## UCSF

UC San Francisco Electronic Theses and Dissertations

## Title

Computational studies characterizing the information encoded in protein structures and sequences

## Permalink

https://escholarship.org/uc/item/7nw686nq

## Author

Aynechi, Tiba
Publication Date
2005

Peer reviewed|Thesis/dissertation


copyright (2004)
by
Tiba Aynechi


What eye can pierce the veil of truths? Or read the riddle of earth's destinies?

Pondered have I for years threescore and ten, But still am baffled by these mysteries


To mom, dad, and Amir. With love and gratitude.



## ACKNOWLEDGEMENTS

The work presented here would not have been possible without the help and support of many people. I consider myself very fortunate to have come to UCSF and the Graduate Group in Biophysics not just for the caliber of science offered, but also for the unique group of people that make up the programs and departments here. It is indeed very rare to find a collection of scientists who are not only renowned but also excel in their humanity and promote a research environment in which collaboration is so highly valued.

I must begin by thanking Julie Ransom for her kindness and support, particularly during those difficult first years. The students in the program owe a great deal of the camaraderie and warmth with which they are received to her.

I can only say that my fortune improved even greater once I joined the laboratory of Dr. Irwin Kuntz, who is known among students, friends and colleagues simply as Tack. Tack is a mentor in every sense of the word; an outstanding advisor whose zeal for science is paralleled by his kindness and humanity. He has always provided me with guidance, yet at the same time indulged my need for independence. I must also thank him for humoring my numerous travels, a privilege most advisors would not grant so generously. I am indebted to him.

My friends in the Kuntz group: Bill, Demetri, Jim, Karla, Kevin, Natasja, Rob \& Terry. I sincerely looked forward to coming to work everyday. Our discussions ranged from science, to politics, to the mundane. This past year, in particular, has been bittersweet for the few of us who moved to Mission Bay, as our small and tightly knit group began to disband and each of us parts for his/her separate way. I owe a special thank you to Dr. Robert Rizzo, who so graciously mentored me through my first Docking project

and later on in our collaboration presented in chapter six and To Dr. David Sullivan my colleague during the work presented in chapter three,

Several of the faculty at UCSF have been instrumental to my success in the past few years; members of my orals exam committee Drs. Patricia Babbitt, Tom Ferrin, Hao Li, and Volker Dotsch, and members of my thesis committee, Drs. Patricia Babbitt and Andrej Sali.

My gratitude also goes to Dr. James E. Hall for accepting me as an undergrad in his lab at UC Irvine and getting me interested in Biophysics. Although I eventually persuaded a topic far different than electrophysiology, Jim and the members of his lab, particularly Dr. Derek Beahm and Mary Hawley allowed me to experience the excitement and satisfaction of scientific research that eventually led me to graduate school.

I need to thank my friends and family outside of UCSF, for their love and support these past few years. My parents, whose sacrifices paved the way towards this Ph.D., I admire them for their exemplary dedication to their children. My sisters, Noyan and Mana for all their love and sharing all the good times as well as the bad times with me. Shirin and Mehdi for being so kind, always. Arash, Armita, Azin, Behrooz, Eric, Farshad, Jeyran, Kamran, Mandana, Nazanin, Peyman, Ramin, Reza, Roshanak, Tabassom, and ... thank you for your kind hearts and the moments we have shared.

Amir. For being my voice of reason, for keeping me calm and collected, for encouraging me, and for making me laugh. I am forever grateful for his unwavering love
 and support.


## COMPUTATIONAL STUDIES CHARACTERIZING THE INFORMATION ENCODED IN PROTEIN STRUCTURES AND SEQUENCES <br> Tiba Aynechi


#### Abstract

Proteins are often responsible for human diseases. Furthermore, their function and biological role is defined by their three dimensional structures. The field of therapeutic design has undergone major leaps in the past two decades, 1) due to our increased understanding of biological systems 2) the availability of large amount of sequence data, and 3) the exponential growth of computing power coupled with advances structure based drug design techniques.

Experimental structure determination is time consuming and not practical for large-scale processing. Comparative modeling, which relies on sequence similarity, is often used to identify relatives of unknown proteins. This thesis begins by examining the role of distance constraints as an alternative metric for similarity when looking for fold relatives. However we find that in the absence of clear definitions for similarity an objective method can not be developed.

We then shift our focus to quantifying the information in distance constraints using information theory. We use sets of exhaustive lattice walks to develop numerical measures of the information content of sets of exact distance constraints applied to specific conformational ensembles. We examine the effects of experimental uncertainties by considering "noisy" constraints.

We extend the use of information theory and simplified models in the following two chapters to quantitatively analyze the protocols involved in comparative modeling.


We begin by deriving the ideal costs of sequence alignments and gap penalties based on gap distributions using exhaustive sequence set with simplified alphabets. We show that there are different gap penalties for different alphabet sizes and that there can be dependencies on the length of the sequences being aligned. In addition we use two dimensional lattice models to quantify the relative resolving power of some commonly used force fields. We show that long-range intra-atomic interaction are the most informative,

The last chapter of this thesis is an investigation of charge models in calculations of free energies of binding. Through the use of a large test set, we show that optimization of parameters, specifically those involved in calculating the non polar contributions to the free energy, can significantly increase correlation of free energies with those obtained from experiment.


## TABLE OF CONTENTS

Chapter One
Introduction ..... 1
Chapter Two: Use of Distance Constraints in the Development of a Sequence Independent Fold Recognition Method
Abstract ..... 14
Introduction ..... 15
Background ..... 17
Methods ..... 20
Developing a fold recognition Algorithm (DP) using distance constraints ..... 20
Test Sets ..... 24
Results and Discussion ..... 27
Acknowledgments ..... 35
References ..... 36
Chapter Three: Information Content of Molecular Structures
Keywords ..... 38
Abstract ..... 39
Introduction ..... 40
Theory and Methodology ..... 43
Ensemble Generation ..... 43
Entropy and Information ..... 45
Nomenclature ..... 46
Example of the information content of a constraint. ..... 47
Discrete noisy systems ..... 48
Conformer distributions ..... 50
Problem Formulation ..... 52
Results. ..... 52
Exact constraints ..... 53
Information loss from uncertainties in distance constraints ..... 64
Information per constraint. ..... 79
Discussion ..... 80
Conclusions ..... 82
Acknowledgments: ..... 84
References ..... 85
CHAPTER FOUR: AN INFORMATION THEORETIC APPROACH ..... то Macromolecular Modeling: I. Sequence Alignments
Keywords ..... 88
Abstract ..... 89
Introduction ..... 90
Methods ..... 91
Overview ..... 91
Entropy and Information ..... 93
Sequence Alignment ..... 93
Overview ..... 93
Gapless Alignments ..... 95
Gapped Alignments ..... 96
Gap Penalties ..... 97
Search Algorithm Methods ..... 97

First Occurrence Alignments ..... 97
Multiple Occurrence Alignments ..... 98
Exhaustive vs. Incomplete Sequence Sets ..... 98
Mapping ..... 98
Correlation of Sequence Alignment and Conformational Resolution ..... 99
Relationship to standard alignment efforts ..... 100
Results ..... 100
Discussion ..... 108
Acknowledgments ..... 110
References ..... 111
Chapter Five: An Information Theoretic Approach to Protein Modeling: II. Force Fields
Keywords ..... 113
Abstract ..... 114
Introduction ..... 115
Methods. ..... 117
Overview ..... 117
Entropy and Information ..... 118
Protein Models ..... 118
Simplified Potentials ..... 120
HP interaction model ..... 120
Solvation Model ..... 121
Electrostatic Interactions ..... 121
GO Type Potentials ..... 122
Results ..... 123
HP Interactions ..... 124
Solvent Interactions ..... 126
Electrostatic Interactions ..... 127
Go-type Potential ..... 129
Discussion ..... 130
Acknowledgments. ..... 131
References ..... 132
Chapter Six: Estimation of Absolute Free Energies of Hydration using Continuum Methods: Accuracy of Partial Charge Models and Optimization of Nonpolar Contributions
Keywords ..... 135
Absract ..... 136
Introduction ..... 137
Computational Methods ..... 141
Free Energies of Hydration ( $\Delta \mathrm{G}_{\text {hyd }}$ ) ..... 141
Computation details ..... 142
Molecular structures and experimental data ..... 143
Partial Charge Models ..... 144
Molecular geometries ..... 145
Results and Discussion ..... 145
Charge Model Evaluation ..... 145
PBSA vs.GBSA ..... 149
G nonpolar from Molecular SA ..... 149
$\mathrm{G}_{\text {nonpolar }}$ from Atomic $\mathrm{SA}_{\mathrm{i}}$ ..... 151
Optimized SA Coefficients ..... 155
Monoatomic ions ..... 157
Conclusions ..... 160
Acknowledgments ..... 163
References ..... 164
Appendix ..... 169
Chapter Seven
Concluding Remarks ..... 177

List OF TABLES
TABLE 1: Library of similar folds as designated by CATH ..... 25
TABLE 2: 68 protein pairs from Fischer et. al ..... 26
TABLE 3: Comparison of a beta strand from 2AK3A to a set of secondary structures derived from the 68benchmark. ..... 31
Chapter Three
TABLE 1: Self-avoiding square-lattice walks ..... 44 ..... 4
TABLE 2: Square Hamilton walks ..... 45
TABLE 3: Information content for lattice walks ..... 54
Chapter Four
,
TABLE 1: Gapless alignments for exhaustive sequence sets Multiple occurrences ..... 101
TABLE 1. Gapless alignments for exhaustive sequence sets Multiple occurrences ..... 1
TABLE 2: Gapless alignments for exhaustive sequence sets - First Occurrence ..... 101
TABLE 3: Gapped alignments for exhaustive sequence sets - Multiple occurrences ..... 102
TABLE 4: Gapped alignments for exhaustive sequence sets - First Occurrence ..... 103
TABLE 5: Gap distributions and gap penalties for the first occurrence model.
TABLE 5: Gap distributions and gap penalties for the first occurrence model. ..... 105 ..... 105
TABLE 6: Gap distributions and gap penalties for the multiple occurrence model
TABLE 6: Gap distributions and gap penalties for the multiple occurrence model ..... 105 ..... 105
Chapter Five
TABLE 1: HP sequences and self avoiding 2D walks ..... 119
TABLE 2: Compact and semi-compact hamilton walks ..... 120
Chapter Six
TABLE 1: Atom type, radii, and GB scaling factor ..... 144
TABLE 2: Correlation coefficients ( $\mathrm{r}^{2}$ ) and average unsigned errors (AUE) for experimental vs. calculated PBSA/GBSA $\Delta \mathrm{G}_{\text {hyd }}$ ..... 147
TABLE 3: Correlation coefficients ( $\mathrm{r}^{2}$ ) and average unsigned errors (AUE) for experimental vs. calculated PBSA/GBSA $\Delta \mathrm{G}_{\text {hyd }}$ ..... 153
TABLE 4: Optimized atomic SA coeffecient from PB ..... 155

## Chapter Two <br> Chapter Two

TABLE 5: Optimized atomic SA coeffecients from GB ..... 156
TABLE 6: Residuals ${ }^{\text {a }}$ for monoatomic ions with experiment. ..... 156
TABLE 7: Optimized atomic SA coeffecients for single ions ..... 158
List of Figures
Chapter Two
FIGURE 1 Similarity matrix $S$ for two proteins ..... 23
FIGURE 2 Behavior of the similarity score for various values of M and $\mathrm{d}_{0}$ ..... 23
FIGURE 3: Distribution of DP Alignment scores ..... 28
FIGURE 4: Distribution of alignment scores ..... 29
FIGURE 5: Coverage versus error plot for DP using 110 member fold library ..... 30
FIGURE 6: Topology diagrams for a pair of related proteins 1aaj, 1 paz and 2 fox ..... 32
FIGURE 7: . Distance constraints derived from cross linking experiments ..... 34
Chapter Three
FIGURE 1: Information content, I, for each distance element $[d]_{i, j}$ for $\mathrm{N}=15$. ..... 54
FIGURE 2: Mean information content as a function of $s$ ..... 55
FIGURE 3: Information content by sequence separation ..... 56
FIGURE 4: Relative information loss $\Delta \mathrm{I} / \mathrm{I}$ is plotted for all distance elements $[d]_{i, j}$ ..... 57
FIGURE 5: Relative information loss, $\Delta \mathrm{I} / \mathrm{I}$, for all pairs of distances ..... 59
FIGURE 6: . Maximum information $\mathrm{I}^{\mathrm{M}}$ for best sets of distance constraints ..... 61
FIGURE 7: Information content dependence on number of constraints. ..... 62
FIGURE 8: Distance constraints for percentage information. ..... 62
FIGURE 9: $\mathrm{I}^{\mathrm{M}}$ with upper bounds on distances. ..... 66
FIGURE 10: Information content, $\mathrm{I}\left[d_{i, j}\right]$ by sequence separation with bounded distance detection. ..... 67
FIGURE 11: Information content of contact/no contact determinations. ..... 68

$$
\frac{1 / 2}{2}
$$

FIGURE 12: . Limiting conformations on 2D square lattices ..... 69
FIGURE 13: Conformational distinguishability ..... 70
FIGURE 14: Conformational Indistinguishability. ..... 71
FIGURE 15: Marginal dimensionality, $n$, plotted as a function of the chain length $N$ ..... 72
FIGURE 16: Mean information content for Hamilton square walks ..... 73
FIGURE 17: Conformational indistinguishability for Hamilton walks ..... 74
FIGURE 18: Conformational indistinguishability for stochastic polyalanine ensembles. 75
FIGURE 19: Relative information loss for [ $d]_{l, N}$, full enumeration ensemble ..... 77
FIGURE 20: Information loss per degree of freedom ..... 78
FIGURE 21: Information loss per degree of freedom for compact 2D lattice structures. ..... 78
Chapter Four
FIGURE 1: Search algorithm for finding all occurrences of $T$ is $S$ ..... 98
FIGURE 2: Distribution of gap initiation and extension penalties. ..... 103
FIGURE 3: Average number of hits for all possible 5mer templates in 32 related 20 mer evolutionary subsets ..... 107
FIGURE 4: Gap length distributions in multiple occurrence alignments for 32 , 5 mer binary templates in a 20 mer evolutionary subset ..... 107
Chapter Five
FIGURE 1: Information of HP and solvent contact ..... 124
FIGURE 2: Information of force fields for compact structures ..... 127
FIGURE 3: Information from coulombic pair-wise interactions ..... 128
Chapter Six
FIGURE 1: Experimental free energies of hydration versus total molecular solvent accessible surface area. ..... 146
FIGURE 2: Comparison of $\mathrm{r}^{2}$ values for calculated versus experimental free energies of hydration from PBSA and GBSA calculations ..... 148


FIGURE 3: Correlation of individual components with experimental free energies of hydration for neutral compounds ..... 151
FIGURE 4: Predicted $\Delta \mathrm{G}_{\mathrm{hyd}}$ vs. experimental $\Delta \mathrm{G}_{\mathrm{hyd}}$ from PBSA calculations with RESP charges for neutral compounds $(\mathrm{N}=460)$ ..... 154
FIGURE 5: Predicted $\Delta \mathrm{G}_{\text {hyd }}$ vs. experimental $\Delta \mathrm{G}_{\mathrm{hyd}}$ from PBSA calculations with RESP charges for charged compounds ( $\mathrm{N}=42$ ) ..... 154
FIGURE 6: $\Delta$ Gpolar for monoatomic ions vs experimental $\Delta$ Ghyd ..... 160


## Chapter One: Introduction



The focus of this dissertation is on the application of information theory (Shannon, 1948) to molecular modeling. There are several reasons as to why this is a timely subject and I will discuss some of these motivations in the paragraphs below.

In the era of genomics, proteomics, and many other 'omics', computation is becoming an increasingly integral part of biology. The impetus for this trend is two-fold, one being the availability of more powerful computing tools and two, the ever increasing volume of available biological data. It is now clear that understanding the three dimensional structure of proteins offers crucial insights into molecular interactions. The formation of groups such as the Structural Genomics Consortium (Burley, 1999), dedicated to large-scale structure determination, highlights the potential impact of the knowledge embedded in protein structures.

From a drug design point of view, the goal is to be able to either predict or design therapeutics for as many target proteins as possible, in the least amount of time possible. Drug discovery is a multi-step process. It begins with the identification of a biomolecular target with therapeutic value followed by numerous rounds of screening chemical compounds in search of drug-like candidates that either interfere or bind the molecular target.

The technological advances of the past two decades have opened the door to structure-based drug design in which the three-dimensional structure of the target protein is used to guide the selection of compounds (Kuntz, 1992) (Jorgensen, 2004). To date, there are many instances of drugs on the market or in clinical trials that were discovered via the structure-based design approach (Malikayil and Hardy, 2003). Most commonly, scientists use atomic structures derived largely from either x-ray crystallography or NMR

experiments as their design templates. However, these experimental methods are time consuming, difficult to perform, and not amenable to large scale automation. In light of the enormous amounts of genomic data that is available today, they seem impractical. Specifically, while to date there are close to 28,000 three-dimensional structures in the Protein Data Bank (Berman, 2000), the number of sequences with no associated structure is several orders of magnitude larger. The problem is also exacerbated by the fact some fold families are over represented while others are yet to be observed (Govindarajan, 1996, 1999). As such, efforts are focusing on developing computational methods that will draw on the available structures in the Protein Data Bank as well our understanding of the laws of physics, towards the development of large-scale protein structure prediction methods in silico (Sanchez, 1998).

Methods in theoretical structure prediction can be divided in to two categories, 1) comparative and 2) ab intio modeling. In the former case, building a model requires that one identify a template, calculate an alignment between the template and the target, and finally build and refine the model based on the template. This process is modular and errors in any one step can not be corrected in subsequent steps (Sanchez and Sali, 1997). Sequence and structure similarity are highly correlated (Chothia and Lesk, 1986). This relationship is, however, only true in one direction, meaning similar sequences have similar structures, but the reverse is not always the case. Homology-based methods exploit this relationship in determining structures to serve as templates for new structures. However, a major bottle neck of this approach is identifying appropriate templates for sequences that fall in the so called similarity 'twilight zone' (Blake and Cohen, 2001) where similarity among sequences falls below 30 percent.


Dynamic programming is the most popular algorithm for sequence alignment (Needleman and Wunsch, 1970; Waterman and Vingron, 1994) and requires two types of parameters, a substitution scoring matrix (Henikoff and Henikoff, 1992) (Dayhoff, 1978) (Tomii and Kanehisa, 1996) and a penalty for introducing gaps, both of which are derived via empirical optimization. The algorithm calculates the highest alignment score possible and produces a single alignment with that score, with no further information on similar or alternative alignments. Exploring all alignments for average size proteins is computationally unfeasible due to combinatorial explosion. This results in sequencesequence or structure-sequence alignments that do not always agree with structurestructure alignments. Furthermore, depending on the scoring method used for aligning a pair of structures one can arrive at varying solutions (Godzik, 1996) (Jaroszewski et al., 2000). Advances in fold recognition and threading algorithms have certainly extended the reach of comparative modeling, but nevertheless alignment errors still persist (Venclovas et al., 2003). Another shortcoming of comparative modeling is prediction of novel folds. In the absence of a template an appropriate model can not be built.

Protein conformation space is vast (Sullivan and Kuntz, 2001) and the prediction
 of structure from sequence would require searching among numerous decoys. Ab initio methods, attempt to understand the free energy landscape of protein space to produce a minimum energy structure given a sequence. Molecular mechanics force fields are the cornerstone of computer simulations for proteins; they make use of simplified potentials and reduced representations to solve otherwise impossible problems (Lee et al., 2001). In recent years, force fields (Cornell et al., 1995; Wang et al., 2001) have been widely used to refine low resolution structures of proteins, however due to computational limitations
it has not yet been possible to extend the simulations to de novo prediction, other than for very small peptides (Duan and Kollman, 1998). Atomic force fields are made up of a potential energy function that takes into account various pairwise atomic interactions. The functions are heavily parameterized using various charge models derived either empirically or quantum mechanically (Bayly et al., 1993; Jorgensen et al., 1996; Li et al., 1998). Although these methods have been successful in the past, full understanding of their failures is not possible without their application to exhaustive and fully enumerated structure sets.

In order to circumvent the sampling problems mentioned, many methods employ models that significantly reduce protein complexity. These models fall into two categories: lattice and off-lattice models. Due to their relative analytical and computational simplicity lattice models have long been used to study the nature of polymers and compact conformations (Chan and Dill, 1989). Energy functions are simple to evaluate on these models and the conformational space can be searched exhaustively. However, they are not without shortcomings, including their failure to accurately model secondary (Park and Levitt, 1995) structures; but it has been shown that their advantages outweigh their inadequacies. Off-lattice models often simplify proteins by eliminating side-chain degrees of freedom; however they are still too complex for exhaustive simulations. To evaluate the energies of these simplified models, energy functions are devised so as to be computationally efficient, yet still representative of the forces responsible for protein structure. Potential functions can either be physics-based as in the molecular dynamics force fields, or they can be empirically derived such as potentials of mean force. While physics-based potentials are far more accurate they are


computationally expensive. Empirical score functions can be much more efficient, however they are derived from protein databases and are biased toward the arrangement of amino acids in known proteins, and hence disfavor the rarer structures.

The approaches above do not always produce structures with desired resolutions. In the absence of an appropriate template comparative modeling methods are not useful and as mentioned before, ab initio methods rely mainly on scoring functions whose forms are derived from the laws of physics but are parameterized by small molecules or protein databases. Although they are successful in narrowing the conformational search, they can not produce structures with resolutions greater than $3 \AA$ (Bonneau and Baker, 2001). The incorporation of experimental data such as contact order and distance constraints from cross linking can remedy some of the inaccuracies (Bonneau et al., 2002; Shakhnovich and Gutin, 1990) .

It is now apparent that biological sequences are molecular messages. As mentioned earlier, the sequences of amino acids determine the structure of protein molecules. However these structures are degenerate with many examples of different sequences leading to similar structures. In addition, there are molecules with no structural similarity that perform the same biochemical functions. These examples highlight the significance of questions regarding the information encoded in biological sequences and lead us to further ask how best we can define and quantify such information.

From a structure prediction point of view, information content is viewed as that which is required to map a sequence to a unique structure. This approach parallels that of classical information theory in which the information content of a string is the bits of data required to transmit the signal. In this thesis, we explore the property of molecules as
molecular messages and aim to draw inferences about the information encoded in them using the basic tenants of information theory.

In chapter two, we begin to qualitatively assess the information value of distance constraints in protein prediction. It is now possible to obtain experimentally derived distance constraints for proteins using chemical cross linkers in conjunction with mass spectrometry (Young, 2000). Such constraints can then be used to reduce the search space for compatible protein structures. In addition, it is reasonable to assume that any three-dimensional shape can be characterized by a unique set of constraints (Havel, 1983). So theoretically, in the absence of sequence similarity among proteins with known structures, we may be able to identify structural homologues using distance constraints alone. Through the use of a dynamic programming algorithm we investigate whether it is possible to identify structures belonging to the same fold family as defined by the CATH database (Orengo CA, 1997). In order to design a benchmark test set containing both related and unrelated structures, one must rely on protein classification schemes of which there are several, and none agree completely (Hadley, 1999). Furthermore, we could never realize a complete test set containing all possible protein folds and hence any conclusions would be inherently biased towards the members of the test set.

In chapter three, we switch from the empirical world of real proteins and structures to the more analytical realm of two-dimensional lattice models. We employ information theory to deduce the information content of distance constraints in fully enumerated lattice walks. We also further examine the effects of experimental uncertainties by considering noisy constraints. Our findings are subsequently expressed in terms of information per degree of freedom in the chain. This approach provides a
quantitative means for comparing various constraint sets and allows us to dissect the results into a form that is independent of chain length.

As previously mentioned, sequence alignments remain a bottle neck in the structure prediction process. In chapter four, our aim is to quantify the difficulty and the information cost of performing sequence alignments using exhaustive sequence sets generated with simplified alphabet models. There are two important parameters involved in aligning two sequences, one is the scoring matrix and second is the choice of gap penalties. At present both these parameters are derived empirically via optimization methods. Using information theory allows us to present an analytical framework for evaluating their proper values by utilizing gap distribution functions derived from exhaustive sets.

In chapter five we shift our focus onto force fields often used in structure minimization. As is the case in sequence alignment protocols, force field parameters and terms are optimized empirically. Although in theory most energy potential functions are evaluated as a sum of pairwise atomic interactions, the informational contributions of the various interaction types are unclear. The simplified lattice models are used to generate both extended and compact conformations. We then evaluate their energies based on simple interaction schemes such as nearest-neighbor contact and long-range coulomb type interactions. Analyzing the distributions of various energy levels using information theory can determine the relative resolving power of these simplified force fields.

Although, an analytical analysis of simple model systems offer an in depth understanding of the system's behavior, computational limitations still prevent us from studying real protein system analytically. In chapter six, we move back to the empirical

optimization regime in order to improve the accuracy of free energy calculations. Most molecular species are charged and consequently electrostatics play an essential role in biological interactions (Honig and Nicholls, 1995). Inter-molecular interactions are driven by favorable changes in free energy. Approximation methods such as the Generalized Born Model (Still et al., 1990) and Poisson-Boltzmann (Sitkoff et al., 1994) calculations have been developed in order to allow free energy calculations suitable for high throughput processing. Our focus in the present study is to evaluate the accuracy of eight different point charge models typically used for structure-based drug design calculations through computation of hydration free energies. These models can easily be assigned, in an automated fashion, to relatively large and diverse data sets. Our results show that further optimization using an extended data set greatly improves the agreement between theoretical experimental calculations.


## REFERENCES:

Bayly, C.I., P. Cieplak, W.D. Cornell, and P.A. Kollman. 1993. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the Resp Model. Journal of Physical Chemistry 97(40):1026910280.

Berman, H.M., Westbrook,J., Feng,Z.,Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., Bourne, P.E. 2000. The Protein Data Bank. Nucleic Acids Research 28:235-242.

Blake, J.D., and F.E. Cohen. 2001. Pairwise sequence alignment below the twilight zone. J Mol Biol 307(2):721-35.

Bonneau, R., and D. Baker. 2001. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct 30:173-89.

Bonneau, R., I. Ruczinski, J. Tsai, and D. Baker. 2002. Contact order and ab initio protein structure prediction. Protein Sci 11(8):1937-44.

Burley, S.K., Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland,T., Lin, D., Sal,i A., Studier, F.W., Swaminathan, S. 1999. Structural genomics: beyond the human genome project. Nature: Genetics 23(2):151-157.

Chan, H.S., and K.A. Dill. 1989. Compact Polymers. Macromolecules(22):4559-4573.
Chothia, C., and A.M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. Embo $J$ 5(4):823-6.

Cornell, W.D., P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman. 1995. A second generation force fiels for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society 117(19):5179-5197.

Dayhoff, M.O. 1978. A model of evolutionary change in proteins. Atlas of protein sequence and structure. Silver Spring, Md.,: National Biomedical Research Foundation. p 345-352.

Duan, Y., and P.A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 282(5389):740-4.

Godzik, A. 1996. The structural alignment between two proteins: is there a unique answer? Protein Science 5(7):1325-38.

Govindarajan, S., Goldstein, R.A. 1996. Why are some proteins structures so common? Proceedings of the National Academy of Sciences of the United States of America 93(8):3341-3345.


Govindarajan, S., Recabarren, R., Goldstein, R.A. 1999. Estimating the total number of protein folds. Proteins: Structure, Function, and Genetics 35(4):408-415.

Hadley, C., Jones, T.,. 1999. A systematic comparison of protein structure classifiactions: SCOP, CATH and FSSP. Structure 7:1099-1112.

Havel, T., Kuntz, I.D., Crippen, G. 1983. The Theory and Practice of Distance Geometry. Bulletin of Mathematical Biology 45(5):665-720.

Henikoff, S., and J.G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89(22):10915-9.

Honig, B., and A. Nicholls. 1995. Classical electrostatics in biology and chemistry. Science 268(5214):1144-9.

Jaroszewski, L., L. Rychlewski, and A. Godzik. 2000. Improving the quality of twilightzone alignments. Protein Sci 9(8):1487-96.

Jorgensen, W.L. 2004. The many roles of computation in drug discovery. Science 303(5665):1813-8.

Jorgensen, W.L., D.S. Maxwell, and J. Tiradorives. 1996. Development and Testing of the Opls All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. Journal of the American Chemical Society 118(45):1122511236.

Kuntz, I.D. 1992. Structure-based strategies for drug design and discovery. Science 257(5073):1078-82.

Lee, M.R., J. Tsai, D. Baker, and P.A. Kollman. 2001. Molecular dynamics in the endgame of protein structure prediction. J Mol Biol 313(2):417-30.

Li, J.B., T.H. Zhu, C.J. Cramer, and D.G. Truhlar. 1998. New Class Iv Charge Model for Extracting Accurate Partial Charges from Wave Functions. Journal of Physical Chemistry 102(10):1820-1831.

Malikayil, A., and L.W. Hardy. 2003. The impact of structure-guided drug design on clinical agents. Current drug discovery:15-20.

Needleman, S.B., and C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3):443-453.

Orengo CA, M.A., Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH--a hierarchic classification of protein domain structures. Structure 5(8):1093-1108.

Park, B.H., and M. Levitt. 1995. The complexity and accuracy of discrete state models of protein structure. J Mol Biol 249(2):493-507.

Sanchez, R., and A. Sali. 1997. Advances in comparative protein-structure modelling. Curr Opin Struct Biol 7(2):206-14.

Sanchez, R., Sali, A. 1998. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proceedings of the National Academy of Sciences of the United States of America 95:13597-13602.

Shakhnovich, E., and A. Gutin. 1990. Enumeration of all compact conformations of copolymers with random sequence of links. Journal of Chemical Physics 93(8):5967-71.

Shannon, C.E. 1948. A mathematical theory of communication. The Bell System Technical Journal(27):379-423 and 623-656.

Sitkoff, D., K.A. Sharp, and B. Honig. 1994. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. Journal of Physical Chemistry 98(7):1978-1988.

Still, W.C., A. Tempczyk, R.C. Hawley, and T. Hendrickson. 1990. Semianalytical Treatment of Solvation For Molecular Mechanics and Dynamics. J. Am. Chem. Soc. 112:6127-6129.

Sullivan, D.C., and I.D. Kuntz. 2001. Conformation spaces of proteins. Proteins 42(4):495-511.

Tomii, K., and M. Kanehisa. 1996. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng $9(1): 27-36$.

Venclovas, C., A. Zemla, K. Fidelis, and J. Moult. 2003. Assessment of progress over the CASP experiments. Proteins 53 Suppl 6:585-95.

Wang, W., O. Donini, C.M. Reyes, and P.A. Kollman. 2001. Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, proteinligand, protein-protein, and protein-nucleic acid noncovalent interactions [Review]. Annual Review of Biophysics \& Biomolecular Structure 30:211-243.

Waterman, M.S., and M. Vingron. 1994. Rapid and accurate estimates of statistical significance for sequence data base searches. Proc Natl Acad Sci US A 91(11):4625-8.

Young, M., Tang, N., Hempel, C., Oshiro, C., Taylor, E.W., Kuntz, I.D., Gibson, B.W., Dollinger, G. 2000. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America 97(11):5802-5806.

# Chapter Two: Use of Distance Constraints in the Development of a Sequence Independent Fold Recognition Method 

Tiba Aynechi, Irwin D. Kuntz


#### Abstract

In the absence of sufficient sequence similarity, it is not possible to identify homologs for proteins of unknown function and structure. It has been demonstrated that a three dimensional structure can be constructed using a set of distance constraints. It has also been demonstrated that for a protein of length $\mathrm{L}, \mathrm{L} / 10$ constraints are sufficient for enriching threading methods. This chapter is a preliminary study of the use of distance constraints for identifying fold relatives. We can identify fold relatives for certain proteins in the absence of any sequence information, However due to the ambiguities in the definition of similarity the development of a proper test set is a challenge that must be overcome prior to further investigation.





## INTRODUCTION

The completion of the genome-wide sequencing projects of many organisms has resulted in a large amount of data awaiting analysis. Specifically, structural genomics hopes to be able to determine the structures of proteins coded by these sequences and ultimately determine their function in order to investigate their biological implications and aid in the development of therapeutic agents. At present, structure determination has become the rate-limiting step in this effort, and as such there is a great need for methods that facilitate fast, high throughput, and reliable structure determination.

The experimental techniques for structure determination, i.e. crystallography and NMR are not suitable as fast, high throughput processes. In spite of scale-up efforts, the time, effort, and limitations involved in protein preparation, size, and type make these methods unfeasible. On the other hand, the current theoretical methods such as homology modeling, threading, and ab initio prediction are data limited. Homology modeling relies heavily on sequence similarity between a probe protein and proteins with experimentally determined structures. Threading methods have a limited search space defined by the number of experimentally observed unique protein structures. While both methods will improve over time, they will fail in cases where a new fold may be involved. Ab initio methods attempt to build models using first principles of folding and energetics, however they are not yet reliable enough due to our limited understanding of the physical principles involved.

Currently the Protein Data Bank (Berman, 2000) contains in excess of 27000 structures, representative of only 800 protein folds according to release 1.65 of SCOP (Murzin, 1995). Although experimental structure determination serves as the gold

standard for all other prediction methods, the current volume of data and the need for efficiency dictate that experimental structure determination efforts focus on those proteins that would most expand the data sets used in theoretical methods, i.e. those with potentially novel folds. Target selection has become the most important strategic issue faced by structural genomics, whose performance will be measured in part by the number of structures determined and the fraction that contain novel folds (Burley, 1999). The aim is to place most protein structures within a 'modeling distance' of at least one known structure. However, due to the fact that many unrelated sequences share the same structure, it is not possible to select novel targets using sequence comparison methods alone. Reliable, rapid fold identification is thus a timely and important task of structural genomics efforts.

It has been shown previously that any configuration of points in space can be characterized by a number of distance constraints among those points (Kuntz, 1979). Thus, in principle, given a set of constraints related to a conformation of points in space and a set of sample conformations, one should be able to determine whether that conformation belongs to the sample set or not. This approach has been extended to protein folds. It has been shown that the determination of a set of constraint for a particular protein fold improves the ability of threading protocols in identifying structurally similar proteins (Young, 2000). We propose that the use of distance information, when applied on a large scale, can serve as both a novel fold filter as well as a sequence-independent fold prediction method.

## BACKGROUND

In order to gain a global and comprehensive view of proteins, it is necessary to compare multiple structures and investigate their fold similarities and evolutionary relationships. The task requires the comparison of large numbers of three- dimensional shapes, currently on the order of 27,000 in the Protein Data Bank and categorizing them based on a set of similarity criteria. What remains an unresolved issue is a definition for similarity. In defining similarity, we are faced with two issues: determining a metric for similarity, such as size, shape, sequence, function, etc.; and defining a spectrum for that metric that spans from similar to dissimilar. While structure and substructure identity is a wellformulated mathematical problem, the definition of structural similarity contains some genuine ambiguities that have resulted in the development of multiple classification schemes each with their merits and shortcomings. The most widely used and comprehensive databases dealing with structural similarity are SCOP, CATH, and FSSP, which represent three unique methods; purely manual, a combination of manual and automated, and fully automated respectively and use different classification schemes.(Hadley, 1999).

SCOP, organizes proteins in a hierarchy, class being at the highest level followed by fold, superfamily and family (Murzin, 1995) The process is carried out by the visual inspection and comparison of protein structures. Proteins in the same class, share the same type of secondary structures, for example all alpha, or all beta (Chothia, 1977). The fold level implies similar packing and chain topologies, while those in the same superfamily have structural and functional features in common. The family level includes those proteins with similar sequences.

CATH, another hierarchical scheme, uses a combination of manual and automated methods. Proteins are grouped into class(C), based on the types of secondary structures, architecture(A) based on the general arrangement and composition of SSEs, topology(T) based on the connectivity, and homologous superfamily $(\mathrm{H})$ based on evolutionary relationships. The A level assignment is done via visual inspection. Domains sharing the same CAT designation have the same fold whereas a shared $H$ level implies an evolutionary relationship (Orengo CA, 1997).

FSSP uses a fully automated method based on structure-structure alignment of proteins. The method does not use the same hierarchical scheme as the previous two. The metric for similarity is a Z-score, which is the number of standard deviations a given structure-structure comparison score lies above the mean score for all comparisons. Given a representative set of proteins, for each member FSSP creates a list of all matches with a Z -score greater than 2. Matches below this score are considered to be dissimilar. A fold tree may be constructed using Z-score cutoffs of 2, 3, 4, 510, and 15 . However these cut-offs do not distinguish between folds and superfamilies accurately (Holm, 1996).

The relationships established among proteins by these databases provide us with a standard set of true positive and true negative fold relationships. As such, they form the basis for identifying and classifying related proteins in the emerging genomes. Since protein function is derived from its structure, in order to fully harness the information contained within these genomes scientists must be able to determine the structures of the newly found proteins and study their relationships to other proteins of known structure
and function. The availability of so many genomes has made this a compelling problem that has resulted in the emergence of structural genomics efforts.

The problem is often approached by seeking sequences that are similar to the sequence of a protein with known structure. This strategy works well for closely related sequences, but structural similarities can go undetected as the level of sequence similarity falls below 25 percent, a level referred to as the "twilight zone" (Doolittle, 1986). However, there are many proteins with similar structures where no obvious sequence homology can be detected (Jaroszewski, 1998). As a result, molecular modeling of proteins is confronted with the problem of finding homologous proteins. Methods developed to identify such structural relationships in the absence of sequence similarity are referred to as fold recognition or threading methods.

