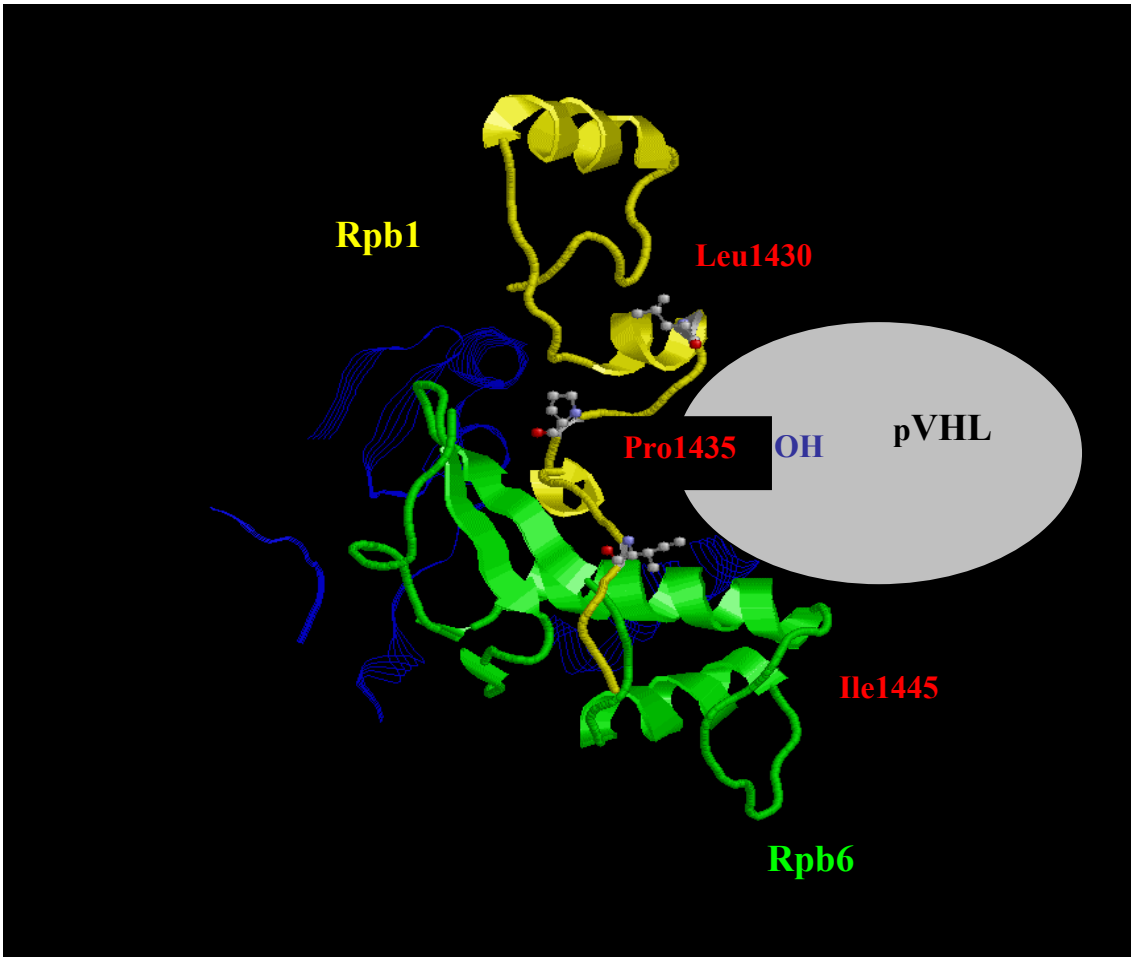
Advances in Protein Structure Prediction:

Algorithms and Applications



Jarosław Meller

Habilitation Thesis

Contents

Preface

Bioinformatics emerged as a separate discipline on the wave of the on-going revolution in biological sciences that has taken place in the last several years. The advent of large-scale sequencing projects such as the Human Genome Project or the whole-genome analysis of gene expression patterns with microarrays as well as numerous other technological breakthroughs generated huge amounts of data to be stored, organized and mined. These challenges required new solutions in terms of databases and their integration and other informatics demands. On the other hand, the need to handle and mine the data in order to generate hypotheses facilitating further biomedical studies triggered a rapid development of theoretical and computational approaches applicable to the problems at hands. Examples of such problems include large-scale sequence assembly from fragments obtained using the so-called *shotgun* approach, gene prediction and annotation, clustering of whole-genome gene expression profiles, large-scale sequence alignment and subsequence search or protein structure prediction.

This dissertation deals with the latter problem. Because of the importance of protein structure and function on one hand and a relatively slow progress in high throughput experimental structural studies, computational protein structure prediction can help to close the gap between large-scale sequencing projects and their actual outcomes in terms of the understanding of molecular machinery of life and mechanisms underlying various disease states. However, despite progress that has been made in the last several years, predicting three-dimensional structure of a protein from its amino acid sequence remains one of the central challenges in computational biology. The problem of predicting protein function is further complicated by the fact that proteins are gregarious and often work together, forming large complexes and interacting with other macromolecules and small ligands.

In this habilitation thesis, several new methods for protein structure prediction and their selected applications are described. These new methods utilize the framework of Linear Programming (LP) for the design of scoring functions for protein recognition. Such optimized scoring functions are then incorporated into effective protocols for sequence-to-structure matching within the framework of sequence alignment and Dynamic Programming (DP) techniques. A novel Maximum Feasibility heuristic for solving infeasible LP problems is presented, allowing one to address the problem of overfitting in the optimization of scoring functions by using LP. A general strategy for obtaining consensus classifiers in the form of linear combination of individual classifiers is proposed, based on a combination of LP and Maximum Feasibility heuristic. This novel strategy is applied to protein membrane domain and secondary structure prediction. Using new algorithms and scoring functions enabled several interesting and highly relevant predictions, as discussed in articles included in this dissertation.

I would like to thank Prof. Ron Elber who guided my initial encounters with the world of proteins and many specific problems discussed in this thesis. I would also like to gratefully acknowledge my other co-authors and collaborators, from whom I learned a great deal about biological systems and the related computer science problems. I truly enjoyed both formal and algorithmic developments as well as fascinating encounters with specific proteins and networks of interactions discussed in this dissertation. I also wish to thank all my teachers and mentors who gave me the fundamentals and ability to absorb quickly new

research challenges. Last but not least, I would like to acknowledge the support from Cincinnati Children's Hospital Research Foundation and the Faculty of Physics, Astronomy and Informatics, Nicholas Copernicus University in Toruń.

PART I:

Introduction

I.1. Overview

In this dissertation we present a description of several algorithms and methods for protein structure prediction and their subsequent applications, published in seven original research articles also included in this thesis. The methods and algorithms presented here are based on Linear Programming (LP) techniques and their novel extensions, such as the Maximum Feasibility (MaxF) heuristic for solving large-scale infeasible LP problems. Using an LP based approach, several scoring functions for protein recognition by sequence-to-structure matching are designed and a novel strategy for obtaining consensus classifiers in the form of a linear combination of individual (weaker) classifiers is proposed.

While some of the methods presented here, such as the MaxF approach for consensus classifiers, are of general applicability, this dissertation focuses on applications to protein structure prediction. Predicting three-dimensional structure of a protein from its amino acid sequence and inferring the functional consequences and possible interactions with other proteins, nucleic acids and membranes is one of the central challenges in the field of computational biology in the post-genomic era. Here, several specific applications to protein structure and protein-protein interaction prediction are discussed.

Using our novel approach to protein recognition by sequence-to-structure matching, as incorporated into the software package and Web server LOOPP [Meller and Elber, 2001, 2002], we were able to make a link between a gene that regulates the size of the tomato fruit and human Ras protein that had been implicated in many types of cancer [Frary et. al., 2000]. This study and our prediction are now mentioned in some textbooks on genomics [e.g. Gibson and Muse, 2002].

Recently, using a combination of our approach and standard computer modeling techniques we were able to make a link between tumor suppressor pVHL and the major transcription enzyme RNA Polymerase II [Kuznetsova, Meller et. al., 2003]. We were also able to identify the binding pocket for the antigen receptors on the surface of the Norwalk-like viruses that cause acute gastroenteritis [Tan, Huang, Meller et. al., 2003]. The first study opens a new avenue in research on the role of hypoxia and pVHL in the regulation of transcription, whereas the latter opens a way to rationally design inhibitors that may prevent common Norwalk virus infections.

The above applications demonstrate the ability of novel methods presented here to provide, based on structural predictions and insights, crucial hypotheses facilitating greatly further experimental studies on highly relevant molecular systems. Perhaps more importantly, however, the LP based approach for the design of scoring functions for sequence-to-structure matching offers a guidance as to how to further improve such scoring functions (also called folding potentials throughout this thesis and papers included here) and how to choose their optimal functional form, as discussed in details in Part II.

In regard to LP and MaxF based strategies for obtaining accurate consensus classifiers, we would like to comment that this approach, due to efficiency and scalability of *interior point* methods for LP, can be applied to very large classification problems involving tens of millions of data points in high dimensional feature spaces. This is one of the critical characteristics of algorithms to be applied to large-scale classification problems that are likely to arise in the near future in genomics and proteomics, for example in the context of expected massive screening of gene expression profiles with microarrays.

The first part of this dissertation contains an introduction to computational problems and applications considered here, presented in a rather informal way meant to introduce the basic biological concepts and algorithmic issues. Following the overview, the protein folding problem and different approaches for protein structure prediction, including protein recognition by sequence-to-structure matching, are briefly discussed. Next, central conceptual and algorithmic issues in the context of the presented extensions and applications of Linear Programming (LP) and Dynamic Programming (DP) techniques to protein structure prediction are discussed. The Maximum Feasibility heuristic for infeasible LP problems is defined and its merits and applications to both: the design of scoring functions for sequence-to-structure matching and consensus classifiers, are presented. Finally, structural predictions pertaining to the tomato fruit size gene fw2.2, interaction between RNA Polymerase II and pVHL, and structural determinants of Norwalk-like viruses binding to antigen receptors are summarized, followed by closing remarks.

The remaining sections include several original research papers covering in details both: methodological aspects (Part II) and applications (Part III) discussed in Part I. The following papers and chapters (in chronological order) are included or discussed in this dissertation:

A. Frary, T. C. Nesbitt, A. Frary, S. Grandillo, E. van der Knaap, B. Cong, J. Liu, **J. Meller**, R. Elber, K. B. Alpert, S. D. Tanksley; *fw2.2: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size*, **Science**, 289: 85-88 (2000)

J. Meller and R. Elber; *Linear Optimization and a Double Statistical Filter for Protein Threading Protocols*, **Proteins: Structure, Function and Genetics**, 45: 241-261 (2001)

J. Meller, Wagner M, Elber R; *Maximum Feasibility Guideline to the Design and Analysis of Protein Folding Potentials*, **Journal of Computational Chemistry**, 23: 111-118 (2002)

J. Meller, R. Elber; *Protein Recognition by Sequence-to-Structure Fitness: Bridging Efficiency and Capacity of Threading Models*, in Computational Methods for Protein Folding: A Special Volume of **Advances in Chemical Physics**, ed. R. A. Friesner, John Wiley & Sons 2002

A. V. Kuznetsova, **J. Meller**, P. O. Schnell, J. A. Nash, Y. Sanchez, J. W. Conaway, R. C. Conaway and M. F. Czyzyk-Krzeska; *VHL Binds Hyperphosphorylated Large Subunit of RNA Polymerase II through a Proline Hydroxylation Motif and Targets It for Ubiquitination*, **PNAS** vol. 100 (5), 2706-2711 (2003)

J. Meller; *Molecular Dynamics*, to appear in **Encyclopedia of the Human Genome**, Nature Publishing Group, Macmillan Publishers Ltd 2003

M. Wagner, **J. Meller** and R. Elber; *Large-Scale Linear Programming Techniques for the Design of Protein Folding Potentials*, **Mathematical Programming**, to appear (2003)

M. Tan, P. Huang, **J. Meller**, W. Zhong, T. Farkas and X. Jiang; *Mutations within P2 Domain of Norovirus Capsid Affect Binding to Human Histo-Blood Group Antigens: Evidence for a Binding Pocket*, **Journal of Virology**, to appear (2003)

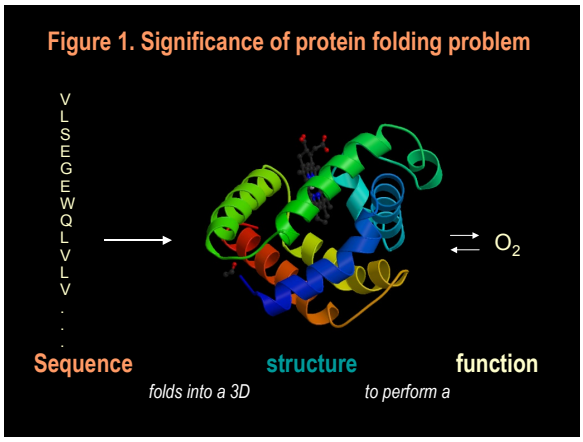
A. Porollo, R. Adamczak, M. Wagner and **J. Meller**; *Maximum Feasibility Approach for Consensus Classifiers: Applications to Protein Structure Prediction*, CIRAS 2003, accepted

Part IV contains several appendices, including abbreviations and definitions regarding basic notions discussed in this dissertation, list of other publications by the author of this thesis, selected materials highlighting the context and importance of some of the findings presented here and a brief description of the LOOPP software package.

I.2. The Protein Folding Problem

The recent unveiling of the human genome marked the transition in the biological sciences towards the post-genomic era, in which the understanding of protein structure and function becomes a crucial extension of the sequencing efforts. Despite recent progress in high throughput techniques, the experimental determination of protein structure by using X-ray crystallography or NMR spectroscopy [Branden and Tooze, 1991; van Holde et. el., 1998] remains a bottleneck in structural genomics. This poses a challenge and an opportunity for computational approaches to complement and facilitate experimental methods.

Proteins are linear polymers composed of a sequence of amino acid residues that are connected by peptide bonds (creating the protein "backbone"). Without accounting for several exotic amino acids and numerous chemical modifications of the basic blocks, there are 20 different amino acids that are characterized by chemically unique side chains (containing from one to approximately 20 atoms) that hang off the backbone chain. Protein molecules consist of several tens to several thousands of amino acids and thus between a few hundred and tens of thousands of atoms [Branden and Tooze, 1991].



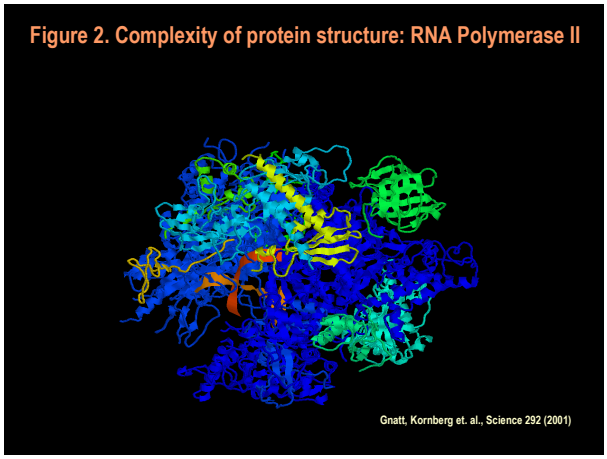
Proteins typically adopt a "unique" three-dimensional structure, meaning that under physiological conditions proteins with identical or nearly identical sequences would adopt very similar backbone conformations that form a well-defined cluster (called protein family), which is different from "structures" (clusters) of other families [Murzin et. al., 1995; Bateman et. al., 2002]. On the other hand, however, there are many examples of inherently unstructured or unstable proteins that may adopt very different conformations, for example depending on specific interactions with other proteins.

The above paragraph contains several somewhat fuzzy qualifiers – something that quantitatively trained readers may find difficult to accept. We will try to explain some of the above statements in the remaining part of this section. Yet, it is important to realize that the nature of biological objects and processes we are dealing with is such that approximate and inherently arbitrary distinctions need to be made when designing mathematical models representing the underlying biology. Here, for the sake of simplicity we will only consider proteins that under specific conditions fold into well-defined stable structures.

The *protein folding* problem consists of predicting the three-dimensional structure of a protein from its amino acid sequence (see Figure 1). The methodology and modeling aspects of protein folding have been vastly discussed in the literature. For excellent and up-to-date

surveys of methods as well as their limitations, the reader is referred to [Schonbrun et. al., 2002; Banavar et. al., 2001; Sternberg et. al., 1999]. In what follows, we briefly discuss several central concepts and ideas that underlie developments in the field.

The overall three-dimensional structure (conformation) of a protein may be hierarchically described first in terms of the conformation of the backbone, with locally ordered structures such as alpha helices and beta strands called secondary structures, and then in terms of side chain conformations given the relatively rigid backbone conformation. Protein structures can be further classified according to their secondary structure content and the relative packing of the secondary structure elements into distinct structural classes called folds. At present, there are well over 20,000 known protein structures, which are deposited in the Protein Data Bank (PDB) [Berman et. al., 2000]. Depending on the classification criteria these structures are divided into several hundred to about one thousand distinct folds. A number of families can be distinguished for each fold, with a total number of about 6,000 distinct families according to the Protein Families (PFAM) database [Bateman et. al., 2002].



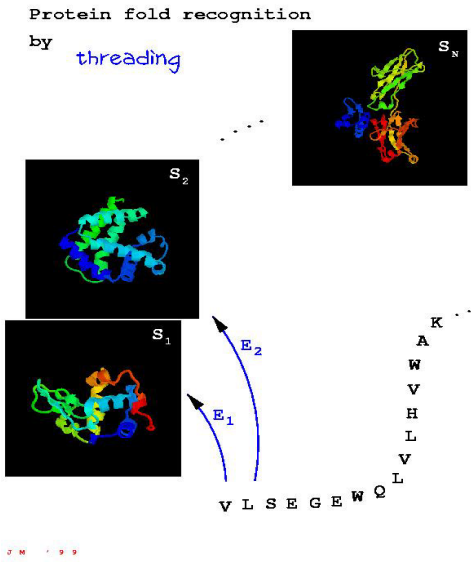
The computational protein structure prediction is a challenging problem. In order to appreciate the difficulty of the problem at hands it is useful to consider a brute force approach based on exhaustive enumerations of all possible conformations that may be adopted by a chain of amino acids. Each residue adds to the backbone two single bonds, which are free to rotate around their axis. However, due to steric constraints (clashes between backbone and side chains atoms), only up to three states (torsional angles) can be adopted around each single bond. Therefore, the number of possible backbone conformations is of the order of 9^N, where N is the number of amino acids in the chain [Branden and Tooze, 1991; van Holde et. el., 1998].

While this estimate is an upper bound, the conformational space to be explored becomes huge even for relatively short proteins, making a straightforward approach of exhaustive search impractical. Of course, even if we could perform an exhaustive search we would still face the problem of finding an appropriate scoring function capable of scoring the native-like structures higher than all the alternative conformations, which is far from trivial as discussed in the subsequent sections.

Except for extremely slow folders, proteins fold under physiological conditions on the time scales of milliseconds to seconds. Thus, the exponential scaling in the size of the conformational space remains in stark contrast with the observed folding rates, an observation known as the Levinthal's paradox [van Holde et. el., 1998]. Clearly, nature does not use a combinatorial approach in order to fold proteins. Consequently, using nature as a guideline may help in designing successful modeling and simulation protocols. In general, the existing computational approaches to protein folding problem may be roughly divided into two classes, based on the underlying principles and the extent of incorporating the physical characteristics of the protein folding process into computational protocols.

The *ab-initio* or *de novo* protein folding simulations attempt to reproduce (or at least to use as a guideline) the actual physical folding process. Such folding simulations are based on the thermodynamical hypothesis, first introduced by Anfinsen [Anfinsen, 1973], in which the unique three-dimensional structure of a protein is postulated to correspond to a global minimum of the free energy function. The free energy function is usually postulated in the form of a conveniently chosen atomistic force field (or folding potential), with parameters fitted to reproduce experimental data, whereas the search for the native conformation entails the solution of a global optimization problem. Some methodological aspects of atomistic models of proteins and computer simulations using Molecular Dynamics, Monte Carlo and other global optimization techniques are discussed in [Meller, 2003].

One might argue that knowing the complete atomistic description of the environment and the underlying physical interaction laws, we should be able to find the structure of a protein given the environment and interaction partners in particular. However, the problem is far from being trivial due to mentioned before size of the conformational space and the resulting sampling problem as well as inherent inaccuracy of atomistic force fields. Therefore, protocols that are in fact effective combinations of the *de novo* and knowledge-based approaches (see below), such as the *Rosetta* method by D. Baker and colleagues [Simons et. al. 1997], are more successful in practice.



The alternative *protein* (or *fold) recognition* approach [Bowie et. al., 1991; Jones et. al., 1992; Sippl et. al., 1992] relies on the fact that a significant fraction of protein structures (folds) have already been determined. The determination of the overall structure is reduced in fold recognition methods to tests of sequence fitness into known and limited number of known folds (thus it cannot be applied to novel folds). In other words, the search for the native conformation is restricted to the set of known structures, as opposed to computationally expensive search in the space of all possible conformations in case of *ab initio* folding simulations.

Since proteins of similar sequences fold into similar structures, sequence alignment (discussed in the next section) is the basic tool for assigning an unknown protein to a family of structurally and functionally characterized proteins. In many cases, sequence identity between 20 to 30% is sufficient to confidently assign a new protein to its family by using family profile based methods for sequence alignment, such as Position Specific Iterative Basic Local Alignment Tool (Psi-BLAST) algorithm [Altschul et. al., 1997] or profile Hidden Markov Models (HMMs) [Durbin et. al., 1998]. High degree of sequence similarity (also called sequence homology) allows one to obtain reliable alignments and effectively overlap new sequences with backbones of known structures. Furthermore, final three-dimensional models may be built by subsequent refinement of the alignment-based initial structure with atomistic force fields and global optimization, an approach known as *homology modeling*.

On the other hand, experiments found a limited set of folds compared to a large diversity of sequences. In other words, while sequence similarity usually implies significant structural similarity, the reverse is not true i.e. structural similarity does not necessarily imply sequence similarity. Because divergent or unrelated sequences may fold into similar structures, it suggests the use of structures to find remote similarities between proteins.

Threading is a fold recognition technique to directly match a sequence with a protein structure and a plausible function [Bowie et. al., 1991]. Protein recognition by sequence-to-structure matching or threading, allows one to find distant homologs that share the same fold without detectable sequence similarity (see for example [Meller and Elber 2001]). Given an appropriate scoring function, which can be thought of as a simplified folding potential, these methods find the "best" template from the library of known folds by evaluating directly sequence-to-structure fitness [Mirny and Shaknovich, 1998].

The scoring functions for threading (threading potentials) may incorporate different measures of sequence to structure fitness, such as compatibility between predicted and observed secondary structures or optimality of the effective inter-residue interactions imposed by overlaying a query sequence with a template structure. Such scoring functions should have a functional form that facilitates efficient computing of optimal alignments (with gaps) of a sequence into known protein structures, as discussed in the next two sections.

I.3. Sequence Alignment and Dynamic Programming

There is an enormous level of redundancy in biological systems [Gibson and Muse, 2002]. For instance, identical or very similar molecules and involving them processes are being used across different cells, tissues and species. On the other hand, it is important to recognize the limits to similarity (for example between analogous protein pathways in human and yeast) in order to identify the most significant (i.e. conserved) features. For these reasons, analogy and comparison between molecular objects and processes is an extremely powerful tool in biology.

Proteins and other important bio-molecules such as nucleic acids and poly-saccharides are linear (with some exceptions) polymers that can be represented as strings or sequences in mathematical terms. For that reason (and in light of the remarks from the previous paragraph) string matching and sequence alignment algorithms play central role in bioinformatics as crucial tools of sequence analysis and comparison. For example, as discussed before, high degree of sequence similarity typically implies similar structure and function and, therefore, new proteins can be assigned to known protein families using sequence alignment tools.

In order to assess the level of similarity between two sequences one may utilize their optimal alignment. The problem of finding the optimal alignment of two sequences with gaps results in a global optimization problem that may be solved efficiently by the *Dynamic Programming* (DP) algorithm. DP is a classical computer science technique to solve combinatorial optimization problems [Gusfield, 1997], and plays an important role in computational biology [Durbin et. al., 1998].

A typical DP problem spawns a search space of potential solutions in a recursive fashion, from which the final answer is selected according to some criterion of optimality. If

an optimal solution can be derived recursively from optimal solutions of subproblems, DP can evaluate a search space of exponential size in polynomial time and space as a function of the length of the sequences to be aligned, provided that a ("local") scoring function leading to piecewise decomposable problem is used [Durbin et. al., 1998]. In the following we will show how DP can be applied to the sequence and sequence-to-structure alignment problem, highlighting these aspects of DP that play an important role in designing effective threading potentials for sequence-to-structure matching.

Formally, the relatedness of two strings or sequences may be defined in terms of the *edit* distance defined as the minimal number of basic edit operations, including substitution, insertion and deletion, that are needed to transform one string into another [Gusfield, 1997]. Edit distance may be further generalized, allowing for character dependent weights (scores) of different substitutions. Alternatively, a notion of similarity between two sequences in terms of the score of their optimal alignment (which corresponds to their minimum weighted edit distance) may be introduced.

Let us consider two strings (or sequences in the dual formalism), $S_1 = a_1 a_2 \ldots a_n$ and $S_2 = b_1 b_2 \ldots b_n$, over certain alphabet A (for example consisting of twenty letters representing different amino acids, $A = \{\alpha_i\}_{i=1}^{20}$), $a_k, b_l \in A \quad \forall k,l$. We also consider an extended alphabet that contains the "space" or "dash" symbol, $\overline{A} = A \cup \{-\}$, representing "gaps" i.e. insertions of unknown ("missing" with respect to other sequences in the family) characters to one of the sequences or equivalently deletions of characters from the other sequence.

A *global alignment* of sequences S_1 and S_2, denoted here as $\Lambda(S_1, S_2)$, is obtained as a result of intercalating the two sequences such that a new sequence of length $n+m$ is obtained and the order of characters in each sequence is preserved. Such intercalated sequence may be conveniently displayed with one of the original sequences above the other so that every character or gap in either string is placed against a unique character or gap in the other sequence (with gap against gap alignments excluded).

As an example of conversion between the two representations of global alignments let us consider an intercalated sequence $a_1 b_1 a_2 a_3 b_2 b_3 a_4 b_4$, which corresponds to an alignment $\Lambda = a_1 b_1 a_2 - a_3 b_2 - b_3 a_4 b_4$ that can be represented in the alternative notation as:

$$\begin{matrix} \mathtt{a_1 a_2 a_3 - a_4} \\ \mathtt{b_1 - b_2 b_3 b_4} \end{matrix} . \qquad (1)$$

We would like to comment that *local alignments*, which are more appropriate when partial similarity (e.g. similarity between protein domains) is considered, are in fact displayed in Figure 3. As opposed to global alignments, only the subsequences that maximize the similarity in terms of the alignment score (defined in equation (2) below) are considered in case of local alignments.

We define a *scoring function* (also referred to as a scoring matrix), $f_s : \overline{A} \times \overline{A} \to R$, that assigns to each pair of characters a score for replacing (substituting) one character by the other, e.g. a score for amino acid substitution, $f_s(\alpha_i, \alpha_j)$. The total score of an alignment $\Lambda(S_1, S_2)$ of length l is defined as the sum of scores for pairs of characters that are aligned against each other, $f_s(x_i, x_{i+1})$:

$$f_{\text{tot}}(\Lambda(S_1, S_2)) = \sum_{i=1}^{l/2} f_s(x_i, x_{i+1}) \;; \quad x_i, x_{i+1} \in \overline{A} \quad \forall i . \qquad (2)$$

We assume here that the scores of individual pairs (substitutions) are not explicitly dependent on the alignment. In other words, the scores are *local* and do not change depending on what characters are aligned at other positions.

There is an extensive literature regarding the design of scoring matrices for sequence alignment (see for example [Henikoff and Henikoff, 1989; Durbin et. al., 1998]). Biologically meaningful alignments can only be obtained when suitable scoring schemes are used and different tasks may require different scoring matrices. One approach is to choose the scores based on the observed frequencies of amino acid substitutions between carefully selected representatives of known protein families.

An example of such derived scores are the BLOSUM scoring matrices, with the number indicating the level of evolutionary relatedness between the representatives included in the training set (for example BLOSUM50 denotes the scoring matrix derived from sequences sharing at least 50% of sequence identity). In addition to BLOSUM scoring matrices for 20 amino acids, one also needs to assign gap penalties. Here, for simplicity gap penalties are assumed to be proportional to the number of spaces that are inserted. More realistic models of gap penalties usually assume different cost of opening and extending a gapped region [Durbin et. al., 1998].

Figure 4. Sequence alignments reveal biological relatedness

```
i..d......1.......i..2.........3.....i.i...iii.....i.....5.    531 - 582
-FKLELVEKLFAEDTEAK-NPFSTQDTDLDLEMLAPY-I-PMD---DDLQL-RSFDQLS   Hif-1a
SFE-ETVEILFEAGASAELDDCRGVSENVILGQMAPIGTGAFDVMIDEESLVKYMPEQK   1i50_A (Rpb1)
... ......1.........2.........3.........4.........5........   1400 - 1458


    i..d......1.......i..2.........3.........4.........5.    531 - 582
    -FKLELVEKLFAEDTEAK-NPFSTQDTDLDLEMLAPYIPMDDDLQLRSFDQLS   Hif-1a
    SFE-ETVDVLMEAAAHGESDPMKGVSENIMLGQLAPAGTGCFDLLLDAEKCKY   Rpb1 (Human)
    ... ......1.........2.........3.........4.........5..    1400 - 1451
```

The size of the search space in the problem of finding the optimal alignment with gaps scales exponentially with the length of the sequences considered. Indeed, the total number of non-redundant global alignments (two alignments are redundant if they result in the same score, f_{tot}) for two sequences of length n and m is given by $(n+m)!/[m!n!]$. This is a simple consequence of the one-to-one correspondence between alignments and intercalated sequences stated in our definition, and it may be easily verified as follows. The order of each of the sequences is preserved when intercalating them, and therefore, we have in fact $n+m$ positions to place m elements of the second sequence (once this is done the position of each of the elements of the first sequence is fixed unambiguously). Hence, the number of intercalated sequences is simply the number of m-element combinations of $n+m$ elements.

Gaps allow one to take into account important evolutionary events that lead to insertions or deletions of stretches of nucleotides (and consequently amino acids) of various length, leading to proteins of similar core structures and functions, but of different lengths. It is the introduction of gaps, however, which makes the problem scaling exponentially. In light of the huge and ever growing size of the biological sequences databases, the importance of efficient solutions to this problem can hardly be overstated. This is exactly why DP is so

important in bioinformatics - using DP the problem may be solved in the order of $O(n \times m)$ steps, i.e. the optimal alignment may be found in polynomial time [Durbin et. al., 1998].

This dramatically less expensive solution is achieved by breaking the problem into subproblems. Only best partial alignments up to a given point are considered and then another pair of characters is added to the alignment, depending on what is the optimal extension of a given partial alignment. For the problem of the global alignment and the linear gap penalty considered here (with a score for aligning a residue with a gap defined as $f_s(-,\alpha_i) = f_s(\alpha_i,-) = -d\,;\ d > 0$), the particular DP solution is known as the Needleman-Wunsch algorithm [Needleman and Wunsch, 1990], which consists of two steps: the construction of the so-called DP representing possible alignments table and the trace back procedure to identify the optimal alignment.

The DP table represents all the possible alignments. However, starting from the first pair of characters, only these partial alignments are traced, which proceed through locally optimal extensions of partial alignments up to a given point, defined using the following recursive rules:

$$f_{\text{tot}}(0,0) = 0; \quad f_{\text{tot}}(k,0) = f_{\text{tot}}(0,k) = -kd$$
$$f_{\text{tot}}(i,j) = \max\{ f_{\text{tot}}(i-1,j-1) + f_s(a_i,b_j), f_{\text{tot}}(i-1,j) - d, f_{\text{tot}}(i,j-1) - d \}. \qquad (5)$$

Therefore, the optimal alignment can be then traced back, starting from the lower right corner of the DP table, as shown in the example included in Figure 5.

Figure 5. Global alignment algorithm

		H	E	A	G	A	W	G	H	E
	0	<-8	<-16	<-24	<-32	<-40	<-48	<-56	<-64	<-72
P	^-8	*-2	*-9	*-17	<-25	*-33	<-41	<-49	<-57	*-65
A	^-16	^-10	*-3	*-4	<-12	*-20	<-28	<-36	<-44	<-52
W	^-24	^-18	^-11	*-6	*-7	*-15	*-5	<-13	<-21	<-29
H	^-32	*-14	*-18	*-13	*-8	*-9	^-13	*-7	*-3	<-11
E	^-40	^-22	*-8	<-16	^-16	*-9	*-12	^-15	*-7	*3

HEAGAWGHE

--P-AW-HE

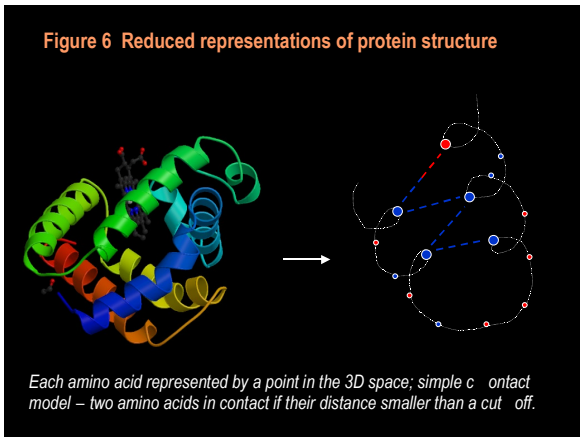
The BLOSUM50 scoring matrix (for instance $f_s(P,H) = -2$) and the gap penalty $d = 8$ were used in this example. Note that symbols *, ^ and < are used to indicate which of the three possible extensions of the alignment was optimal, corresponding to the alignment of an amino acids in the first sequence with an amino acid in the second sequence, a gap in the first sequence with an amino acid in the second sequence, or an amino acid in the first sequence with a gap in the second sequence, respectively.

The above symbols represent in fact pointers that allow one to efficiently trace back the optimal alignment. Since the number of the cells in the DP table is $(n+1)\times(m+1)$ and a fixed number of operations per cell are required, it is easy to see that the overall complexity of the Needleman-Wunsch algorithm is indeed polynomial (quadratic in n assuming for simplicity that $n = m$) in time and space. Further discussion of this algorithm may be found in [Durbin et. al., 1998].

There are many extensions and modifications of this basic scheme, such as the Smith-Waterman [Smith and Waterman, 1981] algorithm for local alignments. Dynamic Programming is truly ubiquitous in sequence analysis [Gibson and Muse, 2002; Pevzner, 2000, Gusfield, 1997]. On the other hand, however, DP with its quadratic polynomial complexity may be computationally too expensive for large-scale applications. Therefore, many heuristic schemes, such as BLAST [Altschul et. al., 1997], which are more efficient but are not guaranteed to find the optimal solution, were devised.

The sequence-to-structure matching may be perceived as a generalized sequence matching, with one of the sequences consisting of amino acids and the other of structural sites characterized in terms of their structural environment (e.g. the number of neighbors to a site). Therefore, DP techniques may be directly applied to solve efficiently the problem of finding optimal sequence-to-structure alignments. In light of considerations included in this section, however, scoring functions for efficient sequence-to-structure matching should enable piecewise approach and decomposition of the problem into "local" subproblems. This observation is the starting point for the developments summarized in the next section.

I.4. Contact Potentials for Protein Recognition



Protein structure is often represented in terms of simplified, reduced models that speed up computation. For example, the commonly used contact model represents each amino acid by just one point, which defines the approximate location (site) of an amino acid. The overall shape of a protein may be characterized in terms of contacts between closely packed amino acid residues, or in other words in terms of effective interactions between the structural sites representing amino acid residues.

Such contact models allow one to capture the packing of hydrophobic residues that are buried in the core of the protein and contribute to the stability of the structure. Hydrophobic residues are represented as blue circles in Figure 3, as opposed to hydrophilic residues that are marked in red and are predominantly found on the surface of globular proteins [Branden and Tooze, 1991].

Let us consider widely used inter-residue folding potentials. In contact pairwise models [Sippl et. al., 1992; Bryant et. al., 1993; Godzik et. al., 1992] the energy of a protein with sequence S and structure $\mathbf{X}$ is a sum of pair energies from all pairs of interacting amino acids:

$$E(S,\mathbf{X};\mathbf{z}) = \sum_{i<j} z_{\alpha_i\beta_j} = \sum_{\gamma} z_\gamma n_\gamma(S,\mathbf{X}) . \qquad (6)$$

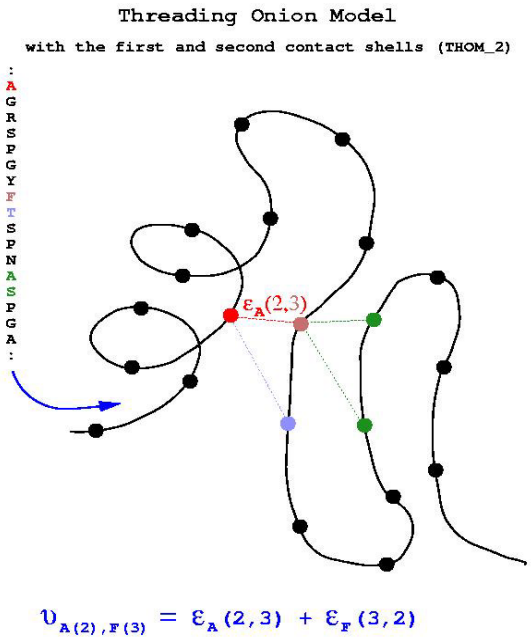
The summation index, $\gamma \equiv \alpha\beta$, runs over 210 different contact types, where α and β denote types of amino acids ($\alpha,\beta \in \{1,2,...,20\}$) at certain sites i and j in contact, and $n_\gamma(S,\mathbf{X})$ denotes the number of contacts of a specific type found in the structure $\mathbf{X}$. Thus, given the effective "pair energies", $z_\gamma \equiv z_{\alpha\beta}$ (also denoted as $\varepsilon_{\alpha\beta}$ throughout papers included in this dissertation), computing the overall energy of a structure reduces to counting of different types of contacts. Sites i and j are said to be in contact, if their distance, r_{ij}, is sufficiently small. In this work we consider the model that was used before to optimize threading potentials [Tobi et. al., 2000], with geometric side chain centers as interaction sites. Two sites are assumed to be in contact if their distance satisfies $1.0 < r_{ij} < 6.4$ Å, which implies that only neighbors from the first contact shell are taken into account. Furthermore, pairs of residues that are separated by fewer than four virtual bonds, i.e. $|i-j| \geq 4$, are excluded.

The effective pair energies for inter-residue interactions can be derived from the analysis of contacts in known structures, with z_γ defined by the frequency of observing contacts of type γ normalized by the so-called background frequencies [Sippl et. al., 1992]:

$$z_{\alpha\beta} = -C \ln[\frac{p_{\alpha\beta}}{p_\alpha p_\beta}]. \quad (7)$$

Here, C is a positive constant that defines the energy scale, $p_{\alpha\beta}$ denotes the probability of observing (in a set of native structures) amino acids of types α and β in contact, whereas p_α and p_β denote probabilities of observing these individual amino acids (again in a set of native structures). Such knowledge-based, pairwise potentials are widely used in fold recognition [Jones et. al., 1992; Bryant et. al., 1992; Mirny and Shaknovich 1998], *ab-initio* folding [Sternberg et. al., 1999; Liwo et. al., 1997; Xia et. al., 2000] and sequence design [Babajide et. al., 1997, 1999]. Alternative strategies to find the effective pair energies (parameters of folding potentials in general) are discussed below.

It is important to realize that such simplified models incorporate the interactions with the solvent in terms of the effective pair energies. Proteins adopt their three-dimensional conformations in specific environments. Soluble proteins fold in an aqueous environment, whereas membrane proteins fold in a lipid environment. Thus, effective pair energies must be derived separately for different environments in order to account for the observed (in a given environment) structure.



As an alternative to pairwise contact models, one may consider the so-called "profile" models [Bowie et. al., 1991; Elofsson et. el., 1998], in which the overall effective energy of a protein takes the form of a sum of individual site contributions, depending on the structural environment of a site. For example, the solvation or burial state or the secondary structure can be used to characterize different local environments.

The advantage of profile models is the simplicity of finding optimal alignments with gaps (deletions and insertions into the aligned sequence) that allow the identification of homologous proteins of different length. As discussed in the previous section, using DP algorithm one may compute optimal alignments with gaps in polynomial time, as compared to the exponential number of all possible alignments, if a "local" scoring function is used.

In contrast to profile models, the potentials based on pair energies do not lead to exact alignments with dynamic programming. The reason for that may be explained by considering how a score for aligning an amino acid residue with a structural site is computed when using pairwise potentials. Namely, all contacts to a site need to be considered, each contributing an effective pair energy that is dependent on the identity of the "other" amino acid in contact.

However, the placement of gaps (i.e. the alignment) may change the identity of the "other" residues and the problem becomes non-local (NP-complete in fact [Lathrop, 1994]).

A number of heuristic algorithms, providing approximate alignments, have been proposed, e.g. [Lathrop and Smith, 1996]. However, they cannot guarantee an optimal solution with less than exponential number of operations. We introduced a novel energy function that employs reduced, contact models of protein structure and blends the contact energies with profile models to achieve computational efficiency and higher accuracy in recognition of native-like structures [Meller and Elber, 2002]. The new model is called THreading Onion Model 2 (THOM2) since it uses information about the first and the second contact shells of an amino acid residue and it incorporates some cooperativity effects that are not included in standard pairwise folding potentials.

In THOM2 one defines the effective energy $z_{\alpha_i}(n_i, n_j)$ (also denoted as $\varepsilon_{\alpha_i}(n_i, n_j)$ in some of the figures and papers included here) of a contact between structural sites *i* and *j*, where n_i is the number of neighbors to site *i* and n_j is the number of neighbors to site *j*. The type of amino acid at site i is α_i. Only one of the amino acids in contact is known. The total contribution to the energy of site *i* is a sum over all contacts to this site $\phi_{i,\mathrm{THOM2}}(\alpha_i, \mathbf{X}) = \sum_j{}' z_{\alpha_i}(n_i, n_j)$. The prime indicates that we sum only over sites *j* that are in contact with *i*, where contact is defined as previously for pairwise models. The total energy is finally given by a double sum over *i* and *j*,

$$E_{THOM2} = \sum_i \sum_j{}' z_{\alpha_i}(n_i, n_j) \ . \tag{8}$$

As was the case for pairwise models defined before, computing the overall energy of a structure reduces to the counting of different types of contacts, $n_\gamma(S, \mathbf{X})$, which are however defined in terms of the number of neighbors to sites involved in contact and identity of the amino acid occupying the "primary" site. Therefore, we may express the overall energy as linear combination with respect to the parameters z_γ:

$$E_{THOM2}(S, \mathbf{X}; \mathbf{z}) = \sum_\gamma z_\gamma n_\gamma(S, \mathbf{X}), \tag{9}$$

where the summation index is defined now in terms of the amino acid type occupying the primary site, its number of neighbors and the number of neighbors to the other site involved in contact, $\gamma \equiv \gamma(\alpha_i, n_i, n_j)$. We use a coarse-grained model leading to a reduced set of structural environments (types of contacts) by merging residues with similar number of neighbors into several classes. Therefore, the number of parameters, which might be very large in principle (assuming up to ten neighbors to a site we would obtain $20 \times 10 \times 10 = 2000$ parameters), is reduced to a number comparable with 210 parameters of the pairwise model (see Paper 1 for details).

Since each contact contributes twice to the overall energy, it is possible to define an effective pair energy using THOM2 as well:

$$V_{ij}^{eff} = z_{\alpha_i}(n_i, n_j) + z_{\alpha_j}(n_j, n_i) \ . \tag{10}$$

Hence, one can formally express the THOM2 energy as a sum of pair energies,

$$E_{THOM2} = \sum_{i<j} V_{ij}^{eff} \ . \tag{11}$$

The effective energy mimics the formalism of pairwise interactions. However, in contrast to the usual pair potential, the optimal alignments with gaps can be computed efficiently with THOM2, since structural features alone determine the "identity" of the neighbor.

