Functional Annotation Screening Technology by Nuclear Magnetic Resonance Spectroscopy

By

Kelly A. Mercier

A Dissertation

Presented to the Faculty of
The Graduate College at the University of Nebraska
In Partial Fulfillment of Requirements
For the Degree of Doctor of Philosophy

Major: Chemistry

Under the Supervision of Professor Robert Powers
Lincoln, Nebraska
August 2008

Functional Annotation Screening Technology by Nuclear Magnetic Resonance Spectroscopy

Kelly A. Mercier, Ph.D

University of Nebraska, 2008

Advisor: Robert Powers

The past decade has brought continued success for genomics, leading to many new areas of research in chemistry, biology, and medicine. While the sequencing the genomes provides a great deal of information, the true wealth of information lies in the structure and function of these gene products. The Protein Structure Initiative was initiated in 1999 to address the need for a unified effort to address the ~6.5 million proteins hypothesized. Since the proteins solved continually have novel folds, bioinformatics programs and databases are frequently unable to assign putative functions for these protein structures.

Functional Annotation Screening Technology by Nuclear Magnetic Resonance Spectroscopy was developed to assign a general biological function to proteins that lack putative function annotation. FAST-NMR is based on the premise that a biological function can be described by a similarity in binding sites and ligand interactions with proteins of known function. Using NMR techniques developed in pharmaceutical research, each protein is screened in a tierred-manner against a library of biologically active compounds. Once the functional ligands and active site are identified experimentally, the active site is then compared to all active sites in the Protein DataBase by the Comparision of Active Site Structures database and program

(CPASS). The work described here encompasses the development of the FAST-NMR method, two function annotations of proteins using FAST-NMR, and a NMR structure of SR211, a *Bacillus subtilis* protein from the PSI.

ACKNOWLEDGMENTS

I would like to thank Dr. Robert Powers for his guidance, patience, and support during the past 5 years. Your expertise and experience in pharmaceutical research and NMR screening has been invaluable, and I have truly felt appreciated for my efforts. I wish you all the best in your tenure review next fall.

My supervisory committee has been very helpful in my research and coursework. Dr. Hage, Dr. Griep, Dr. Takacs, and Dr. Stone have been a wonderful committee that challenged me to be a better chemist.

As my work relied on collaborations, I would have been lost without Dr. Guy Montelione, Dr. Mike Baran, and Dr. John Cort, all from the NorthEast Structural Genomics consortium. Thank you for the protein material, structures, and stimulating discussions on the functions. I appreciate our collaboration with Dr. Somerville and Dr. Sadykov for their work on the SAV1430 metabolomics project and helpful advice and plasmids for the over-expression of recombinant proteins. I would also like to thank Dr. Geoff Mueller of the National Institute of Environmental Health Science for his expertise and assistance with the structure of SR211.

I have been very lucky to have worked with a group of intelligent, motivated, and funny graduate students in the Powers lab: Steve Halouska, Jenni Copeland, Matt Shortridge, Jaime Stark, Visu Ramanathan, and Bo Zhang. Thank you for your assistance during my pregnancy and for our discussions over the past few years. I have learned so much from our discussions. I have also been fortunate to have worked with several undergraduate students: Katherine Germer,

Matt Shortridge, Lisa Ludwig, and Mike Bokemper. I challenged them to help me with some big projects, and they all came through with flying colors. Thank you for your hard work.

I have been blessed to have great friends in Hamilton Hall. I very much appreciate the continued assistance with biochemistry and computer troubleshooting and friendship of Torri and Garhan Attebury. I have also been very lucky to have female friends in the department who are excellent role models. Thank you, Sarah Pettito, Erin Marsh, Annette Moser, Lancia Darville, and JoDell Whittington, for your friendship and support.

My family has been there for me thorough my educational career: my mother, Robin; my father and stepmother, Rex and Helen; and my sisters, Stacey, Jennifer, and Amy. While not always understanding my work, you have always pushed me to keep working.

Lastly, to my husband, Michael: thank you for your love. You have moved around the country to support me in my studies and work. I could not have done any of this without you.