The current approaches to fold recognition differ in at least one of the following components: the representation of the protein, the function used to evaluate fold/target compatibility, the alignment algorithm, the ranking scheme, and methods to evaluate significance. Secondary elements such as alpha helices, and beta sheets, can now be recognized with an accuracy of more than $70 \%$ (Rost, 1993). However, their relative spatial relationship to each other as well the conformations of connecting loops cannot yet be deciphered from the amino acid sequence. Many attempts have been have been made to combine sequence information with predicted secondary structure (Geourjon, 2001), evolutionary information (Fischer, 2000), and 3D-1D profiles (Bowie, 1991), to enhance the underlying sequence similarity, thus pushing the threshold of the twilight zone. A survey of several fold prediction methods reveals that none has a success rate of


$>29 \%$ for recognizing proteins of the same superfamily, and that rate falls to $15 \%$ for proteins of the same fold (Elofsson, 2000).

A few studies have shown that the employment of experimental data in sequencebased structure prediction methods can improve their performance. For example, Jin et. al have used constraints derived from epitope mapping data to screen large numbers of computer-generated structural models (Jin et al., 1994). However, the generation of functional epitopes is limited to small proteins of less than 30 KDa . Dandekar \& Argos include a term for adherence to experimental data, derived from either conserved hydrophobic and catalytic residues, the distribution of cyteinyl S-S bond, or cross links amongst side chains in the fitness function for their genetic algorithm (Dandekar and Argos, 1997). This method too, is only applicable to proteins with less than 100 residues and fewer than eight secondary structures due to computational expense. In addition, both approaches are handicapped by their reliance on sequence as discussed earlier.

Recent advances in cross-linking chemistry and mass spectrometry have made it possible to produce distance constraints at a relatively fast rate for proteins of relatively large size and structural complexity (Young, 2000). With these advances in mind, our aim was to develop a new approach to fold recognition, independent of sequence, and neither limited by protein size nor complexity by utilizing distance constraints as the principal component; hence circumventing the limitations imposed on previous methods. This chapter is preliminary study of the issues involved in such a project.

## METHODS

DeVELOPing a fold recognition Algorithm (DP) using distance constraints

## Definitions:

1. Two proteins are said to have a common fold if their three-dimensional structures share the same overall geometry as defined by CATH.
2. A Novel fold is one that does not have a representative in a given library.

Given a partial set of distance constraints for a protein structure, P (probe), and a set of T (target) structures each with a corresponding set of restraints, we aim to establish whether in the absence of the full set of restraints, i.e. the complete intra-residue distance matrix, it is possible to determine which T has a common fold with P , if any. And if so, what is the limit on the amount of information required for providing satisfactory recognition. The metric of similarity is the pair-wise intra-residue distance, defined by $\mathrm{C}_{\alpha}-\mathrm{C}_{\alpha}$ separations in space. Comparison of these distances will yield insight on the degree of similarity between two structures. In doing so, we reduce P and T from three dimensional shapes to one dimensional strings of intra-molecular pair-wise distances whose comparison is less complex and well studied. A two-step alignment protocol

```
# Seed Alignment
    set n
    For probe, P
        For i=1 to length of protein
            Forj=1 to length of protein
                        j=j+n
                            Create distance matrix, DD*', by calculating Co\alpha-C\mp@subsup{\alpha}{j}{}}\mathrm{ distance
            i=i+n
    For Target,T
            k=1
            while k<length difference P and T
                            For i=k to length of protein
                            For j=k to length of protein
                            j=j+n
                                Create distance matrix, DT,k by calculating C }\mp@subsup{\alpha}{i}{}-C\mp@subsup{\alpha}{j}{
distance
            Compare each matrix, D}\mp@subsup{D}{T,K}{}\mathrm{ with }\mp@subsup{D}{P}{}\mathrm{ , keep one with least difference, call }\mp@subsup{D}{T}{
# Align Distance Matrices
    Collapse D D into one-dimensional arrays, }\mp@subsup{s}{P}{
    For each DT into one-dimensional arrays, }\mp@subsup{s}{T}{
            Create SP,T using formula I and dynamic programming
            Obtain an equivalence score
    Two proteins with highest equivalence score are the most similar
```


(outlined in figure 1) is used to obtain a score for each P and T pair. The scores are then ranked to highlight the best target, T .

As an initial test of our premise, we used the coordinates of structures deposited in the PDB to generate $\mathrm{C}_{\alpha}-\mathrm{C}_{\alpha}$ intra-residue distances for the members of our test sets. However, to parallel the data made available by real experiments we will only use a subset of the full set of restraints. The subset will be generated from residues with sequence separation $n$. It has been shown that for a sequence of length $\mathrm{L}, \mathrm{L} / 10$ distance constraints are sufficient for improving the performance of threading algorithms (Young, 2000). Initial studies are carried out using $n=5,10$, and 20 and probes proteins are always smaller than target proteins. Our aim was to derive a lower bound for the amount of information required for accurate fold recognition.

We generate an intra-residue distance matrix, $\boldsymbol{D}$, for every structure in the library, fig 2. Each distance matrix, $\boldsymbol{D}$, is then collapsed into a one dimensional array, s. A similarity matrix $\boldsymbol{S}_{\boldsymbol{T}, \boldsymbol{P}}$ is created using arrays from a target and probe protein with the application of formula 1 below (Gerstein, 1998). $\mathrm{D}_{\mathrm{ij}}$ is the difference between the $\mathrm{C}_{\alpha}$ $\mathrm{C}_{\alpha}$ distances in the two proteins. M represents the maximum possible score for a pair of distances and is set arbitrarily set to $20 . \mathrm{d}_{0}$ is the distance at which similarity fall to half it value, $\mathrm{S}_{\mathrm{ij}}=\mathrm{M} / 2$ and is set to 2.24 reflect the intrinsic length scale of protein structural similarity. This is about midway between the length of a C-C bond ( $1.54 \AA$ ) and the usual distance between Ca atoms ( $3.8 \AA$ ).

$$
\begin{align*}
& S_{i j}=\operatorname{Max}\left\{\begin{array}{l}
S_{i-1, j-1}+S i j \\
\max \left(S_{(-k, j-1}-W_{k}+S_{j}\right) \\
\max \left(S_{(-1, j-k}-W_{k}+S_{j}\right)
\end{array}\right\} \\
& S_{i j}=\frac{M}{1+\left(\frac{D i j}{d 0}\right)^{2}}  \tag{1}\\
& W_{k}=O+(i-k) E
\end{align*}
$$

FIGURE 1: Similarity matrix $S$ for two proteins is created from the intra-residue distance matrix of a probe and a target protein

## $\mathrm{C} \alpha-\mathrm{C} \alpha$ intra-residue matrix for every protein

## Similarity matrix

 Probe|  | 1 | 11 | 21 | 31 | 41 | 51 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 1 |  |  |  |  |  |  |


|  |  | D1 | D2 | D3 | D4 | $\ldots$ | D14 | D15 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| T | D1 |  |  |  |  |  |  |  |
| a | D2 |  |  |  |  |  |  |  |
| r | D3 |  |  |  |  |  |  |  |
| $g$ | D4 |  |  |  |  |  |  |  |
| e | - |  |  |  |  |  |  |  |
| $t$ | D14 |  |  |  |  |  |  |  |
|  | D15 |  |  |  |  |  |  | $\mathrm{S}_{11}$ |

FIGURE 2:Behavior of the similarity score for various values of M and $\mathrm{d}_{0}$



In order to optimize the values for M and $\mathrm{d}_{0}$ we graph the behavior of Sij vs. $\mathrm{D}_{\mathrm{ij}}$ for various M and $\mathrm{d}_{0}$ values. While M simply shifts the scoring scale, $\mathrm{d}_{0}$ dictates how fast the score will drop for various $\mathrm{D}_{\mathrm{ij}}$ values. Ideally, we would like a score function with heavy gradation in the range of protein similarity, $0-5 \AA$, in order to best differentiate among various degrees of similarity. We observe that $\mathrm{d}_{0}=2$ achieves this purpose, since it falls quickly for $\mathrm{D}_{\mathrm{ij}}$ in the range of 0-5. All other values plateau much later, resulting in only a small difference in similarity scores for related and unrelated proteins, fig 3.

We apply dynamic programming (DP) to the similarity matrix $\boldsymbol{S}_{\boldsymbol{T}, \boldsymbol{P}}$ to align the probe, P , and target, T (see text box algorithmic pseudo code). The alignment procedure is a dynamic programming algorithm inspired by the Needleman-Wunsch alignment protocol used in for sequences. The gap penalty used in the alignment is an affine gap penalty and we used values commonly used in sequence alignments, i.e. gap opening of 12 and gap extension of -2 . Two proteins with the highest equivalence score should show similar three-dimensional structures.

A percent similarity score, $\mathrm{S}_{\mathrm{p}}$, the percentage of the score achieved by aligning P against itself is calculated for each ( $\mathrm{P}, \mathrm{T}$ ) pair and a rank ordered list of $\mathrm{S}_{\mathrm{p}}$ is generated

to create a library of 110 domains, non sharing more than $20 \%$ sequence identity. The list of domains can be found in table 1. In addition, we use the 68 benchmark set used by

Fischer et. al. (Fischer, 1996) in order to test our algorithm against an independent test set, see table 2.

Furthermore, we used the program DSSP (Definition of Secondary Structure of Proteins) (Kabsch and Sander, 1983) to compile a library of SSE's from the structures in the 68 benchmark test set.

TABLE 1: Library of similar folds as designated by CATH

| PDB Code | Length | CATH <br> Classification | PDB Code | Length | CATH <br> Classification | PDB Code | Length | CATH <br> Classification |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $1 \mathrm{f3z00.pdb}$ | 150 | 2.70 .70 | $1 \mathrm{bgf00.pdb}$ | 124 | 1.10 .532 | 1 prcC2.pdb | 171 | 1.10.468 |
| $1 \mathrm{gpr00.pdb}$ | 158 | 2.70 .70 | $2 \mathrm{phy00}$ pdb | 125 | 3.30 .450 | 1waj04.pdb | 175 | 1.20.185 |
| $2 \mathrm{gpr00.pdb}$ | 154 | 2.70 .70 | $1 \mathrm{af500.pdb}$ | 126 | 3.10 .28 | $1 \mathrm{ami03.pdb}$ | 175 | 3.30.499 |
| $1 \mathrm{mup} 00 . \mathrm{pdb}$ | 157 | 2.40.128 | $1 \mathrm{rie} 00 . \mathrm{pdb}$ | 127 | 2.102.10 | 1wer01.pdb | 176 | 1.10.506 |
| 2a2uA0.pdb | 158 | 2.40.128 | $1 \mathrm{ddf00.pdb}$ | 127 | 1.10.533 | 1 fvkA0.pdb | 176 | 3.40.300 |
| 1ew3A0.pdb | 159 | 2.40.128 | 1agrE1.pdb | 127 | 1.10.196 | $1 \mathrm{nfa00.pdb}$ | 178 | 2.60 .71 |
| $10 \mathrm{bpAO} . \mathrm{pdb}$ | 158 | 2.40 .128 | $1 \mathrm{msc} 00 . \mathrm{pdb}$ | 129 | 3.30 .380 | $1 \mathrm{ckmA1} . \mathrm{pdb}$ | 179 | 3.90 .63 |
| $1 \mathrm{rbp00.pdb}$ | 174 | 2.40 .128 | 1 ordA4.pdb | 130 | 3.90.100 | 1bp101.pdb | 180 | 3.15 .10 |
| $1 \mathrm{np} 400 . \mathrm{pdb}$ | 184 | 2.40 .128 | $1 \mathrm{lis} 00 . \mathrm{pdb}$ | 131 | 1.20.150 | $2 \mathrm{vhbB0} 0 \mathrm{pdb}$ | 137 | 1.10.490 |
| 1 bbpA0.pdb | 173 | 2.40 .128 | $2 \mathrm{tct02} . \mathrm{pdb}$ | 133 | 1.10.357 | $1 \mathrm{hbrA0} 0 \mathrm{pdb}$ | 141 | 1.10.490 |
| 1euoA0.pdb | 157 | 2.40.128 | 1 furA1.pdb | 134 | 1.10.275 |  |  |  |
| 1 j jon00.pdb | 140 | 3.50.7 | $1 \mathrm{cfe00.pdb}$ | 135 | 3.40 .33 |  |  |  |
| 1sival.pdb | 145 | 3.50 .7 | 1a2601.pdb | 136 | 1.20.142 |  |  |  |
| $1 \mathrm{ass} 00 . \mathrm{pdb}$ | 152 | 3.50.7 | 2nef00.pdb | 136 | 3.30 .62 |  |  |  |
| 4bp200.pdb | 116 | 1.20.90 | 2end00.pdb | 137 | 1.10.440 |  |  |  |
| 3 p 2 pAO . pdb | 119 | 1.20.90 | $1 \mathrm{ckmA3} . \mathrm{pdb}$ | 137 | 4.10 .87 |  |  |  |
| 1 1poa00.pdb | 118 | 1.20 .90 | 1 preH2.pdb | 139 | 3.90 .50 |  |  |  |
| 1ae700.pdb | 119 | 1.20 .90 | $1 \mathrm{pjr} 04 . \mathrm{pdb}$ | 140 | 1.10.486 |  |  |  |
| 1 vip00.pdb | 121 | 1.20 .90 | 1bucA3.pdb | 141 | 1.20.140 |  |  |  |
| 1bk900.pdb | 124 | 1.20 .90 | $1 \mathrm{rgs} 02 . \mathrm{pdb}$ | 142 | 3.50 .12 |  |  |  |
| 1ppa00.pdb | 121 | 1.20 .90 | 2occD0.pdb | 144 | 1.10.442 |  |  |  |
| $1 \mathrm{poc} 00 . \mathrm{pdb}$ | 134 | 1.20 .90 | 1at000.pdb | 145 | 2.170 .117 |  |  |  |
| $1 \mathrm{bfg} 00 . \mathrm{pdb}$ | 126 | 2.80 .10 | $2 \mathrm{hhmA1.pdb}$ | 146 | 3.30 .540 |  |  |  |
| $1 \mathrm{afcA0} . \mathrm{pdb}$ | 127 | 2.80 .10 | 1aohB0.pdb | 147 | 2.70 .45 |  |  |  |
| $1 q q / A 0 . p d b$ | 131 | 2.80 .10 | 1def00.pdb | 147 | 3.90.45 |  |  |  |
| $1 \mathrm{j} \times \mathrm{xA} 1 . \mathrm{pdb}$ | 159 | 2.80 .10 | $1 \mathrm{tf4A2} . \mathrm{pdb}$ | 147 | 2.60 .43 |  |  |  |
| $111 \mathrm{~b} 00 . \mathrm{pdb}$ | 151 | 2.80 .10 | $1 \mathrm{uxy03.pdb}$ | 150 | 3.30.465 |  |  |  |
| 1ilr $10 . \mathrm{pdb}$ | 145 | 2.80 .10 | 1 sra00.pdb | 151 | 1.10.467 |  |  |  |
| $2 \mathrm{wbc} 00 . \mathrm{pdb}$ | 183 | 2.80 .10 | $1 \mathrm{lul000.pdb}$ | 152 | 2.60 .35 |  |  |  |
| $1 \mathrm{wba00} . \mathrm{pdb}$ | 171 | 2.80 .10 | 190f01.pdb | 153 | 2.60 .50 |  |  |  |
| $1 \mathrm{chmA1.pdb}$ | 155 | 3.40.350 | 1 pruAO.pdb | 154 | 3.40.210 |  |  |  |
| 1a1601.pdb | 171 | 3.40 .350 | $1 \mathrm{lacp00.pdb}$ | 154 | 1.20 .70 |  |  |  |
| $2 \mathrm{aak00.pdb}$ | 150 | 3.10 .110 | $1 \mathrm{pprM1}$. pdb | 155 | 1.40 .10 |  |  |  |
| $1 \mathrm{layzA0.pdb}$ | 153 | 3.10 .110 | $1 \mathrm{rhs01.pdb}$ | 156 | 3.40 .250 |  |  |  |
| 149 aAO .pdb | 159 | 3.10 .110 | $1 \mathrm{hfc} 00 . \mathrm{pdb}$ | 157 | 3.40 .390 |  |  |  |
| $2 \mathrm{ucz00.pdb}$ | 164 | 3.10 .110 | $1 \mathrm{msk02.pdb}$ | 157 | 1.10.288 |  |  |  |
| $1 \mathrm{rdl10.pdb}$ | 111 | 3.10 .100 | 1him00.pdb | 158 | 1.10.490 |  |  |  |
| 1hup00.pdb | 141 | 3.10 .100 | $1 \mathrm{ra900.pdb}$ | 159 | 3.40 .430 |  |  |  |
| $2 \mathrm{msbAO} . \mathrm{pdb}$ | 111 | 3.10 .100 | 3daaA2.pdb | 159 | 3.20 .10 |  |  |  |
| $1 \mathrm{ttm} 10 . \mathrm{pdb}$ | 149 | 3.10 .100 | $1 \mathrm{npc} 02 . \mathrm{pdb}$ | 161 | 1.10.390 |  |  |  |
| $1 \mathrm{~b} 08 \mathrm{AO}, \mathrm{pdb}$ | 152 | 3.10 .100 | $3 \mathrm{pmgA4}$.pdb | 161 | 3,30.530 |  |  |  |
| $1 \mathrm{t} 300 . \mathrm{pdb}$ | 137 | 3.10 .100 | 1apyAO.pdb | 161 | 3.30.426 |  |  |  |
| ${ }_{\text {1 }}^{\text {1 }}$ litoo.pdb | 128 | 3.10 .100 | 1dhx01.pdb | 162 | 2.170 .9 |  |  |  |
| 1bytA0.pdb 2afpoo.pdb | 123 | 3.10 .100 | 1 fuiA2.pdb | 165 | 3.40.275 |  |  |  |
| 1bvp11.pdb | 128 | 3.10 .100 | $1 \mathrm{inp03.pdb}$ | 166 | 3.40.191 |  |  |  |
| 1 mdaL .pdb | 120 | 1.10.250 | $1 \mathrm{clh} 00 . \mathrm{pdb}$ | 166 | 2.40 .100 |  |  |  |
| 1 1dekA2.pdb | 122 | 2.60.30 1.10 .345 | 1936A4.pdb | 168 169 | 1.10 .132 3.90 .76 |  |  |  |
| 1ahuA1.pdb | 123 | 3.30.373 | $1 \mathrm{aihA0.pdb}$ | 170 | 1.10.443 |  |  |  |
| 1bucA1.pdb | 123 | 1.10.540 | 1 anv02.pdb | 170 | 3.90.148 |  |  |  |




TABLE 2: 68 protein pairs from Fischer et. al



里 －
－

$$
0
$$

－

$\checkmark$


$\square$
$\square$
$\square$
$\square$
$\square$ ． $\square$

$\square$
$\square$
$\square$
$\square$
－

ND $\triangle$
ND $\triangle$
ND $\triangle$
ND $\triangle$
ND $\triangle$
ND $\triangle$
ND $\triangle$
ND $\triangle$
ND $\triangle$


$\square$
$\square$
$\square$
$\square$ O $\square$
$\square$
$\square$ －
$\square$


## RESULTS AND DISCUSSION

We begin by performing all against all comparisons of both non-related and related folds in both our own library and the 68 benchmark. The results of our comparisons show that you can indeed separate related and unrelated proteins using distance constraints alone. Figures 4 and 5 show the distribution of similarity scores for the positive matches (same fold family) in green and the non-related pairs in red. There is no clear cutoff between the positive and negative matches; instead we observe an overlapping transition. This overlap is often referred to as either a lack of sensitivity (coverage) or selectivity (error per query (EPQ)) in fold recognition. Sensitivity refers to the number of true matches that can go unrecognized; in this case those structures that rank lower than an unrelated protein while selectivity is related to the number of false positives, in our case unrelated proteins with high similarity scores. As a result, in the absence of a priori knowledge about the proteins in question it is difficult to determine a definitive similarity cutoff score. The results using the 68 benchmark show the same trend. While close to sixty percent of the protein pairs are retrieved with a rank of ten or better (table 2) the score distribution of the unrelated and related domains overlap. In addition there are several cases in which the 'best' match ranks almost at the bottom of the list.



FIGURE 3: Distribution of DP Alignment scores. The $x$-axis is the percent score of each probe-target alignment obtained by dividing the raw alignment score by the alignment score of the probe to itself. Data points depicted by ' + ' show the alignment of non-homologous pairs, as defined by CATH, and those with a '*' are homologous pairs. There are a total of 2609 non-homologous pairs and 137 homologous pairs, excluding the self-matches
 droll



FIGURE 4: Distribution of alignment scores. A test set of 110 folds was used to generate a set of 2746 probe-target comparisons, excluding self-self pairs. Pairs are said to be incorrect if they are not in the same fold family as defined by CATH. The distribution of the correct match scores(solid line) shows that similar structures have an overall higher score than non-similar pairs (dotted line)


To assess the difficulty of our test set, we performed standard sequence alignments on the member of the set, to see whether we can identify fold relatives using

the more common scoring functions such as Blossum62 and (Henikoff and Henikoff, 1992) PAM250 (Dayhoff, 1978). A coverage vs. EPQ plot (figure 6) shows that the Blossum62 matrix performs better than our algorithm, by approximately twenty percent, while the PAM250 scoring matrix is worse by thirty percent.

FIGURE 5: Coverage versus error plot for DP using 110 member fold library. The x -axis, coverage, shows the fraction of all protein pairs with the same fold in the library with an alignment score above the selected threshold (indicated by the numbers on the curve). The $y$-axis shows the number of nonhomologous pairs above the threshold as a fraction of all protein-protein pairs. Coverage is a measure of the sensitivity of the method and EPQ a measure of selectivity. As we find more homologous pairs (increase in sensitivity, the number of pairs found in error increases as well, decrease in selectivity


We also ran the DP algorithm with distance constraints on the secondary structure element (SSE) test set. The set is composed of small helices and beta strands of the 68benchmark test set. As seen in table 3, there is a clear separation of scores among the alpha helices and the beta strands meaning that distance constraints alone can distinguish among these two different geometries.


TABLE 3: Comparison of a beta strand from 2 AK 3 A to a set of secondary structures derived from the 68 benchmark. The scores show a clear gap among the scores of similar beta strands and beta strand - alpha helix comparisons. Highlighted rows are alpha helix elements, while white rows are beta strand elements

| Raw | Similarity |  |
| :--- | ---: | ---: |
| Probe Name | Score | Score |
| 2ak3A_0101 | 60 | 1 |
| 2fox_0104_0 | 59.95575 | 0.9992626 |
| 1aaj_0103_1 | 59.95522 | 0.9992537 |
| 1aaj_0203_0 | 59.88732 | 0.998122 |
| 1gky_0104_0 | 59.86916 | 0.9978193 |
| 1gky_0303_2 | 59.81012 | 0.9968353 |
| 1paz_0202_1 | 59.7737 | 0.9962284 |
| 1paz_0203_0 | 59.76299 | 0.9960499 |
| 1aaj_0101_0 | 59.72995 | 0.9954992 |
| 1aaj_0102_1 | 59.69001 | 0.9948334 |
| 1paz_0101_0 | 59.6709 | 0.994515 |
| 1gky_0304_0 | 59.60995 | 0.9934991 |
| 1aaj_0202_1 | 59.59306 | 0.9932176 |
| 2fox_0101_0 | 59.40849 | 0.9901415 |
| 1paz_0103_1 | 59.33673 | 0.9889455 |
| 2fox_0103_0 | 59.25019 | 0.9875032 |
| 1gky_0103_0 | 59.24475 | 0.9874125 |
| 1paz_0102_1 | 55.46105 | 0.9243508 |
| 1gky_0500_2 | 43.94399 | 0.7323998 |
| 2fox_0400_0 | 43.7066 | 0.7284433 |
| 1gky_0800_1 | 43.69803 | 0.7283004 |
| 2fox_0500_5 | 43.60831 | 0.7268052 |
| 1gky_0600_2 | 43.60489 | 0.7267481 |
| 2fox_0600_0 | 43.53544 | 0.7255906 |
| 1gky_0200_3 | 43.48087 | 0.7246812 |
| 1gky_0900_7 | 43.47443 | 0.7245739 |
| 2fox_0200_8 | 43.4374 | 0.7239566 |
| 1paz_0400_1 | 43.41235 | 0.7235392 |
| 1gky_0400_4 | 43.22084 | 0.7203474 |
| 1gky_0700_2 | 43.16956 | 0.7194926 |
| 1paz_0300_1 | 43.06243 | 0.7177072 |

We also look specifically at an example where DP fails to point out the correct match (fig. 6). The probe is an all $\beta$ protein (PDB: 1aaj) and its designated pair is another all $\beta$ protein (PDB: 1paz). The algorithm can not recognize 1 paz as the best scoring match for 1 aaj, and ranks it as $78^{\text {th }}$ best match. The closest match using intra-residue
distances is determined to be the $\alpha / \beta / \alpha$ sandwich 2 fox. The three major protein classification schemes, SCOP, CATH, and FSSP place these two proteins in unrelated

fold families and the Blossum62 matrix successfully find 1paz to be related to 1aaj. So the question remains as to why a geometric metric such as intra-residue distances fails to recognize this similarity.

FIGURE 6: Topology diagrams for a pair of related proteins laaj, lpaz and an unrelated pair laaj,2fox.Although both laaj and lpaz are classified as an all $\beta$ sandwich the topology of the $\beta$ strands in laaj is much more similar to the $\beta$ strands in 2 fox, an $\alpha / \beta / \alpha$ sandwich

| PDB ID |  | FOLD |  |  | RANK |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | SCOP | CATH | FSSP | DP | Blossum <br> $\mathbf{6 2}$ | $\mathbf{2 5 0}$ |
| 1aaj,1paz | All $\beta$, sandwich | 2.60 .40 .120 <br> $\beta$ sandwich | z-score <br> 9.7 | 28 | 1 | 78 |
| 2fox | $\alpha / \beta$ <br> 3 layer $\alpha / \beta / \alpha$ | 3.40 .50 .3603 <br> layers, $\alpha / \beta / \alpha$ <br> (Rossmann) | - | 1 | 93 | 70 |



When looking at the topology of these proteins (fig 6), it becomes clear that although 1aaj and 1 paz are both composed of $\beta$ strands their topologies are different. The $\beta$ strands of 1aaj are connected in more of an even-odd order where as the order of the strands is continuous in 1 paz. Interestingly, the $\beta$ strands of 2 fox show the same alternating connection where the loops go from strand to helix to strand.

All three of the previously mentioned classification schemes, namely SCOP, CATH, and FSSP rely on the secondary structure make up of the protein as their main metric of comparison. It is also worth noting that the Blossum62 matrix too is derived by aligning similar sequence blocks thus being inherently biased towards secondary structure similarity rather than topological similarity. While secondary structure similarity is a certain necessity for identifying related folds, the current challenge in structure prediction lies in identifying the topological arrangement of SSE's. The use of geometric constraints in the form of experimentally derived bounds for intra-residue distances can aide in determining these arrangements. However, development of computational methods that can properly incorporate this data is not possible in the absence of classification schemes that allow for spatial considerations in addition to secondary structure composition. As realized, it is currently not feasible to asses the success or failure of our method objectively due to the classification bias in our test sets. Similarity must not be viewed as a discrete state but rather a continuum in which the transition from similar to dissimilar is gradual. With the advent of large scale genomics and our current understanding of the issues involved in assessing structural similarities, a re-evaluation of current classification methodologies and the development of more sophisticated criteria seem not just appropriate but also timely. Such an effort however is outside of the scope of this dissertation.

In light of the complications involved in defining protein fold similarity we decided to shift to a more fundamental approach and measure the information content of protein structures. In the following chapters we will utilize reduced protein models to continue our investigation of distance constraints by making a quantitative assessment of


their information content using information theory as our theoretical backbone. Lattice models have previously been used to study protein folding (Chan, 1989). They are advantageous in that for a relatively small number of beads all conformations can be enumerated. The exhaustive enumeration of all or a subset of conformations, such as compact structures, allows the development of precise measures of information content and helps us gain insight into effective definitions of conformational "similarity." In using conformational constraints (i.e. $\mathrm{C}_{\alpha}-\mathrm{C}_{\alpha}$ distances for fold recognition) an important question arises concerning the limit of information contained within those constraints and whether a limited number is sufficient for elucidating the fold of a protein. This information will further help in assessing the utility of a restraint-based approach, since distance constraints for new sequences must be determined experimentally; for a brief outline of such the protocol refer to figures 7 (Young, 2000).

FIGURE 7: . Distance constraints derived from cross linking experiments. Sample experiment from top to bottom, protein is subjected to cross linking agent. Monomers are separated from dimers. Proteolytic digestion, followed by separation of cross linked from non cross linked peptides. The cross linked peptides are analyzed by Mass Spec to determine which residues are x -linked, yielding a set of distance constraints (Taken from Young, 2000)




## ACKNOWLEDGMENTS

We are grateful to Judith Hemple, and Connie Oshiro for helpful discussions.


$\mathcal{N}$

## REFERENCES

Berman, H.M., Westbrook,J., Feng,Z.,Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., Bourne, P.E. 2000. The Protein Data Bank. Nucleic Acids Research 28:235-242.

Bowie, J., Luthy, R., Eisenberg, D. 1991. A method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. Science 253:164-170.

Burley, S.K., Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland,T., Lin, D., Sal,i A., Studier, F.W., Swaminathan, S. 1999. Structural genomics: beyond the human genome project. Nature: Genetics 23(2):151-157.

Chan, H.S., Dill, K.A. 1989. Compact Polymers. Macromolecules 22:4559-4573.
Chothia, C., Levitt, M., Richardson, D. 1977. Structure of proteins: Packing of $\alpha$-helices and pleatred sheets. Proceedings of the National Academy of Sciences of the United States of America 74(10):4130-4134.

Dandekar, T., and P. Argos. 1997. Appyling experimental data to protein fold prediction with the genetic algorithm. Protein Engineering 10(8):877-893.

Dayhoff, M.O. 1978. A model of evolutionary change in proteins. Atlas of protein sequence and structure. Silver Spring, Md.,; National Biomedical Research Foundation. p 345-352.

Doolittle, R.F. 1986. of Urfs and Orfs: A primer on How to Analyze Derived Amino Acid Sequences. University Science Books, Mill Valey, CA

Elofsson, A., Lindahl, E. 2000. Identification of Related Proteins on Family, Superfamily, and Fold Level. Journal of Molecular Biology 295:613-625.

Fischer, D. Hybrid Fold Recognition: Combining Sequence Derived Properties with Evolutionary Information; 2000. p 119-130.

Fischer, D., Elofsson, A., Eisenberg, D. Assessing the Perfomance of Fold Recognition Methods by Means of a Comprehensive Benchmark; 1996. p 300-318.

Geourjon, C., Combet, C, Blanchet, C., Deleage, G. 2001. Identification of related proteins with weak sequence identity using secondary structure information. Protein Science 10:788-797.

Gerstein, M., Levitt, M. 1998. Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. Protein Science 7(2):445-56.



Hadley, C., Jones, T.,. 1999. A systematic comparison of protein structure classifiactions: SCOP, CATH and FSSP. Structure 7:1099-1112.

Henikoff, S., and J.G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89(22):10915-9.

Holm, L., Sander, C. 1996. The FSSP database: Fold classification based on structurestructure alignment of proteins. Nucleic Acids Research 24(1):206-209.

Jaroszewski, L., Rychlewski, L.,Zhang, B.,Godzik, A. 1998. Fold prediction by a hierarchy of sequence, threading, and modeling methods. Protein Science 7(6):1431-40.

Jin, L., F.E. Cohen, and J.A. Wells. 1994. Structure from function: screening structural models with functional data. Proc Natl Acad Sci U S A 91(1):113-7.

Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577-637.

Kuntz, I.D., Crippen, G., Kollman, P.A. 1979. Application of Distance Geometry to Protein Tertiary Structure Calculations. Biopolymers 18:939-957.

Murzin, A. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 247(4):536-540.

Orengo CA, M.A., Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH--a hierarchic classification of protein domain structures. Structure 5(8):1093-1108.

Rost, B., Sander C. 1993. Prediction of protein secondary structure at better than 70\% accuracy. Journal of Molecular Biology 232(2):584-599.

Young, M., Tang, N., Hempel, C., Oshiro, C., Taylor, E.W., Kuntz, I.D., Gibson, B.W., Dollinger, G. 2000. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America 97(11):5802-5806.



# Chapter Three: Information Content of Molecular Structures 

David C. Sullivan ${ }^{*}$, Tiba Aynechi ${ }^{\ddagger}$, Vincent A. Voelz ${ }^{\ddagger}$, Irwin D. Kuntz ${ }^{*}$<br>* Department of Pharmaceutical Chemistry<br>${ }^{*}$ Graduate Group in Biophysics<br>University of California, San Francisco, California, 94143-2240<br>${ }^{\text {8}}$ Present address: Institute of Biomedical Sciences, Academia Sinica<br>Taipe 11529, Taiwan<br>Correspondence should be addressed to Dr. Irwin D. Kuntz, Department of<br>Pharmaceutical Chemistry, University of California, San Francisco, California, 94143-<br>2240<br>Phone: (415) 476-1937<br>Fax: (415) 502-1411<br>Email: kuntz@cgl.ucsf.edu<br>Published in Biophysical Journal, Vol. 85, July 2003 pp. 174-190<br>\section*{KEYWORDS}<br>Shannon Entropy, Conformation Space, Lattice Models, Distance Constraints, Noise,<br>Structure Predictin





#### Abstract

For a completely enumerated set of conformers of a macromolecule or for exhaustive lattice walks of model polymers it is straightforward to use Shannon information theory to deduce the information content of the ensemble. It is also practicable to develop numerical measures of the information content of sets of exact distance constraints applied to specific conformational ensembles. We examine the effects of experimental uncertainties by considering "noisy" constraints. The introduction of noise requires additional assumptions about noise distribution and conformational clustering protocols that make the problem of measuring information content more complex. We make use of a standard concept in communication theory, the "noise sphere", to link uncertainty in measurements to information loss. Most of our numerical results are derived from two-dimensional lattice ensembles. Expressing results in terms of information/degree of freedom removes almost all of the chain length dependence. We also explore off-lattice polyalanine chains that yield surprisingly similar results.




## INTRODUCTION

An important challenge for structural biology is to provide structural and functional information on the same grand scale as the genome sequencing projects. While there are many experimental procedures aimed at the determination of the structures of proteins and nucleic acids, relatively little attention has been paid to measuring the quality of any given method, and a framework for discussing the optimum utility of diverse procedures is lacking. (See, however, (Brunger et al., 1993) for error analysis in crystallography). Further, many experimental efforts combine direct structural data with sequence alignments or molecular refinement techniques, adding to the difficulty of analysis. In this paper, we introduce a protocol to quantify the information content of structural data and we explore some of the many issues that arise in reducing such a protocol to practice.

The process of determining the structure of a macromolecule is largely a matter of specifying the conformational states of highest occupancy for a given physical environment. While we speak of the "structure" of a molecule, we are normally referring to the equilibrium properties of an ensemble of molecules that constitute $a$ thermodynamic state. Individual molecules undergo dynamic transitions among conformations and only time-averaged properties of the ensemble can be measured directly. For biomacromolecules, except at the highest resolution, the lengths of the chemical bonds and the bond angles are taken to be constant. Conformations are essentially established through direct or indirect specification of the dihedral angles as the critical variables. In this paper, we explore how much information must be supplied to fix these angles within a certain tolerance or uncertainty. More precisely, we are interested in

the amount of information needed to discriminate among the different conformations accessible to a macromolecular system in a well-characterized thermodynamic state.

We will make use of information theory (Shannon, 1948; Young, 1971) to link the information content of a particular experiment or procedure (Havel et al., 1983; Sibbald, 1995) to the conformational entropy of a molecular ensemble.

There have been attempts to deduce the entropy of a molecular assembly from the variation of the atomic coordinates (Levy et al., 1984; Luo and Sharp, 2002; Potter and Gilson, 2002; Schlitter, 1993). While this approach works exactly for ideal gases, it is still unclear whether it yields a proper result for systems with conformational degrees of freedom (Schafer et al., 2001; Schafer et al., 2000). A large number of studies on chain entropy for polymer systems have been carried out (Dill et al., 1995; Flory, 1953; Pande et al., 1994; Wang et al., 1999) using a variety of models. Clearly, if it were possible to enumerate all (accessible) conformations and associated occupancies for a molecular ensemble, the total conformational entropy would easily be obtained.

However, an enumeration approach has two major difficulties for proteins or nucleic acids. First, the natural orthogonal variables are the dihedral angles. While such data are available from multidimensional NMR coupling experiments, a full set has not been reported. Instead, experiments typically yield a partial set of labeled (assigned) intramolecular distances and coupling constants from NMR or a set of unlabeled distances/phases from diffraction experiments. These data are strongly self-correlated so that constraints are generally non-orthogonal and the information gained is not a simple linear function of the number of constraints. Such correlations must be accounted for in any assessment of the information content of an experiment. The second problem is that
tive enumeration of the conformations of a macromolecule is not currently feasible ecause of the large numbers involved and because any working definition of a nolecular "conformation" is integrally connected to assumptions about energy s which introduce additional complications.