The energy terms (parameters of the potentials), $z_{\alpha_i}(n_i, n_j)$, could be computed using statistical approach for example, in analogy to knowledge based pairwise potentials

defined in equation (7). However, such statistical potentials "learn" from the native structures ("good" examples) only. In order to increase their power to distinguish misfolded states (the "bad" examples) from native states, more sophisticated protocols incorporate data from decoy structures as well. One approach to designing potentials that improves upon statistical potentials is the so-called Z-score optimization, discussed in Paper 3.

Here, in order to achieve better discrimination of native structures with respect to misfolded decoys, we explicitly demand that the folding potentials mimic the postulate that the native states have the lowest energy. Such formulation leads to a problem of solving linear system of inequalities, which we chose to solve using Linear Programming techniques (for an overview of LP and other techniques and algorithms for solving linear systems of inequalities the reader is referred to [Vanderbei, 1996]).

We used LP methods to design and evaluate several scoring functions (including THOM2) for threading and to optimize their parameters. For example, the site energies $z_{\alpha_i}(n_i, n_j)$ are optimized using the LP protocol to find a solution of a large set of linear inequalities derived from a large set of native and misfolded structures as described in the next section. LP is also used to determine optimal gap penalties. The new model provided an efficient threading approach for annotations of remote homologs that share structural similarity without significant sequence similarity. Applications of this approach are presented in papers included in Part III of this dissertation.

I.5. LP Approach to the Design of Folding Potentials

In both ab-initio folding and protein recognition we are faced with the problem of finding (designing) an appropriate expression for the free energy or scoring function (also called here folding potentials), respectively. The basic requirement for protein folding potentials is their ability to distinguish native-like from non-native structures. This can be achieved by an appropriate choice of the functional form and parameters of the energy function (in the following we will use the "physical" convention according to which well folded structures are expected to yield low energies, as opposed to high scores when using scoring functions).

Assuming that folding potentials are expected to have the lowest energy for the native fold, one may impose that for each pair of native and misfolded structures that are considered the following constraints are satisfied:

$$\Delta E_{\text{mis, nat}} \equiv E_{\text{misfolded}} - E_{\text{native}} \geq \varepsilon \quad . \tag{12}$$

Here, $E_{\text{native}} \equiv E(S, \mathbf{X}_{nat}; \mathbf{z})$ is the energy of the native structure $\mathbf{X}_{nat}$, $\mathbf{z}$ is the vector of parameters to be optimized, $E_{\text{misfolded}} \equiv E(S, \mathbf{X}_{mis}; \mathbf{z})$ represents the energies of the misfolded (non-native) structures $\mathbf{X}_{mis}$ and ε is a positive constant. In other words, we require that the energies of native structures are lower than the energies of misfolded structures.

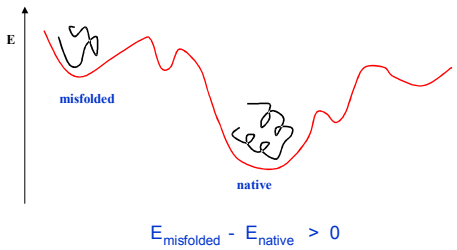
It should be noted that casting the problem of designing folding potentials in terms of optimization of the parameters $\mathbf{z}$, such that correct recognition (classification) of a set of examples (pairs of native and misfolded structures) is imposed in the training, implies that the problem is in fact formulated within the framework of the supervised classification approach. Obviously, as with any other supervised classification protocol, the choice of training set of examples and further validation of the results on independent control sets is critical for the successful optimization of folding potentials. Discussion of different issues involved in making these critical choices is included in Paper 2.

For energy models considered here, such as the contact potentials defined in (6) and (9), one may in general expand the energy as linear combination in terms of their parameters:

$$E(S,\mathbf{X};\mathbf{z}) = \sum_{\gamma} z_{\gamma}\alpha_{\gamma}(S,\mathbf{X}), \qquad (13)$$

with the coefficients of the linear combination, $\alpha_{\gamma}(S,\mathbf{X})$, taking a model specific (structure and sequence dependent) form. In such case, the set of inequalities in equation (12) that impose correct recognition of native structures in the training can be solved for the parameters $\mathbf{z}$ by using Linear Programming approaches. Linear systems of inequalities have simple geometric interpretation and a number of efficient techniques and tools can be used to solve them, as discussed in the next section.

Figure 8. Recognition of native structures by folding potentials



The LP approach for the design of folding potentials was pioneered by Maiorov and Crippen [Maiorov and Crippen, 1992]. Recently, the LP approach has been applied to the design and evaluation of various folding and threading potentials [Tobi et. al., 2000; Vendruscolo and Domany, 1998]. It has been found, for example, that simple contact pairwise potentials are not sufficient for recognition of all types of protein structures (see also the discussion included in Papers 1 and 2).

The set of inequalities in equation (12), that we attempt to solve, proves infeasible (meaning that there is no solution satisfying all the constraints in (12)) for sufficiently large sample of native and misfolded shapes. As discussed in Papers 1, 2 and 3, we used infeasibility of large training sets with different functional models as a guideline to design optimal threading potentials. We seek functional forms for the potentials that achieve high accuracy in recognition of representative (exhaustive) sets of protein structures included in the training set and minimize the number of required potential parameters.

As further discussed in Paper 1 and in the book chapter [Meller and Elber, 2002] (not included in this dissertation), the linear dependence of the potential functions on their parameters is not a major restriction. Any nonlinear function can be expanded or at least approximated as a linear combination of basis functions. The challenge is to find a limited set of basis functions that capture most of the intrinsic complexity of the true energy function and thus make for a reasonable model.

Our ultimate goal is an "optimal" energy model, which balances complexity and accuracy, while avoiding the dangers of over- and under-fitting. Using this approach, a number of scoring functions, based on "structural profiles" of a site and on pairwise interactions, are evaluated and a novel model, blending the prediction capacity of pairwise models with efficiency of profile models, is optimized.

The use of the LP approach to the design of folding potentials usually involves solving very large sets of inequalities derived from large samples of native and misfolded structures (decoys), as discussed in Paper 1. Therefore, the efficiency of LP algorithms is an important issue. In the papers included in Part II of this dissertation we demonstrate that large-scale LP tools based on the so-called *interior point* approach to solving LP problems [Karmakar, 1984; Ye, 1997], which allow for the solution of systems with hundreds of

millions of constraints, result in significant improvements in the quality of scoring functions for protein folding and threading.

We also demonstrate how solving these very large LP problems in conjunction with the recently proposed ``Maximum Feasibility" heuristic [Meller et. al., 2002] may be used to evaluate different functional forms. A brief overview of interior point methods and some other LP techniques is included in the next section.

I.6. Interior Point Methods for LP

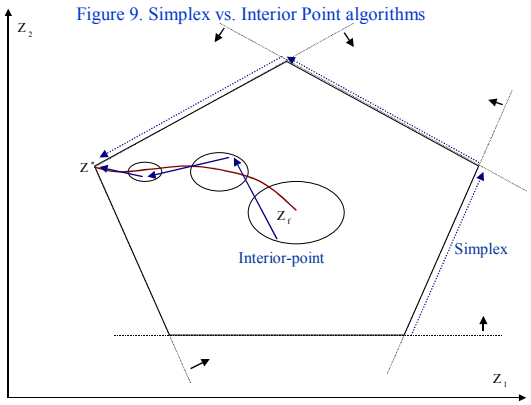
Let us consider a linear programming problem of the form (which includes the set of inequalities in (12) as a special case):

$$\{\min f_0(\mathbf{z});\ \mathbf{z} \in \mathbf{R}^n;\ f_i(\mathbf{z}) \le b_i,\ i = 1, \cdots, m\ \}, \tag{14}$$

where $\mathbf{z}$ is a vector of n variables and the objective function to be minimized, f_0, as well as the constraints functions, f_i, are linear.

Linear inequalities of equation (14) define a set of "cutting" hyperplanes in the parametric space, as discussed in detail in Papers 1, 2 and 3. The intersection of the corresponding feasible (closed) half spaces defines a convex polyhedron (see Figure 9). If there exists a solution satisfying all the constraints in (14), the LP problem is called *feasible* (otherwise the problem is called *infeasible*). For feasible problems LP solvers provide a solution, $\mathbf{z}^*$, which belongs to the feasible polyhedron and optimizes a linear objective function defined in (14).

Figure 9. Simplex vs. Interior Point algorithms



Interior point methods, due to their superior polynomial time complexity [Karmakar, 1984; Ye, 1997] and practical efficiency are nowadays a method of choice for large-scale linear optimization problems. An interior point algorithm generates a series of points away from the boundary of the polyhedron (unlike the simplex algorithm which proceeds along the edges of the feasible region [Vanderbei, 1996], see also Figure 9). These points are near a smooth curve, called the *central path*, which is contained within the interior of the feasible polyhedron and terminates at an optimal and complementary solution on a facet or at the vertex (if the optimal solution is unique) of the polyhedron [Ye, 1997].

In order to formally elucidate the idea of interior point algorithms for LP (and in general for convex programming), one can define a logarithmic barrier function associated with (14) as follows:

$$\phi_B(\mathbf{z}, \mu) = \frac{f_0(\mathbf{z})}{\mu} - \sum_{i=1}^{m} \ln(b_i - f_i(\mathbf{z})), \tag{15}$$

where $\mu > 0$ is the barrier parameter. If the feasible region of (14) is bounded (i.e. all variables, $z_j;\ j = 1, \cdots, n$, are bounded from below and from above by finite numbers) and non-empty, then for each value of μ the barrier function, $\phi_B(\mathbf{z}, \mu)$, achieves the minimal

value at a unique (feasible) point, $\mathbf{z}(\mu)$, which is called the μ-center [Ye, 1997; Adler and Monteiro, 1991].

The *central path* is defined as the set of μ-centers, where μ changes from ∞ to 0. In the limit of $\mu \to 0$, when minimizing the barrier function of equation (3), one obtains the desired optimal and feasible solution of (14) – see Figure 9. Barrier functions of the form specified in equation (15) are used in the interior point methods for inequality constraints [Ye, 1997] in order to reformulate the constrained optimization problem of (14) into unconstrained, nonlinear optimization problem of (15). The advantage of the latter is that the non-linear minimization techniques, such as gradient or Newton methods, can be applied. Such reformulation proved also to be critical for obtaining the polynomial complexity bounds for interior point algorithms.

The unique minimum of the barrier function in the limit of $\mu \to \infty$ is called the *analytic center* of the feasible region [Ye, 1997; Adler and Monteiro, 1991]. The central path always starts at the analytic center and, in the absence of an objective function to optimize, the interior point algorithms converge to the analytic center.

We emphasize that in practice (as in the popular infeasible primal-dual implementations, for example [Czyzyk et. al., 1999]) the functional constraints, f_i, are often initially relaxed and the method proceeds through points away from the central path that may not belong to the feasible polytope. Therefore, the analytic center is reached only upon convergence of the Newton procedure. Moreover, there are many parameterizations of the central path. In particular, different barrier functions (as for example weighted logarithmic barriers) can be applied [Adler and Monteiro, 1991]. Therefore the actual position of the analytic center may vary between different implementations.

Note that solving a set of linear inequalities is equivalent to solving a special case of (14), obtained by setting the objective function in (14) to zero, $f_0(\mathbf{z}) = \mathbf{0} \cdot \mathbf{z}$. Therefore, when solving a set of inequalities by an interior point algorithm one obtains (in principle) the analytic center of the feasible polyhedron as a solution. It is worth noting that solving a set of inequalities (which is by duality theorem equivalent to solving an LP problem [Vanderbei, 1996]) is of the same complexity as the original problem with an objective function to optimize, as defined in (14).

It should be also pointed out that the analytic center does not correspond (in general) to the center of the feasible polytope in the topological sense. Redundant constraints that do not define the boundaries of the polytope contribute to the barrier function in equation (15) as well, "repulsing" the analytic center. However, the analytic center is always located away from any individual cutting hyperplane, due to singularity of the logarithm function at zero.

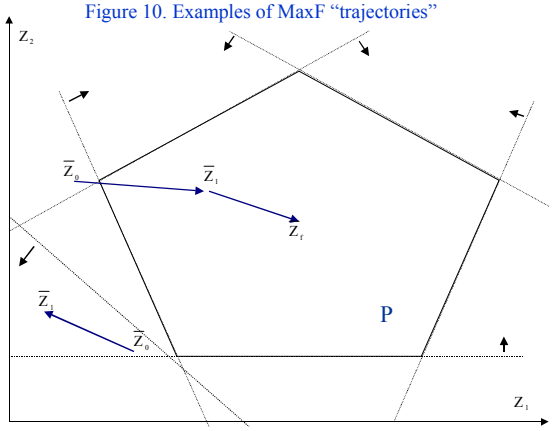
In summary, the use of interior point methods enables solving efficiently large-scale LP problems. While this is obviously a very important and desired feature, the use of interior point methods has additional advantages by providing, in conjunction with MaxF heuristic, approximate solutions to infeasible LP problems. Such solutions are characterized by "wide margin" of separation (i.e. they are away from individual constraints included in the training), bearing a promise of good generalization properties. This additional feature of interior point methods is exploited in several approaches to the design of folding potentials and consensus classifiers described in the next two sections.

I.7. Maximum Feasibility Heuristic

Folding and threading potentials are expected to have the lowest energy for the native shape. The Linear Programming approach achieves exactly that goal for a training set, or indicates that this goal is impossible to obtain. If a solution cannot be found (i.e. the problem is infeasible) one can either choose a new functional form for the potential (which can imply that more parameters are needed, increasing the risk of overfitting), or detect inconsistent constraints and find the best potential with a feasible subset of the data in the training set. Here, we explore the latter option i.e. finding an approximate solution (by which we mean a solution that satisfies most of the inequalities) of an infeasible set of inequalities. Finding a maximal subset of satisfiable constraints is known to be an *NP*-complete problem [Chakravarti, 1994]. We proposed a simple heuristic for finding an approximate solution to an infeasible set of linear inequalities, which is outlined below as well as in Papers 2 and 3.

The "Maximum Feasibility" (MaxF) method is a heuristic approach to find an approximate solution, which satisfies a possibly large subset of an infeasible set of inequalities. The MaxF procedure is based on a special property of interior point algorithms for LP. Namely, the interior point methods provide the so-called analytic center of the feasible polyhedron (defined in terms of logarithmic barriers "repelling" the solution from the constraints) when the objective function is not used to "force" the convergence to an optimal solution on a facet of the polyhedron [Adler and Monteiro, 1991]. For a bounded polyhedron (polytope) the analytic center is unique. However, even if the problem is not bounded and the notion of the analytic center is not defined, the interior point algorithms converge to solutions that are away from individual constraints, owing to the implicit logarithmic barrier function.

Figure 10. Examples of MaxF "trajectories"



An approximately feasible solution is obtained iteratively, starting from a certain initial guess e.g. a statistical potential that can be easily derived for a problem at hand or using an "elastic" LP problem (see the next section for the definition of the elastic LP). Given a set of constraints that are satisfied by the initial guess of the solution, a series of "maximally feasible" approximations is computed. The set of constraints that are satisfied by the initial guess defines a feasible LP problem, which is solved using an interior point method, providing the next approximate solution. The new approximation satisfies at least as many constraints as the previous partial solution. The newly satisfied constraints are added to the problem, which is solved to obtain the next "maximally feasible" approximation. If no further constraints can be satisfied the procedure stops.

The idea behind the MaxF heuristic is that a solution that is close to the center (or at least away from the constraints) of the feasible polyhedron, is likely to satisfy more constraints than an off-centered guess. Clearly, the success of the MaxF procedure depends on the choice of the initial solution, as discussed in details in Papers 2 and 3. Examples of

successful and unsuccessful MaxF "trajectories" are included in Figure 10 (feasible half spaces are indicated by the arrows pointing out from the cutting hyperplanes). For the problem at hands, however, approximate solutions of desired characteristics (such as the statistical folding potentials) are available and may be further improved by MaxF owing to a "wide margin" and good generalization features of its solutions, as demonstrated in Papers 2 and 3.

Using the MaxF guideline allows us to go beyond a simple feasibility test when assessing the quality of a given model and may provide a better insight for improving the functional models of folding potentials. It also provides a simple way to improve potentials that are not optimized to satisfy inequality constraints of the type of equation (12), for example the commonly used statistical potentials. Finally, MaxF is useful in pointing flows and cross-validation of training sets, which is a critical step in all supervised learning protocols.

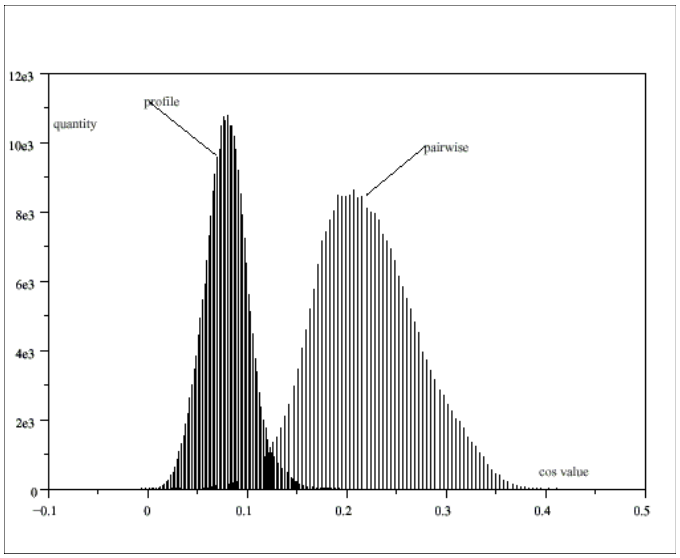


Figure 11. The distribution of cosines of angles between a vector in the parametric space representing a folding potential and normals to the cutting hyperplanes, defined by a set of 200,000 inequalities from the TE problem, is shown. The profile THOM1 (with 200 parameters) and pairwise TE (with 210 parameters) potentials are compared (see Paper 1 for details). Note, that the distribution for the pairwise potential is wider and shifted to the right, indicating a larger volume of the feasible polyhedron in this case.

For example, in Paper 2 we discuss the problem of solving the Hinds-Levitt (HL) set of proteins using reduced (meaning that chemically similar types of amino acids are grouped together to reduce the number of parameters) pairwise potentials. The HL set includes, in addition to soluble proteins, a number of membrane proteins, which are characterized by different folding principles and effective inter-residue interactions that account for different environments (aqueous vs. lipid). The so-called gapless threading protocol (see Papers 1 and 2 for detailed description of this protocol) was used to generate a large number of decoys (misfolded alternatives). Using the feasibility test we showed that perfect recognition of the proteins in the HL set is impossible without at least ten types of amino acids (55 parameters). The MaxF procedure, when applied to the HL problem, results in a potential with only four types of amino acids (10 parameters) that recognizes all but membrane proteins.

Other examples of successful applications of the MaxF algorithm are included in Paper 3. For instance, the THOM1 threading model, which is a simple profile model with 200 parameters corresponding to different types of structural sites, each contributing an energy term dependent on the type of amino acid occupying this site and the number of

neighbors to this site, was shown to results in an infeasible problem when applied to the Tobi-Elber (TE) set of proteins. Using MaxF procedure, however, one can find an approximate solution that satisfies all but 905 inequalities (out of 30 million). Furthermore, in order to achieve perfect recognition for the THOM2 model on the TE set it was necessary to increase the number of parameters to 300 by defining more structural classes for types of contacts. However, with a coarser definition of the structural classes and only 180 parameters one may get a solution that satisfies all but 233 constraints.

The results discussed above demonstrate clearly the dangers of overfitting when imposing a perfect solution on the training set. Of course, by finding a nearly perfect solution when using MaxF approach can still result in overfitting, since we are attempting to find a solution satisfying as many constraints in the training as possible. Thus, the choice of the training set and the initial solution for the MaxF iterations as well as careful cross-validation of the results will play a critical role in successful applications of the new approach. For example, identifying constraints that were added during the MaxF iteration and "pushed" the solution into an undesirable subset of the feasible polyhedron would allow us to remove them from the training and re-train. Such cross-validation procedures are commonly used in supervised learning to achieve better generalization.

The analysis of the geometry and the volume of a feasible polyhedra defined by linear constraints is a difficult problem. However, some simple measures may help to elucidate the structure of the problem at hands. For example, feasible volume histograms, showing the distribution of angles between an arbitrary vector in the parametric space (e.g. a folding potential) and the normals to the hyperplanes defined by the constraints included in the LP problem, are easy to compute and allow one to compare directly different models, irrespective of the energy scale. In particular, they may be used to measure the capacity of different models and difficulty of different training sets as illustrated in Figure 10. Moreover, using geometric techniques one may identify and remove from the training constraints that significantly decrease the feasible volume (that may result from the contamination of the training set by homologous structures for instance). Redundant constraints may be removed as well, resulting in a more centered solution and better generalization properties.

The last issue that we discuss here is of technical nature, yet it is critical for the success of the new heuristic. In order to solve large LP problems described in Paper 1 we used an iterative approach, solving a subsystem fitting into the memory and selecting constraints difficult to satisfy (or violated) to be included in the next iteration. Although such an approach is not guaranteed to converge, in practice we were able to obtain solutions fro problems an order of magnitude larger than the size of the subsystem that we were able to solve in “one shot”. However, in order to use the MaxF heuristic we need to be able to load all currently satisfied inequalities into memory. For approximate solutions of a good quality most of the constraints should be satisfied, which means that large LP problems must be solved using “one shot” approach. Therefore, we used a parallel implementation of the PCx primal-dual interior point LP solver [Czyzyk et. al., 1999, Wagner et. al., 2003] to obtain the results of MaxF in large-scale training of folding potentials described in Paper 3.

In the next section we will discuss further applications of MaxF heuristic, which we applied to develop novel strategy for optimizing consensus classifiers in the context of general supervised classification problem.

I.8. Maximum Feasibility Approach for Consensus Classifiers

Ensemble classifiers are an active area of research in the field of machine learning [Hastie et.al., 2001; Krogh and Vedelsby, 1995]. Many strategies, such as simple voting, linear combination based methods or boosting [Mulgrew and Cowan, 1988; Fruend and Schapire, 1996; Breiman, 1996], have been proposed to find an improved consensus classifier, given a number of individual classifiers. Consensus classifiers are often able to improve significantly on the classification accuracy. Some important and relevant in bioinformatics examples include applications of neural network based classifiers for protein secondary structure prediction or combining various individual scores into a consensus score for gene prediction (see for example [Baldi and Brunak, 1998]).

Recently, we proposed a novel strategy to optimize consensus classifiers for large-scale problems, using LP techniques in conjunction with the Maximum Feasibility heuristic. For a set of classifiers and their normalized class dependent scores one postulates that the consensus score is a linear combination of individual scores. Such defined total score is required to satisfy a set of linear constraints, imposing that the consensus score for the true class is higher than for any other class for each data point in the training. The formulation of the problem resembles the LP based approach for the design of scoring functions for protein folding, where for each protein family two classes (native and non-native structures) can be formally defined and the folding potential is supposed to assign a higher score (lower energy) to the true class (native states).

Let us consider a supervised classification problem with N real vectors from a certain feature space X, divided into K classes. A discrete set of class labels, conveniently chosen as $1,\ldots,K$, will be referred to as Y. A classifier Q is then a mapping from X to Y. For clarity of notation the k-th class will be alternatively labelled as C_k - $\mathbf{x} \in X$ is classified as belonging to class C_k, if $Q(\mathbf{x}) = k$.

Consider now a number of individual models, M_i, $i = 1,\ldots,p$, that provide estimates for conditional probabilities of class C_k given the model and a vector in the feature space, $P(C_k \mid \mathbf{x}; M_i)$. For each model we define an individual classifier Q_i as:

$$Q_i(\mathbf{x}) = \underset{k=1,\ldots,K}{\arg\max}\ P(C_k \mid \mathbf{x}; M_i)\,. \tag{16}$$

In other words, a data point **x** is assigned to the class with the highest probability**.** The goal is then to combine the individual models into a mixture (consensus) model.

We define a consensus classifier in the form of a linear combination of individual classifiers:

$$P(C_k \mid \mathbf{x}; M_c) = \sum_{i=1}^{p} \alpha_i P(C_k \mid \mathbf{x}; M_i)\,. \tag{17}$$

Note that the coefficients of the linear combination, which will be a target for optimization, are class independent here (as opposed to more general models with class dependent coefficients). Linear decision boundaries for the consensus classifier are defined using again the simple rule:

$$Q_c(\mathbf{x}) = \underset{k=1,\ldots,K}{\arg\max}\ P(C_k \mid \mathbf{x}; M_c)\,. \tag{18}$$

In supervised classification problem each training vector is assigned to its "true" class, which will also be called its "native" state in the context of applications to protein structure

prediction. The true (or native) class will be referred to as C_n, where $Q^*(\mathbf{x}) = n$ is the true classifier (with the implicit dependence of index n on $\mathbf{x}$).

In order to impose correct consensus predictions in the training, the following inequality constraints (with one inequality per data point) are used:

$$\sum_{i=1}^{p} \alpha_i P_i(C_n) \geq \sum_{i=1}^{p} \sum_{k \neq n} \alpha_i P_i(C_k), \tag{19}$$

where coefficients $P_i(C_k) = P(C_k \mid \mathbf{x}; M_i)$ of the constraint matrix are obtained by applying individual classifiers. Thus, for each data point an inequality as defined in (19) is used to impose that consensus classifier of equation (17) assigns the highest (and larger than 0.5) probability to the true class of that point. A solution to the set of inequalities defined in (19) provides the coefficients α_i, and thus, a linear combination based classifier as defined in (17).

The resulting LP problems are infeasible for classification problems that are not linearly separable in the feature space of individual classifiers scores. Our strategy to find an approximate solution is to identify a possibly large subset of inequalities that can be satisfied by combining the elastic LP and MaxF heuristic. In other words, we identify a subset of data points that can be classified using linear decision boundaries, with points difficult to classify excluded from the training. Such approximate solutions that achieve high accuracy and have good generalization properties may be found efficiently using interior point methods for LP. The linear decision boundaries are optimized for a subset of data points that are separable. In addition, due to the "central" properties of interior point methods, discussed before, the solutions that we obtain are away from any individual constraint, providing (at least in principle) a wide margin of separation and a good generalization.

Formulating the problem in terms of linear optimization with constraints opens a way for flexible generalizations. For example, one may impose that the margin of separation between the true and other classes should be at least as wide for the consensus classifier as for the individual classifier, which achieves best separation for a given point. This can be achieved by imposing (again for each vector in the training) additional inequalities of the following form:

$$P_c(C_n) - P_c(C_k) \geq \max_{i=1,p} [P_i(C_n) - P_i(C_k)]. \tag{20}$$

Moreover, instead of considering positive and normalized conditional probabilities one may introduce a generalized classification problem in terms of real scores. One may also weaken the condition of equation (19) by decoupling inequalities for classes other than native. Replacing conditional probabilities for the i-th model by the corresponding score, S_i, and introducing one inequality for each non-native state we obtain the following set of inequalities:

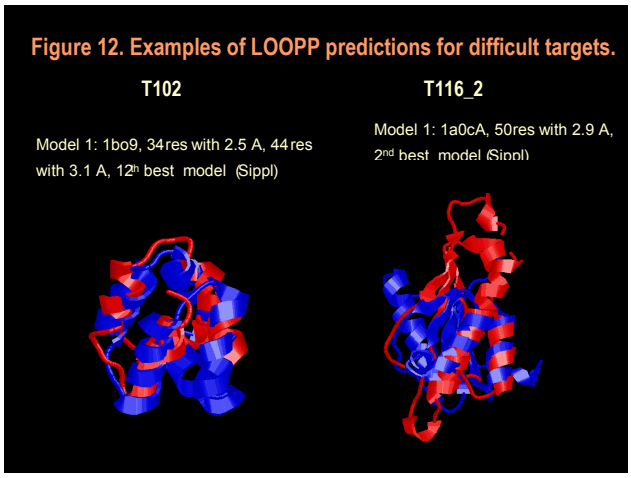
$$\sum_{i=1}^{p} \alpha_i S_i(C_n, \mathbf{x}) \geq \sum_{i=1}^{p} \alpha_i S_i(C_k, \mathbf{x}) \quad \forall k \neq n \ \ \forall \mathbf{x}. \tag{21}$$

The decision is made as previously: the class with the highest score is assigned to each data point. Some preliminary applications of the new approach to protein structure prediction are discussed in Paper 4.

I.9. Biological Applications

Several novel folding potentials, including in particular the THOM2 scoring function defined before and optimized by large-scale LP approach, are incorporated in a software package called Learning, Observing and Outputting Protein Pattern (LOOPP) that we developed. By

blending the contact energies with profile models, in conjunction with large-scale optimization of parameters of the model, THOM2 provides an efficient and accurate threading approach for annotations of remote homologs that share structural similarity without significant sequence similarity.



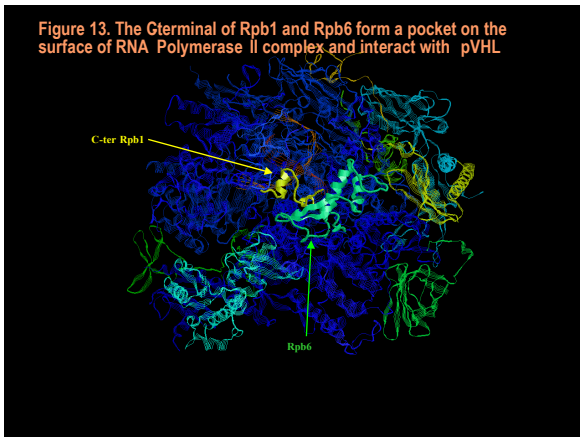
During the second edition of the Critical Assessment of Fully Automated protein Structure Prediction (CAFASP2) the LOOPP server, which is based on threading with THOM2 potential and novel statistical filters, provided best predictions among the servers for three difficult targets (including two targets shown in Figure 12 in blue, for which LOOPP provided one of the best models, which are shown in red in the figure, in the overall CASP competition as well) and was ranked as the third best server in the category of targets with distant homology to known proteins.

The LOOPP server was also included among the eight servers described in the paper summarizing CAFASP2 competition. The overall performance (including "easy" targets) was about average though. This is consistent with earlier observations that protein structure prediction based solely on sequence-to-structure alignments is difficult. Therefore, many fold recognition methods use for example effective combinations of threading and sequence alignments in order to make the overall performance more robust.

LOOPP can be used not only for protein recognition but also to design new potentials and scoring functions for protein folding and protein threading. The LOOPP package and based on it Web server are available from the Cornell Theory Center (http://www.tc.cornell.edu/CBIO/loopp). There were well over 25,000 submissions to the server from more than 5,000 different researchers from all over the world since the server was launched in the summer of 2000. In addition, the program LOOPP has been downloaded by nearly two thousand researches and is used in several research groups.

An early version of our code was able to identify a novel tomato gene that was shown experimentally to control the size of the tomato fruit. We suggested an interesting evolutionary link between this gene and a small GTPase of the Rax (Ras, Ran etc.) family that controls cell division and growth in humans. Mutations and the resulting malfunction of Rax proteins in human can cause uncontrolled cell division and were implicated in many types of cancer. A plant gene, which regulates the tomato fruit size, has been predicted to share a 3D shape with human Ras (oncogene) proteins.

The article reporting this study in the Science magazine (included here as Paper 5 in Part III of this dissertation) and the discovery of the plausible relationship between cancer and tomato fruit growth was discussed in many popular publications and was also described in a textbook: "*A Primer of Genome Science*" by G. Gibson and S. V. Muse, Sinauer Associates, Inc. Publishers, 2002.



Another example of a successful application of threading protocols is the prediction of the interaction between the von Hippel - Lindau protein (pVHL) and the RNA Polymerase II complex. The pVHL tumor suppressor, which plays a central role in post-translational regulation of Hypoxia-Inducible Factor 1-a (HIF-1a) in response to the oxygen stress, was first computationally predicted and then experimentally found to bind to the largest subunit of the major transcription complex of RNA Polymerase II.

The interaction with pVHL was found to regulate ubiquitination and accumulation of RNA Polymerase II. The prediction was possible due to consistency of weak threading matches between a fragment of Hif-1a sequence that contains the pVHL binding motif and two adjacent units (Rpb1 and Rpb6 – see Figure 9) of RNA Polymerase II, as obtained by LOOPP and 3D-PSSM [Sternberg et. al., 1999] threading programs. This discovery of previously unknown mechanism that may regulate the function of the major transcription complex opens a new avenue in the studies on the role of hypoxia in the regulation of transcription processes. This study was published in the Proceedings of the National Academy of Science and included as Paper 6 in Part III of this dissertation.

Furthermore, using homology modeling and structural insights we were able to shed light onto structural basis of histo-blood type dependent susceptibility to Norwalk-like viruses, which cause acute gastroenteritis. Norwalk-like viruses bind to different histo-blood group antigens in a strain specific manner. We were able to identify computationally a putative binding pocket on the surface of the P2 domain of Norwalk-like virus capsid proteins. The residues forming the predicted pocket were subsequently shown by using mutational studies to be indeed involved in binding to histo-blood type antigen receptors. This study was published in the Journal of Virology and included as Paper 7 in Part III of this dissertation.

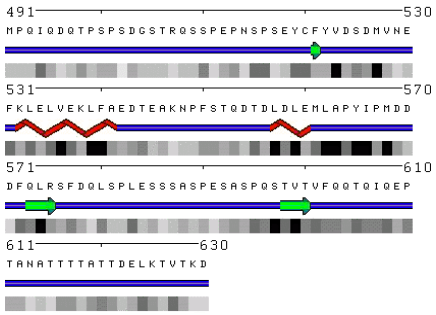
I.10. Future Directions

Our current research efforts focus on interdisciplinary studies involving both development of new methods and algorithms for bioinformatics and computational biology and applications to highly relevant problems that bear a promise of advancing understanding and therapeutic potential for gastroenteritis, AIDS and cancer.

The underlying idea is to blend computational approaches used in the fields of machine learning and data mining as well as functional genomics and proteomics with the expertise in protein structure and protein chemistry in order to enable generation of hypotheses that may facilitate experimental and clinical studies. Such interdisciplinary efforts can help to close the gap between large-scale sequencing projects and their actual outcomes, advancing the understanding of the molecular machinery of life and the mechanisms underlying various disease states, as illustrated by examples discussed in this dissertation. We strongly believe that the post-genomic era paradigm shift will make the kinds of

computational studies that we undertake one of the major sources of hypotheses for biomedical research in the next several years.

Figure 14. Predicted secondary structures and solvent accessibilities for Hif-1a Oxygen -Dependent Domain



We continue to develop new machine learning techniques for large-scale classification problems, novel methods for gene finding and annotation, protein structure and function prediction and protein-protein interactions. In particular, we have recently developed an accurate method for predicting relative solvent accessibilities and secondary structures (http://sable.cchmc.org) that may be used to enhance protein structure prediction and folding simulations as well as computational identification of interaction interfaces in protein complexes. The latter is of particular interest in light of the efforts to understand and subsequently control protein pathways and networks of interactions in the cell.

An example of secondary structures and relative solvent accessibility prediction (using the SABLE server) for the Oxygen-Dependent Domain of HIF-1a is included in Figure 14. This prediction provides further support for the postulated structural similarity between ODD domain of HIF-1a and fragments of Rpb1 and Rpb6 units of RNA Polymerase II discussed in Paper 6. Our predictions were also used to provide structural interpretation and putative functional consequences of polymorphisms in several medically relevant proteins, including Rho GTPases involved in inflammatory responses to pathogens and ENaC epithelium sodium channel implicated in hypertension (see Appendix). It is truly rewarding to see our computational protocols applied to such highly relevant problems in biomedical research.

References

Adler I., Monteiro R.D.C., "Limiting behavior of the affine scaling continous trajectories for linear programming problems", Math. Program. 1991; 50:29-51

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acid Res. 1997, 25:3389-3402

Anfinsen C., "Principles that govern the folding of protein chains", Science, 1973, 181: 223-230

Babajide A., Farber R., Hofacker I.L., Inman J., Lapedes A.S., Stadler P.F., "Exploring protein sequence space using knowledge based potentials", J. Comp. Biol. 1999;

Babajide A., Hofacker I.L., Sippl M.J., Stadler P.F., "Neural networks in protein space: a computational study based on knowledge-based potentials of mean force", Folding Design 1997: 2: 261-269

Baldi P. and Brunak S., "Bioinformatics, the machine learning approach", MIT PressCambridge Massachusetts, London England 1998

Banavar J.R and Maritan A., "Computational Approach to the Protein-Folding Problem", Proteins: Structure, Function, and Genetics, 2001, 42: 433-435

Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M., Sonnhammer E.L., "The Pfam Protein Families Database", Nucleic Acids Research 2002, 30(1): 276-280

Bauer F. and Kohavi R., "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants", Machine Learning 36: 105-139 (1999)

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

Bonneau R., Chivian D., Strauss C.E.M., Carol Rohl C., Baker D., "De Novo Prediction of Three Dimensional Structures for Major Protein Families", J. Mol. Biol. 322 (1): 65, 2002

Bowie J.U., Luthy R., Eisenberg D., "A method to identify protein sequences that fold into a known three-dimensional structure", Science, 1991; 253:164-170

Branden C. and Tooze J., "Introduction to Protein Structure", Garland Publishing, New York and London 1991

Breiman L., "Bagging predictors", Machine Learning 24: 123-140 (1996)

Brown G. and Graves G., "Elastic Programming: A New Approach to Large-Scale Mixed Integer Optimization", presented at ORSA/TIMS conference, Las Vegas, 1975

Bryant S.H., Lawrence C.E., "An empirical energy function for threading protein sequence through folding motif", Proteins 1993; 16:92-112

Chakravarti N., "Some results concerning post-infeasibility analysis", Eur. J. Oper. Res. 73: 139, 1994

Czyzyk J., Mehrotra S., Wagner M. and Wright S., "PCx: An Interior--Point Code for Linear Programming", Optimization Methods and Software 1999, 12: 397-430

Durbin R., Eddy S., Krogh A. and Mitchinson G., "Biological Sequence Analysis", Cambridge University Press 1998

Elofsson A., Fischer D., Rice D.W., Le Grand S., Eisenberg D., "A study of combined structure-sequence profiles", Folding & Design 1998; 1:451-461

Fischer D., "3D-SHOTGUN: A novel, cooperative, fold-recognition meta-predictor", Proteins: Structure, Function, and Genetics 51: 434-441, 2003

Fischer D., Elofsson A., Rychlewski L., Pazos F., Valencia A., Rost B., Ortiz A. R., and Dunbrack R. L., "CAFASP2: The second critical assessment of fully automated structure prediction methods", Proteins: Structure, Function, and Genetics, Suppl. 5: 171-183, 2001

Frary A., Nesbitt T. C., Frary A., Grandillo S., van der Knaap E., Cong B., Liu J., Meller J., Elber R., Alpert K. B., Tanksley S. D., "fw2.2: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size", Science 2000, 289: 85-88

Freund Y. and Schapire R. E., "Experiments with a new boosting algorithm", in L. Saitta, ed., Machine Learning: Proceedings of the Thirteenth National Conference, Morgan Kaufman, pp. 148-156 (1996)

Garey M.R. and Johnson D.S., "Computers and Intractability: A Guide to the Theory of NP-Completeness", W.H. Freeman and Company, New York, 1979

Gibson G. and Muse S.V., "A Primer of Genomic Science", Sinauer Associates Inc. 2002

Godzik A., Kolinski A., Skolnick J., "Topology fingerprint approach to the inverse folding problem", J. Mol. Biol. 1992; 227:227-238

Gusfield D., "Algorithms on Strings, Trees and Sequences. Computer Science and Computational Biology", Cambridge University Press, 1999

Hastie T., Tibshirani R. and J. Friedman J., "The Elements of Statistical Learning", Springer, New York 2001

Henikoff S. and Henikoff J.G., "Amino acid substitution matrices from protein blocks", PNAS USA 1989, 89: 10915-10919

Jones D.T., "GenTHREADER: An Efficient and Reliable Protein Fold Recognition Method for Genomic Sequenecs", J. Mol. Biol. 1999; 287:797-815

Jones D.T., Taylor W.R., Thornton J.M., "A new approach to protein fold recognition", Nature 1992; 358:86-89

Karmakar N. K., "A new polynomial-time algorithm for linear programming", Combinatorica 4: 373-395, 1984

Krogh A. and Vedelsby J., "Neural Network Ensembles, Cross Validation and Active Learning", Advances in Neural Information Processing Systems, MIT Press, 7: 231-238 (1995)

Kuznetsova A. V., Meller J., Schnell P. O., Nash J. A., Sanchez Y., Conaway J. W., Conaway R. C. and Czyzyk-Krzeska M. F., "VHL binds hyperphosphorylated large subunit of RNA Polymerase II through a proline hydroxylation motif and targets it for ubiquitination", PNAS 2003, vol. 100 (5): 2706-2711

Lathrop R.H., "The protein threading problem with sequence amino-acid interaction preferences is NP-complete", Protein Eng. 1994; 7:1059-1068

Lathrop R.H., Smith T.F., "Global optimum protein threading with gapped alignment and empirical pair score functions", J. Mol. Biol. 1996; 255: 641-665

Liwo A., Oldziej S., Pincus M.R., Wawak R.J., Rackovsky S., Scheraga H.A., "A united-residue force field for off-lattice protein structure simulations: functional forms and parameters of long range side chain interaction potentials from protein crystal data", J. Comp. Chem. 1997; 18: 849-873

Maiorov V.N. and Crippen G.M., "Contact potential that recognizes the correct folding of globular proteins", J. Mol. Biol. 1992; 227:876-888

Matsuo Y., Nishikawa K., "Protein structural similarities predicted by a sequence-structure compatibility method", Protein Sci. 1994; 3:2055-2063

Meller J. and Elber R., "Linear Programming optimization and a double statistical filter for protein threading protocols", Proteins: Structure Function and Genetics 2001, 45: 241-261

Meller J. and Elber R., "Protein Recognition by Sequence-to-Structure Fitness: Bridging Efficiency and Capacity of Threading Models", in Computational Methods for Protein Folding: A Special Volume of Advances in Chemical Physics, ed. R. A. Friesner, John Wiley & Sons 2002

Meller J., "Molecular Dynamics", in Encyclopedia of the Human Genome, Nature Publishing Group, Macmillan Publishers Ltd 2003

Meller J., Wagner M. and Elber R., "Maximum feasibility guideline in the design and analysis of protein folding potentials", J. Comp. Chem. 2002, 23: 111-118

Mirny L.A. and Shakhnovich E.I., "Protein structure prediction by threading. Why it works and why it does not", J. Mol. Biol. 1998; 283:507-526

Mjolsness E. and DeCoste D., "Machine learning for science: state of the art and future prospects", Science 2001, 293: 2051-5

Mulgrew B. and Cowan C. F. N., "Adaptive Filters and Equalisers", Kluwer Academic Publ., Boston 1988

Murzin A.G., Brenner S.E., Hubbard T., Chothia C., "SCOP: a structural classification of proteins database for the investigation of sequences and structures", J. Mol. Biol. 1995, 247:536-540

Needleman S.B. and Wunsch C.D., "A general method applicable to the search for similarities in the amino acid sequences of two proteins", J. Mol. Biol. 1970; 48:443-453

Ouzounis C., Sander C., Scharf M., Schneider R., "Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from 3D structures", J. Mol. Biol. 1993; 232:805-825

Panchenko A.R., Marchler-Bauer A., Bryant S.H., "Combination of threading potentials and sequence profiles improves fold recognition", J. Mol. Biol. 2000; 296: 1319-1331

Parker M. and Ryan J., "Finding the minimum weight IIS cover of an infeasible system of linear inequalities", Annals of Mathematics and Artificial Intelligence 1996, 17: 107-126

Pevzner P.A., "Computational Molecular Biology: An Algorithmic Approach", MIT Press, Cambridge, Massachusetts 2000

R. Bonneau, J. Tsai, I. Ruczinski and D. Baker, "Functional inferences from blind *ab initio* protein structure predictions", J. Struct. Biol. 134 (2-3): 186-90, 2001

Schonbrun J., Wedemeyer W. J. and Baker D., "Protein structure prediction in 2002", Curr. Opin. Struct. Biol., Vol. 12, pp. 348-354, 2002

Simons K. T., Kooperberg C., Huang E. and Baker D., "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions", J. Mol. Biol. 268: 209-25, 1997.

