

Data-mining Protein Structure Databanks for Crystallization patterns of Proteins

Homayoun Valafar*, James H. Prestegard**, Faramarz Valafar***

- * Southeast collaboratory for Structural Genomics, 220 Riverbend rd., Athens, GA 30602
Email: homayoun@ccrc.uga.edu
- ** James H. Prestegard, University of Georgia, CCRC, 220 Riverbend rd., Athens, GA 30602
- *** Faramarz Valafar, Department of Computer Science, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182

Keyword: X-ray, crystallography, NMR, crystallization, pattern, PDB, datamining, secondary, structure.

Abstract

A study of 345 protein structures selected among 1500 structures determined by Nuclear Magnetic Resonance (NMR) methods, revealed useful correlation between crystallization properties and several parameters for the studied proteins. NMR methods of structure determination do not require the growth of protein crystals, and hence allow comparison of properties of proteins that have or have not been the subject of crystallographic approaches.

One and two-dimensional statistical analyses of the data confirmed a hypothesized relation between the size of the molecule and its crystallization potential. Furthermore, two-dimensional Bayesian analysis revealed a significant relationship between relative ratio of different secondary structures and the likelihood of success for crystallization trials. The most immediate result is an apparent correlation of crystallization potential with protein size. Further analysis of the data revealed a relationship between the unstructured fraction of proteins and the success of its crystallization. Utilization of Bayesian analysis on the latter correlation resulted in a prediction performance of ~64% while a two-dimensional Bayesian analysis succeeded with performance of ~75%.

Introduction

Traditional research in the field of structural biology consists of isolation and purification of a protein (or any other macromolecule) based on interest in function; followed by relating the function of the protein to a high-resolution structure determined by NMR spectroscopy or X-ray crystallography. Current rapid advancements in the field of genomics/bioinformatics have provided a nearly orthogonal path of research in the field of structural biology. Also, structure determination is motivated more by the need to characterize proteins of unknown functions. The new field of structural/functional genomics often utilizes the genetic information in order to predict the structure and function of an unknown protein¹⁻⁴.

The task of structure or function prediction purely based on the genetic or primary sequence information is a complex, yet promising one that may fall into one of the two categories of research endeavors⁴. These two categories consist of the sequence-to-function or sequence-to-structure-to-function paradigms. The first paradigm attempts to extract functional information based on sequence homologies. So far, the use of this approach has been limited to proteins with homologies of at least 30%³⁻⁵. The second paradigm incorporates additional information such as the protein fold or formation of the active site by first obtaining the three dimensional description of a protein fold. While for small proteins, structures may be predicted by ab initio methods⁴, for the more common case of large proteins, structure prediction is possible by using a threading method³. Use of algorithms such as the threading is only possible if a suitable database of protein folds exists. The advancement of prediction methods have imposed an enormous pressure for efficiency in experimental determination of protein structures.

NMR spectroscopy and X-ray crystallography are the prevalent methods of structure determination that can contribute to the rapid production of structures or development of a protein fold database. Although, there is general consensus that the larger molecules are better candidates for crystallization and smaller molecules are better candidates for NMR studies, the size threshold for making this distinction is poorly determined. There are many small and intermediate proteins, for which a choice of approach must be made. An approach that requires no prior decisions is to subject all proteins to an array of conditions in order to yield crystals for X-ray crystallography. At the end of this procedure the protein in question may still not crystallize and thus become a candidate for study by NMR spectroscopy. This trial-and-error based structure determination impedes the rate of the structure determination process. In this paper we examine parameters that could be related to crystallization properties of proteins in an effort to develop an algorithm to predict the most suitable method of structure determination without the need for time-consuming trials.