We envision two general approaches for measuring information content using ar conformational constraints. First, correlated constraints can be mapped to an nal space. For example, distances can be mapped to dihedral angles, although the hip can be significantly error-prone. The second approach, explored in this $s$ to use model systems where exhaustive enumeration of conformational es is feasible.
n previous work, Dill and co-workers and Wang et al., among others, used lattice $s$ to probe the statistical properties of ensembles of protein structures (Crippen, 11 et al., 1995; Dobson et al., 1998; Wang et al., 1999). Choy and Gregoret (Choy nan-Kay, 2001; Gregoret and Cohen, 1991) have also reported off-lattice models ed states. We will use the Dill ensembles to examine the information content of distance constraints and to explore the degradation of information as noise is


## RY AND METHODOLOGY

## ble Generation

## nensional lattice walks

n this initial study, we primarily use two-dimensional square lattice structures. of beads, each bead representing one "residue", are arranged in self-avoiding cording to the following rules. The elementary step, the distance between ive beads, $d_{i, i+1}$, is fixed at unit length. The move set is limited to a single step onal moves disallowed. Beads cannot overlap. This set of walks is the same as ustive ensembles of Chan and Dill (Chan and Dill, 1989) that count all tions not related by translation, rigid rotation, or reflection. These latter as are readily accomplished without loss of generality by limiting the first move ng the positive $y$ axis and by restricting the first turn to the positive $(x, y)$ The N -terminus to C -terminus directionality of proteins is preserved in these s. This directionality permits discrimination between "retro-inverso" tional pairs (Chorev and Goodman, 1995) - two conformations that become pon reflection and reversal of the bead numbering. sembles of unconstrained self-avoiding lattice walks and a separate subset of milton lattice walks were enumerated exhaustively up to $N=28$ ( $N$ is number the chain) (Table 1) and $N=49$ (Table 2) respectively. Enumerations of up to ave been published by (Chan and Dill, 1991; Irback and Troein, 2002) for ned walks. Square Hamilton walks of up to $N=36$ have been enumerated by Dill, 1989). Our values for $\boldsymbol{W}$, the number of distinguishable walks, agree in all cases.


Ve studied longer self-avoiding chain ensembles $(N=49,100)$ using stochastic n. During stochastic generation, conformations with the first turn outside of the $(x, y)$ quadrant were terminated and removed. For these ensembles, simple king from a point of chain overlap produces an over-representation of compact mpared to the exhaustive results (Rosenbluth and Rosenbluth, 1955). Instead, discard the run leading to failure and start a new walk from its beginning.

## Self-avoiding square-lattice walks

| $N^{*}$ | $W^{\#}$ | $I^{S \S}$ |
| :---: | :--- | :---: |
| 2 | 1 | 0.000 |
| 3 | 2 | 1.000 |
| 4 | 5 | 2.322 |
| 5 | 13 | 3.700 |
| 6 | 36 | 5.170 |
| 7 | 98 | 6.615 |
| 8 | 272 | 8.087 |
| 9 | 740 | 9.531 |
| 10 | 2034 | 10.990 |
| 11 | 5513 | 13.429 |
| 12 | 15037 | 15.310 |
| 13 | 40617 | 16.750 |
| 14 | 110188 | 18.179 |
| 15 | 296806 | 19.613 |
| 16 | 802075 | 21.040 |
| 17 | 2155667 | 22.470 |
| 18 | 5808335 | 23.893 |
| 19 | 15582342 | 25.320 |
| 20 | 41889578 | 26.742 |
| 21 | 112212146 | 28.166 |
| 22 | 301100754 | 29.585 |
| 23 | 805570061 | 31.007 |
| 24 | 2158326727 | 32.425 |
| 25 | 5768299665 | 33.846 |
| 26 | 15435169364 | 35.262 |
| 27 | 41214098278 | 36.681 |
| 28 | 110164686454 |  |



* The number of beads in a chain of length $\mathrm{N}-1$
\# The number of conformations (see text)
${ }^{8}$ Calculated as $\log _{2}(\mathrm{~W})$


## : Square Hamilton walks

| $N$ | $W$ | $I^{S}$ |
| :--- | :--- | :--- |
| 4 | 1 | 0.000 |
| 9 | 5 | 2.322 |
| 16 | 69 | 6.109 |
| 25 | 1081 | 10.078 |
| 36 | 57337 | 15.807 |
| 49 | 3383820 | 21.690 |

## ally-constrained 3-D Polyalanine Ensembles with Excluded Volume

he program YARN (Gregoret and Cohen, 1991) was used to generate random 3anine conformations that obey excluded volume constraints. In the default ombinations of $\phi$ and $\psi$ are chosen based on statistics from a reference set of described by Gregoret and Cohen (Gregoret and Cohen, 1990). An ellipsoid s the size of generated conformations to gyration radii consistent with entally derived structures (Gregoret and Cohen, 1991).

## and Information

iven a set of constraints, $\boldsymbol{X}$, the information content of the constraint set can be 1 in bits by its partitioning effect on the structural ensemble using Shannon's on (Shannon, 1948):

$$
\begin{equation*}
I(X)=-\Sigma\left[p_{k} \log _{2}\left(p_{k}\right)\right] \tag{1}
\end{equation*}
$$

is the population of cluster $\boldsymbol{k}$ expressed as a fraction of the ensemble, summed clusters. These clusters are subsets of the population of conformers that are ishable under a particular constraint. direct connection with classical statistical mechanics is available if it is possible $y$ the conformations which belong to a specific thermodynamic microstate and if 1 information is provided about the relative energy of each conformation (Wang

999). For this paper, we will assume that all lattice conformations have the same and hence the same occupancy. This assumption is equivalent to an "infinite ture" limit.

The measured information content of a particular constraint set, $\boldsymbol{X}$, can be ed to the theoretical information content of the ensemble defined as:

$$
\begin{equation*}
I^{S}=\log _{2}(W) \tag{2}
\end{equation*}
$$

$V$ is the ensemble size. $\boldsymbol{I}^{S}$ is referred to as the "source" information (Shannon, Other terms we will use are: $\boldsymbol{I}^{\boldsymbol{M}}$, defined as the maximum amount of information be recovered using a given set of measurements and $\boldsymbol{I}^{\boldsymbol{L}}$, the information lost at e of an experiment (see section IV for further discussion.)

## clature

Ve use the Cartesian (through-space) distance, $\boldsymbol{d}$, between beads $\boldsymbol{i}, \boldsymbol{j}$ as:

$$
\begin{equation*}
d_{i, j}=\left(\left(x_{i}-x_{j}\right)^{2}+\left(y_{i}-y_{j}\right)^{2}\right)^{1 / 2} \tag{3}
\end{equation*}
$$

$[d]_{i, j}$ will represent the $(i, j) t h$ element of the distance matrix, which can take on values, and $\boldsymbol{d}_{i, j}$ will represent the specific value of this element in a particular ation. The sequential separation, $s_{i, j}$, is defined as:

$$
\begin{equation*}
s_{i, j}=|i-j| \tag{4}
\end{equation*}
$$

he city-block sequence distance, $\boldsymbol{B}$, for two pairs of beads $(\boldsymbol{i}, \boldsymbol{j})$ and $\left(\boldsymbol{i}^{\prime}, \boldsymbol{j}^{\prime}\right)$, is

$$
\begin{equation*}
B=\left|i-i^{\prime}\right|+\left|j-j^{\prime}\right| \tag{5}
\end{equation*}
$$

here are several measures of determining the difference, $\delta(a, b)$, between a pair mations, $\boldsymbol{a}$ and $\boldsymbol{b}$. The most popular are the minimal root mean square difference of the coordinates after rigid translation and rotation and the closely related
ion of the difference of the distance matrices (Levitt, 1976). We will also make new measure of distance uncertainty based on examination of the distance-
ce matrix, $\Delta$ :

$$
\begin{equation*}
\Delta_{i j}(a, b)=\left|\left(d_{i j}\right)^{a}-\left(d_{i j}\right)^{b}\right| \tag{6}
\end{equation*}
$$

and $\boldsymbol{b}$ refer to specific conformations and $i, j$ are taken over all bead numbers, $j>$ fically, we focus on the maximum element in $\Delta$ defined as:

$$
\begin{equation*}
\mathrm{e}^{\mathrm{a}, \mathrm{~b}}=\max \left(\Delta_{\mathrm{i}, \mathrm{j}}(\mathrm{a}, \mathrm{~b})\right) \tag{7}
\end{equation*}
$$

inition is motivated by the simplicity of some results when formulated this way ults section.) We note that most of these measures are not proper metrics because not obey the triangle inequality.

For an $N$-mer lattice walk, the full set of constraints for any conformation is
as $\boldsymbol{\Xi}$ uch that:

$$
\begin{gathered}
\Xi=\left\{[d]_{i, j} \mid 1 \leq i \leq N, \quad 1 \leq j \leq N, i \neq j\right\} \text { and } \\
M \subset \Xi
\end{gathered}
$$

te the size of $\boldsymbol{M}$ as $|\mathbf{M}|$.
$\geq$ of the information content of a constraint
he information content of a distance element, $[d]_{i, j}$, for a given ensemble is 1 by partitioning the ensemble based on the distribution of distance values, $\boldsymbol{d}_{i j}$, $(i, j)$ in the ensemble. The fraction of the ensemble having a particular distance $[d]_{i, j}$, defines the value for $\boldsymbol{p}_{k}$. The indexing length for $\boldsymbol{k}$ is determined by the f accessible distance values for $[d]_{i, j}$.
r example, for a chain of length $N=3$, the information encoded in $[d]_{1,3}$, s determined as follows: The number of conformations in the ensemble, $\boldsymbol{W}$, is


he only allowed conformations are $\operatorname{straight}(s)$ and $\operatorname{bent}(b)$ which results in $\boldsymbol{k}=2$ $=p_{b}=0.5$. For one-half the conformations $\boldsymbol{d}_{1,3}=2$ and for the other half $\boldsymbol{d}_{1,3}=$
ng Shannon's equation:

$$
I\left([\mathrm{~d}]_{1,3}\right)=-2\left[0.5\left(\log _{2}(0.5)\right)\right]=1 \mathrm{bit}
$$

he information content of sets of distances is calculated in a similar manner. nembers share the same distance values across all distance elements of the set. is important to recognize that this protocol measures the amount of information d with knowledge of the full set of distance values for each distance element, an the (different) amount of information contained in knowing a specific value ticular distance element. Further, while this formulation is useful for any lattice would need to be altered for systems where internal distances vary in a as fashion. For example, in our studies of the polyalanine models, we will make aptions that each structure generated represents a different conformation and that ampling is done to provide reliable estimates of the distance distributions (see since we do not impose any force fields on the polyalanine ensembles, these are not related to discrete local minima on an energy landscape.

## noisy systems

ur discussion so far has assumed that the constraint set is noise-free and exact. this is not the general case. In order to study the effects of inexact nents and the addition of noise to the system, we will use a simplified cation model. It has the following components:
nation source: the set of noise-free messages that can be communicated - in this he set of fully enumerated conformations.
mission system: the set of constraints that select conformations for "broadcast". ources in transmission can give rise to "noisy" or inexact constraints.
tion system: reconstruction of the messages from the transmitted signal. The uction process may use filters (prior knowledge about the messages) or ing algorithms to recover the signal. Additional noise sources may be associated reception process.
ation loss from noise

We consider a conformational ensemble to be a set of $W$ independent, distinct s, $\left\{w_{i}\right\}$, of equal probability. The information content of the ensemble is defined W) (Shannon, 1948). As noise is introduced in the constraint sets some messages e distinct in a noise free environment become indistinguishable. A set of n probabilities, $p_{i}(j)$, the probability of message $i$ being received as message $j$, s this behavior. We denote the information of the source and the received signal $I^{M}$ respectively. In a noiseless case $I^{S}=I^{M}$ whereas in the noisy case $I^{M}<I^{S}$ (see


information loss due to noise, averaged across the ensemble, is:

$$
\begin{equation*}
\left\langle I^{L}\right\rangle=-\sum_{j=1}^{W} p(j) \sum_{i=1}^{W} p_{j}(i) \log _{2} p_{j}(i) \tag{8}
\end{equation*}
$$

j) is the probability of transmitting a particular symbol $w_{j}$.

To derive numerical results in the lattice model system, we assume that each ymbol is transmitted with equal probability $p(i)=p(j)=1 / W$. We use the "noisehere" model (Young, 1971) for the transmission loss, in which conformations $\mathrm{w}_{j}$ that within a hyper-sphere of radius $r$ centered about conformation $w_{i}$ are distinguishable. Let $u_{i}(r)$ be the number of conformations about $w_{i}$, inclusive, within a lius $r$ : The model of the transmission error probability, for a particular $r$, can thus be ressed as:

$$
p_{j}(i)= \begin{cases}0 & \text { if } \delta\left(w_{i}, w_{j}\right)>r  \tag{9}\\ 1 / u_{i} & \text { if } \delta(w i, w j) \leq r\end{cases}
$$

er this model, Eq. 8 simplifies to:

$$
\begin{equation*}
\left\langle I^{L}\right\rangle=\frac{1}{W} \sum_{i=1}^{W} \log _{2}\left[\left(u_{i}\right)^{-1}\right] \tag{10}
\end{equation*}
$$

an use the same approach to calculate the loss of information for noise in individual ce constraints. The noise sphere will contain all conformations $\left(u_{i}\right)$ whose $d_{i, j}$ is $r$ of the $d_{i, j}$ of the reference conformation, $i$. The calculation uses each conformer as the reference. $\boldsymbol{I}^{L}$ is obtained from Eq. 10.

## rmer distributions

To calculate how many conformers lie within a fixed interval, we will use the of Sullivan and Kuntz (Sullivan and Kuntz, 2001). We assume a conformational a which individual conformations are points and whose axes are the true cal degrees of freedom. We are interested in two situations. In the first case, we an ensemble that can, in principle, be generated exhaustively, although we may


to stochastic enumeration for long chains. In the second case, we assume that we tarry out exhaustive enumeration, but that we do have some prior knowledge the conformer distribution, e.g. that conformations are distributed uniformly in the rmational) space. In either case, we can develop a geometric model for the mation space as an appropriately dimensioned hyper-sphere and define the ated radial pair conformational density function, $v(r)$, as the fraction of the ble within a given radius, $r$, averaged over all conformations:

$$
v(r)=\frac{1}{W} \sum_{i=1}^{W} \frac{1}{W-1} \sum_{j=1}^{W} \begin{cases}1 & \delta^{i, j} \leq r \text { and } i \neq j  \tag{11}\\ 0 & \text { otherwise }\end{cases}
$$

ating the conformation space as an hypersphere with volume,

$$
\begin{equation*}
v(r)=C r^{n} \tag{12}
\end{equation*}
$$

us to identify $n$ as the marginal number of dimensions of the hyper-sphere and $C$ astant that depends on the value of $n$. We solve for $n$ as a function of $r$ by equating rithms:

$$
\begin{equation*}
\log (v(r))=\log (C)+n \log (r) \tag{13}
\end{equation*}
$$

$n$ as the slope in a plot of $\log (r)$ versus $\log (v(r))$.
n our previous work (Sullivan and Kuntz, 2001), we studied protein and polymer ith C $\alpha$-RMSD as the measure of conformational distance. In this paper, we will RMSD and $\boldsymbol{e}^{a, b}$, the maximal difference distance element, as defined earlier.
he concept of the marginal or "effective" dimensionality of conformation space larified with an example (Sullivan and Kuntz, 2001). Consider a conformation aped as a long, solid, cylindrical rod. The marginal dimensionality depends on scale being explored. On average, for any point surrounded by a sphere of

Is $r$ the sphere volume (i.e. the number of conformations if uniformly distributed) ases as the cube $(n=3)$ of the probe radius for $r$ much less than the diameter of the jut for large probe lengths, the number of conformers can only increase linearly ( $n=$ his same behavior is seen in molecular dynamics simulations of proteins where the inal dimensionality is equal to the total number of mechanical degrees of freedom or very small displacement. Larger displacements are limited to only a few degrees edom and/or correlated degrees of freedom (Sullivan and Kuntz, 2001).

## BLEM FORMULATION

Individual conformations of an N -mer bead can be characterized by their distance es, each composed of a unique $\boldsymbol{d}_{i, j}$ set for the corresponding $[d]_{i, j}$. Distance matrices 1 enough information to resolve all conformers except those related by a global on or handedness (Crippen and Havel, 1988). The problem we pose is to measure ormation contained in arbitrary sets of exact and "noisy" distance constraints. We ch this problem by:
entifying the information content, $\boldsymbol{I}$, of each $[d]_{i j}$.
suring the reduction in information resulting from correlation among exact
 ance elements.
nining various routes to useful sets of constraints, $\boldsymbol{M}$, of size $|\mathbf{M}|$, that iminate among all conformers. idering the reduction in information content arising from noise in $\boldsymbol{d}_{i, j}$.

## -TS

Ve begin by exploring the information content of a set of constraints consisting ied distances between numbered (i.e. "labeled") beads for lattice walks that
rve as models of molecular conformers. We start with the assumption that all these stances are known exactly and are free from "assignment" errors. We will call such nstraints "exact labeled constraints".

We first calculate the number of 2-D self-avoiding conformers as a function of ain length (Table 1). In table 2 we calculate the number of conformers that form perfect uares (see below). For convenience we also summarize these results in approximate alytical functions (Table 3). Given the simple dependence on chain length, we can lculate the (average) information content of adding a bead to the chain for different ttices and different chain constraints (Table 3). For comparison, we also include entries duced from entropic considerations for globular proteins.

## kact constraints

## formation content associated with individual labeled constraints

Information content varies in a predictable way for distance elements. It is also pendent on the particular lattice and move set under study (Table 3). For example, our priori decision to fix $\boldsymbol{d}_{i, i+1}$ to unit length means that knowledge of this distance carries partitioning information. In contrast, distance matrix elements with sequence paration, $s>1$, can assume multiple distance values and knowledge of these distances rtitions the ensemble. Establishing the rules for lattice walks is analogous to defining ference states in thermodynamics. Changes in entropy or information content based on w constraints are calculated with respect to the appropriate reference state which can, principle, be related to other reference states.

TABLE 3: Information content for lattice walks

| Lattice | Constraints | $\mathbf{W}(\mathbf{N})^{*}$ | Choices/residue | Bits/residue |
| :---: | :---: | :---: | :---: | :---: |
| 2D Square | None | $4^{N}$ | 4 | 2 |
|  | No Reversal | $3^{N}$ | 3 | 1.58 |
|  | Self Avoiding | $0.103\left(2.691^{N}\right)$ | 2.69 | 1.43 |
|  | Square Hamilton self-avoiding | 0.269 (1.399 ${ }^{N}$ ) | 1.4 | 0.48 |
| 3D Cubic | None | $6^{N}$ | 6 | 2.58 |
|  | No Reversal | $5^{N}$ | 5 | 2.32 |
|  | Self-avoiding (Chan and Dill, 1990) | 0.293 (4.782 ${ }^{\text {N }}$ ) | 4.78 | 2.26 |
|  | Hamilton-walk (Pande et al., 1994) | $\mathrm{e}^{-4.3 \pm 1.2}\left(1.86^{\mathrm{N}}\right)$ | 1.86 | 0.90 |
|  | Flory, mean field (Flory 1953) | $\left(\frac{z-1}{e}\right)^{N}$ | 1.84 | 0.88 |
| 3D Tetrahedral | None |  | 4 | 2 |
|  | No Reversal |  | 3 | 1.58 |
|  | Self Avoiding (Wang et al. 1999) |  | 1.72 | 0.78 |
| Stochastic Chains | Fit to extreme value distribution (Feldman and Hogue, 2002) |  |  | 1.2-2.0 |
| Protein Backbone | Native -> Compact (Dill 1985) |  | 1.7 | 0.76 |
| Protein Backbone + Sidechain | Native ->Unfolded (Cooper 1999) |  | 7.5-20.5 | 2.9-4.4 |

$\because: W(N)$ for $N \gg 1$.

FIGURE 1: Information content, I , for each distance element $[d]_{i, j}$ for $\mathrm{N}=15$. Color coded as indicated



All conformations of a 15 bead chain were enumerated, and the information content of each $[d]_{i, j}, \mathrm{I}\left(\left[d_{i, j}\right]\right)$, calculated according to Eq. 1, is shown in Fig. 1. As expected, information content increases for $[d]_{i, j}$ off the diagonal (Chan and Dill, 1990). This trend is seen more clearly in Fig. 2 which re-plots the information content for the
exhaustive ensemble of $N=16$ and the stochastic ensemble of $N=100$ as a function of $s$. There is a near-monotonic increase of information with sequence separation that is essentially independent of the chain length (Figs. 2, 3). For large $N$, the increase in information with $s$ is well approximated by a logarithmic function (Eq. 14) similar to the Jacobson-Stockmayer equation (Jacobson and Stockmayer, 1950) which computes the loss of entropy for loop closures as a function of loop size.

$$
\begin{equation*}
\mathrm{I}\left([\mathrm{~d}]_{\mathrm{i}, \mathrm{j}}\right)=1.36 \times \log _{2}\left(\mathrm{~s}_{\mathrm{i}, \mathrm{j}}\right)-0.92 \tag{14}
\end{equation*}
$$

Exhaustive enumerations of self-avoiding walks for $N=3$ to $N=16$, shows the tendency for even sequence separations to be slightly more informative than odd sequence separations (Fig. 3a). This observation is consistent with even-odd oscillations in other structural features on square lattices (Chan and Dill, 1989) and has no obvious implication for protein structures.

FIGURE 2: Mean information content as a function of $s$ for single distance elements for ensembles from chains of $\mathrm{N}=16(\mathrm{O})$ (exhaustive enumeration ) and $\mathrm{N}=100(\square)$ (stochastic enumeration of 10,000 conformations). The line fit is given by Eq. 14.



FIGURE 3: Information content by sequence separation. (a) Mean I $[d]_{i, j}$ as a function of $s_{i j}$. for single distance elements $[d]_{i, j}$, plotted for exhaustive ensembles of $N=4$ to $N=16$. (b) Independence of information content on chain length or chain position for fixed $\mathrm{s}_{\mathrm{i}, \mathrm{j}}=5$.


## Correlation of Constraints

Although the single most informative distance element is the "end-to-end" sequence separation (1, $N$ for odd $\mathrm{N} ; 1, N-1$ for even $N$ ) (Fig. 3a), finding the most informative set of distance elements is a more complex problem. The principal issue is the overlapping information contained in the distance elements. We begin by examining pairs of distance elements. A related problem has been considered in depth by Chan \& Dill (Chan and Dill, 1990), who calculated the entropic losses associated with pairs of pre-specified contacts for two and three dimensional lattices. In contrast, we examine the nonadditivity (loss) of information for all pairs of distance elements. We develop a numerical relationship that summarizes the average relative loss as a function of the separation of the distance elements. We quantify the correlation by the relative pair-wise information reduction for two distance elements $[d]_{i, j}$ and $[d]_{i^{\prime}, j^{\prime}}$ defined as:

$$
\begin{equation*}
\{\Delta I I I\}=\left\{\left[I\left([d]_{i, j}\right)+\mathrm{I}\left([d]_{i^{\prime}, j}\right)\right]-\left[\mathrm{I}\left([d]_{i, j},[d]_{i^{\prime}, j}\right)\right]\right\} /\left[\mathrm{I}\left([d]_{i, j},[d]_{i^{\prime}, j}\right)\right] \tag{15}
\end{equation*}
$$



This measure is bounded by zero (no loss), if there is no correlation, and unity for complete correlation. In Fig. 4(a-c) the relative loss of information is plotted as a function of $\left(i^{\prime}, j\right)$ for particular reference values of $(i, j)$ for $N=16$.

FIGURE 4: Relative information loss $\mathrm{II} / \mathrm{I}$ is plotted for all distance elements $[d]_{i j}$, assuming prior knowledge $[d]_{i, j}$. Reference $[d]_{i, j}:(1,11)$ for (a) and (d); $(1,16)$ for (b) and (e); $(4,13)$ for (c) and (f). For (a-c), the absolute information loss is plotted, equal to $\left\{\left[I(i, j)+I\left(i^{\prime}, j^{\prime}\right)\right]-\left[I\left(i, j ; i^{\prime}, j^{\prime}\right)\right]\right\} /\left[I\left(i, j ; i^{\prime}, j^{\prime}\right)\right]$. In $(d-f)$, the decimal logarithm of the information loss is plotted.


As expected, the loss is greater between elements close to each other in the distance matrix (Chan and Dill, 1990). Fig. 4(d-f) re-plots the information reduction logarithmically for the same reference distance elements. As the contour lines appear to lie more on the matrix diagonals than on circles about the reference point, we replot the
 $\log$ of the information loss as a function of the city-block sequence distance, $\boldsymbol{B}$, for all pairs of sequence distance elements for $N=14$ (Fig. 5a). This simple equation explains
much of the information loss behavior, with the correlation constant $\mathbf{r}^{2}=-0.882$ for the best-fit line. However, the individual sequence separations, $\boldsymbol{s}=s_{i, j}$ and $s^{\prime}=s_{i^{\prime} j^{\prime}}$, also influence the information reduction, where proximal distances with larger $s$ (and thus inherently more information) are reduced relatively more than distances with smaller $s$. Dividing $B$ by the sum of the sequence separation $(S S D)$, where $S S D=s+s^{\prime}$, tightens the correlation (Fig. 5b), bringing $\mathbf{r}^{2}=-0.920$. Most of the scatter is in the low informationloss (weak correlation) region of the plot. When considering only the points with $\Delta I / I>$ $0.001, \mathbf{r}^{\mathbf{2}}=-0.972$. While the scatter in information loss as a function of these simple distance element transformations appears significant on a logarithmic scale, it is much less significant on a linear scale. In Fig. 5(c, d), (1- $\Delta I / I)$ vs. $\boldsymbol{B}$ shows that at worse, $90 \%$ of the joint information is available at a city block separation of 4 and $95 \%$ of the information is available (worst case) at a $\boldsymbol{B}$ of 6 .

In summary, while we have no simple analytical statement of the information correlation of pairs of Cartesian distances, information loss is dominated by the sequence proximity (loop size) of the beads involved in the two distances, with the loss dropping rapidly for "loops" whose ends are separated by more than 4 beads.


FIGURE 5: Relative information loss, $\mathrm{DI} / \mathrm{I}$, for all pairs of distances, shown on a $\log _{10}$ scale, calculated by Eqn. 14 as a function of transformations of the distance element distances. In (a), the x -axis is the block element identity distance, $B$, equal to $\left|\mathrm{i}-\mathrm{i}^{\prime}\right|+\left|\mathrm{j}-\mathrm{j}^{\prime}\right|$. In (b), the x -axis is $\mathrm{B} / \mathrm{SSD}$, where $\operatorname{SSD}=\left(\mathrm{s}_{\mathrm{i}, \mathrm{j}}+\mathrm{s}_{\mathrm{i}^{\mathrm{i}}, \mathrm{j}}\right)$. (c) plots ( $1-[\mathrm{DI} / \mathrm{I}]$ ) versus B. (d) plots ( $1-\mathrm{DI} / \mathrm{I}$ ) versus B/SSD.





## Finding the Optimal Constraint Set

The optimal constraint set is defined as the smallest number of exact constraints that partition all the conformers uniquely. Distance-distance correlation makes the problem a difficult one. However, efficient procedures have been developed to construct any specific conformation on 2D and 3D lattices from distance data. Faulon et al show that $O(n)$ distances are sufficient for $n$ sites (Faulon et al., 2002). In this paper we wish to compare arbitrary constraint sets using the Shannon information to quantify the constraint set quality. Specifically, we examine three constraint sets:

1. The globally optimal constraint set. For a pre-specified set size, $|\mathbf{M}|$, the globally optimal set of distance constraints, $\boldsymbol{M}_{\text {global }}$, is determined by measuring $\boldsymbol{I}(\boldsymbol{M})$ for all
possible constraint combinations. Because of computational limitations, this calculation is only possible for small $N$ and small constraint set size $|\mathbf{M}|$.
2. The greedy algorithm constraint set. A less resource-intensive method is a "greedy" algorithm. The constraint set, $\boldsymbol{M}_{\text {greedy }}$, is calculated by first finding the single most informative distance constraint, $[d]_{\max }$ and then iteratively finding additional maximal constraints. In the case of our lattice models $[d]_{\max }$ is $[d]_{1, N}$ or $[d]_{1, N-I}$ for odd and even length chains respectively. Of course, this approach has the usual limitations of greedy algorithms (Cormen et al., 2001).
3. The random constraint set. Finally, as a simple control, we measure the information contained in sets of randomly selected distance constraints (Shakhnovich and Gutin, 1990).

Method 1: We calculate $\boldsymbol{I}(\boldsymbol{M})$ for all possible element combinations

$$
t!/(t-|\mathbf{M}|)!
$$

where $\boldsymbol{t}$ is the number of all possible pairings for a bead of length $N$, equal to $((N-2) \times(N-$ $\mathbf{1}) / \mathbf{2}$ ). As noted repeatedly, $\boldsymbol{I}(\boldsymbol{M})$ is not additive as $|\mathbf{M}|$ increases (Fig. 6). One element sets $\left(\left|\mathbf{M}_{\text {global }}\right|=1\right)$ are the most informative, per constraint, for all chain lengths. For large $\boldsymbol{N}$ and small $|\mathbf{M}|$, the information content of distance constraints approaches simple additivity; e.g. $I\left(M\left|\left|\mathbf{M}_{\text {global }}\right|=2\right)\right.$ is $81 \%$ greater than $I\left(\boldsymbol{M}\left|\left|\mathbf{M}_{\text {global }}\right|=1\right)\right.$, for $N=16$. Progressively more constraints yield less information per constraint. Combinatorial exploration of optimum constraints up to $\left|\mathbf{M}_{g l o b a l}\right|=5$ is shown in Fig. 6. For reference, $\boldsymbol{I}^{S}$ for each chain length is also given.

FIGURE 6: . Maximum information $\mathrm{I}^{\mathrm{M}}$ for best sets of distance constraints for $\left|\mathrm{M}_{\text {global }}\right|=1$ to 5 as a function of $N$. The line represents the maximum information per chain length based on the number of selfavoiding lattice walks (Table 1). $O: 1$ distance, $\square: 2$ distances, $\diamond: 3$ distances, $\triangle: 4$ distances, $\triangleleft: 5$ distances, $-\mathrm{I}^{\mathrm{s}}$.


Method 2: The best set of constraints found with the greedy algorithm for the 12 -mer chain shows a similar trend (Fig. 7). Fig. 7a illustrates a problem: the relatively small amount of information contained in the later choices makes the results very pathdependent. Fig. 7b shows the complex evolution of choices as the greedy algorithm explores the distance matrix. Interestingly, much of the information content can be realized with fewer constraints than the $N-2$ true degrees of freedom. For example, in a 15-mer chain, $95 \%$ of $\boldsymbol{I}^{S}$ can be encoded through a set of 8 distance elements $\left(\left|\mathbf{M}_{\text {greedy }}\right|=\right.$ 8) (Fig. 8). The difference between the number of constraints needed to achieve the maximum information and the number needed for a fixed percentage of the information increases exponentially with chain length. To recover $\boldsymbol{I}^{S}$ completely with the greedy alo orithm requires significantly more than $N-2$ distance constraints. This discrepancy
derives in large part from the imperfect search by such algorithms over all constraint combinations.

FIGURE 7: Information content dependence on number of constraints. $\mathrm{I}\left(M_{\text {greedy }}\right)$ was calculated using a greedy algorithm for $\mathrm{N}=12$. 17 distance constraints are required to obtain $\mathrm{I}^{M}$ by this method. (a) $\mathrm{I}\left(M_{g r e e d y}\right)$ versus number of distances, $|\mathrm{M}|$. The continuous line serves only to guide the eye. Dashed line $\mathrm{I}\left(M_{\text {random }}\right)$, averaged over 100 random constraint sets per $|\mathrm{M}|$, with standard deviation given by upper/lower bars. ( $b$ ) The greedy algorithm choices for $|\mathrm{M}|=17$ and $\mathrm{N}=12$ plotted by i,j identity.



FIGURE 8: Distance constraints for percentage information. The minimum number of distance constraints necessary to retain a given percentage, P , of the ensemble $\mathrm{I}^{\mathrm{S}}$, conformational information, is plotted as a function of chain length. A greedy algorithm was used to calculate the minimum number of distance constraints. P = $\mathrm{O}: 100 \%, \square: 95 \%, \diamond: 90 \%, \triangle: 80 \%, \triangleleft: 70 \%$.


Method 3: Random selection of constraints performs much worse than the previous

two strategies (Fig. 7a). Nearly twice as many randomly selected constraints are required to achieve the same level of information as those selected by the greedy algorithm.

There are practical issues raised by this analysis. Our calculations are limiting values for the information per constraint. Real systems will be less efficient for many reasons. First, only experiments that can report a range of distance values (e.g. fluorescence labeling, diffraction) can return the maximum amount of information per measurement. Second, only systems in which a significant fraction of all conformers are being sampled can approach the limits shown. More typically, in an experiment on compact states (e.g. native structures of proteins) with a method that is only sensitive to distances within a narrow range (e.g. NMR NOEs) one would expect considerably less information per measurement. Finally, we have been assuming that data are available to sufficient precision to discriminate all distance values for any distance element; "noise" in distance values will reduce the information even further. We explore these points more quantitatively in a later section.

## Information content of unlabeled distance constraints

One interesting difference between typical diffraction and NMR experiments on proteins is the "unlabeled" nature of the diffraction data until the "chain tracing" and "phasing" steps occur, while, in the NMR studies, assignment of the peaks can be carried out in a largely orthogonal manner to the calculation of tertiary structure. A simple assessment of the information contained in the assignments is available from lattice models of compact states as representatives of folded proteins. We ask what fraction of the total number of conformers have the maximal number of contacts for a given chain len gth. The ensemble of maximally compact structures contains the contacts that could give rise to (unassigned) NOEs. Each structure contains the same number of contacts. Adaitional information, beyond just the contact number, is needed to select an individual
structure from this set and can be taken as the information to be gained via the assignment procedure for well-folded structures. Values of the maximal number of contacts are simply calculated for square and rectangular Hamilton walks (see Chan \& Dill and below). For example, the number of square Hamilton walks is approximated by $W_{\text {SHW }}=1.40^{N-3.90}$ (Tables 2,3), so the additional information to find a unique structure from this set can be estimated as $\log _{2}(1.40)$ or .48 bits/bead (Cejtin et al., 2002; Pande et al., 1994). Attempts have been made to do "real space" assignments from NMR data (Grishaev and Llinas, 2002; Oshiro and Kuntz, 1993). This analysis indicates that any procedural or time-saving advantages of such approaches will carry a cost associated with the loss of orthogonal assignment information.

## Information loss from uncertainties in distance constraints

There are three major sources of uncertainty that affect distance measurements: 1) upper/lower bounds on the distance measurements, 2) imprecise distance measurements, and 3) misassignment of distances through incorrect labeling/assignment. Berger et al (Berger et al., 1996, 1999) have studied this last category of error, which we will not discuss here.

Uncertainty gives rise to information loss by preventing discrimination among d ifferent conformations. This information loss can often be attributed to the transmission stage of information transfer (Cole, 1993) and is defined as:

$$
\begin{equation*}
I^{L}=\left(I^{S}-I^{M}\right) \tag{16}
\end{equation*}
$$

## Bo ennd limitations

Consider an upper bound on distances, $\boldsymbol{D}_{\boldsymbol{u}}$, such that,

$$
d_{i, j} \leq D_{u}
$$

and $\boldsymbol{D}_{u}$ depends on the physical principles of the experiment and the experimental conditions. For example, since the magnitude of an NOE is proportional to $d^{6}$, NOEs are typically only determined for hydrogen atoms separated by $<5 \AA$. For our lattices, assuming a one-bead to one-residue mapping, detecting an NOE would be equivalent to knowing that two beads are in contact, i.e. separated by the lattice unity distance. Fluorescence energy transfer and chemical cross-link data have longer distance limits. Crystallographic structures have upper bounds set by the smallest diffraction angle that can be observed and lower bounds related to the limit of resolution. We want to calculate the dependence of the information content on the distance detection limit, $\boldsymbol{D}_{\boldsymbol{u}}$.

If the particular experiment provides a monotonic relationship between "signal intensity" and "distance", we can proceed in a straightforward manner to assign distances greater than $\boldsymbol{D}_{\boldsymbol{u}}$ a lower bound of $\boldsymbol{D}_{\boldsymbol{u}}$. For example, it is common practice in some experiments and calculations to report atom pairs as either a 'contact' $\left(d_{i, j} \leq \boldsymbol{D}_{u}\right)$ or 'nocontact $\left(d_{i, j}>\boldsymbol{D}_{u}\right)$. However, in NMR and FRET the measured signal is a product of both a distance term and an angular correlation term which can drive the signal close to zero regardless of the distance. To be logically consistent with the underlying physics we must allow for this possibility and give all distances the same lower bound, $D_{l,}$ for such experiments.

In the first case, where the distance magnitudes are unambiguous, $\boldsymbol{I}^{M}$ increases linearly for all values of $\boldsymbol{D}_{u}$ (Fig. 9).


FIGURE 9: $\mathrm{I}^{\mathrm{M}}$ with upper bounds on distances. For the unfilled symbols ( $\mathrm{O}, \square, \square: \mathrm{u}=1,1.42,2$ units respectively), $\mathrm{I}^{\mathrm{M}}$ is calculated from all interbead distances encoded as $\mathrm{d}_{\mathrm{i}, \mathrm{j}}$ for $\mathrm{d}_{\mathrm{i}, \mathrm{j}} \leq \mathrm{u}$ and as equal to u for $\mathrm{d}_{\mathrm{i}, \mathrm{j}}$ $>u$. For the filled symbols ( $\square, \boldsymbol{A}: u=1.42$, 2units repectively), as above, except distances longer than $u$ are treated as unknown. Solid line: No limit.

$\boldsymbol{I}^{\boldsymbol{M}}$ for the most limiting contact/no-contact detection limit $\left(\boldsymbol{D}_{\boldsymbol{u}}=1\right)$ retains nearly half the value of $\boldsymbol{I}^{S}$. However, the second case, where we are not allowed to use "negative" data, the information content of the experiment is much less. $\boldsymbol{I}^{\boldsymbol{M}}$ equals zero for simple
 contact/no contact decisions. Only for $\boldsymbol{D}_{u} \geq 2$, i.e., "next-nearest neighbors", does such an experiment yield information on the 2D lattice ensemble.