Sippl M.J. and Weitckus S., "Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations", Proteins 1992; 13:258-271

Smith T.F. and Waterman M.S., "Identification of common molecular subsequences", J. Mol. Biol. 1981; 147:195-197

Sternberg M.J.E, Bates P.A., Kelley L.A., MacCallum R.M., "Progress in protein structure prediction: assessment of CASP3", Curr. Opin. Struct. Biol. 1999; 9:368-373

Tobi D., Shafran G., Linial N., Elber R., "On the design and analysis of protein folding potentials", Proteins: Structure Function and Genetics 2000, 40: 71-85

Vanderbei R.J., "Linear Programming: Foundations and Extensions", Kluwer Academic Publishers, New York, 1996

van Holde K.E., Johnson W.C. and Ho P.S., "Principles of Physical Biochemistry", Prentice-Hall International , Inc. 1998

Venclovas C., Zemla A., Fidelis K., and Moult J., "Comparison of performance in successive CASP Experiments", Proteins: Structure, Function, and Genetics, Suppl. 5: 163-170, 2001

Vendruscolo M. and Domany E., "Pairwise contact potentials are unsuitable for protein folding", J. Chem. Phys. 1998; 109:11101-11108

Wagner M., Meller J. and Elber R., "Large-Scale Linear Programming Techniques for the Design of Protein Folding Potentials", Mathematical Programming, to appear (2003)

Xia Y., Huang E.S., Levitt M., Samudrala R., "Ab initio construction of protein tertiary structures using a hierarchical approach", J. Mol. Biol. 2000; 300: 171-185

Ye Y., "Interior Point Algorithms: Theory and Analysis", Wiley, 1997

PART II:

Methods and Algorithms

PROTEINS: Structure, Function, and Genetics 45:241–261 (2001)

Linear Programming Optimization and a Double Statistical Filter for Protein Threading Protocols

Jaroslaw Meller[1,2] and Ron Elber[1]*
[1]*Department of Computer Science, Cornell University, Ithaca, New York*
[2]*Department of Computer Methods, Nicholas Copernicus University, Torun, Poland*

ABSTRACT **The design of scoring functions (or potentials) for threading, differentiating native-like from non-native structures with a limited computational cost, is an active field of research. We revisit two widely used families of threading potentials: the pairwise and profile models. To design optimal scoring functions we use linear programming (LP). The LP protocol makes it possible to measure the difficulty of a particular training set in conjunction with a specific form of the scoring function. Gapless threading demonstrates that pair potentials have larger prediction capacity compared with profile energies. However, alignments with gaps are easier to compute with profile potentials. We therefore search and propose a new profile model with comparable prediction capacity to contact potentials. A protocol to determine optimal energy parameters for gaps, using LP, is also presented. A statistical test, based on a combination of local and global *Z*-scores, is employed to filter out false-positives. Extensive tests of the new protocol are presented. The new model provides an efficient alternative for threading with pair energies, maintaining comparable accuracy. The code, databases, and a prediction server are available at http://www.tc.cornell.edu/CBIO/loopp. Proteins 2001;45:241–261.**

Key words: linear programming; potential optimization; decoy structures; threading; gaps

INTRODUCTION

The threading approach[1–8] to protein recognition is an alternative to the sequence-to-sequence alignment. Rather than matching the unknown sequence S_i to another sequence S_j (one-dimensional [1D] matching), we match the sequence S_i to a shape $\mathbf{X}_j$ (three-dimensional [3D] matching). Experiments found a limited set of folds compared with a large diversity of sequences, suggesting the use of structures to find remote similarities between proteins. Thus, the determination of overall folds is reduced to tests of sequence fitness into known and limited number of shapes.

Sequence-to-structure compatibility is commonly evaluated using reduced representations of protein structures. Points in 3D space represent amino acid residues, and an effective energy of a protein is defined as a sum of interresidue interactions. The effective pair energies can be derived from the analysis of contacts in known structures. Such knowledge-based pairwise potentials are widely used in fold recognition,[2,3,6,9–11] ab initio folding,[11–13] and sequence design.[14,15]

Alternatively, one may define the so-called "profile" energy,[1,5,16] taking the form of a sum of individual site contributions, depending on the structural environment of a site. For example, the solvation/burial state or the secondary structure can be used to characterize different local environments. The advantage of profile models is the simplicity of finding optimal alignments with gaps (deletions and insertions into the aligned sequence) that permit identification of homologous proteins of different length. Using the dynamic programming (DP) algorithm,[17–20] optimal alignments with gaps in polynomial time can be computed, as compared with the exponential number of all possible alignments.

In contrast to profile models, the potentials based on pair energies do not lead to exact alignments with dynamic programming. A number of heuristic algorithms, providing approximate alignments, have been proposed.[21] However, they cannot guarantee an optimal solution with a less than exponential number of operations.[22] Another common approach is to approximate the energy by a profile model, the so-called frozen environment approximation (FEA), and to perform the alignment using DP.[23]

In this article, we evaluate several different scoring functions for sequence-to-structure alignments, with parameters optimized by linear programming (LP).[24–26] The capacity of the energies is explored in terms of a number of protein shapes that are recognized with a certain number of parameters. We propose a novel profile model, designed to mimic pair energies, which is shown to have accuracy comparable to that of other contact models. We discuss gap energies and introduce a double *Z*-score measure (from global and local alignments) to assess the results. The proposed threading protocol emphasizes structural fitness (as opposed to sequence similarity) and can be made a part of more complex fold recognition algorithms that use

Grant sponsor: National Institutes of Health; Grant sponsor: DARPA; Grant sponsor: Polish State Committee for Scientific Research; Grant number: 6-P04A-06614).

*Correspondence to: Ron Elber, Department of Computer Science, Cornell University, Upson Hall 4130, Ithaca, NY 14853. E-mail: ron@cs.cornell.edu

Received 11 December 2001; Accepted 4 June 2001

TABLE I. Definitions of Different Groups of Amino Acids Used in the Present Study*

Hydrophobic (HYD)	ALA CYS HIS ILE LEU MET PHE PRO TRP TYR VAL
Polar (POL)	ARG ASN ASP GLN GLY LYS SER THR
Charged (CHG)	ARG ASP GLU LYS
Negatively charged (CHN)	ASP GLU

*Note that 10 types of amino acids are found to be sufficient to solve the Hinds–Levitt set either by pairwise interaction models or by THOM2. The amino acid types are HYD POL CHG CHN GLY ALA PRO TYR TRP HIS. The list implies that when an amino acid appears explicitly, it is excluded from other groups that may contain it. For example, HYD includes in this case CYS, ILE, LEU, MET, and VAL, while CHG includes ARG and LYS only, since the negatively charged residues form a separate group.

family profiles, secondary structures, and other patterns relevant for protein recognition.

THEORY AND COMPUTATIONAL PROTOCOLS

Functional Form of the Energy

In this section, we formally define the families of pairwise and profile models. We also introduce a novel threading onion model (THOM), which is investigated in subsequent sections. In the widely used pairwise contact model, the score of the alignment of a sequence S into a structure $\mathbf{X}$ is a sum of all pairs of interacting amino acids:

$$E_{\text{pairs}} = \sum_{i<j} \phi_{ij}(\alpha_i, \beta_j, r_{ij}) \tag{1}$$

The pair interaction model ϕ_{ij} depends on the distance between sites i and j and on the types of the amino acids, α_i and β_j. The latter are defined by the alignment, as certain amino acid residues are placed in sites i and j, respectively.

Let us consider the widely used contact potential. If the geometric centers of the side-chains are closer than 6.4 Å; the two amino acids are then considered in contact. The total energy is a sum of the individual contact energies:

$$\phi_{ij}(\alpha_i, \beta_j, r_{ij}) = \left\{ \begin{array}{cc} \varepsilon_{\alpha\beta} & 1.0 < r_{ij} < 6.4\,\text{Å} \\ 0 & \text{otherwise} \end{array} \right\} \tag{2}$$

where i, j are the structure site indices (contacts due to sites in sequential vicinity are excluded, $i + 3 < j$), α, β are indices of the amino acid types (we drop subscripts i and j for convenience) and $\varepsilon_{\alpha\beta}$ is a matrix of all the possible contact types. For example, it can be a 20×20 matrix for the 20 amino acids. Alternatively, it can be a smaller matrix if the amino acids are grouped together to fewer classes. Different groups used in the present study are summarized in Table I. The entries of $\varepsilon_{\alpha\beta}$ are the target of parameter optimization.

The second type of energy function assigns "environment," or a profile, to each of the structural sites.[1] The total energy E_{profile} is written as a sum of the energies of the sites:

$$E_{\text{profile}} = \sum_i \phi_i(\alpha_i, \mathbf{X}) \tag{3}$$

As previously, α_i denotes the type of an amino acid that was placed at site i of $\mathbf{X}$. For example, if α_i is a hydrophobic residue, and x_i is characterized as a hydrophobic site, the energy $\phi_i(\alpha_i, \mathbf{X})$ will be low (high score). If α_i is charged, the energy will be high (low score). The total score is given by a sum of the individual site contributions.

We consider two profile models. In threading onion model 1 (THOM1), which was used in the past as an effective solvation potential,[1,2] the total energy of the protein is a direct sum of the contributions from structural sites and can be written as

$$E_{\text{THOM1}} = \sum_i \varepsilon_{\alpha_i}(n_i) \tag{4}$$

The energy of a site depends on two indices: (1) the number of neighbors to the site—n_i (a neighbor is defined by a cutoff distance—eq. 2); and (2) the type of the amino acid at site i—α_i. For 20 amino acids and a maximum of 10 neighbors, we have 200 parameters to optimize, a number comparable to that of the detailed pairwise model.

THOM1 provides a nonspecific interaction energy, which has relatively low prediction ability as compared with pairwise interaction models (see section, Application to Potential Design and Analysis). Threading onion model 2 (THOM2) attempts to improve the accuracy of the environment model, making it more similar to pairwise interactions.

We define the energy $\varepsilon_{\alpha_i}(n_i, n_j)$ of a contact between structural sites i and j, where n_i is the number of neighbors to site i, and n_j is the number of neighbors to site j. The type of amino acid at site i is α_i. Only one of the amino acids in contact is known. The total contribution to the energy of site i is a sum over all contacts to this site

$$\phi_{i,\text{THOM2}}(\alpha_i, \mathbf{X}) = {\sum_j}' \varepsilon_{\alpha_i}(n_i, n_j)$$

The prime indicates that we sum only over sites j that are in contact with i (i.e., over sites j satisfying the condition $1.0 < r_{ij} < 6.4$ Å and $|i - j| \geq 4$). The total energy is finally given by a double sum over i and j:

$$E_{\text{THOM2}} = \sum_i {\sum_j}' \varepsilon_{\alpha_i}(n_i, n_j) \tag{5}$$

It is possible to define an effective contact energy using THOM2:

$$V_{ij}^{\text{eff}} = \varepsilon_{\alpha_i}(n_i, n_j) + \varepsilon_{\alpha_j}(n_j, n_i) \tag{6}$$

TABLE II. Definitions of Contact Types for the THOM2 Energy Model*

Type of site	$n' = 1,2;\bar{1}$	$n' = 3,4,5,6;\bar{5}$	$n' \geq 7;\bar{9}$
$n = 1,2;\bar{1}$	$(\bar{1},\bar{1})$	$(\bar{1},\bar{5})$	$(\bar{1},\bar{9})$
$n = 3,4;\bar{3}$	$(\bar{3},\bar{1})$	$(\bar{3},\bar{5})$	$(\bar{3},\bar{9})$
$n = 5,6;\bar{5}$	$(\bar{5},\bar{1})$	$(\bar{5},\bar{5})$	$(\bar{5},\bar{9})$
$n = 7,8;\bar{7}$	$(\bar{7},\bar{1})$	$(\bar{7},\bar{5})$	$(\bar{7},\bar{9})$
$n \geq 9;\bar{9}$	$(\bar{9},\bar{1})$	$(\bar{9},\bar{5})$	$(\bar{9},\bar{9})$

*There are 15 types of energy terms in THOM2 that are based on contacts in the first and the second contact layers. A contact between two amino acids is "on" if the distance is <6.4 Å. We consider five types of sites in the first layer and three in the second layer. Thus, there are $20 \times 15 = 300$ different energy terms for 20 different amino acids. A reduced set of amino acids is associated with a smaller number of parameters to optimize (for 10 types of amino acids, the number of parameters is $10 \times 15 = 150$). The notation we used for each type of site is based on a representative number of neighbors. The number of neighbors n in a given class and its representative are given in the first column (for different classes of sites in the first layer) and in the first row (for different classes of sites in the second layer). The intersections between columns and rows correspond to contacts of different types: a contact between two sites of medium number of neighbors is denoted by $(\bar{5},\bar{5})$, for example.

Hence, we can formally express the THOM2 energy as a sum of pair energies

$$E_{\mathrm{THOM2}} = \sum_{i<j} V_{ij}^{\mathrm{eff}}$$

The effective energy mimics the formalism of pairwise interactions. However, in contrast to the usual pair potential, the optimal alignments with gaps can be computed efficiently with THOM2, as structural features alone determine the "identity" of the neighbor.

We use a coarse-grained model that leads to a reduced set of structural environments (types of contacts), as outlined in Table II. The use of a reduced set makes the number of parameters (300 when all 20 types of amino acids are considered) comparable to that of the contact potential. Further analysis of the new model is included in the section, Application to Potential Design and Analysis.

Linear Programming Protocol for Optimization of the Energy Parameters

Consider the alignment of a sequence S of length n, into a structure $\mathbf{X}$ of length m. In order to optimize the energy parameters for the amino acid interactions (the gap energies are discussed in the section, Protocol for Optimization of Gap Energies), we employ the so-called gapless threading, in which sequence S is fitted into the structure $\mathbf{X}$ with no deletions or insertions. Hence, the length of the sequence must be shorter than, or equal to, the length of the protein chain. If n is shorter than m, we may try $m - n + 1$ possible alignments varying the structural site of the first residue (and the following sequence).

The energy (score) of the alignment of S into $\mathbf{X}$ is denoted by $E(S, \mathbf{X}, \mathbf{p})$, where $\mathbf{X}$ stands (depending on the context) either for the whole structure or only for a substructure of length n. The energy function $E(S, \mathbf{X}, \mathbf{p})$ depends on a vector $\mathbf{p}$ of q parameters (thus far undetermined).

Consider the sets of structures $\{\mathbf{X}_i\}$ and sequences $\{S_j\}$. There is an energy value for each of the alignments of the sequences $\{S_j\}$ into the structures $\{\mathbf{X}_i\}$. A good potential will make the alignment of the "native" sequence into its "native" structure the lowest in energy. Let $\mathbf{X}_n$ be the native structure. A condition for an exact recognition is

$$E(S_n, \mathbf{X}_j, \mathbf{p}) - E(S_n, \mathbf{X}_n, \mathbf{p}) > 0 \qquad \forall\ j \neq n \tag{7}$$

In the set of inequalities (Eq. 7), the coordinates and sequences are given, and the unknowns are the parameters we need to determine.

The LP protocol makes it possible to measure the difficulty of a training set. The number of parameters of the energy function necessary to satisfy all the inequalities is derived from the set of structures, as defined in eq. 7. Whereas the statistical potentials are based on the analysis of native structures only, the LP protocol is using sequences threaded through misfolded structures during the process of learning. As a result, the LP has the potential for accumulating more information, in an attempt to place the energies of the misfolded sequence as far as possible from the energy of the native state. In fact, the LP protocol can be used to optimize the Z-score of the distribution of energy gaps.[27] In the remainder of this section, we describe the technique to solve the inequalities of eq. 7.

The "profile" and pairwise interaction models considered in this work can be written as a scalar product:

$$E = \sum_{\gamma} n_{\gamma}(\mathbf{X}) p_{\gamma} \equiv \mathbf{n}(\mathbf{X}) \cdot \mathbf{p} \tag{8}$$

where $\mathbf{p}$ is the vector of parameters we wish to determine. The index of the vector γ, is running over the types of contacts or sites. For example, in the pairwise interaction model, the index γ denotes the types of the amino acid pairs, whereas in THOM1, it denotes the types of sites characterized by the identity of the amino acid at the site and the number of its neighbors. $n_{\gamma}(\mathbf{X})$ is the number of contacts, or sites of a specific type found in the structure $\mathbf{X}$.

Using the representation of eq. 8, we may rewrite eq. 7 as follows:

$$\mathbf{p} \cdot \Delta\mathbf{n}_j = \sum_{\gamma} p_{\gamma}[n_{\gamma}(\mathbf{X}_j) - n_{\gamma}(\mathbf{X}_n)] > 0 \qquad \forall\ j \neq n \tag{9}$$

Standard linear programming tools can solve Eq. 9 for $\mathbf{p}$. We use the BPMPD program of Meszaros,[28] which is based on the interior point algorithm. In the present computations, we seek a point in parameter space that satisfies the constraints, and we do not optimize a function in that space. Without a function to optimize the interior point, the algorithm places the solution at the "maximally feasible" point, which is at the analytic center of the feasible polyhedron that defines the "accessible" volume of parameters.[27,29]

The set of inequalities that we wish to solve includes tens of millions of constraints that could not be loaded directly into the computer memory (we have access to machines with 2–4 Gigabytes [Gb] of memory). Therefore,

the following heuristic approach was used. Only a subset of the constraints is considered:

$$\{\mathbf{p}\cdot\Delta\mathbf{n} < C\}_{j=1}^{J}$$

where the threshold of C is chosen to restrict the number of inequalities to a manageable size (~500,000 inequalities for 200 parameters). Hence, during a single iteration, we considered only the inequalities that are more likely to be relevant for further improvement by being smaller than cutoff C. This subset is sent to the LP solver "as is." If proven infeasible, the calculation stops (no solution possible). Otherwise, the result is used to test the remaining inequalities for violations of the constraints (Eq. 9). If no violations are detected, the process is stopped (a solution was found). If negative inner products are found in the remaining set, a new subset of inequalities below C is collected. The process is repeated, until it converges. Sometimes convergence is difficult to achieve, necessitating human intervention in the choices of the inequalities. For example, mixing subsets of inequalities from previous runs with the lowest inequalities obtained with the new parameters helps avoid the problem of "oscillating" between solutions.

Protocol for Optimization of Gap Energies

In this section, we discuss the derivation of the energy terms for gaps and deletions that enable the detection of homologs. We introduce an "extended" sequence, $\bar{S}$, which may include gap "residues" (spaces, or empty structural sites) and deletions (removal of an amino acid, or an amino acid placed at a virtual structural site).

The gap residue, —, is considered another amino acid. We assign to it a score (or energy), $\varepsilon(\underline{\mathbf{X}})$, according to its environment. Gap training is similar to the training of other amino acid residues, in contrast to the usual ad hoc treatment of gap energies. The proposed treatment is also more symmetric than the different penalties for opening and extending gaps.

The database of "native" and decoy structures is different, however, for gapless and gap training. To optimize the gap parameters, we need "pseudo-native" structures that include gaps. We construct such "pseudo-native" conformations by removing the true native shape $\mathbf{X}_n$ of sequence S_n from the coordinate training set and by replacing it with a homologous structure, $\mathbf{X}_h$. The best alignment of the native sequence, S_n, into the homologous structure, $\mathbf{X}_h$, with an initial guess of gap penalties, defines $\bar{S}_n$. The extended sequence, $\bar{S}_n$, with gap residues at certain (fixed from this point on) positions becomes our new (pseudo-) native sequence of the structure $\mathbf{X}_h$.

We require that the alignment of $\bar{S}_n$ into the homologous protein will yield the lowest energy compared with all other alignments of the set. Hence, our constraints are

$$E(\bar{S}_n, \mathbf{X}_j, \mathbf{p}) - E(\bar{S}_n, \mathbf{X}_h, \mathbf{p}) > 0 \qquad \forall\ j \neq h, n \qquad (10)$$

Equation 10 is different from eq. 7 because we consider the extended set of amino acids, $\bar{S}$ instead of S, and the native-like structure is $\mathbf{X}_h$ instead of $\mathbf{X}_n$.

To limit the scope of the computations, we optimize the scores of the gaps only. Thus, we do not allow the amino acid energies optimized separately (see the section, Linear Programming Training of "Minimal" Models) to change while optimizing parameters for gaps. Moreover, the sequence $\bar{S}$, obtained by a certain prior (e.g., structure-to-structure) alignment, or from the experimental data, if available, is held fixed. In other words, threading of the extended sequence with fixed positions and number of gap residues (treated now as any other residue), $\bar{S}$, against all other structures in the training set is used, in order to generate a corresponding set of inequalities, (eq. 10). This optimization, although limited, and clearly not the final word on the topic, is still expected to be better than a guess. Further studies of gap penalties are in progress (T. Galor, J. Meller, and R. Elber, unpublished data). Optimization of gaps has been attempted in the past.[23,30]

In principle, one could optimize deletion penalties using a similar protocol. In this article, we exploit an assumed symmetry between insertion of a gap residue to a sequence and the placement of a "delete" residue in a virtual structural site. The deletion penalty is set equal to the cost of insertion averaged over the two nearest structural sites. No explicit dependence on the amino acid type is assumed.

Double *Z*-Score Filter for Gapped Alignments

In later sections on these assessments we consider optimal alignments of an extended sequence $\bar{S}$ with gaps into the library structures $\mathbf{X}_j$. We focus on the alignments of complete sequences to complete structures (global alignments[17]) and alignments of continuous fragments of sequences into continuous fragments of structures (local alignment[18]). In global alignments, opening and closing gaps (gaps before the first residue and after the last amino acid) reduce the score. In local alignments, gaps or deletions at the C- and N-terminals of the highest-scoring segment are ignored. Only one local segment, with the highest score, is considered.

Threading experiments that are based on a single criterion (the energy) are usually unsatisfactory.[26,31] Although it is our goal that the (free) energy function that we design is sufficiently accurate that the native state (the native sequence threaded through the native structure) is the lowest in energy, this is not always the case. Our exact training is for the training set and for gapless threading only (see the section, Application to Potential Design and Analysis). The results were not extended to include exact learning with gaps, or exact recognition of structures of related proteins that are not the native. Such extensions are difficult, as the number of inequalities for $\bar{S}$ is exponentially larger than the number of inequalities without gaps.

Other investigators use the Z-score as an additional filter or as the primary filter,[19,32,4,6] and we follow their steps. The novelty in the present protocol is the combined use of global and local Z-scores to assess the accuracy of the prediction. This filtering mechanism, in addition to the initial energy filter, provides improved discrimination as compared with a single Z-score test.

The Z-score is a dimensionless "normalized" score, defined as

$$Z = \frac{\langle E \rangle - E_p}{\sqrt{\langle E^2 \rangle - \langle E \rangle^2}} \quad (11)$$

The energy of the current "probe," i.e., the energy of the optimal alignment of a query sequence into a target structure is denoted by E_p. The averages, $\langle \ldots \rangle$, are over "random" alignments. The Z-score measures the deviation of our "hits" from random alignments (alignments with scores far from the "random" average value are more significant). Following common practice,[32–34] we generate the distribution of random alignments numerically, employing sequence shuffling. That is, we consider the family of sequences obtained by permutations of the original sequence. The set of shuffled sequences has the same amino acid composition and length as the native sequence, and all the shuffled sequences have the same energy in the unfolded state (the energy of an amino acid with no contacts is set to zero).

In the section, Assessing the Distribution of Z-Scores for Gapped Alignments, we estimate numerically the probability $P(Z_p)$ of observing a Z-score of greater than Z_p by chance for local threading alignments. The relatively high likelihood of observing large Z-scores for false-positives makes predictions based on the Z-score test problematic. Therefore, we propose an additional filtering mechanism, based on a combination of Z-scores for global and local alignments. The double Z-score filter eliminates false-positives, missing a smaller number of correct predictions.

Global alignments (in contrast to local alignments) are influenced significantly by a difference in the lengths of the structure and the threaded sequence. The matching of lengths was considered too restricted in previous studies.[35] Nevertheless, using environment-dependent gap penalty, the Z-score of the global alignment proved a useful independent filter (see later sections on these assessments). We observe that good scores are obtained for length differences (between sequence and structure) that are on order of 10%. By contrast, when the differences in length are profound the global alignment fails. Large differences imply identification of domains and not a whole protein. This is a different problem, not addressed in the present work.

APPLICATION TO POTENTIAL DESIGN AND ANALYSIS

In this section, we analyze and compare several pairwise and profile potentials, optimized using the LP protocol. Given the training set, either we obtain a solution (exact recognition on the training set), or the LP problem proves infeasible.

We use the infeasibility of a set to test the capacity of an energy model. We compare the capacity of alternative energy models by inquiring how many native folds they can recognize (before hitting an infeasible solution). The larger the number of proteins that are recognized with the same number of parameters, the better the energy model. We find that, in general, the "profile" potentials have lower capacity than that of the pairwise interaction models.

Training and Test Sets

Two sets of protein structures and sequences are used for the training of parameters in the present study. Hinds and Levitt developed the first set,[31] which we call the HL set. It consists of 246 protein structures and sequences. Gapless threading of all sequences into all structures generated the 4,003,727 constraints (i.e., the inequalities of eq. 7). The gapless constraints were used to determine the potential parameters for the 20 amino acids. Because the number of parameters does not exceed a few hundred, the number of inequalities is larger than the number of unknowns by many orders of magnitude.

The second set of structures consists of 594 proteins and was developed by Tobi et al.,[25] which we call the TE set. This set is considerably more demanding. It includes structures chosen according to diversity of protein folds, but also some homologous proteins ($\leq$60% sequence identity), and poses a significant challenge to the energy function. For example, the set is infeasible for threading onion model 1 (THOM1), even when using 20 types of amino acids (see the next section). The total number of inequalities that were obtained from the TE set using gapless threading was 30,211,442. The TE set includes 206 proteins from the HL set.

We developed two other sets that are used as testing sets to evaluate the new potentials in terms of both gapped and gapless alignments. These test sets contain proteins that are structurally dissimilar to the proteins included in the training sets, specified by the root-mean-square deviation (RMS) between the structures. A structure-to-structure alignment algorithm, based on the overlap of the contact shells defined for the superimposed side-chain centers in analogy with THOM2 (disregarding however the identities of amino acids), was used (J. Meller and R. Elber, unpublished results).

The first testing set, referred to as S47, consists of 47 proteins: 25 proteins from the CASP3 competition[36] and 22 related structures, chosen randomly from the list of DALI[37] relatives of the CASP3 targets. Using CASP3-related structures is a convenient way of finding protein structures that are not sampled in the training. None of the 47 structures has homologous counterparts in the HL set, and only six (representing three different folds) have counterparts in the TE set, with a cutoff for structural (dis)-similarity of 12 Å RMS (between the superimposed side-chain centers).

The second test set, referred to as S1082, consists of 1,082 proteins that were not included in the TE set and that are different by $\geq$3 Å RMS (measured, as previously, between the superimposed side-chain centers) with respect to any protein from the TE set and with respect to each other. Thus, the S1082 set is a relatively dense (but nonredundant at $\leq$3 Å RMS) sample of protein families. The training and testing sets are available from the web.[38]

TABLE III. Comparing the Capacity of Different Threading Potentials*

Potential	Hinds–Levitt set	Tobi–Elber set
Pairwise, HP model, par. free	200	456
Pairwise, 10 aa, 55 par	246[a]	504
Pairwise, 20 aa, 210 par	246[a]	530
Pairwise, 20 aa, 210 par	237	594[a]
THOM1, 20 aa, 200 par	246[a]	474
THOM2, 10 aa, 150 par	246[a]	478
THOM2, 20 aa, 300 par	246[a]	428
THOM2, 20 aa, 300 par	236	594[a]

*Capacity for recognition of pairwise and profile threading potentials measured by gapless threading on Hinds–Levitt (HL) and Tobi–Elber (TE) representative sets of proteins (see the section, Training and Test Sets). Threading onion model 1 (THOM1) performs significantly worse than pairwise potentials. THOM2 shows a comparable performance and is able to learn the TE set (see also Table X). For each potential, the number of amino acid (aa) types used and the resulting number of parameters are reported. The number of correct predictions for structures in HL and TE sets is given in the second and third columns, respectively.
[a]The training set used (either HL or TE).

Linear Programming Training of "Minimal" Models

This section addresses the question: What is the minimal number of parameters required to obtain an exact solution for the HL and for the TE sets? By "exact" we mean that each of the sequences picks the native fold as the lowest in energy using a gapless threading procedure. Hence, all the inequalities in eq. 7, for all sequences S_n and structures $\mathbf{X}_j$, are satisfied.

The pairwise model requires the smallest number of parameters (i.e., 55) to solve the HL set exactly (Table III). Only 10 types of amino acids were required: HYD POL CHG CHN GLY ALA PRO TYR TRP HIS (see also Table I). THOM1 and THOM2 require 200 and 150 parameters, respectively, to provide an exact solution on the same (HL) set (Table III). It is impossible to find an exact potential (of the functional forms we examined) for the HL set without (at least) 10 types of amino acids. The potentials optimized on the HL set are then used to predict the folds of the proteins of the TE set. Again, we find that the pairwise interaction model performs better than threading onion models.

An indication that THOM2 is a better choice than THOM1 is included in the next comparison. It is impossible to find parameters that will solve the TE set exactly using THOM1 (the inequalities form an infeasible set). The infeasibility is obtained even if 20 types of amino acids are considered. In contrast, both THOM2 and the pairwise interaction model lead to feasible inequalities if the number of parameters is 300 for THOM2 and 210 for the pairwise potential. The set of parameters that solved the TE set exactly does not solve exactly the HL set, as the latter set includes proteins not included in the TE set.

We have also attempted to solve the TE set using pair energies and THOM2 with a smaller number of parameters. The problem proved infeasible even for 17 different types of amino acids and only very similar amino acids grouped together (Leu and Ile, Arg and Lys, Glu and Asp). Similarly, we failed to reduce the number of parameters by grouping together structurally determined types of contacts in THOM2. Enhancing the range of a "dense" site to be a site of seven neighbors or more also results in infeasibility.

Analysis of THOM2

As discussed earlier, in the section, Theory and Computational Models, the THOM1 potential provides a new set of parameters for an effective solvation model that was used in the past. Because in applying the LP protocol we can only solve the HL set, the solution for that set gives our optimal THOM1 energies, as included in Table IVA. In this section, we analyze THOM2 in detail, which has significantly higher capacity than THOM1. However, the double layer of neighbors makes the results more difficult to understand.

Figure 1 presents a contour plot of the total contributions of different types of contacts to the native energies of the native alignments in the TE set. The plots show the energy contributions as a function of the number of neighbors of the primary site (with known amino acid identity) and the number of contacts to a secondary site, n', respectively. The results for two types of residues, lysine and valine, are presented. The contribution of a given type of contact is defined as $f \cdot \varepsilon_\alpha(n, n')$, where $\varepsilon_\alpha(n, n')$ is the energy of a given type of contact, and f is the frequency of that contact, observed in the TE set.

It is possible to find a very attractive (or repulsive) site that makes only negligible contribution to the native energies, since it is extremely rare (i.e., f is small). Table V displays specific examples. By plotting $f \cdot \varepsilon_\alpha(n, n')$, we emphasize the important contributions. Hydrophobic residues with a large number of contacts stabilize the native alignment, as opposed to polar residues that stabilize the native state only with a small number of neighbors.

It has been suggested that pairwise interactions are insufficient to fold proteins, and higher-order terms are necessary.[26] It is of interest to check whether the environment models that we use catch cooperative, many-body effects. As an example, we consider the cases of valine–valine and lysine–lysine interactions. We use eq. 6 to define the energy of a contact. In the usual pairwise interaction, the energy of a valine–valine contact is a constant and is independent of other contacts that the valine may have.

Table VI lists the effective energies of contacts between valine residues as a function of the number of neighbors in the primary and secondary sites. The energies differ widely from -1.46 to $+3.01$. The positive contributions refer to very rare type of contact. The plausible interpretation is that these rare contacts are used to enhance recognition in some cases, due to specific "homologous features." Significant differences are observed also for the frequently occurring types of contacts that contribute in accord with the "general principle" of rewarding contacts between hydrophobic sites. For example, the effective energies of contacts between valine of five neighbors with

TABLE IV. Parameters of Some Threading Potentials Trained Using the LP Protocol*

A: THOM1[a]

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
(1)	−0.02	0.10	−0.22	0.02	−0.13	0.02	0.05	−0.05	−0.15	−0.17	−0.04	0.13	−0.40	−0.52	0.29	−0.02	0.02	−0.20	−0.23	−0.16
(2)	−0.06	−0.23	−0.07	0.20	−0.37	0.21	−0.03	−0.06	−0.05	−0.30	−0.22	0.12	−0.20	−0.25	0.24	−0.01	−0.10	−0.57	−0.27	−0.25
(3)	−0.02	−0.01	−0.01	0.43	−0.72	0.09	0.10	0.05	−0.25	−0.48	−0.37	0.19	−0.66	−0.58	0.06	0.05	−0.12	−0.77	−0.37	−0.38
(4)	−0.17	0.12	0.29	0.37	−0.70	0.22	0.40	0.14	−0.31	−0.64	−0.41	0.60	−0.50	−0.68	0.22	0.00	0.21	−0.36	−0.39	−0.36
(5)	−0.13	0.22	0.20	0.68	−1.13	0.33	0.45	0.38	0.24	−0.53	−0.50	0.37	−0.39	−0.65	0.31	0.31	0.02	−0.65	−0.78	−0.51
(6)	0.02	0.32	0.17	0.43	−1.16	0.02	0.70	0.42	0.36	−0.57	−0.58	0.63	−0.80	−0.82	0.75	0.27	0.24	−0.46	−0.72	−0.51
(7)	0.12	−0.10	0.30	0.43	−1.27	0.46	0.39	0.20	0.27	−0.76	−0.54	0.73	−0.44	−0.40	0.42	0.09	0.36	0.12	−0.39	−0.78
(8)	−0.07	0.91	−0.12	−0.01	−1.60	0.51	0.83	0.29	−0.71	−1.37	−0.72	0.57	−0.66	0.25	0.02	0.36	0.15	−0.26	−0.74	−0.59
(9)	0.83	1.36	0.11	0.35	−1.71	0.82	10.00	2.12	3.38	−0.33	1.03	10.00	1.66	−1.03	1.13	2.23	−0.57	10.00	−0.38	−0.13
(10)	1.57	10.00	10.00	10.00	10.00	10.00	10.00	0.83	10.00	−0.93	−0.47	10.00	10.00	0.40	10.00	10.00	10.00	−0.78	10.00	0.71

B. THOM2[b]

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
(1,1)	0.23	−0.03	−0.03	−0.08	−0.82	−0.26	0.09	0.29	0.07	−0.12	−0.16	−0.02	0.21	−0.20	0.03	0.05	−0.07	−0.50	−0.64	−0.28
(1,5)	−0.21	−0.26	−0.10	0.20	−1.11	0.00	−0.08	0.00	0.03	−0.31	−0.23	−0.13	−0.15	−0.29	−0.23	0.07	−0.09	−0.60	−0.40	−0.36
(1,9)	−6.01	−4.09	−5.42	−6.14	−7.27	−5.88	−5.80	−5.81	−4.75	−5.46	−5.85	−4.91	−4.97	−5.83	−6.17	−5.89	−5.89	−5.25	−6.79	−6.99
(3,1)	−0.01	−0.10	−0.17	0.02	−0.50	−0.09	0.11	0.31	0.04	−0.10	−0.10	0.11	−0.20	−0.17	−0.02	0.40	0.06	−0.31	−0.29	−0.05
(3,5)	−0.08	0.18	0.15	0.13	−0.69	0.12	0.24	0.04	−0.03	−0.29	−0.21	0.14	0.08	−0.32	−0.05	0.06	0.08	−0.36	−0.28	−0.17
(3,9)	−0.29	0.06	−0.33	0.08	−0.78	0.18	0.02	−0.13	−0.47	−0.60	−0.49	0.09	−0.85	−0.07	0.19	0.23	0.15	−0.15	0.03	−0.27
(5,1)	0.13	−0.21	0.04	0.22	−0.15	−0.11	0.08	0.48	0.19	−0.15	−0.32	−0.06	−0.15	−0.27	0.17	0.19	0.34	−0.07	0.02	0.19
(5,5)	0.06	0.16	0.20	0.17	−0.60	0.04	0.13	0.18	−0.04	−0.25	−0.19	0.26	−0.26	−0.28	0.09	0.11	0.02	−0.36	−0.30	−0.27
(5,9)	−0.65	0.68	−0.26	−0.19	−0.82	−0.09	0.43	−0.36	−0.19	−0.47	−0.42	0.34	0.32	0.07	0.55	0.22	0.01	0.04	−0.46	−0.58
(7,1)	6.29	5.50	5.56	6.02	5.09	5.55	5.68	6.10	5.70	5.59	5.26	6.08	5.64	5.80	5.82	5.23	5.48	6.42	5.17	5.53
(7,5)	0.17	0.29	0.36	0.39	−0.28	0.28	0.45	0.33	0.28	−0.08	−0.01	0.50	0.24	−0.16	0.42	0.13	0.34	0.04	−0.08	−0.03
(7,9)	0.08	0.41	0.00	−0.15	−0.30	0.04	−0.27	0.05	0.69	0.04	−0.17	0.67	0.06	0.03	−0.71	0.82	0.24	−0.36	0.14	−0.25
(9,1)	10.00	4.50	6.05	5.21	4.00	5.94	10.00	10.00	10.00	10.00	6.22	5.59	4.91	6.02	9.61	10.00	10.00	5.88	10.00	10.00
(9,5)	0.26	0.30	0.26	0.71	0.41	−0.02	0.32	0.83	−0.09	1.26	−0.15	0.52	−0.19	0.43	3.07	0.43	0.52	−0.08	0.08	0.21
(9,9)	0.20	0.04	−0.37	−1.34	−1.19	0.47	1.37	−1.36	1.06	−1.99	−0.25	−0.29	1.41	−1.33	6.94	3.22	−0.54	0.81	−0.53	−0.52

[a]Numerical values of the energy parameters for threading onion model 1 (THOM1) potential trained on the Hinds–Levitt (HL) set of proteins.
[b]Numerical values of the energy parameters for threading onion model 2 (THOM2) potential trained on the Tobi–Elber (TE) set of proteins.
*Rows correspond to either different types of sites (THOM1) or contacts (THOM2). Columns correspond to different types of amino acids. See text for details.

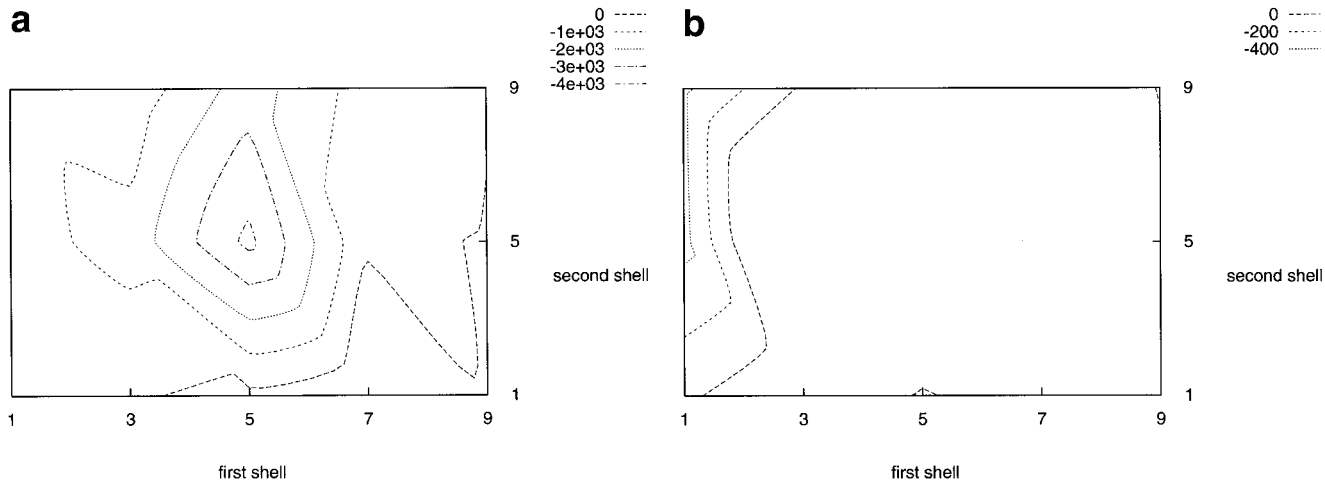


Fig. 1. Contour plots of the total energy contributions to the native alignments in the Tobi–Elber (TE) set for valine and lysine residues as a function of the number of neighbors in the first and second shells. **a:** Contacts involving valine residues with five to six neighbors with other residues of medium number of neighbors stabilize most of the native alignments. **b:** Only contacts involving lysine residues with a small number of neighbors stabilize native alignments.

another valine of three, five, or seven neighbors are equal to −0.44, −0.54, −0.61, respectively. Hence, THOM2 includes significant cooperativity effects. The optimal parameters for THOM2 potential are provided in Table IVB.

Training of Gap Energies

In this section, we apply the linear protocol for the optimization of gap energies described earlier. Training concerns the gap energies for THOM2 only, and it is limited to a small set of carefully chosen homologous pairs. Despite the limited scope of our training, we obtain satisfactory results in terms of recognition of remote homologues, as discussed subsequently.

Pairs of homologous proteins from the following families were considered: globins, trypsins, cytochromes, and lysozymes (Table VII). The families were selected to represent different folds. The globins are helical, trypsins are mostly β-sheets, and lysozymes are α/β proteins. Note also that the number of gaps differs appreciably from a protein to a protein. For example, $\bar{S}_n$ includes only one gap for

TABLE V. Characterization of Native and Decoy Structures*

A: THOM1[a]

Type of site[a]	Native (HYD/POL)	Decoys (HYD/POL)
(1)	16.97 (4.89/12.09)	24.20 (11.72/12.48)
(2)	17.30 (6.06/11.24)	21.72 (10.52/11.20)
(3)	17.72 (8.29/9.43)	18.70 (9.06/9.64)
(4)	16.60 (9.68/6.92)	15.00 (7.28/7.73)
(5)	14.62 (10.16/4.47)	10.79 (5.24/5.55)
(6)	9.96 (7.66/2.30)	6.04 (2.94/3.10)
(7)	4.95 (4.02/0.92)	2.63 (1.28/1.35)
(8)	1.57 (1.32/0.25)	0.77 (0.38/0.40)
(9)	0.26 (0.21/0.05)	0.12 (0.06/0.06)
(10)	0.04 (0.04/0.00)	0.02 (0.01/0.01)

B: THOM2[b]

Type of contact	Native (HYD/POL)	Decoys (HYD/POL)
$(\bar{1},\bar{1})$	5.09 (1.59/3.50)	11.34 (5.48/5.85)
$(\bar{1},\bar{5})$	9.02 (2.99/6.04)	12.69 (6.15/6.54)
$(\bar{1},\bar{9})$	0.41 (0.15/0.26)	0.35 (0.17/0.18)
$(\bar{3},\bar{1})$	6.25 (2.88/3.37)	9.51 (4.60/4.91)
$(\bar{3},\bar{5})$	24.09 (13.01/11.08)	26.59 (12.91/13.68)
$(\bar{3},\bar{9})$	3.23 (1.88/1.35)	2.29 (1.12/1.18)
$(\bar{5},\bar{1})$	2.77 (1.81/0.96)	3.18 (1.54/1.64)
$(\bar{5},\bar{5})$	28.36 (20.96/7.40)	22.09 (10.75/11.34)
$(\bar{5},\bar{9})$	6.85 (5.11/1.74)	3.84 (1.87/1.96)
$(\bar{7},\bar{1})$	0.40 (0.31/0.09)	0.34 (0.16/0.17)
$(\bar{7},\bar{5})$	9.56 (8.00/1.56)	5.84 (2.85/3.00)
$(\bar{7},\bar{9})$	3.21 (2.60/0.61)	1.54 (0.75/0.79)
$(\bar{9},\bar{1})$	0.01 (0.01/0.00)	0.01 (0.01/0.01)
$(\bar{9},\bar{5})$	0.52 (0.44/0.08)	0.29 (0.15/0.14)
$(\bar{9},\bar{9})$	0.23 (0.19/0.04)	0.09 (0.05/0.05)

*Overall site and contact distributions are split into distributions for hydrophobic and polar residues (as defined in Table I), given in the parentheses.