Materials and Methods

Database construction:

A total of 12000 protein structures present in the RCSB protein data bank (PDB database at the time of analyses)⁶ were downloaded and statistically analyzed. The set of 12000 structures was partitioned into 1504 structures determined by NMR spectroscopy and 7964 structures determined by X-ray crystallography (after selecting proteins of size 50 amino acids and larger). The proteins present in the NMR set were partitioned into two sets. The first set consisted of protein structures, for which only the structures had been determined by NMR spectroscopy, while the second set consisted of protein structures for which structures had been determined by

both NMR spectroscopy and X-ray crystallography. To accomplish the partitioning, the primary sequence of each of the proteins in the initial NMR set was compared to every protein sequence for which X-ray structures existed.

Since we want to consider proteins homologous in sequence and ensure that we are working with structurally well defined proteins, more detailed criteria for the separation and preparation of the partitioned data sets had to be developed.

The following criteria were used in determining the similarity and uniqueness of proteins:

- Two peptide sequences were considered to be similar if and only if:
 1. Both sequences match the minimum size requirement (50 residues).
 2. The two sequences are less than 12 residues different in size.
 3. Sequence alignment of the two sequences results in less than 4 unmatched residues.
- Two peptide sequences are considered to be dissimilar if all of the following conditions are met:
 1. Both sequences match the minimum size requirement (50 residues).
 2. The best sequence alignment of the two sequences (without any gaps) produces more than 42 miss-aligned residues.
 3. The two sequences produce a blast alignment homology score of less than 43⁷.

The effects of imposing these criteria are as follows.

1. The NMR database of 1504 structures was reduced to 918 structures by isolating the proteins with at least 50 amino acid residues.

2. Of the 918 isolated proteins, 182 structures with similar primary sequences in the X-ray database (of 7964 structures) were isolated and put in a database called Xray/NMR-set.
3. Of the 918 isolated proteins, 163 structures which uniquely existed in the NMR database (nothing similar in the X-ray database) were isolated and put in a database called NMRonly-set.

A fundamental assumption that allows interpretation of correlation seen on examining the database is that proteins with only a NMR structure were proved to be uncrystallizable. Since amounts of material needed for crystallization trials are similar to that needed for NMR and since determination of structures for small proteins is straightforward once crystals are found, this assumption seems to be reasonable. However, there may exist a potential for lack of interest in studying small proteins by crystallographers that could contribute to a bias in data. All of the conclusions of this paper are based on the absence of the mentioned bias.

Non-parametric density function estimation:

Bayesian analysis is a widely used and powerful classification tool⁸. This method utilizes the a-posteriori probabilities of an event in order to predict an a-priori probability of another related event. This prediction method highly relies on the accurate density estimation of events under study. In other words, more accurate probability density functions can lead to a more accurate prediction routine. Therefore it is important to try to estimate the probability density functions as accurately and precisely as possible.

Although histograms or other models can be used in constructing the probability density functions, Parzen density estimation⁸ remains one of the best non-parametric methods. This

method estimates the true probability density function of a stochastic event as shown in equations 1 and 2.

$$\hat{f}(X) = \frac{1}{N} \sum_{i=1}^N K\left(X_i, \frac{\hat{\sigma}}{\sqrt{N}}\right) \quad \text{Eq. (1)}$$

$$K(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} \quad \text{Eq. (2)}$$

In Equation 1, \hat{f} indicates the estimated probability density function, N corresponds to the number of data points. K indicates the unit kernel function with $\hat{\sigma}$ corresponding to the estimated sample standard deviation. The unit kernel used during the Parzen density estimation is simply a Gaussian kernel as described by Equation 2.

Parameters chosen for evaluation:

In our experiments we chose to estimate probability density functions to describe the crystallization dependence of proteins on certain parameters including size, percentage secondary structure and structural rigidity along a protein's backbone. As a measure of structural rigidity we use the rmsd of atomic positions along the protein backbone when structures are reported as sets of acceptable structures rather than a single structure. This is common for NMR structures. We define the following entity as the rmsd measure of any atomic position.