The dependence of $\boldsymbol{I}\left([d]_{i, j}\right)$ on $\boldsymbol{D}_{\boldsymbol{u}}$ varies with sequence separation (Fig. 10). Information content decreases the most for large sequence separations and low values of $\boldsymbol{D}_{\boldsymbol{u}}$ - In general, the most informative distance elements have sequence separations of $\boldsymbol{D}_{\boldsymbol{u}}-42$. For example, the most informative contact/no contact $\left(\mathbf{D}_{u}=1\right)$ distance element occurs at a sequence separation of three, and only yields 0.53 bits. Thus, the information
content of knowing that a contact exists, which generally increases with sequence separation (as contacts become more rare with increasing sequence separation) is offset by the loss of the information potential of knowing the distances associated with longer sequence separations. The rarity of contacts at larger sequence separations means that knowing two highly separated residues are in contact is very informative. This is seen in Fig. 11 which plots the information content of knowing two beads $(i, j)$ are in contact $\left(\boldsymbol{d}_{i, j}\right.$ $=1)$ as a function of $s_{i j}$.

FIGURE 10: Information content, $\mathrm{I}\left[d_{i, j}\right]$ by sequence separation with bounded distance detection. Mean information content as a function of sequence separation for single distance constraints is plotted for $\mathrm{N}=15$ with given distance detection limits, u . Distances are encoded as $\mathrm{d}_{\mathrm{i}, \mathrm{j}}$ for $\mathrm{d}_{\mathrm{i}, \mathrm{j}} \leq \mathrm{u}$ and as equal to u for $\mathrm{d}_{\mathrm{i}, \mathrm{j}}>\mathrm{u}$. O: 1 unit, 1.42 units, $\diamond: 2$ units, $\triangle: 6$ units, $\triangleleft:$ No Limit


FIGURE 11: Information content of contact/no contact determinations. Information content of knowing a contact exists $(\mathrm{d}=1)$ is plotted averaged over distance identities of the given sequence separations. Values for even-value sequence distances are not given since these contacts are geometrically unfeasible. O: 10mer, $\square: 16$-mer, $\triangle: 49$-mer.


## Uncertainty due to limitations in precision of measurements

An issue common to all experiments is the magnitude of the "noise" or imprecision in the measurements. To explore the impact of random noise on the ability to distinguish conformations from one another, we consider two limiting cases for fully enumerated conformational ensembles from 2D lattices. First, we identify conformations most resistant to noise, defined as pairs of conformations which are maximally different and second, in the same ensemble, we find which conformational pairs are most similar. The conventional measures for conformational difference (see Methods section) are the RMS atom-position difference after superposition and the RMS of the distance-difference
matrix elements (Levitt, 1976). We will also use the largest element in distancedifference matrix, $\boldsymbol{e}^{a, b}$ (see Methods).

FIGURE 12: . Limiting conformations on 2D square lattices. Top pair (a) has the largest $e^{a, b}$ value (even$N$ ) and the bottom pair (b) illustrates the lowest $e^{a, b}$ pair ( $N \geq 7$ ).





The use of $\boldsymbol{e}^{a, b}$ yields unexpectedly simple comparisons among chains of different lengths, especially when $\boldsymbol{e}^{a, b}$ is normalized through division by the chain length, $N$. The value $\boldsymbol{e}^{a, b}$ can assume has natural limits. The largest possible distance differences, over all conformational pairs, is in element $[d]_{1, N}$ (Fig. 12a). The smallest possible non-zero difference elements likewise occur near $[d]_{1, N}$ for cases where the bead displacement between two conformations is nearly orthogonal to the inter-bead vector. For $N \geq 7$, the smallest $\boldsymbol{e}^{a, b}$ over all pairs of $\boldsymbol{d}^{a}$ and $\boldsymbol{d}^{b}$ for the ensemble of 2D conformers is in the single conformational pair in Fig. 12b for which

$$
\begin{equation*}
\Delta_{i, N}(a, b)=\left((N-3)^{2}+4\right)^{1 / 2}-(N-3) \tag{17}
\end{equation*}
$$

To provide an overview of the distribution of conformer-conformer differences we plot, in Fig. 13, the fraction of distinguishable conformational pairs compared to all conformational pairs, $(1-v(r))$, as a function of $e^{a, b} / \boldsymbol{N}$ for the fully enumerated square conformational pairs, $(1-\boldsymbol{v}(\boldsymbol{r}))$, as
lattice walks of size up to $N=13$.

FIGURE 13: Conformational distinguishability. The fraction of distinguishable conformational pairs compared to the total number of conformational pairs, equal to $1-v(r)$, (see text) is plotted as a function of the relative noise, equal to $r=e / N$. $\bigcirc: 3$-mer, $\triangle: 6$-mer, $\triangle: 9-\mathrm{mer}, *: 12$-mer, $O: 13$-mer.



In addition to these complete distributions, we also show the limiting values for the most similar and most different pairs of conformers for $3 \leq N \leq 25$. A related plot shows the fraction of indistinguishable conformational pairs compared to all pairs (Fig. 14). Both plots show a remarkable independence from chain length.

FIGURE 14: Conformational Indistinguishability. $\mathrm{v}(\mathrm{r})$ is plotted as a function of the relative uncertainty, $r$ $=e / N$. The limiting threshold noise levels for ensembles $N=7$ to 23 are given by ensemble identity, N , and are placed at $x=\left\{\left[(N-3)^{2}+4\right]^{1 / 2}-(N-3)\right\} / N, y=2 /\left[W^{*}(W-1)\right]$ which are the limiting relative noise levels and inverse of total number of conformational pairs, respectively. $O$ : 3-mer, $\square: 6$-mer, $\diamond: 9$-mer, $\triangle: 12$ mer, $\nabla$ : $15 \mathrm{mer}, 7-23$ : Limiting errors for 7-23mers.


There are several features of Fig. 14 that are useful for our analysis. First, as noted earlier, $\boldsymbol{v}(\boldsymbol{r})$ can be thought of as a cumulative distance distribution function for pairs of ensemble within a specific error, or conformational distance, of a given conformation, averaged over all ensemble members. It also provides a visualization of the impact of noise on the ability to discriminate one conformer from all the others. Additionally, we can calculate the marginal dimensionality, $n$, from Fig. 14 by computing the slope of the line passing through the points for limiting conformational pairs from the ensembles of length $N-1$ and $N+1$. We find, for 2D square lattices, that the limiting marginal
dimensionality is nearly equal to $N-2$, the true number of mechanical degrees of freedom for these walks (Fig. 15).

FIGURE 15: The marginal dimensionality, $n$, is plotted as a function of the chain length, $N$. The marginal dimensionality for N was calculated from the logarithmic slope between the two points for $\mathrm{N}-1$ and $N+1$ in Figure 14.


This value for the slope can also be derived directly from the formulas given in the legend of Fig. 14, assuming $N \gg 1$. Following this idea one step further, we can interpret the slope at all points on Fig. 14 as the number of degrees of freedom that are "effective" in producing the conformational differences associated with a particular (normalized) displacement.

## Effect of uncertainty on compact lattice structures

The properties of the fully enumerated ensembles are dominated by extended conformers analogous to denatured states of proteins. To provide insight into arguably

more biologically relevant ensembles such as the native and molten globule protein states (Chan and Dill, 1989), we studied the subset of compact conformers by generating perfect-square Hamilton walks where every lattice site is occupied. We exhaustively enumerated square Hamilton walks up to $N=49$ (Table 2).

The dependence of information content on bead sequence separation is fundamentally different in square Hamilton walk ensembles compared to full enumeration ensembles (Fig. 16; compare to Fig. 2,3).

FIGURE 16: Mean information content for Hamilton square walks as a function of sequence separation for single distance constraints ( $N=9,16,25,49$ ), a full enumeration ensemble (FE) $(N=16)$ and a stochastic, non-exhaustive ensemble of unconstrained conformations (FES) ( $N=49$ ). $\mathrm{O}: 16$-mer FE unit, $\square: 49$-mer FES, $\diamond: 9$-mer HW, $\triangle$ : 16 -mer HW, $\checkmark: 25$-mer HW, $\triangleright: 49$-mer HW.
 correlation between beads $i$ and $j$ is constant for sequence separations greater than the diagonal distance, which increases as $\sim N^{1 / 2}$.


We also calculated $v(r)$ as a function of $\boldsymbol{e}^{a b} / \boldsymbol{N}$ for the Hamilton walk ensembles (Fig. 17). The curves are surprisingly similar to the fully enumerated walks (Fig. 14), even though the Hamilton walk ensembles sample only a small subset of full enumeration conformational space and have additional degeneracy. For example, in the $\mathrm{N}=36$ Hamilton walk ensemble, 3,608 pairs of conformations become indistinguishable with an absolute uncertainty of 1.24 (equal to $\sqrt{5}$-1) or relative uncertainty of 0.0343 .

FIGURE 17: Conformational indistinguishability for Hamilton walks. $v(r)$ is plotted as a function of the relative noise, $e / N$, for the Hamilton walk constrained ensembles. See Fig. 14 O: 9-mer, $\square$ : 16-mer, $\diamond: 25$ mer, $\Delta: 36$-mer, $\nabla: 49$-mer.


The full enumeration ensembles, discussed previously, have limiting characteristics largely governed by simple relationships among extended conformations. None of these situations arises when the ensemble of interest is restricted to compact conformers. Thus it is not clear at this point whether the similarity in (normalized) pair distributions arises from some fundamental principle or from specific geometric constraints.


We note that the pair distributions show multi-modal character (notice the small break in the curve near $e^{a b} / N=0.03$ in Fig. 14), as we saw in our earlier work on nonlattice chains (Sullivan and Kuntz, 2001). Very similar $\boldsymbol{v}(\boldsymbol{r})$ distributions are obtained using off-lattice polyalanine chains (Fig. 18). Note that the 30 residue chains with 58 dihedral degrees of freedom closely approximate the distribution from a stochastic sampling of a 60 bead ( 58 degrees of freedom) 2D lattice walk.

FIGURE 18: Conformational indistinguishability for stochastic polyalanine ensembles (see text and Fig. 14) $O$ : Yarn 30 Extended, $\square$ : Yarn 30 Compact, $\leqslant$ : 2D Lattice Stochastic $N=60$, Numbered points 7-15: Limiting Distances for $\mathrm{N}=7-15$ for 2D extended walks( see Fig. 14.) Yarn $e$ values are divided by 3.8 N . 2D lattice $e$ values are divided by $N$.


## Relating information loss to noise

Extracting a relationship between information loss and noise requires a detailed model of how noisy messages are misread. One such model uses the "noise sphere" concept outlined in the Methods section. Briefly, a set of $W$ distinct messages, $\left\{w_{i}\right\}$, becomes scrambled as noise is introduced and some messages become indistinguishable. Note that this approach requires that after the noise has been introduced, every conformer still be a proper member of the set - distorted (off-lattice) geometries are not allowed. We define a "noise-sphere" in which conformations $w_{j}$ that are within a hyper-sphere of radius $r$ centered about conformation $w_{i}$ are indistinguishable. The radius $r$ can be associated with any measure of "noise" and formulated with any explicit error distribution function: we use either RMSD or $\boldsymbol{e}^{a, b}$ and assume a uniform distribution of noise.

This procedure can be used for entire conformations, but as noted in the methods section, it is also directly applicable to information loss for individual constraints or sets of constraints. In Fig. 19 we show the fractional information loss for the $[d]_{l, N}$ distance element for 2D chains as a logarithmic function of the noise magnitude which we take as the noise sphere radius. While there is some small dependence of normalized loss on the chain length, the curves indicate a smooth relationship with half of the information lost when the noise magnitude is equal to the lattice spacing.
ismenemase

FIGURE 19: Relative information loss for $[d]_{l, N}$, full enumeration ensemble, for $\mathrm{N}=9(\mathrm{O}), 11(\square), 13(\diamond)$, 15( $\Delta$ ).


For 2D lattice walk ensembles, the relationship between information loss per degree of freedom and $e^{a, b}$ (derived via Eq. 10) is shown in Fig. 20a. The curves relating information loss and coordinate RMSD for the same ensembles are shown in Fig. 20 b. Fig. 21 shows similar plots for Hamilton walks. At very low noise magnitudes, there is no information loss, as expected for a set of discrete conformers. As the noise increases beyond a critical value, there is a region of barely perceptible loss as the most similar conformers are merged. At some point, increasing error causes major information loss because many conformations populate the average noise sphere. Finally at large noise levels, there is a slow loss of information because only the most different conformers are left to merge.

FIGURE 20: Information loss per degree of freedom for full enumeration, for $\mathrm{N}=7(\triangleleft), 8(\nabla), 9(\triangleright)$, $10(+), 11(\times), 12(*), 13(\bullet), 14\left(\begin{array}{|}\end{array}\right), 15(\bullet)$. The factor $(\mathrm{N}-2.29)$ comes from $\mathrm{I}^{\mathrm{S}}=1.43(\mathrm{~N}-2.29)$, a recasting of the self-avoiding walk equation for $W(N)$ in Table 3. (a) plotted against $e$. (b) plotted against coordinate RMSD.


FIGURE 21: Information loss per degree of freedom for compact two dimensional lattice structures for N $=16(\square), 25(\diamond), 36(\Delta)$. (a) plotted against $e$. (b) plotted against coordinate RMSD.


To summarize this section: the noise sphere model allows a straightforward treatment of the effect of noise on information content for individual distance elements, sets of distance constraints, and full enumeration conformational ensembles. Not surprisingly, the information loss/noise curves are the steepest when the noise magnitude
is near the lattice spacing. Most of the chain length dependence can be removed by reporting information per residue which is sensibly constant at longer chain lengths.

## Information per constraint

As a practical matter, experimentalists are interested in how much information can be extracted from a necessarily limited set of measurements. This question has been addressed at various levels of sophistication by many authors. For example, Gutin and Shakhnovich have studied a model of polymer chains where the entropy loss on random cross-linking yields a leading term proportional to the number of cross-links per residue (Shakhnovich and Gutin, 1990). Our analysis of exact constraints on fully enumerated conformers also yields some limiting answers. For a 2D self-avoiding walk on a square lattice a single optimal measurement can provide $\sim \log _{2} N$ bits (Fig. 2, 3) while $N$ beads can be fixed on the lattice with $N-2$ constraints for any given conformer, or $\sim 1.5$ bits/constraint. Compact structures, such as the 2D Hamilton walks, can yield even more information per constraint (Fig.11). If the optimal set of constraints is not available, more measurements are needed. For example, for $N=12,17$ constraints are needed to supply 14 bits or 0.8 bits/constraint (Fig. 7). If constraints are chosen randomly, many more
 would be required to supply the same information. Thus, for exact (noise free) constraints, one might expect ca. 0.5 bits/constraint over a random set of measurements. Gutin and Shakhnovich give similar numerical results when converted into the same units. They report $0.5-1.5$ bits/constraint over the chain lengths we consider.

If we turn to constraints that contain random noise, the information content decreases further. Using Fig. 19 we can estimate that noise levels of ca. 0.1 lattice units for individual distance constraints cost less than $10 \%$ of the information per residue,
while noise levels comparable to the lattice spacing would require doubling the number of constraints to achieve the same information content as noise-free measurements. For noise levels greater than the lattice spacing, the information content per residue diminishes very rapidly. Fig. 19 suggests that at noise levels of twice the lattice length, five times as many constraints would be needed compared to the exact constraints. These numbers are, of course, very approximate guides. Presumably, an analogous estimate would apply to non-lattice models of polymers as long as discrete conformers can be enumerated. For polypeptide chains, the results of Trover and Cohen (Trover and Cohen, 1995) imply an absolute minimum separation of ca. $0.1 \AA$ per residue or a relative separation of $0.001 \AA$ per residue ${ }^{2}$ for a 100 residue protein. These limiting "conformational radii" are quite comparable to those for the most similar conformers in 2D lattice walks of the same length derived from Eq. 17.

In summary, by considering the effects of noise on single distances, we are able to make estimates of how much additional effort is required, in a best-case scenario, to overcome the information loss due to random noise in measurements.

## DISCUSSION

Developing a general and quantitative treatment of information content for macromolecular ensembles raises both fundamental and practical issues. One serious concern is the need for enumeration of the conformations. Exhaustive enumeration will always be limited by computational resources and is not applicable to off-lattice models for the foreseeable future (Sullivan and Kuntz, 2001). Feldman and Hogue's more optimistic view (Feldman and Hogue, 2002) is based on the extreme value distribution function which may overestimate the number of structures at small RMSD. The real goal

for off-lattice structures is an analytic distribution function with sufficient accuracy to derive thermodynamic properties. The relative simplicity of the $v(r)$ vs. $e$ curves offers some hope that such functions can be devised, although the multimodal character of the curves indicates that direct stochastic sampling may not suffice to probe the most closely related conformers.

The data for various lattice and off-lattice systems (Table 3) raises the question of what reference state is most appropriate for comparisons among different models. The most obvious choice is an unconstrained ideal gas. This is roughly analogous to measuring thermodynamic energies using $\mathrm{E}=\mathrm{mc}^{2}$ - it gives the right answers in a very awkward form. The important point is that the choice of lattice and lattice move set (or any other prior constraints) influences the information content of the resulting ensemble, with varying amounts of residual information (entropy) being associated with the set of choices.

The application of noise theory requires the development of parametric noise models and a set of choices for parameter values. There is currently little guidance from physical principles for choosing error metrics and clustering methods. We elected to use a very simple formulation of the problem based on the application of the noise sphere model to fully enumerated lattice ensembles. We postpone a treatment of energetic differences among conformers, although they could be put directly into Eq. 10 as population weights. We assume the noise to be white noise which implies uniform probability of "scrambling" for all conformers within the noise sphere. More realistic, distance-dependent noise functions could also be readily incorporated. We chose displacement measures pragmatically rather than attempting a full physical analysis. We

noted earlier that the noise sphere model is formally adapted to accept other displacement metrics. More sophisticated entropic clustering models are available from information theory (Guiasu, 1977). However, their computational complexity is extremely high, and they are not practicable even for small 2D ensembles.

While our specific results for the information per constraint and information lost as a function of noise are limited to the ensembles studied, the general features of these curves can provide useful insight into experimental design. It certainly should be possible to extend these ideas to proteins and nucleic acid polymers. In situations where diverse types of data are used and noise propagation is poorly understood, maximum-information optimization using hypothetical models of transmission errors could help determine which combinations of various measurements are most informative. This would be a first attempt toward improving the utility of measurements in such systems, a critical step if we are to improve the quality and speed of current structure determination methods (Rabitz, 1989).

## CONCLUSIONS

1. Information content of distance constraints increases as the log of the sequence separation for all systems studied except square Hamilton walks where a limiting value is reached as the sequence separation reaches $\sqrt{ } \mathrm{N}$.
2. While a single noise-free distance constraint, namely the end to end distance, can select individual conformers from an ensemble and construction methods exist that use as few as $N-2$ distance constraints per conformer, the size of the set of constraints needed to uniquely partition the entire ensemble is not known in a general way. The problem is inherently complex (Chan and Dill, 1990) arising

from correlations among distance elements that are largely local in sequence space. We show that a simple greedy algorithm can supply an arbitrarily high percentage of the total information (e.g. 95\%) with many fewer than $N-2$ constraints. On a practical level, randomly selected exact constraints provide much less information, which we estimate to be 0.5 bits/constraint, on the average, for 2D lattice ensembles.
3. Using the "noise sphere" model, we show that noise reduces information content in a surprisingly universal way for fully enumerated lattice walks and maximally compact Hamilton square walks. It is not possible to use the same model for offlattice ensembles without some method of estimating the total number of conformations.
4. The slope of the information loss vs. noise curves can be directly related to the number of active or "effective" degrees of freedom for the ensemble.
5. A complete quantitative treatment of information content is surprisingly difficult. Many technical issues arise that involve additional assumptions which influence the numerical results. These issues include: choice of potential functions, clustering methods, and noise distribution functions among others. There is currently little guidance from physical principles or experiment for this selection. More work is needed to clarify the best way to extend these studies to off-lattice ensembles.


## ACKNOWLEDGMENTS:

We are grateful to Gordon Crippen, Ken Dill, and Scott Pegg for helpful comments.
Supported by a grant from the National Science Foundation (NSF CHE-0118481); R.L.

Guy, Principal Investigator.

## REFERENCES:

Berger, B., J. Kleinberg, and T. Leighton. Reconstructing a three-dimensional model with arbitrary errors.; 1996; Philadelphia, PA. p 449-458.

Berger, B., J. Kleinberg, and T. Leighton. 1999. Reconstructing a three-dimensional model with arbitrary errors. Journal of the ACM 46(2):212-235.

Brunger, A.T., G.M. Clore, A.M. Gronenborn, R. Saffrich, and M. Nilges. 1993. Assessing the Quality of Solution Nuclear Magnetic Resonance Structures by Complete Cross-Validation. Science 261(5119):328-331.

Cejtin, H., J. Edler, A. Gottlieb, R. Helling, L. Hao, J. Philbin, W. Ned, and T. Chao. 2002. Fast tree search for enumeration of a lattice model of protein folding. Journal of Chemical Physics 116(1):352-9.

Chan, H.S., and K.A. Dill. 1989. Compact Polymers. Macromolecules(22):4559-4573.
Chan, H.S., and K.A. Dill. 1990. The effect of internal constraints on the configurations of chain molecules. Journal of Chemical Physics 92(5):3118-35.

Chan, H.S., and K.A. Dill. 1991. Polymer principles in protein structure and stability. Annu Rev Biophys Biophys Chem 20:447-90.

Chorev, M., and M. Goodman. 1995. Recent Developments in Retro Peptides and Proteins - an Ongoing Topochemical Exploration. Trends in Biotechnology 13(10):438-445.

Choy, W.Y., and J.D. Forman-Kay. 2001. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. J Mol Biol 308(5):1011-32.

Cole, C. 1993. Shannon revisited: Information in terms of uncertainty. Journal of the American Society for Information Science 44(4):204-211.

Cormen, T.H., C.E. Leiserson, and R.L. Rivest 2001. Introduction to Algorithms. McGraw-Hill

Crippen, G.M. 2000. Enumeration of cubic lattice walks by contact class. Journal of Chemical Physics 112(24):11065-11068.

Crippen, G.M., and T.F. Havel 1988. Distance Geometry and Molecular Conformation. Wiley, New York

Dill, K.A., S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. 1995. Principles of protein folding--a perspective from simple exact models. Protein Sci 4(4):561-602.


Dobson, C.M., A. Sali, and M. Karplus. 1998. Protein folding: A perspective from theory and experiment. Angewandte Chemie International Edition 37(7):868-893.

Faulon, J.L., M.D. Rintoul, and M.M. Young. 2002. Constrained walks and self-avoiding walks: implications for protein structure determination. Journal of Physics AMathematical \& General 35(1):1-19.

Feldman, H.J., and C.W. Hogue. 2002. Probabilistic sampling of protein conformations: new hope for brute force? Proteins 46(1):8-23.

Flory, P.J. 1953. Principles of Polymer Chemistry. Cornell University Press, New York
Gregoret, L.M., and F.E. Cohen. 1990. Novel method for the rapid evaluation of packing in protein structures. J Mol Biol 211(4):959-74.

Gregoret, L.M., and F.E. Cohen. 1991. Protein folding. Effect of packing density on chain conformation. J Mol Biol 219(1):109-22.

Grishaev, A., and M. Llinas. 2002. Protein structure elucidation from NMR proton densities. Proceedings of the National Academy of Sciences of the United States of America 99(10):6713-6718.

Guiasu, S. 1977. Information Theory with Applications. McGraw-Hill
Gutin, A.M., and E.I. Shakhnovich. 1994. Statistical Mechanics of Polymers with Distance Constraints. Journal of Chemical Physics 100(7):5290-5293.

Havel, T.F., I.D. Kuntz, and G.M. Crippen. 1983. The theory and practice of distance geometry. Bulletin of Mathematical Biology(45):665-720.

Irback, A., and C. Troein. 2002. Enumerating Designing Sequences in the HP Model. Journal of Biological Physics(28):1-15.

Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104(1):59-107.

Levy, R.M., M. Karplus, J. Kushick, and D. Perahia. 1984. Evaluation of configurational entropy for proteins: Application to molecular dynamics simulations of an $\alpha$ helix. Macromolecules(17):1370-1374.

Luo, H.B., and K. Sharp. 2002. On the calculation of absolute macromolecular binding free energies. Proceedings of the National Academy of Sciences of the United States of America 99(16):10399-10404.

Oshiro, C.M., and I.D. Kuntz. 1993. Application of distance geometry to the proton assignment problem. Biopolymers 33(1):107-115.

Pande, V.S., C. Joerg, A.Y. Grosberg, and T. Tanaka. 1994. Enumerations of the Hamiltonian walks on a cubic sublattice. Journal of Physics A (Mathematical and General) 27(18):6231-6.

Potter, M.J., and M.K. Gilson. 2002. Coordinate systems and the calculation of molecular properties. Journal of Physical Chemistry A 106(3):563-566.

Rabitz, H. 1989. Systems analysis at the molecular scale. Science 246:221-226.
Rosenbluth, M.N., and A.W. Rosenbluth. 1955. Monte Carlo Calculation of te Average Extension of Molecular Chains. Journal of Chemical Physics 23(2):356-359.

Schafer, H., X. Daura, A.E. Mark, and W.F. van Gunsteren. 2001. Entropy calculations on a reversibly folding peptide: Changes in solute free energy cannot explain folding behavior. Proteins: Structure, Function, \& Genetics 43(1):45-56.

Schafer, H., A.E. Mark, and W.F. Van Gunsteren. 2000. Absolute entropies from molecular dynamics simulation trajectories. Journal of Chemical Physics 113(18):7809-17.

Schlitter, J. 1993. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. Chemical Physics Letters 215(6):617-21.
Shakhnovich, E., and A. Gutin. 1990. Enumeration of all compact conformations of copolymers with random sequence of links. Journal of Chemical Physics 93(8):5967-71.
Shannon, C.E. 1948. A mathematical theory of communication. The Bell System Technical Journal(27):379-423 and 623-656.
Sibbald, P.R. 1995. Deducing protein structures using logic programming: exploiting minimum data of diverse types. J Theor Biol 173(4):361-75.

Sullivan, D.C., and I.D. Kuntz. 2001. Conformation spaces of proteins. Proteins 42(4):495-511.

Troyer, J.M., and F.E. Cohen. 1995. Protein conformational landscapes: energy minimization and clustering of a long molecular dynamics trajectory. Proteins 23(1):97-110.

Wang, Z.H., M.B. Luo, and J.M. Xu. 1999. Conformational entropy of self-avoiding polymer chains. European Polymer Journal 35(5):973-975.

Young, J.F. 1971. Information Theory. Butterworth \& Company Ltd., Bristol


# Chapter Four: An Information Theoretic Approach to Macromolecular Modeling: I. Sequence Alignments 

Tiba Aynechi ${ }^{\ddagger}$, Irwin D. Kuntz ${ }^{*}$<br>* Department of Pharmaceutical Chemistry<br>${ }^{*}$ Graduate Group in Biophysics<br>University of California, San Francisco<br>San Francisco, California, 94143-0446

Correspondence should be addressed to Dr. Irwin D. Kuntz, Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, California, 94143-0446
Phone: (415) 476-1937
Fax : (415) 502-1411
Email: kuntz@cgl.ucsf.edu

Submitted for publication to Biophysical Journal

## KEYWORDS

Information Theory, Shannon Entropy, Lattice models, Sequence Alignment, Gap Penalty.



#### Abstract

We are interested in applying the principles of information theory to structural biology calculations. In this paper, we explore the information content of an important computational procedure: sequence alignment. Using a reference state developed from exhaustive sequences, we measure alignment statistics and evaluate gap penalties based on first principle considerations and gap distributions. We show that there are different gap penalties for different alphabet sizes and that the gap penalties can depend on the length of the sequences being aligned. In a companion paper we examine the information content of molecular force fields.


## INTRODUCTION

Structural biology uses many different experimental and computational procedures to deduce the structure and function of biomacromolecules. A partial list includes sequence analysis, crystallography, magnetic resonance, spectroscopy, homology modeling, and molecular dynamics. Despite the quantitative nature of such undertakings, there is no unifying model of information content and error analysis for the field as a whole. While we do not discount the importance of several specialized forays (Berger et al., 1999; Brunger, 1992; Carothers et al., 2004; Levitt and Gerstein, 1998; Luzzati, 1952; Park and Levitt, 1996; Stroud and Fauman, 1995) there is a need to build on these efforts to seek a broader approach that would permit evaluation and comparison of the wide range of existing methods. Another concern is to understand the additivity of information when different techniques are combined.

To examine these issues, consider the information content of a protein structure. In the simplest form, it is related to the number of conformations the molecule can adopt. Knowing this value for a specific protein would allow us to set an upper bound on the difficulty of predicting the structure of the native state from, say, the amino acid sequence; the classic "protein folding" problem. Experiments and/or modeling, from this point of view, generate constraints that reduce the allowed conformations. None of the current experimental or computational approaches offers a perfect solution to finding structures (Kuntz and Agard, 2003). Instead, our focus is the quality of the "best" structures consistent with the information and techniques available. In previous work (Sullivan et al., 2003) we used information theory to quantify information derived from a knowledge of internal distances. In these papers, we wish to extend our methods to two
other aspects of computational biology that underpin methodologies for structure prediction: sequence analysis and "force fields". This paper considers sequence alignment issues, while the companion paper deals with the information content of molecular force fields.

Aside from ab initio methods (Bonneau and Baker, 2001) theoretical structure prediction is generally approached in two steps; 1) Given an amino acid sequence, find an appropriate structural template (using homology modeling and/or threading), and 2) refine the structural model to produce an energetically minimized or "best scoring" conformation. The first step requires sequence alignment algorithms which rely heavily on the use of empirical parameters i.e. gap penalties and scoring matrices (Vingron and Waterman, 1994); consequently, it is difficult to either evaluate or improve their over-all performance except in the context of specific training sets. The second step, determining the lowest energy conformation from among all possible three-dimensional structures of a given sequence, requires an energy function capable of discrimination among native and decoy conformations. Force fields are considered in a second paper (Aynechi and Kuntz, 2005).

In this paper we examine the fundamental linkage between sequences and information theory from which we can derive the ideal costs of alignment procedures, scoring matrices, and gap penalties.

## METHODS Overview

Our basic approach is to write out all possible occurrences of the set of interest (i.e. sequences) and then ask what the informational consequence of performing an

operation that combines or clusters some of the objects would be. The information required to select one object from a set of $W$ objects is $\log _{2} W$. If during any stage of the procedure some objects become indistinguishable from others, the set can be considered to be "clustered" into $W^{\prime}$ ' distinguishable subsets. Such clustering reduces the effective number of objects and hence reduces the information required to select subsets from the transformed set. It is important to note that the normal use of sequence alignment procedures is to increase the information associated with a given probe sequence. In such usage, one queries a database of sequences and assigns properties (structure, function) to the probe sequence based on the statistically valid matches that are found. This information "increase" arises as a single sequence is placed into a cluster smaller than the full set of sequences from which it was indistinguishable before the alignment procedure. However, as we shall see, there are circumstances where it is useful to consider alignment from the other end of the telescope, as a clustering procedure in which a number of sequences are grouped. From this point of view, information is lost as a number of sequences that were distinct from one another are now considered as the same subset. In this context, gap penalties are directly related to this information reduction.

Of course, it is not feasible to write out all possible protein or nucleic acid sequences. Our usual strategy will be to uncover general properties by making use of model systems and simplified alphabets (Solis and Rackovsky, 2000). However, we are also interested in obtaining results of practical interest whenever possible. To do so, we must examine the relationship of model results to the properties of the "real" world of sequences and conformations. This relationship is not a formal part of information theory and will involve additional assumptions or hypotheses, whose information content can be

evaluated, but whose truth must be established by other methods. For example, it is straightforward to evaluate the informational consequences of the proposition that the known sequences are a random subset of all possible sequences. But, information theory, alone, cannot determine its validity.

## Entropy and Information

Given a metric set, $\boldsymbol{M}$, the information content of the set can be measured in bits by its partitioning effect on the ensemble of structures and sequences using Shannon's formulation (Shannon, 1948):

$$
I(M)=-\Sigma\left[p_{k} \log _{2}\left(p_{k}\right)\right]
$$

in which $\boldsymbol{p}_{\boldsymbol{k}}$ is the population of cluster $\boldsymbol{k}$ expressed as a fraction of the ensemble, summed over all clusters. These clusters are subsets of the population of conformers that are indistinguishable under a particular constraint.

The information required to select an individual entity from $\boldsymbol{W}$ is defined as:

$$
\begin{equation*}
I^{S}=\log _{2}(W) \tag{2}
\end{equation*}
$$

where $\boldsymbol{W}$ is the ensemble size. $\boldsymbol{I}^{S}$ is referred to as the "source" information (Shannon, 1948).

## Sequence Alignment

## Overview

Sequences of proteins or nucleic acids of unknown structure and function are sources of information through association with other sequences where the function/structure is already known. The most widely used associative process is "alignment". Alignment algorithms can be divided into two categories, global and local. A global alignment (Needleman and Wunsch, 1970) looks for the best overall similarity

among sequences while a local alignment (Smith and Waterman, 1981) searches for similar subsequences among two proteins. Both of these algorithms make use of a variety of scoring matrices and gap penalties (Altschul, 1998; Apostolico and Giancarlo, 1998; Benner et al., 1993; Koretke et al., 1996; Lesk et al., 1986; Qian and Goldstein, 2002). Sequence alignment problems are underdetermined, having multiple optimal solutions depending on the parameters used. Thus far there has not been a quantitative analysis of the parameter dependence, one reason being the absence of a standard comparison metric. With an information theoretic approach, we are able to formalize the effects of these parameters. We consider the sets of sequences of length $N$, drawn from an alphabet of $A$ characters. Assume that the characters are used with equal frequency. Then each sequence has equal weight and there are $A^{N}$ unique sequences. The information content of the set is simply $N \log _{2} A$. Alignment procedures require the definition of a template of length $T<N$. The template may contain gaps - that is the string for the template may contain one or more positions that match any character. Alternatively, the template may be considered continuous and the sequences with which it is being compared can contain gaps. The critical question for clustering is how many sequences of an exhaustive list match a specific template. Most generally, because there is nothing of special interest for any given template, we are interested in the information content averaged over all templates of a certain type. We begin with the case of gapless pair-wise alignments and then move to multiple-gapped alignments. Our goal is to use both exhaustive and stochastic data sets, along with simple alignment models, to provide insight into the informational issues associated with sequence comparisons.


Statistics from alignments are gathered under two scenarios. 1) For every sequence in the data set, a single (first) occurrence of the template, $T$, is sufficient for cluster assignment during sequence alignment, and 2) All possible occurrences of a template are sought in each sequence of the data set (henceforth referred to as 'complete alignments').

## Gapless Alignments

For an $A$-letter alphabet, the total number of possible $N$-letter sequences is $A^{N}$.

$$
\begin{equation*}
W=A^{N} \tag{3}
\end{equation*}
$$

In the simplest case, we consider those template sequences of length $T$ whose elements are found in contiguous positions in probe sequences of length $N$. Defining $M=N-T$, the templates can be anchored in $M+1$ positions and $A^{M}$ choices will match, leading to an estimate of $(M+1) A^{M}$ sequences if there are no duplicate sequences. Consequently, the information required to distinguish among the ungapped clusters in an exhaustive set is:

$$
\begin{align*}
& W=(M+1) A^{M}  \tag{4}\\
& I^{M}=M \log _{2} A+\log _{2}(M+1)
\end{align*}
$$

This formula counts exactly all occurrences of the template in complete (multiple occurrence) alignments. For a two letter alphabet consisting of 0 's and 1 's the templates are of the form $01,001,0001, \ldots$.

For single occurrence, ungapped alignments, we have an alternative approach using the i) contiguous string, and ii) the standard formula for the probability of failure to match, $p_{F}$, given the probability of occurrence of the template, $p_{T}$, and the number of independent attempts taking $p_{T}=(1 / A)^{T}$ and $p_{F}=\left(1-p_{T}\right)^{M+I}$. The probability of a hit, $p_{H}$, is then $\left(1-p_{F}\right)$ and the number of hits is $A^{N} * p_{H}$. (Equation 5). This formula normally

underestimates when compared to exact counting. However, Equation 5 provides useful values for I for the full range of $M$ for single occurrence of templates (see results).

$$
\begin{equation*}
I^{M}=N \log _{2} A+\log _{2}\left[1-\left(1-p_{T}\right)^{M+l}\right] \tag{5}
\end{equation*}
$$

## Gapped Alignments

For more general gap distributions, where all templates of length $T=N-M$ are aligned against a probe of length $N$, we need to consider the combinatorial arrangement of gaps of varying length. For gaps of minimum length one, there will be

$$
C(N, M)=N!/ M!(N-M)!
$$

ways to arrange the gaps in an $N$-long sequence. However, if we require the minimum gap size, $G_{m i n}$, to be greater than one, then the effective length of the sequence is reduced to $N_{\text {effective }}=\mathrm{N}-M * \mathrm{G}_{\text {min }}+M$. There are $A^{M}$ sequences for each arrangement.

$$
\begin{align*}
& W=C\left(N_{e}, M\right) A^{M}  \tag{6}\\
& I^{M}=M \log _{2} A+\log _{2} C\left(N_{e}, M\right)
\end{align*}
$$

Results for $G_{\text {min }}=1$ are exact for complete alignments (see results).
We have also found an alternative formulation leading to an exact solution of the number of gapped matches for single occurrences of the template. The number of hits to match a given template of length $T$ where $M=N-T$, against an exhaustive set of sequences becomes:

$$
\begin{equation*}
W_{\text {hits }}=\sum_{1}^{M}\left[C(N, M)^{*}(A-1)^{M}\right] \tag{7}
\end{equation*}
$$

This equation (developed empirically from the counting data) provides exact counts over the complete range of $N, T$. When converting to bits of information the right-hand side of Eq. 7, generally can not be reduced to a simpler form; however, when A $>2$ and $\mathrm{T}<95 \%$

of N , the second term sufficiently dominates so that the summation is no longer needed. Under these circumstances, the information change is, to a good approximation:

$$
\begin{equation*}
I^{M}=\log _{2} C(N, M)^{*}(A-1)^{M} \tag{8}
\end{equation*}
$$

## Gap Penalties

The formulas above quantify the amount of information associated with successful alignments when an exhaustive basis set of all possible sequences is available and can be used to set bounds on gap penalty values (see results). Gap penalties have also been derived by directly examining the length distributions of gaps in systems where structural alignment is possible (Qian and Goldstein, 2001). To compare our exhaustive reference state to these values, we can use our counting experiments to determine the distribution of gap lengths for sets of sequences and templates of varying length. We then calculate gap initiation $\left(\gamma_{\mathrm{I}}\right)$ and extension $\left(\gamma_{\mathrm{E}}\right)$ penalties based on the $\log$ of probabilities following Qian and Goldstein (Qian and Goldstein, 2001):

$$
\begin{align*}
& \gamma_{I}=\log _{2}\left[\frac{P_{g}}{1-\exp (-1 / \lambda)}\right]-2 / \lambda  \tag{9}\\
& \gamma_{E}=-1 / \lambda
\end{align*}
$$

Here, $P_{g}$ is the probability of opening a gap, and $\lambda$ is the half-decay of the exponential representing the gap length distribution.

