Using Databases and Computational Techniques to Infer the Function of Novel Proteins

Viswanathan Ramanathan[1], Kelly A. Mercier[2], Robert Powers[3], and Peter Z. Revesz[4]
[1,4]*Department of Computer Science & Engineering,* [2,3]*Department of Chemistry*
University of Nebraska-Lincoln, Lincoln, NE 68588, USA
[1]ramanath@cse.unl.edu, [2]kannedo@bigred.unl.edu, [3]rpowers3@unl.edu, [4]revesz@cse.unl.edu

Abstract

The Human Genome Project and similar efforts have resulted in the identification of an abundance of novel proteins. There is a need to expedite the process of assigning function to novel proteins. Nuclear magnetic resonance (NMR) spectroscopy can be used to infer a general biological function for a protein of unknown function by identifying compounds that preferentially bind the protein and comparing these results against proteins with defined structure and function. The Functional NMR screen generates hundreds of data sets and a manual analysis of these data sets is laborious and time-consuming. It is hypothesized that several sub-tasks of the Functional NMR can be automated successfully using an integrated database and data analysis system. Our database system integrates NMR data collection, processing, analysis, and data archiving into a unified user interface. An NMR spectra comparison algorithm is designed and implemented to compare NMR data in the presence and absence of a protein to ascertain if any compound-protein binding occurred.

1. Introduction

The Human Genome Project [13], Protein Structure Initiative [10], and other large-scale research projects have generated a vast amount of biological information, creating the need for database systems that can organize and analyze this information effectively [1, 4, 6]. Development of database systems to support biological research has gained widespread momentum over the last five years [2, 3, 4]. An important criticism of these biological databases has been a disregard for the analyses that may need to be performed on the data [3]. Overall, databases have not been efficiently designed to address biological research needs. There have been recent efforts to design databases (e.g., SPINE; [3, 6]) that incorporate analytic capabilities to help answer relevant biological research questions.

The current paper presents the design of one such database system used to assign general biological function to an increasingly expanding repository of novel proteins. Fundamental to our database design is the inclusion of strategies for subsequent data analyses. We propose that a comprehensive database system coupled with data analysis techniques can significantly advance biological research.

1.1. Protein-compound binding

Proteins are basic constituents in all living organisms that perform diverse functions essential for the survival and proliferation of the organism. Small errors in protein structure often cause human diseases, making the study of proteins an important scientific enterprise. Intrinsic to the functional activity of a protein is its interaction with other biomolecules and small molecules such as carbohydrates, nucleic acids and vitamins. Thus, the structure of a protein is optimized to specifically bind these components as it performs its function.

The Human Genome Project has generated a considerable amount of genetic information that is inundating the scientific community. Approximately 30,000–90,000 proteins are predicted to be encoded from the human genome alone, where sequencing of other model organisms is adding to this tremendous wealth of knowledge [14]. Using traditional biochemical approaches to obtain functional information for this immense collection of proteins is not feasible, as years of research are typically required to identify the function of a single protein [5].

Given that "function follows form", the Protein Structure Initiative has focused on inferring function from protein structures and has embarked on an ambitious effort to determine a structure for all proteins. Ex-

-7803-9232-9/05/$20.00 2005 IEEE.

isting protein structures have been grouped into hierarchical clusters of families based on structural similarities. Typically, proteins with similar structural characteristics have similar or related biological functions. Thus for a large number of novel proteins, function will be inferred based on the observation that proteins of similar structure and sequence have related functions. Determining accurate structures of proteins can take several months to a year. A further extension of the analysis of the function of a protein is by understanding and identifying the biomolecules and small molecules the protein binds. In a comparable manner, proteins can be assigned to similar functional groups based on similar binding interactions to small molecules and other biomolecules.

1.2. Functional NMR

Nuclear magnetic resonance (NMR) spectroscopy is routinely employed to study the physical, chemical, and biological properties of proteins at atomic resolution [7, 11, 15]. A one-dimensional (1D) ^{1}H NMR spectrum for a specific small molecule or protein will contain a peak corresponding to the resonance absorbance for each unique ^{1}H nucleus in the molecule. Depending on the complexity of the molecule, there will be numerous ^{1}H resonance peaks where clusters of overlapping peaks are very common. The complexity of an NMR spectra can be simplified by spreading out the information into 2D, 3D or even 4-dimensions.