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_{xx}^2 & \hat{\sigma}_{xy}^2 & \hat{\sigma}_{xz}^2 \\ \hat{\sigma}_{xy}^2 & \hat{\sigma}_{yy}^2 & \hat{\sigma}_{yz}^2 \\ \hat{\sigma}_{xz}^2 & \hat{\sigma}_{yz}^2 & \hat{\sigma}_{zz}^2 \end{bmatrix} \quad \text{Eq. (3)}$$

$$rmsd = \sqrt{\hat{\sigma}_{xx}^2 + \hat{\sigma}_{yy}^2 + \hat{\sigma}_{zz}^2} \quad \text{Eq. (4)}$$

$\hat{\Sigma}$ in Equation 3 denotes the covariance matrix of the positional variation of any atom position. In this paper the rmsd values were calculated for the backbone amide proton of the protein in question.

Results

One Dimensional Analysis

The probability density functions (PDF) of the parameters speculated to be involved in crystallization properties of proteins were calculated and compared in order to establish the existence of any relationship. Figures 1, 2 and 3 illustrate the PDF for the fraction of proteins involved in α -helical (number of residues in α helices divided by the total number of residues), β -sheet and α/β structures respectively.

Another parameter potentially correlated to the crystallizability of a protein is its size. Size may not be a direct factor, but because size correlates with more complex issues such as the fraction of surface residues, it is a relevant parameter to observe⁹. Figure 4 illustrates the probability density function of the size of proteins present in the NMRonly-set as well as the Xray/NMR-set. There clearly are differences with proteins present in the Xray/NMR-set and NMRonly-set. If we ignore the influence of any interest bias, we could seek a threshold beyond which X-ray becomes the preferred method of structure determination. The simplified, optimal threshold determined by Bayesian analysis is estimated to be ~100 amino acid residues. The employment of this threshold as the criterion for the determination of the most successful method of structure determination provided the results shown in Table 1. The impact of the absence of

such decision-making criteria can easily be assessed by examining the distribution of protein size in a given organism such as *E. coli* shown in figure 5. Approximately 1573 proteins out of the 4290 of the entire genome of *E. coli* fall in an ambiguous region of 50-220 amino acids. This constitutes more than 1/3 of the entire *E. coli* genome.

Finally a more useful parameter that is speculated to be involved as a determinant of crystallization of a protein is the level of order (or structure) in a protein. Although it is reasonable to assume that a more flexible and dynamic protein will be more difficult to crystallize⁹, a correlation between flexibility and crystallizability has not been documented.

Unstructured regions of a protein can be tentatively identified by observing the backbone rmsd across the different models reported by the NMR spectroscopy for both the NMR-set and Xray-set. Counting the number of residues that exhibit a rmsd of larger than 4 Å can be used to quantify the level of disorder or unstructured portion of a protein. The comparison of the distribution of this measure (illustrated in Figure 6) for both the NMR-set and Xray-set provides the means of studying this phenomenon. Application of Bayesian algorithm to the distribution functions shown in Figure 6 provides a threshold of 0.05 fraction of a protein. A protein with unstructured region of more than 5% can therefore be considered to be a better candidate for NMR structure determination. Similarly a protein with unstructured region of less than 5% can then be considered to be a better candidate for X-ray structure determination. Application of this threshold to the task of method of structure determination produces the performances listed in Table 2. The utility in other case is enhanced because it seems possible to rapidly experimentally determine the percentage of unstructured residues prior to attempts at structure determination¹⁰.

Two Dimensional Analysis

Often times, the performance of statistical prediction engines improve as the dimensionality of the problem increases. Therefore, it is possible to develop a fully functional, two-dimensional predictor when the one-dimensional equivalent of the same predictor fails to perform. As an illustration we examined the α helical and β sheet parameters that showed little predictive value when examined separately.

Figures 7 and 8 illustrate the difference between the probability density function for the NMR-set and the Xray-set ($\text{PDF(Xray)} - \text{PDF(NMR)}$). In these figures, the two independent dimensions are the fraction of the molecule occupied in the α -helix and β -sheet secondary structures. This two-dimensional map can then be used to perform a more complex discrimination analyses.