[a]Frequencies of different types of sites, relevant for training of threading model 1 (THOM1), found in the native structures of the Hinds–Levitt (HL) set, as opposed to decoy structures generated using the HL set. In THOM1, the type of site is defined by number of its neighbors (n). Frequencies are defined by the percentage from the total number of 53,012 native sites in the HL set and 556.14 millions of decoy sites generated using the HL set, respectively.

[b]Frequencies of different types of contacts, appropriate for training of threading onion model 2 (THOM2), found in the native structures of the Tobi–Elber (TE) set, as opposed to decoy structures generated using TE. Different classes of contacts are specified in Table II. Frequencies are defined by the percentage from the total number of 439,364 native contacts in the TE set and 10089.19 millions of decoy contacts generated using the TE set, respectively.

TABLE VI. Cooperativity in Effective Pairwise Interactions of the THOM2 Potential*

A: VAL residues[a]

	$V(\bar{1})$	$V(\bar{3})$	$V(\bar{5})$	$V(\bar{7})$	$V(\bar{9})$
$V(\bar{1})$	−0.56	−0.41	−0.17	−1.46	3.01
$V(\bar{3})$	−0.41	−0.34	−0.44	−0.30	−0.07
$V(\bar{5})$	−0.17	−0.44	−0.54	−0.61	−0.38
$V(\bar{7})$	−1.46	−0.30	−0.61	−0.49	−0.76
$V(\bar{9})$	3.01	−0.07	−0.38	−0.76	−1.03

B: LYS residues[b]

	$K(\bar{1})$	$K(\bar{3})$	$K(\bar{5})$	$K(\bar{7})$	$K(\bar{9})$
$K(\bar{1})$	−0.03	−0.03	−0.19	1.18	0.69
$K(\bar{3})$	−0.03	0.28	0.40	0.58	0.61
$K(\bar{5})$	−0.19	0.40	0.52	0.83	0.86
$K(\bar{7})$	1.18	0.58	0.83	1.34	0.38
$K(\bar{9})$	0.69	0.61	0.86	0.38	−0.59

*For a pair of two amino acids α and β in contact, we have 25 different possible types of contacts (and consequently 25 different effective energy contributions) as α and β may occupy sites that belong to one of the five different types characterized by the increasing number of contacts in the first contact shell (see Table II). Moreover, the 5×5 interaction matrix will be in general asymmetric.

[a]Effective energies of contact between two VAL residues with a different number of neighbors.

[b]Effective energies of contacts between two LYS residues.

alignment of 1ccr (sequence) versus 1yea (structure), and 22 gaps for 1ntp versus 2gch. The structures of the lysozymes 1lz5 and 1lz6 include engineered insertions that allow us to sample experimentally observed gap locations.

For the remaining families, the process of generating pseudo-native sequences is as follows. For each pair of native and homologous proteins, the alignment of the native sequence $\bar{S}_n$ into the homologous structure $\mathbf{X}_h$ is constructed using the THOM1 potential, with an initial guess for the gap energies, provided in Table VIIIA. The ad hoc gap penalties favor gaps at sites with few neighbors, and they satisfy the following constraints: (1) the gap penalty should increase with the number of neighbors; (2) the energy of a gap with n contacts must be larger than the energy of an amino acid with the same number of contacts (the gap energy must be higher; otherwise, gaps will be preferred to real amino acids); and (3) the energy of amino acids without contacts is set to zero, and therefore the gap energy is greater than zero. Given these constraints, the initial gap penalties are tuned up to minimize the discrepancies with the DALI[37] structure-to-structure alignments (we choose not to use the DALI alignments directly, since they involve deletions that are not trained explicitly at this stage; see the section, Protocol for Optimization of Gap Energies.

The "pseudo-native" structures with extended sequences, obtained as described above, are added to the HL set (while removing the original native structures). The energy functional form we used for the gaps is the same as for other amino acids in THOM2. "Gapless" threading into other structures of the HL set generates about 200,000 constraints for the gap energies, which are solved using the LP solver. The resulting gap penalties for THOM2 are given in Table VIIIB. The value of 10 is the maximal penalty allowed by the optimization protocol we used. The maximal penalty is assigned to gaps found only in decoy states and that have no native states to bound the penalty at lower values. For example, using our initial guess for gap penalties, we do not observe gaps at the hydrophobic cores of pseudo-native structures. Gaps are usually found in loops with significant solvent exposure, and we have no information in our training set on "native" gaps in sites that are neighbor-rich.

Table IX presents the results of optimal threading with gaps (using dynamic programming) for myoglobin (1mba) against leghemoglobin (1lh2) structure. We show the

TABLE VII. Pairs of Homologous Structures Used for Training of Gap Penalties

Native[a]	Homologous[a]	Similarity[b]
1mba (myoglobin, 146)	1lh2 (leghemoglobin, 153)	20%, 2.8 Å, 140 res
1mba (myoglobin, 146)	1babB (hemoglobin, chain B, 146)	17%, 2.3 Å, 138 res
1ntp (β-trypsin, 223)	2gch (γ-chymotrypsin, 245)	45%, 1.2 Å, 216 res
1ccr (cytochrome c, 111)	1yea (cytochrome c, 112)	53%, 1.2 Å, 110 res
1lzl (lysozyme, 130)	1lz5 (1lzl + 4 res insert, 134)	99%, 0.5 Å, 130 res
1lzl (lysozyme, 130)	1lz6 (1lzl + 8 res insert, 138)	99%, 0.3 Å, 129 res

[a]For each pair, the native and the homologous structures are specified by their Protein Data Bank (PDB) codes, names, and lengths, respectively.
[b]The similarity between the native and the homologous proteins is defined in terms of the sequence identity (%), root-mean-square (RMS) distance (Å), and length (number of residues), as defined by structure-to-structure alignments obtained by submitting the corresponding pairs to the DALI server.[37]

TABLE VIII. Gap Penalties for THOM2 Model as Trained by the LP Protocol*

A: THOM1[a]	
Type of site	Penalty
(0)	0.1
(1)	0.3
(2)	0.6
(3)	0.9
(4)	2.0
(5)	4.0
(6)	6.0
(7)	8.0
(8)	9.0
(9)	10.0
B: THOM2[b]	
Type of contact	Penalty
(0)	1.0
($\bar{1}$,$\bar{1}$)	8.9
($\bar{1}$,$\bar{5}$)	5.7
($\bar{1}$,$\bar{9}$)	10.0

*The limited set of homologous structures presented in Table VII is used.
[a]Initial guess of gap penalties for different types of sites in threading onion model 1 (THOM1).
[b]Optimized gap penalties for different types of contacts in threading onion model 2 (THOM2). Penalties that are not specified explicitly are equal to the maximum value of 10.0. Note that the training is limited and will be extended in a future work.

initial alignment (with the ad hoc gap parameters in Table VIIIA), defining the pseudo-native state, and the results for optimized gap penalties for THOM2. The location of gaps in the initial alignment is largely consistent with the DALI[37] structure-to-structure alignment. Four out of seven insertions coincide with the DALI superposition of the two structures, two insertions are shifted by three residues (see footnote to Table IX). The THOM2 alignment (different from the initial setup) is less consistent with the DALI alignment. Interestingly, however, it provides a better superposition of α-helices. The gaps appear (as expected) in loop regions (e.g., the CD, EF, and GH loops). An exception is the gap at position 9 (in 1lh2), located in the middle of the α-helix instead of position 2, as suggested by the DALI alignment. Further tests of alignments with gaps are presented in the section we present threading results for the pairwise TE potential (see the section, Tests of the Model).

To compute optimal alignments with the FEA, we need to set gap penalties for the TE potential. Pairwise models are not the focus of our study, and we do not attempt to optimize gap energies for the TE potential. Therefore, for the sake of fair comparison, we introduce ad hoc gap penalties based on a similar functional model, for both the TE and THOM2 potentials.

After some experimentation, the insertion penalties are chosen to be proportional to the number of neighbors to a site, $\varepsilon_{-}^{\mathrm{TE}}(n) = 0.2 \cdot (n + 1)$ and $\varepsilon_{-}^{\mathrm{THOM2}}(n) = 1.0 \cdot (\langle n \rangle + 1)$, for the TE and THOM2 potentials (the averaged number of neighbors, $\langle n \rangle$, in a class n belongs to, is used for THOM2; see Table II), respectively. This choice is consistent with the trained THOM2 gap energies, which also penalize sites of no neighbors. The proportionality coefficients were gauged using the same families used to train THOM2 gap energies. However, no LP training was attempted. The deletion penalties are also consistent with the THOM2 model, and they are defined as described in the section, Protocol for Optimization of Gap Energies. For further comparisons with sequence-to-sequence alignments, we also introduce environment-dependent gap penalties that are used for family recognition in conjunction with the BLOSUM50[39] substitution matrix,

$$\varepsilon_{-}^{B50}(n) = (5 - n) - 8$$

(see the section, Assessing the Specificity of the Protocol).

Assessing the Distribution of *Z*-Scores for Gapped Alignments

In this section, we compute numerical distributions of the *Z*-scores for local and global threading alignments, using THOM2 and the gap penalties shown in Table VIIIB. On the basis of these distributions, we derive empirical cutoffs for the double *Z*-score test (discussed in the section, Double *Z*-Score Filter for Gapped Alignments) that filters out all the incorrect predictions observed in our benchmark. Further tests of the specificity, as well as sensitivity of the double *Z*-score filter, are included in the following sections.

TABLE IX. Comparison of Alignments of Myoglobin (1mba) Sequence into Leghemoglobin (1lhz) Structure*

A: THOM1[a]	
.....1.....2.....3.....4.....5.....	1–59
SLS**AAEADLAGKSWAPVFA**N**KNANGLDFLVALFEK**F**PDSANF**FADFKGKS**VADIK**ASP**K**	1mba
GAL**TESQAALVKSSWEEFNANIPKHTHRFFILVLEI**A**PAAKDL**FSFLKGTSEVPQN**NPE**	1lh2
.....1.....2.....3.....4.....5.....	1–59
6.....7.....8.....ii...9.....0.....1.....	60–116
LRDVSSRIFTRLNEFVNNAAN**AGKMSA--MLSQFAKEHVG**FGVG**SAQFENVRSMFPGFV**	1mba
LQAHAGKVFKLVYEAAIQLEVTGVVVTD**ATLKNLGSVHVSKGVADAHFPVVK**EA**ILKTI**	1lh2
6.....7.....8.....9.....0.....1.....	60–118
...2..i..i.....3.....4..i...i.i	117–146
ASVAAP-PA-G**ADAAWTKLFGLIIDALK-AA**G-A-	1mba
KEVVGAKW**SEELNSAWTIAYDELAIVIKKEMDDA**A	1lh2
.2.....3.....4.....5...	119–153
B: THOM2[b]	
.....i.1.....2.....3.....4.....i....i.i.	1–55
SLS**AAEAD-LAGKSWAPVFA**N**KNANGLDFLVALFEK**F**PDSANF**FADFKGK-S**VAD-I-K**	1mba
GAL**TESQAALVKSSWEEFNANIPKHTHRFFILVLEI**A**PAAKDL**FSFLKGTSEVPQN**NPE**	1lh2
.....1.....2.....3.....4.....5.....	1–59
....6.....7.....i.8.i.....9.....0.....1..	56–112
ASP**KLRDVSSRIFTRLNEFVN**NA-AN**A-GKMSAMLSQFAKEHVG**FGVG**SAQFENVRSMF**	1mba
LQAHAGKVFKLVYEAAIQLEVTGVVVTD**ATLKNLGSVHVSKGVADAHFPVVK**EA**ILKTI**	1lh2
6.....7.....8.....9.....0.....1.....	60–118
.....2.i.....3.....4.....	113–146
PGFVASVAA-PPAG**ADAAWTKLFGLIIDALKA**AGA	1mba
KEVVGAKW**SEELNSAWTIAYDELAIVIKKEMDDA**A	1lh2
.2.....3.....4.....5...	119–153

*The location of insertions in the initial alignment (which is used for training of gap energies) is largely consistent with the DALI structure-to-structure alignment,[37] which aligns: residues 2–50 of 1mba to 3–51 of 1lh2, residues 53–56 of 1mba to 52–55 of 1lh2 (implying deletions at positions 51 and 52 in 1mba), residues 59–80 of 1mba to 56–77 of 1lh2, residues 81–86 of 1mba to 82–87 of 1lh2, residues 87–121 of 1mba to 89–123 (with the implied insertion at position 88 in 1lh2), residues 122–139 of 1mba to 126–143 of 1lh2 (implying two insertions at positions 124 and 125 in 1lh2), and residues 140–145 of 1mba to 145–150 of 1lh2 (with an insertion at position 144 in 1lh2), respectively. α-Helices in both structures are marked in boldface. Note that F- and G-helices are shifted considerably in the DALI alignment (there is no counterpart for the D-helix in 1lh2). The initial THOM1 alignment is in perfect agreement with the DALI superposition between residues 88 and 150 of 1lh2, except for two insertions at positions 128 and 147 (shifted by three residues with respect to the DALI alignment). The insertions at positions 88, 125, 151, and 153 coincide with the DALI alignment. The THOM2 alignment, with trained gap penalties of table 9.B, is in perfect agreement with the DALI superposition for residues 10–50 of 1lh2 (including A-, B-, and C-helices) and then departs from the DALI alignment, overlapping E-, F-, and G-helices with a smaller shift.

[a]Threading onion model 1 (THOM1) alignment with the initial gap penalties.

[b]Threading onion model 2 (THOM2) alignment with trained gap penalties.

To establish a cutoff for the Z-score (and not the energy itself) that eliminates false-positives, we estimate numerically the probability $P(Z_p)$ of observing a Z-score larger than Z_p by chance. The distribution of Z-scores for random alignments is generated by threading sequences of the S47 set through structures included in the HL set. The probe sequences of known structures were selected to ensure no structural similarity between the HL set and the structures of the probe sequences (see the section, Training and Test Sets). Therefore, any significant hit in this set may be regarded as a false-positive. Z-scores of local alignments are employed to estimate $P(Z_p)$. The number of local alignments with "good" energies (significantly lower than zero) is large, underlying the need for an additional selection mechanism to eliminate false-positives.

In local alignments, a contribution due to a given contact should be only included if it belongs to the alignment (which is not known to start with). This implies a "structural" FEA (see also the section, Assessing Protein Family Signals and the Sensitivity of the Protocol). When counting contacts, we assume that the sites in contact in the original structure belong to the aligned part of the structure. This may result in spuriously low energies of local matches, making the Z-score of the local threading alignment an important filter.

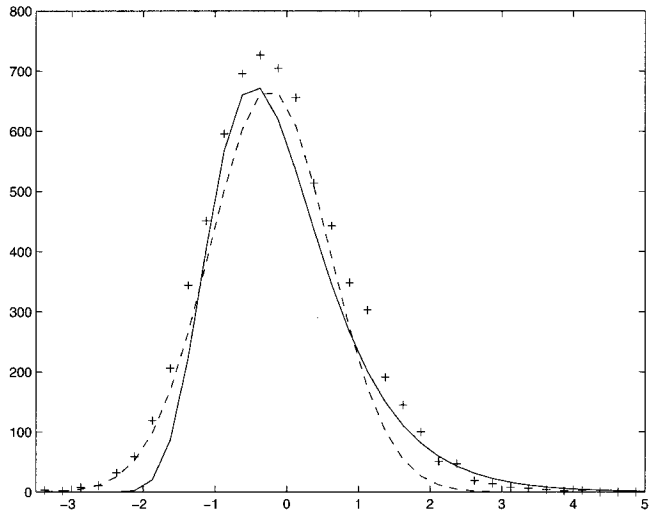


Fig. 2. Probability distribution function of the *Z*-scores computed with local threading alignments for the population of false-positives. A set of 47 sequences of proteins included in the S47 set is used to sample the distribution of the *Z*-scores for false-positives (proteins of the S47 set have no homologues in the Hinds–Levitt (HL) set; see text for details). Each of the sequences is aligned to all the structures included in the HL set. The *Z*-scores are calculated for the 200 best matches (according to energy), using 100 shuffled sequences. The observed distribution of *Z*-scores for 6,813 local threading alignments is represented by +. Note the significant tail to the right, indicating a relatively high likelihood of observing false-positives with large *Z*-scores. The dotted line shows the attempted analytical fit to the gaussian distribution, whereas the solid line the attempted fit to the extreme value distribution (EVD). Note that actual distribution deviates significantly from both. According to the analytical fit to the EVD, the probability of observing a *Z*-score larger than Z_P by chance is equal to $P(Z_P) = 1 - \exp\{-\exp[-1.313 \cdot (Z_P + 0.466)]\}$ with the 98% confidence intervals: 1.313 ± 0.112 and 0.466 ± 0.079. For example, the probability of observing by chance a *Z*-score of >4 = 0.003. We emphasize, however, that the analytical fit to the extreme value distribution provides an upper bound for the observed number of observed false-positives.

As can be seen in Figure 2, the attempted analytical fit to the gaussian distribution underestimates the tail of the observed distribution. The analytical fit to the extreme value distribution,[40] in turn, provides an upper bound for the tail. In the realm of sequence comparison, the extreme value distribution has been used to model scores of random sequence alignments for both local ungapped alignments,[41] as well as local alignments with gaps.[42,43] However, we establish our thresholds on the basis of the numerical distribution.

The number of random alignments with a *Z*-score of >3, for example, is non-negligible (see the tail in Fig. 2 as well as the analytical estimate in the legend of Fig. 2). The expected number of false-positives observed in N trials is $N \cdot P(Z_p)$. Therefore, only relatively high *Z*-scores (that would miss, at the same time, many correct predictions) may result in significant predictions, when searching large databases. Restricting the *Z*-score test only to best matches (according to energy) is insufficient. We find that the double *Z*-score filter performs better, eliminating false-positives with a smaller number of correct predictions that are dismissed as insignificant.

Figure 3 displays the joint probability distribution for global and local *Z*-scores for a population of false-positives versus a population of correct predictions. The squares at the upper right corner represent correct predictions, resulting from 331 native alignments (of a sequence into its native structure) and homologous alignments (of a sequence into a homologous structure) of the HL set proteins. The circles at the lower left corner are incorrect predictions (false-positives) obtained from the alignments of the sequences of the S47 set against all structures in the HL.

The procedure is the same as the one used previously to generate the probability density function for the *Z*-scores of local alignments. However, the *Z*-scores are computed using 1,000 shuffled sequences for both global and local alignments, which is sufficient for convergence. The converged results somewhat reduce the tails of the distribution. For example, the number of false-positives with a global *Z*-score greater than 2.5 and a local *Z*-score greater than 1.0 is equal to 3, as compared with 7 with only 100 shuffled sequences.

Figure 3 shows that the thresholds of 3.0 for global *Z*-scores and of 2.0 for local *Z*-scores are sufficient to eliminate all the false predictions. These cutoffs result in a number of misses (see also the next section). However, this is the price we have to pay for high confidence in our predictions. The total number of pairwise alignments for which we compute the global and the local *Z*-scores, and subsequently test for the presence of false-positives, is about 10,000. Hence, we estimate that the probability of observing a single false-positive with a global *Z*-score and a local *Z*-score greater than 3.0 and greater than 2.0 than that of the thresholds is <0.0001.

TESTS OF THE MODEL

We perform four tests in this section on the THOM2 potential. First, we compare the performance of the THOM2 and pair potentials from the literature, using gapless alignments and the S1082 set of proteins. Next, we consider alignments with gaps. We test the specificity and sensitivity of the double *Z*-score filter employed to assess the statistical significance of gapped alignments. Using the double *Z*-score filter, we analyze self-recognition for the S47 set of proteins that contains representatives of folds not sampled in the training. Next, tests of family recognition are presented, including comparison of THOM2 results with those of a pairwise model, using the FEA.

Evaluation of THOM2 and Pair Potentials by Gapless Threading

To make a comparison with pairwise potentials, and to test, at the same time, the generalization capacity of THOM2, we use the S1082 set. This set does not contain proteins included in the training set. However, as discussed in the section, Training and Test Sets, the threshold of 3 Å RMS for global structure-to-structure alignments (using side-chain centers) excludes only close structural homologues. Therefore, the S1082 set includes many structural variations of the folds used in the training. In general, it is difficult to find completely independent test sets when using training sets covering essen-

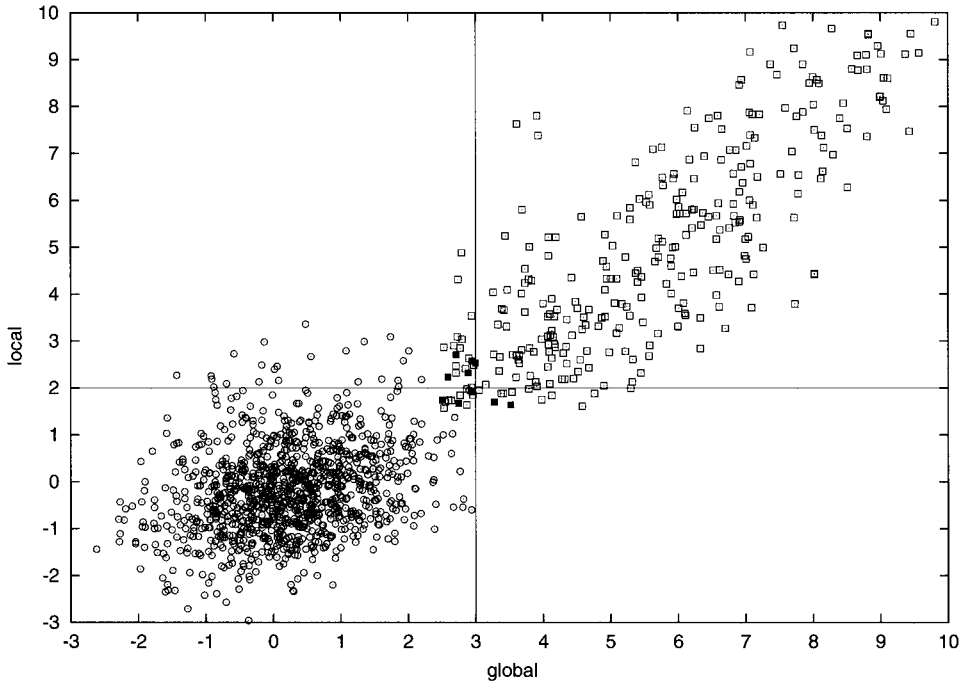


Fig. 3. The joint probability distribution for the *Z*-scores of global and local alignments. Circles at the lower left corner represent a population of 1,081 false-positives, resulting from the alignments of the S47 set sequences (see Fig. 4) against all structures in the Hinds–Levitt (HL) set (100 best global and 200 best local matches are considered, disregarding matches with positives energies of global alignments). The best pair scoring false-positive is slightly below the threshold (3,2). The population in the right upper corner represents (□) 331 pairs of HL sequences aligned to HL structures with global *Z*-scores of >2.5 and local *Z*-scores of >1. This set includes 236 native alignments and 95 non-native alignments; 10 matches are false-positives (■), and they are all below the threshold (3,2). Stiffer energy constraints were employed with only the 10 best global and 200 best local alignments considered. There is a population of true-positives below (2.5,1.0), which are not shown (including 10 native alignments). However, the number of false-positives below this threshold makes predictions within this range difficult.

tially all the known folds. This problem concerns all the knowledge-based potentials considered in this discussion.

Using gapless threading, we compare the performance of THOM2 with the performance of five knowledge-based pairwise potentials. As can be seen in Table X, the Godzik–Skolnick–Kolinski (GSK) potential[44] is the best in terms of the number of inequalities that are not satisfied, followed by the Betancourt–Thirumalai (BT),[45] Tobi–Elber (TE),[25] THOM2, Miyazawa–Jernigan (MJ),[46] and the Hinds–Levitt (HL)[47] potentials. However, in terms of the number of proteins recognized exactly (i.e., proteins with native energies lower than energies of all the decoys generated by gapless threading into all the structures in the S1082 set), the HL potential is the best, followed by TE, MJ, THOM2, BT, and GSK potentials.

The lack of correlation between the above two criteria is related to the fact that some of the above potentials, while recognizing very well many proteins, fare quite poorly for some of the proteins included in the S1082 set. Reducing the number of violated inequalities becomes important when applying some additional filters to select correct predictions from a small subset of energetically favorable matches (e.g., the *Z*-score test; see the section, Assessing the Distribution of *Z*-Scores for Gapped Alignments). Therefore, it would be desirable to satisfy both criteria at same time (also maximizing the *Z*-score of the distribution

TABLE X. Comparison of THOM2 and Knowledge-Based Pairwise Potentials, Using Gapless Threading*

Potential	Recog structs[a]	N_{sat} ineqs (M)[b]	*Z*-score[c]
HL	915	1.84	1.14
TE	914	0.20	1.45
MJ	902	0.28	1.23
THOM2	877	0.24	1.35
BT	861	0.17	1.26
GSK	819	0.08	1.35

*Results of gapless threading on the S1082 set (see text for the details). The results of threading onion model 2 (THOM2) potential are compared with five other knowledge-based pairwise potentials: Betancourt and Thirumalai (BT),[44] Hinds and Levitt (HL),[46] Miyazawa and Jerningan (MJ),[45] Godzik, Kolinski, and Skolnick (GKS),[43] and Tobi and Elber (TE).[25] The latter potential was trained using the linear programming (LP) protocol and the same (TE) training set. Note lack of correlation between the number of proteins that are missed and the number of inequalities, which are not satisfied. See text for further details.

[a]Potentials are ordered according to the number of proteins recognized exactly (out of 1,082).

[b]The number of inequalities that are not satisfied, out of approximately 95 million inequalities generated from the S1082 set (in units of millions).

[c]*Z*-scores (i.e., the ratios of the first and the square root of the second moments) for the distributions of energy differences between the native and misfolded structures.

of energy gaps). From this point of view, the TE, MJ, and THOM2 potentials seem to be somewhat better than the other four potentials. Gapless training of energies remains difficult problem, as reflected in Table X. None of the widely used potentials has a better than 90% success rate. In a set of 1,000 proteins, this translates into many errors.

The conclusion, which is important for the present work, is that the performance of the THOM2 potential is comparable to the performance of pairwise potentials, including the TE potential trained on the same set using a similar LP protocol. Since the proteins used in this test either were not included in the training or represent at least considerable variations of the structures included in the training, we conclude that the exact learning on the training set does not result in overfitting. This is further supported by the results (presented in the next section) for the S47 set of proteins that represent folds not sampled during the training.

Self-recognition by Gapped Alignments

We summarize first the performance of the THOM2 potential in terms of self-recognition of the HL set proteins by optimal alignments and *Z*-score filters. The HL set was partially learned (using gapless threading). However, our training did not include the *Z*-score or the possibility of gaps. Successful predictions based on the *Z*-score only are useful tests, even if performed on the training set of structures. Additionally, there are 40 proteins in the HL set that were not included in the learning (TE) set.

For each sequence, we generate all the global and local alignments into all the structures in the HL set. Energy and *Z*-score filters are considered. Of the total of 246 proteins, 234 native (global) alignments obtain the lowest energy and the highest *Z*-score. There are four native alignments resulting in weak *Z*-scores. The four failures are membrane proteins (from the photosynthetic reaction centers) that were not included in the training set. Only 5 of the remaining 242 native alignments obtain *Z*-scores of <3 (four alignments with *Z*-scores of >2.5 and one alignment with a *Z*-score of <2.5).

For the local alignments, we use the *Z*-score as the main filter, as there are many incorrect alignments with low energies. There are 226 local native alignments with *Z*-scores of >2 (177 of them of rank 1 and 35 of them of rank 2). Among the remaining 20 local native alignments, 9 result in very low *Z*-scores ($Z < 1.0$), including six structures from the training set. Using the double *Z*-score filter with the conservative threshold of 3 for global *Z*-scores and of 2 for local *Z*-scores results in dismissing 23 native alignments as insignificant.

In order to assess further the generalization capacity of THOM2 in terms of self-recognition by optimal alignments, we use the S47 set again. The structures of S47 proteins were embedded in the structures of the TE set, and the sequences of 25 proteins representing different folds in the S47 set were aligned into all the structures of this extended set. We observe that the native structures are found with high probability. A total of 20 of 25 structures result in native alignments with global *Z*-scores of >3 and local *Z*-scores of >2 (Table XI).

A less encouraging observation is the sensitivity of the results to structural fluctuations. THOM2 can identify related structures only if their distance is not too large. Seven out of 14 homologous structures with the DALI[37] *Z*-score for a structure-to-structure alignment of >10 are detected with high confidence. Only one homologous structure with the DALI *Z*-score of <10 is detected.

We would like to point out that only six structures (three pairs of structures representing three folds) of the S47 set had homologous counterparts in the training set. It is therefore reassuring that most of the native structures and significant fraction of the relatives are recognized in terms of both their energies and the *Z*-scores. Moreover, there are no further significant hits into other structures from the TE set. Hence, no false-positives above our confidence thresholds are observed in this test. We conclude that our nearly exact learning (on a training set) preserves significant capacity for identification of new folds using optimal alignments with gaps.

Assessing the Specificity of the Protocol

We present examples of family recognition (i.e., identification of homologues) in terms of energy and double *Z*-score filter. Only a few homologues are identified in a large set of (decoy) structures. This allows us to assess the specificity of the protocol, providing a limited analysis of the sensitivity as well (see the next section for an extended assessment of the sensitivity). The test set S1082 is used. Eight families that have at least three representatives in the S1082 set are chosen to illustrate various aspects of THOM2 threading alignments, as compared with DALI[37] structure-to-structure alignments, as well as sequence-to-sequence alignments. The latter ones are generated using Smith–Waterman algorithm,[18] with the BLOSUM50[39] substitution matrix and structurally biased gap penalties (see the section, Training of Gap Energies). Since we do not incorporate family profiles in our threading protocol, we consider only pairwise sequence alignments for comparison in this discussion.

Similarly to threading, the confidence of sequence matches is estimated using *Z*-scores, defined by the distribution of scores for shuffled sequences. We find that structurally biased gap penalties improve the recognition in case of weak sequence similarity. We do not observe false-positives with more than 50% of the query sequence aligned and with a *Z*-score larger than 8 (for sequence alignment). If there is no clear evidence of sharing a common ancestor and a common function, the structural dissimilarity is used to define false-positives. Note that the distribution of *Z*-scores for sequence substitution matrices is different from that of threading potentials, with a very high *Z*-score for highly homologous sequences.

Regarding the specificity of threading results for the families considered in this discussion, we point out that there are only two energy-based predictions with relatively high global and local threading *Z*-scores that are false. They are still below our thresholds. The highest-

TABLE XI. Self-Recognition for Folds That Were Not Learned*

Name (len)[a]	DALI[b] *Z*-sc (RMS)	THOM2[c] Glob *Z*-sc	THOM2[c] Loc *Z*-sc
1hka (158)	33.0 (0.0)	**7.1**	**7.1**
1vhi (139)	4.3 (5.2)	0.2	0.3
2a2u (158)[d]	33.8 (0.0)	**2.5**	**4.0**
1bbp (173)[d]	11.6 (3.3)	**3.5**	**3.0**
2ezm (101)	55.3 (0.0)	**3.7**	**3.2**
1qgo (257)	46.0 (0.0)	**5.6**	**7.6**
1abe (305)	6.4 (3.4)	0.5	0.4
1byf (123)	29.5 (0.0)	1.8	2.8
1ytt (115)	16.4 (2.2)	−0.1	1.4
1jwe (114)	26.9 (0.0)	2.6	2.3
1b79 (102)	18.7 (1.3)	0.3	1.3
1b7g (340)	61.5 (0.0)	**8.7**	**8.8**
1a7k (358)	25.1 (2.9)	−0.4	−0.9
1eug (225)	43.0 (0.0)	**3.4**	**3.0**
1udh (244)	30.8 (1.7)	−1.0	2.9
1d3b (72)	18.4 (0.0)	**3.5**	**2.8**
1b34 (118)	13.4 (1.1)	1.9	2.0
1dpt (114)	24.8 (0.0)	**6.2**	**6.0**
1ca7 (114)	18.7 (1.2)	**4.0**	**2.5**
1bg8 (76)	19.1 (0.0)	**3.4**	**3.5**
1dj8 (79)	16.2 (0.7)	**5.1**	**3.9**
1qfj (226)	42.7 (0.0)	**8.1**	**8.4**
1vid (214)	7.1 (3.1)	−2.0	0.5
1bkb (132)	25.1 (0.0)	2.7	1.5
1eif (130)	17.4 (1.6)	**3.5**	**2.0**
1b0n (103)	19.5 (0.0)	**4.7**	**5.0**
1lmb (87)	8.0 (5.3)	0.3	0.1
1bd9 (180)	38.8 (0.0)	**4.5**	**5.8**
1beh (180)	36.0 (0.3)	**7.4**	**5.8**
1bhe (376)	70.2 (0.0)	6.7	0.6
1rmg (422)	36.9 (2.2)	0.9	—
1b9k (237)	39.7 (0.0)	**8.1**	**8.2**
1qts (247)	36.1 (0.7)	**3.5**	**6.4**
1eh2 (95)	24.3 (0.0)	**6.0**	**6.5**
1qjt (99)	7.6 (2.5)	**3.6**	**3.7**
1bqv (110)	20.9 (0.0)	**3.5**	**2.3**
1b4f (82)	3.2 (3.3)	0.0	1.7
1ck2 (104)	26.0 (0.0)	**5.2**	**4.3**
1cn8 (104)	14.3 (2.2)	**5.3**	**2.0**
1b10 (116)	24.9 (0.0)	0.5	0.5
1jhg (101)	3.4 (6.6)	1.1	1.0
1bnk (100)	24.9 (0.0)	**5.4**	**6.3**
1b93 (148)	31.4 (0.0)	**4.0**	**3.2**
1mjh (143)	6.1 (3.4)	0.3	1.3
1bk7 (190)	37.2 (0.0)	**7.7**	**9.0**
1bol (222)	19.7 (2.3)	0.1	−1.0
1bvb (211)	37.3 (0.0)	**5.3**	**4.3**

*The S47 set of proteins is used in order to test the self-recognition. It is also a test of the sensitivity of the results to structural fluctuations for 25 different folds (of which 22 were not represented in the training set), using the double *Z*-score test.

[a]Pairs of homologous structures belonging to the S47 set are specified (three folds are represented by a single structure, for 2a2u its structural relative from the training set is included), using their Protein Data Bank (PDB) codes and lengths (specified in parentheses). If the domain is not specified and one refers to a multidomain protein, the A (or first) domain is used. High confidence predictions (global *Z*-score of >3.0 and local *Z*-score >2.0) are indicated in bold. Query sequences are indicated in italics (for each pair, the first line describes the native alignment and the second line an alignment into a homologous structure). Two of 25 native alignments gave weak signals (DNA binding protein 1blo and glycosidase 1bhe). Four other native alignments (2a2u, 1byf, 1jwe, and 1bkb) result in global *Z*-scores of somewhat <3.

[b]DALI[37] *Z*-scores and root-mean-square deviations (RMS) for structure-to-structure alignments into native and homologous structures. Low DALI *Z*-scores indicate that only short fragments of the respective structures are aligned and the resulting RMS may not be representative. Most of the homologous structures with a DALI *Z*-score of >10 are recognized with high confidence.

[c]Results of global and local THOM2 threading alignments of the 25 query sequences into an extended TE + S47 set.

[d]Alignment of the 2a2u sequence into the 1bbp structure was the only significant hit of any of the query sequences into the structures included in the training (Tobi–Elber [TE]) set. Thus, no false-positives with scores greater than our confidence cutoffs were observed.

scoring false-positive, namely the alignment of the aspartyl protease 1htrB into the xylanase 1clxA (*Z*-scores of 3.7 and 1.5, when converged using 1,000 shuffled sequences; see Table XIID), is still below our cutoffs. The alignment of the zinc-finger protein 1meyC into the Adrl DNA-binding domain 2adr is potentially the highest-scoring false-positive among the sequence-based matches. However, even though 1meyC and 2adr are structurally dissimilar according to DALI (RMS of 7.9 Å for 40 residues), they share very high sequence similarity (42% for 55 residues), have similar function, and are classified as related folds (zinc-finger design and classic zinc finger, respectively) by SCOP.[52] Other false-positives due to the sequence-to-sequence alignments obtain *Z*-scores of 5–7, which may cause difficulties in making predictions based on weak sequence similarity.

Regarding the sensitivity of the protocol, one finds first that all the native structures are with the lowest energies and are recognized with high confidence in terms of the double *Z*-score filter. We observe a varying degree of success in the recognition of family members and structural homologs, as illustrated in Table XIIA–H. Threading predictions are very robust for RAS, lactoglobulin, and glutathione transferase families. In the case of the RAS family (Table XIIA), a number of matches into remote structural relatives that share certain structural motifs with the RAS fold are observed. The structural similarity between lactoglobulins and bilin-binding proteins (that do not share detectable sequence similarity) is recognized (see alignment of 2blg into 2apd in Table XIIB). Glutathione transferases 1aw9 and 1axdA, with very weak signals from sequence alignments, are recognized as well.

By contrast, there are families for which threading performance is erratic, including phosphotransferase, cytochrome, and zinc-finger families that include matches recognizable by sequence alignment, of similar length and significant structural similarity, yet not recognized by threading (Table XIIC–F). The results for the pepsin-like acid proteases (Table XIIG) demonstrate missing matches attributable to significant differences in length, which are difficult to account for in global alignments. Local sequence and threading alignments for proteases 1pfzA and 1lyaB result in high *Z*-scores, but no signal from global threading alignment is observed. The family of small toxins is an example of relatively weak signals (both from threading and sequence alignment) that are below our universal cutoffs for false positives (Table XIIH).

Assessing Protein Family Signals and the Sensitivity of the Protocol

Three families are considered: globins (92 proteins), immunoglobins (Fv fragments, 137 proteins), and the DNA-binding, POU-like domains (26 proteins). Sequences of all family members are aligned optimally to all the structures in the family. Both the local and global alignments are generated for each sequence–structure pair and the results are compared in terms of a simplified version of the double *Z*-score filter discussed earlier. Ideally, all the scores should be above the thresholds we presented. The

scores should also correlate with the RMS. The THOM2 results are compared with the results of the TE pairwise potential, which was trained on the same (TE) set using the LP protocol.[25]

The alignments due to the TE potential are computed using the first iteration of the FEA.[23] In THOM2, the number of neighbors to a secondary site determines its identity, whereas in FEA it is approximated by the identity of the native residue at that site. In principle, the FEA should be iterated until self-consistency is achieved.[23] Alternative to FEA are global optimization techniques[22] that are computationally expensive and difficult to use at the scale of testing presented here. Purely structural characterization of contact types in THOM2 avoids this problem, making the THOM2 potential amendable to dynamic programming, at least for global alignments (see the section, Self-recognition by Gapped Alignments).

Figures 4a–c shows the joint histograms of the sum of Z-scores for local and global THOM2 threading alignments (with trained gap penalties of Table VIIIB) versus the RMS between the superimposed side-chain centers (see the section, Training and Test Sets), for globins, immunoglobins, and POU-like domains, respectively. The vertical lines in the Figure 4 correspond to the sum of global and local Z-scores equal to 5, which approximately discriminates the high confidence matches (with the sum of local and global Z-scores of >5) and lower confidence matches that might be obscured by the false-positives. Nearly all pairs differing by <3 Å RMS can be identified by THOM2 threading alignments. Most of the matches within the range of 3–5 Å can still be identified with high confidence. Overall, 60%, 90%, and 95% of homologues with RMS <5 Å are recognized, for POU, globin and immunoglobin families, respectively. However, the number of matches with high confidence quickly decreases with the growing RMS.

The population of matches that are difficult to identify by pairwise sequence-to-sequence alignments, with structurally biased gap penalties (see the section, Training of Gap Energies and the section, Assessing the Specificity of the Protocol) is represented by the filled squares. All the matches represented by circles can be identified with high confidence by sequence-to-sequence alignments (i.e., they result in Z-scores of >8.0). Essentially all the pairs with RMS of <3 Å are identified by sequence alignments as well. Below this threshold, we observe many matches that can be still identified by threading, but not by sequence alignment (filled rectangles with the sum of threading Z-scores of >5). We also found examples of matches detected with high confidence by threading and not detected by PsiBLAST[48] (with default parameters and the PDB database) in many of the families considered: globins 1flp and 1ash, immunoglobin 2hfm and T-cell receptor 1cd8, toxins 1acw and 1pnh, lactoglobulin 2blg and bilin-binding protein 2apd, pheromones 2erl and 1erp, and POU-like proteins 1akh and 1mbg. By contrast, many sequence alignment matches are not detected by threading.

The performance of THOM2 and TE potentials is compared using 1D histograms for the sum of Z-scores for local and global threading alignments. For the sake of fair comparison, the ad hoc gap penalties, as defined in the section, Training of Gap Energies are used for both potentials. As can be seen in Figure 4d,e for globins and POU-like domains, the number of low Z-scores for THOM2 is smaller than the number of low Z-scores obtained with the TE potential and FEA. For example, the number of low confidence matches (which can still be roughly defined as matches below the cutoff of 5) for globins increases from 2,401 in the case of THOM2 to 3,350 (out of 8,558 matches) in the case of the TE potential. It can also be seen that the distribution of Z-scores is different. The TE potential yields many high Z-scores for alignments into very close homologues, as opposed to lower scores for more divergent pairs.

The somewhat worse performance of the pairwise model for these two families may result from the suboptimality of the alignments that we generate using the FEA. Interestingly, FEA with the TE potential also fails for a larger number of native alignments. For example, in the family of DNA binding proteins, the number of native alignments with very low Z-scores (<4) is equal to 7 for TE and only 2 for THOM2.

By contrast, there are families for which the TE potential works better. An example is the family of the immunoglobins (Fig. 4f). The FEA is expected to perform well when the sequence similarity is sufficiently high, since the information about the native sequences is used to generate optimal alignments. The divergence in terms of what can be detected by sequence similarity is larger for globins and POU-like proteins than for immunoglobins. For example, contrary to other families considered here, all the immunoglobins with an RMS of <4 Å can be detected by sequence alignments (Fig. 4c). Therefore, good performance of the FEA with the TE potential is expected in this case.

The above observation is further supported by the results of the FEA with the TE potential for eight families from the S1082 set, considered in the previous section. We do not include detailed results in this discussion. Instead, we summarize them. The threading results with the FEA and the TE potential are robust (and comparable to the THOM2 results) for RAS, SH3, and acid protease families that are represented by proteins of high sequence similarity. The results of the FEA are considerably worse for lactoglobulins and glutathione transferase families that are characterized by much lower success of sequence-based recognition (Table XII). At the same time, the FEA performs as poorly as THOM2 for cytochrome and zinc-finger families. An exception is observed for the toxin family, for which the FEA performs considerably better than THOM2, although there is no (or low) sequence similarity for some of the matches.