## Search Algorithm Methods

First Occurrence Alignments
For every sequence, $S$, in the set, given a template $T$, we look for the first occurrence of the symbol in position one of the template. Looking forward, we then look for the first instance of the symbol in the second position of the template and so forth, until the last position in the $T$. If a symbol is not observed in $S$ in order of appearance in

$T$ the search is terminated. Indices of each hit in the sequence are tabulated in order to determine the length of the gap among instances of each symbol present in the template.

## Multiple Occurrence Alignments

Figure 1 shows how the occurrences of a template T are sought in a sequence, $S$, consisting of an alphabet of size $A$. The final list contains all the occurrences of $T$ in $S$ by specifying the indices of the symbols in $S$. The positional indices for each occurrence are used to determine the distribution of gap lengths.

FIGURE 1: Search algorithm for finding all occurrences of $T$ is $S$.

> for each symbol, $\alpha$, in the alphabet pos = [position indices of $\alpha$ in the sequence $S]$
> $\mathrm{t}_{0}=\operatorname{pos}_{\mathrm{T}(0)}$ (first symbol of the template, $T$ )
> for $\mathrm{i}=0$ to (length $T-2$ ):
> for each $s$ in $\mathrm{t}_{\mathrm{i}}$
> for each $p$ in $\operatorname{pos}_{\mathrm{T}(\mathrm{i}+1)}$ if $s<p$ add $\{s, p\}$ to $\mathrm{t}_{\mathrm{i}+1}$

## Exhaustive vs. Incomplete Sequence Sets

## Mapping

We turn to the question of how to compare results from the exhaustive list of sequences with those generated from a (sub)set of observed sequences. There are several issues. First, the set of observed sequences is not fixed but is continually updated with new sequences being added and old sequences being modified or even deleted. We are not concerned with these dynamics. For our purposes we can take a "snapshot" of the observed data. A second concern is that the observed sequences show unequal utilization of the characters. Such variable weightings were part of Shannon's initial formulation and Eq. 1 yields a single correction term equal to 0.12 bits/amino acid based on the nonuniform composition of amino acids in real proteins (Strait and Dewey, 1996). Higher

order terms dealing with joint probability of multiple characters can also be considered as "corrections" to the simple assumption of equal frequencies (Cline et al., 2002).

Of more interest, and complexity, is whether there is a simple model that describes the relationship of the observed sequences to the exhaustive set. Many hypotheses can be put forward. We will consider two: 1) the observed sequences are a random subset of the exhaustive sequences and 2) the observed sequences are an "evolutionary" subset of the exhaustive sequences. Our question is how the basic formulas for information content of sequence alignments, Eqs. 4-7, are changed for these two hypotheses.

To examine the nature of alignment information for a random subset of the exhaustive sequences we generated sets of $10,000-100,000$ random sequences of lengths $\mathrm{N}=10,20,30$ for $\mathrm{A}=4$. These were scanned with templates of various lengths and gap lengths. The number of hits was recorded with each of the sequences as a starting point and the probabilities of clustering were calculated. The information for each alignment procedure was tabulated.

We also explored a simple evolutionary model based on the constraint that L of the N positions in the exhaustive set did not vary. The resulting subset of sequences is thus in exact correspondence to sequences from the exhaustive list for $\mathrm{N}^{\prime}$ where $\mathrm{N}^{\prime}=\mathrm{N}$ L. The methods of equations 4-7 can then be applied.

## Correlation of Sequence Alignment and Conformational Resolution

Of course, random and evolutionary models do not exhaust the list of sequence "constraints". Another important set of limits on the database of observed sequences is that many of the sequences of nucleic acids and presumably most of the protein
sequences arise from sequence subsets that provide stable 3-dimensional (tertiary) structures for some range of physical variables. We do not attempt such a model in this paper, but others have approached the problem (Helling et al., 2001; Lau and Dill, 1990).

## Relationship to standard alignment efforts

To conclude this section, we return to the difference between our approach and the standard use of sequence alignment procedures. As noted above, the normal use of alignment is to increase the "information" associated with a given probe sequence, by associating it with one or more sequences of known properties. This information increase corresponds to a shift in the reference state from that we have used (the information lost as sequences are clustered) to the information gained as a single sequence is placed into a cluster smaller than the full set of sequences, from which it was indistinguishable before the alignment procedure. A successful alignment decreases the uncertainty below that of the entire ensemble.

$$
\begin{equation*}
I_{g a i n}=I^{S}-I^{M} \tag{10}
\end{equation*}
$$

This is a measure of the information gained by aligning sequences.

## RESULTS

We have studied two alignment models, treating "ungapped" or gapped templates. We have both analytic formulas and statistical results for the information. In addition, we have collected statistics on gap frequencies and the probability distribution of gap lengths. As noted earlier, Eq. 4 and Eq. 5 describe the ungapped data exactly for multiple hits (Table 1) and within an average of $3 \%$ for single occurrences (Table 2) respectively.

TABLE 1: Gapless alignments for exhaustive sequence sets - Multiple Occurrences as described by Eq. 4.

| Alphabet Size (A) | Template Length | Sequence Length | M | $A^{M}$ | $(\mathrm{M}+1) \mathrm{A}^{M}$ | Number of Hits (Actual Count) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 2 | 3 | 20 | 17 | 131072 | 2359296 | 2359296 |
| 2 | 4 | 20 | 16 | 65536 | 1114112 | 1114112 |
| 2 | 5 | 20 | 15 | 32768 | 524288 | 524288 |
| 2 | 6 | 20 | 14 | 16384 | 245760 | 245760 |
| 2 | 7 | 20 | 13 | 8192 | 114688 | 114688 |
| 2 | 8 | 20 | 12 | 4096 | 53248 | 53248 |
| 2 | 9 | 20 | 11 | 2048 | 24576 | 24576 |
| 2 | 10 | 20 | 10 | 1024 | 11264 | 11264 |
| 2 | 11 | 20 | 9 | 512 | 5120 | 5120 |
| 2 | 12 | 20 | 8 | 256 | 2304 | 2304 |
| 2 | 13 | 20 | 7 | 128 | 1024 | 1024 |
| 2 | 14 | 20 | 6 | 64 | 448 | 448 |
| 2 | 15 | 20 | 5 | 32 | 192 | 192 |
| 2 | 16 | 20 | 4 | 16 | 80 | 80 |
| 2 | 17 | 20 | 3 | 8 | 32 | 32 |
| 2 | 18 | 20 | 2 | 4 | 12 | 12 |
| 2 | 19 | 20 | 1 | 2 | 4 | 4 |
| 2 | 20 | 20 | 0 | 1 | 1 | 1 |
| 3 | 3 | 12 | 9 | 19683 | 196830 | 196830 |
| 3 | 4 | 12 | 8 | 6561 | 59049 | 59049 |
| 3 | 5 | 12 | 7 | 2187 | 17496 | 17496 |
| 3 | 6 | 12 | 6 | 729 | 5103 | 5103 |
| 3 | 7 | 12 | 5 | 243 | 1458 | 1458 |
| 3 | 8 | 12 | 4 | 81 | 405 | 405 |
| 3 | 9 | 12 | 3 | 27 | 108 | 108 |
| 3 | 10 | 12 | 2 | 9 | 27 | 27 |
| 3 | 11 | 12 | 1 | 3 | 6 | 6 |
| 3 | 12 | 12 | 0 | 1 | 1 | 1 |

TABLE 2: Gapless alignments for exhaustive sequence sets - 1st Occurrence as described by Eq. 5 .

| Alphabet <br> Size (A) | Template Length | $\mathrm{P}_{\text {T }}$ | Sequence Length* | M | $\mathrm{P}_{\mathrm{H}}{ }^{\text {g }}$ | Eq. 5: Number of Hits | Number of Hits (actual) | $\Delta l$ | $\begin{gathered} \hline \% \\ \text { Difference } \\ \text { \# Hits } \end{gathered}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 2 | 3 | 0.125 | 20 | 17 | $9.10 \mathrm{E}-01$ | 953790 | 1019920 | 19.96 | 6.48 |
| 2 | 4 | 0.0625 | 20 | 16 | $6.66 \mathrm{E}-01$ | 698541 | 782497 | 19.58 | 10.73 |
| 2 | 5 | 0.03125 | 20 | 15 | $3.98 \mathrm{E}-01$ | 417637 | 458495 | 18.81 | 8.91 |
| 2 | 6 | 0.015625 | 20 | 14 | 2.10E-01 | 220618 | 234280 | 17.84 | 5.83 |
| 2 | 7 | 0.007813 | 20 | 13 | $1.04 \mathrm{E}-01$ | 109042 | 112896 | 16.78 | 3.41 |
| 2 | 8 | 0.003906 | 20 | 12 | $4.96 \mathrm{E}-02$ | 52018 | 53008 | 15.69 | 1.87 |
| 2 | 9 | 0.001953 | 20 | 11 | 2.32E-02 | 24314 | 24552 | 14.58 | 0.97 |
| 2 | 10 | 0.000977 | 20 | 10 | $1.07 \mathrm{E}-02$ | 11209 | 11263 | 13.46 | 0.48 |
| 2 | 11 | 0.000488 | 20 | 9 | 4.87E-03 | 5109 | 5120 | 12.32 | 0.22 |
| 2 | 12 | 0.000244 | 20 | 8 | $2.20 \mathrm{E}-03$ | 2302 | 2304 | 11.17 | 0.10 |
| 2 | 13 | 0.000122 | 20 | 7 | $9.76 \mathrm{E}-04$ | 1024 | 1024 | 10.00 | 0.04 |
| 2 | 14 | $6.1 \mathrm{E}-05$ | 20 | 6 | $4.27 \mathrm{E}-04$ | 448 | 448 | 8.81 | 0.02 |
| 2 | 15 | $3.05 \mathrm{E}-05$ | 20 | 5 | $1.83 \mathrm{E}-04$ | 192 | 192 | 7.58 | 0.01 |
| 2 | 16 | $1.53 \mathrm{E}-05$ | 20 | 4 | $7.63 \mathrm{E}-05$ | 80 | 80 | 6.32 | 0.00 |
| 2 | 17 | 7.63E-06 | 20 | 3 | $3.05 \mathrm{E}-05$ | 32 | 32 | 5.00 | 0.00 |
| 2 | 18 | $3.81 \mathrm{E}-06$ | 20 | 2 | $1.14 \mathrm{E}-05$ | 12 | 12 | 3.58 | 0.00 |
| 2 | 19 | $1.91 \mathrm{E}-06$ | 20 | 1 | $3.81 \mathrm{E}-06$ | 4 | 4 | 2.00 | 0.00 |
| 2 | 20 | $9.54 \mathrm{E}-07$ | 20 | 0 | $9.54 \mathrm{E}-07$ | 1 | 1 | 0.00 | 0.00 |
| 3 | 3 | 0.037037 | 12 | 9 | $3.14 \mathrm{E}-01$ | 167064 | 176957 | 17.43 | 5.59 |
| 3 | 4 | 0.012346 | 12 | 8 | $1.06 \mathrm{E}-01$ | 56215 | 57835 | 15.82 | 2.80 |
| 3 | 5 | 0.004115 | 12 | 7 | $3.25 \mathrm{E}-02$ | 17246 | 17442 | 14.09 | 1.12 |
| 3 | 6 | 0.001372 | 12 | 6 | $9.56 \mathrm{E}-03$ | 5082 | 5102 | 12.32 | 0.39 |
| 3 | 7 | 0.000457 | 12 | 5 | $2.74 \mathrm{E}-03$ | 1456 | 1458 | 10.51 | 0.11 |
| 3 | 8 | 0.000152 | 12 | 4 | $7.62 \mathrm{E}-04$ | 405 | 405 | 8.66 | 0.03 |
| 3 | 9 | $5.08 \mathrm{E}-05$ | 12 | 3 | $2.03 \mathrm{E}-04$ | 108 | 108 | 6.75 | 0.01 |
| 3 | 10 | $1.69 \mathrm{E}-05$ | 12 | 2 | $5.08 \mathrm{E}-05$ | 27 | 27 | 4.75 | 0.00 |
| 3 | 11 | 5.65E-06 | 12 | 1 | 1.13E-05 | 6 | 6 | 2.58 | 0.00 |
| 3 | 12 | $1.88 \mathrm{E}-06$ | 12 | 0 | $1.88 \mathrm{E}-06$ | 1 | 1 | 0.00 | 0.00 |

Tables 3 and 4 show that Eq. 4 and Eq. 5, respectively, provide an exact numerical result for complete alignments.


TABLE 3: Gapped alignments for exhaustive sequence sets - multiple occurrences as described by Eq. 6.

| Alphabet Size(A) | Template Length | Sequence Length* | M | Number of Hits (actual) | $\Delta 1$ | $\mathrm{C}(\mathrm{N}, \mathrm{T})$ | $(A)^{M}$ | $\begin{gathered} \text { Eq. 6: } \\ \mathrm{C}(\mathrm{~N}, \mathrm{~T})(\mathrm{A})^{\mathrm{M}} \\ \hline \end{gathered}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 2 | 3 | 20 | 17 | 149422080 | 27.15 | 1140 | 131072 | 149422080 |
| 2 | 4 | 20 | 16 | 317521920 | 28.24 | 4845 | 65536 | 317521920 |
| 2 | 5 | 20 | 15 | 508035072 | 28.92 | 15504 | 32768 | 508035072 |
| 3 | 3 | 12 | 9 | 4330260 | 22.05 | 220 | 19683 | 4330260 |
| 3 | 4 | 12 | 8 | 3247695 | 21.63 | 495 | 6561 | 3247695 |
| 3 | 5 | 12 | 7 | 1732104 | 20.72 | 792 | 2187 | 1732104 |

One of our primary concerns is the implication of these results for gap penalties. We can get estimates of these penalties by examining the equations directly, or we can calculate the distribution of gap lengths. Equations 4-7 contain terms sensitive to $M$ (the total length of all gaps), as well as terms that depend on the size of the alphabet, $A$, the length of the template, T , and the length of the sequence, $N$. These formulations do not yield cleanly separable gap initiation and gap extension terms. However, they are generally consistent with a gap extension "penalty" that costs information at the rate of $\log _{2} A$ per unit of gap length. The full loss (initiation + extension) at $M=1$ is $\log _{2}(A * N)$. For $\mathrm{N}=100$, such a penalty would be equivalent to -6.0 for $\mathrm{A}=4$ and -7.6 for $\mathrm{A}=20$ in the units normally used for sequence alignment programs (i.e. $\ln A$ ). These values are model-dependent. We also note that, for equivalent coding, the nucleic acid model, $\mathrm{A}=$ 4, would have an N of 300 , yielding a penalty term of -7.1 . These results are in reasonable agreement with the range of empirical gap initiation penalties reported by Qian and Goldstein, see fig. 2 (Qian and Goldstein, 2002).

We can glean additional insight into gap penalties by examining the probability distribution of gap lengths (Qian and Goldstein, 2001). We gathered these data either from short exhaustive binary sequences or from samples of longer sequences with larger alphabets. There are two ways that we can count gap frequencies and gap lengths (see methods). First, for the equations given above, we have used a "first occurrence" model
in which the gap length data are taking from the initial successful match of a template to a sequence.

TABLE 4: Gapped alignments for exhaustive sequence sets - 1st Occurrence model as described by Eq. 7.

| Alphabet Size(A) | Template Length | Sequence Length* | M | $\begin{gathered} \text { Number } \\ \text { of Hits } \\ \text { (actual) } \end{gathered}$ | \# <br> Failures | $\Delta 1$ | $\mathrm{C}(\mathrm{N}, \mathrm{T})(\mathrm{A}-1)^{M}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 2 | 3 | 20 | 17 | 1048365 | 1140 | 20.00 | 1140 |
| 2 | 4 | 20 | 16 | 1047225 | 4845 | 20.00 | 4845 |
| 2 | 5 | 20 | 15 | 1042380 | 15504 | 19.99 | 15504 |
| 2 | 6 | 20 | 14 | 1026876 | 38760 | 19.97 | 38760 |
| 2 | 7 | 20 | 13 | 988116 | 77520 | 19.91 | 77520 |
| 2 | 8 | 20 | 12 | 910596 | 125970 | 19.80 | 125970 |
| 2 | 9 | 20 | 11 | 784626 | 167960 | 19.58 | 167960 |
| 2 | 10 | 20 | 10 | 616666 | 184756 | 19.23 | 184756 |
| 2 | 11 | 20 | 9 | 431910 | 167960 | 18.72 | 167960 |
| 2 | 12 | 20 | 8 | 263950 | 125970 | 18.01 | 125970 |
| 2 | 13 | 20 | 7 | 137980 | 77520 | 17.07 | 77520 |
| 2 | 14 | 20 | 6 | 60460 | 38760 | 15.88 | 38760 |
| 2 | 15 | 20 | 5 | 21700 | 15504 | 14.41 | 15504 |
| 2 | 16 | 20 | 4 | 6196 | 4845 | 12.60 | 4845 |
| 2 | 17 | 20 | 3 | 1351 | 1140 | 10.40 | 1140 |
| 2 | 18 | 20 | 2 | 211 | 190 | 7.72 | 190 |
| 2 | 19 | 20 | 1 | 21 | 20 | 4.39 | 20 |
| 3 | 3 | 12 | 9 | 435185 | 112640 | 16.78 | 112640 |
| 3 | 4 | 12 | 8 | 322545 | 126720 | 16.95 | 126720 |
| 3 | 5 | 12 | 7 | 195825 | 101376 | 16.63 | 101376 |
| 3 | 6 | 12 | 6 | 94449 | 59136 | 15.85 | 59136 |
| 3 | 7 | 12 | 5 | 35313 | 25344 | 14.63 | 25344 |
| 3 | 8 | 12 | 4 | 9969 | 7920 | 12.95 | 7920 |
| 3 | 9 | 12 | 3 | 2049 | 1760 | 10.78 | 1760 |
| 3 | 10 | 12 | 2 | 289 | 264 | 8.04 | 264 |
| 3 | 11 | 12 | 1 | 25 | 24 | 4.58 | 24 |

FIGURE 2: Distribution of gap initiation and extension penalties. Medium hashed bars designate gap initiation. Solid bars: gap extension. Dense hash bars: gap initiation + gap extension


Alternatively, we can identify all matches of a template with a specific sequence, and for each match, tabulate the gap length information. We call this model a "multiple occurrence" model. It might be closer to the empirical data reported by Qian and Goldstein (Qian and Goldstein, 2001).

Our basic findings are: 1) Most of the gap length distributions can be approximated by an exponential, but those arising from larger alphabets and longer templates clearly have more complex character. The distributions can be numerically fit as multiple exponentials similar to those found by Qian and Goldstein for sequence alignments of proteins of known structures. They can also be fit with polynomial functions. It is not obvious whether these expanded functions carry any physical significance. 2) For the first occurrence model, the exponential decay increases strongly with alphabet size and slowly with sequence length (Table 5).

However, for the "multiple-occurrence" model, the exponential decay is independent of alphabet size, although it still increases with N and decreases with T (Table 6). 3) If we use the single exponential approximation and the treatment of Qian and Goldstein (see Eq. 9 above), we get the range of gap penalties shown in figure 2. Our values are consistently on the low end of the empirical range.

To make the comparison with the values in the literature (Qian and Goldstein, 2002) more useful, we must consider two additional issues: first, how does the set of observed sequences relate to the exhaustive reference state we have been using. Second,

TABLE 5: Gap distributions and gap penalties for the $1^{\text {st }}$ occurrence model

| N | Alphabet <br> Size | Template <br> Length | Total \# <br> Gaps | Total \# <br> hits | Pg | $\lambda$ | $(-) \gamma_{\text {gap-I }}$ | $(-) \gamma_{\text {gap-E }}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 10 | 2 | 2 | 1004 | 1013 | 1.043 | 1.425 | 0.676 | 0.702 |
| 10 | 2 | 3 | 1398 | 968 | 1.219 | 1.352 | 0.632 | 0.739 |
| 10 | 2 | 4 | 1528 | 848 | 1.695 | 1.209 | 0.552 | 0.827 |
|  |  |  |  |  |  |  |  |  |
| 20 | 2 | 2 | 1048536 | 1048555 | 1.000 | 1.443 | 0.693 | 0.693 |
| 20 | 2 | 3 | 1572291 | 1048365 | 1.001 | 1.442 | 0.693 | 0.693 |
| 20 | 2 | 4 | 2092512 | 1047225 | 1.004 | 1.441 | 0.692 | 0.694 |
|  |  |  |  |  |  |  |  |  |
| 50 | 2 | 2 | 1009981 | 1000000 | 0.997 | 1.444 | 0.694 | 0.692 |
| 50 | 2 | 3 | 1514903 | 1000000 | 0.996 | 1.445 | 0.694 | 0.692 |
| 50 | 2 | 4 | 2020116 | 1000000 | 0.999 | 1.443 | 0.694 | 0.693 |
| 50 | 2 | 5 | 2524154 | 100000 | 0.997 | 1.444 | 0.694 | 0.692 |
|  |  |  |  |  |  |  |  |  |
| 100 | 2 | 2 | 999853 | 1000000 | 1.000 | 1.442 | 0.693 | 0.693 |
| 100 | 2 | 3 | 1501252 | 100000 | 1.000 | 1.443 | 0.693 | 0.693 |
| 100 | 2 | 4 | 2000332 | 100000 | 0.999 | 1.443 | 0.693 | 0.693 |
| 100 | 2 | 5 | 2501670 | 1000000 | 0.998 | 1.444 | 0.693 | 0.693 |
|  |  |  |  |  |  |  |  |  |
| 20 | 4 | 4 | 28163173 | 774544 | 0.170 | 2.928 | 1.215 | 0.341 |
| 20 | 4 | 5 | 54233834 | 585323 | 0.228 | 2.589 | 1.112 | 0.386 |
| 20 | 4 | 6 | $1.08 \mathrm{E}+08$ | 383398 | 0.316 | 2.264 | 1.004 | 0.442 |
| 20 | 4 | 7 | $2.16 \mathrm{E}+08$ | 214460 | 0.447 | 1.972 | 0.897 | 0.507 |
|  |  |  |  |  |  |  |  |  |
| 50 | 4 | 4 | 2998161 | 999517 | 0.111 | 3.477 | 1.386 | 0.288 |
| 50 | 4 | 5 | 3740043 | 997900 | 0.112 | 3.471 | 1.384 | 0.288 |
| 50 | 4 | 6 | 4461340 | 992992 | 0.112 | 3.460 | 1.381 | 0.289 |
| 50 | 4 | 7 | 5126973 | 980293 | 0.114 | 3.439 | 1.374 | 0.291 |
|  |  |  |  |  |  |  |  |  |
| 100 | 4 | 4 | 2999727 | 1000000 | 0.111 | 3.472 | 1.385 | 0.288 |
| 100 | 4 | 5 | 3750300 | 1000000 | 0.111 | 3.473 | 1.386 | 0.288 |
| 100 | 4 | 6 | 4500088 | 999998 | 0.111 | 3.472 | 1.386 | 0.288 |
| 100 | 4 | 7 | 5248490 | 999998 | 0.111 | 3.474 | 1.386 | 0.288 |
| 100 | 20 | 2 | 1826955 | 963111 | 0.003 | 19.056 | 2.958 | 0.052 |
| 100 | 20 | 3 | 2500674 | 881617 | 0.003 | 17.958 | 2.894 | 0.056 |
| 100 | 20 | 4 | 2792228 | 741973 | 0.004 | 16.435 | 2.804 | 0.061 |
| 100 | 20 | 5 | 2637250 | 564189 | 0.005 | 14.822 | 2.703 | 0.067 |
| 100 | 20 | 6 | 2133962 | 383374 | 0.006 | 13.287 | 2.597 | 0.075 |
|  |  |  |  |  |  |  |  |  |

TABLE 6: Gap distributions and gap penalties for the multiple occurrence model

| N | $\begin{gathered} \text { Alphabet } \\ \text { Size } \\ \hline \end{gathered}$ | Template Length | $\begin{aligned} & \text { Total \# } \\ & \text { Gaps } \end{aligned}$ | $\begin{aligned} & \text { Total \# } \\ & \text { Hits } \end{aligned}$ | Pg | $\lambda$ | $(-) \gamma_{\text {gap-I }}$ | $(-) \gamma_{\text {gap }-\mathrm{E}}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 20 | 2 | 3 | $2.99 \mathrm{E}+08$ | $1.49 \mathrm{E}+08$ | 0.031 | 5.974 | 1.944 | 0.167 |
| 20 | 2 | 4 | $9.53 \mathrm{E}+08$ | $3.18 \mathrm{E}+08$ | 0.058 | 4.298 | 1.737 | 0.233 |
| 20 | 2 | 5 | $2.03 \mathrm{E}+09$ | $5.08 \mathrm{E}+08$ | 0.096 | 3.330 | 1.592 | 0.300 |
| 20 | 3 | 3 | 84355136 | 42177568 | 0.028 | 6.301 | 1.990 | 0.159 |
| 20 | 3 | 4 | $1.79 \mathrm{E}+08$ | 59655795 | 0.052 | 4.542 | 1.783 | 0.220 |
| 20 | 3 | 5 | $2.54 \mathrm{E}+08$ | 63554348 | 0.085 | 3.528 | 1.637 | 0.283 |
| 20 | 3 | 7 | 2.12E +08 | 35334238 | 0.217 | 2.222 | 1.414 | 0.450 |
| 20 | 4 | 3 | 35598996 | 17799498 | 0.031 | 5.976 | 1.945 | 0.167 |
| 20 | 4 | 4 | 56714217 | 18904739 | 0.052 | 4.545 | 1.784 | 0.220 |
| 20 | 4 | 5 | 60481940 | 15120485 | 0.085 | 3.529 | 1.637 | 0.283 |
| 20 | 4 | 7 | 28234680 | 4705780 | 0.187 | 2.367 | 1.456 | 0.423 |
| 50 | 2 | 3 | $4.95 \mathrm{E}+09$ | $2.47 E+09$ | 0.004 | 15.899 | 2.774 | 0.063 |
| 50 | 2 | 4 | $4.36 \mathrm{E}+10$ | $1.45 \mathrm{E}+10$ | 0.008 | 11.713 | 2.529 | 0.085 |
| 50 | 2 | 5 | $2.65 \mathrm{E}+11$ | $6.62 \mathrm{E}+10$ | 0.012 | 9.488 | 2.362 | 0.105 |
| 50 | 3 | 3 | $1.45 \mathrm{E}+09$ | $7.26 E+08$ | 0.004 | 15.902 | 2.774 | 0.063 |
| 50 | 3 | 4 | $8.54 \mathrm{E}+09$ | $2.85 \mathrm{E}+09$ | 0.008 | 11.717 | 2.529 | 0.085 |
| 50 | 3 | 5 | $3.49 \mathrm{E}+10$ | $8.72 \mathrm{E}+09$ | 0.012 | 9.292 | 2.342 | 0.108 |
| 50 | 3 | 7 | $3.2 \mathrm{E}+11$ | $4.57 \mathrm{E}+10$ | 0.023 | 6.661 | 2.087 | 0.150 |
| 50 | 4 | 3 | $6.12 \mathrm{E}+08$ | $3.06 E+08$ | 0.004 | 16.223 | 2.794 | 0.062 |
| 50 | 4 | 4 | 2.7E+09 | $8.99 E+08$ | 0.008 | 11.715 | 2.529 | 0.085 |
| 50 | 4 | 5 | $8.26 \mathrm{E}+09$ | $2.07 \mathrm{E}+09$ | 0.012 | 9.291 | 2.342 | 0.108 |
| 50 | 4 | 7 | $3.65 \mathrm{E}+10$ | $6.08 \mathrm{E}+09$ | 0.023 | 6.659 | 2.087 | 0.150 |

how do the results depend on the type of gap model chosen? We can imagine that the experimental sequences map onto the exhaustive sequences via random selection or via evolutionary selection. A random model is easily constructed. Within the expected statistical variation, the numerically-derived probabilities are equal to those from the formulas. Evolutionary models for the observed sequences can also be constructed. Two such models would be: 1) use of a full alphabet for a subset of the sequence positions with the other positions fixed; 2) restricted alphabets at all sequence positions. In the former case, we would expect the equations to apply directly, but with a reduced chain length. In the second example, we can approximate the effects by using a reduced alphabet. We obtained numerical data on the fixed position model, using fully variable 15 mers embedded in 20 mers with the first 5 positions invariant. The (exhaustive) results (figure 3) for first occurrence gapped hits closely correspond to the exhaustive 15 mer data, suggesting that Equations 6 and 7 are good approximations for sequences generated by evolutionary relationships. However, when we compute the full occurrence gap length distributions for the same data set, the situation is more complicated. The distribution functions are intermediate between the 15 mer and 20 mer data, with the distributions closer to the 20 mer results (Figure 4).

FIGURE 3: Average number of gapped hits for all possible 5 mer templates in 32 related 20 mer evolutionary subsets using the single occurrence model. Average hits for the 15 mer and 20 mer exhaustive sets are shown on the axis ends.


FIGURE 4: Gap length distributions in multiple occurrence alignments for 32 , 5 mer binary templates in a 20 mer evolutionary subset. Colored lines indicate gap length distribution for each template in the 20 mer subset. For comparison we also show the behavior of 5 mer templates in $20 \mathrm{mer}(\nabla)$ and $15 \mathrm{mer}(+)$ exhaustive sets.


The other question raised above is what type of gap simulation best captures "normal" alignment procedures, as, for example, in the Needleman-Wuntsch algorithm. The essential issue is that the "real world" data are drawn from a highly heterogeneous sequence set. The sequences and templates are of variable lengths, and the alphabets, while nominally of 4 or 20 letters, have unequal utilization of the letters in a sequence/structure dependent manner. Furthermore, the results depend on whether "first occurrence" or "multiple occurrence" statistics are used. While our "penalties" cannot be used directly for any empirical data set, two practical suggestions are that gap penalties should differ for nucleic acid vs. amino acid sequence alignments, and that it would be useful to generate sequence-length dependent penalty corrections.

## DISCUSSION

Most scoring matrices are heavily parameterized. It is common practice to develop reliable test sets of either sequence or structure alignments in order to derive empirical statistics for parameterization. Although many of these functions work well when applied to known sequences and structures they often fall short when dealing with novel targets. The protein databank (PDB) is the largest available test set, however it has not yet been possible to verify any of the empirical methods developed to date using the PDB in its entirety. And even if it were feasible to test the entire PDB, we do not understand the scope of the PDB as subset of the entire protein universe. Our approach in this paper has been to use simple, exhaustive models whenever possible. This approach sheds light on current practices and provides guidelines for the development of future methodologies by benchmarking the information content of the available data.

Score matrices are valued for their discriminating power when faced with choices among similar sequences. Their resolution is determined by both sensitivity and selectivity thresholds. For sequence alignments, the choices of gap penalties as well as the scoring matrix are critical components. The root question is whether there is a "best" set of gap penalties. As mentioned above, the empirical gap penalties currently in use are obtained from training sets on homologous proteins. By analyzing exhaustive sequence sets, we have derived formulas for sequence clustering under a variety of gap types. These formulas suggest length and alphabet size dependencies that are not directly included in current methods. We also have examined gap occurrence probabilities, finding that they (approximately) follow a geometric distribution. As noted in the results section, the differing approaches to calculating gap penalties yield a range of answers. Most of our results are at the low end of reported range of gap penalties. One reasonable explanation is that the set of known sequences is heavily weighted to higher than random similarity because of some combination of evolutionary and structural constraints (e.g. reduced gap probabilities inside secondary structure elements). It is an open question whether our values will lead to better alignments in a biased data set. However, it is likely that including specific length and alphabet size dependencies in empirical studies would be helpful.

Although our exhaustive results are limited ( $\mathrm{N}=20 ; \mathrm{A}=2$ ), we observe the same general behavior and trends for numerically sampled sequences using four-letter as well as 20 -letter alphabets. The move from simple to more complete models will shift the reference states but not the general trend. Such quantitative assessments are critical for improving the effectiveness of scoring functions used in various alignment protocols.

## ACKNOWLEDGMENTS

We are grateful for helpful discussions with Scott Pegg \& Kevin Masukawa. This work was supported by a grant from the National Science Foundation (NSF CHE-0118481); R.

Kip Guy, Principal Investigator and a University of California president's dissertation year fellowship (to T. Aynechi).

## REFERENCES

Altschul, S.F. 1998. Generalized affine gap costs for protein sequence alignment. Proteins 32(1):88-96.

Apostolico, A., and R. Giancarlo. 1998. Sequence alignment in molecular biology. $J$ Comput Biol 5(2):173-96.

Aynechi, T., and I.D. Kuntz. 2005. An information theoretic approach to molecular modeling: II. Force fields. In preparation.

Benner, S.A., M.A. Cohen, and G.H. Gonnet. 1993. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J Mol Biol 229(4):1065-82.

Berger, B., J. Kleinberg, and T. Leighton. 1999. Reconstructing a three-dimensional model with arbitrary errors. Journal of the ACM 46(2):212-235.

Bonneau, R., and D. Baker. 2001. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct 30:173-89.

Brunger, A.T. 1992. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature Vol 355(6359):472-475.

Carothers, J.M., S.C. Oestreich, J.H. Davis, and J.W. Szostak. 2004. Informational complexity and functional activity of RNA structures. J Am Chem Soc 126(16):5130-7.

Cline, M.S., K. Karplus, R.H. Lathrop, T.F. Smith, R.G. Rogers, Jr., and D. Haussler. 2002. Information-theoretic dissection of pairwise contact potentials. Proteins 49(1):7-14.

Helling, R., H. Li, R. Melin, J. Miller, N. Wingreen, C. Zeng, and C. Tang. 2001. The designability of protein structures. J Mol Graph Model 19(1):157-67.

Koretke, K.K., Z. Luthey-Schulten, and P.G. Wolynes. 1996. Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. Protein Sci 5(6):1043-59.

Kuntz, I.D., and D.A. Agard. 2003. Assessment of the role of computations in structural biology. Adv Protein Chem 66:1-25.

Lau, K.F., and K.A. Dill. 1990. Theory for protein mutability and biogenesis. Proc Natl Acad Sci U S A 87(2):638-42.

Lesk, A.M., M. Levitt, and C. Chothia. 1986. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. Protein Eng 1(1):77-8.

Levitt, M., and M. Gerstein. 1998. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci U S A 95(11):5913-20.

Luzzati, V. 1952. Traitement Statistique des Erreurs dans la Determination des Structures Cristallines. Acta Cryst 5(6):802-810.

Needleman, S.B., and C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3):443-453.

Park, B., and M. Levitt. 1996. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 258(2):367-92.

Qian, B., and R.A. Goldstein. 2001. Distribution of Indel lengths. Proteins 45(1):102-4.
Qian, B., and R.A. Goldstein. 2002. Optimization of a new score function for the generation of accurate alignments. Proteins 48(4):605-10.

Shannon, C.E. 1948. A mathematical theory of communication. The Bell System Technical Journal(27):379-423 and 623-656.

Smith, T.F., and M.S. Waterman. 1981. Identification of common molecular subsequences. J Mol Biol 147(1):195-7.

Solis, A.D., and S. Rackovsky. 2000. Optimized representations and maximal information in proteins. Proteins 38(2):149-64.

Strait, B., and T. Dewey. 1996. The Shannon information entropy of protein sequences. Biophys. J. 71(1):148-155.

Stroud, R.M., and E.B. Fauman. 1995. Significance of structural changes in proteins: expected errors in refined protein structures. Protein Sci 4(11):2392-404.

Sullivan, D.C., T. Aynechi, V.A. Voelz, and I.D. Kuntz. 2003. Information Content of Molecular Structures. Biophys. J. 85(1):174-190.

Vingron, M., and M.S. Waterman. 1994. Sequence alignment and penalty choice.
Review of concepts, case studies and implications. J Mol Biol 235(1):1-12.
Vingron, M., and M.S. Waterman. 1994. Sequence alignment and penalty choice.
Review of concepts, case studies and implications. J Mol Biol 235(1):1-12.

# Chapter Five: An Information Theoretic Approach to Protein Modeling: II. Force Fields 

Tiba Aynechi ${ }^{\ddagger}$, Irwin D. Kuntz ${ }^{*}$<br>*Department of Pharmaceutical Chemistry<br>*Graduate Group in Biophysics<br>University of California, San Francisco<br>San Francisco, California, 94143-0446

Correspondence should be addressed to Dr. Irwin D. Kuntz, Department of
Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, California, 94143-0446

Phone: (415) 476-1937
Fax : (415) 502-1411
Email: kuntz@cgl.ucsf.edu

Submitted for publication to Biophysical Journal

## KEYWORDS

Information Theory, Shannon Entropy, Lattice models, Force Field, Electrostatic Interactions, HP Interactions.