In Figures 7 and 8, the positive/zero regions correspond to the area that X-ray crystallography would be the method of choice, while the negative regions correspond to the regions more amenable to NMR spectroscopy. This information can be utilized in order to calculate the performance of this classifier. These results are tabulated in Table 3.

Discussion and Conclusion

Based on the information presented in Figure 4, one can conclude that both NMR spectroscopy and X-ray crystallography can be utilized in the structure determination of proteins with the size of less than 220 amino acids. More detailed statistical analyses may however help refine this choice by including other information. Approximately 1573 proteins out of the 4290 of the entire genome of *E. coli* fall in the ambiguous region of 50-220 amino acids. This constitutes more than 1/3 of the entire *E. coli* genome. The data in Figure 4 suggests that the

optimal method of structure determination for proteins larger than 100 amino acids, is X-ray crystallography while proteins smaller than 100 amino acids are better candidates for NMR studies.

While careful examination of Figures 1, 2 and 3 on secondary structure correlation may not reveal significant differences between the two methods of structure determination; the combined two-dimensional analyses of the same data provide more definite results. Based on the results displayed on Figures 7 and 8, one can observe a bias towards the crystallization of proteins with high fraction of specific combinations of secondary structures. Furthermore, one can make the observation that proteins with β sheets are less likely to crystallize than ones with α helical structures. As a result, X-ray crystallography can then be concluded to be marginally more successful in the structure determination of highly α helical proteins. Conversely, proteins with high fraction of β sheets are more amenable for NMR spectroscopy.

The examination of Figures 1, 2 and 3 can conclude that in general, the fraction of any type of secondary structure alone can not be used in elucidating the method of structure determination. However the combined fractions of two types of structure determination can provide a classifier with performance of 75% accuracy in determining the likelihood of crystallization of a given protein.

Another interesting parameter that can correlate with crystallizability of a protein is the percentage of unstructured regions of that protein. Figure 6 clearly illustrates the difference in the distribution of the fraction of proteins with rmsd larger than 4 Å. There are currently convenient experimental methods for determining the unstructured fraction of a protein¹¹ and one can foresee the further development of rapid NMR methods of measuring (directly or indirectly)

this parameter. The experimental determination of this parameter can in the future provide a basis of initiating a more likely method of structure determination.

It must be kept in mind that the performance of any of the Bayesian classifiers mentioned in this paper may significantly increase by increasing the number of variables in the analysis. For example, the construction of a three dimensional space where the independent variables consist of fraction α helix, fraction β sheet and backbone rmsd may provide yet more powerful classifiers.

References

1. Brenner, S.E. and M. Levitt, *Expectations from structural genomics*. Protein Science, 2000. **9**(1): p. 197-200.
2. Orengo, C.A., A.E. Todd, and J.M. Thornton, *From protein structure to function*. Current Opinion in Structural Biology, 1999. **9**(3): p. 374-382.
3. Jaroszewski, L., et al., *Fold prediction by a hierarchy of sequence, threading, and modeling methods*. Protein Science, 1998. **7**(6): p. 1431-1440.
4. Fetrow, J.S. and J. Skolnick, *Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T-1 ribonucleases*. Journal of Molecular Biology, 1998. **281**(5): p. 949-968.
5. Skolnick, J. and J.S. Fetrow, *From genes to protein structure and function: novel applications of computational approaches in the genomic era*. Trends in Biotechnology, 2000. **18**(1): p. 34-39.