Table of Contents

CHAPTER 1

Introduction

1.1 Introduction

Genomics in the past decade has spurred a plethora of research in biology, chemistry and drug discovery. Genetic mapping is of the utmost importance for whole organisms to obtain biological information and alleles for drug discovery. Fred Sanger first solved the primary sequence, or linear arrangements of amino acids, of insulin in 1955.[1, 2] He also developed the chain termination for this sequencing in 1977, which he then used to solve the genome of Phi X 174, a bacteriophage of 5386 bases.[3] The techniques that Sanger developed spearheaded the efforts that have led to today's successes. With more than 686 completed genomes and 1270 on-going projects, there are ~6.5 million proteins projected.[4]

The central dogma of molecular biology connects the DNA sequences to RNA and proteins through the processes of transcription and translation.[5] These gene products are instrumental in all cellular processes. In order to understand how these proteins function, structural information is sought. As mentioned previously, the primary sequence of proteins is most commonly determined by chain termination. Secondary structures of proteins were discovered by Pauling in 1951.[6] The φ,ψ angles of the peptide bonds that link each amino acid together in the primary sequence dictate the secondary structure elements, like α-helices and β-sheets. The three dimensional arrangement of these secondary elements form the tertiary structures or fold of the protein. The first tertiary structures of proteins were solved in 1958 and 1959 for myoglobin and hemoglobin, for which Kendrew and Perutz won Nobel prizes in 1962.[7,

[8] Both structures demonstrated how the domains interact and stabilize the heme groups for their essential functions *in vivo*.

1.2 Protein Structure Initiative

The Protein Structure Initiative (PSI) was initiated in 1999.[9] There were four main goals outlined by the National Institute of General Medical Sciences: 1), organize known protein sequences into families; 2), select family representatives as targets; 3), solve the 3D structure of targets by X-ray crystallography or NMR spectroscopy; and 4,) build models for other proteins by homology to the solved 3D structures. Protein selection is based on the goal of maximizing the coverage of structural space by choosing targets that are representative of a family of proteins that lack a known structure.[10] A single structure may serve as a model for the other members in the protein family.[11] As the protein structural space is expanded through the PSI, several benefits have come to fruition. The structures solved have given a better understanding as to the relationship between structure and function, and therefore better predictions can be made. These structures are also being used as starting points for drug-discovery projects and for providing key information on protein folding and evolution. The technical advances through the PSI also have wide-reaching benefits in chemistry, medicine, and physics.

Protein structures are determined using two methods, X-ray crystallography and NMR spectroscopy. X-ray crystallography is used far more frequently than NMR because of its versatility, but there are inherent issues with both techniques. The process of solving protein structures by X-ray crystallography has become high-throughput.[12-14] X-ray crystallography requires that the protein be crystallized, which is not a trivial process for some systems.[15] The

process of crystallization can cause crystal packing problems and may not lead to reliable structures.[16, 17] There have also been multiple instances where the use of synchrotrons have altered the protein in a random, unpredictable manner through excess radiation.[18-20] While there are clearly some issues that can arise from the use of crystallography, the sample molecular weight is not an issue for X-ray methods, as recently demonstrated by virus structures that have been solved by this method.[21, 22]

NMR Spectroscopy, on the other hand, has not been used as extensively for protein structures until recently and thus is still being developed into a more high throughput method for protein structural determination. Whereas the rate-limiting step of X-Ray crystallography is the obtainment of a quality crystal that gives good reflections, NMR spectroscopy requires 1) isotopic labeling all the carbon and nitrogen atoms to be NMR active,[23] and 2) the chemical shift assignments of all atoms in a series of 3D experiments.[24] There is a much tighter restriction on the size of proteins that are able to be solved by NMR spectroscopy than X-ray crystallography. The increased number of dimensions (3D,4D,5D NMR)[25], the advent of TROSY-type experiments,[26] and selective isotopic-labeling[27-29] have helped significantly increase the molecular weight limit, which is greater than 25 kDa for a monomer.[30]

1.3 Proteins of Unknown Function

A current analysis of the protein structures from the PSI indicate that up to 50% of the structures have novel folds.[31] These structures emerging from structural genomics correspond to folds that provide little insight into function.[32] There are currently 2,755 orphaned proteins classified as “unknown function” in the Protein Database (PDB).[33] Valuable information is

hidden among this multitude of unannotated proteins that could be associated with cell viability, biofilm formation, infection, and pathogenesis. These proteins may provide key information for developing new antibiotics, where drug discovery efforts would benefit greatly from new functional annotations methodology.