The Functional NMR screen is based on the knowledge that a protein by nature has been optimized to interact with a unique and specific target [12]. Such a target that binds specifically to a protein is called a *ligand* and the site of interaction is called the *active site* or *binding site*. The following information is obtained from the Functional NMR screen: (1) the identity of the ligands that preferentially bind with the unknown protein; (2) the protein's active binding site; and (3) the structure of the protein-ligand binding interaction. By comparing these results against databases of proteins with known function, a general biological function is assigned to the protein of unknown function. This information is obtained by screening a chemical library composed of ligands with defined biological functions such as amino-acids, carbohydrates, co-factors, fatty acids, hormones, inhibitors, known drugs, metabolites, nucleic-acids, substrates, and vitamins.

A crucial step in the Functional NMR screen is identifying the ligand(s) from the chemical library that binds the protein of unknown function and ranking their relative binding affinity. To minimize utilization of resources, mixtures comprising three to four compounds are used in the Functional NMR experiments to screen the entire chemical library [9]. To identify ligands that bind the protein of unknown function, 1D NMR spectrum for each mixture of three-four compounds (reference spectrum) is compared with the 1D NMR spectrum for the mixture of the same set of compounds with the addition of the target protein (compound–protein spectrum). A change in the intensity of the peaks, specifically a decline in the compound-protein spectrum relative to the reference spectrum is indicative of protein binding (Figure 1).

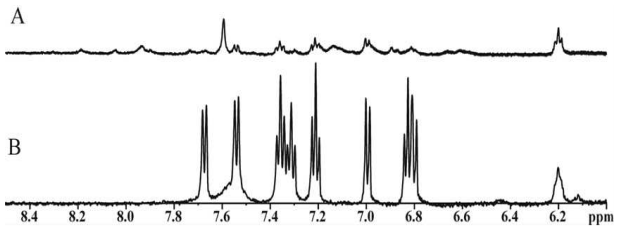

Figure 1. Example shows expanded regions of NMR spectra. (A) Compound-protein spectrum; (B) Reference spectrum. The decrease in the peak intensities in (A) relative to (B) indicates compound-protein binding.

Spectra comparison is complicated due to several data issues. These include the challenges of lack of a common scale between spectra, differentiating peaks from noise, the presence of protein NMR signals in the compound-protein spectrum, the presence of water and buffer NMR peaks, the variability in peak position and intensity due to instrument instability, multiple peaks attributed to each compound in the NMR spectra, overlap of NMR peaks, and the need to properly assign the NMR peaks to each of the three-four compounds in a mixture. Hundreds of NMR spectra pairs will eventually need to be compared as the NMR screening experiments are repeated for each protein of unknown function, adding to the difficulty of manual spectra comparison.

2. Database design

A database system has been designed and is actively being developed to automate several components of the Functional NMR screening process. This database system design integrates NMR data collection, processing, analysis and data archiving, and includes a unified user interface. Figure 2 shows the flow diagram of the Functional NMR screening process.

The three databases that are integrated in this de-

-7803-9232-9/05/$20.00 2005 IEEE.