CONCLUSIONS AND FINAL REMARKS

We propose and apply an automated procedure for the design of threading models. The strength of the procedure, which is based on linear programming tools, is the automation and the ability of continuous exact learning. The LP protocol was used to evaluate different energy functions for accuracy and recognition capacity. Keeping in mind the

TABLE XII. Examples of Predictions for Eight Families of Homologous Proteins*

A: RAS family[a]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
121p	**5.9**	**9.5**	**1**	**166**	**74.8**	**36.1/0.0/166/100**
1kao	6.1	5.9	2	167	46.4	28.5/1.4/166/49
3rabA	4.8	3.2	3	169	17.1	27.5/1.4/165/31
1ftn	3.7	3.8	10	193	21.7	22.9/1.8/161/35
1hurA	2.8	3.6	4	180	9.3	14.8/2.5/147/15
1kevA	—[k]	3.6	—[k]	351	—[k]	3.1/4.0/83/11
1mioB	—[k]	3.5	—[k]	458	—[k]	3.5/3.6/99/10[j]
1hdeA	—[k]	3.4	—[k]	310	—[k]	2.5/3.6/111/9
1ksaA	—[k]	*2.7*	—[k]	*366*	—[k]	—[k]
1cbf	—[k]	—	—[k]	*285*	*5.1*	*1.8/3.9/74/9*

B: Lactoglobulin family[a]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
2blg	**8.2**	**10.0**	**1**	**162**	**79.9**	**35.1/0.0/162/100**
1a3yA	4.7	3.8	5	149	10.6	17.6/2.4/140/17
1bj7	3.0	3.1	4	150	4.4	17.8/2.4/142/18
2apd	3.0	2.1	2	169	—[k]	11.8/3.0/138/15
1mup	1.7	2.5	3	157	8.9	19.2/2.2/146/16
2pcfB	—[k]	*3.0*	—[k]	*250*	—[k]	—[k]
1lgbC	—[k]	—[k]	—[k]	*159*	*6.0*	—[k]
1nglA	—[k]	—[k]	—[k]	179	5.8	12.0/3.5/136/14

C: Glutathione S-transferase family[a]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
2gsq	**7.0**	**7.3**	**1**	**202**	**87.3**	**37.5/0.0/202/100**
1axdA	2.0	5.2	3	209	3.3	18.1/2.9/190/17
1gsdA	3.2	3.7	4	221	16.5	25.1/2.1/200/29
1aw9	4.3	2.5	2	216	4.9	18.4/3.1/194/19
1gnwA	—[k]	4.0	—[k]	211	5.0	17.1/3.1/187/17
1clxA	—[k]	*3.7*	—[k]	*347*	—[k]	*0.5/3.9/50/9*
1fhe	—[k]	—[k]	—[k]	217	11.5	20.9/2.3/195/25
2ljrA	—[k]	—[k]	—[k]	244	5.1	15.7/3.1/195/18
1ao7E	—[k]	—k	—[k]	*245*	*4.9*	—[k]

D: Phosphotransferase (SH3 domain) family[b]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
1aww	3.4	4.6	2	67	19.3	8.1/1.7/56/36[j]
2semA	3.9	2.6	3	58	13.8	10.2/1.5/56/31[j]
1fynA	**4.3**	**2.1**	**1**	**62**	**49.8**	**9.5/1.7/56/47[j]**
4hck	3.2	3.5	5	72	28.4	8.1/2.0/55/40[j]
1hsq	2.7	4.0	6	71	10.8	9.8/1.4/54/26[j]
1a3k	—[k]	*3.1*	—[k]	*137*	*3.2*	—[k]
1gbrA	—[k]	—[k]	—[k]	74	13.9	7.7/2.0/57/34[j]
1ark	—[k]	—[k]	—[k]	60	12.3	7.6/1.9/56/20[j]
lnksA	—[k]	—[k]	—[k]	*194*	*5.5*	—[k]

E: Cytochrome c family[b]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
2cxbA	**6.8**	**6.0**	**1**	**124**	**57.4**	**28.1/0.0/123/100**
1co6	—[k]	—[k]	—[k]	107	15.9	14.4/1.7/99/36
1dsn	—[k]	*3.4*	—[k]	*333*	—[k]	—[k]
1crxA	—[k]	*3.1*	—[k]	*322*	—[k]	*0.9/3.3/50/8*
1ndoA	—[k]	—[k]	—[k]	*449*	*4.9*[k]	—[k]
451c	—[k]	—[k]	—[k]	82	3.9[k]	4.9/2.1/64/19
3cyr	—[k]	—[k]	—[k]	*107*	*2.9*[k]	—[k]

TABLE XII. (Continued)

F: Zinc-finger family[b]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
1meyC	**5.5**	**3.6**	**1**	**87**	**36.8**	**8.9/1.8/82/51**[j]
1jhb	—[k]	*3.4*	—[k]	*106*	—[k]	—[k]
1iml	—[k]	*2.9*	—[k]	*76*	*6.5*	—[k]
2adr	—[k]	—[k]	—[k]	*60*	*11.4*	*1.2/7.9/40/35*[j]
2drpA	—[k]	—[k]	—[k]	66	9.9	4.9/2.6/58/33[j]

G: Aspartyl protease family[c]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
1htrB	**9.6**	**8.3**	**1**	**329**	**95.0**	**56.8/0.0/329/100**
4cms	5.0	5.7	3	323	47.5	39.6/1.7/301/39[j]
2jxrA	5.6	3.7	2	329	43.9	37.0/2.1/307/41
1clxA	*3.7*	*1.5*	*4*	*347*	—[k]	—[k]
1egzA	—[k]	*3.6*	—[k]	*291*	—[k]	—[k]
1pfzA	—[k]	2.9	—[k]	380	32.2	31.7/2.4/298/29
1lyaB	—[k]	2.4	—[k]	241	32.0	9.3/2.4/83/59
2pia	*1.3*	—[k]	*6*	*321*	*6.0*	*0.5/4.3/49/6*

H: Scorpion toxin-like family[d]

Name[e]	GT[g]	LT[g]	Ene[b]	Len[f]	LS[g]
1pnh	2.8	2.9	2	31	—[k]
1acw	**2.8**	**2.1**	**1**	**29**	**15.6**
1mea	*2.5*	*1.3*	*4*	*28*	—[k]
1bh4	1.1	2.5	5	30	—[k]
1mtx	1.4	—[k]	10	39	6.0
2pta	—[k]	—[k]	—[k]	35	5.9
1ica	—[k]	—[k]	—[k]	40	4.8
1ilmA	—[k]	—[k]	—[k]	*61*	*3.3*

*Eight families, with a number of representatives included in the S1082 set, illustrating various degrees of success of our threading protocol in terms of sensitivity and specificity. Results are presented for global and local threading alignments using the threading onion model 2 (THOM2) potential, together with the results for (structurally biased) local sequence-to-sequence alignments and DALI structure-to-structure alignments. Representatives used as query sequences aligned to all the structures in the S1082 set are marked in boldface. Matches are ordered according to the sum of global and local threading *Z*-scores and according to *Z*-scores of the local sequence alignments if no threading signal is detected. False-positives (defined as matches with DALI *Z*-scores of <2.0) are indicated in italics. The highest-scoring false-positives for both: threading and sequence alignments are reported for each family.

[a]Example of family with successful threading predictions that do not share a detectable sequence similarity or that have a weak signal from sequence-to-sequence alignment (*Z*-score of <8.0).

[b]Example of family for which threading is less successful, missing a number of family members (of similar length) that can be detected by sequence-to-sequence alignment.

[c]Lack of detection when the difference in length is significant is expected, and it is one of the limitations of the present protocol.

[d]Example of family for which the DALI results could not be retrieved; therefore, the SCOP classification is used to define structural relatives (i.e., proteins that do share the knottins fold).

[e]Names of proteins (Protein Data Bank [PDB] codes).

[f]Lengths of proteins.

[g]*Z*-scores are computed using 50 shuffled sequences for a number of alignments with the lowest energies: 20 best matches in case of global threading (GT) alignments, 500 best matches in case of local threading (LT) alignments, and 50 best matches in case of local sequence (LS) alignments.

[h]Rank of the energy of global threading alignments is reported in the 4th column.

[i]DALI[37] alignments between the (known) structure of a query and the structure of a match are characterized in the last column: *Z*-score, RMS, length of the aligned fragment, and the identity for this fragment are provided.

[j]Comparisons with the FSSP representative of the query structure are used instead of a direct DALI alignment.

[k]Lack of a detectable (threading, sequence, or structural) similarity.

necessity for efficient threading algorithms with gaps, we selected the THOM2 as our best choice.

Statistical filters based on local and global *Z*-scores were outlined. We observe that, while using conservative *Z*-scores that essentially exclude false-positives, the new protocol recognizes correctly (without any information about sequences) most of the family members with the RMS between the superimposed side-chain centers of ≤ 5 Å

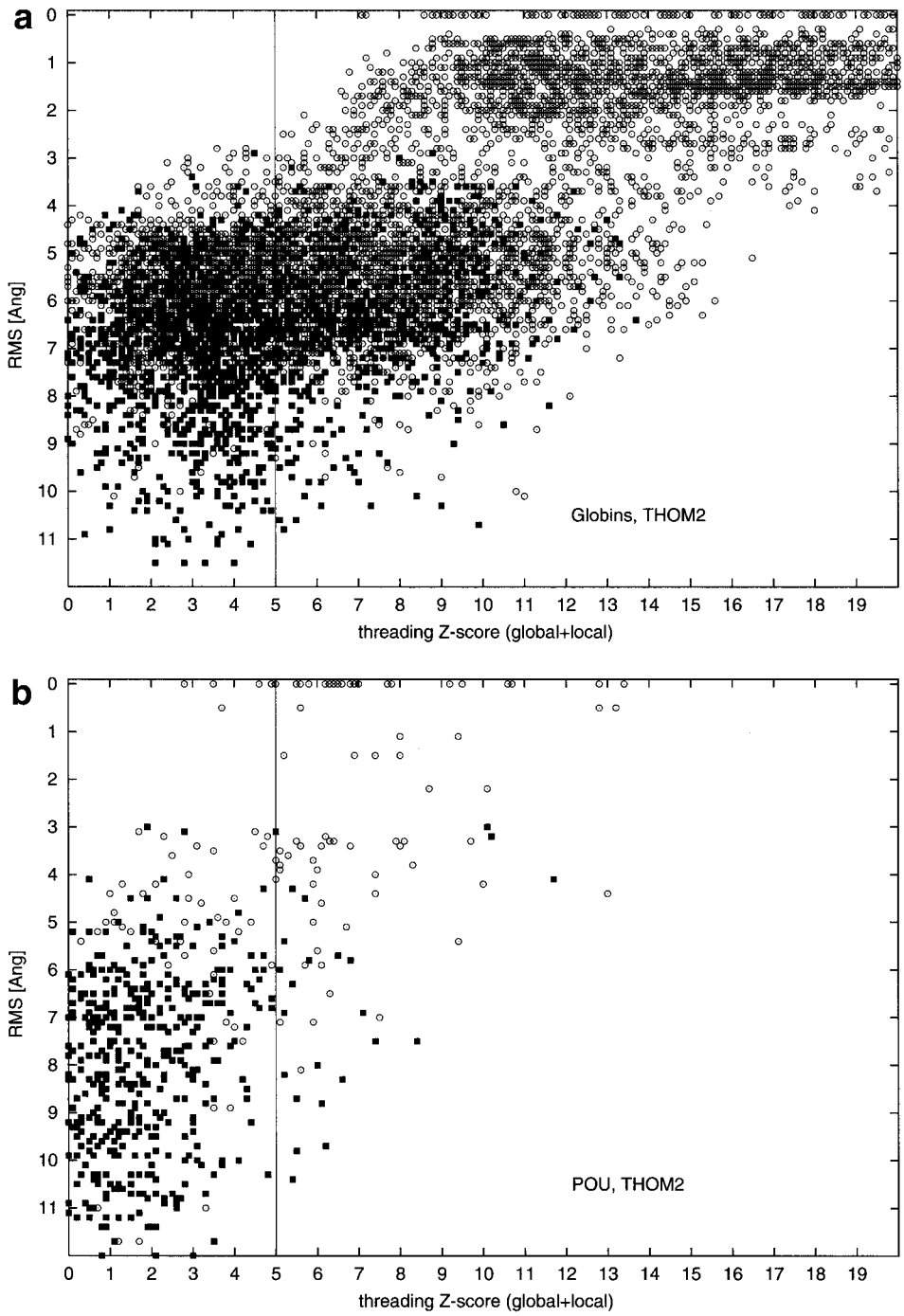


Fig. 4. Comparison of family recognition by THOM2 and pair energies. The results of THOM2 (with the trained gap penalties of Table VIIIB) for families of globins (**a**), POU-like domains (**b**), and immunoglobins (Fv fragments) (**c**). The joint histograms of the sum of *Z*-scores for local and global threading alignments versus the root-mean-square deviations (RMS) between the superimposed (according to structure-to-structure alignments) side-chain centers are presented. The population of matches that are difficult to identify by sequence-to-sequence alignments is represented by the filled squares. Next, the THOM2 results are compared to the results of Tobi–Elber (TE) pairwise potential,[25] using ad hoc gap penalties defined in text. The TE potential was optimized using the LP protocol and the same training set. The first iteration of the so-called frozen environment approximation (FEA)[23] is performed to obtain approximate alignments for the TE potential. One-dimensional histograms of the sum of *Z*-scores for local and global threading alignments for the globins (**d**), POU (**e**), and immunoglobin (**f**) families. Note, that the number of low THOM2 *Z*-scores (<5) is smaller for families of globins and POU-like proteins. By contrast, the TE potential and the FEA perform better for the family of immunoglobins, which is also easier for sequence alignment methods (see text for details).

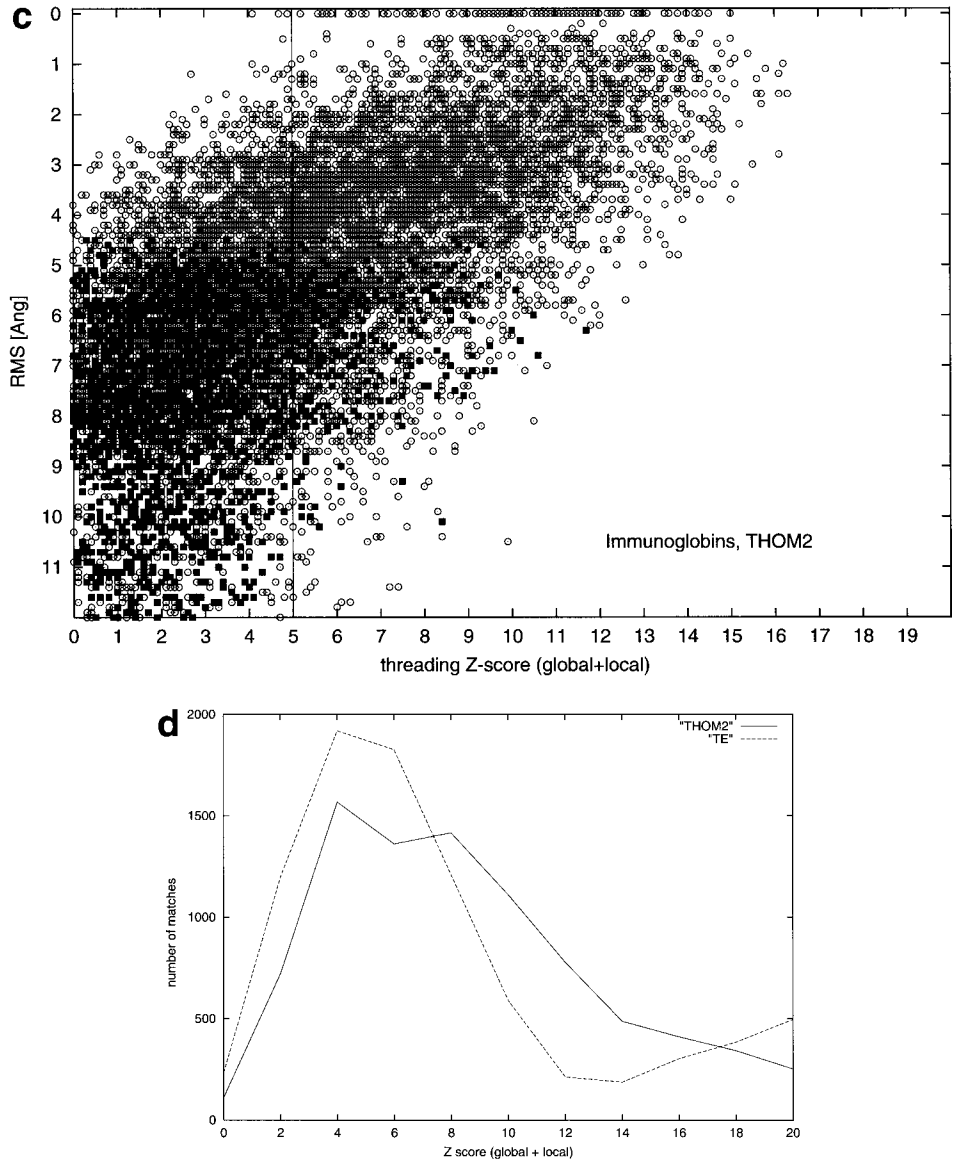


Figure 4. (Continued.)

and differences in length of ≤10%. We also observe many instances of successful recognition of family members that are not recognized by pair energies with the so-called frozen environment approximation.

The present approach is based on fitness of sequences into structures. Nevertheless, it is easily extendable to include sequence similarity, family profiles, or secondary structures as well. Such complementary "signals" are often employed in conjunction with pairwise potentials.[9–11,16] Threading protocols that are based exclusively on contact models were shown (consistent with our observations) to be quite sensitive to variations in structures.[49] THOM2 provides an alternative comparable in performance to pairwise potentials. Therefore, it can be used as a fast component of fold recognition methods employing pair energies, which is the target of a future work.

Despite the limitations of the threading protocol that is based on the THOM2 potential and the double *Z*-score filter (in terms of range of variations in structure and length that can be recognized), we found a number of useful predictions for remote homologues (e.g., ref. 50). Therefore, we decided to take part (group 280) in the recently held critical assessment of fully automated protein structure prediction methods (CAFASP),[51] even at the preliminary phase, without using additional information as secondary structures or family profiles. The performance of the LOOPP server[30] was about average for all fold recognition targets (e.g., LOOPP missed some targets recognizable by Psi-BLAST). How-

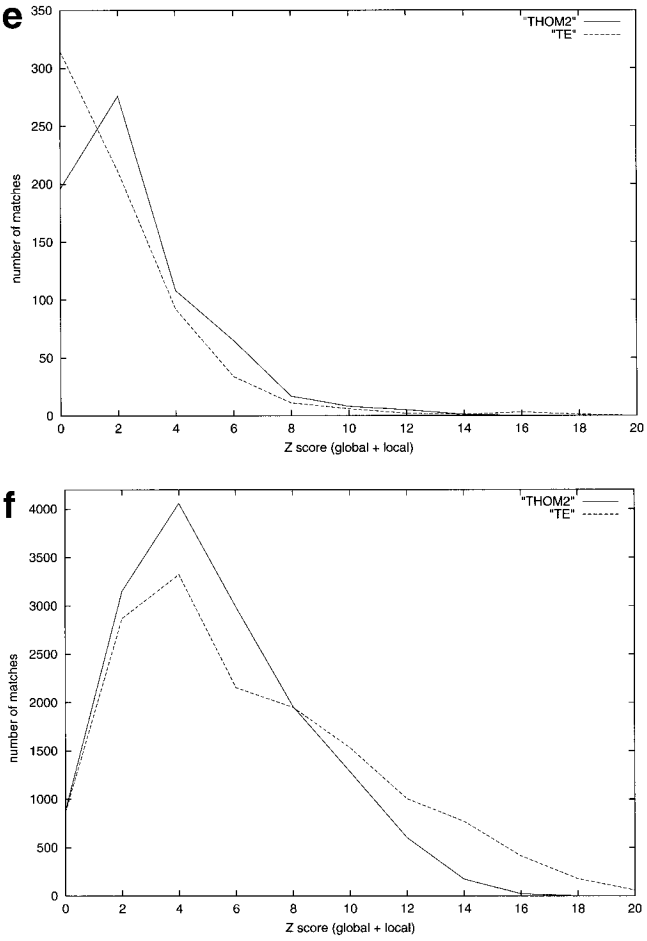


Figure 4. (Continued.)

ever, in the category of difficult-to-recognize targets, it was ranked among the best servers (rank 4 in the MaxSub 5.0 A evaluation), providing the best predictions among the servers for two difficult targets (T0097 and T0102).[51]

ACKNOWLEDGMENTS

This research was supported by a grant from the National Institutes of Health National Center for Research Resources (NCRR) to the Cornell Theory Center (acting director Ron Elber) for the development of Computational Biology Tools. It was further supported by a seed grant from DARPA (to R.E.). Jaroslaw Meller acknowledges also partial support from the Polish State Committee for Scientific Research, grant 6 P04A-066-14.

REFERENCES

1. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
2. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
3. Sippl MJ, Weitckus S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. Proteins 1992;13:258–271.
4. Godzik A, Kolinski A, Skolnick J. Topology fingerprint approach to the inverse folding problem. J Mol Biol 1992;227:227–238.
5. Ouzounis C, Sander C, Scharf M, Schneider R. Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from 3D structures. J Mol Biol 1993;232:805–825.
6. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through folding motif. Proteins 1993; 16:92–112.
7. Matsuo Y, Nishikawa K. Protein structural similarities predicted by a sequence–structure compatibility method. Protein Sci 1994;3: 2055–2063.
8. Mirny LA, Shakhnovich EI. Protein structure prediction by threading. Why it works and why it does not. J Mol Biol 1998;283:507–526.
9. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
10. Panchenko AR, Marchler-Bauer A, Bryant SH. Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 2000;296:1319–1331.
11. Sternberg MJE, Bates PA, Kelley LA, MacCallum RM. Progress in protein structure prediction: assessment of CASP3. Curr Opin Struct Biol 1999;9:368–373.
12. Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein structure simulations: functional forms and parameters of long range side chain interaction potentials from protein crystal data. J Comp Chem 1997;18:849–873.
13. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 2000;300:171–185.
14. Babajide A, Hofacker IL, Sippl MJ, Stadler PF. Neural networks in protein space: a computational study based on knowledge-based potentials of mean force. Folding Design 1997;2:261–269.
15. Babajide A, Farber R, Hofacker IL, Inman J, Lapedes AS, Stadler PF. Exploring protein sequence space using knowledge based potentials. J Compar Biol 1999;(in press).
16. Elofsson A, Fischer D, Rice DW, Le Grand S, Eisenberg D. A study of combined structure-sequence profiles. Folding Design 1998;1: 451–461.
17. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J Mol Biol 1970;48:443–453.
18. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
19. Johnson MS, Overington JP, Blundell TL. Alignment and searching for common protein folds using a data bank of structural templates. J Mol Biol 1993;231:735–752.
20. Croman HT, Leiserson CE, Rivest RL. Introduction to algorithms. Cambridge, MA: MIT Press; 1985.
21. Lathrop RH, Smith TF. Global optimum protein threading with gapped alignment and empirical pair score functions. J Mol Biol 1996;255:641–665.
22. Lathrop RH. The protein threading problem with sequence amino-acid interaction preferences is NP-complete. Protein Eng 1994;7: 1059–1068.
23. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. The statistical mechanical basis of sequence alignment algorithms for protein structure prediction. In: Elber R, editor. Recent developments in theoretical studies of proteins. Singapore: World Scientific; 1996. pp 110–140.
24. Maiorov VN, Crippen GM, Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
25. Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins 2000;40:71–85.
26. Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys 1998;109:11101–11108.
27. Meller J, Wagner M, Elber R. Maximum feasibility guideline in the design and analysis of protein folding potentials. J Comp Chem 2001;(in press).
28. Meszaros CS. Fast Cholesky factorization for interior point methods for linear programming. Computer Math Applications 1996;31: 49–51.
29. Adler I, Monteiro RDC. Limiting behavior of the affine scaling continuous trajectories for linear programming problems. Math Program 1991;50:29–51.

30. Taylor WR, Munro RE. Multiple sequence threading: conditional gap placement. Folding Design 1997;2:S33–S39.
31. Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol 1994;243:668–682.
32. Bryant SH, Altschul SF. Statistics of sequence–structure threading. Curr Opin Struct Biol 1995;5:236–244.
33. Fitch WM. Random sequences. J Mol Biol 1983;163:171–176.
34. Altschul SF, Erickson BW. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. Mol Biol Evol 1985;2:526–538.
35. Fischer D, Elofsson A, Rice D, Eisenberg D. Assessing the performance of fold recognition methods by means of a comprehensive benchmark. In: Pacific Symposium on Biocomputing, Hawaii, 1996; p 300–318.
36. CASP3. Third community wide experiment on the critical assessment of techniques for protein structure prediction, Proteins 1999;Suppl 3; see also http://predictioncenter.lnl.gov/casp3.
37. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res 1994;22:3600–3609; see also DALI server; http://www2.embl-ebi.ac.uk/dali.
38. Meller J, Elber R. Learning, Observing and Outputting Protein Patterns (LOOPP)—a program for protein recognition and design of folding potentials; http://www.tc.cornell.edu/CBIO/loopp.
39. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1989;89:10915–10919.
40. Gambel EJ. Statistics of extremes. New York: Columbia University Press; 1958.
41. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 1990;87:2264–2268.
42. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–2448.
43. Pearson WR. Empirical statistical estimates for sequence similarity searches. J Mol Biol 1998;276:71–84.
44. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci 1995;4:2107–2117.
45. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;2:361–369.
46. Miyazawa S, Jernigan RL. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. J Mol Biol 1996;256:623–644.
47. Hinds DA, Levitt M. A lattice model for protein structure prediction at low resolution. Proc Natl Acad Sci USA 1992;89:2536–2540.
48. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
49. Bryant SH. Evaluation of threading specificity and accuracy. Proteins 1996;26:172–185.
50. Frary A, Nesbitt TC, Frary A, Grandillo S, van der Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KP, Tanksley SD. Cloning transgenic expression and function of fw2.2: a quantitative trait locus key to the evolution of tomato fruit. Science 2000;289:85–88.
51. Fischer D, et al. CAFASP-2: the second critical assessment of fully automated structure prediction methods. Proteins CASP4 issue. 2001;(in press).
52. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of protein data base for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.

Maximum Feasibility Guideline in the Design and Analysis of Protein Folding Potentials

JAROSLAW MELLER,[1,2] MICHAEL WAGNER,[3] RON ELBER[1]
[1]*Department of Computer Science, Upson Hall 4130, Cornell University, Ithaca, New York 14853*
[2]*Department of Computer Methods, Nicholas Copernicus University, 87-100 Torun, Poland*
[3]*Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia 23529-0011*

Received 23 February 2001; Accepted 2 August 2001

Abstract: Protein folding potentials are expected to have the lowest energy for the native shape. The Linear Programming (LP) approach achieves exactly that goal for a training set, or indicates that this goal is impossible to obtain. If a solution cannot be found (i.e., the problem is infeasible) two possible routes are possible: (a) choosing a new functional form for the potential, (b) finding the best potential with a feasible subset of the data, and (or) detecting inconsistent subset of the data in the training set. Here, we explore option (b). A simple heuristic for finding an approximate solution to an infeasible set of linear inequalities is outlined. An approximately feasible solution is obtained iteratively, starting from a certain initial guess, by computing a series of analytic centers of the polyhedra defined by all the inequalities satisfied at the subsequent iterations. Standard interior point algorithms for Linear Programming can be used to compute efficiently the analytic center of a polyhedron. We demonstrate how this procedure can be used for the design of folding potentials that are linear in their parameters. The procedure shows an improvement in the quality of the potentials and sometimes points to flaws in the original data.

 J Comput Chem 23: 1–8, 2002

Key words: linear programming; interior-point methods; folding potentials

Introduction

The basic requirement for protein folding potentials is their ability to distinguish native-like from nonnative shapes. This can be achieved by an appropriate choice of the potential (or energy) function, such that for each pair of native and misfolded structures the following constraints are satisfied:

$$\Delta E_{\text{mis, nat}} = E_{\text{misfolded}} - E_{\text{native}} \geq \varepsilon. \quad (1)$$

Here, $E_{\text{native}} \equiv E(\mathbf{X}_{\text{nat}}; \mathbf{z})$ is the energy of the native structure $\mathbf{X}_{\text{nat}}$, $\mathbf{z}$ is the vector of parameters, $E_{\text{misfolded}} \equiv E(\mathbf{X}_{\text{mis}}; \mathbf{z})$ represents the energies of the misfolded (nonnative) structures $\mathbf{X}_{\text{mis}}$ and ε is a positive constant. In other words, we require that the energies of native structures are lower than the energies of misfolded structures.

For energy models linear in their parameters, the set of inequalities in eq. (1) can be solved for the parameters $\mathbf{z}$ by standard Linear Programming (LP) tools. Note that the inequalities of eq. (1) define a set of cuts (hyperplanes) in the parametric space. The intersection of the corresponding feasible (closed) half spaces defines a convex polyhedron (see Fig. 1). LP solvers provide a feasible solution $\mathbf{z}^*$ that belongs to the feasible polyhedron [i.e., $\mathbf{z}^*$ satisfies all the constraints in (1)] and optimizes certain linear objective function. The LP approach for the design of protein folding potentials, which was pioneered by Maiorov and Crippen,[1] usually involves solving very large sets of inequalities, and the efficiency of LP algorithms is an important issue.

Recently, the LP approach has been applied to the design of various folding and threading potentials.[2–7] It has been found, for example, that simple contact pairwise potentials are not sufficient for recognition of all types of protein shapes.[2, 3] The set of inequalities in eq. (1), which we attempt to solve, proves infeasible for a sufficiently large sample of native and misfolded shapes. Infeasibility of large training sets with different functional models was also used as a guideline to design optimal threading potentials. We seek functional forms for the potentials that preserve the exact recognition of all the proteins in the training set and minimize the number of required potential parameters.[5, 6]

Here, we present a heuristic approach to find an approximate solution, which satisfies a possibly large subset of an infeasible set

Correspondence to: R. Elber; e-mail: ron@cs.cornell.edu

Contract/grant sponsors: NIH NCRR grant to the Cornell Theory Center and DARPA (to R.E.)

Contract/grant sponsor: Polish State Committee for Scientific Research; contract/grant number: 6 P04A 066 14

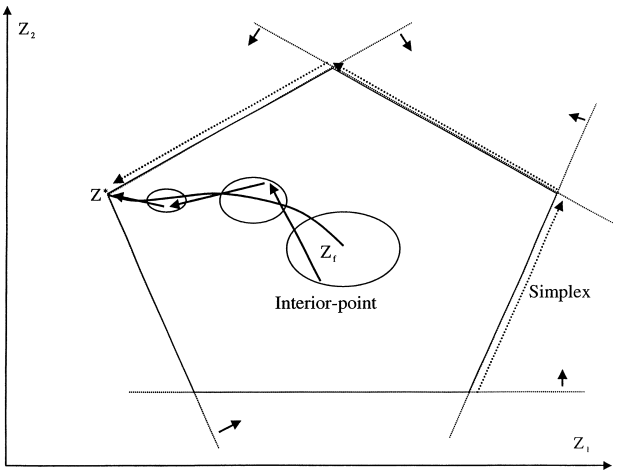


Figure 1. A schematic plot of a polyhedron representing the feasible volume defined by cuts in the parametric space. Given certain feasible set of linear inequalities with bounded variables, the intersection of feasible half spaces (indicated by arrows) takes a form of a polytope. Central path starts at the analytic center of this polytope ($\mathbf{z}_f$) and terminates at the optimal solution ($\mathbf{z}^*$) of the LP problem with an objective function to optimize. The interior point methods proceed through a series of interior points (usually obtained by the subsequent steps of the Newton method) that are located near the central path (arrows in the figure). In practical implementations steps out of the feasible polytope may be allowed, and only the converged solution is guaranteed to be feasible. If there is no function to optimize the interior point algorithms converge to the analytic center.

of inequalities. We call it the "maximum feasibility" (MaxF) guideline. The MaxF procedure is based on a special property of interior point algorithms for LP. Namely, the interior point methods provide the so-called analytic center of the feasible polyhedron (defined in terms of logarithmic barriers "repelling" the solution from the constraints) when the objective function is not used to "force" the convergence to an optimal solution on a facet of the polyhedron. For a bounded polyhedron (which is called the *polytope*) the analytic center is unique. We consider here only bounded problems.

Starting from a set of constraints that are satisfied by a certain initial guess of the solution, a series of "maximally feasible" approximations is computed. The subset of all the inequalities satisfied by the previous approximation, which defines a feasible polytope, is solved using an interior point method. The analytic center of the feasible polytope, obtained as a solution, becomes our next "maximally feasible" approximation. The new approximation satisfies at least as many constraints as the previous partial solution. If no further constraints can be satisfied the procedure stops. The idea behind this heuristic is that the analytic center, which is usually located close to the center (in the topological sense) of the feasible polytope, is likely to satisfy more constraints than an off-centered guess.

Using the MaxF guideline allows us to go beyond a simple feasibility test when assessing the quality of a given model, and may provide a better insight for improving the functional models of folding potentials. It also provides a simple way to improve potentials that are not optimized to satisfy inequality constraints of the type of eq. (1), for example, the commonly used statistical potentials.[8–10] The method is outlined in the Methods section, whereas the results of numerical examples (for a series of medium size problems) are demonstrated in the Results section.

Methods

Interior Point Algorithms

Interior point methods, due to their polynomial time complexity[11, 12] and practical efficiency are nowadays a method of choice for large-scale linear optimization problems.[13–16] An interior point algorithm generates a series of points away from the boundary of the polyhedron (unlike the simplex algorithm, which proceeds along the edges of the feasible region;[14] see also Fig. 1). These points are near a smooth curve, called the *central path*, which is contained within the interior of the feasible polyhedron and terminates at an optimal and complementary solution on a facet or at the vertex (if the optimal solution is unique) of the polyhedron.[17]

Let us consider a linear programming problem [which will be referred to as (LP)] of the form:

$$\left\{\min f_0(\mathbf{z});\ \mathbf{z} \in \mathbf{R}^n;\ f_i(\mathbf{z}) \le b_i,\ i = 1, \ldots, m\right\}, \qquad (2)$$

where $\mathbf{z}$ is a vector of n variables and the objective function to be minimized, f_0, as well as the constraints functions, f_i, are linear. One can define the logarithmic barrier function associated with (LP) as:

$$\phi_B(\mathbf{z}, \mu) = \frac{f_0(\mathbf{z})}{\mu} - \sum_{i=1}^{m} \ln\bigl(b_i - f_i(\mathbf{z})\bigr), \qquad (3)$$

where $\mu > 0$ is the barrier parameter. If the feasible region of (LP) is bounded (i.e., all variables, $z_j;\ j = 1, \ldots, n$, are bounded from below and from above by finite numbers) and nonempty [otherwise (LP) is called *infeasible*], then for each value of μ the barrier function, $\phi_B(\mathbf{z}, \mu)$, achieves the minimal value at a unique (feasible) point, $\mathbf{z}(\mu)$, which is called the μ-center.[13, 16]

The *central path* is defined as the set of μ-centers, where μ changes from ∞ to 0. In the limit of $\mu \to 0$, when minimizing the barrier function of eq. (3), one obtains the desired optimal and feasible solution of (LP)—see Figure 1. Barrier functions of the form specified in eq. (3) are commonly used in the interior point methods for inequality constraints.[13–17] The advantage of reformulating the constrained optimization problem of (2) into unconstrained, nonlinear optimization problem of (3) is that the nonlinear minimization techniques (e.g., gradient or Newton methods) can be applied.

The unique minimum of the barrier function in the limit of $\mu \to \infty$ is called the *analytic center* of the feasible region.[13, 16] The central path always starts at the analytic center and, in the absence of an objective function to optimize, the interior point algorithms converge to the analytic center. We emphasize that, in practice (as in the popular infeasible primal-dual implementations, for example[17]) the functional constraints, f_i, are often initially relaxed and the method proceeds through points away from the central path that may not belong to the feasible polytope. Therefore, the analytic center is reached only upon convergence of the Newton

procedure. There are many parameterizations of the central path. In particular, different barrier functions (as, e.g., weighted logarithmic barriers) can be applied.[16, 18] Therefore, the actual position of the analytic center may vary between different implementations.

Note that solving a set of linear inequalities is equivalent to solving a special case of (LP), obtained by setting the objective function in (2) to zero, $f_0(\mathbf{z}) = \mathbf{0} \cdot \mathbf{z}$. Therefore, when solving a set of inequalities by an interior point algorithm we obtain the analytic center of the feasible polyhedron as a solution. We comment that just solving a set of inequalities (which is by duality theorem equivalent to solving an LP problem[14]) is of the same complexity as the original (LP) problem, with an objective function to optimize.

It should also be pointed out that the analytic center does not correspond (in general) to the center of the feasible polytope in the topological sense. Redundant constraints that do not define the boundaries of the polytope contribute to the barrier function in eq. (3) as well, "repulsing" the analytic center. However, the analytic center is always located away from any individual cutting hyperplane, due to singularity of the logarithm function at zero.

"Maximum Feasibility" Guideline

So far, we were assuming that the problem was feasible, for instance, that there exists a solution to (LP). If the problem proves infeasible it is useful to understand the source of the infeasibility and to assess the "hardness" of the problem. In other words, we would like to know what is the largest subset of constraints that can be satisfied simultaneously, which will be referred to as the Largest Feasible Subset (LFS). The LFS (in the mathematical literature often referred to as the maximum cardinality satisfiable subset) can be used to generate an approximate solution to our problem. Moreover, the analysis of the constraints that cannot be satisfied may help in suggesting a new functional form, new parameters for the potential, or perhaps points to problems in the database.

Unfortunately, finding the LFS is an NP-hard problem.[19] Several heuristic approaches that provide approximate solutions at a low computational cost have been proposed in the past.[20, 21] Such heuristics are of theoretical interest as well, because the problem of finding the LFS is closely related to the so-called satisfiability problem, which is at the origin of complexity theory.[22] Below, we define a simple, iterative procedure, referred to as the "maximum feasibility" guideline. The MaxF heuristic can be advantageous when a reasonable partial solution to an infeasible problem is available, which is usually the case in the design of folding potentials.

Let $\mathbf{z}_0 \in \mathbf{R}^n$ be our initial guess of the solution, which satisfies certain a subset of inequalities in (LP). We will denote this set a $\mathbf{P}(\mathbf{z}_0) = \{f_i;\ f_i(\mathbf{z}_0) \le b_i\}$. Because we assumed that (LP) is infeasible, there are some inequalities in (LP) that are not satisfied by $\mathbf{z}_0$. Let $\mathbf{z}_1$ be the analytic center of the set of inequalities satisfied by the initial guess, $\mathbf{P}(\mathbf{z}_0)$. As described in the previous section, $\mathbf{z}_1$ can be obtained by an interior point method when solving the following set of inequalities:

$$\{f_i(\mathbf{z}) \le b_i;\ f_i \in \mathbf{P}(\mathbf{z}_0)\}. \tag{4}$$

In other words, we solve a feasible LP problem with the inequalities satisfied by the initial guess and without a function to optimize (the objective function is set to zero).

The analytic center of the initial polytope becomes our new guess of the solution. Let $\mathbf{P}(\mathbf{z}_1) = \{f_i;\ f_i(\mathbf{z}_1) \le b_i\}$ be the set of inequalities satisfied by $\mathbf{z}_1$. The new solution satisfies all the constraints of the initial problem, and therefore, $\mathbf{P}(\mathbf{z}_0) \subseteq \mathbf{P}(\mathbf{z}_1)$. In general, let $\mathbf{z}_{k+1}$ be the analytic center of the polytope defined by $\mathbf{P}(\mathbf{z}_k)$, i.e., $\mathbf{z}_{k+1}$, is obtained as the solution of the following set of inequalities:

$$\{f_i(\mathbf{z}) \le b_i;\ f_i \in \mathbf{P}(\mathbf{z}_k)\}. \tag{5}$$

Obviously, $\mathbf{P}(\mathbf{z}_k) \subseteq \mathbf{P}(\mathbf{z}_{k+1})$, that is at each iteration we solve at least all the constraints included in the previous iteration. If no improvement is observed, i.e., when $\mathbf{P}(\mathbf{z}_k) = \mathbf{P}(\mathbf{z}_{k+1})$, the procedure stops.

The analytic center of the final polytope, which we denote as $\mathbf{z}_f$, defines our best approximate (partial) solution to an infeasible set of inequalities, which is our goal here. Alternatively, if our original problem involves a function to optimize, it can be now optimized over the final feasible polytope. The set of inequalities defining the final polytope, $\mathbf{P}(\mathbf{z}_f)$, becomes our approximation to the LFS in (LP). Note that this is not an approximation in the topological sense because just one inequality may dramatically change the shape of the polytope. The number of inequalities in the problem that do not belong to $\mathbf{P}(\mathbf{z}_f)$ can be used to measure the quality of the approximation.

The level of success of the MaxF procedure is critically dependent on the choice of the initial guess and the structure of the problem (in practice a reasonable guess can be obtained from a statistical potential). Imagine, for example, a feasible problem with its feasible polytope **P**. Let us now define a new problem by adding one more constraint, such that the intersection of the polytope and the feasible half space of the new cut is empty, for instance, the new problem is infeasible. The LFS of the new problem is defined by the initial polytope **P**. If we start from a point in the feasible half space of the new cut our procedure will fail to provide a reasonable approximation to **P**. On the other hand, however, if we start from a point in the infeasible half space, then we observe an improvement of the initial guess (see the MaxF "trajectories" in Fig. 2).

Linear Programming Protocol for the Optimization of Folding Potentials

In the next section we demonstrate the numerical performance of the MaxF procedure using realistic examples, relevant for the design of folding potentials. The LP problem in the design of folding potentials is concerned with the exact recognition of the native structures with respect to misfolded shapes in a training set.

Any potential energy function $E(\mathbf{X}; \mathbf{z})$ can be expanded in terms of a basis set (say $\{n_\gamma(\mathbf{X})\}_{\gamma=1}^{\infty}$), in which the coefficients are unknown parameters:

$$E(\mathbf{X}; \mathbf{z}) = \sum_{\gamma=1}^{\infty} z_\gamma n_\gamma(\mathbf{X}). \tag{5'}$$

The information on the protein structure **X** (and implicitly on its sequence S) is "buried" in $n_\gamma(\mathbf{X})$. A good choice of the basis set will converge the sum to the right solution with only a few terms. Of course, such a choice is not trivial to find and one of the advantages

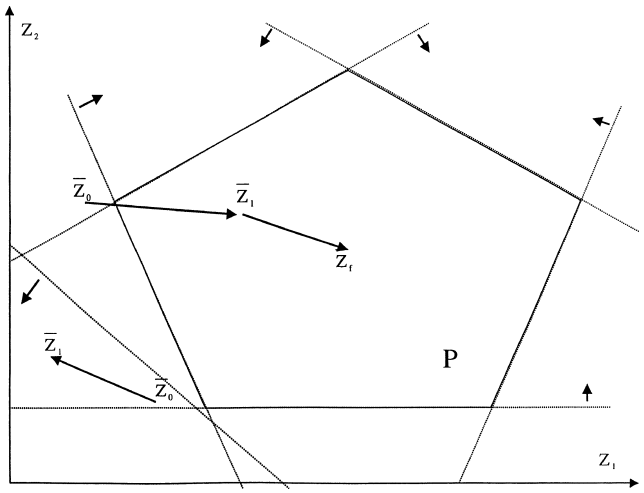


Figure 2. A pictorial representation of a series of analytic centers obtained by applying the "maximum feasibility" guideline to an infeasible LP problem (feasible half spaces are indicated by arrows pointing out from the cutting hyperplanes). Bad and good scenarios are illustrated by two trajectories starting from different initial solutions. In practice, a reasonable guess that provides a good starting point for MaxF can be obtained from statistical potentials.

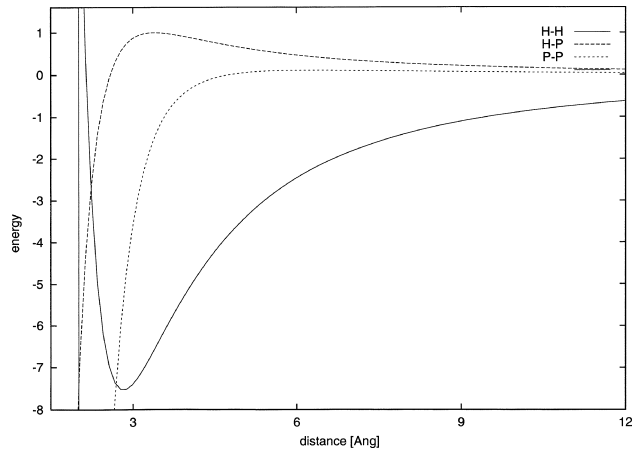


Figure 3. A Lennard–Jones-like potential for two types of amino acids obtained using the MaxF procedure. The functional form is $A_{\alpha\beta}/r_{ij}^6 + B_{\alpha\beta}/r_{ij}^2$, where the indices α and β denote the amino acid types, the indices i and j are the positions along the chain and the coefficients $A_{\alpha\beta}$, $B_{\alpha\beta}$ are optimized using the LP approach coupled with the MaxF guideline. Interactions of different types are denoted as HH, HP, and PP, where H stands for hydrophobic and P for polar residues, respectively. The coefficients A and B are given in Table 3. Note that, similar to the contact HP potentials, the HH interactions are highly favorable, whereas the HP and PP interactions contribute little to the energy because there are only few contacts corresponding to very short distances (251 with distances shorter than 3 Å and 9 with distances shorter than 2.5 Å, respectively, out of the 291,651 native contacts used in the training). There are no contacts in the training set with distances shorter than 2 Å, which corresponds to an infinite wall at that distance (represented as a vertical line in the figure).

of the LP approach to the design of folding potentials is that it allows exploring different possibilities and assess them using the infeasibility test.