A recent analysis of various bacterial genomes from the August 2007 Gold release shows that, even with improved computational methods, approximately 40% of bacterial proteins have not been assigned to a functional category (Figure 1.1)[34]. There are more than 11,000 proteins from the ten bacterial organisms listed in Figure 1.1 that lack a functional annotation. Considering this list is only from a small segment of currently sequenced genomes and ~6.5 million proteins projected for all organisms, the prospect of obtaining experimental functional information for all hypothetical proteins identified from completed and ongoing sequencing efforts is a daunting proposition. Current biochemical methods for functional assignments of proteins include the yeast 2-hybrid assay,[35] protein microarrays,[36] and immunoaffinity chromatography and mass spectrometry.[37] These methods, however, require some preexisting knowledge about each system under investigation and cannot be used *de novo*. The yeast 2-hybrid screening, for instance, locates protein-protein interactions and protein-DNA interactions. There are two reasons why the yeast 2-hybrid assay does not work in an exploratory manner, though. The function of the binding domain is usually known previous to the assay,[38] and the incidence of false positives is extremely high.

The methods with which the proteomics groups define function are diverse in their interpretation of protein function. The proteins are classified through cellular, biochemical, or physiologically definitions.[39] Biologists and chemists use computational comparative programs and databases to make these putative assignments based on sequence and structural homology.