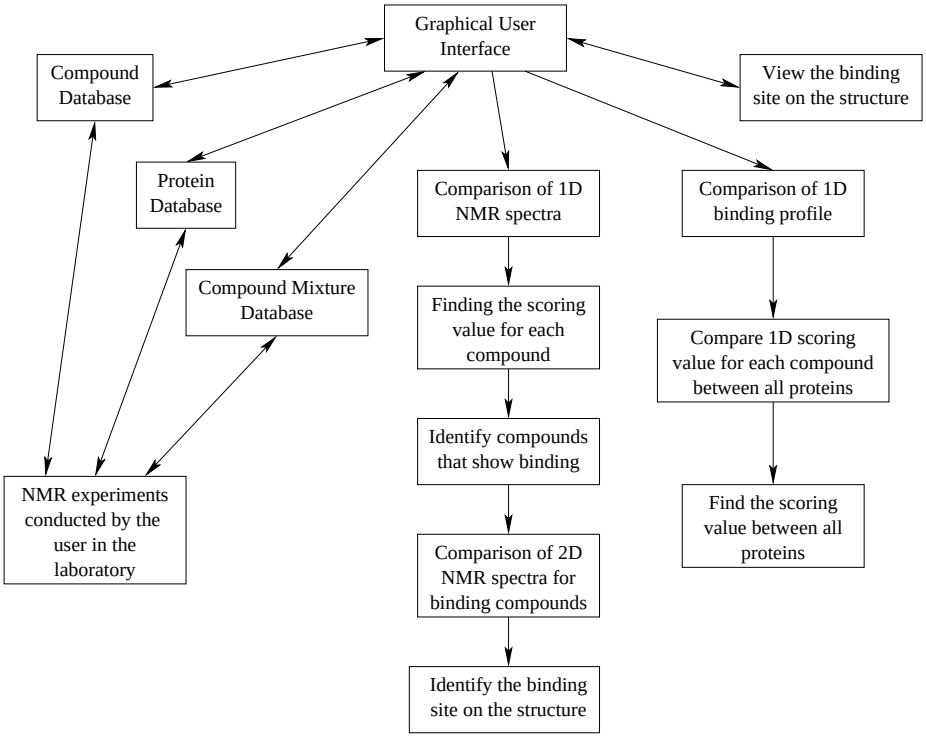

Figure 2. Flow diagram of the Functional NMR screening process.

sign are the *Compound Database*, the *Protein Database*, and the *Compound Mixture Database*. The Compound Database is a database of more than 300 (and increasing) compounds that have defined biological functions and utility as ligands in the Functional NMR screen. The database contains the name, structure, function, and NMR reference spectra of these ligand compounds. The Protein Database consists of the proteins of known structure and function. It contains the name, sequence, structure, function, and the similarity score for each of the proteins against the entire Compound Database. The Compound Mixture Database comprises information about the ligands that make up each mixture, the 1D and 2D NMR spectra of each mixture, and the *similarity score* for each ligand in each mixture for each protein of unknown function. The NMR experiments are conducted using the compounds, the proteins and the compound mixtures whose information is stored in these databases, where the resulting NMR data is added back to the respective databases for each execution of the Functional NMR screen.

A critical consideration in developing the database system was to incorporate the ability to analyze data stored in the databases. Initial NMR experiments are conducted to obtain 1D NMR spectra for each individual compound in the library, for the ligand mixture(s) and for the same mixture(s) in the presence of each protein. The database design includes a computational program to compare the 1D NMR spectra acquired for the compound ligand mixture and the compound ligand-protein mixture to identify ligands that show binding to the protein target. A scoring function is used to give a similarity score for each ligand in the compound mixture. The scoring function identifies the ligands that bind the protein and provides a means to rank the relative binding affinity of the ligands. This data is stored in the Compound Mixture Database. For ligands that are identified as binding the protein, a second 2D NMR experiment is required to determine the binding-site on the protein. These potential active-sites can be viewed through a graphical user interface.

2.1. System components

An effective database design needs thoughtful consideration of the organization and relationship among its system components. Figure 3 shows the interaction among the system components of the proposed database system.

The Compound, Protein, and Compound Mixture databases are implemented on the MySQL 4.1 data-

-7803-9232-9/05/$20.00 2005 IEEE.