## ABSTRACT

In this paper, we explore the information content of molecular force field calculations. We make use of exhaustive lattice models of molecular conformations and reduced alphabet sequences to determine the relative resolving power of pair-wise interactionbased force fields. We find that sequence-specific interactions that operate over longer distances offer greater amounts of information than nearest neighbor or non-sequence specific interactions. In a companion paper we explored the information content of sequence alignments and gap penalties. We find that the general trends in both papers can be extended to real proteins and nucleic acids.
par parsing inter

## INTRODUCTION

Predicting the three-dimensional structure of macromolecules from their sequences remains a fundamental problem of great interest in biology and a most challenging theoretical puzzle; and to date, most attention has been focused on proteins. Current protein folding hypotheses assume that the native structure is the conformational set of lowest free energy when compared to all other accessible conformational ensembles. To test this assumption, the free energy of all conformations must be calculated, an intractable undertaking at present. A first step to solve this problem is to develop a "force field", typically based on atomic interactions, which can, in principle, be evaluated, for all conformers. Force fields are developed from experimental geometries and energies of small molecules, often with the help of high-level quantum mechanical calculations (Kollman, 1993). The accuracy and validity of force fields is tested by comparison against experiments. But only in the simplest cases of very small systems can conformational space be explored exhaustively (Kuntz and Agard, 2003). Our aim here is to present a framework for understanding the discriminatory power of various force fields using information theory. We emphasize, however, information theory, alone, cannot be used to verify the physical correctness of force field terms.

In a companion paper (Aynechi and Kuntz, 2005) we use information theory to quantify the information gained during sequence alignment procedures and show how to extract gap penalty values based on actual gap distributions rather than empirical optimization. In this work, we wish to extend our methods towards the understanding of force fields.

Determining the lowest energy conformation from among all possible threedimensional structures of a given sequence requires an energy function capable of discriminating among native and decoy conformations. Many force fields and scoring functions have been developed for this purpose (Bahar and Jernigan, 1997; Halgren, 1995; Jones et al., 1992; Park and Levitt, 1996; Sippl, 1995; Tenette et al., 1996). Some are physics-based, concentrating on pair-wise or higher order atomic interactions while others, such as potentials of mean force, incorporate properties of the ensemble. Alternatively, one can define an empirical score function using training data and then apply these functions to sequences/structures to test their efficacy.

However, there is no agreed measure for evaluating the above methods nor is there consensus on what these methods should accomplish. It has been shown that most statistically derived potentials are inadequate in the representation of real world proteins (Park et al., 1997; Thomas and Dill, 1996). In addition, because the structures in the Protein Databank (Berman, 2000) represent only a fraction of the protein conformational space, the question of whether parameter-based methods, or those relying on potentials of mean force, will fail when a new structure lies outside the training data remains unanswered. Physics-based methods avoid this particular extrapolation problem, but, as noted above, can fall short due to the computational costs of conformational sampling and the complexity of comprehensive atomic interaction models (Feldman and Hogue, 2002; Sullivan and Kuntz, 2004). The utility of simplified models depends on the level of resolution required (Huber and Torda, 1998). Thus efficient use and development of future and current methodologies require an in-depth understanding of their behavior and dependencies. As previously demonstrated (Aynechi and Kuntz, 2005; Sullivan et al.,
2003), exhaustive 2D lattice models (Chan and Dill, 1989) and information theory (Young, 1971) allow us to move from an empirical regime towards an analytical formulation. We can then quantitatively measure the discriminating power of various scoring functions and force fields, develop metrics for performance analysis and draw inferences that cross over to real proteins.

In this paper we will examine various force fields for their ability to distinguish among conformations given a set of all possible conformations and all possible sequences.

## METHODS

## Overview

Our basic approach is to write out all possible occurrences of the set of interest (conformations, sequences, etc.) and then ask what the informational consequence of performing an operation that combines or clusters some of the objects. In particular, we will score a set of conformations using various force fields and cluster the conformations based on the degeneracy of their scores (energies). The information content of the original set of $W$ objects is $\log _{2} W$. An "object" in this context will be a specific conformer with a specific sequence. Any clustering will reduce the effective number of objects and hence reduce the information content of the transformed set. Of course, it is not feasible to write out all possible protein or nucleic acid sequences or all possible macromolecular conformations. Instead, we will make use of walks on two-dimensional lattices and simplified alphabets (Chan and Dill, 1989; Dill et al., 1995; Solis and Rackovsky, 2002). However, we are also interested in obtaining results of practical interest whenever possible. Also, we must add the caution that information theory is
analogous to statistical mechanics: it provides formal consequences of initial assumptions, but the truth of those assumptions has to be established from additional context. For example, the information content of an electrostatic theory in which like charges attract is indistinguishable from the information content of a theory in which like charges rebel, but only one of these models describes our experiential universe.

## Entropy and Information

As before (Aynechi and Kuntz, 2005) the information content of an ensemble recovered by a constraint set $\boldsymbol{M}$ in bits is defined to be the Shannon entropy (Shannon, 1948):

$$
\begin{equation*}
I(M)=-\Sigma\left[p_{k} \log _{2}\left(p_{k}\right)\right] \tag{1}
\end{equation*}
$$

in which $\boldsymbol{p}_{\boldsymbol{k}}$ is the population of cluster $\boldsymbol{k}$ expressed as a fraction of the ensemble, summed over all clusters. The population clusters become indistinguishable when the constraints are imposed on the ensemble $W$.

The theoretical information content of the ensemble of size, $\boldsymbol{W}$, is defined as:

$$
\begin{equation*}
I^{S}=\log _{2}(W) \tag{2}
\end{equation*}
$$

$\boldsymbol{I}^{\boldsymbol{S}}$ is referred to as the "source" information (Shannon, 1948).

## Protein Models

We will employ the two-dimensional lattice models of Chan and Dill (Chan and Dill, 1989) and focus our discussion on protein-like conformation space. We will extend our results to nucleic acids only at the level of the alphabet size and will not consider nucleic acid conformers, per se. We will use the Cartesian (through-space) distance, $\boldsymbol{d}$, between beads $\boldsymbol{i}, \boldsymbol{j}$ is defined as:

$$
\begin{equation*}
d_{i, j}=\left(\left(x_{i}-x_{j}\right)^{2}+\left(y_{i}-y_{j}\right)^{2}\right)^{1 / 2} \tag{3}
\end{equation*}
$$

Chains of beads, each bead representing one "residue", are arranged in selfavoiding walks according to the following rules. The elementary step, the distance between consecutive beads, $d_{i, i+l}$, is fixed at unit length. The move set is limited to a single step with diagonal moves disallowed. Beads cannot overlap (the excluded volume constraint). This set of walks is the same as the exhaustive ensembles of Chan and Dill (Chan and Dill, 1989) that count all conformations not related by translation, rigid rotation, or reflection. The N -terminus to C -terminus directionality of proteins is preserved in these ensembles.

We will explore two types of conformational sets: exhaustive, which contains all conformers allowed by the rules (Table 1), above, and compact, the set in which all vertices of an $i \times j=N$ two-dimensional lattice must be occupied (Table 2). All conformations of up to length 26 have been enumerated. We also generate semi-compact structures by fitting the $N$ mer to the next smallest perfect-square lattice (Table 2). Compact lattices were obtained in an efficient manner by modifying the generation program to terminate whenever the $i$ or $j$ limits were exceeded.

TABLE 1: HP sequences and self avoiding 2-dimensional walks on a square lattice

| Chain Length ( N ) | \# HP sequences generated | \# Fully Enumerated Structures | \# Structures Generated | Total pairwise (W) | $\left.\right\|^{\text {Source }}$ (bits) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 4 | 16 | 5 | 5 | 80 | 6.322 |
| 5 | 32 | 13 | 13 | 416 | 8.700 |
| 6 | 64 | 36 | 36 | 2,304 | 11.170 |
| 7 | 128 | 98 | 98 | 12,544 | 13.615 |
| 8 | 256 | 272 | 272 | 69,632 | 16.087 |
| 10 | 1,024 | 2,034 | 2,034 | 2,082,816 | 20.990 |
| 12 | 4,096 | 15,037 | 15,037 | 61,591,552 | 25.876 |
| 16 | 10,000* | 802,075 | 10,000* | 10,000 ${ }^{\text {a }}$ | 13.288 |
| 25 | 10,000** | 5,768,299,665 | 10,000* | 10,000 ${ }^{\text {z }}$ | 13.288 |

TABLE 2: Compact and semi-compact hamilton walks

| Chain Length (N) | Lattice <br> dimensions | \# HP sequences | \# Compact <br> Structure | Total pairwise <br> $(W)$ | $1^{\text {Source }}$ (bits) |
| :---: | :---: | ---: | ---: | ---: | ---: |

## Simplified Potentials

## HP interaction model

The sequence space of proteins grows exponentially as $20^{N}$ if the natural amino acids are used. In order to exhaustively explore sequence space, the HP model (Lau and Dill, 1989) of residues is used to generate all possible sequences of length $N$ (Table 1). The residues are typed as H (nonpolar) and P (polar). For every sequence, all possible conformations are generated on the two-dimensional lattice and the interaction energy is calculated for each structure according to some interaction rule set. For example, residues can be said to interact if their geometric distance, $\boldsymbol{d}_{i j}$, within the lattice is one unit length or less, and they do not occupy adjacent positions in the chain. The energy of a structure is lowered by an amount $\varepsilon$ if the interacting residues are both of the type $\boldsymbol{H}$. In the original Dill formulation, $\boldsymbol{H P}$ and $\boldsymbol{P P}$ interactions do not lower the energy of the conformations. The interaction energy, $\boldsymbol{e}_{H H}$, is:

$$
e_{H H}=\sum_{1}^{k} \varepsilon ; k \text { is \# of interacting HH pairs }
$$

The partitioning power of the interaction energy will be determined in bits using Shannon's entropy. In our implementations we assume that all energy differences are resolvable, In effect, we are measuring the diversity of the energy landscape, that is the distribution of conformer-sequence pairs that have specific energies. We do not get
formal entropies from this characterization because we are not examining the conformer distribution at specific temperatures. Nor do we have to be concerned about the interconvertibility of conformers and sequences. Extension to other interaction rules and larger alphabets is straightforward.

## Solvation Model

Interactions among solvent and protein can be difficult to model with any realism.
We employ a simple procedure that again uses all possible conformations of Nmer chains on the two-dimensional lattice (Table 1). Residues are classified as buried, exposed, or partially buried. On a two-dimensional lattice, each vertex has a coordination number $z=$ 4. A buried residue is one whose three surrounding vertices (other than its immediate predecessor in the chain) in the lattice are occupied by other residues. An exposed residue is surrounded by three unoccupied vertices and a partially buried residue has one or two of the vertices filled. The same sequence/structure pairs used above are used here as well. However, in addition to the $\boldsymbol{H} \boldsymbol{H}$ contact score a solvation score is also added to each residue's energy contribution. For every filled coordination site a contribution $\xi$ is added. A fully buried residue would incur an additional term equaling $3 \boldsymbol{\xi}$, while an exposed residue would have no added term. Again, we will assume that each energy level is fully resolvable, and we count the number of objects in each energy level.

## Electrostatic Interactions

Electrostatic interactions within molecules play an important role in defining their properties (Honig and Nicholls, 1995). Pair-wise attractive and repulsive interactions between atoms are governed by Coulomb's law whose energy, $\boldsymbol{e}$, depends on charge and Cartesian distance, $e \sim q_{i} q_{j} / d_{i j}$. In our simplified model, we allow residues to have unit
positive ( + ) or negative ( - ) charges. All possible + - sequences of length $N$ are subsequently generated and for each one the coulomb energy, $E_{c}$, of all possible two dimensional conformations is calculated. We assume the residues to be spherical point charges laid on a two dimensional lattice yielding the familiar $1 / \mathrm{r}$ dependence on separation. Alternative two dimensional representations use disc charges, yielding $\ln d$ dependence (Dill, 2004; Dill and Bromberg, 2003). Thus we will take the coulomb energy of each conformation as:

$$
\begin{align*}
& e_{i j}=\frac{q_{i} q_{j}}{d_{i j}} \\
& E=\sum_{i=1}^{N} \sum_{j>i}^{N} e_{i j} \tag{5}
\end{align*}
$$

We again use Shannon's equations for entropy to quantify the partitioning power of electrostatic interactions. For a more realistic model one could define sequences where the sum of $+/$ - residues is a percentage of the total residues.

## GO Type Potentials

Go-type potentials were proposed by Go and Taketomi (Taketomi et al., 1975) (Go and Taketomi, 1978) to elucidate the effects of long and short range interactions during protein folding. In their work, interactions that involve bond lengths and angles are called "short range" and often associated with secondary structure formation, while long range interactions are defined as residues nearest neighbors in space. Protein structures were stabilized by specific long and short range interactions of varying weights during Monte Carlo simulations. It was concluded that native state stability is achieved through long range interactions while folding rates were affected by the short range interactions. We investigate whether knowledge of long range interactions, as defined by the Go potential,
alone is sufficient for discriminating among various conformations when averaged over the ensemble of structures. All possible $W$ compact and semi-compact conformations of $N$ mer chains were generated (Table 2). We suppose residues $i$ and $j$ to be interacting if $j>$ $i+1$ in the chain and they occupy nearest neighbor vertices on the lattice. The energy of the interacting units is assumed to be identical with a value of $-\varepsilon$. We define the potential for every structure in the ensemble as the number of interacting pairs in units of $\varepsilon$. The energies of all other conformations are evaluated based on the potential for a target structure. Conformations are scored as follows: Given a target structure with a set of interacting pairs, $S_{T}$, the energy of the conformation is lowered by $\varepsilon$ for every interacting pair present in both the target structure and the conformer being evaluated. Thus the total score of any structure is:

$$
\begin{equation*}
E=\sum_{k} \varepsilon_{k}, k=\text { total \# of shared pairs } \tag{6}
\end{equation*}
$$

All conformations are scored as above, averaged over $W$ target structures. The resolving power is determined by the partitioning effect of the resulting scores on the conformer set.

## RESULTS

Folded states of real proteins are characterized by their free energy differences. We examine the energy distributions of simplified force fields applied to sets of twodimensional conformations threaded onto sets of sequences. Using Shannon's entropy for the energy distributions, we measure each force field's classifying power. To do so, exhaustive and stochastic sets of fully enumerated and compact conformers are generated as described in the methods section. Our models assume that all distances and sequences
are known exactly and are free from errors. In general, the most informative force fields are those whose energy functions produce the least degenerate set of values for a given set of conformer/sequence pairs. Our results indicate that force fields with terms that include long-range inter-atomic distances yield much more information than force fields which just make use of pair-wise contact potentials.

## HP Interactions

Self-avoiding 2D conformations of $N$ bead chains were enumerated (Table 1) and threaded with all possible HP sequences (see methods). Each conformer/sequence pair was subsequently scored according to Eq. 4 and the Shannon entropy was calculated using Eq. Error! Reference source not found.. Figure 1 shows that there is a steady increase in the amount information retrieved $\left(I^{M}\right)$ as $N$ becomes larger.

FIGURE 1: Information of HP and solvent contacts. - HP contacts, $\boldsymbol{\Delta}$ : Solvent with weigh scale $0.5, \mathbf{v}$ : Solvent with weigh scale $1.0, \star$ : Solvent with weight scale 0.2 . Inset: $\circ$ : $I^{\text {S. Filled symbols: exhaustive }}$ set, open symbols stochastic sample.


The information increase for the fully enumerated sets is approximated by:

$$
\begin{equation*}
I\left(H P_{N}\right)=0.61 \times \log _{2} N-0.99 \tag{5}
\end{equation*}
$$

With increasing $N$ the conformation space becomes exponentially larger, creating more energy states. However, the rise in information with increasing $N$ is much slower than the rise in the bits of information required to fully classify the ensemble, referred to as $I^{\text {Source }}$. Recall that for an ensemble of size $W, I^{\text {Source }}=\log _{2} W$ and $W$ grows exponentially with $N$. For 2-D lattice models $W=0.103 * 2.691^{N}$ (Sullivan et al., 2003). Consequently, the percentage of the total information retrieved diminishes as the number of conformers grows (see Fig. 1 inset), making simple $H P$ contact potentials less informative for chains whose lengths near the size of real proteins.

The properties of the fully enumerated ensembles are dominated by the extended structures, analogous to the denatured state of proteins. In order to better resemble native and molten globule states (Chan and Dill, 1989), we also study subsets consisting of compact and semi-compact conformers only. Compact conformers are generated as perfect-square Hamilton walks where every lattice site is occupied. For semi-compact structures, the lattice is restricted to the smallest square which fits a chain of length $N$ (Table 2). The percentage of information recovered is higher in the compact and semicompact structures, $\sim 10-15 \%$ compared to $\sim 5 \%$ for the fully enumerated. The $H P$ potential quantifies the number of $H H$ contacts. Since the beads in compact and semicompact (depending on the tightness of the lattice fit) are limited to a square, there is a higher occurrence of $H H$ contacts leading to finer partitioning of the ensemble by the potential. However, pair-wise contact potentials offer only a small amount of the information required to discriminate amongst individual conformers.

## Solvent Interactions

Protein interactions of interest to biochemistry do not occur in a vacuum, but in a matrix of interactions with some form of solvent which influences their energetics. Atomsolvent interactions have been modeled both explicitly (Kollman et al., 2000; Levy and Gallicchio, 1998) and in more simplified models (continuum models) (Bashford and Case, 2000). Although, it has been shown that solvation terms improve theoretical calculations (Kollman et al., 2000), there has never been a quantitative analysis of their contribution.

In order to explore the information value of solvent interactions, we assume that each conformer is immersed in a uniform solvent. We determine the burial state of each residue within a conformer by counting the number empty lattice vertices around it. For two-dimensional lattice walks, the coordination number $z=4$, resulting in at most 3 additional solvent/bead interactions per residue.

For the fully enumerated sets there is a $\sim 2.5$ fold increase in the amount of recovered information (Figure 1). We also experiment with varying the weight scale for the solvent score. The amount of information is larger for weight scales that increase the score separation among conformers of varying salvation states. This occurs due the fact that the many of the scores resulting from the solvation scoring are a subset of the nonsolvated scores. On the other hand, the scores resulting from 0.2 and 1.0 offer either finer partitioning or enlarge the superset of scores by adding larger values. The separation of the solvation curves (Figure 1) for varying weight scales becomes larger with increasing N. Larger more diverse conformer sets reduce the degeneracy of the score set leading to better enrichment and information recovery.

As expected, the compact structures show only marginal rises in information when the energy score includes a solvation score (Figure 2). By design, the compact structures minimize the number of exposed residues by maximally filling all lattice vertices. The only exposed residues are those placed on the four lattice edges. We observe a consistent but small rise in information with increasing $N$. More information is recovered in the semi-compact structures which may be physically relevant to protein globular states. The smaller the ratio of $N$ to the lattice dimensions the larger the enrichment.

FIGURE 2: Information of force fields for compact structures. $\star$ : $I^{\mathrm{s}},>$ : Solvent with weight $0.2, ~ \square$ : Coulomb interactions, $\bullet$ : HP contacts,: *: Go-type potential


## Electrostatic Interactions

Biological processes are often governed by long-range electrostatic interactions (Honig and Nicholls, 1995). These distance dependent energies are modeled by Eq. 5 (methods section). While the HP force fields mimic short distance interactions, the distance dependent function can be used to illustrate the discrimination power of longrange (large space separations) pair-wise interactions.

For a set of two-dimensional conformers, we assign every lattice point as either a positive or negative charge. All possible $+/$ - combinations are explored for a given set of conformers. The energy of each lattice and subsequent entropy per set are evaluated according to equations 5 and 1 respectively. For exhaustive sets of fully enumerated conformers we observe that, on average, close to $80 \%$ of the maximum information is retrieved. Furthermore, electrostatic screening among residues, modeled by a $1 / \mathrm{r}^{2}$ potential, offers nearly the same amount of information (Figure 3). It is also worth mentioning that the information connected with various terms in common force fields is not additive. Instead it is only as informative as its most discriminating descriptor. In the case where both a coulomb energy function and a solvent potential function are used, there is no significant additional information to that supplied by the latter term (Figure 3). For compact structures, the amount of recovered information is slightly more than $50 \%$. Comparison of the performance of the pair-wise energy function on compact versus the fully enumerated set seems to indicate that extended structures are better described by long-range electrostatic terms than compact structures. Because the coulombic energy term is a sum over all residue pairs in the lattice, this function will fail to discriminate among compact conformers with the same sequence elements placed in different chain positions. In effect, the constraint for compactness reduces the effective number of conformers to one and highlights the energy differences among sequences.

FIGURE 3: Information from coulombic pair-wise interactions. $\star$ : $\mathrm{I}^{\mathrm{S}}, \triangleright$ : Coulomb energy + solvent,: *: $1 / \mathrm{r}$ coulomb energy, $\diamond 1 / \mathrm{r}^{2}$ dialectric screening, $\bullet: \mathrm{HP}$ contacts


## Go-type Potential

The force fields, above, utilize two different descriptors for each conformer/sequence pair: 1) A geometric descriptor, i.e. Cartesian coordinates and 2) A bead descriptor, i.e. residue type, charge. Consequently, two pairs of residues occupying the same chain position and lattice points can yield different energy contributions. In dealing with a Go-type potential we only consider geometric descriptors, in particular pairs of interacting residues, not adjacent in chain position but occupying neighboring lattice vertices (Eq. 6). For every target conformer (see methods) we consider all possible interactable pairs, the equivalent of nonspecific interactions. Although all compact conformations adhere to the square shape, the chain trace on the lattice is different and as a result most conformers share few similar contacts. As a results partitions can be either highly populated or sparse; to be part of a partition, each of members must exhibits the same degree of dissimilarity to the target structure as others, resemblance to the other cluster members is neither necessary nor required. Figure 2 shows that for compact conformers Go-type potentials offer only a small amount of information; similar to the
$H P$ contact and solvent potential. This is consistent with the absence of unique clusters containing similar structures. Given a target structure, the Go potential is capable of identifying " $a$ " similar structure or structures, based on the relative energies. However it can not effectively describe an ensemble by differentiating among its members.

## DISCUSSION

Score matrices and force fields, used in sequence and structure alignments respectively, are valued for their discriminating power when faced with choices among similar sequences or conformers. Their resolution is determined by both sensitivity and selectivity thresholds. By using information theory we are able to quantify the resolving power of several basic force fields in bits. Although our metric cannot comment on the correctness of the physical assumptions, such measures could indicate whether there is enough information in the FF to serve a particular purpose. Depending on the particular task at hand, one might ask that a force field discriminate well between "open" and "closed" conformations or that it distinguish among closed states. Our data show that contact-based interactions (i.e. solvent/solute, HP, and Go-type potentials) have much lower resolving power than interactions over larger distances, such as coulomb forces, even when the number of parameters is small. Further, distance-dependent potentials retain this advantage as the number of parameters (e.g. specific amino acid interactions) increases. However, subdividing the energy landscape more finely yields higher resolution of energies, which in turn, dictates a more elaborate treatment of entropy if one is to calculate accurate free energies. Another issue is how orthogonal are the terms included in force fields. For example, we see that adding solvation terms and/or distance dependent terms greatly expands the information content compared to simple square well
representations. We have certainly over-estimated the coulombic contributions in our simple model by treating every residue as ionic, but the principle remains clear that any distance dependence expands the resolving power of the force field.

Conclusions on the relative resolving power of force fields, the information content of various interactions, and the additivity of information appear extendible to real proteins. The move from simple to exact models will shift the reference states but not the general trend. Such quantitative assessments are critical for improving the effectiveness of current force fields and score functions used in various alignment protocols.

## ACKNOWLEDGMENTS

We are grateful for helpful discussions with Scott Pegg \& Kevin Masukawa. This work was supported by a grant from the National Science Foundation (NSF CHE-0118481); R. Kip Guy, Principal Investigator and a University of California president's dissertation year fellowship (to T. Aynechi).

## REFERENCES

Aynechi, T., and I.D. Kuntz. 2005. An information theoretic approach to molecular modeling: I. Sequence alignments. In preparation.

Bahar, I., and R.L. Jernigan. 1997. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol 266(1):195-214.

Bashford, D., and D.A. Case. 2000. Generalized born models of macromolecular solvation effects. Аппи Rev Phys Chem 51:129-52.

Berman, H.M., Westbrook,J., Feng,Z.,Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., Bourne, P.E. 2000. The Protein Data Bank. Nucleic Acids Research 28:235-242.

Chan, H.S., and K.A. Dill. 1989. Compact Polymers. Macromolecules(22):4559-4573.
Dill, K.A. 2004. Private Communication.
Dill, K.A., and S. Bromberg 2003. Molecular driving forces : statistical thermodynamics in chemistry and biology. Garland Science, New York

Dill, K.A., S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. 1995. Principles of protein folding--a perspective from simple exact models. Protein Sci 4(4):561-602.

Feldman, H.J., and C.W. Hogue. 2002. Probabilistic sampling of protein conformations: new hope for brute force? Proteins 46(1):8-23.

Go, N., and H. Taketomi. 1978. Respective roles of short- and long-range interactions in protein folding. Proc Natl Acad Sci U S A 75(2):559-63.

Halgren, T.A. 1995. Potential energy functions. Curr Opin Struct Biol 5(2):205-10.
Honig, B., and A. Nicholls. 1995. Classical electrostatics in biology and chemistry. Science 268(5214):1144-9.

Huber, T., and A.E. Torda. 1998. Protein fold recognition without Boltzmann statistics or explicit physical basis. Protein Sci 7(1):142-9.

Jones, D.T., W.R. Taylor, and J.M. Thornton. 1992. A new approach to protein fold recognition. Nature 358(6381):86-9.

Kollman, P. 1993. FREE ENERGY CALCULATIONS - APPLICATIONS TO CHEMICAL AND BIOCHEMICAL PHENOMENA [Review]. Chemical Reviews 93(7):2395-2417.

Kollman, P.A., I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, and T.E. Cheatham, 3rd. 2000. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res 33(12):889-97.

Kuntz, I.D., and D.A. Agard. 2003. Assessment of the role of computations in structural biology. Adv Protein Chem 66:1-25.

Lau, K.F., and K.A. Dill. 1989. A Lattice Statistical-Mechanics Model of the Conformational and Sequence-Spaces of Proteins. Macromolecules 22(10):39863997.

Levy, R.M., and E. Gallicchio. 1998. Computer simulations with explicit solvent: recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. Annu Rev Phys Chem 49:531-67.

Park, B., and M. Levitt. 1996. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 258(2):367-92.

Park, B.H., E.S. Huang, and M. Levitt. 1997. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 266(4):831-46.

Shannon, C.E. 1948. A mathematical theory of communication. The Bell System Technical Journal(27):379-423 and 623-656.

Sippl, M.J. 1995. Knowledge-based potentials for proteins. Curr Opin Struct Biol 5(2):229-35.

Solis, A.D., and S. Rackovsky. 2002. Optimally informative backbone structural propensities in proteins. Proteins 48(3):463-86.

Sullivan, D.C., T. Aynechi, V.A. Voelz, and I.D. Kuntz. 2003. Information Content of Molecular Structures. Biophys. J. 85(1):174-190.

Sullivan, D.C., and I.D. Kuntz. 2004. Distributions in protein conformation space: implications for structure prediction and entropy. Biophys $J$ 87(1):113-20.

Taketomi, H., Y. Ueda, and N. Go. 1975. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. Int J Pept Protein Res 7(6):445-59.

Tenette, C., J.C. Smith, and C.M. Topham. 1996. Force field development and conformational search strategy in the simulation of biomolecular recognition processes. Biochem Soc Trans 24(1):268-74.

Thomas, P.D., and K.A. Dill. 1996. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 257(2):457-69.

Young, J.F. 1971. Information Theory. Butterworth \& Company Ltd., Bristol

## Chapter Six: Estimation of Absolute Free Energies of

Hydration using Continuum Methods: Accuracy of
Partial Charge Models and Optimization of Nonpolar

## Contributions

Robert C. Rizzo, ${ }^{\S}$ Tiba Aynechi, ${ }^{\dagger}$ David A. Case, ${ }^{\ddagger}$ and Irwin D. Kuntz*<br>Contribution from the Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, California 94143-2240 and the Department of Molecular Biology TPC-15, The Scripps Research Institute, La Jolla, CA 92037<br>${ }^{\text {§ }}$ Current Address: Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York 11794-3600<br>${ }^{\dagger}$ Graduate Group in Biophysics<br>${ }^{\ddagger}$ The Scripps Research Institute<br>*Corresponding author. E-mail: kuntz@cgl.ucsf.edu<br>In preparation for publication.

## KEYWORDS

Free energy of hydration, MMPBSA, MMGBSA, Partial charge, Solvation, :


#### Abstract

ABSRACT

Calculation of binding free energies is essential to understanding molecular recognition and binding. Current theoretical methods such as Poisson-Boltzmann (PB) and Generalized Born (GB) rely on the use charge parameters that are obtained either using empirical or quantum mechanical methods. Using a large test set of 500 compounds to evaluate the accuracy of eight commonly used charge models.using both PB and GB methods. We show that the results from the faster GB method are highly correlated with PB results. In addition we show that estimations of solvation free energy can be improved by optimizing the constants used in the non-polar solvent accessible surface area (SASA) calculations.


## INTRODUCTION

The quantification of how a solute will partition into two different phases, A and B , is widely used in drug design.(Leo et al., 1971; Lipinski et al., 2001) Notable examples include using $n$-octanol to water partitioning ( $\log \mathrm{P}_{\text {octanol/water }}$ ) as a surrogate for cell membrane permeability and gas to water partitioning ( $\log \mathrm{P}_{\mathrm{gas} / \text { water }}$ ) to estimate desolvation penalties associated with protein-ligand binding. The two quantities are related, from the perspective of continuum models of solvation, in that they quantify partitioning between phases with low (gas $\sim 1$, octanol $\sim 17$ ) and high (water $\sim 80$ ) dielectric constants. Experimental $\log \mathrm{P}_{\mathrm{gas} / \text { water }}$ measurements, often expressed as free energies of hydration $\left(\Delta \mathrm{G}_{\text {hyd }}=-2.3 \log \mathrm{P}_{\text {gas/water }}\right)$, have been compiled by several research groups for both neutral and charged species (see Table S1 supporting information).(Abraham et al., 1990; Chambers et al., 1996; Gerber, 1998; Li et al., 1999; Marcus, 1994) These experimental data make computation of $\Delta \mathrm{G}_{\mathrm{hyd}}$ an attractive thermodynamic property for validating continuum simulation methods and can be used to guide the choice of parameters employed in such calculations.

Historically, the most accurate $\Delta \mathrm{G}_{\text {hyd }}$ calculations have employed free energy perturbation (FEP) or thermodynamic integration (TI) simulations incorporating explicit models of water.(Jorgensen, 1989; Kollman, 1993) This was first done in 1985 by Jorgensen and Ravimohan(Jorgensen and Ravimohan, 1985) who used FEP methods to compute the relative free energy of hydration $\left(\Delta \Delta \mathrm{G}_{\text {hyd }}\right)$ for ethane and methanol in excellent agreement with experiment using Monte Carlo simulations. The FEP and TI methods yield $\Delta \Delta \mathrm{G}_{\text {hyd }}$ (or $\Delta \mathrm{G}_{\text {hyd }}$ ) directly without the need for partitioning the free energy into separate components, as in other more approximate approaches described below,
however the simulations can be tedious to setup, computationally prohibitive for highthroughput structure-based design, and absolute free energies of hydration can be difficult to obtain. Although numerous alternative and diverse techniques for computation of absolute $\Delta \mathrm{G}_{\text {hyd }}$ have been devised, REF REF REF continuum theories which treat solvent as a bulk macroscopic quantity(Cramer and Truhlar, 1999) are of particular interest given the direct connection with recently reported protein-ligand binding calculation methods as described below. In particular, Poisson-Boltzmann (PB)(Sitkoff et al., 1994) and Generalized Born (GB)(Still et al., 1990) are two widely used continuum methods which may be used to evaluate the polarization energy associated with bringing any species from the gas-phase to the bulk solvent phase. PB and GB calculation results are typically augmented by a solvent accessible Surface Area term (SA) to account for non-polar contributions to the total free energy of hydration.

The recently reported Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) and Molecular Mechanics Generalized Born Surface Area (MM-GBSA) methods(Kollman et al., 2000; Massova and Kollman, 2000; Srinivasan et al., 1998) use continuum methods to compute a $\Delta \mathrm{G}_{\text {hyd }}$-like term as a measure of the change in desolvation $\left(\Delta \Delta \mathrm{G}_{\text {hyd }}\right)$ for the receptor-ligand binding event.(Donini and Kollman, 2000; Huo et al., 2002; Kuhn and Kollman, 2000a; Masukawa et al., 2003; Suenaga et al., 2003; Wang et al., 2001a; Wang et al., 2001b) In MM-PBSA and MM-GBSA analysis,(Kollman et al., 2000; Massova and Kollman, 2000; Srinivasan et al., 1998) the resulting PBSA or GBSA continuum energy terms for a given species (complex, receptor, or ligand) are formally equivalent to an absolute $\Delta \mathrm{G}_{\text {hyd }}$ if, as is commonly done, dielectric constants of 1 (gas-phase) and 80 (water-phase) are specified. Unfortunately,
experimental free energies of hydration are not available for proteins, most drugs, or protein-drug complexes which would allow for a direct comparison with the computational results. A reasonable alternative is to verify that the calculation methods and parameters yield good results for small organic molecules, for which experimental absolute free energies of hydration are available,(Abraham et al., 1994; Abraham et al., 1990; Marcus, 1991, 1994) prior to using such methods for estimating $\Delta \Delta \mathrm{G}_{\mathrm{hyd}}$ for protein-ligand binding.

The focus of the present study is to evaluate the accuracy of eight different point charge models typically used for structure-based drug design calculations through computation of hydration free energies. Here, we have evaluated point charge models based on ab initio, semiempirical, and empirical calculations. The results revealed a surprising lack of correlation of non-polar energy contributions with experiment and led us to pursue an optimization protocol which employed atom-based SA's in an attempt to improve the overall $\Delta \mathrm{G}_{\text {hyd }}$ for both neutral and charged molecules. It should be noted that a comprehensive study which compares the performance of various GB implementations to PB reference calculations has recently been reported by Feig et al.(Feig et al., 2004) In this report, we instead focus on evaluating which commonly used partial charge models yield GBSA and PBSA absolute hydration free energies in agreement with experiment.

Two earlier studies that directly compared continuum $\Delta \mathrm{G}_{\mathrm{hyd}}$ results with experiment include the original GBSA report by Still et al.(Still et al., 1990) and the Sitkoff et al.(Sitkoff et al., 1994) PARSE (Parameters for Solvation Energy) study designed for use with PBSA methods. Both studies reported excellent continuum results for $\Delta \mathrm{G}_{\text {hyd }}$ versus experiment however the number of molecules tested were relatively
small (between 20-67 molecules).(Sitkoff et al., 1994; Still et al., 1990) The Still GBSA and Sitkoff PBSA studies employed OPLS(Jorgensen and Tirado-Rives, 1988) and PARSE(Sitkoff et al., 1994) atomic partial charge models respectively which yielded calculated $\Delta \mathrm{G}_{\text {hyd }}$ results in good agreement with experiment for the compounds tested. However, OPLS and PARSE partial charge models are primarily based on "functional groups" assignment which can be difficult to assign to molecules typically found in databases used for structure-based drug design and high throughput virtual screening calculations. Here, we have focused on evaluating partial charge models that could be easily be assigned, in an automated fashion, to relatively large and diverse data sets. In the present work, we have compared experimental versus calculated $\Delta \mathrm{G}_{\text {hyd }}$ for more than 500 compounds ( 460 neutral compounds, 42 polyatomic ions, and 11 monoatomic ions). To our knowledge this is the largest number of reference compounds employed for continuum model calculations.

The aim of this work is: (1) to use GBSA and PBSA methods to compute absolute free energies of hydration ( $\Delta \mathrm{G}_{\text {hyd }}$ ) for comparison with experiment in order to evaluate the accuracy of eight different partial charge models and (2) to optimize non-polar contributions to $\Delta \mathrm{G}_{\text {hyd }}$ using atom-based solvent accessible surface areas for each charge model and theoretical method tested. Notwithstanding the inherent theoretical differences between GB and PB methods, it is our view that parameter set validation is critical since the use of different atomic partial charge models, atomic radii, and nonpolar SA parameters will lead to different calculated $\Delta \mathrm{G}_{\mathrm{hyd}}$ results.