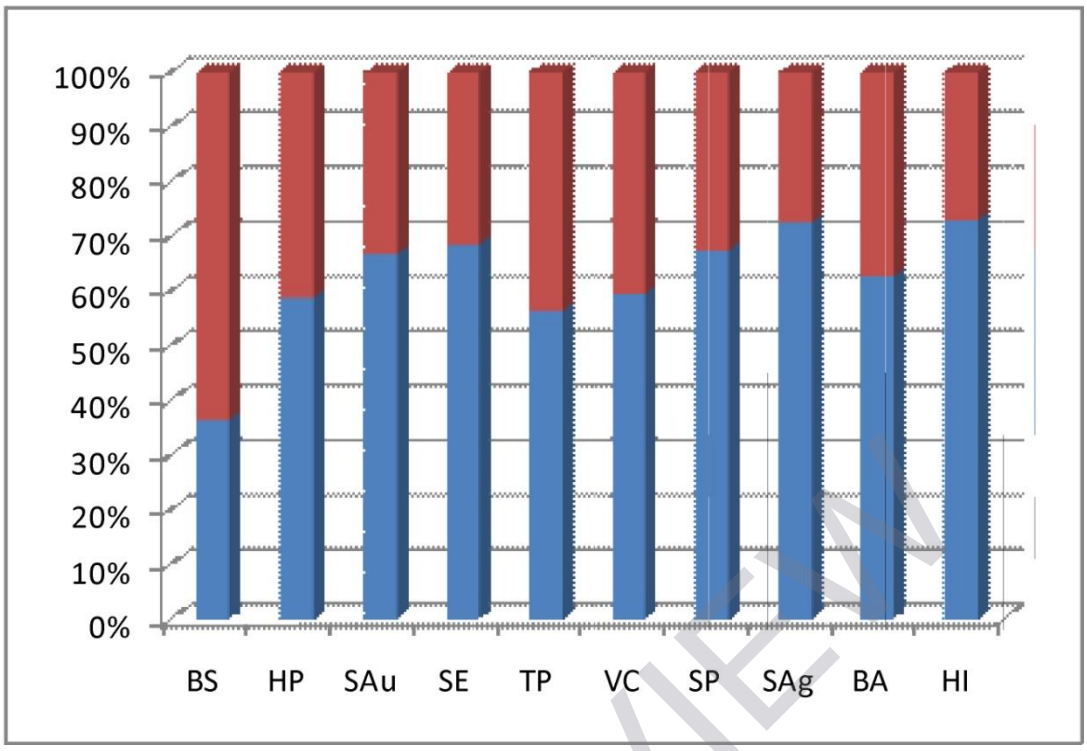

Figure 1.1. Functional analysis of bacterial genomes. The blue partitions are the percentage of proteins that have assigned functional categories; the red partitions are the percentage of unannotated proteins. *Bacillus subtlus* (BS); *Heliobactor pylori* (HP); *Staphylococcus aureus* (SAu); *Staphylococcus epidermis* (SE); *Treponema pallidum* (TP); *Vibrio cholerae* (VC); *Streptococcus pneumonia* (SP); *Streptococcus agalactiae* (SAg); *Bacillus anthracis* (BA); *Haemophilus influenzae* (HI).[34]

[40-42] Even in the hands of experienced scientists, these techniques can be teeming with errors. It has been noted previously that results in common programs, such as SWISS-PROT, are biased or under-sampled, and BLAST scores with seemingly high probability for proteins being related doesn't necessarily tie to the same protein classification.[39]At the most basic definition, protein function is intrinsically linked to the protein active site and a ligand. The identification of the ligand and active site is the root of the problem, but defining a protein function based on active site comparisons has been discussed since the start of the PSI.[43] The residues in the active site must be conserved for evolutionary stability. There have been several examples where sited-directed mutagenesis has verified the functional relationship to conserved amino acids, as in the mammalian alkaline phosphatases[44] and glutamine phosphoribosylpyrophosphate amidotransferase.[45] These clusters of amino acids correspond to sites integral to the function and stability of the whole protein.[46] Divergent evolution has shown that while the global fold can vary, a protein can have comparable function based on the active site and resulting transition states and intermediates.[47] As seen in Figure 1.2, the ligand size can vary quite a bit, but the size of the ligand is inconsequential in this definition. Each of these proteins are in specific pathways where their interactions are highly regulated. The ligands that they come into contact in these pathways with structurally similar. Each of these ligands has a known role in the pathway, as well. Databases like LigBase allow for active site comparisons based on the ligand-defined active sites. This information can then be used for functional assignments based on ligand and active site similarities.

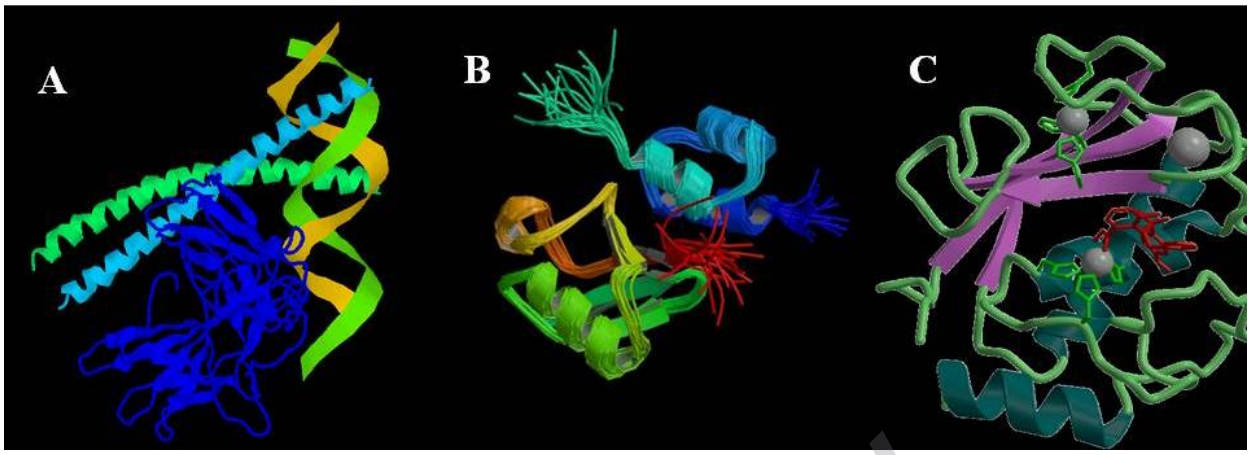

Figure 1.2: The function of a protein is intrinsically linked to the identity of its active-site and ligand. A. Protein-DNA complex (PDB ID 1A02);[48] B. Protein-protein complex (PDB ID 1OTR);[49] C. Protein-ligand complex (PDB ID 1FM1).[50, 51]