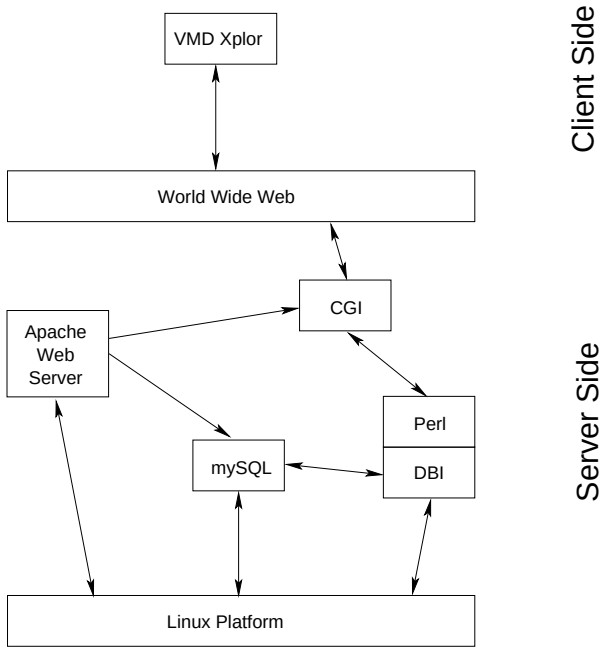

Figure 3. Interaction among System Components.

base server. MySQL is a free, fast, robust, reliable open source relational database that supports all known platforms and requires less hardware resources. The computational analysis techniques for NMR spectra comparison are implemented using Perl 5.8.6. Perl is a free programming and scripting language that is available for most operating systems. It has extensive programming flexibility and is very suitable for file and string handling. Perl uses the Perl Database Interface (DBI) to facilitate low level database interaction with the MySQL database engine. The Perl DBI.pm module enables Perl applications to access the MySQL database transparently. The core of the user interface was developed using CGI. The Perl CGI.pm module provides a simple interface for parsing and interpreting query strings passed to CGI scripts. The user interface is hosted on the Apache 1.3.33 HTTP Web Server. The mod_perl module can be used to manage the Apache web server, and respond to requests for web pages. It gives a persistent Perl interpreter embedded in the web server. Users can access the information through a graphical user interface. The binding site on the protein structures can be viewed using VMD-XPLOR. The whole system is implemented on a Red Hat Linux 7.3 platform.

3. Spectra comparison algorithm

The usefulness of the Functional NMR screening database system lies in its inclusion of data analytic techniques to perform spectra comparison in an efficient manner. The NMR screen assigns a general function to a protein of unknown function based on its similarity with proteins of known function. A key criterion for similarity comparison is the specific ligands that bind the novel and known proteins. The identification of ligands that preferentially bind is in turn dependent on comparison of the reference spectrum and the compound-protein spectrum. A decline in peak intensities in the compound-protein spectrum relative to the reference spectrum indicates binding. The subsequent task is to assign a similarity score for the ligands in the mixture based on the strength of their binding with the protein.

Manual comparison of NMR spectra is laborious and time-consuming. Also, the nature of NMR spectra gives rise to a number of data analyzability issues. The spectra comparison algorithm was written and implemented to automate NMR spectra comparison. For any meaningful analysis of NMR data, the algorithm needs to address key data analyzability issues, such as:

1. *Scale*: The scale of the peak intensities in a NMR spectrum differs from spectrum to spectrum. The lack of a common scale makes spectra comparison difficult.

2. *Noise*: The problem of noise in scientific data is well-acknowledged and NMR spectra are no exceptions [8]. Data analysis of NMR spectra requires separation of noise from actual data, especially relatively smaller peaks from noise.

3. *Water and Buffer*: Sample solutions used in NMR experiments contain water, in which the compound is dissolved, and a buffer that maintains constant sample conditions. The NMR spectrum of the solution will therefore contain peaks that correspond to the water and the buffer in the solution. The water and buffer regions need to be distinguished from the remaining peak regions in the NMR spectrum and removed before data analysis.

4. *Peaks shift or merge*: After the addition of the protein to the compound mixture containing the potential ligands, some of the peaks in the resulting NMR spectrum shift from their original position in the reference spectrum. In some cases, the peaks merge together to form a new peak. These shifts or changes in peaks introduce additional complex-

-7803-9232-9/05/$20.00 2005 IEEE.

ity to the problem of finding changes in peak intensities relative to the reference spectrum.

5. *Overlapping peaks*: The NMR spectra sometimes contain peaks that have overlapping regions. These peaks are often difficult to discern from each other in the graphical display of the spectra. On the other hand, manual inspection of the actual data points is not feasible (there are a total of 8192 points per spectrum).