6. H.M.Berman, et al., *The Protein Data Bank*. Nucleic Acids Research, 2000. **28**: p. 235-242.
7. Durbin, R., et al., *Biological sequence analysis, Probabilistic models of proteins and nucleic acids*. 1998: Cambridge University Press.
8. Fukunaga, K., *Introduction to Statistical Pattern Recognition*. 2nd ed. 1990: Academic Press, Incorporated. 591pp.
9. Kwong, P.D., et al., *Probability analysis of variational crystallization and its application to gp120, the exterior envelope glycoprotein of type 1 human immunodeficiency virus (HIV-1)*. Journal of Biological Chemistry, 1999. **274**(7): p. 4115-4123.
10. Prestegard, J.H., et al., *Nuclear magnetic resonance in the era of structural genomics*. Biochemistry, 2001. **40**(30): p. 8677-8685.
11. Mori, S., et al., *Water exchange filter with improved sensitivity (WEX II) to study solvent-exchangeable protons. Application to the consensus zinc finger peptide CP-I*. Journal of Magnetic Resonance Series B, 1996. **110**(1): p. 96-101.

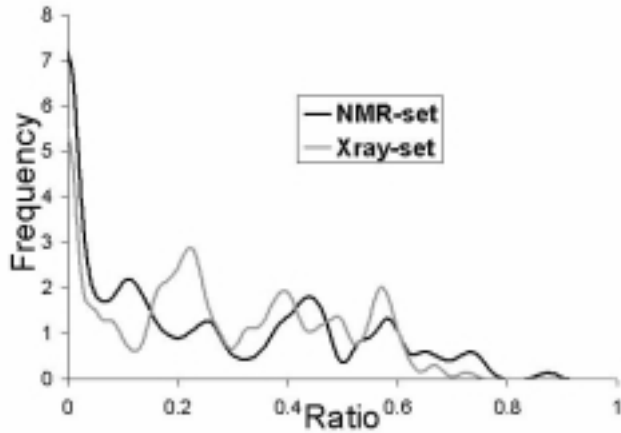


Figure 1. Distribution of the fraction of proteins in α helix structure.

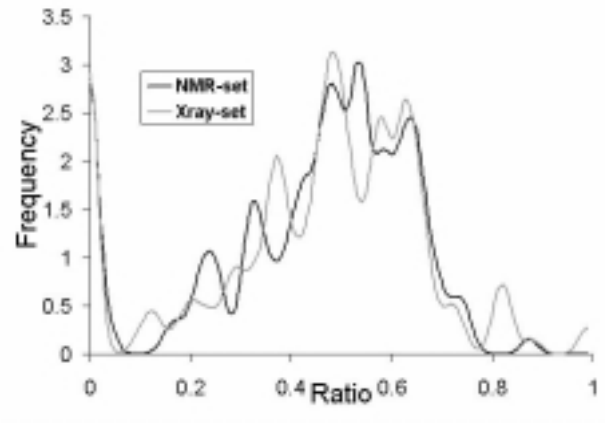


Figure 3. Distribution of the fraction of proteins in α -helix or β -sheet structures.

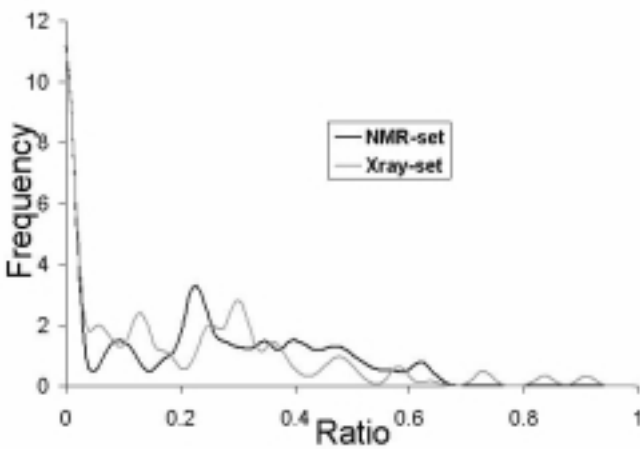


Figure 2. Distribution of the fraction of protein in β sheet structure.