1.4 NMR as a screening tool

High throughput screening (HTS) using NMR spectroscopy has become a common component of the drug discovery effort and is widely used throughout the pharmaceutical industry because of the unique ability of NMR to provide direct evidence of a specific binding interaction between a potential chemical lead and the protein of interest.[52-58] Additionally, NMR may be used to evaluate the physical properties of a chemical lead, measure disassociation constants (K_D's),[59] identify ligand binding sites,[55] and determine a co-structure.[55, 60-62] A diverse number of NMR screening approaches have been developed, which include SAR by NMR, [63-67] SHAPES, [68-70] and MS/NMR.[71] NMR spectroscopy is a relatively insensitive technique requiring higher amounts of material and acquisition time compared to standard methods used in traditional HTS assays. Thus, a fundamental issue with NMR screens is a need to optimize the efficiency of sample throughput by achieving a balance between information content and resource utilization. As a result, NMR-based assays screened in an HTS assay.

An approach used to address the fundamentally lower throughput of NMR has been the development of small, directed compound libraries that are more amenable to NMR based screens.[53, 72-74] The SHAPES library is a typical example of the fragment based approach to NMR screening, where the library consists of a small, structurally diverse set of water soluble compounds that correspond to fragments or molecular frameworks of known drugs.[67-70, 75] A comparable approach to reduce the size of screening libraries is to use NMR-based assays as a secondary screen to validate hits from HTS assays. In this manner, the HTS assays reduce a large corporate library to a small, focused list composed of a few hundred to a few thousands

compounds that is more applicable to an NMR screen.[76, 77] In these assays, the biological function of the compounds are unknown, though. While designed to interact in specific locations on the surface or in a cleft, it is not known in these experiments how the compounds will react *in vivo*.

In general, the receptor of interest in an HTS or NMR-based screen is already well characterized before a drug discovery effort is initiated. Nevertheless, pharmaceutical companies are constantly searching for new protein targets that may lead to the development of novel therapeutics or alleviate adverse side-effects.[78] The success of various genomic projects has identified a wealth of potential therapeutic targets.[79, 80] There is clearly a critical need for a method that identifies these functional ligands and the active site for drug discovery.

1.5 Functional Annotation Screening Technology by NMR (FAST-NMR)

NMR-based screening methodologies can be applied to assign a biological function to these novel proteins identified from PSI. Functional Annotation Screening Technology by NMR (FAST-NMR) provides experimental data for the functional annotation of novel proteins by combining ligand affinity screens and structural biology with bioinformatic approaches.[81, 82] Functional ligands that bind to the protein are identified using a tiered NMR ligand screen.[83] The active site of the ligand and protein are then found in the second tier of the screening through chemical shift perturbations which are mapped onto the surface of the protein. A co-structure is rapidly determined using the chemical shift perturbations[84] and then used to drive the bioinformatics analysis. The Comparison of Active Site Structures (CPASS) program and database compares the AutoDock co-structure and active site to all the active sites in the PDB.[82]

A general biological function is assigned through the comparative similarity of the experimentally determined ligand binding sites.

This dissertation will encompass the development of FAST-NMR methodology, describe two complete annotations of novel proteins, and the structure of SR211, a unannotated protein from *Bacillus subtilis*. Chapter 2 describes FAST-NMR as a screening method and computational tool for functional annotation of proteins from the PSI. Chapter 3 of this manuscript fully describes how mixture size was considered in the screening process of the function ligand library, which is an important modification of NMR screening (Chapter 4). The annotation of SAV1430, a novel protein from *Staphylococcus aureus*, is described in Chapter 5, and the annotation of PA1324, a putative carbohydrate-binding protein from *Pseudomonas aeruginosa*, is described in Chapter 6. Chapter 7 contains the NMR structure of SR211 and describes future work for this protein.