## COMPUTATIONAL METHODS

## Free Energies of Hydration ( $\Delta \mathrm{G}_{\text {hyd }}$ ).

As in prior continuum studies,(Sitkoff et al., 1994; Still et al., 1990) the free energy of hydration is partitioned into two terms, polar and nonpolar, according to eq 1 .

$$
\begin{equation*}
\Delta \mathrm{G}_{\text {hyd }}=\mathrm{G}_{\text {polar }}+\mathrm{G}_{\text {nonpolar }} \tag{1}
\end{equation*}
$$

Polar energies $\left(\mathrm{G}_{\text {polar }}\right)$ for PB calculations were obtained using a grid-based finite difference solution to the Poisson-Boltzmann equation with zero salt concentration (eq 2), where $\rho(r)$ is the charge distribution of the molecule and $\varepsilon(r)$ is the dielectric constant. Solution of the PB equation for systems described by a classical force field yields the electrostatic potential at every grid point and $\mathrm{G}_{\mathrm{polar}}$ is then evaluated as a sum over all atoms (eq 3) where the partial atomic charge for atom i is multiplied by the difference in the computed grid-point potential $\phi_{i}$ for the transfer from gas-phase $(\varepsilon=1)$ to water $(\varepsilon=$ 80).

$$
\begin{align*}
& \nabla \mathcal{E}(r) \nabla \phi(r)+4 \pi \rho(r)=0  \tag{2}\\
& \mathrm{G}_{\text {polar }}=\frac{1}{2} \sum_{i}^{N} q_{i}\left(\phi_{i}^{80}-\phi_{i}^{1}\right) \tag{3}
\end{align*}
$$

For GB calculations, $\mathrm{G}_{\text {polar }}$ contributions were obtained using eq $4-5$. Here, $\varepsilon$ is the dielectric constant ( 80 for water-phase), $\mathrm{r}_{\mathrm{ij}}$ is the interatomic distance, and $\alpha_{\mathrm{i}}$ are the Born radii which are computed according to the pairwise descreening algorithm of Hawkins et al.(Hawkins et al., 1995, 1996)

$$
\begin{equation*}
G_{\text {polar }}=-166\left(1-\frac{1}{\varepsilon}\right) \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq 1}}^{N} \frac{q_{i} q_{j}}{f_{G B}}-166\left(1-\frac{1}{\varepsilon}\right) \sum_{i=1}^{N} \frac{q_{i}^{2}}{\alpha_{i}} \tag{4}
\end{equation*}
$$

$$
\begin{equation*}
f_{G B}=\left(r_{i j}^{2}+\alpha_{i j}^{2} e^{-r_{i j}^{2} /\left(4 \alpha_{i j}^{2}\right)^{0.5}}\right)^{0.5} \tag{5}
\end{equation*}
$$

Nonpolar contributions ( $\mathrm{G}_{\text {nonpolar }}$ ) to $\Delta \mathrm{G}_{\text {hyd }}$ were estimated using either total molecular SA (eq 6) or atomic-based $\mathrm{SA}_{\mathrm{i}}$ (eq 7 ). Prior MM-PBSA and MM-GBSA binding energy protocols typically employed molecular SA (eq 6 ) with $\gamma=0.00542$, and $\beta=0.92$ as recommended by Kollman and coworkers.(Kollman et al., 2000; Massova and Kollman, 2000) An alternative procedure, which was pursued in the present work, is to compute atom-based $\mathrm{SA}_{\mathrm{i}}$ and optimize each SA constant using multiple linear regression to improve agreement with experiment (eq 7). Using atom-based $\mathrm{SA}_{\mathrm{i}}$ contributions to estimate free energies of solvation was first proposed by Eisenberg and McLachlan,(Eisenberg and McLachlan, 1986) and Scheraga and coworkers.(Ooi et al., 1987)

$$
\begin{gather*}
G_{\text {nonpolar }}=\left(\gamma^{*} S A\right)+\beta  \tag{6}\\
\Delta G_{\text {hyd }}(\text { exptl })-G_{\text {polar }}=G_{\text {norpolar }}=\sum_{i} c_{i} S A_{i} \tag{7}
\end{gather*}
$$

For a given set of calculations, PBSA or GBSA, the same structures, partial charges, and atomic radii were employed. Any differences in the final calculation results in this report will therefore be only a function of the two different continuum theories.

## Computation details

The atomic radii used in the calculations were assigned using the mbondi (modified bondi)(Tsui and Case, 2000a, b) scheme as in AMBER7.(2002a) In the mbondi scheme, hydrogen atoms connected to carbon, sulfur, nitrogen, or oxygen (types HC, HS, HN, or HO respectively) have unique radii (Table 1). Dielectric constants for all calculations ( PB and GB ) were set to 1 representing gas-phase and 80 representing
water-phase. PB calculations were performed using the program Delphi4(Rocchia et al., 2001; Rocchia et al., 2002) with the following parameters:, boundary conditions $=4$, internal dielectric constant $=1.0$, external dielectric constant $=80.0$, scale $=4$ grids $/ \AA$. Other Delphi parameters were assigned automatically using default values. Generalized Born calculations were performed using an in-house version of the Hawkins et al.(Hawkins et al., 1995, 1996) pairwise de-screening model with scaling parameters (Sx values Table 1) adopted from Tsui and Case.(Tsui and Case, 2000a)

The DMS program was used for all the SA calculations.(2003) In addition to the total SA value for a compound, the DMS program reports individual atom-based $\mathrm{SA}_{\mathrm{i}}$ estimates that were used to derive atom-based constants (eq 7) for the non-polar component of the free energy of hydration.

## Molecular structures and experimental data

Bordner et al.(Bordner et al., 2002) have generously made available 410 neutral molecular structures along with the corresponding experimental $\log \mathrm{P}_{\text {gas/water }}$ partition coefficients from the tabulated work of Abraham and coworkers(Abraham et al., 1990) (converted to free energies at $25{ }^{\circ} \mathrm{C}$ using $\Delta \mathrm{G}_{\text {hyd }}=-2.3 \log \mathrm{P}_{\text {gas } / \text { water }}$ ). However, the Bordner set did not contain compounds with polar hydrogens connected to sulfur (HS, Table 1) or charged species . We augmented the 410 neutral set with 50 additional neutral compounds (including compounds containing HS), as well as 42 charged ( $\pm 1$ ) polyatomic compounds and 11 ionic monoatomic species (see Table S1 supporting information). These 103 additional compounds were constructed using the MOE program.(2002b)

TABLE 1: Atom type, atomic radii, GB scaling factor, and total number of each atom type in the total dataset. ${ }^{\text {a }}$

| Type | mbondi radii | Sx value | No. atoms |
| :---: | :---: | :---: | :---: |
| HC | $1.30{ }^{\text {b }}$ | $0.85{ }^{\text {d }}$ | 4215 |
| HN | $1.30{ }^{\text {b }}$ | $0.85{ }^{\text {d }}$ | 98 |
| HO | $0.80{ }^{\text {b }}$ | $0.85{ }^{\text {d }}$ | 93 |
| HS | $0.80{ }^{\text {b }}$ | $0.85{ }^{\text {d }}$ | 13 |
| C | $1.70{ }^{\text {b }}$ | $0.72{ }^{\text {d }}$ | 2678 |
| N | $1.55{ }^{\text {b }}$ | 0.79 d | 128 |
| O | $1.50{ }^{\text {b }}$ | $0.85{ }^{\text {d }}$ | 299 |
| F | $1.50{ }^{\text {b }}$ | $0.88{ }^{\text {d }}$ | 53 |
| P | 1.85 | $0.86{ }^{\text {d }}$ | 6 |
| S | 1.80 | $0.96{ }^{\text {d }}$ | 26 |
| Cl | 1.70 | $0.80{ }^{\text {c default }}$ | 114 |
| Br | 1.85 | $0.80{ }^{\text {c default }}$ | 27 |
| I | 1.98 | $0.80{ }^{\text {c default }}$ | 12 |
| Li+ | $1.82{ }^{\text {c }}$ | $0.80{ }^{\text {c default }}$ | 1 |
| $\mathrm{Na}+$ | $2.27{ }^{\text {c }}$ | $0.80{ }^{\text {e default }}$ | 1 |
| K+ | $2.75{ }^{\text {c }}$ | $0.80{ }^{\text {c default }}$ | 1 |
| $\mathrm{Mg}^{2+}$ | $1.73{ }^{\text {c }}$ | $0.80{ }^{\text {e default }}$ | 1 |
| $\mathrm{Ca}^{2+}$ | $1.70{ }^{\text {c }}$ | 0.72 | 1 |
| $\mathrm{Fe}^{2+}$ | $1.50{ }^{\text {c }}$ | $0.80{ }^{\text {e default }}$ | 1 |
| $\mathrm{Zn}^{2+}$ | $1.39{ }^{\text {c }}$ | $0.80{ }^{\text {c default }}$ | 1 |

## Partial Charge Models

Eight charge models were evaluated in this study: Gasteiger-Marsili (Gast),(Gasteiger and Marsili, 1980) MMFF94,(Halgren, 1996) AM1BCC,(Jakalian et al., 2000; Jakalian et al., 2002) AM1CM2,(Li et al., 1998) PM3CM2,(Li et al., 1998) Merz-Singh-Kollman (MSK),(Besler et al., 1990) Restrained Electrostatic Potential (RESP),(Bayly et al., 1993; Cornell et al., 1993) and ChelpG.(Breneman and Wiberg, 1990) While the preceding list is not exhaustive, it does include methods currently implemented in several molecular modeling packages which allow for the rapid calculation of atomic partial charges. Here, Gast and MMFF94 charges were assigned using the program MOE.(2002b) AM1BCC charges were determined using the ANTECHAMBER module in AMBER7(2002a) from MOPAC(Stewart et al., 1999) calculations. AM1CM2 and PM3CM2 partial charges(Li et al., 1998) were computed using the program AMSOL(Hawkins et al., 1999) with the SM5.42R(Li et al., 1999) water solvent model specified. MSK, RESP, and ChelpG charges were computed at the
$\mathrm{HF} / 6-31 \mathrm{G}^{*} / / \mathrm{HF} / 6-31 \mathrm{G}^{*}$ level of theory using the program Gaussian98.(Frisch et al., 1998) The ANTECHAMBER module in AMBER7 was used for the two-stage RESP fittings. It should be noted that different software programs may yield slight variations in molecular charge distributions due to differences in implementation of a particular partial charge model. Only the above named program implementations were evaluated in this report.

## Molecular geometries

For each compound, the partial charges obtained using the eight different methods were mapped back to one set of standard geometries. Using one set of conformations allows for a direct comparison of the accuracy of the partial charge models and removes the possibility that different geometries would affect the results. Here, the standard geometries were defined as that obtained from a gas-phase geometry optimization using the MMFF94 force-field as implemented in the MOE program. Other geometries could have been used although this was not explored in the present work. Given that the data set contains mostly rigid compounds the effect of including multiple conformations on the computed free energies of hydration was not investigated; averaging over multiple conformations in the previous Bordner study(Bordner et al., 2002) changed the computed free energies by only a trivial amount.

## RESULTS AND DISCUSSION

## Charge Model Evaluation

Free energies of hydration were computed for comparison with experiment for 460 neutral and 42 charged compounds employing one of eight different partial charge models (Gast, MMFF94, AM1BCC, AM1CM2, PM3CM2, MSK, RESP, and ChelpG).

Table 2 lists the correlation coefficients ( $\mathrm{r}^{2}$ ) and average unsigned errors (AUE) between experiment and calculated $\Delta \mathrm{G}_{\text {hyd }}$ obtained from both PBSA and GBSA calculations. In Table 2 the $\mathrm{G}_{\text {nonpolar }}$ term is computed from molecular SA (eq 6) using the standard MMPBSA and MM-GBSA constants $(\gamma=0.00542, \beta=0.92)$. It should be noted that in every case the correlations between the experimental and theoretical free energies in Table 2 are due to solely to the $\mathrm{G}_{\text {polar }}$ term; molecular SAs show no correlation with experiment (Figure 1).

FIGURE 1: Experimental free energies of hydration versus total molecular solvent accessible surface area. The best fit line to the 27 linear and branched alkanes $\left(\mathbb{O}\right.$ ) yields a correlation coefficient $\mathrm{r}^{2}=0.85$, slope $\gamma$ $=0.00538$, and intercept $\beta=0.92$. Other compounds are represented as filled squares $(\boldsymbol{\square})$.
 using both species together. This is primarily due to the large difference in magnitude of the experimental data for charged versus neutral species.

TABLE 2: Correlation coefficients ( $\mathrm{r}^{2}$ ) and average unsigned errors (AUE) for experimental ${ }^{\mathrm{a}}$ vs. calculated ${ }^{b}$ (PBSA or GBSA) free energies of hydration $\left(\Delta \mathrm{G}_{\text {hyd }}\right)$. Nonpolar contributions obtained using molecular solvent accessible surface areas with standard constants. ${ }^{\text {c }}$ Energies in $\mathrm{kcal} / \mathrm{mol}$

| Charge Model | Neutral molecules, $\mathrm{N}=460$ Part I |  |  |  | Charged ( $\pm 1$ ) molecules, $\mathrm{N}=42$ <br> Part II |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\mathrm{r}^{2}$ PBSA | AUE | $\mathrm{r}^{2}$ GBSA | AUE | $\mathrm{r}^{2}$ PBSA | AUE | $\mathrm{r}^{2}$ GBSA | AUE |
| Gast | 0.53 | 3.20 | 0.49 | 3.36 | 0.68 | 7.52 | 0.67 | 8.15 |
| MMFF94 | 0.29 | 3.26 | 0.26 | 3.41 | 0.73 | 7.44 | 0.72 | 8.27 |
| AM1BCC | 0.74 | 1.36 | 0.70 | 1.38 | 0.56 | 8.28 | 0.53 | 9.64 |
| AM1CM2 | 0.71 | 3.09 | 0.67 | 2.81 | 0.39 | 11.67 | 0.34 | 13.63 |
| PM3CM2 | 0.69 | 2.79 | 0.64 | 2.61 | 0.62 | 10.84 | 0.62 | 11.90 |
| MSK | 0.77 | 1.54 | 0.72 | 1.63 | 0.74 | 6.42 | 0.72 | 7.30 |
| RESP | 0.77 | 1.47 | 0.72 | 1.51 | 0.75 | 6.34 | 0.73 | 7.20 |
| ChelpG | 0.73 | 1.61 | 0.69 | 1.67 | 0.74 | 6.36 | 0.72 | 7.28 |

The correlation coefficients for neutral compounds in Table 2 (Part I) track with the eight different charge schemes in roughly the following order; ab initio (MSK, RESP, ChelpG) > semiempirical (AM1BCC, AM1CM2, PM3CM2) > empirical (Gast, MFF94). Ab initio charges yield PBSA and GBSA $r^{2}$ values from 0.69 to 0.77 , semiempiricle $r^{2}$ from 0.64 to 0.74 , and empirical $\mathrm{r}^{2}$ from 0.26 to 0.53 . Average unsigned errors (AUE) follow the $\mathrm{r}^{2}$ trends; ab initio charges yield smaller errors ( 1.47 to $1.67 \mathrm{kcal} / \mathrm{mol}$ ), than do semiempiricle ( 1.36 to $3.09 \mathrm{kcal} / \mathrm{mol}$ ), or empirical ( 3.20 to $3.41 \mathrm{kcal} / \mathrm{mol}$ ).

Surprisingly the three semiempirical methods yield poorer agreement with experiment than do the two empirical methods for charged $( \pm 1)$ molecules (Table 2, Part II). Ab initio charges yield the strongest correlations with $\mathrm{r}^{2}$ values from 0.72 to 0.75 , semiempiricle $\mathrm{r}^{2}$ from 0.34 to 0.62 , and empirical $\mathrm{r}^{2}$ from 0.72 to 0.73 . As was the case for neutral compounds the AUE errors also track with the correlation coefficients. Again ab initio partial charges yield the lowest errors ( 6.34 to $7.30 \mathrm{kcal} / \mathrm{mol}$ ) but for the charged species, semiempiricle yield the largest errors ( 8.28 to $13.63 \mathrm{kcal} / \mathrm{mol}$ ). Empirical AUEs are in the middle ( 7.44 to $8.27 \mathrm{kcal} / \mathrm{mol}$ ).

Thus, using MSK, RESP, and ChelpG partial charges for neutral and charged species consistently yield the strongest correlations and lowest average unsigned errors
with experimental free energies of hydration regardless of which continuum method was employed for the computation (Table 2). The $r^{2}$ values for these three ab initio methods cluster around 0.75 for both neutral and charged species (Figure 2). Using semiempicircle charges with continuum methods for computation of $\Delta \mathrm{G}_{\text {hyd }}$ appear to yield good agreement with experiment only for neutral compounds.

In an attempt to increase $\mathrm{r}^{2}$ correlations with experiment for poorly performing charge methods and reduce average AUEs across the board for neutral and charged species, we investigated optimizing nonpolar energy terms using atom-based solvent accessible surface areas $\left(\mathrm{SA}_{\mathrm{i}}\right)$ instead of molecular surface areas (eq 7) as described below.

FIGURE 2: Comparison of correlation coefficients ( $r^{2}$ values) for calculated versus experimental free energies of hydration from PBSA and GBSA calculations. For each partial charge model two $\mathrm{r}^{2}$ values are plotted representing results for 460 neutral compounds (filled symbols) and 42 charged compounds (open symbols) compounds (Table ?). The overall correlation between the total PBSA and GBSA results is $\mathrm{r}^{2}=$ 0.94 .


## PBSA vs.GBSA

The PBSA and GBSA results are highly correlated and independent of the charge model used for the calculations (Table 2, Figure 2). The strong agreement between PBSA and GBSA $r^{2}$ values (obtained from computed versus experimental results) suggest that a given partial charge model will influence the final free energies much more than which continuum method (PBSA or GBSA) is used for the calculations. Correlation coefficients between PB and GB polar energies are always very strong, $\mathrm{r}^{2}>0.94$, and independent of which partial charge model or data set (neutral or charged compounds) was employed in the calculations. These trends continue to provide support for using GBSA methods as a reasonable alternative to the more computationally demanding PBSA calculations for free energy calculations

## G nonpolar from Molecular SA

The constants $(\gamma=0.00542, \beta=0.92)$ typically used(Donini and Kollman, 2000; Huo et al., 2002; Kuhn and Kollman, 2000a; Masukawa et al., 2003; Rizzo et al., 2004; Suenaga et al., 2003; Wang et al., 2001a; Wang et al., 2001b) in MM-PBSA and MMGBSA calculations to convert SA $\left(\AA^{2}\right)$ to $G_{\text {nonpolar }}(\mathrm{kcal} / \mathrm{mol})$ are based on fitting molecular SA results to experimental $\Delta \mathrm{G}_{\text {hyd }}$ for small straight-chain alkanes.(Sitkoff et al., 1994) The rational for this procedure exploits the fact that alkanes have low dipole moments and nonpolar contributions will therefore dominate $\Delta \mathrm{G}_{\mathrm{hyd}}$. Figure 1 shows the molecular SA for the 460 neutral molecules studied here versus experimental $\Delta \mathrm{G}_{\text {hyd }}$ along with the best fit regression line using only the 27 linear and branched alkanes. The constants obtained from this linear regression fit (Figure 1, open circles, $\mathrm{r}^{2}=0.85, \mathrm{~m}=$ $0.00538, b=0.92)$ are essentially identical to the standard constants $(\gamma=0.00542, \beta=$
0.92).(Donini and Kollman, 2000; Huo et al., 2002; Kuhn and Kollman, 2000a; Masukawa et al., 2003; Rizzo et al., 2004; Suenaga et al., 2003; Wang et al., 2001a; Wang et al., 2001b) However, as a group, molecular SAs have no correlation with experiment (Figure 1, filled squares). Although the results will be charge model dependent, in general, $\mathrm{G}_{\text {polar }}$ contributions are linearly correlated with $\Delta \mathrm{G}_{\mathrm{hyd}}$. This is illustrated in Figure 3 for RESP charged neutral compounds in which the polar energies ( $\mathrm{G}_{\text {polar }} \mathrm{r}^{2}=0.77$, filled squares) were computed using PB calculations. Given that $\Delta \mathrm{G}_{\text {hyd }}$ is estimated from the linear sum of two terms (eq 1), the sum of $G_{\text {polar }}$ and $G_{\text {nonpolar }}$ and both terms individually should be linear with experiment if there is to be agreement. However, using standard constants (eq $6, \gamma=0.00542, \beta=0.92$ ) to compute nonpolar contributions yield no correlation ( $\mathrm{G}_{\text {nonpolar }} \mathrm{r}^{2}=0.00$, open circles) and will therefore not contribute to any improvement or diminishment in the total correlation coefficient with experiment $\left(\Delta \mathrm{G}_{\text {hyd }} \mathrm{r}^{2}=0.77\right)$. In short, converting molecular-based $S A$ to $G_{\text {nonpolar }}$ energies (eq 6) for use in computing absolute free energies of hydration yields no advantage over using $\mathrm{G}_{\text {polar }}$ energies alone; correlations listed in Table 2 are due solely to the $\mathrm{G}_{\mathrm{polar}}$ term.

FIGURE 3: Correlation of individual components with experimental free energies of hydration for neutral compounds ( $\mathrm{N}=460$ ) using RESP derived partial charges. Polar ( - ) energies $\mathrm{G}_{\text {polar }}$ from PB calculations. Nonpolar (O) energies from molecular solvent accessible surface area calculations $\mathrm{G}_{\text {nonpolar }}=$ $\left(\mathrm{SA}_{\text {total }} * 0.00542\right)+0.92$


## $\mathrm{G}_{\text {nonpolar }}$ from Atomic $\mathrm{SA}_{\mathrm{i}}$.

The lack of correlation between $G_{\text {nonpolar }}$ with $\Delta G_{\text {hyd }}$ is troublesome, especially given that molecule-based SA $G_{\text {nonpolar }}$ energies are routinely added to $G_{\text {polar }}$ to estimate $\Delta \mathrm{G}_{\text {hyd }}$ (or alternatively $\Delta \Delta \mathrm{G}_{\text {hyd }}$ for protein-ligand binding).(Donini and Kollman, 2000; Huo et al., 2002; Kuhn and Kollman, 2000a; Masukawa et al., 2003; Rizzo et al., 2004; Suenaga et al., 2003; Wang et al., 2001a; Wang et al., 2001b) In an attempt to improve agreement with experiment we explored a procedure to re-compute $\mathrm{G}_{\text {nonpolar }}$ which includes calculation of atom-based $\mathrm{SA}_{\mathrm{i}}$ and makes use of multiple linear regression fitting to determine an optimal coefficient for each SA type.(Eisenberg and McLachlan, 1986; Ooi et al., 1987) For a given compound, the total solvent accessible surface area should be equivalent to the sum of each atom-based solvent accessible surface area $\left(\mathrm{SA}=\Sigma \mathrm{SA}_{\mathrm{i}}\right)$. Atomic $\mathrm{SA}_{\mathrm{i}}$ for each mbondi type (Table 1) were obtained from calculations using the DMS program.(2003) We optimized $\mathrm{C}_{\mathrm{i}}$ coefficients for each mbondi type (HC, HN, HO,

HS, C, N, O, F, P, S, Cl, Br) using continuum results ( $\mathrm{G}_{\mathrm{polar}}$ from GB or PB ) for each of the eight charge models using eq 7. After the fittings, new $\mathrm{G}_{\text {nonpolar }}$ contributions were recomputed using the atom-based constants $\left(\mathrm{C}_{\mathrm{i}}\right.$ 's) so that computed $\Delta \mathrm{G}_{\text {hyd }}$ could be compared with experiment.

Initially, fits were pursued using only the 460 neutral compounds. However, using these $\mathrm{SA}_{\mathrm{i}}$ constants to compute $\Delta \mathrm{G}_{\text {hyd }}$ for the 42 charged $( \pm 1)$ species lead to poor agreement with experiment. Our initial tests lead us to conclude that including neutral $(\mathrm{N}=460)$ and charged $(\mathrm{N}=42)$ compounds together in the fitting procedure would yield the best overall agreement with experiment. Therefore, this protocol was adopted for subsequent parameter optimizations.

In most cases, utilizing optimized $\mathrm{SA}_{\mathrm{i}}$ constants to estimate nonpolar terms improves agreement with $\Delta \mathrm{G}_{\text {hyd }}$ experiment (Table 3 versus 2). However, substantial improvement in AUE and correlations for charged species are coupled to diminishment in $\mathrm{r}^{2}$ for neutral compounds that have utilized semiempirical partial charges (AM1BCC, AM1CM2, PM3CM2). This diminishment is not surprising given that the semiempirical models originally performed quite poorly for charged compounds (Table 2, Part II, $\mathrm{r}^{2}=$ $0.34-0.62$ ). Optimization of the $\mathrm{SA}_{\mathrm{i}}$ constants using all data (neutral and charged) attempts to corrects for differences between experiment and theory in an average sense.

TABLE 3: Correlation coefficients $\left(r^{2}\right)$ and average unsigned errors (AUE) for experimental ${ }^{\text {a }}$ vs. calculated ${ }^{b}$ (PBSA or GBSA) free energies of hydration $\left(\Delta \mathrm{G}_{\text {hyd }}\right)$. Nonpolar contributions obtained using atomic solvent accessible surface areas with optimized constants from both neutral and charged species. ${ }^{\text {c }}$ Energies in $\mathrm{kcal} / \mathrm{mol}$.

| Charge Model | Neutral molecules, $\mathrm{N}=460$ |  |  |  | Charged ( $\pm 1$ ) molecules, $\mathrm{N}=42$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Part I |  |  |  | Part II |  |  |  |
|  | $\mathrm{r}^{2}$ PBSA fit | AUE | $\mathrm{r}^{2}$ GBSA fit | AUE | $\mathbf{r}^{2}$ PBSA fit | AUE | $\mathrm{r}^{2}$ GBSA fit | AUE |
| Gast | 0.67 | 1.43 | 0.56 | 1.62 | 0.69 | 8.60 | 0.69 | 8.99 |
| MMFF94 | 0.36 | 1.91 | 0.28 | 2.07 | 0.70 | 8.24 | 0.68 | 8.60 |
| AM1BCC | 0.68 | 1.26 | 0.58 | 1.49 | 0.61 | 6.71 | 0.60 | 6.83 |
| AM1CM2 | 0.62 | 1.71 | 0.54 | 1.83 | 0.55 | 7.35 | 0.58 | 7.55 |
| PM3CM2 | 0.61 | 1.66 | 0.52 | 1.83 | 0.68 | 7.24 | 0.71 | 7.47 |
| MSK | 0.81 | 0.99 | 0.69 | 1.32 | 0.79 | 4.46 | 0.77 | 4.68 |
| RESP | 0.80 | 1.02 | 0.69 | 1.33 | 0.80 | 4.45 | 0.78 | 4.69 |
| ChelpG | 0.81 | 0.99 | 0.70 | 1.30 | 0.79 | 4.46 | 0.77 | 4.67 |

Figures 4 and 5 highlight favorable cases where atom-based constants can be useful even in cases where a particular charge model leads to good agreement with experiment. Ab initio charges (MSK, RESP, ChelpG) appear to yield $\mathrm{G}_{\mathrm{polar}}$ energies in strong correlation with experiment for neutral and charged compounds in all cases. However, using molecule-based constants (grey crosses) to compute $\mathrm{G}_{\text {nonpolar }}$ can lead to a systematic overestimate (absolute error) of the hydration free energies for species with ab initio charges in the experimental range from -11 to $-2 \mathrm{kcal} / \mathrm{mol}$ for neutrals and -90 to $-60 \mathrm{kcal} / \mathrm{mol}$ for charged species. As an example, for neutrals, Figure 4 show that using $C_{i}$ constants optimized from PBSA-RESP fits leads to an improvement $r^{2}$ from 0.77 ( $\Delta \mathrm{G}_{\text {hyd }}$ std, grey crosses $)$ to $0.80\left(\Delta \mathrm{G}_{\text {hyd }}\right.$ fit, black squares) and the AUE error with experiment drops from 1.47 to $1.02 \mathrm{kcal} / \mathrm{mol}$ (Figure 4). More dramatic results are observed for charged species; PBSA correlations increases from $0.75\left(\Delta \mathrm{G}_{\mathrm{hyd}}\right.$ std, grey crosses) to 0.80 ( $\Delta \mathrm{G}_{\mathrm{hyd}}$ fit, black squares) and the AUE error with experiment drops dramatically from 6.34 to $4.45 \mathrm{kcal} / \mathrm{mol}$ (Figure 5).

FIGURE 4: Predicted free energies of hydration ( $\Delta \mathrm{G}_{\text {hyd }}$ calcd) vs experiment ( $\Delta \mathrm{G}_{\text {hyd }}$ exptl) from PBSA calculations with RESP charges for neutral compounds ( $\mathrm{N}=460$ ). Nonpolar energies from molecular SA's using standard constants ( $\times$ ) or atom-based SA's using fitted constants ( $\quad$ ).


FIGURE 5: Predicted free energies of hydration ( $\Delta \mathrm{G}_{\text {hyd }}$ calcd) vs experiment ( $\Delta \mathrm{G}_{\text {hyd }}$ exptl) from PBSA calculations with RESP charges for charged compounds ( $\mathrm{N}=42$ ). Nonpolar energies from molecular SA's using standard constants $(\times)$ or atom-based SA's using fitted constants ( $\quad$ ).


The primary motivation for using atom-based $\mathrm{SA}_{\mathrm{i}}$ instead of molecule-based SA procedures is to reduce errors with respect to experiment in three ways: (1) remedy gross deficiencies a particular charge model may have ( $r^{2}$ and AUE), (2) fine tune already reasonable agreement with experiment (primarily AUE), or (3) account for minor
differences between PB and GB results. And, on a case-by-case basis atom-based constants (Figures 4-5, black squares) can correct for systematic errors.

## Optimized SA Coefficients

Tables 5 and 6 list "sets" of optimized $\mathrm{SA}_{\mathrm{i}}$ constants $\left(\mathrm{C}_{\mathrm{i}}\right)$ obtained from multiple linear regressions using PB and $\mathrm{GB} \mathrm{G}_{\text {polar }}$ results for all eight charge models employed in the calculations. For a new calculation that employes a particular charge model atombased $C_{i}$ values can be used to estimate $G_{\text {nonpolar }}$ energies that should lead to improved $\Delta \mathrm{G}_{\text {hyd }}$ calculations. Despite the fact that $\mathrm{G}_{\text {polar }}$ results from both continuum methods show strong correlation (Table 2 and 3 ), for completeness, separate fits were performed for PB (Table 5) or GB (Table 6) derived $\mathrm{G}_{\text {polar }}$ energies.

TABLE 4: Optimized atomic SA coeffecients (Ci values) ${ }^{\text {a }}$ obtained using Poisson-Boltzmann (PB) derived $\mathrm{G}_{\text {polar }}$ energies.

| Type | Gast | MMFF94 | AM1BCC | AM1CM2 | PM3CM2 | MSK | RESP | ChelpG |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Hc | 0.00093 | -0.00002 | 0.00355 | 0.00962 | 0.00827 | 0.00679 | 0.00687 | 0.00649 |
| Ho | -0.00434 | -0.11172 | 0.25999 | 0.12379 | 0.11210 | 0.37414 | 0.36422 | 0.36037 |
| Hs | 0.28952 | 0.23307 | 0.33731 | -0.53475 | -0.45896 | 0.05493 | 0.06772 | 0.09424 |
| Hn | -0.04103 | -0.02779 | -0.01058 | -0.01094 | -0.01857 | -0.00574 | -0.00436 | -0.00813 |
| Hp | -0.12342 | 0.00990 | 0.02589 | 0.47729 | 0.38605 | -0.02415 | -0.00164 | -0.01025 |
| C | -0.01634 | -0.01610 | 0.02001 | 0.04395 | 0.03708 | 0.01765 | 0.01468 | -0.00278 |
| N | -0.00798 | -0.01032 | 0.07251 | 0.05061 | 0.08398 | 0.04518 | 0.04440 | 0.05156 |
| O | 0.00759 | 0.04621 | 0.02409 | 0.09277 | 0.08863 | 0.03592 | 0.03292 | 0.04072 |
| F | 0.02036 | 0.02024 | 0.02256 | 0.02661 | 0.01954 | 0.01755 | 0.01643 | 0.01873 |
| P | 2.12323 | 0.36337 | 0.98863 | -2.44577 | -2.59507 | 0.92016 | 0.61608 | 0.79176 |
| S | 0.01477 | 0.02908 | 0.05082 | 0.15426 | 0.13041 | 0.04414 | 0.04145 | 0.03315 |
| Cl | 0.00336 | 0.00302 | 0.00384 | 0.00330 | 0.00662 | 0.00560 | 0.00527 | 0.00657 |
| Br | -0.00532 | -0.00455 | 0.00139 | -0.00410 | 0.00415 | 0.00681 | 0.00550 | 0.00479 |
| I | -0.00635 | -0.00609 | 0.01495 | -0.01134 | -0.00775 | 0.00656 | 0.00562 | -0.00116 |

[^0]TABLE 5: Optimized atomic SA coeffecients (Ci values) ${ }^{\text {a }}$ obtained using Generalized Born (GB) derived $\mathrm{G}_{\text {polar }}$ energies.

| Type | Gast | MMFF94 | AM1BCC | AM1CM2 | PM3CM2 | MSK | RESP | ChelpG |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Hc | -0.00049 | -0.00146 | 0.00129 | 0.00719 | 0.00588 | 0.00489 | 0.00484 | 0.00436 |
| Ho | 0.02765 | -0.07826 | 0.25937 | 0.13669 | 0.12418 | 0.36583 | 0.35732 | 0.35663 |
| Hs | 0.29006 | 0.23719 | 0.30314 | -0.48392 | -0.40299 | 0.07870 | 0.09415 | 0.13314 |
| Hn | -0.04816 | -0.02950 | -0.02299 | -0.02113 | -0.02386 | -0.01374 | -0.01218 | -0.01520 |
| Hp | -0.07575 | 0.06191 | 0.07913 | 0.54968 | 0.46106 | 0.01841 | 0.03414 | 0.02993 |
| C | -0.01537 | -0.01529 | 0.01715 | 0.03967 | 0.03379 | 0.02164 | 0.01859 | 0.00328 |
| N | 0.01065 | 0.00709 | 0.10707 | 0.07294 | 0.10361 | 0.06938 | 0.06810 | 0.07333 |
| O | 0.01100 | 0.04920 | 0.02952 | 0.09624 | 0.09423 | 0.04269 | 0.03965 | 0.04760 |
| F | 0.02353 | 0.02559 | 0.02941 | 0.02826 | 0.02085 | 0.02082 | 0.01948 | 0.02374 |
| P | 1.50762 | -0.30940 | 0.71401 | -1.75879 | -2.78904 | 0.47251 | 0.25635 | 0.40528 |
| S | 0.01889 | 0.03237 | 0.05437 | 0.15530 | 0.13185 | 0.04452 | 0.04131 | 0.03165 |
| Cl | 0.00536 | 0.00489 | 0.00662 | 0.00515 | 0.00913 | 0.00878 | 0.00784 | 0.00787 |
| Br | -0.00329 | -0.00275 | 0.00466 | -0.00130 | 0.00710 | 0.01492 | 0.01301 | 0.00786 |
| I | -0.00419 | -0.00384 | 0.02054 | -0.00733 | $-0.00400$ | 0.01865 | 0.01703 | 0.00294 |

TABLE 6: Residuals ${ }^{\text {a }}$ for monoatomic ions with experiment.

|  |  | PBSA residuals |  | GBSA residuals |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Ion | $\Delta \mathrm{G}_{\text {hyd }} \operatorname{exptl}^{\text {b }}$ | std constants ${ }^{\text {c }}$ | $\text { fit } C_{i}{ }^{d}$ | std constants | fit $\mathrm{C}_{\text {i }}$ |
| F and F | -107 | 0.92 | $0.67{ }^{\text {e }}$ | 7.77 | $7.20^{\text {e }}$ |
| Cl and $\mathrm{Cl}^{-}$ | -78 | 16.96 | $17.90^{\text {e }}$ | 22.24 | $22.87^{\text {c }}$ |
| Br and $\mathrm{Br}^{-}$ | -72 | 15.07 | $15.98{ }^{\text {e }}$ | 19.5 | $19.41^{\text {e }}$ |
| $I$ and $\mathrm{I}^{-}$ | -63 | 18.19 | $19.08^{\text {e }}$ | 22.04 | $21.29^{\text {e }}$ |
| $\mathrm{Li}^{+}$ | -111 | -22.45 | 0.0 | -17.87 | 0.0 |
| $\mathrm{Na}^{+}$ | -87 | -16.54 | 0.0 | -13.64 | 0.0 |
| $\mathrm{K}^{+}$ | -71 | -13.42 | 0.0 | -11.47 | 0.0 |
| $\mathrm{Mg}^{2+}$ | -437 | -59.13 | 0.0 | -38.77 | 0.0 |
| $\mathrm{Ca}^{2+}$ | -360 | 24.57 | 0.0 | 45.69 | 0.0 |
| $\mathrm{Fe}^{2+}$ | $-440$ | $-3.86$ | $0.0$ | 23.54 | 0.0 |
| $\mathrm{Zn}^{2+}$ | -467 | 1.15 | 0.0 | 35.93 | 0.0 |

${ }^{\text {b }}$ Residuals $=\Delta \mathrm{G}_{\text {hyd }}$ exptl $-\Delta \mathrm{G}_{\text {hyd }}$ calcd, calculated values obtained using eq 1. ${ }^{\text {b }}$ See supporting information Table S1 for experimental references. ${ }^{\mathrm{c}} \mathrm{G}_{\text {nonpolar }}=\left(\mathrm{SA}^{*} 0.00542+0.92\right),{ }^{\mathrm{d}} \mathrm{G}_{\text {nonpolar }}=\sum \mathrm{SA}_{\mathrm{i}}{ }^{*} \mathrm{C}_{\mathrm{i}} . \quad{ }^{\mathrm{C}}$ Constants for ions $\mathrm{F}^{-}, \mathrm{Cl}^{-}, \mathrm{Br}^{-}$, and $\Gamma^{-}$from RESP fits, Tables 4 and 5.