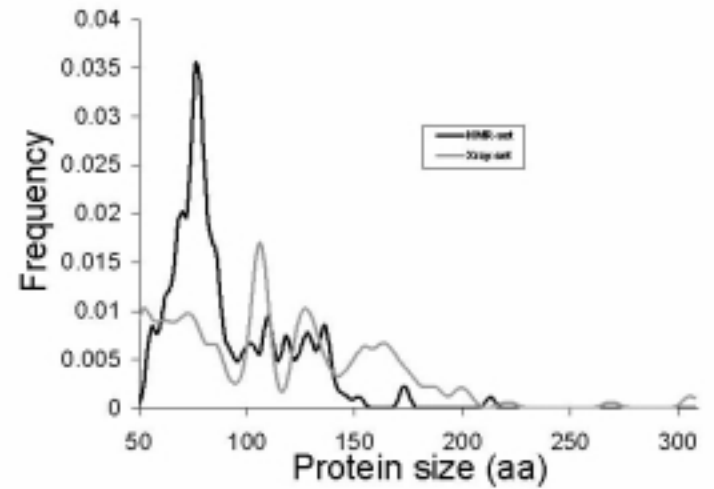


Figure 4. Distribution of protein size.

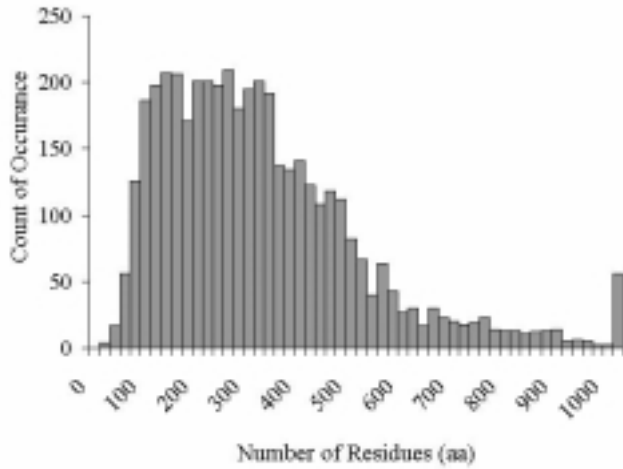


Figure 5. Distribution of the protein size in E. coli.

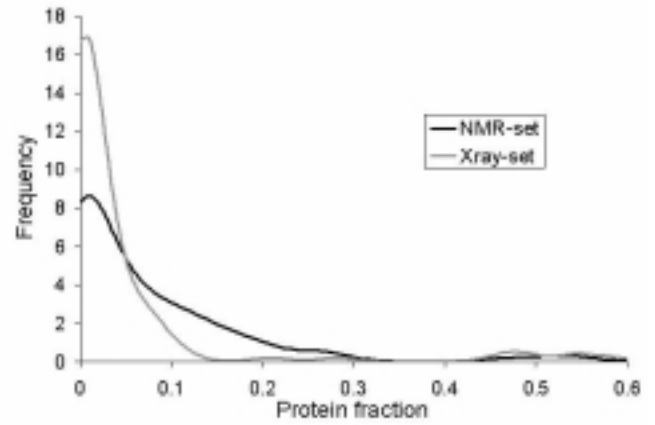


Figure 6. Unstructured fraction of proteins with rmsd > 4 Å.