1.6 References

1. Ryle, A. P.; Sanger, F.; Smith, L. F.; Kitai, R., The disulphide bonds of insulin. *Biochemical Journal* **1955,** 60, (4), 541-56.

2. Sanger, F.; Thompson, E. O.; Kitai, R., The amide groups of insulin. *Biochemistry Journal* **1955,** 59, (3), 509-18.

3. Sanger, F.; Air, G. M.; Barrell, B. G.; Brown, N. L.; Coulson, A. R.; Fiddes, C. A.; Hutchison, C. A.; Slocombe, P. M.; Smith, M., Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **1977,** 265, (5596), 687-95.

4. Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Wheeler, D. L., GenBank. *Nucleic Acids Res* **2008,** 36, (Database issue), D25-30.

5. Crick, F. H., On protein synthesis. *Symposia of the Society for Experimental Biology* **1958,** 12, 138-63.

6. Pauling, L.; Corey, R. B., Configuration of polypeptide chains. *Nature* **1951,** 168, (4274), 550-1.

7. Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C., A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **1958,** 181, (4610), 662-6.

8. Perutz, M. F., Structure of hemoglobin. *Brookhaven Symposium in Biology* **1960,** 13, 165-83.

9. Stevens, R. C.; Yokoyama, S.; Wilson, I. A., Global efforts in structural genomics. *Science* **2001,** 294, (5540), 89-92.

10. Mirkovic, N.; Li, Z.; Parnassa, A.; Murray, D., Strategies for high-throughput comparative modeling: applications to leverage analysis in structural genomics and protein family organization. *Proteins* **2007,** 66, (4), 766-77.

11. Cort, J. R.; Koonin, E. V.; Bash, P. A.; Kennedy, M. A., A phylogenetic approach to target selection for structural genomics: solution structure of YciH. *Nucleic Acids Res* **1999,** 27, (20), 4018-27.

12. Kambach, C., Pipelines, robots, crystals and biology: what use high throughput solving structures of challenging targets? *Current Protein and Peptide Science* **2007,** 8, (2), 205-17.

13. Manjasetty, B. A.; Shi, W.; Zhan, C.; Fiser, A.; Chance, M. R., A high-throughput approach to protein structure analysis. *Genetic Engineer (N Y)* **2007,** 28, 105-28.

14. Manjasetty, B. A.; Turnbull, A. P.; Panjikar, S.; Bussow, K.; Chance, M. R., Automated technologies and novel techniques to accelerate protein crystallography for structural genomics. *Proteomics* **2008,** 8, (4), 612-25.

15. Hickman, A. B.; Davies, D. R., Principles of macromolecular X-ray crystallography. *Current Protocols in Protein Science* **2001,** Chapter 17, Unit 17 3.

16. Andrec, M.; Snyder, D. A.; Zhou, Z.; Young, J.; Montelione, G. T.; Levy, R. M., A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. *Proteins* **2007,** 69, (3), 449-65.

17. Eyal, E.; Gerzon, S.; Potapov, V.; Edelman, M.; Sobolev, V., The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *Journal of Molecular Biology* **2005,** 351, (2), 431-42.

18. Carugo, O.; Djinovic Carugo, K., When X-rays modify the protein structure: radiation damage at work. *Trends Biochemical Sciences* **2005,** 30, (4), 213-9.

19. Ravelli, R. B.; Garman, E. F., Radiation damage in macromolecular cryocrystallography. *Current Opinion in Structural Biology* **2006,** 16, (5), 624-9.

20. Massover, W. H., Radiation damage to protein specimens from electron beam imaging and diffraction: a mini-review of anti-damage approaches, with special reference to synchrotron X-ray crystallography. *Journal of Synchrotron Radiation* **2007,** 14, (Pt 1), 116-27.

21. Zhang, Y.; Kaufmann, B.; Chipman, P. R.; Kuhn, R. J.; Rossmann, M. G., Structure of immature West Nile virus. *Journal of Virology* **2007,** 81, (11), 6141-5.