The proposed algorithm addresses these data analyzability issues in the following ways:

1. *Scale*: The algorithm transforms the raw intensities of the NMR spectra to standardized z-scores, thus allowing for the comparison of intensities between different NMR spectra.
2. *Noise*: Several methods of accounting for noise were attempted. One of the methods involved finding the median of the z-scores belonging to each NMR spectrum. This median was used as a baseline to differentiate real peaks from noise. Points having z-scores below the median were considered noise and made zero for further analyses. This method consistently gave the best results for comparing NMR spectra.
3. *Water and Buffer*: It is known that the water peaks occur between the frequency positions of 4.3 ppm and 5.0 ppm and the buffer peaks occur between the frequency positions of 3.6 ppm and 3.9 ppm. The algorithm is designed to disregard the regions of the spectrum that correspond to the water and buffer peaks.
4. *Peaks shift or merge*: If the peak in the compound-protein spectrum shifts from its original position in the reference spectrum, then the peak with the highest intensity in the same region is found with an error tolerance of $\pm$ 0.03 ppm. This peak intensity is then used for comparison with the reference spectrum. If peaks merge to form a new single peak in the compound-protein spectrum, then the intensity of this merged peak is compared to all the peaks in the same region of the reference spectrum.
5. *Overlapping Peaks*: The algorithm uses actual data points of the spectra for analysis as opposed to analyzing the graphical display of the spectra, thus minimizing the problem of overlapping peaks. While not manually feasible, an automated program can easily deal with 8192 data points per spectrum.

The spectra comparison method used in this study is described in Algorithm SPECTRA COMPARISON.

Algorithm SPECTRA COMPARISON

Input: Reference ligand mixture spectrum ASCII data file *RefMix*; Compound mixture+protein spectrum ASCII data file *CompProt*; Unknown protein spectrum ASCII data file *UnProt*; List of peaks to be monitored in the reference spectrum *PRef*; Peak assignments for each ligand *PLig*.

Output: Score for each compound; Compound that binds the strongest.

1. For *RefMix*, *CompProt*, and *UnProt*:
   (a) Convert the ASCII data from 4 decimals to 2 decimals.
   (b) Discard the TMSP, water, and buffer peaks.
   (c) Calculate the z-score for each point in *RefMix*, *CompProt*, and *UnProt*.
2. For each point in *CompProt*:
   (a) new z-score in *ligand_mix* := z-score in *CompProt* $-$ z-score in *UnProt*.
3. Find the median of *ligand_mix* and make it as the new baseline (zero). Points having a z-score smaller than the median have a new value of zero.
4. Repeat Step 3 for *RefMix*.
5. For each peak in *PRef*:
   (a) Find the highest z-score in *ligand_mix* corresponding to the peak position with an error tolerance of $\pm$ 0.03 ppm.
   (b) *scoring value* := z-score of peak position in *RefMix* / (z-score of peak position in *RefMix* $-$ z-score in *ligand_mix*).
   (c) Save the peak and it's *scoring value*.
6. Assign the saved peaks to their respective ligand using *PLig*.
7. For each ligand find the average *scoring value*.
8. Return the compounds with their scores.
9. Return the compound that binds the strongest.

Spectra Comparison Algorithm is implemented in Perl. The Perl program takes as input the reference spectrum ASCII data file, the compound-protein ASCII data file, the unknown protein ASCII data file, the list of peaks to be monitored, and the lists of peak assignments for each compound. The total time required to come up with a score for each ligand compound using the Perl program was around two seconds as compared to the numerous hours of effort that would be needed to generate the same data manually.

-7803-9232-9/05/$20.00 2005 IEEE.

4. Discussion

The availability of fully sequenced genomes containing an extensive number of unknown proteins challenges biological researchers to elucidate the structure and function of these proteins at a faster pace. We have proposed a database design to assign biological function to a vast number of proteins. Our database design integrates the various components of the Functional NMR screen. A valuable feature of this design is the inclusion of the Spectra Comparison Algorithm. This database is critical for organizing and analyzing NMR data, effectively streamlining the process of assigning protein function using NMR, and maximizing throughput of this process. This database will be available over the World Wide Web using a graphical user interface. We hope this will help link researchers engaged in similar efforts, make results readily available, and guide future research efforts.