As averaged over the entire dataset of 502 molecules, the magnitude and sign for each $\mathrm{C}_{\mathrm{i}}$ value can give some indication as to the error with experiment (and direction) associated with a particular charge model for a given atom (mbondi) type. However, caution should be exercised when trying to ascribe too much physical significance to any given SA coefficient. For some atom types listed in Table 1, HS (N=13), P (N=6), and I $(\mathrm{N}=12)$, a lack of experimental data could potentially lead to SA optimizations that are underdetermined. Nevertheless, given that related charge methods such as

AM1CM2/PM3CM2 or MSK/RESP often yield similar fitted SA constants (Table 5 and 6), the multiple linear regression results appear to be robust. As an example, phosphorus (mbondi type P) coefficients from GB fits for AM1CM2 and PM3CM2 charged compounds are relatively large in magnitude compared with other types (Tables 5 and 6). Here, the negative coeffecient are always in the range -1.8 to -2.8 . The large negative sign indicates that on average $G_{\text {polar }}$ terms computed using AM1CM2 and PM3CM2 charges underestimate the experimental $\Delta \mathrm{G}_{\mathrm{hyd}}$ values. Nonpolar contributions computed using atom-based $\mathrm{SA}_{\mathrm{i}}$ will yield a favorable free energy to correct for this underestimation given that SA is always a positive value and, in this case, the $\mathrm{C}_{\mathrm{i}}$ for P atoms are negative. On the other hand, the $G B C_{i}$ coefficients for atom types $P$ for $a b$ initio-based methods (MSK, RESP, and ChelpG) are positive and in the much smaller at about 0.26 to 0.47 . The variation in the optimized coefficients in Table $4-5$ are a direct result of the differences that are obtained from the different partial charge methods used for computation of $\mathrm{G}_{\mathrm{polar}}$. Because of this fact, these optimized constants can be viewed as a SA-based correction factor to account for errors in any particular charge model in an average sense.

## Monoatomic ions

We have also pursued free energy of hydration calculations for 11 monatomic ions using the same PBSA and GBSA protocols for comparison with experiment (Tables 7). Monoatomic ions are a unique case given that only a "single" atom is present and therefore not charge-model dependant; only the formal charge and radius needs to be specified. In general, nonpolar contributions to the total $\Delta \mathrm{G}_{\text {hyd }}$ for monatomics ions would be negligible given the large polarization energy ( -63 to $-467 \mathrm{kcal} / \mathrm{mol}$, Table 6)
compared to the small solvent accessible surface area. The solvent accessible surface area for a monatomic species is simply SA $=4 \pi(r+1.4)^{2}$ where $1.4 \AA$ represents the standard probe radius for water and $r$ is the radii. A number of prior studies have concluded that for monatomics, different radii should be used for anions versus cations if using a simple Born model of ion hydration. REF Our goal was to evaluate the recommended mbondi values for ions by testing agreement with experiment. It should be noted that atom types $\mathrm{F}, \mathrm{Cl}, \mathrm{Br}$, and I may exist as single ions or as part of a polyatomic species (neutral or charged). In this report we have considered only one type of radius for both bound and unbound elements.

TABLE 7: Optimized atomic SA coeffecients ( Ci values) ${ }^{\text {a }}$ for single ions.

| Ion | PB | GB |
| :--- | ---: | :---: |
| $\mathrm{Li}^{+}$ | -0.16017 | -0.12495 |
| $\mathrm{Na}^{+}$ | -0.08693 | -0.06984 |
| $\mathrm{~K}^{+}$ | -0.05223 | -0.04323 |
| $\mathrm{Mg}^{2+}$ | -0.46716 | -0.30181 |
| $\mathrm{Ca}^{2+}$ | 0.21647 | 0.39131 |
| $\mathrm{Fe}^{2+}$ | -0.02233 | 0.23696 |
| $\mathrm{Zn}^{2+}$ | 0.02654 | 0.38155 |
| ${ }^{2} \mathrm{G}_{\text {nonpolar }}=\sum \mathrm{SA}_{\mathrm{i}}{ }^{*} \mathrm{C}_{\mathrm{i}}$. |  |  |
|  |  |  |

Free energy of hydration results for the 11 monoatomics are presented in Table 6. The experimental and difference in the computed values with experiment (residuals $=$ $\Delta \mathrm{G}_{\text {hyd }}$ exptl $-\Delta \mathrm{G}_{\text {hyd }}$ calcd) are shown for PBSA and GBSA results using both standard and fitted SA constants. As emphasized above, SA contributions to hydration free energies for monoatomic species are assumed to be small. For this reason, the fitted constants for $\mathrm{Li}, \mathrm{Na}, \mathrm{K}, \mathrm{Mg}, \mathrm{Ca}, \mathrm{Fe}$, and Zn , are in effect a correction factor that (1) accounts for deficiencies in the implicit hydration model, (2) non-optimal radii, or (3) differences in polarization energies between PB and GB calculations.


It should be noted that all PB calculations in this work originally employed a grid spacing of $0.5 \AA$ (scale $=2$ grids $/ \AA$ ) as was reported in numerous prior MM-PBSA studies.(Chong et al., 1999; Gouda et al., 2003; Huo et al., 2002; Kuhn and Kollman, 2000a, b; Massova and Kollman, 1999, 2000; Masukawa et al., 2003; Wang et al., 2001b) However, poor $\Delta \mathrm{G}_{\text {hyd }}$ results for monoatomic ions computed using PB results obtained with 2 grids $/ \AA$ prompted us to redo all the calculations reported here using a much finer grid spacing of $0.25 \AA$ (scale $=4$ grids $/ \AA$ ). Figure 6 highlights the fact that the PB results are highly dependent on the grid spacing resolution used. Such convergence behavior of PB was also noted in the original MM-PBSA/GBSA study.(Srinivasan et al., 1998) The GB method used here does not use a grid and therefore the results will not affected by grid-based convergence issues as in $\mathrm{PB} ; \mathrm{GB} \mathrm{G}_{\text {polar }}$ results show good accord with experimental $\Delta \mathrm{G}_{\text {hyd }}$ (Figure 6) for the monoatomic ions. Given that current computational limits prohibit the use of PB grids much finer than $\sim 0.5 \AA$ (scale $=2$ grids $/ \AA$ ) for protein-ligand binding calculations, the lack of convergence for small highly charged species noted here remains an issue.

FIGURE 6: Polar energies ( $\square$ Gpolar) for monoatomic ions vs experiment ( $\square$ Ghyd exptl). GB and PB calculations using either 2.0 or 4.0 grids $/ \AA$ are shown for mono, di, and tri-valent species.


It should be emphasized that PB results obtained using a given grid spacing might lead to reasonable cancellation of errors for relative $\Delta \Delta \mathrm{G}_{\text {hyd }}$ calculations. If absolute agreement with experiment is important, an alternative approach would be to employ optimized atom-based SA coefficients as reported in this work (Tables 4,5, and 7) to correct for low-resolution PB results in order to yield $\Delta \mathrm{G}_{\text {hyd }}$ values in better absolute agreement with experiment.

## CONCLUSIONS

Absolute free energies of hydration have been estimated using continuum PBSA and GBSA methods for comparison with experiment (Table S1 supporting information) for 460 neutral compounds, 42 polyatomic ions, and 11 monoatomic ions. A systematic evaluation of eight different models have revealed that continuum results which employ partial charge based on one of three ab initio methods (MSK, RESP, and ChelpG) consistently lead to better agreement with experiment for neutral and charged species
-
$\qquad$
(Table 2, Figure 2). Use of semiempirical (AM1BCC, AM1CM2, PM3CM2) and empirical (Gast, MFF94) charge schemes yielded mixed results dependant on whether the compounds were charged or neutral.

The results presented here clearly show that correlations with experimental $\Delta \mathrm{G}_{\text {hyd }}$ are independant of which implicit solvation model (PBSA or GBSA) is employed in the calculations. In all cases, the Hawkins pair-wise GB results are strongly correlated ( $\mathrm{r}^{2}=$ 0.94 ) with the much more expensive Delphi PB calculations provided that identical coordinates, radii, and atomic charges are used (Figure 2). It should be noted that calculations for monoatomic ions revealed that the PB results are highly dependant on the grid spacing used in the computations (Figure 3) and should therefore be closely monitored if absolute free energies of hydration be required.

Examination of polar and non-polar energy components revealed that $\Delta \mathrm{G}_{\text {nonpolar }}$ energies derived from molecule-based SA's with standard conversion constants have no correlation with experimental results (Figure 3). The problem stems from the erroneous assumption that all exposed atoms will contribute equally to the non-polar energy term. To remedy this fact, re-optimization of SA's constants on an atom-type basis was pursued through multiple linear regression fittings to the difference in experimental free energies and polar energy terms obtained from continuum calculations. For completeness these optimizations were performed for all eight charge models and both continuum methods (Tables 4, 5, and 7). Use of atom-based $\mathrm{SA}_{\mathrm{i}}$ instead of molecule-based SA constants reduces both relative $\left(\mathrm{r}^{2}\right)$ and absolute unsigned errors (AUE) with experiment with respect to experiment by eliminating any gross deficiencies a particular charge model may have (Tables 2 vs 3 ). In particular, AUE errors for charged species were

examarsw =rnel
substantially reduced (Tables 3 and 6). On a case-by-case basis use of atom-based constants (Figures 4-5, black squares) correct for systematic errors. In the case of monatomic ions atom-based constants may be primarily viewed as correcting for deficiencies in the implicit model or to correct for non-optimal radii. As noted above, convergence of PB for monatomic ions can be problematic.

Levy et al. have recently reported a novel non-polar method which represent solvent-solute van der Waals interactions more explicitly than does molecule-based SA to $\Delta \mathrm{G}_{\text {hyd }}$ to. However their calculation are more difficult and the parameters can not be applied universally, but rather require individual calculations for molecules. Our approach is an in between solution which recognize the inadequacy of a universal SA constant and show that by optimizing the constants to individual atom types rather than molecules we can get enrichment in the correlation between experimental and computational methods with out any added time cost.

This is important in larges scale drug discovery efforts and such improvements are expected to be important for protein-ligand binding calculations which include $\Delta \mathrm{G}_{\text {hyd }}{ }^{-}$ like terms (e.g. MM-GBSA and MM-PBSA methods) and could help to facilitate the discovery and design of novel chemotherapeutic compounds.

In this report we have emphasized two distinct types of errors, For neutral species error estimates for experimental values are approximately $0.5 \mathrm{kcal} / \mathrm{mol}$ but for ions are 5 $\mathrm{kcal} / \mathrm{mol}$. REF Cramer/Trulhar papers

IR


$$
]
$$

## ACKNOWLEDGMENTS

Gratitude is expressed to the Department of Defense for support of this research (Award
Number DAMD17-00-1-0192, Modification P00001). TA was supported by the UC president's dissertation year fellowship

81 U
$i_{0}$
$A R i_{6}$
$A R$
$42 n$
2

## REFERENCES

2002a. AMBER Version 7. San Francisco, CA: University of California at San Francisco.
2002b. MOE Version 2002.03. Montreal, Canada: Chemical Computing Group.
2003. DMS. San Francisco, CA: University of California at San Francisco Computer Graphics Laboratory.

Abraham, M.H., J. Andonianhaftvan, G.S. Whiting, A. Leo, and R.S. Taft. 1994. Hydrogen Bonding. Part 34. The Factors That Influence the Solubility of Gases and Vapours in Water at 298 K , and a New Method for Its Determination. Journal of the Chemical Society-Perkin Transactions 2(8):1777-1791.

Abraham, M.H., G.S. Whiting, R. Fuchs, and E.J. Chambers. 1990. Thermodynamics of Solute Transfer from Water to Hexadecane. Journal of the Chemical SocietyPerkin Transactions 2(2):291-300.

Bayly, C.I., P. Cieplak, W.D. Cornell, and P.A. Kollman. 1993. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints For Deriving Atomic Charges - the RESP Model. Journal of Physical Chemistry 97(40):1026910280.

Besler, B.H., K.M. Merz, and P.A. Kollman. 1990. Atomic Charges Derived from Semiempirical Methods. Journal of Computational Chemistry 11(4):431-439.

Bordner, A.J., C.N. Cavasotto, and R.A. Abagyan. 2002. Accurate Transferable Model for Water, n-Octanol, and n-Hexadecane Solvation Free Energies. J. Phys. Chem. B 106:11009-11015.

Breneman, C.M., and K.B. Wiberg. 1990. Determining Atom-Centered Monopoles from Molecular Electrostatic Potentials - the Need for High Sampling Density in Formamide Conformational-Analysis. Journal of Computational Chemistry 11(3):361-373.

Chambers, C.C., G.D. Hawkins, C.J. Cramer, and D.G. Truhlar. 1996. Model for aqueous solvation based on class IV atomic charges and first solvation shell effects. Journal of Physical Chemistry 100(40):16385-16398.

Chong, L.T., Y. Duan, L. Wang, I. Massova, and P.A. Kollman. 1999. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. Proc Natl Acad Sci USA 96(25):14330-5.

Cornell, W.D., P. Cieplak, C.I. Bayly, and P.A. Kollman. 1993. Application of Resp Charges to Calculate Conformational Energies, Hydrogen-Bond Energies, and Free-Energies of Solvation. Journal of the American Chemical Society 115(21):9620-9631.

Cramer, C.J., and D.G. Truhlar. 1999. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. Chem Rev 99(8):2161-2200.

Donini, O.A.T., and P.A. Kollman. 2000. Calculation and prediction of binding free energies for the matrix metalloproteinases. Journal of Medicinal Chemistry 43(22):4180-4188.

Eisenberg, D., and A.D. McLachlan. 1986. Solvation Energy in Protein Folding and Binding. Nature 319(6050):199-203.

Feig, M., A. Onufriev, M.S. Lee, W. Im, D.A. Case, and C.L. Brooks, 3rd. 2004. Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. J Comput Chem 25(2):265-84.

Frisch, M.J., G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, V.G. Zakrzewski, J.A. Montgomery, Jr., R.E. Stratmann, J.C. Burant, S. Dapprich, J.M. Millam, A.D. Daniels, K.N. Kudin, M.C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G.A. Petersson, P.Y. Ayala, Q. Cui, K. Morokuma, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J. Cioslowski, J.V. Ortiz, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P.M.W. Gill, B.G. Johnson, W. Chen, M.W. Wong, J.L. Andres, M. Head-Gordon, E.S. Replogle, and J.A. Pople. 1998. Gaussian 98, revision A.9. Pittsburgh PA: Gaussian Inc.

Gasteiger, J., and M. Marsili. 1980. Iterative Partial Equalization of Orbital Electronegativity - a Rapid Access to Atomic Charges. Tetrahedron 36(22):32193228.

Gerber, P.R. 1998. Charge distribution from a simple molecular orbital type calculation and non-bonding interaction terms in the force field MAB. J Comput Aided Mol Des 12(1):37-51.

Gouda, H., I.D. Kuntz, D.A. Case, and P.A. Kollman. 2003. Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods. Biopolymers 68(1):16-34.

Halgren, T.A. 1996. Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94. Journal of Computational Chemistry 17(5-6):490519.

Hawkins, G.D., C.J. Cramer, and D.G. Truhlar. 1995. Pairwise Solute Descreening of Solute Charges from a Dielectric Medium. Chemical Physics Letters 246(1-2):122-129.

Hawkins, G.D., C.J. Cramer, and D.G. Truhlar. 1996. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. Journal of Physical Chemistry 100(51):19824-19839.

Hawkins, G.D., D.J. Giesen, G.C. Lynch, C.C. Chambers, I. Rossi, J.W. Storer, J. Li, P. Winget, D. Rinaldi, D.A. Liotard, C.J. Cramer, D.G. Truhlar, D.A. based in part on AMPAC version 2.1 by Liotard, E.F. Healy, J.M. Ruiz, M.J.S. Dewar, and F. and on EF by Jensen. 1999. AMSOL version 6.6. Minneapolis, Minnesota: University of Minnesota.

Huo, S., J. Wang, P. Cieplak, P.A. Kollman, and I.D. Kuntz. 2002. Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: insight into structure-based ligand design. $J$ Med Chem 45(7):1412-9.

Jakalian, A., B.L. Bush, D.B. Jack, and C.I. Bayly. 2000. Fast, efficient generation of high-quality atomic charges. AM1- BCC model: I. Method. Journal of Computational Chemistry 21(2):132-146.

Jakalian, A., D.B. Jack, and C.I. Bayly. 2002. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J Comput Chem 23(16):1623-41.

Jorgensen, W.L. 1989. Free Energy Calculations: A Breakthrough For Modeling Organic Chemistry in Solution. Accounts of Chemical Research 22(5):184-189.

Jorgensen, W.L., and C. Ravimohan. 1985. Monte Carlo Simulation of Differences in Free Energies of Hydration. Journal of Chemical Physics 83(6):3050-3054.

Jorgensen, W.L., and J. Tirado-Rives. 1988. The Opls Potential Functions For Proteins Energy Minimizations For Crystals of Cyclic-Peptides and Crambin. Journal of the American Chemical Society 110(6):1657-1671.

Kollman, P. 1993. Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. Chemical Reviews 93(7):2395-2417.

Kollman, P.A., I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, and T.E. Cheatham. 2000. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res 33(12):889-97.

Kuhn, B., and P.A. Kollman. 2000a. Binding of a diverse set of ligands to avidin and streptavidin: An accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. Journal of Medicinal Chemistry 43(20):3786-3791.

Kuhn, B., and P.A. Kollman. 2000b. A ligand that is predicted to bind better to avidin than biotin: Insights from computational fluorine scanning. Journal of the American Chemical Society 122(16):3909-3916.

Leo, A., C. Hansch, and D. Elkins. 1971. Partition Coefficients and Their Uses. Chemical Reviews 71(6):525-616.

Li, J.B., T.H. Zhu, C.J. Cramer, and D.G. Truhlar. 1998. New class IV charge model for extracting accurate partial charges from wave functions. Journal of Physical Chemistry A 102(10):1820-1831.

Li, J.B., T.H. Zhu, G.D. Hawkins, P. Winget, D.A. Liotard, C.J. Cramer, and D.G. Truhlar. 1999. Extension of the platform of applicability of the SM5.42R universal solvation model. Theoretical Chemistry Accounts 103(1):9-63.

Lipinski, C.A., F. Lombardo, B.W. Dominy, and P.J. Feeney. 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46(1-3):3-26.

Marcus, Y. 1991. Thermodynamics of Solvation of Ions, Part 5. Gibbs Free Energy of Hydration at 298.15 K. Journal of the Chemical Society-Faraday Transactions 87(18):2995-2999.

Marcus, Y. 1994. A Simple Empirical Model Describing the Thermodynamics of Hydration of Ions of Widely Varying Charges, Sizes, and Shapes. Biophysical Chemistry 51(2-3):111-127.

Massova, I., and P.A. Kollman. 1999. Computational alanine scanning to probe proteinprotein interactions: A novel approach to evaluate binding free energies. Journal of the American Chemical Society 121(36):8133-8143.

Massova, I., and P.A. Kollman. 2000. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. Perspectives in Drug Discovery and Design 18:113-135.

Masukawa, K.M., P.A. Kollman, and I.D. Kuntz. 2003. Investigation of NeuraminidaseSubstrate Recognition Using Molecular Dynamics and Free Energy Calculations. Journal of Medicinal Chemistry 46:5628-5637.

Ooi, T., M. Oobatake, G. Nemethy, and H.A. Scheraga. 1987. Accessible Surface-Areas as a Measure of the Thermodynamic Parameters of Hydration of Peptides. Proceedings of the National Academy of Sciences of the United States of America 84(10):3086-3090.

Rizzo, R.C., S. Toba, and I.D. Kuntz. 2004. A Molecular Basis for the Selectivity of Thiadiazole Urea Inhibitors with Stromelysin-1 and Gelatinase-A from Generalized Born Molecular Dynamics Simulations. J Med Chem 47:3065-3074.

Rocchia, W., E. Alexov, and B. Honig. 2001. Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. Journal of Physical Chemistry B 105(28):6507-6514.

Rocchia, W., S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera, and B. Honig. 2002. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. J Comput Chem 23(1):128-37.

Sitkoff, D., K.A. Sharp, and B. Honig. 1994. Accurate Calculation of Hydration FreeEnergies Using Macroscopic Solvent Models. Journal of Physical Chemistry 98(7):1978-1988.

Srinivasan, J., T.E. Cheatham, P. Cieplak, P.A. Kollman, and D.A. Case. 1998. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate DNA helices. Journal of the American Chemical Society 120(37):9401-9409.

Stewart, J.J.P., I. Rossi, W.-P. Hu, G.C. Lynch, Y.-P. Liu, Y.-Y. Chuang, J. Li, C.J. Cramer, P.L. Fast, and D.G. Truhlar. 1999. MOPAC version 5.09 mn . Minneapolis, Minnesota: University of Minnesota.

Still, W.C., A. Tempczyk, R.C. Hawley, and T. Hendrickson. 1990. Semianalytical Treatment of Solvation For Molecular Mechanics and Dynamics. Journal of the American Chemical Society 112(16):6127-6129.

Suenaga, A., M. Hatakeyama, M. Ichikawa, X. Yu, N. Futatsugi, T. Narumi, K. Fukui, T. Terada, M. Taiji, M. Shirouzu, S. Yokoyama, and A. Konagaya. 2003. Molecular dynamics, free energy, and SPR analyses of the interactions between the SH2 domain of Grb2 and ErbB phosphotyrosyl peptides. Biochemistry 42(18):5195200.

Tsui, V., and D.A. Case. 2000a. Molecular dynamics simulations of nucleic acids with a generalized born solvation model. Journal of the American Chemical Society 122(11):2489-2498.

Tsui, V., and D.A. Case. 2000b. Theory and applications of the generalized Born solvation model in macromolecular simulations. Biopolymers 56(4):275-91.

Wang, J., P. Morin, W. Wang, and P.A. Kollman. 2001a. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MMPBSA. J Am Chem Soc 123(22):5221-30.

Wang, W., W.A. Lim, A. Jakalian, J. Wang, R. Luo, C.I. Bayly, and P.A. Kollman. 2001b. An analysis of the interactions between the Sem-5 SH3 domain and its ligands using molecular dynamics, free energy calculations, and sequence analysis. J Am Chem Soc 123(17):3986-94.
APPENDIX

 ethene propene
but-1-ene pent-1-ene (Z)-2-pentene 3-methyl-1-butene 2-methylbut-2-ene hex-1-ene 2-methylpent-1-ene 1-heptene (E)-2-heptene 1 -octene 1-nonene buta-1,3-diene 2-methylbuta-1,3-diene
 penta-1,4-diene hexa-1,5-diene cyclopentene cyclohexene 1-methylcyclohexene cyclohepta-1,3,5-triene propyne 1-butyne

 | $\stackrel{0}{5}$ |
| :--- |
| $\stackrel{0}{0}$ |
| I |


Table S1. Experimental Free Energies of Hydration $\left(\Delta \mathrm{G}_{\text {hyd }}\right)$ in $\mathrm{kcal} / \mathrm{mol}^{\mathrm{a}}$
y⿹弓ū



[^1]


| $123-38-6$ | -3.43 | 166 | n-propyl_acetate |
| :--- | :--- | :--- | :--- |
| $123-72-8$ | -3.18 | 167 | isopropyl_acetate |
| $110-62-3$ | -3.03 | 168 | n-butyl_acetate |
| $66-25-1$ | -2.81 | 169 | isobutyl_acetate |
| $111-71-7$ | -2.67 | 170 | n-pentyl_acetate |
| $124-13-0$ | -2.29 | 171 | isoamyl_acetate |
| $124-19-6$ | -2.07 | 172 | n-hexyl_acetate |
| $4170-30-3$ | -4.22 | 173 | methyl_propanoate |
| $6728-26-3$ | -3.68 | 174 | ethyl_propanoate |
| $2548-87-0$ | -3.43 | 175 | n-propyl_propanoate |
| $67-64-1$ | -3.80 | 176 | n-pentyl_propanoate |
| $78-93-3$ | -3.71 | 177 | methyl_butanoate |
| $107-87-9$ | -3.52 | 178 | ethyl_butanoate |
| $96-22-0$ | -3.41 | 179 | n-propyl_butyrate |
| $563-80-4$ | -3.24 | 180 | methyl_pentanoate |
| $591-78-6$ | -3.28 | 181 | ethyl_pentanoate |
| $108-10-1$ | -3.05 | 182 | methyl_hexanoate |
| $110-43-0$ | -3.04 | 183 | ethyl_hexanoate |
| $123-1-3$ | -2.92 | 184 | isobutyl_isobutanoate |
| $111-13-7$ | -2.88 | 186 | propanenitrile |
| $821-55-6$ | -2.49 | 187 | butanenitrile |
| $502-56-7$ | -2.64 | 188 | pentanenitrile |
| $693-54-9$ | -2.34 | 189 | ammonia |
| $112-12-9$ | -2.15 | 190 | methylamine |
| $120-92-3$ | -4.70 | 191 | ethylamine |
| $108-94-1$ | -4.91 | 192 | n-propylamine |
| $107-31-3$ | -2.78 | 193 | n-butylamine |
| $109-94-4$ | -2.56 | 194 | n-pentylamine |
| $110-7-7$ | -2.48 | 195 | n-hexylamine |
| $625-55-8$ | -2.02 | 196 | n-heptylamine |
| $542-55-2$ | -2.22 | 197 | n-octylamine |
| $110-45-2$ | -2.13 | 198 | cyclohexylamine |
| $79-20-9$ | -3.13 | 199 | dimethylamine |
| $141-78-6$ | -2.94 | 200 | diethylamine |

[^2]
$142-84-7$
$108-18-9$
$11-92-2$
$75-50-3$
$121-44-8$
$75-52-5$
$7-24-3$
$1-8-03-2$
$79-46-9$
$627-5-4$
$628-05-7$
$119-4-9$
$68-12-2$
$64-19-7$
$79-09-4$
$107-92-6$
$109-52-4$
$503-74-2$
$142-62-1$
$7732-18-5$
$67-56-1$
$64-17-5$
$71-23-8$
$67-63-0$
$71-3-3$
$78-83-1$
$78-92-2$
$75-65-0$
$71-41-0$
$6032-29-7$
$584-02-1$
$137-32-6$
$123-51-3$
$75-85-4$


$541-73-1$
$106-46-7$
$87-61-6$
$120-82-1$
$108-70-3$
$634-66-2$
$63-9-2-2$
$95-94-3$
$95-49-8$
$108-86-1$
$106-38-7$
$59-50-4$
$100-66-3$
$103-73-1$
$100-52-7$
$104-7-0$
$98-86-2$
$122-00-9$
$93-58-3$
$43-89-0$
$100-47-0$
$95-53-4$
$106-49-0$
$87-62-7$
$95-51-2$
$108-42-9$
$106-47-8$
$90-04-0$
$536-90-3$
$10-44-9$
$88-74-4$
$99-09-2$
$100-01-6$
$134-32-7$

| $95-47-6$ | -0.90 | 304 | 1,3 -dichlorobenzene |
| :--- | :--- | :--- | :--- |
| $108-38-3$ | -0.83 | 305 | 1,4 -dichlorobenzene |
| $106-42-3$ | -0.80 | 306 | $1,2,3$-trichlorobenzene |
| $103-65-1$ | -0.53 | 307 | 1,2, -trichlorobenzene |
| $98-82-8$ | -0.30 | 308 | $1,3,5$-trichlorobenzene |
| $526-73-8$ | -1.21 | 309 | $1,2,3,4$-tetrachlorobenzene |
| $95-63-6$ | -0.86 | 310 | $1,2,3,5$-tetrachlorobenzene |
| $108-67-8$ | -0.90 | 311 | $1,2,4,5$-tetrachlorobenzene |
| $611-14-3$ | -1.04 | 312 | 2-chlorotoluene |
| $622-96-8$ | -0.95 | 313 | bromobenzene |
| $104-51-8$ | -0.40 | 314 | 4-bromotoluene |
| $538-93-2$ | 0.16 | 315 | iodobenzene |
| $135-98-8$ | -0.45 | 316 | anisole |
| $98-06-6$ | -0.44 | 317 | ethyl_phenyl_ether |
| $99-87-6$ | -0.68 | 318 | benzaldehyde |
| $538-68-1$ | -0.23 | 319 | 4-methylbenzaldehyde |
| $1077-16-3$ | -0.04 | 320 | acetophenone |
| $100-42-5$ | -1.24 | 321 | 4-methylacetophenone |
| $98-83-9$ | -1.24 | 322 | methyl_benzoate |
| $92-52-4$ | -2.66 | 323 | ethyl_benzoate |
| $91-20-3$ | -2.40 | 324 | benzonitrile |
| $90-12-0$ | -2.44 | 325 | o-toluidine |
| $575-41-7$ | -2.47 | 326 | p-toluidine |
| $571-58-4$ | -2.82 | 327 | 2,6-dimethylaniline |
| $581-40-8$ | -2.78 | 328 | 2-chloroaniline |
| $581-42-0$ | -2.63 | 329 | 3-chloroaniline |
| $1127-76-0$ | -2.40 | 330 | 4-chloroaniline |
| $496-11-7$ | -1.46 | 331 | 2-methoxyaniline |
| $83-32-9$ | -3.15 | 332 | 3-methoxyaniline |
| $86-73-7$ | -3.35 | 333 | 4-methoxyaniline |
| $462-06-6$ | -0.80 | 334 | 2-nitroaniline |
| $98-08-8$ | -0.25 | 335 | 3-nitroaniline |
| $108-90-7$ | -1.12 | 336 | 4-nitroaniline |
| $95-50-1$ | -1.36 | 337 | 1-naphthylamine |

[^3]




2-naphthylamine
N -methylaniline
N,N-dimethylaniline
nitrobenzene
2-nitrotoluene
3-nitrotoluene
benzamide
phenol
o-cresol
p-cresol
2,3-dimethylphenol
2,4-dimethylphenol
2,5-dimethylphenol
2,6-dimethylphenol
3,4-dimethylphenol
3,5-dimethylphenol
3-ethylphenol
4-ethylphenol
4-n-propylphenol
4-tert-butylphenol
2-fluorophenol
4-fluorophenol
2-chlorophenol
3-chlorophenol
4-chlorophenol
4-chloro-3-methylpheno
4-bromophenol
2-iodophenol
2-methoxyphenol
3-methoxyphenol
3-hydroxybenzaldhyde
4-hydroxybenzaldhyde
3-cyanophenol
4-cyanophenol 338
339
340
341
342
343
344
345
344
347
348
349
345
351
352
353
354
355
336
357
358
359
360
361
362
363
364
365
366
367
368
369
397
371




2-isobutylpyrazine
thiophene
2-methylthiophene
dimethyl_ether
methyl_ethyl_ether
methl_-tert-butyl_ether
3,3-dimethylbutan-2-one
2,4-dimethylpentan-3-one
methyl_cyclopropyl_ketone
methyl_cyclohexyl_ketone
methyl_trimethylacetate
methyl_cyclopropylcarboxylate
methyl_cyclohexylcarboxylate
dimethyl_sulfide
methyl_ethyl_sulfide
dimethyl_sulfoxide
4-methoxyacetophenone
anthracene
phenanthrene
pyrene
2-methylpropene
1,2-ethanediol
m-cresol
methyl-propyl-ether
methyl-isopropyl-ether
1,2-dimethoxyethane
methyl-octanoate
azetidine
pyrrolidine
piperazine
N-methylpiperazine
piperidine
aniline
ethanamide


| -101.00 | 1041 | Br - | -72.00 |
| :---: | :---: | :---: | :---: |
| -73.00 | 1042 | I- | -63.00 |
| -78.00 | 1044 | PH4+ | -73.00 |
| -70.00 | 1045 | CH3PH3+ | -63.00 |
| -66.00 | 1046 | (CH3)2PH2+ | -57.00 |
| -59.00 | 1047 | (CH3)3PH+ | -53.00 |
| -64.00 | 1048 | H2PO4- | -68.00 |
| -58.00 | 1049 | HCO2- | -94.00 |
| -68.00 | 1051 | H3S+ | -87.00 |
| -81.00 | 1052 | Li+ | -111.00 |
| -75.00 | 1053 | $\mathrm{Na+}$ | -87.00 |
| -73.00 | 1054 | K+ | -71.00 |
| -66.00 | 1055 | (CH3)4N+ | -38.00 |
| -74.00 | 1059 | Mg++ | -437.00 |
| -74.00 | 1060 | Ca++ | -360.00 |
| -61.00 | 1061 | Fe++ | -440.00 |
| -76.00 | 1062 | Zn++ | -467.00 |
| -74.00 | 1065 | OCN- | -87.00 |
| -76.00 | ${ }^{\text {a }}$ Codes 1-426 from reference (Abraham et al., 1990), codes 500-534 from reference (Chambers et al., 1996), codes 535-538 from reference (Gerber, 1998), codes 1000-1048 from reference (Li et al., 1999) , codes 1049-1051 from reference (Chambers et al., 1996), codes 1052-1065 from reference (Marcus, 1994). Energies in $\mathrm{kcal} / \mathrm{mol}$. |  |  |
| -65.00 |  |  |  |
| -107.00 |  |  |  |
| -70.00 |  |  |  |
| -78.00 |  |  |  |
| -66.00 |  |  |  |




## Chapter Seven: Concluding Remarks

This thesis offers a brief glimpse of the challenges in protein structure prediction using computational methods. Chapter two deals with the issues faced in comparative modeling in the absence of sequence similarity and highlights the need for improved protein classification systems. In chapters three, four, and five we take a more fundamental approach and introduce a framework for understanding protein structures and the protocols of comparison in a quantitative form. Chapter six, is an effort to improve the parameterizations used in calculating the free energies of binding; a key step in structure base drug design.

Our efforts above were limited not just by sheer computing power, but also by the available knowledge on proteins, specifically their interactions and relationships to one another. Homology detection offers great potential in classifying the large numbers of unknown sequences that have become available from genomics projects. However, as mentioned earlier, the finite number of observed folds and the lack of high enough similarity has lead to the emergence of orphan proteins whose relatives can not be identified computationally. Despite past efforts in utilizing geometric constraints to narrow the search area for suitable sequence/structure pairs these methods have yet to reach their full potential. To date, constraints have been used in conjunction with other methods all which have relied on sequence similarity. Our alternative approach to use restraints as the primary metric were hampered by the lack of clear guidelines and definitions for structure similarity.

Our knowledge of the protein universe is derived mainly from experimental data. The nature of much of this data is however often not understood. In particular, there are no analytical methods for quantifying the information they offer nor for assessing the
inherent errors contained within them. In addition, proteins amenable to experimental study represent only a subset of the entire universe. As a result, it is difficult to draw conclusions about the properties of the entire protein universe.

We applied the basic tenets of information theory to reduced representations such as two-dimensional lattice models and limited alphabets to study exhaustive sets whose properties can be transferred on to the protein universe. Currently, exhaustive studies are limited to short chain length of no more than twenty residues, however we see that stochastic sets of longer length exhibit similar behavior and as such it is reasonable to assume that conclusions from these models can be extended to world of real proteins and sequences as well.

The end-game of structure-prediction as a discipline is to be able to target proteins responsible for human disease by designing ligands that inhibit their activity through intermolecular interactions. The field of computational chemistry at present is born out of a marriage of computing power and the increased availability of protein structure data. As of today, it is not yet possible to represent protein interactions in their entirety in computer simulations; hence the use of approximation methods, or once again reduced representations, in calculations of free energies of binding. We have compiled a rather large test set of molecules to use in optimizing the input parameters of these algorithms in order to increase their correlation with experimental results.

Computers are playing larger and larger roles in biology. Their purpose and utility appears circular in that with greater computing power we are able to study more complex systems and subsequently enhance our knowledge base, and yet as we increase our
understanding of the laws of molecular recognition, we can design more efficient and reliable algorithms for simulation. The work presented here is a step towards these goals


$$
\begin{aligned}
& \begin{array}{c}
68 \\
88
\end{array}
\end{aligned}
$$


[^0]:    ${ }^{2} \mathrm{G}_{\text {nonpolar }}=\sum \mathrm{SA}_{\mathrm{i}}{ }^{*} \mathrm{C}_{\mathrm{i}}$ optimized using neutral $(\mathrm{N}=460)$ and charged $(\mathrm{N}=42)$ compounds.

[^1]:    1-octyne
    tetrafluoromethane
    chloromethane
    dichloromethane
    trichloromethane
    tetrachloromethane
    chloroethane
    1,1-dichloroethane
    1,2-dichloroethane
    1,1,1-trichloroethane
    1,1,2-trichloroethane
    1,1,2,2-tetrachloroethane
    1,1,1,2-tetrachloroethane
    pentachloroethane
    1-chloropropane
    2-chloropropane
    1,2-dichloropropane
    1,3-dichloropropane
    1-chlorobutane
    2-chlorobutane
    2-chloro-2-methylpropane
    1,4-dichlorobutane
    1-chloropentane
    1-chlorohexane
    1-chloroheptane
    1,1-dichloroethene
    (Z)-1,2-dichloroethene
    (E)-1,2-dichloroethene
    trichloroethene
    tetrachloroethene
    3-chloropropene
    bromomethane
    dibromomethane
    tribromomethane
    

[^2]:    propionaldehyde边 त $=$ $=$ 11 2
    
     0
    む.
    ㅡ․
    을 pentan-2-one pentan-3-one 3-methylbutan-2-one 3-methylbutan-2-one
    hexan-2-one 4-methylpentan-2-one
     octan-2-one
     nonan-5-one decan-2-one undecan-2-one cyclopentanone methyl_formate ethyl_formate n-propyl_formate isopropyl_formate isobutyl_formate isoamyl_formate methyl_acetate
    
    

[^3]:    o-xylene
    m-xylene
    p-xylene
    n-propylbenzene
    isopropylbenzene
    1,2,3-trimethylbenzene
    1,2,4-trimethylbenzene
    1,3,5-trimethylbenzene
    2-ethyltoluene
    4-ethyltoluene
    n-butylbenzene
    isobutylbenzene
    sec-butylbenzene
    tert-butylbenzene
    4-isopropyltoluene
    n-pentylbenzene
    n-hexylbenzene
    styrene
    alpha-methylstyrene
    biphenyl
    naphthalene
    1-methylnapthalene
    1,3-dimethylnapthalene
    1,4-dimethylnaphtalene
    2,3-dimethylnapthalene
    2,6-dimethylnaphtalene
    1-ethylnapthalene
    indane
    acenaphthene
    fluorene
    fluorobenzene
    benzotrifluoride
    chlorobenzene
    1,2-dichlorobenzene
    

