQVEQI
150 160 170 180 190

Figure 4. (a) Cartoon illustration of the NifU domain structure and the sequence overlap with SAV1430 and SAV0936. (b) Sequence alignment of the C-terminal NifU domain from *Brucella melitensis* with SAV1430 (30% identity) and SAV0936 (47% identity).

SE0630 complex was generated using Hex (Figure 3h).[42] Hex determines a protein−protein complex by identifying complementary binding sites on the basis of surface shape and electrostatic charge and potential distributions. SE0630 was preferentially docked to SAV1430 in the same binding site determined by the NMR ligand binding studies. This further supports a protein−protein binding interaction as part of the biological activity of SAV1430.

SAV0936 exhibits 47% sequence identity with the N-terminal region of the C-terminal NifU domain (Figure 4). SAV0936 contains a surface exposed CXXC loop motif that aligns with conserved cysteines in NifU and has been identified as [Fe−S] ligands, enriched in Ser and Thr residues (typical of regulatory phosphorylation sites) and required for growth.[43] A more exhaustive sequence analysis of SAV1430, based on the results with SAV0936, indicates the protein shares ∼30% sequence identity with the C-terminal region of the C-terminal domain of the NifU multidomain structure (Figure 4). The observed ferredoxin-like fold for SAV1430 is consistent with the electron-transfer activity in the [Fe−S] cluster assembly pathway.[44] Yet, the lack of any cysteine residues in the SAV1430 sequence and an inability to bind iron indicate SAV1430 would not play a direct role in the electron-transfer activity. SAV1430 and SAV0936 sequences account for the majority of the full length NifU domain. These results imply that SAV1430 would interact with SAV0936 to form a complex that exhibits activity similar to that of the full length NifU domain.

The apparent requirement for prokaryotic SAV1430 to form a protein−protein complex, instead of an intact single domain observed in eukaryotic organisms, may suggest a protein assembly or regulation mechanism for SAV1430 in *S. aureus*. The proper assembly of the SAV1430−SAV0936 complex may be required for activity of the NifU-like domain or to provide a mechanism to regulate the [Fe−S] cluster assembly pathway. Potentially, phosphorylation of SAV0936 in the CXXC loop motif may regulate the SAV1430−SAV0936 complex formation in a manner similar to SH2 domains. Precedence for staphylococcus utilizing heterodimeric proteins to accomplish what other bacteria accomplish using monomeric proteins can be found in the enzyme serine dehydratase. In most bacteria, a single gene encodes the serine dehydratase; however, in staphylococci two genes (*sdhA* and *sdhB*) encode serine dehydratase.[45]

Conclusion

Our analysis of hypothetical *S. aureus* protein SAV1430 in the FAST-NMR screen suggests SAV1430 is part of a multiprotein complex within the [Fe−S] cluster assembly network, where SAV1430 may interact with SAV0936 to form a complex

[42] Ritchie, D. W. *Proteins* **2003**, *52*, 98.
[43] Dos Santos, P. C.; Smith, A. D.; Frazzon, J.; Cash, V. L.; Johnson, M. K.; Dean, D. R. *J. Biol. Chem.* **2004**, *279*, 19705.
[44] Bertini, I.; Luchinat, C.; Provenzani, A.; Rosato, A.; Vasos, P. R. *Proteins* **2001**, *46*, 110.
[45] Ogawa, H. *Trends Comp. Biochem. Physiol.* **2001**, *6*, 1.

that exhibits activity similar to that of the intact C-terminus domain of NifU. Formation of the SAV1430−SAV0936 complex may be regulated in a manner similar to SH2 domains through the phosphorylation of SAV0936. It is also plausible that the SAV1430−SAV0936 complex may be used to regulate the [Fe−S] cluster assembly network. It is important to stress that the only information available for hypothetical protein SAV1430 at the beginning of the FAST-NMR screen was a novel fold and a sequence similar to other hypothetical proteins. Although the results of the FAST-NMR screen are not definitive, it provides important direction for further exploration of the biological function of SAV1430, SAV0936, and other similar hypothetical proteins. The success of structural genomics has resulted in hundreds of unannotated protein structures being deposited in the PDB. Inevitably, the inherent value of data generated by genomics sequencing and structural efforts is significantly reduced without new methodologies to provide annotation information.

Acknowledgment. This work was supported by grants from the Protein Structure Initiative of the National Institutes of Health (U54 GM074958), Nebraska Tobacco Settlement Biomedical Research Development Funds, Nebraska EPSCoR, and an Eloise Gerry Fellowship from Sigma Delta Epsilon. The research was performed in facilities renovated with support from the NIH (RR015468-01).

JA0651759