This project has important implications for computational biology and bioinformatics. It demonstrates that databases can answer relevant biological research questions by assimilating analytical requirements into database design. This project also indicates that databases in bioinformatics need to be thoroughly informed by current biological research needs to be truly effective. Such a database is the result of true collaboration between biological experimental researchers and computational researchers. The development of this database system is an ongoing project. The fully developed and integrated database system will prove highly valuable for inferring protein biological function using NMR.

Efforts are ongoing to incorporate more NMR spectra into the database to examine its scalability. In the future, more aspects of the Functional NMR screening process will be integrated into the database design. Peak-picking algorithms, for instance, can be added to the database system design to eliminate the need for peak lists from the user. An alternative approach to comparing NMR peak intensities would be to directly compare the graphical displays of NMR spectra. Although such an approach would generate more information, it requires a higher level of computational sophistication to avoid problems related to overlapping peaks and peak-shifts. Finally, future work should also thoroughly examine the potential data mining capabilities of the Functional NMR database.

4.1. Conclusion

It is well-acknowledged that NMR spectroscopy can play a key role in the determination of structure and function of proteins. The laborious expert reasoning needed for data analysis in NMR significantly limits its applications. The use of databases and automated data analyses can maximize the throughput of using NMR in protein research.

References

[1] H. M. Berman, T. N. Bhat, P. E. Bourne, Z. Feng, G. Gilliland, H. Weissig, and J. Westbrook. The Protein Data Bank and the challenge of structural genomics. *Nature Structural Biology*, 7(11:Suppl.):959–959, 2000.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[3] P. Bertone, Y. Kluger, N. Lan, D. Zheng, D. Christendat, A. Yee, A. M. Edwards, C. H. Arrowsmith, G. T. Montelione, and M. Gerstein. SPINE: An integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Research*, 29(12):2884–2898, 2001.

[4] S. E. Brenner, D. Barken, and M. Levitt. The PRESAGE database for structural genomics. *Nucleic Acids Research*, 27(1):251–253, 1999.

[5] S. K. Burley. An overview of structural genomics. *Nature Structural Biology*, 7(11:Suppl.):932–934, 2000.

[6] C.-S. Goh, N. Lan, N. Echols, S. M. Douglas, D. Milburn, P. Bertone, R. Xiao, L.-C. Ma, D. Zheng, Z. Wunderlich, T. Acton, G. T. Montelione, and M. Gerstein. SPINE 2: A system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Research*, 31(11):2833–2838, 2003.

[7] W. Kemp. *NMR in Chemistry*. Macmillan, 1986.

[8] Y. Lin, P. Hodgkinson, M. Ernst, and A. Pines. A novel detection-estimation scheme for noisy NMR signals: Applications to delayed acquisition data. *Journal of Magnetic Resonance*, 128(1):30–41, 1997.

[9] K. A. Mercier and R. Powers. Determining the optimal size of small molecule mixtures for high throughput NMR screening. *Journal of Biomolecular NMR*, 31(3):243–258, 2005.

[10] National Institute of General Medical Sciences. *Protein Structure Initiative*. http://www.nigms.nih.gov/psi/, 2000.

[11] National Institute of General Medical Sciences. *The structures of life*. Technical Report 01-2778, National Institutes of Health, 2000.

[12] J. M. Thornton, A. E. Todd, D. Milburn, N. Borkakoti, and C. A. Orengo. From structure to function: Approaches and limitations. *Nature Structural Biology*, 7(11:Suppl.):991–994, 2000.

[13] U.S. Department of Energy, Office of Science. *Human Genome Project*. http://www.ornl.gov/hgmis/, 1992.

[14] C. Venter, M. D. Adams, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[15] K. Wuthrich. *NMR of Proteins and Nucleic Acids*. Wiley, 1986.

-7803-9232-9/05/$20.00 2005 IEEE.