Let us consider the widely used pairwise folding potentials.[8–10] The energy of the protein of a sequence S and a structure $\mathbf{X}$ is a sum of all pairs of interacting amino acids,

$$E_{\text{pairs}} = \sum_{i<j} \phi'_{ij}(\alpha_i, \beta_j, r_{ij}). \tag{6}$$

The pair interaction model—ϕ_{ij} depends on the distance between sites i and j, and on the types of the amino acids, α_i and β_j at sites i and j, respectively. We consider both: a simple contact potential and a continuous pairwise potential.

In the case of the contact potential, two amino acids are considered in contact if the geometric centers of the side chains are closer than 6.4 Å. The interaction model reads:

$$\phi_{ij}(\alpha_i, \beta_j, r_{ij}) = \begin{Bmatrix} \varepsilon_{\alpha\beta} & 1.0 < r_{ij} < 6.4 \text{ Å} \\ 0 & \text{otherwise} \end{Bmatrix}, \tag{7}$$

where $\varepsilon_{\alpha\beta}$ is a matrix of all the possible contact types (we drop the subscripts i and j for convenience). For example, it can be a 20×20 matrix for the 20 amino acids. Alternatively, it can be a smaller matrix if the amino acids are grouped together to fewer classes. The entries of $\varepsilon_{\alpha\beta}$ are the target of parameter optimization.

An example of a more realistic interaction model is the "distance power" potential:

$$\phi_{ij}(\alpha_i, \beta_j, r_{ij}) = \frac{A_{\alpha\beta}}{r_{ij}^m} + \frac{B_{\alpha\beta}}{r_{ij}^n}. \tag{8}$$

Two matrices of parameters are determined: $A_{\alpha\beta}$ and $B_{\alpha\beta}$. The indices m and n are predetermined in advance. We consider here the ($m = 6$, $n = 2$) model, which we found more accurate for the reduced representation of protein structure than the atomic Lennard–Jones (LJ) potential.[6] Hence, the index of the vector in eq. (5′), $\gamma \equiv \alpha\beta$, runs in our case over the types of contacts, whereas n_γ is the number of contacts of a specific type found in $\mathbf{X}$. In case of the LJ model, the "number" includes an additional geometric weight hidden in a continuous "number" function—$n_\gamma \propto 1/r^m$.

The set of eq. (1) can be rewritten now as follows:

$$\begin{aligned} E(\mathbf{X}_j; \mathbf{z}) - E(\mathbf{X}_n; \mathbf{z}) &= \sum_\gamma z_\gamma \left(n_\gamma(\mathbf{X}_j) - n_\gamma(\mathbf{X}_n)\right) \\ &= \mathbf{z} \cdot \Delta\mathbf{n}_{j,n} \geq \varepsilon \qquad \forall(j,n), \end{aligned} \tag{9}$$

where index j runs over the misfolded structures of a given protein, and index n runs over the native structures in the training set. The difference in contacts vector, $\Delta\mathbf{n}_{j,n}$, is a result of counting contacts of specific types in both native and misfolded structures. We solve the set of eq. (9) for $\mathbf{z}$, without optimizing an objective function. We use the BPMPD program of Cs. Meszaros,[23] which is based on the primal-dual interior point algorithm and allows us to compute a series of analytic centers according to the MaxF procedure. In practice, the right hand sides of the inequalities in (9) are set to be equal to a small positive number, $\varepsilon = 10^{-6}$. We also bound the variables, $-10 \leq z_\gamma \leq 10$, for each γ.

A convenient way to generate a set of misfolded structures is the so-called gapless threading. Consider a set of proteins $\{(\mathbf{X}_{n_k}, S_{n_k});\ k = 1, \ldots, N\}$. Each native sequence S_{n_i} is fitted without deletions and insertions into the other (longer) structures in the training set, $\mathbf{X}_{n_j}$, $n_j \neq n_i$, which provide alternative (misfolded) packing of the protein chain. Thus, each gapless alignment of a native sequence into an alternative structure provides one misfolded (decoy) structure and the corresponding inequality, as defined in eq. (9).

We use the Hinds and Levitt (HL) set of 246 proteins.[24] Gapless threading of all sequences into all structures generated the set of 4,003,727 inequalities (note that there are many ways a shorter sequence can be aligned to a longer structure), which we will refer to as the HL problem. We also use a subset of 627,567 constraints (referred to as the HLs problem) that result from aligning all the sequences into structures that are less than 33% longer. Thus, many alignments of very short sequences into long structures are excluded from the training set, reducing the size of the problem. Tobi and Elber's (TE) set of 594 proteins is used as a control set.[4] Gapless threading of all sequences into all structures in the TE set generates about 30 million of inequalities. We use the program LOOPP[25] to generate the inequalities for the LP training.

Results

In a previous work[6] we addressed the question of the minimal number of parameters that is required to obtain an exact solution for the HL problem. We found that the HL problem proves infeasible when using pairwise potentials with less than 10 types of amino acids (i.e., with less than 55 types of contacts between amino acids). Here, we revisit this problem using the "maximum feasibility" guideline.

We consider two reduced alphabets of amino acids: first of two letters only, namely H and P (for hydrophobic and polar residues, respectively), and the second of four letters, namely H, P, C_+, and C_- (C_+ standing for positively charged and C_- for negatively charged residues, respectively). The assignment of the different amino acids to the letters of the reduced alphabets is in Table 1. The HL and HLs problems are infeasible when formulated in terms of the four-letter alphabet. In other words, even the smaller (HLs) set of inequalities cannot be solved exactly with four types of amino acids, corresponding to 10 types of amino acid contacts.

Contact Model with Four Types of Amino Acids

We first apply the MaxF rule to the HLs problem in terms of a contact pairwise model, as defined in eq. (7). Four types of amino acids are employed. Results for a number of different starting points are discussed. The first initial guess is the statistical potential derived from the HL set of native structures. Notice that such a potential can always be generated for the problem at hands. Statistical potentials employ contact energies defined as logarithm of the properly normalized probabilities of observing a given type of contact.[8] With a proper choice of the sample of native shapes, the statistical potentials proved to be quite successful in distinguishing native from misfolded structures.[9, 10, 15, 16]

In our case, the statistical potential derived from the HL set of proteins for the four-letter alphabet (see Table 2) performs poorly. It does not satisfy 57,211 inequalities, and fails to recognize 144 proteins (that is for 144 proteins there are decoy structures with energies lower than the native energy). However, the dominant (stabilizing) contributions to the native energies come from the HH interactions. Therefore, one may expect that our initial guess still captures important characteristics of a good solution, with a significant room for improvement. Indeed, as we can see from Table 4, just the first iteration of MaxF procedure dramatically improves the initial solution. The analytic center of the first polytope, defined by all the inequalities satisfied by the initial guess, misses only 6,800 constraints and 22 proteins.

To characterize the shape of the distribution of energy differences in eq. (1), $\Delta E_{\mathrm{mis,\,nat}}$, we compute the so-called Z-score, which is defined as the ratio of the average over the standard deviation of the distribution, $Z = \langle \Delta E_{\mathrm{mis,\,nat}} \rangle / \sigma$. The Z-score of the distribution obtained with the statistical potential is equal to 1.22, and increases to 1.98 after the first iteration of MaxF. Hence, not only the tail of the distribution is corrected, but also the whole distribution is shifted away from the native energies. This is expected because the analytic center provides in general a more uniform distribution of energy differences (distances to the cutting hyperplanes), as compared to an off-centered guess.

We observe further improvement in the subsequent iterations. The converged solution, which we will refer to as 4HLs potential (short for the four-letter potential, trained on the HLs problem), misses only 1928 inequalities and 11 proteins. The inspection of the constraints that are not satisfied reveals that 1922 of them are due to six membrane proteins included in the HL set (1prcC, 1prcL, 1prcM, 4rcrL, 4rcrM, 2por). The remaining six constraints refer to five other proteins that are not recognized (1pp2R, 2bbkB, 2ltnA,

Table 1. Definitions of Different Groups of Amino Acids That Are Used in the Present Study.

Hydrophobic (H, HYD)	ALA CYS HIS ILE LEU MET PHE PRO TRP TYR VAL
Polar (P, POL)	ARG ASN ASP GLN GLY LYS SER THR
Positively Charged (C_+, CHG)	ARG LYS
Negatively Charged (C_-, CHN)	ASP GLU

When the charged residues are included explicitly the group of polar residues is reduced correspondingly. In a previous study we found that 10 types of amino acids were necessary to solve exactly the Hinds–Levitt set of proteins by pairwise interaction models.[6] Using the MaxF procedure we find that four types of amino acids are essentially sufficient to recognize all but membrane proteins in the HL set.

Table 2. Parameters for Contact Pairwise Potentials with Four Types of Amino Acids.

					MaxF				
	HYD	POL	CHG	CHN		HYD	POL	CHG	CHN
Init1									
HYD	−0.57	−0.55	−0.28	−0.17	HYD	−0.34	0.11	0.17	0.29
POL	−0.55	0.16	−0.22	−0.23	POL	0.11	−0.07	0.24	0.36
CHG	−0.28	−0.22	1.01	−0.97	CHG	0.17	0.24	1.00	−0.40
CHN	−0.17	−0.23	−0.97	0.82	CHN	0.29	0.36	−0.40	0.12
Init2									
HYD	−0.46	0.04	0.41	0.27	HYD	−0.45	0.08	0.37	0.35
POL	0.04	0.03	0.13	0.09	POL	0.08	0.08	0.32	0.14
CHG	0.41	0.13	0.60	−0.41	CHG	0.37	0.32	0.98	−0.65
CHN	0.27	0.09	−0.41	0.39	CHN	0.35	0.14	−0.65	0.12

A statistical potential resulting from the Hinds–Levitt set of proteins (denoted as Init1) and the converged MaxF potential obtained when using Init1 as a starting guess are given in the two upper blocks. A projection of 10-letter potential trained previously (denoted as Init2) and the converged MaxF potential obtained when using Init2 as an initial guess are included in the lower blocks.

2mev3, 3sdpA). Removing the membrane proteins as well as two other proteins (which were not recognized due to the presence of structural relatives) from the training set results in a feasible problem.

The quality of the 4HLs potential is comparable to the previously trained 10-letter potential,[6] despite the fivefold decrease in the size of the parametric space. When 4HLs potential is applied to the full HL problem, 23 proteins and 3652 inequalities are missed (3465 of them due to the membrane proteins). However, when applied to the larger TE set, both potentials recognize correctly the same number of proteins (504 out of 594). Hence, we were forced to use as many as 55 parameters just to solve the full HL problem exactly, without significant improvement in the performance on the TE set. The MaxF procedure effectively reduces the number of parameters by filtering out "hard" constraints due to inherently different protein environments.

Our second initial guess is adopted from the 10-letter potential that solves the HL problem exactly. Because the 10-letter alphabet contains our reduced model, we simply take the relevant 4×4 block from the table of energy parameters (see Table 2). Such a "guess" (which is a projection of the exact solution) is expected to perform well and indeed it only misses 3508 inequalities and 18 proteins. The converged solution is very similar to the previously obtained 4HLs potential. The new approximation misses the same set of 11 proteins. A slightly smaller number of constraints is now violated—1865, including 1858 due to the membrane proteins. When applied to the TE set the same 504 proteins are recognized.

We tried several different perturbations of the original statistical potential to further test the convergence of MaxF with different starting points. Physically motivated initial solutions converge to potentials resembling (numerically and in terms of performance) the 4HLs potential. On the other hand, MaxF procedure fails when starting from nonphysical potentials. Inverting the signs of the diagonal elements in Table 2, for example, results in a nonphysical potential that penalizes the HH contacts and misses 625,444 inequalities and 138 proteins. MaxF procedure yields in this case a potential that is still trapped in the subspace of the parametric space, in which the HH interactions are penalized. The final solution misses 624,466 inequalities and 138 proteins. The Z-score of the initial distribution is negative and remains essentially unchanged during iterations.

We remark that as many as 15 iterations may be needed until the procedure stops, and no further improvement is obtained. However, a nearly converged solution is found already after four to six iterations. We would also like to point out that we do not use a "warm" start at the subsequent iterations, for instance, we do not use the current solution to restart the LP solver in the next iteration. With the promise of better warm start strategies for interior point methods[28] we expect that our problem, with the subsequent iterations being only a perturbation over the previous problem, would be solved in a much smaller number of iterations. Each iteration of MaxF procedure for the HLs problem with 10 parameters (including formulating and solving the problem) takes several minutes on a SUN Ultra Sparc2 machine.

Contact Model with Two Types of Amino Acids

Can we further reduce the number of parameters, without deteriorating the quality of the potentials? Motivated by the relative success of the HP model advocated by Dill,[29] we consider only two types of amino acids. In the original Dill potential the interactions of pairs of amino acids other than HH are set to zero, $\varepsilon_{HP} = \varepsilon_{PP} = 0$, whereas $\varepsilon_{HH} = -\lambda$. The positive parameter λ determines the scale of the energy. For the HL problem, the Dill potential fails to predict the correct fold of 46 out of 246 proteins, violating only 29,129 inequalities. For the larger TE set, the Dill potential recognizes 456 of the 594 proteins. This result is remarkable considering the simplicity of the model.

We applied our procedure to the full HL problem, which contains significantly more constraints than the HLs problem. When starting with the Dill potential as the initial guess, we obtain only a modest improvement. The converged MaxF solution ($\varepsilon_{HH} = -0.57$, $\varepsilon_{HP} = 0.02$, $\varepsilon_{PP} = 0.05$) misses 41 proteins and 22,220

inequalities, including 20,336 inequalities due to the membrane proteins. The Z-score improves only slightly: from 1.91 to 1.94. A minor improvement is also observed for the TE set—472 proteins are recognized. When trying different perturbations of the Dill potential or an HP statistical potential, as the initial guess, we converge to very similar solutions (although the quality of the starting point may be much worse).

Thus, the physically motivated, effective projection of the problem into one-dimensional subspace is close to the best solution in the three-dimensional parametric space (for the sampling of misfolded structures by the gapless threading). Significantly better results are only obtained when the polar residues are further differentiated. This is additionally confirmed by the fact that the reduced HLs problem (that was solved exactly using four types of amino acids) proves infeasible with two types of amino acids.

Continuous Model with Two Types of Amino Acids

The last example we consider is the continuous model of the LJ(6,2) type, defined in eq. (8). We apply it to the smaller HLs problem, using two types of amino acids only, which corresponds to six energy parameters to be optimized. Despite the fact that we are using the same training set, this is a very different problem now, with real coefficients of the constraint matrix, n_γ [see eq. (9)].

To obtain an initial guess we take advantage of the LJ(6,2) potential in terms of 10-letter alphabet that solved the full HL problem.[6] This potential, with just 110 parameters, was shown to be comparable in performance to the best contact potential with 210 parameters and significantly better than 10-letter contact potential trained on the HL set.[6] Therefore, one might expect that, similar to the contact model, using the projection of such a potential into two-letter alphabet would provide a very good starting point.

The projected potential (see Table 3) performs poorly, however, missing 55,894 inequalities and 59 proteins. The MaxF procedure results in only a minute improvement—the converged MaxF potential is numerically very similar to the initial guess, and it misses 55 proteins and 50,405 inequalities. Setting parameters for HP and PP interactions to zero, while keeping A_{HH} and B_{HH} the same as previously, provides even worse guess that misses 84 proteins and 61,150 constraints. The MaxF procedure again fails to improve it significantly, resulting in a potential that violates 54,067 constraints and does not recognize 81 proteins. MaxF solutions are trapped in the neighborhood of the starting point.

Motivated by the relatively better performance of the contact HP model, we next start from a potential with a much softer repulsion term (denoted as Init2 in Table 3). As can be seen from Table 4, the new guess indeed performs much better. Only 18,985 inequalities and 49 proteins are missed, which is further reduced (after applying MaxF) to 12,362 inequalities (11,822 of them due to the membrane

Table 3. Parameters for LJ(6,2) Potentials with Two Types of Amino Acids.

	Init1			Init2			MaxF	
A_{ij}	HYD	POL	A_{ij}	HYD	POL	A_{ij}	HYD	POL
HYD	9.32	1.45	HYD	1.00	0.00	HYD	2.61	−1.06
POL	1.45	−1.19	POL	0.00	0.00	POL	−1.06	−4.26
B_{ij}	HYD	POL	B_{ij}	HYD	POL	B_{ij}	HYD	POL
HYD	−2.34	0.47	HYD	−2.34	0.00	HYD	−9.99	1.94
POL	0.47	0.01	POL	0.00	0.00	POL	1.94	0.69

A projection of 10-letter LJ(6,2) potential from ref. 6 (denoted as Init1) and its modification with a smoother HH repulsion term and HP, PP interactions set to zero (denoted as Init2), as well as the converged MaxF potential obtained when starting from Init2 are presented. Init1 provides a much worse initial guess, which is not improved significantly by MaxF. Note that the "repulsive" coefficients A are given first, followed by the "attractive" coefficients B. The coefficients are expressed in terms of the unit distance of 3 Å.

Table 4. Results of the MaxF Procedure for the Design of Reduced Folding Potentials.

	(N_ineq/N_prot/Z-score)		
Iteration	Contact, 4-lett, Init1	Contact, 4-lett, Init2	LJ(6,2), 2-lett, Init2
Initial guess	57,211/144/1.22	3508/18/1.99	18,985/49/1.74
First iteration	6800/22/1.98	3125/14/1.99	18,022/43/1.76
Converged MaxF	1928/11/2.01	1865/11/2.01	12,362/27/1.92

Gapless threading on the Hinds–Levitt set of 246 proteins[23] is used to generate inequalities for training. An infeasible (in terms of reduced alphabets) set of 627,567 inequalities is used (HLs problem—see the second section). Two types of folding potentials are considered to illustrate how the "maximum feasibility" guideline improves the initial solution, which satisfies certain subset of the constraints. The results for the contact pairwise model and four types of amino acids are presented in the second and third columns, using as the initial guess the statistical potential of Table 2 (Init1) and the projected 10-letter potential of Table 3 (Init2), respectively. The results for the continuous pairwise model of a Lennard–Jones 6-2 type, using as a starting guess a "soft repulsion" potential denoted as Init2 in Table 3, are included in the last column. For each potential the number of inequalities that are not satisfied (N_ineq), the number of proteins that are not recognized (N_prot) and the Z score at a given iteration are reported. Note that in each of the cases reported here a significant improvement with respect to the initial guess is achieved.

proteins) and 27 proteins only. The initial Z-score of 1.74 reaches 1.92 upon convergence.

However, when compared to the simple contact potential with two types of amino acids, the continuous pairwise model is not advantageous. Applied to the full HL problem, the LJ(6-2) potential violates 26,583 inequalities, and does not recognize 40 proteins. Applied to the larger TE problem, the new LJ(6,2) potential recognizes 476 out of 594 proteins, that is only four proteins more than the best contact HP potential we obtained and 20 proteins more than the simple Dill potential.

The two types of amino acids enforce a common distance law for side-chain centers of amino acids of very different volume. Pairs of the type Gly–Ala and Arg–Leu, for example, have the same interaction law. The difficulty with obtaining significant improvement using MaxF procedure and two types of amino acids, together with the success of the 10-letter LJ(6,2) potential (which treats explicitly small residues), may suggest the importance of differentiating amino acids of a different size.

Discussion

The problem of identifying the sources of infeasibility in LP problems (that often come simply from errors in the formulation of the problem) is of significant practical importance, and promotes development of heuristic methods for finding approximations to LFS.[19, 20] Probably the most popular method, implemented in some LP packages, is based on the idea of "elastic programming."[20, 30] Instead of solving the original (infeasible) problem one solves a modified problem, with "elastic" variables added first to ensure that the "elastic" problem has a solution and then iteratively removed until infeasibility is reached again.[20]

In principle, such an "elastic filter" could be used as well to remove the "hard" constraints. The MaxF guideline, applied to the resulting feasible subset of inequalities, would provide a partial solution of the problem. Unfortunately, numerical tests suggest that the elastic filter heuristic is not very effective for problems that require a removal of a large number of constraints to obtain a feasible subset.[31] Moreover, using the elastic filter approach implies that an elastic variable is added for each constraint, increasing dramatically the size of the LP problem when solving millions of inequalities. Therefore, such an approach is rather impractical.

Finally, we comment that the examples considered here are much smaller than those required to train folding potentials of sufficient accuracy. Our experience shows that a much more complete sampling of native and misfolded structures, resulting in a huge number of inequalities, is necessary. Such problems are very likely to be infeasible with simple functional models of folding potentials and a limited number of parameters.

However, due to the underlying physical principles, most of the constraints should be satisfied by commonly used statistical potentials. The MaxF procedure provides a simple way to improve further such potentials, both in terms of the number of inequalities that are not satisfied and in terms of the overall shape of the distribution of energy gaps, as defined in eq. (1).

References

1. Maiorov, V. N.; Crippen, G. M. J Mol Biol 1992, 227, 876.
2. Vendruscolo, M.; Domany, E. J Chem Phys 1998, 109, 11101.
3. Tobi, D.; Elber, R. Proteins Struct Funct Genet 2000, 41, 40.
4. Tobi, D.; Shafran, G.; Linial, N.; Elber, R. Proteins Struct Funct Genet 2000, 39, 71.
5. Meller, J.; Elber, R., submitted.
6. Meller, J.; Elber, R., submitted.
7. Akutsu, T.; Tashimo, H. Proc. Pacific Symposium on Bio-computing, 1998, p. 413.
8. Sippl, M. J.; Weitckus, S. Proteins 1992, 13, 258.
9. Miyazawa, S.; Jernigan, R. L. J Mol Biol 1996, 256, 623.
10. Godzik, A.; Kolinski, A.; Skolnick, J. Proteins Struct Funct Genet 1996, 4, 363.
11. Khachiyan, L. G. Doklady Akad Nauk USSR 1979, 244, 1093.
12. Karmakar, N. K. Combinatorica 1984, 4, 373.
13. Ye, Y. Interior Point Algorithms: Theory and Analysis; Wiley: New York, 1997.
14. Vanderbei, R. J. Linear Programming: Foundations and Extensions; Kluwer Academic Publishers: New York, 1996.
15. Wright, S. J. Primal-Dual Interior-Point Methods; SIAM Publications: 1997.
16. den Hertog, D. Interior Point Approach to Linear, Quadratic, and Convex Programming; Kluwer Academic Publishers: New York, 1994.
17. Adler, I.; Monteiro, R. D. C. Math Program 1991, 50, 29.
18. Monteiro, R. D. C.; Adler, I. Math Program 1989, 44, 43.
19. Chakravarti, N. Eur J Oper Res 1994, 73, 139.
20. Parker, M.; Ryan, J. Ann Math Artif Intell 1996, 17, 107.
21. Chinneck, J. W. INFORMS J Comput 1997, 9, 164.
22. Garey, M. R.; Johnson, D. S. Computers and Intractability: A Guide to the Theory of NP-Completeness; W.H. Freeman and Company: New York, 1979.
23. Meszaros, C. S. Comput Math Appl 1996, 31, 49.
24. Hinds, D. A.; Levitt, M. J Mol Biol 1994, 243, 668.
25. Meller, J.; Elber, R. http://www.tc.cornell.edu/CBIO/loopp.
26. Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. J Comput Chem 1997, 18, 849.
27. Xia, Y.; Huang, E. S.; Levitt, M.; Samudrala, R. J Mol Biol 2000, 300, 171.
28. Yildrim, E. A.; Wright, S. J. Technical Report 1258, School of Operations Res. and Industrial Eng., Cornell University.
29. Chan, H. S.; Dill, K. A. Proteins Struct Funct Genet 1998, 30, 2.
30. Brown, G.; Graves, G. Presented at ORSA/TIMS conference, Las Vegas, 1975.
31. Chinneck, J. W. Ann Math Artif Intell 1996, 17, 127.

Mathematical Programming manuscript No.
(will be inserted by the editor)

Michael Wagner[1] · Jarosław Meller[2] · Ron Elber[3]

Large-Scale Linear Programming Techniques for the Design of Protein Folding Potentials

April 15, 2003

Abstract. We present large-scale optimization techniques to model the energy function that underlies the folding process of proteins. Linear Programming is used to identify parameters in the energy function model, the objective being that the model predict the structure of known proteins correctly. Such trained functions can then be used either for *ab-initio* prediction or for recognition of unknown structures. In order to obtain good energy models we need to be able to solve dense Linear Programming Problems with tens (possibly hundreds) of millions of constraints in a few hundred parameters, which we achieve by tailoring and parallelizing the interior-point code `PCx`.

Key words. protein folding, interior-point algorithm, PCx, Linear Programming, linear feasibility, parallel processing

1. Introduction

The recent unveiling of the human genome marked the transition in the biological sciences toward the post-genomic era, in which the understanding of protein structure and function becomes a crucial extension of sequencing efforts. Despite recent progress in high throughput techniques, the experimental determination of protein structure remains a bottleneck in structural genomics. This poses a challenge and an opportunity for computational approaches to complement and facilitate experimental methods.

The protein folding problem consists of predicting the three-dimensional structure of a protein from its amino acid sequence. The methodology and modeling aspects of protein folding have been vastly discussed in the literature (for excellent and up-to-date brief surveys of methods as well as their limitations, see, e.g., [6] and [14]). In order to characterize the existing computational approaches to this problem one may distinguish two underlying principles.

The so-called *ab-initio* protein folding simulations attempt to reproduce the actual physical folding process using the thermodynamical hypothesis which was

Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529–0077 and Pediatric Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH 45229–3039. e-mail: `mwagner@cchmc.org`.

Pediatric Informatics, Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH 45229–3039 and Department of Informatics, Nicholas Copernicus University, 87-100 Toruń, Poland. e-mail: `jmeller@cchmc.org`.

Department of Computer Science, Cornell University, Ithaca, NY 14853–7501. e-mail: `ron@cs.cornell.edu`.

first introduced by Anfinsen [5] in the early 1970's. The unique three-dimensional structure of a protein is postulated to correspond to a global minimum of the free energy function. The search for the native conformation thus entails solving a global optimization problem.

The *protein recognition* approach, in turn, relies on the fact that large numbers of native protein folds have already been determined. Given an appropriate scoring function, which can be thought of as a simplified folding potential, these methods find the "best" template from the library of known folds. In other words, the search for the native conformation is restricted to the set of known structures, as opposed to an expensive search in the space of all possible conformations. The scoring functions for protein recognition can be based on amino acid sequence similarity, or they may incorporate measures of sequence-to-structure fitness. The latter approach, known as *threading*, allows finding distant homologs that share the same fold without detectable sequence similarity [8].

In both *ab-initio* folding and protein recognition we are faced with the problem of finding (designing) an appropriate expression for the free energy or scoring function, respectively. While optimization tools are certainly crucial for finding the native conformation [12], they also play an important role in the energy modeling stage [17]. This paper introduces new, tailored optimization tools for the design and evaluation of folding potentials with superior prediction and recognition capabilities.

The energy functions we consider here depend linearly on parameters. As discussed in [22], the linear dependence of the potential functions on their parameters is not a major restriction. Any nonlinear function can be expanded or at least approximated as a linear combination of basis functions. The challenge is to find a set of basis functions of small cardinality that captures most of the intrinsic complexity of the true energy function and thus makes for a reasonable model. The tools we present here allow us to *evaluate* the power of different modeling approaches (basis functions), so that over time we expect these to become increasingly more sophisticated. Our purpose in this paper is not so much to find *the* "optimal" model, but rather to illustrate the usefulness of large–scale optimization in this context.

The requirement of perfect recognition of known structures results in a linear feasibility problem (pioneered by Mairov and Crippen [19]), which we solve using Linear Programming techniques. We show that our large-scale tools, which allow for the solution of systems with hundreds of millions of constraints, result in significant improvements in the quality of potentials. We also demonstrate how solving these very large Linear Programming problems in conjunction with the recently proposed "Maximum Feasibility" heuristic [23] may be used to evaluate different functional forms. By enabling a comparison of the power of different approaches we aim to obtain insight into the question of how complex models need to be in order to provide reasonable fold recognition capabilities. Our ultimate goal is an "optimal" energy model which balances complexity and accuracy, while avoiding the dangers of over- and underfitting. We believe this work is a step in that direction.

The structure of this paper is as follows: In Sections 2 and 3 we present the parameter identification problem and a Linear Programming solution to it, respectively. Section 4 describes some computational results and their biological interpretation for several commonly used models. We conclude with an assessment of the power and usefulness of our tools and by pointing to future research directions.

2. Potential Function Modeling for Protein Folding

2.1. Designing the Functional Form of the Potential Model

Proteins are linear polymers composed of a sequence of amino acid residues that are connected by peptide bonds (creating the protein "backbone"). There are 20 different amino acids that are characterized by chemically unique side chains (containing from one to approximately 20 atoms) that hang off the backbone chain. Protein molecules consist of several tens to several thousands of amino acids and thus between a few hundred and tens of thousands of atoms.

Protein structure is often represented in terms of simplified, reduced models that speed up computation. For example, the commonly used contact model represents each amino acid by just one point in $\mathbb{R}^3$, which defines the approximate location of an amino acid. The overall shape of the protein is characterized in terms of contacts between closely packed amino acid residues. Such contact models allow us to capture the packing of hydrophobic residues that are buried in the core of the protein and contribute to the stability of the structure.

In the present work we consider energy functions that employ reduced, contact models of protein structure. We will use the terms structure, fold, or conformation to mean the three-dimensional structure of the protein as defined by a set of coordinates of the geometric centers of the amino acid side chains. Also, the terms side chains (centers), amino acids, and residues are meant to be synonymous. Finally, following our earlier discussion, we will use the terms energy function, scoring function and potential function interchangeably.

We will denote our models of the potential function by E, and we will write it as a function of a sequence of amino acids s and a three-dimensional structure (a triplet of coordinates) x. The energy models considered here may be expressed in terms of functions $\varphi_i(s, x)$:

$$E_y(s, x) = \langle \Phi(s, x), y \rangle,$$

where y is a vector of parameters that are to be determined, $\Phi = (\varphi_1, \varphi_2, \ldots, \varphi_n)$, and $\langle \cdot, \cdot \rangle$ denotes an inner product. The set of functions $\{\varphi_i\}$ may be thought of as a set of *basis functions*. The "right" choice of basis functions is critical for the quality of the model, and the tools presented here allow one to explore different possibilities.

We illustrate this general approach with a few examples of energy models that have been used in the field of protein folding. The interested reader is referred to [14] for a much more in-depth presentation.

For example, in the *pairwise contact potential* two amino acids of type α and β ($\alpha, \beta \in \{1, 2, \ldots, 20\}$), respectively, are said to be in contact if the distance of their geometric centers is less than a certain threshold (here we use the distance of 6.4 Å[21]). The energy as a function of a given sequence of amino acids s and a given three-dimensional structure x can thus be expressed in the following way:

$$E_y(s, x) = \sum_{\alpha \leq \beta} y_{\alpha,\beta} N_{\alpha,\beta}(s, x), \tag{1}$$

where $N_{\alpha,\beta}(s, x)$ represents the number of α-β contacts when sequence s is folded into structure x, and $y_{\alpha,\beta}$ are (unknown) weight parameters which represent the contribution such a contact makes toward the overall energy of the molecule.

In Section 4.2 we will refer to two other models, discussed in detail in [21]. We call them *threading onion models* (THOM) since they characterize the structural environment ("profile") of an amino acid in terms of its contact shells. Profile models, contrary to pairwise models, have the advantage that the optimal alignment with gaps can be efficiently computed using dynamic programming [18].

THOM1 models define the type of a residue using the first contact shell only. They are meant to capture the solvent exposure of amino acid residues. The nature of a given contact is disregarded, and one simply counts the number of times a side chain of type α has a given number of neighbors (contacts):

$$E_y(s, x) = \sum_{\alpha} \sum_{m=1}^{k} y_{m,\alpha} N_{m,\alpha}(s, x). \tag{2}$$

$N_{m,\alpha}(s, x)$ represents the number of times a side chain of type α has m contacts, k is the maximum number of contacts and $y_{m,\alpha}$ are parameters to be determined. THOM2 models, which include the second contact shell (neighbors of neighbors), are meant to mimic pairwise interactions while preserving the efficiency of profile models (see [21] for details).

2.2. Optimization of the Parameters of the Potential Function

One traditional and widely used approach to finding values for the parameter vector y has been to derive them from statistical information about native folds that are already determined. For example, for the contact potential (1), a statistical potential would be defined by setting

$$y_{\alpha,\beta} = -C \ln(p_{\alpha,\beta}/(p_\alpha p_\beta)). \tag{3}$$

C is a constant that defines the energy units, p_α and p_β are the respective frequencies with which the amino acids appear in the chain, and $p_{\alpha,\beta}$ is the frequency of contacts of that type [27].

These statistical, knowledge-based potentials learn form the native structures ("good" examples) only. In order to increase their power to distinguish misfolded states (the "bad" examples) from native states, more sophisticated protocols

incorporate data from decoy folds. To achieve this, we demand that the models mimic the postulate that the native state have the lowest energy. If we denote the native structure of a given sequence s by x_s^*, then the perfect potential function should satisfy:

$$E_y(s,x) > E_y(s,x_s^*) \qquad \forall s, \forall x \neq x_s^*,$$

or, using the expansion in terms of basis functions,

$$\Delta E_y = \langle \Phi(s,x) - \Phi(s,x_s^*), y \rangle > 0 \qquad \forall s, \forall x \neq x_s^*. \tag{4}$$

A slight but meaningful generalization arises when we introduce the notion of a distance between structures in order to distinguish between "close to native" (but misfolded) versus radically different structures. By demanding that the energy gap for the latter be larger than for the former we achieve hierarchical ordering of misfolded states (known as "funnel" in the protein folding literature). In this case we have reason to demand that

$$\langle \Phi(s,x) - \Phi(s,x_s^*), y \rangle \geq b_{x,x^*} \qquad \forall s, \forall x \neq x_s^* \tag{5}$$

for appropriately chosen numbers $b_{x,x^*} > 0$ which in general should be proportional to the distance between the native and misfolded structure.

One approach to designing potentials that improves upon statistical potentials is z-score optimization [15]. Here the quality of a parameter vector y is measured using the distribution of energy gaps ΔE_y defined in (4). In particular, the goal is to maximize the dimensionless ratio of the first and second moments of the distribution (the "z-score")

$$z(y) = \frac{\mu(\Delta E_y)}{\sigma(\Delta E_y)}, \tag{6}$$

where μ and σ denote the mean and standard deviation of the energy gap distribution, respectively. The quantity z is, of course, nonlinear in its arguments. While z-score optimization may lead to remarkable improvements in the quality of the trained potentials, it is heuristic in nature and does not rule out negative energy gaps.

Our goal is to attempt to adapt the models by choosing the parameters y such that (4) holds explicitly. In other words, we would like the models of the energy function to *perfectly recognize* native structures. To this end, we sample misfolded conformations to form a finite system of linear inequalities. The prediction and recognition capability of the resulting model will depend greatly on the number and type of misfolded structures that are included in the consideration. We employ a simple procedure to generate decoy structures called *gapless threading*, in which sequences are (imaginatively) folded into structures that are known not to be their native states [21].

In general the number of parameters in the models that are of relevance to us is on the order of a few hundred. We aim to allow for the solution of problems with hundreds of millions of constraints, resulting from extensive sampling of misfolded structures. Given sufficient diversity of sampled types of proteins and

a large set of inequalities (one per decoy), one may hope that an appropriate set of basis functions $\{\varphi_i\}$ would capture the essential features of the energy function so that the model E recognizes the structures in the database correctly. In Section 4.2 we will use training sets of decoy to find parameters y and then verify the prediction capability of the resulting model on a different test set of decoys and structures.

There are a number of techniques to solve linear systems of inequalities (see, e.g., [31] for alternatives in the protein folding context); we focus on Linear Programming here. Linear Programming is equivalent to solving linear inequality systems, and the modern algorithms we use allow for the efficient solution of problems with the dimensions we are interested in.

3. Linear Programming Solutions

The requirement that the parameters y define a model that satisfies the inequalities (4) for a set of decoy structures can be written as a system of strict linear inequalities

$$A^T y > 0, \tag{7}$$

where $A^T \in \mathbb{R}^{m \times n}$. Typically n is on the order of a few hundred (one column per basis function φ_i) and m is on the order of tens of millions or more (one row per generated decoy fold). We note that if a solution to (7) exists, then it can be scaled to satisfy the system

$$A^T y \geq \rho \mathbb{1}, \tag{8}$$

which is a problem more amenable to computation. $\rho > 0$ is an arbitrary constant and $\mathbb{1}$ is the vector of ones, which is chosen merely for convenience. Our specific choices for ρ will be discussed in Section 4.2. In the more general case (5), which we will refer to from now on, we get a system

$$A^T y \geq b, \tag{9}$$

where $b > 0$ is the vector of desired energy gaps.

3.1. Modeling Techniques and Choice of Algorithm

There are a number of ways to cast (9) as Linear Programming problems. Since any feasible y can be scaled with a positive constant one might think of imposing a constraint on the norm of y in order to bound the feasible region (see, e.g., [29] for an example of this approach). However it is not a priori clear that the resulting system is feasible since we have just introduced an (arbitrarily scaled) right-hand side. Hence for now we refrain from introducing this scaling of y explicitly and instead rely on the quality of the software used to produce a well-scaled parameter vector whenever possible.

Our first approach lies in adding a trivial objective function to get

$$(P) \qquad \begin{array}{l} \min\ 0^T y \\ \text{s.t. } A^T y \geq b. \end{array} \tag{10}$$

It is instructive to look at the corresponding dual problem:

$$(D) \qquad \begin{array}{l} \max\ b^T z \\ \text{s.t. } Az = 0 \\ \qquad\quad z \geq 0. \end{array} \tag{11}$$

We see immediately that the dual problem is always feasible, and in fact that either the origin is the only feasible and hence optimal solution or the dual problem is unbounded (which implies an infeasible primal constraint system). Both of these cases are obviously of interest to us. We discuss their relevance to us in the following section.

There are two prevalent types of software for Linear Programming: those codes which are based on the simplex method and those based on the more recent interior-point methods. Although we don't want to rule out that a sophisticated implementation of the simplex method (with column generation techniques) might be successful in this case, we note that simplex-based methods are not easily parallelized and are likely to run into difficulties due to the degeneracy of the problems.

Instead we focus on using *interior-point* algorithms to solve (P). The interested reader is referred to [32] for an excellent in-depth introduction to these methods. We constrain ourselves to pointing out some of the features that are important in this context. Interior-point methods are Newton-like iterative methods that solve a sequence of perturbed KKT systems. Most importantly, they enjoy polynomial-time convergence properties and have been implemented in very efficient software that is competitive with implementations of the celebrated simplex method. Usually (i.e., for reasonably sized problems) the major computational effort required in each iteration lies in forming a matrix of the form AD^2A^T and then solving a linear system with this matrix using a modified Cholesky factorization. (A is the matrix of the linear equality constraints given to the solver, and D^2 is an iteration dependent diagonal matrix.)

Interior methods have another feature which is beneficial in the context of our application (besides being amenable to parallel computation). Ideally we would like the energy gaps ΔE_y from (4) to be as large as possible. This would mean that the native structures have significantly less energy than any misfolded shapes, something that is generally conjectured to be the case for the true energy function also. This corresponds to having a solution that is, in some sense, "centered," i.e., where the distance to the boundary of the polyhedron is maximized. Interior-point algorithms are known to converge to the *analytic center* of the primal-dual optimal face. While the analytic center in general is not identical with the geometric center, this nevertheless bodes well for the solution being away from the boundary of the polyhedron. Since the system $A^T y \geq b$ is unbounded (and thus also the optimal face of (10)), the notion of an analytic

center is not well-defined in this context. Nevertheless, and even though there is no theoretical guarantee that the algorithm will produce nicely scaled and centered solutions, our experience has never produced examples where this is not the case.

A more sophisticated LP-modeling approach which we mention here avoids the aforementioned unboundedness of the optimal face by minimizing the norm of the parameter sought. Additionally, it deals explicitly with infeasibility by introducing slack variables z and minimizing their norm:

$$(P') \qquad \begin{array}{ll} \min & \|y\|_1 + \gamma\|z\|_1 \\ \text{s.t.} & A^T y + z \geq b \end{array} \tag{12}$$

Here γ is a tradeoff parameter that must be chosen in advance. The dual problem can be written in the following way:

$$(D') \qquad \begin{array}{ll} \max & b^T x \\ \text{s.t.} & -\mathbb{1} \leq Ax \leq \mathbb{1} \\ & 0 \leq x \leq \gamma\mathbb{1} \end{array} \tag{13}$$

The advantage of this formulation is that both problems are guaranteed to be feasible and their respective optimal faces are guaranteed to be bounded, which implies that their analytic center is well-defined.

This formulation is reminiscent of Support Vector Machines (see, e.g., [13]), except that the 1-norm is used for the minimization of $\|y\|$. Support Vector Machines are quadratic programming problems with vast applications in data mining and data classification. Our particular case can be interpreted as finding a separating hyperplane between the energy gaps and the origin, so that one of the two data classes effectively just consists of a single vector (the zero vector). We conjecture that an efficient implementation of a massive support vector machine (such as the one presented in [13]) will be a viable alternative to our linear programming approach.

Turning now to our specific application: we note that if problem (P) or (P') were to be fed to any of the interior solvers we are aware of, then slack variables would be introduced to transform the (primal) constraints into equalities. As a consequence the resulting system AD^2A^T would have millions of rows and columns (for problems of the size we are interested in) and be completely dense, making any computation with it unrealistic. However, the respective dual problems are already in the standard form which solvers use internally, and the system to form and solve in this case has row and column dimensions of a few hundred (and would thus be comparatively trivial!). We conclude that if we can hold the constraint matrix in a distributed computing environment and allow for matrix-matrix and matrix-vector multiplications, we can use standard interior-point algorithms to solve these problems. We also note that the dimensionality of (D') is only marginally larger than that of (D); the computational effort required to solve either one will essentially be the same.

3.2. Dealing with Infeasibility and Insufficient Memory

In the previous section we alluded to the case where the system of inequalities (7) (or (9)) admita no solution. If this is the case then this simply means that the model characterized by the set of basis functions $\{\varphi_i\}$ in question is not sufficiently sophisticated to correctly recognize all the proteins in the database (with the chosen desired energy gaps b). From a conceptual point of view, this outcome is certainly valuable information and an important conclusion when a given model is to be evaluated. For example, [30] and [28] show this way that the simple contact potential is in fact not generally good enough to recognize all structures that are already known.

However, the issue is more subtle than a simple decision of whether the linear inequality system is feasible or not. Not including enough native and misfolded structures in the training set can result in "underfitting" of the parameters for a given model, which is likely to result in poor performance on a larger test set and in real applications. On the other hand, with more extensive sampling the chances of introducing inconsistent constraints increase, which might lead one to resort to smaller training sets to avoid infeasibilities. Again, the resulting potential may again be significantly underfitted. Alternatively, in order to obtain perfect recognition on the training set, one may be inclined to increase the number of parameters (basis functions) in the model, risking significant overfitting. A striking example of this type is discussed in Section 4.2.

In [23] we discuss a case in which adding membrane proteins to a database of soluble proteins, which are characterized by different folding principles, makes the problem infeasible. In order to find a potential which recognizes this augmented set of proteins correctly the number of parameters and basis functions needs to be increased by an order of magnitude compared to the potential for the problem without the membrane proteins.