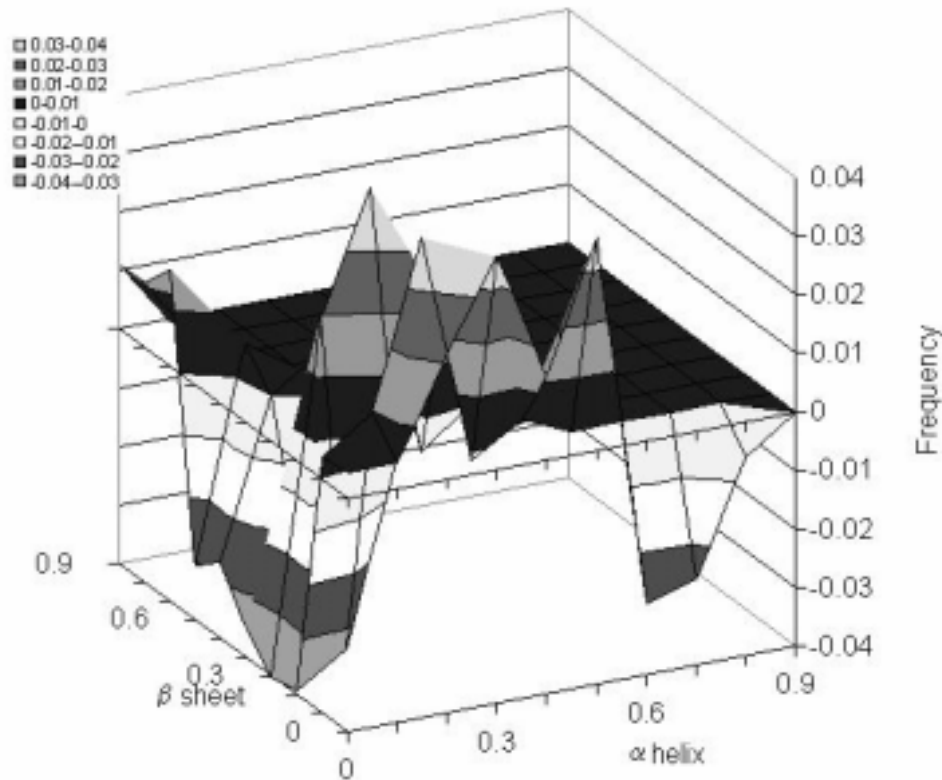


Figure7. Two dimensional plot of the difference between the PDF of Xray-set and NMR-set.

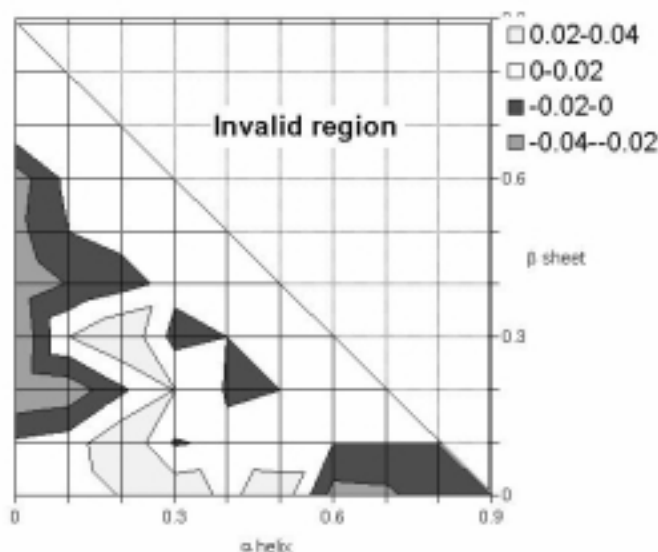


Figure8. Contour map of the difference between the X-ray-set PDF and NMR-set PDF..

Table 1. Performance of Bayesian classifier in determining the most likely method of success for structure determination by using 100 amino acid as the size threshold.

	Correct	Incorrect
NMR-set	70.1	29.9
X-ray-set	60.1	39.9

Table 2. Performance of Bayesian predictor in determining the most likely method of success for structure determination by using 0.05% fraction of protein backbone with rmsd larger than 4.0 Å.

	Correct	Incorrect
NMR-set	61.2	38.8
X-ray-set	66.0	34.0

Table 3. Performance of two-dimensional Bayesian predictor in determining the crystallization likelihood by using both information regarding α helical and β sheet fractions.

	Correct	Incorrect
NMR-set	55.2	44.8
X-ray-set	75.7	24.3