22. Yu, I. M.; Zhang, W.; Holdaway, H. A.; Li, L.; Kostyuchenko, V. A.; Chipman, P. R.; Kuhn, R. J.; Rossmann, M. G.; Chen, J., Structure of the immature dengue virus at low pH primes proteolytic maturation. *Science* **2008,** 319, (5871), 1834-7.

23. Bax, A.; Ikura, M.; Kay, L. E.; Barbato, G.; Spera, S., Multidimensional triple resonance NMR spectroscopy of isotopically uniformly enriched proteins: a powerful new strategy for structure determination. *Ciba Found Symp* **1991,** 161, 108-19; discussion 119-35.

24. Oldfield, E., Chemical shifts and three-dimensional protein structures. *Journal of Biomolecular NMR* **1995,** 5, (3), 217-25.

25. Zhang, Q.; Atreya, H. S.; Kamen, D. E.; Girvin, M. E.; Szyperski, T., GFT projection NMR based resonance assignment of membrane proteins: application to subunit C of E. coli F(1)F (0) ATP synthase in LPPG micelles. *Journal of Biomolecular NMR* **2008,** 40, (3), 157-63.

26. Pervushin, K.; Riek, R.; Wider, G.; Wuthrich, K., Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci U S A* **1997,** 94, (23), 12366-71.

27. Staunton, D.; Schlinkert, R.; Zanetti, G.; Colebrook, S. A.; Campbell, I. D., Cell-free expression and selective isotope labelling in protein NMR. *Magnetic Resonance in Chemistry* **2006,** 44 Spec No, S2-9.

28. Tugarinov, V.; Kanelis, V.; Kay, L. E., Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. *Nat Protocols* **2006,** 1, (2), 749-54.

29. Xu, Y.; Zheng, Y.; Fan, J. S.; Yang, D., A new strategy for structure determination of large proteins in solution without deuteration. *Nature Methods* **2006,** 3, (11), 931-7.

30. Kay, L. E.; Gardner, K. H., Solution NMR spectroscopy beyond 25 kDa. *Current Opinion in Structural Biology* **1997,** 7, (5), 722-31.

31. Levitt, M., Growth of novel protein structural data. *Proceedings of the National Academy of Sciences U S A* **2007,** 104, (9), 3183-8.

32. Kumaran, D.; Eswaramoorthy, S.; Gerchman, S. E.; Kycia, H.; Studier, F. W.; Swaminathan, S., Crystal structure of a putative CN hydrolase from yeast. *Proteins: Structure, Function, and Genetics* **2003,** 52, (2), 283-291.

33. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res* **2000,** 28, (1), 235-42.

34. Bernal, A.; Ear, Y.; Kyrpides, N., Genomes online database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Research* **2001,** 29, (1), 126-127.

35. Lentze, N.; Auerbach, D., The yeast two-hybrid system and its role in drug discovery. *Expert Opinion on Theraputic Targets* **2008,** 12, (4), 505-15.

36. Tao, S. C.; Chen, C. S.; Zhu, H., Applications of protein microarray technology. *Combinatorial Chemistry and High Throughput Screening* **2007,** 10, (8), 706-18.

37. Ackermann, B. L.; Berna, M. J., Coupling immunoaffinity techniques with MS for quantitative analysis of low-abundance protein biomarkers. *Expert Review of Proteomics* **2007,** 4, (2), 175-86.

38. Gietz, R. D.; Triggs-Raine, B.; Robbins, A.; Graham, K. C.; Woods, R. A., Identification of proteins that interact with a protein of interest: applications of the yeast two-hybrid system. *Molecular and Cellular Biochemistry* **1997,** 172, (1-2), 67-79.

39. Rost, B.; Liu, J.; Nair, R.; Wrzeszczynski, K. O.; Ofran, Y., Automatic prediction of protein function. *Cell Mol Life Sci* **2003,** 60, (12), 2637-50.

40. Koonin, E. V., Computational genomics. *Current Biology* **2001,** 11, (5), R155-8.