This motivates the need to deal with infeasible (or near-feasible) problems in an efficient way in order to still obtain meaningful models, e.g., by attempting to correctly recognize a maximum number of proteins. One idea to approximately achieve this is to try to find a maximal subset of satisfiable constraints. This problem, which is known as the maximum feasible subsystem problem (MAX FS), turns out to be not only NP-hard to solve to optimality ([9], [26]). Additionally it has been shown that it does not admit a polynomial time approximation scheme (unless P = NP) [3]). Some exact and heuristc algorithms have been proposed (see, e.g., [24], [10], [2], [25]), but none of these has been tested on problems of the dimensions with which we are concerned. For more details and additional references the reader is referred to [4], [25] or [16]. In [23] we introduced a *Maximum Feasibility* (MaxF) heuristic that aims at finding a "maximally feasible" parameter y, i.e., a parameter that satisfies the largest number of constraints possible. We summarize it here as Algorithm 1.

We stress that this is only a heuristic and one whose performance will depend critically on the choice of a good starting point. Nevertheless, as we show in Section 4.2, we have found it to be very useful in our application. Starting, for example, from a statistical potential (3), which can always and easily be

1: Set $k = 0$, start with an initial approximate solution y_0.
2: **loop**
3: Form A_k^T and b_k by finding all rows of A^T such that $A_k^T y_k \geq b_k$ holds.
4: **if** no new rows are added **then**
5: STOP.
6: **end if**
7: Compute a centered solution by running an interior-point algorithm.
8: Let y_{k+1} be the solution obtained. Set $k = k + 1$
9: **end loop**

Algorithm 1: The MaxF heuristic.

computed, the interior-point solutions to the subproblems in the heuristic each result in further improvement of the quality of the solution. Another plausible initial solution can be obtained by carefully selecting a subset of proteins for which we want to impose perfect recognition, and which is sufficiently diverse to capture the underlying, dominating physical characteristics of the folding process.

Note that in order to use the MaxF heuristic we need to be able to load all currently satisfied inequalities into memory. For approximate solutions of a good quality most of the constraints should be satisfied, which again motivates the need for parallel solvers that can handle very large problems.

If the number of generated inequalities that are of interest causes the problem to be too large to fit into the available memory, then we have little choice but to resort to an iterative scheme in order to try to find a feasible solution (or prove infeasibility). In particular this was often the case when we were constrained to a single-processor environment [28] [21]. We heuristically choose a subsystem that is small enough to fit into memory, try to find a feasible point and check whether the solution satisfies the rest of the constraints. If some of the inequalities are violated, they are used to replace some of the constraints of the original subproblem, and the procedure is repeated. It may be necessary to intervene manually to get this process to converge in reasonable time.

If the number of degrees of freedom is small compared to the number of inequalities that can be solved in one shot, then this approach has proven to be fairly successful if the problem was feasible. It is not difficult to see that, regardless of the constraint selection procedure, this procedure is not guaranteed to terminate if the original system is infeasible to start with. Even though our applications do not seem to pose great difficulties in finding infeasible subsets of inequalities in case the whole system is infeasible, we really would like to avoid having to resort to these iterative heuristics, and being able to solve large problems in one shot becomes crucial.

Table 1 summarizes our discussion. There are essentially two ways of avoiding the undesirable case of needing to deal with an infeasible system which is too large to fit into memory. First, by implementing a parallel code for a distributed memory environment, we are able to solve larger problems. Second, with increasing sophistication, the models tested are more likely to be able to

recognize increasingly larger numbers of protein structures and are hence less likely to produce infeasible systems. We conclude that the challenge is addressed to both computational scientists *and* biochemists to increase the quality of the models and the scalability of the software.

	problem fits into memory	problem is too large
feasible problem	"easy"	· heuristic iterative scheme · works if subproblems are large enough
infeasible problem	· get proof of infeasibility · use MaxF heuristic	· heuristic might cycle · **want to avoid this case**

Table 1. Strategies of dealing with infeasible or large problems

3.3. `pPCx`*: A Tailored Parallel Dense Implementation*

The problem given by (10) with tens of millions of inequalities cannot be solved by conventional and readily available software. Given the dimensions and the fact that the constraint matrix A is in general almost completely dense we need to be able to resort to a distributed memory environment in order to have a chance of solving these problems without having to use the heuristic iterative schemes outlined previously. As mentioned before, the dual problem (11) is more amenable for solvers since it is already in the commonly used standard form. Hence we will always let the solvers work on (11), and the variables of interest to us will be the dual variables. The formulation (13) has not been implemented yet; this will happen in a future version.

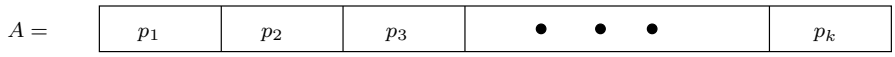


Fig. 1. Distribution of the constraint matrix

Our approach was to tailor the interior-point software `PCx` [11] to fit our needs. `PCx` is a publicly available, state-of-the-art serial implementation of a primal-dual predictor-corrector interior-point algorithm which enjoys widespread popularity in the optimization community. We replaced the sparse serial data structures and basic linear algebra routines by parallel dense counterparts. In particular, the constraint matrix A as well as all long vectors (vectors of length m) like the variables z are held in distributed form only. The distribution is done in the obvious way, with each of the k processors holding m/k columns of the matrix A (see Figure 1). This way, we expect the formation of the matrix AD^2A^T to speed up linearly with the number of processors. We can easily avoid having to store both the matrix A and its transpose by forming AD^2A^T as a

sum of outer products (rank-one updates):

$$AD^2A^T = \sum_{i=1}^{m} d_i^2 a_i a_i^T, \tag{14}$$

where a_i denotes the ith column of A and d_i the ith diagonal entry of D. Note that the matrix AD^2A^T is small and that the effort to solve the associated system is comparatively negligible. Hence we expect the bottleneck in this case to lie in forming the matrix. As a consequence of the expected linear speedup for the formation of the matrix AD^2A^T we expect the overall code to scale well with increasing problem size and number of nodes.

The computations also require several matrix vector products in each iteration, both involving the matrix A and its transpose A^T. Since the short vectors of length n are kept in serial (i.e., each processor owns a copy of each short vector), forming Az requires nontrivial communication among the processors, which does not scale as well.

In order to leverage off of existing parallel linear algebra packages we chose to use the data structures provided by the package PLAPACK [1]. However, the overhead associated with the PLAPACK routines (e.g., matrix-vector multiplications) is so significant that we chose to re-implement all necessary BLAS routines in order to speed up the code (an earlier implementation using the PLAPACK routines turned out to be impractically slow). The solution of the linear system is done using a modified version of the parallel Cholesky solver provided by PLAPACK (see [11] for details on that modification).

At present we can store approximately 220,000 inequalities in 210 parameters per GB of RAM. Note that, for convenience, we store the matrix entries in double precision format and do not yet exploit the fact that these entries typically are small integers. This will change in a future version of `pPCx`.

The data is generated using a package called `loopp` developed by Meller and Elber [20] which performs the threading of the sequences into structures in order to generate the decoys. This process is currently done in serial for simplicity; parallelizing this should be straightforward and will be done in the near future. Each process then reads the local portion of the matrix A from files, the data is put in the appropriate data structures and the core optimization code is called. Preprocessing of the constraint matrix is turned off since it would require accessing and comparing entire columns and rows of the constraint matrix and thus require significant communication overhead. Our experience shows that the models investigated do not require preprocessing in the sense that the code does not fail because of linear dependencies.

We ran our code on the Microsoft Windows 2000 based Velocity Cluster at the Cornell Theory Center. This machine consists of 64 nodes with 4 Pentium III-based processors per node running at 500 MHz and with 4 GB of main memory and 50 GB of disk space per node. For optimal performance we ran the code on at most 2 processors per node. The largest problem we solved so far consisted of approximately 60 million inequalities with 180 parameters. Since the implementation is entirely written in C with MPI extensions it is entirely

portable to other platforms. Specifically, one could imagine running on a large network of (possibly heterogeneous) workstations as long as the communication between them is not too much of a bottleneck.

4. Results

4.1. Parallel Performance

Our main interest is to find a feasible solution to the parameter identification problem (9), which corresponds to finding a dual feasible solution for the problem that is given to the Linear Programming solver. We modified the termination criteria in `pPCx` slightly to reflect this somewhat special case. Our experience with problems of different sizes is that typically between 5 and 20 iterations are necessary to find an optimal solution, and up to 60 if the problem is infeasible. The number of iterations obviously depends on the particular choice of right-hand side in (10). In particular, the number of iterations will depend on the choice of the constant ρ in (8). For the experiments presented here we chose $\rho = .01$ since this seems to represent a reasonable balance between computation time and feasibility of the resulting parameter vector.

The solution times vary from a few minutes (for problems with only a few hundred thousand constraints) to about 2.5 hours for a feasible problem with ca. 60 million constraints, solved on 128 processors.

	34 processors	64 processors
`InitTime`	176.30 (1.7 % of total)	97.62 (1.7 % of total)
`LoopTime`	10158.95 (98.1 % of total)	5664.12 (98.2 % of total)
`FormADATTime`	8040.23 (79.1 % of loop)	4424.49 (78.1 % of loop)
`PredictorTime`	677.52 (6.7 % of loop)	397.58 (7.0 % of loop)
`CorrectorTime`	693.08 (6.8 % of loop)	403.68 (7.1 % of loop)
`Factorization`	42.50 (0.4 % of loop)	43.43 (0.8 % of loop)
`TotalTime`	10351.72	5770.57

Table 2. Scalability on a problem with 30 million constraints

Table 2 shows the performance on a problem with 30,211,442 constraints and 200 parameters. The somewhat unorthodox choice of numbers of processors is solely due to memory requirements: the matrix does not fit onto just 32 processors. The problem is infeasible, which for the purposes of this evaluation is irrelevant. The Linear Programming solver took 57 iterations to terminate. Note that these solution times refer to the Linear Programming part only and do not include the data generation performed by the threading software `loopp`.

The `TotalTime` figure is the sum of `InitTime` (setup time plus time to find an initial point) and `LoopTime` (the main loop in the interior-point algorithm). `LoopTime`, on the other hand, is the sum of `FormADATTime` (the time it takes to

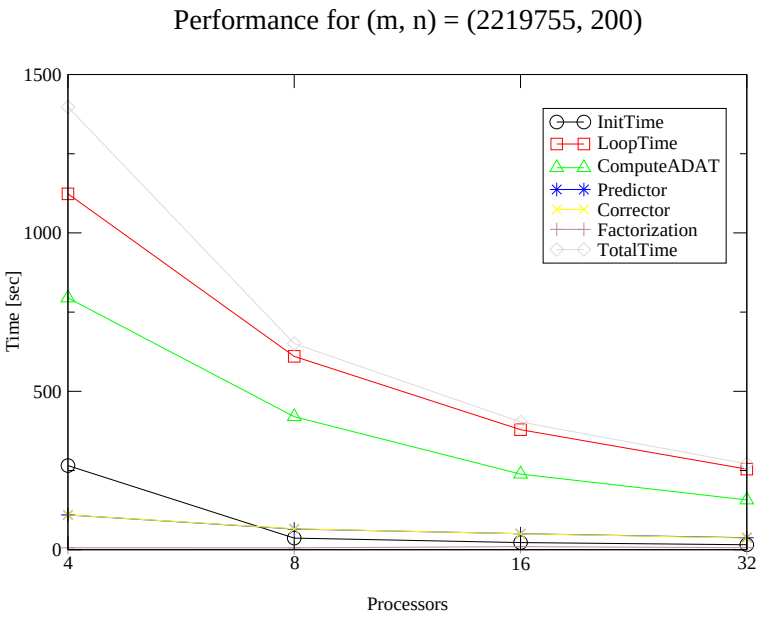


Fig. 2. Scalability on a problem with 2.2 million constraints

form the Schur complement matrix), `PredictorTime` and `CorrectorTime` (time to compute the components of the search direction) and `Factorization`. As expected, the computation is largely dominated by forming the Schur complement matrix (14), and thus the speedup for a problem of this size is linear, as expected. The other parts of the computations don't speed up as well, due to more communication overhead associated with the matrix-vector products of the form $y = Az$. The factorization of the 200×200 matrix is done using the PLAPACK code and does not speed up, probably because the matrix involved is too small. At this stage we are not concerned about that since the computation time spent on the factorization is negligible.

For a more comprehensive demonstration of the code scalability we present results of an experiment done with an arbitrarily chosen subset of $2,219,755$ inequalities from the original constraint set of 30M. Figure 2 shows the overall speedup as well as the speedups of the various components of the algorithms. `pPCx` took 17 iterations to find a solution.

We see an overall speedup factor of roughly 1.8, i.e., doubling the number of processor results in a reduction of about 4/9 computation time. It is not surprising that this is somewhat less impressive than for the larger problem presented earlier, since the percentage of computation spent in forming AD^2A^T is smaller. The speedup factor for `ComputeADAT` is closer to 2.

4.2. Applications to the Design of Folding Potentials

We applied `pPCx` in conjunction with the `loopp` threading program [20] to evaluate and design several folding potentials. To keep the scope of this paper contained we present a few representative computational results that are meant to illustrate the power and value of the algorithms discussed in the previous sections. The first set of experiments we present here consists of applying the MaxF heuristic discussed in Section 3.2 to potentials introduced in Section 2 to see whether we can obtain improvements in their performance. We seek solutions that recognize as many native structures as possible when presented with a population of misfolded structures. We first discuss profile models (THOM1 and THOM2) and then a pairwise interaction model.

To test the recognition capabilities of a particular model of a potential energy function we train its parameters on a set of inequalities, derived from the training set of proteins using gapless threading. We first attempt to solve the whole training problem. If it proves to be infeasible, we perform a number of iterations of the MaxF procedure, starting from a certain initial guess, which is either a statistical potential derived using the `loopp` program or was taken from the literature. The number of inequalities that are still not satisfied at convergence is used as a measure of the difficulty of a given training set and the quality of the model. The performance of the trained potentials is further evaluated on a control set of inequalities, derived from a disjoint set of proteins.

We used three different sets of proteins, developed before to train and test folding potentials. These sets were drawn from the Protein Data Bank PDB [7], which currently contains (as of January 2002) about 17,000 structures, with about 5000 distinct homology modeling targets that form a non-redundant subset of the PDB on the family level. The first set, which will be referred to as the TE set, was developed by Tobi et. al. [29] and includes 594 structures chosen according to the diversity of protein folds but also some homologous proteins (up to 60 percent sequence identity). Therefore it poses a significant challenge to the energy function. The total number of inequalities that were obtained from the TE set using gapless threading was 30,211,442. The second and third sets, referred to as S1082 and S2844, consist of 1082 and 2844 proteins, respectively. These were chosen to be relatively dense and non-redundant subsets of the databank. The S1082 set is used as control, whereas S2844 is used to retrain the pairwise model on a much larger set of proteins. The TE and S1082 sets are disjoint, although the S1082 set includes many structures that are relatively similar to representatives in the TE set [21]. All training and testing sets are available from the web [20].

As demonstrated before, the TE set is infeasible for the THOM1 model (2), and the parameters published in [21] were optimized on a smaller (feasible) set of proteins. Here, we take this potential as the initial solution y_0, and apply the MaxF procedure. The improvement is shown in Table 3, with Iteration 0 corresponding to the initial potential. The results on the training set are included in the left panel, whereas the results on the S1082 set, which is used as a control, are included in the right panel.

There are several ways to evaluate the quality of solutions in this context. Their differences are especially relevant in the infeasible case. One obvious measure of success is the number of violated inequalities at the solution (reported in columns 3 and 6). A more biologically relevant measure is the number of native structures the resulting model fails to recognize (columns 2 and 5). Finally, we also report the z-score as defined in (6). We point out that none of these measures is directly optimized by our procedure.

	—TE training set—			—S1082 control set—		
MaxF iteration	not recog.	viol. ineqs.	z-score	not recog.	viol. ineqs.	z-score
0	120	162,274	1.58	415	539,664	1.42
1	59	1,217	1.87	360	249,854	1.66
2	54	905	1.93	358	250,850	1.66

Table 3. Results using MaxF procedure on a THOM1 potential

The results show a significant qualitative improvement on the training set. The initial solution violates more than 160,000 constraints out of approximately 30 million in the TE problem. After just two iterations (with bulk of the improvement in the first iteration - note that we did not attempt to achieve full convergence) an approximate solution that violates only 905 constraints was obtained. The increase in z-score, from 1.58 to 1.93, indicates also the desired change in the overall shape of the distribution of energy gaps. The performance of the potential on the control set improved significantly as well, although to a smaller extent, still violating approximately 250,000 constraints (out of 95 million included in the control set). As reported previously [21], various folding potentials from the literature reach only a limited accuracy on the S1082 set, which appears to be a demanding test, mostly because of many short proteins included in this set.

The TE set also proved to be infeasible for the THOM2 model with two contact shells if fewer than 300 parameters were used [21]. Here, we consider a simplified THOM2 model with only 180 parameters. The reduction in number of parameters results from a different coarse graining of structural environments. Namely, we group together the second and the third, as well as the fourth and the fifth classes of primary sites that were used before, reducing the number of different types of primary sites from five to three (see [21] for detailed definition of contact types in the THOM2 energy model). Using a statistical potential derived from the TE set as a starting point and applying MaxF we obtain a significantly improved reduced THOM2 potential with just 180 parameters, which approximately solves the TE problem that required as many as 300 parameters to solve exactly.

As evidenced in Table 4, just one iteration of MaxF reduces the number of violated inequalities from approximately 265,000 to 267,000. The increase in z-score from 1.22 to 1.87 is also impressive. A significant improvement in terms of the number of violated inequalities (from 1.6 million to 217,000) and overall

shape of the distribution of energy gaps (z-score increasing from 1.03 to 1.53) is also observed on the control set. However, only minor improvement is observed in terms of the proteins that are not recognized exactly (from 364 to 335). The previously published THOM2 potential with 300 parameters when applied to the S1082 set misses only 205 proteins, although on the other hand, it violates more constraints (240 thousand) with a lower z-score of 1.35.

	—TE training set—			—S1082 control set—		
MaxF iteration	not recog.	viol. ineqs.	z-score	not recog.	viol. ineqs.	z-score
0	102	265,284	1.22	364	1,600,612	1.03
1	34	267	1.87	335	216,623	1.52
2	34	233	1.87	335	216,955	1.53

Table 4. Results using MaxF procedure on a THOM2 potential

Of course, a simultaneous increase in the number of recognized proteins and satisfied inequalities cannot be guaranteed, and in fact discrepancies between these two quality measures have been observed for a number of potentials from the literature on the S1082 set [21]. Since in practice additional filters, such as statistical significance estimates, are applied to a number of low energy matches, the solution with a smaller number of violated constraints may be advantageous.

The next potential we discuss here is a pairwise model (1). The TE set was used by Tobi et. al. [29] for parameter optimization, and the problem proved to be feasible. The solution had to be obtained iteratively by solving subproblems which fit into the memory of a single processor (see the scheme described in Section 3.2, and also Table 1). An additional objective function was used to skew the solution toward maximizing the z-score and thus to improve the quality of the energy gap distribution [29]. We attempted to improve this potential further. First, we solved the TE problem in one shot. The resulting solution does not show improvement over the Tobi et. al. potential (the z-score on the training set was 1.73, compared to 1.75). Second, and in order to further assess the effects of the training set and to sample more extensively from structural variations in protein families, we used the S2844 set for training. To keep the size of the training set manageable we only derived decoys for pairs of sequences and structures that are similar in length (the sequence must be not shorter than 80% of the structure it is aligned to, in order to generate a decoy and a constraint in effect). This results in an infeasible problem of approximately 16 million constraints. The well trained Tobi et. al. potential violates 64,000 inequalities, missing as many as 600 proteins. When applying MaxF (with the Tobi et. al. potential as starting point) only a marginal improvement is observed: 71 additional proteins and approximately 2000 additional inequalities are satisfied after 5 iterations; the z-score of 1.83 remained unchanged.

Although MaxF's failure to improve does not constitute a definitive answer (and may simply have occurred due to the specific structure of the problem at hand), we conjecture that the observed results are an indication of the limits of

the capacity of pairwise models. In light of the above, it is suggestive that the infeasibility reached before on various sets of native and misfolded structures with pairwise models [30] [28] was not due to some rare constraints, but rather due to the low resolution of the model. While non-redundant, the S2844 set includes a number of structural variations of certain folds with a distance of only 3 Ångstrom RMS[1] between the superimposed side chain centers [21]. This threshold of dissimilarity is apparently below the resolution of pairwise folding potentials.

5. Discussion and Conclusions

We described our efforts to provide practical large-scale Linear Programming based tools for the design and evaluation of potential functions that underlie the folding process of proteins. The interplay of biological and physical insights on the one hand and optimization and modeling techniques, large-scale computing and heuristics on the other hand is used to facilitate the design of accurate and efficient potentials for protein folding. The results presented here support the claim that biologically relevant results may be obtained using the new techniques. A systematic application of these techniques is expected to yield a significant improvement in the quality of folding potentials. We also expect to gain new insights to guide the selection of decoys to be included in the training process. The choice of a training set is a critical component of any successful learning procedure that extrapolates from examples and avoids both under- and overfitting of the parameters.

With the present incarnation of `pPCx` we are able to solve problems with a few hundred parameters and tens of millions of constraints in a one-shot approach in a matter of minutes. Development of `pPCx` is ongoing. An extension of the current code will include an implementation of the alternative formulation of the potential modeling problem, defined in Section 3.1. The introduction of slack variables is expected to provide a more satisfactory solution for the infeasible case, while avoiding an increase in the size of the dual problem that we solve. It remains to be seen whether column generation or sampling techniques can be reliably used to speed up computation time. The software itself is application independent. We submit that any problem with similar dimensions (primal dimension several orders of magnitude smaller than dual dimension) and a dense constraint matrix can be efficiently solved using our code. In the future we plan to incorporate the parallel machinery developed for `pPCx` into a Support Vector Machine implementation that would handle large classification problems arising in genomics.

Acknowledgements. The authors gratefully acknowledge support from the Cornell Theory Center in the form of computing time on the Velocity Cluster. We also thank Ron Bryson for carefully reading and editing the manuscript. MW and JM acknowledge support from the Cincinnati Children's Hospital Research Foundation. RE acknowledges the support of an NSF grant on 'Multiscale Hierarchical Analysis of Protein Structure and Dynamics'.

[1] root means square distance

References

1. P. Alpatov, G. Baker, C. Edwards, J. Gunnels, G. Morrow, J. Overfelt, R. van de Geijn, and Y.-J. J. Wu. PLAPACK: parallel linear algebra package. In *Proceedings of the Eighth SIAM Conference on Parallel Processing for Scientific Computing (Minneapolis, MN, 1997)*, page 8 pp. (electronic), Philadelphia, PA, 1997. SIAM.
2. E. Amaldi and R. Hauser. Randomized relaxation methods for the maximum feasible subsystem problem. Technical Report 2001-90, DEI, Politecnico di Milano, 2001.
3. E. Amaldi and V. Kann. The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theor. Comp. Sci.*, 147:181–210, 1995.
4. E. Amaldi, M. E. Pfetsch, and J. Leslie E. Trotter. On the maximum feasible subsystem problem, IISs and IIS-hypergraphs. *to appear in Math. Program.*, 2002.
5. C. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.
6. J. R. Banavar and A. Maritan. Computational approach to the protein-folding problem. *Proteins*, 42:433–435, 2001.
7. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H.Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000. See also `http://www.rcsb.org/pdb/`.
8. J. U. Bowie, R. Luthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170, 1991.
9. N. Chakravarti. Some results concerning post-infeasibility analysis. *Eur. J. Oper. Res.*, 73:139–143, 1994.
10. J. W. Chinneck. Fast heuristics for the maximum feasible subsystem problem. *INFORMS J. Comput.*, 13:210–223, 2001.
11. J. Czyzyk, S. Mehrotra, M. Wagner, and S. J. Wright. PCx: an interior-point code for linear programming. *Optim. Method Softw.*, 11/12(1-4):397–430, 1999.
12. V. A. Eyrich, D. M.Standley, and R. A. Friesner. Ab initio protein structure prediction using a size dependent tertiary folding potential. In R. A. Friesner, editor, *Computational Methods for Protein Folding*, volume 120 of *Advances in Chemical Physics*, pages 223–264. John Wiley & Sons, 2002.
13. M. C. Ferris and T. S. Munson. Interior point methods for massive support vector machines. Technical Report 00-05, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, 2000. To appear in *SIAM J. Optim.*
14. R. A. Friesner and J. R. Gunn. Computational studies of protein folding. *Annu. Rev. Bioph. Biom.*, 25:315–42, 1996.
15. R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes. The statistical mechanical basis of sequence alignment algorithms for protein structure prediction. In R. Elber, editor, *Recent Developments in Theoretical Studies of Proteins*, chapter 6. World Scientific, Singapore, 1996.
16. H. J. Greenberg. Consistency, redundancy, and implied inequalities in linear systems. *Ann. Math. Artif. Intel.*, 17:37–83, 1996.
17. J. L. Klepeis, H. D. Schafroth, K. M. Westerberg, and C. A. Floudas. Ab initio protein structure prediction using a size dependent tertiary folding potential. In R. A. Friesner, editor, *Computational Methods for Protein Folding*, volume 120 of *Advances in Chemical Physics*, pages 265–457. John Wiley & Sons, 2002.
18. R. Lathrop. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.*, 7:1059–1068, 1994.
19. V. Mairov and G. Crippen. Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.*, 227:876–888, 1992.
20. J. Meller and R. Elber. *LOOPP: Learning, Observing and Ouputting Protein Patterns*. Department of Computer Science, Cornell University, available at `http://www.tc.cornell.edu/CBIO/loopp`. A new version of the code and the data sets are available from `http://folding.chmcc.org`.
21. J. Meller and R. Elber. Linear programming optimization and a double statistical filter for protein threading protocols. *Proteins*, 45:241–261, 2001.
22. J. Meller and R. Elber. Protein recognition by sequence-to-structure fitness: Bridging efficiency and capacity of threading models. In R. A. Friesner, editor, *Computational Methods for Protein Folding*, volume 120 of *Advances in Chemical Physics*. John Wiley & Sons, 2002.

23. J. Meller, M. Wagner, and R. Elber. Maximum feasibility guideline in the design and analysis of protein folding potentials. *J. Comp. Chem.*, 23:111–118, 2002.
24. M. Parker and J. Ryan. Finding the miminum weight iis cover of an infeasible system of linear inequalities. *Ann. Math. Art. Intel.*, 17:107–126, 1996.
25. M. E. Pfetsch. *The Maximum Feasible Subsystem Problem and Vertex-Facet Incidences of Polyhedra*. PhD thesis, ZIB, Technische Universität Berlin, 2002.
26. J. Sankaran. A note on resolving infeasibility in linear programs by constraint relaxation. *OR Letters*, 13:19–20, 1993.
27. M. J. Sippl and S. Weitckus. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. *Proteins*, 13:258–271, 1992.
28. D. Tobi and R. Elber. Distance-dependent pair potential for protein folding: Results from linear optimization. *Proteins*, 41:40–46, 2000.
29. D. Tobi, G. Shafran, N. Linial, and R. Elber. On the design and analysis of protein folding potentials. *Proteins*, 40:71–85, 2000.
30. M. Vendruscolo and E. Domany. Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.*, 109:11101–11108, 1998.
31. M. Vendruscolo, L. A. Mirny, E. I. Shakhnovich, and E. Domany. Comparison of two optimization methods to derive energy parameters for protein folding: perceptron and z-score. *Proteins*, 41:192–201, 2000.
32. S. J. Wright. *Primal-dual interior-point methods*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

Maximum Feasibility Approach for Consensus Classifiers: Applications to Protein Structure Prediction

Aleksey Porollo[1], Rafal Adamczak[1], Michael Wagner[1] and Jaroslaw Meller[1,2]

[1]Pediatric Informatics, 3333 Burnet Avenue, Children's Hospital Research Foundation, Cincinnati, OH 45229, USA
{aporollo, radamczak, mwagner}@chmcc.org
[2]Department of Informatics, Nicholas Copernicus University, 87-100 Torun, Poland
jmeller@chmcc.org

Abstract

A novel strategy to optimize consensus classifiers for large classification problems is proposed, based on Linear Programming (LP) techniques and the recently introduced Maximum Feasibility (MaxF) heuristic for solving infeasible LP problems. For a set of classifiers and their normalized class dependent scores one postulates that the consensus score is a linear combination of individual scores. We require this consensus score to satisfy a set of linear constraints, imposing that the consensus score for the true class be higher than for any other classes.. Additional constraints may be added in order to impose that the margin of separation (difference between the true class score and false classes scores) for the consensus classifier be larger than that of the best individual classifier. Since LP problems defined this way are typically infeasible, approximate solutions with good generalization properties are found using interior point methods for LP in conjunction with the MaxF heuristic. The new technique has been applied to a number of classification problems relevant for protein structure prediction.

1. Introduction

Ensemble classifiers are an active area of research in the field of machine learning [1,2]. Many strategies, such as simple voting, linear combination based methods or boosting [3-6], have been proposed to find an improved consensus classifier, given a number of individual classifiers. Consensus classifiers are often able to improve significantly on the classification accuracy. Some important and relevant in bioinformatics examples include applications of neural network based classifiers for protein secondary structure prediction [7] or combining various individual scores into a consensus score for gene prediction [8].

Here, we introduce a novel strategy to optimize consensus classifiers for large problems, using LP techniques and the Maximum Feasibility heuristic for solving infeasible LP problems [9,10]. For a set of classifiers and their normalized class dependent scores one postulates that the consensus score is a linear combination of individual scores. Such defined total score is required to satisfy a set of linear constraints, imposing that the consensus score for the true class is higher than for any other class for each data point in the training.

The resulting LP problems are infeasible for classification problems that are not linearly separable in the feature space of individual classifiers scores. Our strategy to find an approximate solution is to identify a possibly large subset of inequalities that can be satisfied. In other words, we identify a subset of data points that can be classified using linear decision boundaries, with points difficult to classify excluded from the training. Such approximate solutions that achieve high accuracy and have good generalization properties may be found

efficiently using interior point methods for LP in conjunction with the MaxF heuristic.

Here, we briefly revisit the MaxF heuristic and then formally introduce the new approach for finding linear combination based classifiers and discuss strategies for solving the resulting infeasible LP problems (Methods section). The new technique is then applied to a number of classification problems relevant for protein structure prediction, including secondary structure and membrane domains prediction (Results section).

The protein folding problem, which is one of the central challenges in computational biology, consists of predicting the three-dimensional structure of a protein from its amino acid sequence. The methodology and modeling aspects of protein folding have been vastly discussed in the literature [11]. For the sake of completeness it suffice to say here that predicting secondary structures, i.e. locally ordered conformations taking shape of helices or beta strands, greatly facilitates fold recognition and functional annotations. The same concerns membrane domains prediction.

2. Methods

2.1. Maximum Feasibility Heuristic

The Maximum Feasibility (MaxF) [9,10] heuristic aims at finding an approximate solution, which satisfies a possibly large subset of an infeasible set of inequalities. The MaxF procedure is based on a special property of **interior point** algorithms for LP. Without a function to optimize the interior point algorithm places the solution at the "maximally feasible" point, which is away from any individual constraint. For problems with bound feasible polyhedra interior point algorithms converge to the so-called analytic center, when no objective function is used [12]. The idea behind MaxF heuristic is that the "maximally feasible" partial solution is likely to satisfy more constraints than an off-centered guess.

The MaxF heuristic starts from a certain initial guess of the solution and the subset of all the constraints that are satisfied by this initial guess. A series of "maximally feasible" approximations is then computed. The subset of all the inequalities satisfied by the previous approximation, which defines a feasible polyhedron, is solved using an interior point method. The new solution becomes our next "maximally feasible" approximation and satisfies at least as many constraints as the previous partial solution. If no further constraints are satisfied the procedure stops.

The choice of the initial guess of the solution is critical for the success of the MaxF heuristic. Finding the largest feasible subset of an infeasible problem is a NP-hard problem [13] and obtaining a good approximation cannot be guaranteed. However, in practice we observe significant improvement with respect to initial approximate solutions that are carefully chosen using a priori knowledge [9,10].

Another way to obtain an appropriate initial guess is to solve an elastic LP (eLP) problem, with a positive slack variable z_i added to each constraint:

$$\min \sum z_i \text{ subject to } \mathbf{A}\boldsymbol{\alpha} + \mathbf{z} \geq \varepsilon, \quad \varepsilon > 0, \ \mathbf{z} \geq 0 \ . \qquad (1)$$

Here, $\boldsymbol{\alpha}$ denotes the vector of unknowns that are target of optimization and **A** denotes the constraint matrix. The LP problem defined in (1) is always feasible and, by adding the sum of slack variables as the objective function, allows one to find approximate solutions of the original infeasible problem. We applied here the latter strategy.

The eLP finds a solution that effectively minimizes the misclassification error (sum of slacks), and might be influenced by outliers. Nevertheless, we observe in practice that it provides good initial approximations for

the problems considered here. These initial solutions are then improved in terms of margin of separation by subsequent MaxF iterations. Starting from a subset of separable data points, for which the slack variables are equal to zero, MaxF places the separating hyperplanes away from all the data points that are correctly classified by the initial guess.

The pPCx package by one of us (MW), which is a parallel interior point LP solver, was used to obtain results presented in this paper. We would like to comment that interior point methods for LP have superior, polynomial complexity and are very efficient. Problems with millions of constraints and hundreds of variables may be solved, e.g., in a few minutes on a cluster of Xeon CPU's, using the pPCx package [10].

2.2. MaxF based consensus classifiers

Let us consider a supervised classification problem with *N* real vectors from a certain feature space *X*, divided into *K* classes. A discrete set of class labels, conveniently chosen as $1,\ldots,K$, will be referred to as *Y*. A classifier *Q* is then a mapping from *X* to *Y*. For clarity of notation the *k*th class will be alternatively labelled as C_k - $\mathbf{x} \in X$ is classified as belonging to class C_k, if $Q(\mathbf{x}) = k$.

Consider now a number of individual models, M_i, $i = 1,\ldots, p$, that provide estimates for conditional probabilities of class C_k given the model and a vector in the feature space, $P(C_k \mid \mathbf{x}; M_i)$. For each model we define an individual classifier Q_i as:

$$Q_i(\mathbf{x}) = \underset{k=1,\ldots,K}{\arg\max}\ P(C_k \mid \mathbf{x}; M_i)\,. \qquad (2)$$

In other words, a data point **x** is assigned to the class with the highest probability**.** The goal is then to combine the individual models into a mixture (consensus) model.

We define a consensus classifier in the form of a linear combination of individual classifiers:

$$P(C_k \mid \mathbf{x}; M_C) = \sum_{i=1}^{p} \alpha_i P(C_k \mid \mathbf{x}; M_i)\,. \quad (3)$$

Note that the coefficients of the linear combination, which will be a target for optimization, are class independent here (as opposed to more general models with class dependent coefficients – see Results section). Linear decision boundaries for the consensus classifier are defined using again the simple rule:

$$Q_C(\mathbf{x}) = \underset{k=1,\ldots,K}{\arg\max}\ P(C_k \mid \mathbf{x}; M_C)\,. \qquad (4)$$

In supervised classification problem each training vector is assigned to its "true" class, which will also be called its "native" state in the context of applications to protein structure prediction. The true (or native) class will be referred to as C_n, where $Q^*(\mathbf{x}) = n$ is the true classifier (with the implicit dependence of index *n* on **x**).

In order to impose correct consensus predictions in the training, the following inequality constraints (with one inequality per data point) are used:

$$\sum_{i=1}^{p} \alpha_i P_i(C_n) \geq \sum_{i=1}^{p} \sum_{k \neq n} \alpha_i P_i(C_k)\,, \quad (5)$$

where coefficients $P_i(C_k) = P(C_k \mid \mathbf{x}; M_i)$ of the constraint matrix are obtained by applying individual classifiers. Thus, for each data point an inequality as defined in (5) is used to impose that consensus classifier of equation (3) assigns the highest (and larger than 0.5) probability to the true class of that point. A solution to the set of inequalities defined in (5) provides the coefficients α_i, and thus, a linear combination based classifier as defined in (3).

If the problem is feasible, i.e. when the data is linearly separable, the set of inequalities in (5) may be solved efficiently using LP techniques. Typically however, the problem is infeasible and heuristic approaches, such as combination of the elastic LP and

MaxF need to be applied. MaxF was shown before, in the context of protein structure recognition, to effectively filter out outliers that make impossible to separate exactly data points belonging to different classes [9].

The basic idea here is similar. By finding an approximate solution to an infeasible problem defined in (4) we identify points that are difficult to classify. Subsequent iterations of MaxF include only those data points that can be classified correctly (i.e. points that result in inequalities that are satisfied by current guess of the solution). Thus, the linear decision boundaries are optimized for a subset of data points that are separable. In addition, due to the "central" properties of interior point methods, discussed in the Introduction, the solutions that we obtain are away from any individual constraint, providing (at least in principle) a wide margin of separation and a good generalization.

Formulating the problem in terms of linear optimization with constraints opens a way for flexible generalizations. For example, one may impose that the margin of separation between the true and other classes should be at least as wide for the consensus classifier as for the individual classifier, which achieves best separation for a given point. This can be achieved by imposing (again for each vector in the training) additional inequalities of the following form:

$$P_{\mathrm{c}}(C_n) - P_{\mathrm{c}}(C_k) \geq \max_{i=1,\,p} [P_i(C_n) - P_i(C_k)]. \qquad (6)$$

Moreover, instead of considering positive and normalized conditional probabilities one may introduce a generalized classification problem in terms of real scores. One may also weaken the condition of equation (5) by decoupling inequalities for classes other than native. Replacing conditional probabilities for the *i*-th model by the corresponding score, S_i, and introducing one inequality for each non-native state we obtain the following set of inequalities:

$$\sum_{i=1}^{p} \alpha_i S_i(C_n, \mathbf{x}) \geq \sum_{i=1}^{p} \alpha_i S_i(C_k, \mathbf{x}) \quad \forall k \neq n \ \forall \mathbf{x}. \qquad (7)$$

The decision is made as previously: the class with the highest score is assigned to each data point.

3. Results

Preliminary results obtained using the new eLP/MaxF-based approach for protein membrane domain and secondary structure prediction are summarized in Table 1 and Tables 2 and 3, respectively. A set of inequalities defined in equation (7) is solved for each problem using the approach defined in section 2.1. The results are compared to that of several machine learning techniques, including decision trees (SSV [14] and C4.5), k-Nearest Neighbors, adaptable radial basis functions Neural Networks (FSM) [14], Support Vector Machines (SVMs) [15] and Linear Discriminat Analysis (LDA) [16].

Method	Training	Control	Software
Majority	72.1%	67.0%	-
kNN k=10	86.8%	71.8%	Tooldiag
SSV D. tree	85.4%	70.5%	GhostMiner
FSM	85.1%	71.1%	GhostMiner
SVM	86.7%	74.0%	SVMLight
LDA	83.8% (CV)	74.0%	Tooldiag
eLP/MaxF	86.7 (86.1)%	73.1 (72.8)%	pPCx

Table 1. Accuracy for membrane domain prediction.

For membrane domain prediction we used as the training set a curated set of 68 proteins that contained membrane domains and an additional set of 25 proteins as control. Out of the total number of 19,404 residues in the training, 7,704 were in membrane domains. The goal of the prediction is to assign to each amino acid residue one of the two states: membrane or non-membrane. We used as individual weak classifiers (or rather features in this case) twelve statistical scores, each of them assigning a score to a different type of profile (e.g. triplet of residues around the central residue) according to

observed frequency of this profile in a given class in the training set. These individual scores have low prediction accuracy (worse than the baseline). Nevertheless, as can be seen from Table 1, linear discrimination methods (linear SVM, LDA and eLP/MaxF) perform relatively well. Despite the fact that finding a large feasible subset could be potentially hindered by the low quality of individual "classifiers" (features), the LP based approach finds a solution close to that of LDA in terms of accuracy (73.1% when using 24 class dependent coefficients of linear expansion (7) and 72.8 with only 12 class independent coefficients). By combining predictions for adjacent residue the accuracy may be further elevated by about 10%, making this kind of simple predictor an attractive component of a more accurate membrane domain prediction system.

Method	Training:Pfam	Control:S174	Software
Majority	68.3%	67.5%	-
kNN k=10	71.8% (CV)	69.5%	Tooldiag
C4.5 D. tree	95.3%	64.3%	C4.5
LDA	73.5% (CV)	71.0%	Tooldiag
eLP/MaxF	78.8 (73.2)%	70.3 (69.7)%	pPCx

Table 2. Accuracy for coil vs. non-coil prediction.

The second problem that we consider is considerably larger. The training was derived from the Protein Families (Pfam) database and consists of 174,792 residues, which are divided into two classes: coil (no regular secondary structures) and non-coil (helices, beta strands). The feature space consists of 22 different statistical profiles, derived similarly to those for membrane proteins. Despite the still rather moderate size of the problem, we were unable to use either SVMLight or GhostMiner. Again, despite the fact that individual scores have low predictive power, their eLP/MaxF-optimized linear combination achieves accuracy close to that of LDA on the control set of 174 proteins with no homology to proteins in the training.

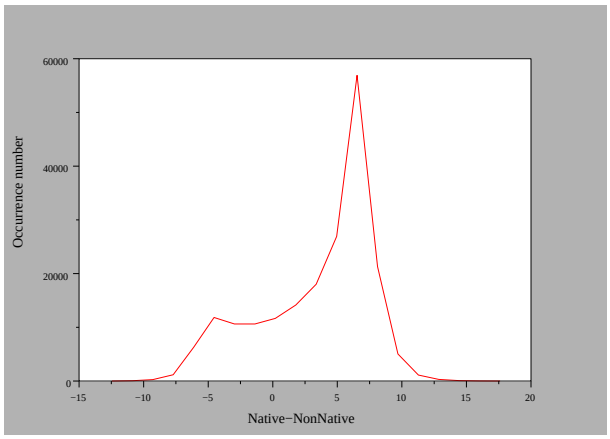


Figure 1. Distribution of differences between consensus scores for native and highest scoring non-native states.

The third problem deals with a consensus of 19 well-trained, NN-based classifiers for the three state (coil, helix, beta strand) secondary structure prediction. These individual predictors achieve accuracy between 71 and 74% in terms of the Q3 measure (three-state per residue accuracy), as opposed to about 78% for the state of the art PsiPRED method, which is itself a consensus of several classifiers [7]. Figure 1 shows the distribution of margins of separation between the native and the highest scoring non-native state for the eLP/MaxF consensus classifier obtained by solving a set of inequalities defined by equations (5) and (6). The use of constraints defined in (6) helps to provide solutions with wide separation margins. Indeed, most of the correctly predicted points (i.e. those with positive margin) are away from the decision boundary with a median separation of about 7. Therefore, by combining the eLP/MaxF consensus with a weighted majority voting for points with a small margin between the two highest scoring classes, we were able to obtain highly accurate predictions (that became part of our SABLE system: http://sable.cchmc.org), as shown in Table 3.

Control sets :	CASP	S174	S189
PsiPRED	80.4%	79.4%	78.7%
eLP/MaxF	81.0%	77.5%	78.8%

Table 3. Accuracy of the secondary structure prediction system obtained using LP-based consensus.

4. Conclusions

A new approach to optimize linear combination based classifiers is introduced. The Maximum Feasibility heuristic for finding approximate solutions to infeasible LP problems is applied to eliminate points that are difficult to classify from the training and to obtain a separating hyperplane for a feasible subset of the data. This approach can be applied to large classification problems with millions of data points and hundreds of variables. In particular, it may advantageous for optimizing consensus classifiers that are postulated as a linear combination of well-trained individual classifiers, while preserving the margin of separation for best classifier in a given region of the feature space. Using this novel strategy we were able to obtain highly accurate consensus classifiers for secondary structure predictions.

In light of the above, the proposed method appears to provide a general and flexible approach to large-scale, multiclass supervised classification problem. Compared to linear perceptron approach, which also produces separating hyperplanes but does not converge for infeasible problems, the present algorithm will efficiently find an approximate solution. Other linear discriminant methods, such as linear regression or LDA focus on centroids of the classes. MaxF based classifiers, similarly to SVM, focus on points close to decision boundaries. Contrary to SVM, though, points that are difficult to classify are first removed from the training. It is worth noting, however, that our strategy is consistent with attempts to achieve a better accuracy by using SVM iteratively, with separating hyperplanes computed for subsets of data points that may result in more robust decision boundaries [15,17].

It is also worth noticing that the standard formulation of the SVM algorithm involves solving a Quadratic Programming (QP) problem [17], which is numerically more expensive than LP. Moreover, multiclass generalizations of SVM are cumbersome [1,17] and the present approach may be an efficient alternative as long as linear discrimination is sufficient. While we present only few examples in the present work, we would expect that linear separation is sufficient in most cases when considering a consensus of well-trained individual classifiers.

References

[1] T. Hastie, R. Tibshirani and J. Friedman; *The Elements of Statistical Learning*, Springer, New York 2001

[2] A. Krogh and J. Vedelsby; *Neural Network Ensembles, Cross Validation and Active Learning*, Advances in Neural Information Processing Systems, MIT Press, 7: 231-238 (1995)

[3] L. Breiman; Bagging predictors, Machine Learning 24: 123-140 (1996)

[4] Y. Freund and R. E. Schapire; Experiments with a new boosting algorithm, in L. Saitta, ed., Machine Learning: Proceedings of the Thirteenth National Conference, Morgan Kaufman, pp. 148-156 (1996)

[5] F. Bauer and R. Kohavi; *An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants*, Machine Learning 36: 105-139 (1999)

[6] B. Mulgrew and C. F. N. Cowan; *Adaptive Filters and Equalisers*, Kluwer Academic Publ., Boston 1988

[7]] D. T. Jones; *Protein secondary structure prediction based on position-specific scoring matrices*, J. Mol. Biol. 292, 195-202 (1999)

[8] S. Salzberg, A. Delcher, K. Fasman and J. Henderson; *A Decision Tree System for Finding Genes in DNA*, J. Comp. Biol. 5: 667-680 (1998)

[9] J. Meller, M. Wagner, R. Elber; *Maximum Feasibility Guideline to the Design and Analysis of Protein Folding Potentials*, Journal of Computational Chemistry, **23**: 111-118 (2002).

[10] M. Wagner, J. Meller, R. Elber; *Large-Scale Linear Programming Techniques for the Design of Protein Folding Potentials*, Mathematical Programming, to appear (2003)

[11] R. A. Friesner and J. R. Gunn; *Computational Studies of Protein Folding*, Annu. Rev. Bioph. Biom. 25: 315-342 (1996)

[12] R. D. C. Monteiro and I. Adler; *Interior path following primal-dual algorithms: Convex quadratic programming*, Math. Program. 44: 43-66 (1989)

[13] N. Chakravarti; *Some results concerning post-infeasibility analysis*, Eur. J. Oper. Res. 73: 139-143 (1994)

[14] GhostMiner; W. Duch, R. Adamczak and K. Grabczewski; *A new methodology of extraction, optimization and application of crisp and fuzzy logical rules*, IEEE Transactions on Neural Networks, Vol. 11 (2): 277-306 (2001)

[15] T. Joachims; *Making large-Scale SVM Learning Practical*, in Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press 1999

[16] T.W Rauber, M.M. Barata and A.S. Steiger-Garcao; *A Toolbox for Analysis and Visualization of Sensor Data in Supervision*, Proceedings of the International Conference on Fault Diagnosis, Toulouse, Toulouse, France, 1993

[17] N. Cristianini and J. Shawe-Taylor; *An Introduction to Support Vector Machines*, Cambridge University Press 2002

PART III:

Applications

54. M. Steiniger-White and W. S. Reznikoff, *J. Biol. Chem.*, in press.
55. T. W. Wiegand and W. S. Reznikoff, *J. Bacteriol.* **174**, 1229 (1992).
56. D. York and W. S. Reznikoff, *Nucleic Acids Res.* **24**, 3790 (1996).
57. R. R. Isberg and M. Syvanen, *J. Biol. Chem.* **260**, 3645 (1985).
58. Z. Otwinowski and W. Minor, *Methods Enzymol.* **276**, 307 (1997).
59. A. T. Brunger *et al.*, *Acta Crystallogr. D* **54**, 905 (1998).
60. A. Roussel and C. Cambilau, in *Silicon Graphics Geometry Partners Directory* (Silicon Graphics, Mountain View, CA, 1991), vol. 86.
61. N. S. Pannu and R. J. Read, *Acta Crystallogr. A* **52**, 659 (1996).
62. P. D. Adams, N. S. Pannu, R. J. Read, A. T. Brünger, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 5018 (1997).
63. M. Carson, *Methods Enzymol.* **277**, 493 (1997).
64. R. J. Read, *Acta Crystallogr. A* **42**, 140 (1986).
65. G. H. Cohen, *J. Mol. Biol.* **190**, 593 (1986).
66. _____, *J. Appl. Crystallogr.* **30**, 1160 (1997).
67. We thank L. Mahnke for providing the first samples of purified transposase, J. B. Thoden for help in collecting the x-ray data, and T. Naumann, A. Bhasin, M. Steiniger-White and R. Saecker for helpful discussions. We gratefully acknowledge the help of N. Duke, F. Rotella, and A. Joachimak at the Structural Biology Center Beamline, Argonne National Laboratory, in collecting the data. This research was supported in part by NIH grants AR35186 (I.R.) and GM50692 (W.S.R.). W.S.R. is a recipient of a Vilas Associates Award and is the Evelyn Mercer Professor of Biochemistry and Molecular Biology. D.R.D. was supported by the NIH Biotechnology Training Grant. The use of the Advanced Photon Source was supported by the U.S. Department of Energy, Basic Energy Sciences, Office of Energy Research, Contract W-31-109-Eng-38. The PDB accession number for the coordinates and structure factors is 1F3I for the transposase/DNA complex.

26 April 2000; accepted 5 June 2000

fw2.2: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size

Anne Frary,[1]* T. Clint Nesbitt,[1]* Amy Frary,[1]† Silvana Grandillo,[1]‡ Esther van der Knaap,[1] Bin Cong,[1] Jiping Liu,[1] Jaroslaw Meller,[2] Ron Elber,[2] Kevin B. Alpert,[1] Steven D. Tanksley[1]§

Domestication of many plants has correlated with dramatic increases in fruit size. In tomato, one quantitative trait locus (QTL), *fw2.2*, was responsible for a large step in this process. When transformed into large-fruited cultivars, a cosmid derived from the *fw2.2* region of a small-fruited wild species reduced fruit size by the predicted amount and had the gene action expected for *fw2.2*. The cause of the QTL effect is a single gene, *ORFX*, that is expressed early in floral development, controls carpel cell number, and has a sequence suggesting structural similarity to the human oncogene c-H-*ras* p21. Alterations in fruit size, imparted by *fw2.2* alleles, are most likely due to changes in regulation rather than in the sequence and structure of the encoded protein.

In natural populations, most phenotypic variation is continuous and is effected by alleles at multiple loci. Although this quantitative variation fuels evolutionary change and has been exploited in the domestication and genetic improvement of plants and animals, the identification and isolation of the genes underlying this variation have been difficult.

Conspicuous and important quantitative traits in plant agriculture are associated with domestication (*1*). Dramatic, relatively rapid evolution of fruit size has accompanied the domestication of virtually all fruit-bearing crop species (*2*). For example, the progenitor of the domesticated tomato (*Lycopersicon esculentum*) most likely had fruit less than 1 cm in diameter and only a few grams in weight (*3*). Such fruit was large enough to contain hundreds of seeds and yet small enough to be dispersed by small rodents or birds. In contrast, modern tomatoes can weigh as much as 1000 grams and can exceed 15 cm in diameter (Fig. 1A). Tomato fruit size is quantitatively controlled [for example, (*4*)]; however, the molecular basis of this transition has been unknown.

Most of the loci involved in the evolution and domestication of tomato from small berries to large fruit have been genetically mapped (*5*, *6*). One of these QTLs, *fw2.2*, changes fruit weight by up to 30% and appears to have been responsible for a key transition during domestication: All wild *Lycopersicon* species examined thus far contain small-fruit alleles at this locus, whereas modern cultivars have large-fruit alleles (*7*). By applying a map-based approach, we have cloned and sequenced a 19-kb segment of DNA containing this QTL and have identified the gene responsible for the QTL effect.

Genetic complementation with *fw2.2*. A yeast artificial chromosome (YAC) containing *fw2.2* was isolated (*8*) and used to screen a cDNA library (constructed from the small-fruited genotype, *L. pennellii* LA716). About 100 positive cDNA clones were identified that represent four unique transcripts (cDNA27, cDNA38, cDNA44, and cDNA70) that were derived from genes in the *fw2.2* YAC contig. A high-resolution map was created of the four transcripts on 3472 F_2 individuals derived from a cross between two nearly isogenic lines (NILs) differing for alleles at *fw2.2* (Fig. 2A) (*8*). The four cDNAs were then used to screen a cosmid library of *L. pennellii* genomic DNA (*9*). Four positive, nonoverlapping cosmids (cos50, cos62, cos69, and cos84) were identified, one corresponding to each unique transcript. These four cosmid clones were assembled into a physical contig of the *fw2.2* region (*10*) (Fig. 2B) and were used for genetic complementation analysis in transgenic plants.

The constructs (*11*) were transformed into two tomato cultivars, Mogeor (fresh market–type) and TA496 (processing-type) (*12*). Both tomato lines carry the partially recessive large-fruit allele of *fw2.2*. Because *fw2.2* is a QTL and the *L. pennellii* allele is only partially dominant, the primary transformants (R0), which are hemizygous for the transgene, were self-pollinated to obtain segregating R1 progeny. In plants containing the transgene (*13*), a statistically significant reduction in fruit weight indicated that the plants were carrying the small-fruit allele of *fw2.2* and that complementation had been achieved. This result was only observed in the R1 progeny of primary transformants fw71 and fw107, both of which carried cos50 (Fig. 1B and Table 1) (*14*). That the two complementing transformation events are independent and in different tomato lines (TA496 and Mogeor) indicates that the cos50 transgene functions similarly in different genetic backgrounds and genomic locations. Thus, the progeny of plants fw71 and fw107 show that *fw2.2* is contained within cos50.

Most QTL alleles are not fully dominant or recessive (*5*). The small-fruit *L. pennellii* allele for *fw2.2* is semidominant to the large-fruit *L. esculentum* allele (*7*). R2 progeny of fw71 were used to calculate the gene action [d/a = dominance deviation/additivity; calculated as described in (*5*)] of cos50 in the transgenic plants.

[1]Department of Plant Breeding and Department of Plant Biology, 252 Emerson Hall, Cornell University, Ithaca, NY 14853, USA. [2]Department of Computer Science, Cornell University, Ithaca, NY 14853, USA.

*These authors contributed equally to this work.
†Present address: Department of Biological Sciences, Clapp Laboratory, Mount Holyoke College, South Hadley, MA 01075, USA.
‡Present address: Research Institute for Vegetable and Ornamental Plant Breeding, IMOF-CNR, Via Universita 133, 80055 Portici, Italy.
§To whom correspondence should be addressed.

The transgene had a *d*/*a* of 0.51; in previous work with nearly isogenic lines (NILs), *fw2.2* had a *d*/*a* of 0.44. This similarity of gene action is consistent with the conclusion that the cos50 transgene carries *fw2.2*.

***fw2.2* corresponds to *ORFX* and is expressed in pre-anthesis floral organs.** Sequence analysis of cos50 (*15*) revealed two open reading frames (ORFs) (Fig. 2C): one corresponding to cDNA44, which was used to isolate cos50, and another 663-nucleotide (nt) gene, *ORFX*, for which no corresponding transcript was detected in the initial cDNA library screen. The insert also contains a highly repetitive, AT-rich (80%) region of 1.4 kb (Fig. 2C). Previous mapping of *fw2.2* had identified a single recombination event that delimited the "rightmost" end of the *fw2.2* candidate region [XO33 in (*8*)]. Comparison of genomic DNA sequence from this recombinant plant with that of the two parental lines indicated that XO33 is within 43 to 80 nt 5′ from the end of *ORFX* (Fig. 2C). Because genetic mutation(s) causing change in fruit size must be to the left of XO33, cDNA44 cannot be involved, and *ORFX* or an upstream region is the likely cause of the *fw2.2* QTL phenotype.

ORFX is transcribed at levels too low to be detected through standard Northern hybridization protocols in all pre-anthesis floral organs (petal, carpels, sepals, and stamen) of both large- and small-fruited NILs; however, semi-quantitative reverse transcriptase–polymerase chain reaction (RT-PCR) analysis indicated that the highest levels were expressed in carpels (*16*) (Fig. 3A). In addition, comparison of the relative levels of *ORFX* transcript in the carpels of the NILs showed significantly higher levels in the small-fruited NIL (TA1144) than in the large-fruited NIL (TA1143) (TA1143/TA1144 carpel transcript levels, mean ratio = 0.51; for the null hypothesis mean = 1, $P = 0.02$). The observation of *ORFX* transcription in pre-anthesis carpels suggests that *fw2.2* exerts its effect early in development. To test this hypothesis, we compared the floral organs from the small- and large-fruited NILs. Carpels (which ultimately develop into fruit), styles, and sepals of the large-fruited NIL were already significantly heavier at anthesis ($P = 0.0007$, 0.001, and 0.001, respectively) than their counterparts in the small-fruited NIL. Stamen and petals

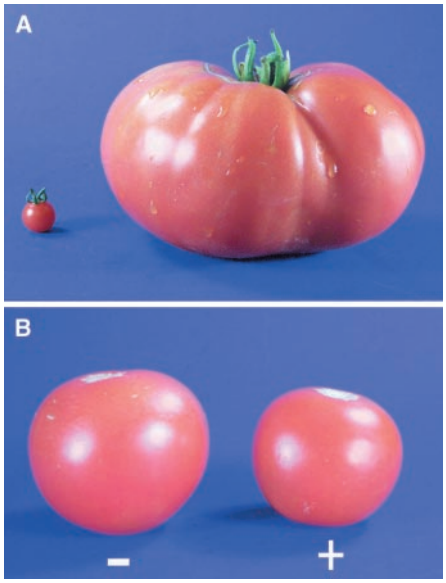


Fig. 1. (**A**) Fruit size extremes in the genus *Lycopersicon*. On the left is a fruit from the wild tomato species *L. pimpinellifolium*, which like all other wild tomato species, bears very small fruit. On the right is a fruit from *L. esculentum* cv Giant Red, bred to produce extremely large tomatoes. (**B**) Phenotypic effect of the *fw2.2* transgene in the cultivar Mogeor. Fruit are from R1 progeny of fw107 segregating for the presence (+) or absence (−) of cos50 containing the small-fruit allele.

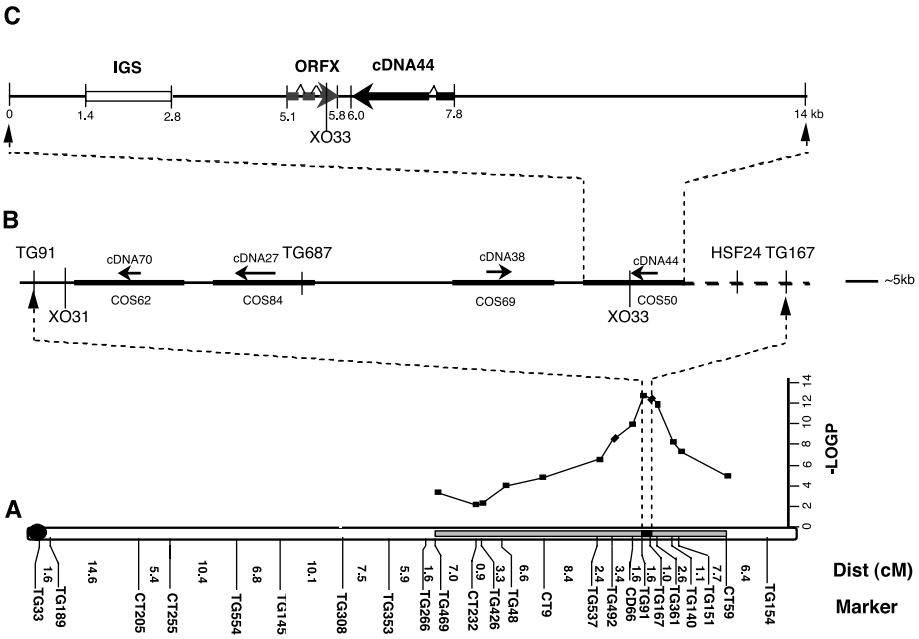


Fig. 2. High-resolution mapping of the *fw2.2* QTL. (**A**) The location of *fw2.2* on tomato chromosome 2 in a cross between *L. esculentum* and a NIL containing a small introgression (gray area) from *L. pennellii* [from (*8*)]. (**B**) Contig of the *fw2.2* candidate region, delimited by recombination events at XO31 and XO33 [from (*8*)]. Arrows represent the four original candidate cDNAs (70, 27, 38, and 44), and heavy horizontal bars are the four cosmids (cos62, 84, 69, and 50) isolated with these cDNAs as probes. The vertical lines are positions of restriction fragment length polymorphism or cleaved amplified polymorphism (CAPs) markers. (**C**) Sequence analysis of cos50, including the positions of cDNA44, ORFX, the A-T–rich repeat region, and the "rightmost" recombination event, XO33.

Table 1. Average fruit weights and seed numbers (*23*) for R1 progeny of several primary transformants. Unless otherwise noted, progeny are from independent R0 plants. Numbers in parentheses are the numbers of R1 individuals tested.

Cosmid	Cultivar	R0 plant no.	Average fruit weight (g)		*P* value	Average seed number		*P* value
			+Transgene	−Transgene		+Transgene	−Transgene	
50*	TA496	fw71	41.6 (18)	56.4 (7)	<0.0001	32.6 (18)	28.3 (7)	0.40
50*	TA496	fw71	47.7 (23)	68.1 (12)	<0.0001	31.4 (23)	27.4 (12)	0.44
50	Mogeor	fw107	25.4 (21)	40.9 (7)	<0.0001	24.1 (21)	28.2 (7)	0.34
62	Mogeor	fw59	46.5 (18)	48.0 (9)	0.70	36.1 (18)	36.5 (9)	0.94
62	TA496	fw70	51.0 (21)	51.3 (3)	0.94	28.3 (21)	39.8 (3)	0.04
69	Mogeor	fw51	50.0 (14)	51.7 (10)	0.58	29.8 (14)	34.8 (10)	0.15
84	Mogeor	fw95	49.4 (18)	47.9 (5)	0.71	33.0 (18)	35.5 (5)	0.62

*R1 progeny of the same primary transformant.

showed no significant difference ($P = 0.63$ and 0.74, respectively). Cell sizes at anthesis were similar ($P = 0.98$ and $P = 0.85$) in the NILs (Fig. 3, B to E); hence, carpels of large-fruited genotypes contain more cells. Therefore, we conclude that allelic variation at *ORFX* modulates fruit size at least in part by controlling carpel cell number before anthesis.

ORFX has homologs in other plant species and predicted structural similarity to human oncogene RAS protein. Sequence analysis of *ORFX* (*17*) revealed that it contains two introns and encodes a 163–amino acid polypeptide of ~22 kD (Fig. 4). Comparison of the predicted amino acid sequence of the *ORFX* cDNA against sequences in the GenBank expressed sequence tag (EST) database found matches only with plant genes. Matches (up to 70% similarity) were found with ESTs in both monocotyledonous and dicotyledonous species. In addition, a weaker match (56.7% similarity) was found with a gymnosperm, *Pinus* (Fig. 4). In tomato, at least four additional paralogs of *ORFX* were identified in the EST database. Although only one *Arabidopsis* EST is represented in the database, eight additional homologs of *ORFX* appear in *Arabidopsis* genomic sequence, often in two or three-gene clusters and having intron-exon arrangements similar to those of *ORFX*. None of the putative homologs of *ORFX* has a known function. Thus, *ORFX* appears to represent a previously uncharacterized plant-specific multigene family.

Analysis of the predicted amino acid sequence of ORFX indicates that it is a soluble protein with alpha/beta-type secondary structure. The threading program LOOPP (*18*) assigns ORFX to the fold of 6q21, domain A, which is human oncogene RAS protein. The *Z* scores for global and local alignments of ORFX are high (3.2 and 4, respectively). Such scores were never observed in false positives and suggest an overall shape similar to that of heterotrimeric guanosine triphosphate–binding proteins. The detailed comparison of ORFX sequence with that of the RAX (where X can be S, N, or D) family reveals conserved fingerprints at RAX-binding domains (*19*). The RAX family includes proteins with wide regulatory functions, including control of cell division (*20*).

The basis for allelic differences at *fw2.2*. In order to understand the basis for allelic differences at *fw2.2*, we compared the *L. pennellii* and *L. esculentum ORFX* alleles by amplifying and sequencing an 830-nt fragment containing *ORFX* [including 55 nt from the 3′ untranslated region (UTR) and 95 nt from the 5′ UTR] from both NILs (Fig. 4). Of the 42 nt differences between the two alleles, 35 fell within the two predicted introns, 4 represent silent mutations, and only 3 cause amino acid changes. All three of the substitutions occurred within the first nine residues of the ORF (asterisks in Fig. 4). Although the start methionine cannot be determined with certainty, if the second methionine in the ORF (M12 in Fig. 4) were used, this would place all three potential substitutions in the 5′ UTR. Conservation between the alleles

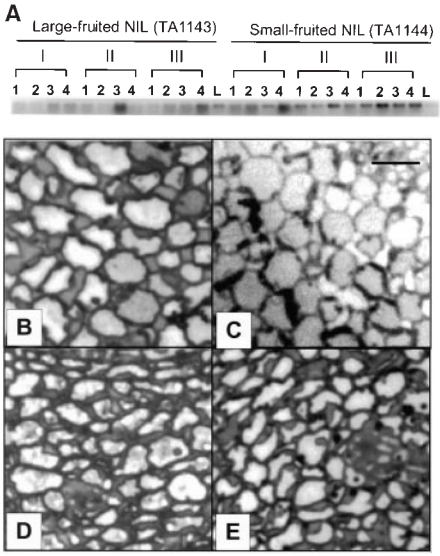


Fig. 3. Reverse transcriptase and histological analyses of the large- and small-fruited NILs (TA1143 and TA1144, respectively). (**A**) RT-PCR detection of *ORFX* transcript in floral organs. Gel showing RT-PCR products for *ORFX* in various stages and organs. Stage I, 3- to 5-mm floral buds; Stage II, 5 mm to anthesis; Stage III, anthesis; lane 1, sepals; lane 2, petals; lane 3, stamen; lane 4, carpels; L, leaves. (**B** to **E**) Transverse thick sections (1 μM) of tomato carpels at anthesis. Top sections (B and C) display cortical cells from carpel septum. Bottom sections (D and E) display pericarp cells from carpel walls. Sections on the left (B and D) are derived from carpels of NIL homozygous for large-fruit allele. Sections on the right (C and E) are derived from carpels of NIL homozygous for small-fruit allele. TA1143 and TA1144 were not significantly different for cell size in either carpel walls (cells per millimeter squared = 17,600 ± 700 versus 17,700 ± 1000; $P = 0.98$) or carpel septa (cells per millimeter squared = 10,100 ± 500 versus 10,300 ± 900; $P = 0.85$) (statistical analysis based on 144 cell area counts from 48 sections). Carpels were fixed in 2.5% glutaraldehyde, 2% paraformaldehyde, and 0.1 M Na cacodylate buffer (pH 6.8) and embedded in Spurr plastic. Bar, 20 μM.

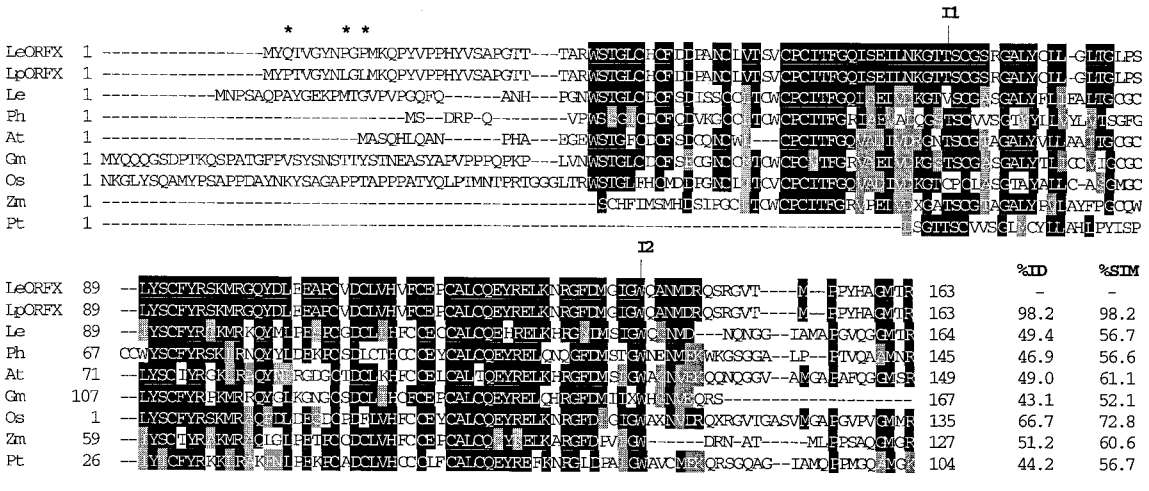


Fig. 4. The results of CLUSTALW alignment of LpORFX (*L. pennellii*, AF261775) and LeORFX (*L. esculentum*, AF261774) with 7 representatives of 26 matched from the GenBank EST and nucleotide databases and the contigs assembled from the TIGR (The Institute for Genomic Research) tomato EST database (*24*). LpORFX and LeORFX residues are shaded black when identical to at least 73% of all the genes included in the analysis. Shading in the other genes represents residues identical (black) or similar (gray) to the black residues in LpORFX, and the dashes are gaps introduced to optimize alignment. Percentages of identical (%ID) or similar (%SIM) amino acid residues over the length of the available sequence are noted (some ESTs may be only partial transcripts). ESTs included in the list are Ph (*Petunia hybrida*, AF049928), Gm (*Glycine max*, AI960277), Os (*Oryza sativa*, AU068795), Zm (*Zea mays*, AI947908), and Pt (*Pinus taeda*, AI725028). The *L. esculentum* EST is contig TC3457 from the TIGR EST database. At represents a predicted protein from *Arabidopsis* genomic sequence (AB015477.1). The positions of the introns in ORFX are indicated as I1 and I2, and the three residue differences between LpORFX and LeORFX are denoted by asterisks. Abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

suggests that the *fw2.2* phenotype is probably not caused by differences within the coding region of *ORFX*, but by one or more changes upstream in the promoter region of *ORFX*. Variation in upstream regulatory regions of the *teosinte branched1* gene has also been implicated in the domestication of maize (*21*). However, differences in fruit size imparted by the different *fw2.2* alleles may be modulated by a combination of sequence changes in the coding and upstream regions of *ORFX* (*22*).

A reduction in cell division in carpels of the small-fruited NIL is correlated with overall higher levels of ORFX transcript, suggesting that ORFX may be a negative regulator of cell division. Whether the ORFX and RAX proteins share common properties other than predicted three-dimensional structure and control of cell division awaits future experimentation. An affirmative result may reflect an ancient and common origin in the processes of cell cycle regulation in plants and animals.

References and Notes

1. J. Doebley, A. Stec, J. Wendel, M. Edwards, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 9888 (1990).
2. J. Smartt and N. W. Simmonds, *Evolution of Crop Plants* (Longman, London, 1995).
3. C. M. Rick, R. W. Zobel, J. F. Fobes, *Proc. Natl. Acad. Sci. U.S.A.* **71**, 835 (1974).
4. A. H. Paterson *et al.*, *Genetics* **127**, 181 (1991).
5. S. D. Tanksley, *Annu. Rev. Genet.* **27**, 205 (1993).
6. S. Grandillo, H. M. Ku, S. D. Tanksley, *Theor. Appl. Genet.* **99**, 978 (1999).
7. K. B. Alpert, S. Grandillo, S. D. Tanksley, *Theor. Appl. Genet.* **91**, 994 (1995).
8. K. B. Alpert and S. D. Tanksley, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 15503 (1996).
9. Details are available at *Science* Online at www.sciencemag.org/feature/data/1050401.shl.
10. The Expand Long Template PCR System (Boehringer Mannheim) was used.
11. Constructs were electroporated into *Agrobacterium tumefaciens* strain ABI-A208 (Monsanto, St. Louis, MO).
12. A. Frary and E. D. Earle, *Plant Cell Rep.* **16**, 235 (1996).
13. The presence of the transgene was assayed by PCR and Southern hybridization analyses.
14. A total of 11 primary transformants were generated for cos50. Although all of these plants carried *nptII*, only two individuals (fw71 and fw107) contained the *L. pennellii* portion of the transferred DNA, as determined by PCR analysis with primers designed from the *L. pennellii* sequence of cos50.
15. Cosmid sequencing is described at *Science* Online at www.sciencemag.org/feature/data/1050401.shl.
16. RT-PCR is described at *Science* Online at www.sciencemag.org/feature/data/1050401.shl.
17. 5′ and 3′ rapid amplification of cDNA ends (RACE) is described at *Science* Online at www.sciencemag.org/feature/data/1050401.shl.
18. The predicted ORFX protein was compared to a training set of 594 structures (chosen from the Protein Data Base to eliminate redundancy) by using the LOOPP algorithm (J. Meller and R. Elber, in preparation). See also www.tc.cornell.edu/reports/NIH/resource/CompBiologyTools/loopp/.
19. The three-dimensional structure of c-H-ras p21 (6q21) is shown at *Science* Online at www.sciencemag.org/feature/data/1050401.shl.
20. Reviewed in S. R. Sprang, *Curr. Opin. Struct. Biol.* **7**, 849 (1997).
21. R Wang, A. Stec, J. Hey, L. Lukens, J. Doebley, *Nature* **398**, 236 (1999).
22. P. C. Phillips, *Trends Genet.* **15**, 6 (1999).

23. Seed number is included in the analysis because reduced fertility, as evidenced by reduced seed per fruit, can decrease fruit size. Thus, these data show that the change in fruit size associated with cos50 is not a byproduct of reduced fertility.
24. The alignment of LpORFX and LeORFX with a total of 26 genes is shown at *Science* Online at www.sciencemag.org/feature/data/1050401.shl.

25. We thank J. Nasrallah, C. Aquadro, J. Doebley, K. Schmid, and W. Swanson for critical review of the manuscript. We also thank C. Lewis and N. van Eck for technical assistance. Supported by grants to S.D.T. from the National Research Initiative Cooperative Grants Program, U.S. Department of Agriculture Plant Genome Program (No. 97-35300-4384); the National Science Foundation (No. DBI-9872617); and the Binational Agricultural Research and Development Fund (No. US 2427-94) and by a grant from the NIH NCRR (National Center for Research Resources) to R.E. for development of LOOPP at Cornell Theory Center. We dedicate this paper to the memory of Dr. Kevin Alpert whose research inspired this work.

14 March 2000; accepted 4 May 2000

Stellar Production Rates of Carbon and Its Abundance in the Universe

H. Oberhummer,[1]* A. Csótó,[2] H. Schlattl[3]

The bulk of the carbon in our universe is produced in the triple-alpha process in helium-burning red giant stars. We calculated the change of the triple-alpha reaction rate in a microscopic 12-nucleon model of the ^{12}C nucleus and looked for the effects of minimal variations of the strengths of the underlying interactions. Stellar model calculations were performed with the alternative reaction rates. Here, we show that outside a narrow window of 0.5 and 4% of the values of the strong and Coulomb forces, respectively, the stellar production of carbon or oxygen is reduced by factors of 30 to 1000.

The formation of ^{12}C through the triple-alpha process takes place in two sequential steps in the He-burning phase of red giants. In the first step, the unstable ^{8}Be with a lifetime of only about 10^{-16} s is formed in a reaction equilibrium with the two alpha particles, $\alpha + \alpha \rightleftharpoons {}^{8}\mathrm{Be}$. In the second step, an additional alpha particle is captured, $^{8}\mathrm{Be}(\alpha,\gamma)^{12}\mathrm{C}$. Without a suitable resonance in ^{12}C, the triple-alpha rate would be much too small to account for the ^{12}C abundance in our universe. Hoyle (*1*) suggested that a resonance level in ^{12}C, at about 300 to 400 keV above the three-alpha threshold, would enhance the triple-alpha reaction rate and would explain the abundance of ^{12}C in our universe. Such a level was subsequently found experimentally when a resonance that possessed the required properties was discovered (*2*, *3*). It is the second 0^+ state in ^{12}C, denoted by 0_2^+. Its modern parameters (*4*) are $\varepsilon = (379.47 \pm 0.18)$ keV, $\Gamma = (8.3 \pm 1)$ eV, and $\Gamma_\gamma = (3.7 \pm 0.5)$ meV, where ε is the resonance energy in the center-of-mass frame relative to the three-alpha threshold, and Γ and Γ_γ are the total width and radiative width, respectively.

The isotope ^{12}C is synthesized further in the He burning in red giants by alpha capture to the O isotope ^{16}O, leading to an abundance ratio in the universe of $^{12}\mathrm{C}:{}^{16}\mathrm{O} \approx 1:2$ (*5*). If the carbon abundance in the universe were suppressed by orders of magnitude, no carbon-based life could have developed in the universe. But the production of O is also necessary because no spontaneous development of carbon-based life is possible without the existence of water.

Here, we investigated the abundance ratios of C and O by starting from slight variations of the strength of the nucleon-nucleon (N-N) interaction with a microscopic 12-nucleon model. In previous studies, only hypothetical ad hoc shifts of the resonance energy of the 0_2^+ state were considered (*6*). Some preliminary results of our calculations are reported elsewhere (*7*).

The resonant reaction rate for the triple-alpha process ($r_{3\alpha}$) proceeding via the ground state of ^{8}Be and the 0_2^+ resonance in ^{12}C is given approximately by (*5*)

[1]Institute of Nuclear Physics, Vienna University of Technology, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria. [2]Department of Atomic Physics, Eötvös University, Pázmány Péter Sétány 1/A, H-1117 Budapest, Hungary. [3]Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, D-85741 Garching, Germany.

*To whom correspondence should be addressed. E-mail: ohu@kph.tuwien.ac.at

von Hippel–Lindau protein binds hyperphosphorylated large subunit of RNA polymerase II through a proline hydroxylation motif and targets it for ubiquitination

Anna V. Kuznetsova*†, Jaroslaw Meller†‡§, Phillip O. Schnell*, James A. Nash*, Monika L. Ignacak*, Yolanda Sanchez¶, Joan W. Conaway‖, Ronald C. Conaway‖**, and Maria F. Czyzyk-Krzeska*††**

*Department of Molecular and Cellular Physiology, ‡Pediatric Informatics, Children's Hospital Research Foundation, ¶Department of Molecular Genetics, University of Cincinnati College of Medicine, Cincinnati, OH 45267-0576; §Department of Informatics, Nicholas Copernicus University, 87-100, Torun, Poland; ‖Stowers Institute for Medical Research, Kansas City, MO 64110; and **Department of Biochemistry and Molecular Biology, University of Kansas Medical Center, Kansas City, KS 66160

Edited by Robert G. Roeder, The Rockefeller University, New York, NY, and approved January 7, 2003 (received for review October 7, 2002)

The transition from transcription initiation to elongation involves phosphorylation of the large subunit (Rpb1) of RNA polymerase II on the repetitive carboxyl-terminal domain. The elongating hyperphosphorylated Rpb1 is subject to ubiquitination, particularly in response to UV radiation and DNA-damaging agents. By using computer modeling, we identified regions of Rpb1 and the adjacent subunit 6 of RNA polymerase II (Rpb6) that share sequence and structural similarity with the domain of hypoxia-inducible transcription factor 1α (HIF-1α) that binds von Hippel–Lindau tumor suppressor protein (pVHL). pVHL confers substrate specificity to the E3 ligase complex, which ubiquitinates HIF-α and targets it for proteasomal degradation. In agreement with the computational model, we show biochemical evidence that pVHL specifically binds the hyperphosphorylated Rpb1 in a proline-hydroxylation-dependent manner, targeting it for ubiquitination. This interaction is regulated by UV radiation.

The von Hippel–Lindau tumor suppressor protein (pVHL)-associated complex, which contains elongin B, elongin C, cullin-2, and Rbx-1 (1–3) is a primary ubiquitin ligase for ubiquitination of the α subunits of the hypoxia-inducible transcription factors (HIFs) (4–6). During normoxia, translated HIF-αs are hydroxylated on conserved proline residues located within L(XY)LAP motifs by the O_2, Fe(II), and 2-oxoglutarate-regulated Egl-9 family of prolyl hydroxylases (7, 8), resulting in their ubiquitination and degradation. During hypoxia, proline hydroxylation is inhibited; HIF-αs are not ubiquitinated, and they accumulate and regulate transcription of the HIF-responsive genes (4–6, 9–12). Loss of pVHL function in VHL disease leads to the accumulation of HIF-αs during normoxic conditions, causing constitutive induction of HIF-responsive genes, including angiogenic vascular endothelial growth factor (VEGF) (13, 14). This functioning, in turn, contributes to the formation of highly vascular tumors such as hemangioblastomas, angiomas, and renal clear cell carcinomas (RCCs) (15).

von Hippel–Lindau disease is also associated with pheochromocytomas, nonmalignant tumors of adrenal medulla chromaffin cells, which synthesize and release large quantities of catecholamines and produce cardiovascular pathologies (16, 17). The molecular mechanism of the augmented catecholamine production is unknown. Recently, we presented evidence that pVHL regulates expression of the rate-limiting enzyme in catecholamine biosynthesis, tyrosine hydroxylase (TH), and in pheochromocytoma-derived (PC12) cells (18, 19). Low levels of pVHL, resulting from expression of *VHL* antisense RNA, correlate with more efficient transcription of the full-length *TH* transcripts (19). In contrast, high levels of overexpressed pVHL block transcript elongation between exons 6 and 8 of the *TH* gene (18). The presence of the elongation arrest site within this region of the *TH* gene has been confirmed by using *in vitro* transcriptional analysis (20).

Processive elongation of the initiated transcripts involves reversible hyperphosphorylation of tandemly repeated heptapeptides on the carboxyl-terminal domain (CTD) of subunit 1 of RNA polymerase II (Rpb1) within the RNA polymerase II complex (21). This elongation-competent, hyperphosphorylated Rpb1 is ubiquitinated in a transcription-dependent manner (22, 23). In particular, ubiquitination of the hyperphosphorylated Rpb1 is induced by UV radiation and DNA damage (24–26), suggesting that Rpb1 ubiquitination may play a role in the transcription-coupled repair (27). In yeast, ubiquitination is mediated by a HECT-class Rsp5 ubiquitin ligase (28); however, the nature of the E3 ligase in mammalian cells is unknown. We hypothesized that the hyperphosphorylated Rpb1 may be a substrate for pVHL-associated E3 ubiquitin-ligase activity.

Here, we identify a region of the Rpb1/Rpb6 subunits of RNA polymerase II that shares sequence and structural similarity with the pVHL binding domain of HIF-1α, and show that the pVHL-associated complex interacts specifically with the hyperphosphorylated Rpb1, leading to its ubiquitination.

Materials and Methods

Cell Cultures and Reagents. PC12 cell clones (18, 19) and 786-O RCC cells were described (1), and were used at the cell density of 1.5–2.5 $\times$ 10^5 per cm^2. UV irradiations (15 J/m^2) were performed in a UV Crosslinker (FB-UVXL-1000, Fisher Biotech, Pittsburgh). *N*-Cbz-L-Leu-L-Leu-L-norvalinal (CbzLLn; 10 μM) was added 30 min before UV irradiation. This medium was removed immediately before the irradiation and the same medium was returned after irradiation.

Iron chloride, cobalt chloride, zinc chloride, desferrioxamine, 2,2′-dipyridyl, ascorbic acid, 2-oxoglutarate, and CbzLLn were purchased from Sigma. Reagents used in ubiquitination reaction were purchased from Boston Biochem (Boston) or Affiniti Research Products (Hamhead, Exeter, Devon, U.K.). The following antibodies were used as follows: H14 (Research Diagnostics, Flanders, NJ); C21 and anti-cullin-2 (Santa Cruz Biotechnology); Ig32 anti-pVHL (PharMingen); 12CA5 anti-hemagglutinin (HA) (Roche Molecular Biochemicals); anti-Rbx1 (Zymed); anti-elongin C (Signal Transduction, Lexington, KY); anti-elongin B polyclonal (custom made by Alpha

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: pVHL, von Hippel–Lindau protein; TH, tyrosine hydroxylase; PC12, phoechromocytoma cell line; CTD, carboxyl-terminal domain; Rpb1 or Rpb6, subunit 1 or 6 of RNA polymerase II; HA, hemagglutinin; CbzLLn, *N*-Cbz-L-Leu-L-Leu-L-norvalinal; HIF, hypoxia-inducible factor; RCC, renal cell carcinoma; ODDD, oxygen-dependent degradation domain.

†A.V.K. and J.M. contributed equally to this work.

††To whom correspondence should be addressed at: Department of Molecular and Cellular Physiology, University of Cincinnati College of Medicine, P.O. Box 670576, Cincinnati, OH 45267-0576. E-mail: Maria.Czyzykkrzeska@uc.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0436037100

Diagnostic, San Antonio, TX); anti-HIF-2α (Novus Biologicals, Littleton, CO); anti-ubiquitin (StressGen Biotechnologies, Victoria, Canada); and mouse anti-rabbit IgG (clone RG-96, Sigma). Anti-mouse secondary antibody-bound agarose was from Sigma. Synthetic biotinylated peptides were made by Alpha Diagnostic.

***In Vitro* pVHL-Peptide Binding Reaction.** Ten micrograms of biotinylated peptide was incubated with streptavidin-coated Dynabeads (M-280, Dynal, Great Neck, NY) in a buffer (25 μl) containing 20 mM Tris at pH 8, 100 mM NaCl, 0.5% Nonidet P-40, and 1 mM EDTA for 1 h at room temperature. Washed beads were incubated with WT [pRC-cytomegalovirus (CMV) expression vector; Invitrogen] or mutated pVHL (pCI-neo-CMV expression vector; Promega), translated *in vitro* by using [^{35}S]methionine and TNT reticulocyte lysate (Promega). Binding reaction products were washed extensively in the same buffer and analyzed for bound [^{35}S]pVHL by using SDS/PAGE. For the peptide hydroxylation step, immobilized peptide was first incubated in the hypotonically prepared cellular extract from PC12 cells as described below in the presence of 100 μM each of $FeCl_2$, ascorbic acid, and 2-oxoglutarate for 1–2 h at 30 or 37°C.

Preparation of Extracts. Intact nuclei were isolated as described (18, 19). The nuclei were resuspended in a half-pellet volume of low-salt buffer (20 mM Hepes, pH 7.9/20 mM NaCl/1 mM EDTA/20% glycerol), to which a half-pellet volume of high-salt buffer (20 mM Hepes, pH 7.9/1 M NaCl/1 mM EDTA/20% glycerol) was added. Proteins were extracted for 30 min at 4°C, followed by digestion of genomic DNA and RNA with DNase and micrococcal nuclease (15 and 88 units, respectively, per 100 μl of nuclear pellet volume) for 60 min, to release DNA-bound RNA polymerase II complexes. The digestion produced DNA fragments of <600 bp, as estimated by ethidium bromide staining on agarose gel. Extracts were centrifuged twice at 21,000 $\times$ *g* for 30 min at 4°C and dialyzed.

Total cellular extracts were prepared by using hypotonic lysis (20 mM Tris, pH 7.5/5 mM KCl/1.5 mM $MgCl_2$/1 mM DTT and standard protease inhibitors) at 4°C and were homogenized by using 40 strokes of a tight-pestle Dounce homogenizer. The lysates were digested with DNase and micrococcal nuclease as described above, and centrifuged twice at 21,000 $\times$ *g*.

Denatured lysates were obtained by boiling pellets in 3 vol of SDS lysis buffer (1% SDS/50 mM Tris, pH 7.5/0.5 mM EDTA/1 mM DTT) for 10 min. The lysates were then diluted with immunoprecipitation buffer (see below) and centrifuged at 21,000 $\times$ *g* for 30 min.

Immunoprecipitations. For all immunoprecipitation reactions, the agarose beads were precoated with BSA, and the primary antibodies were preconjugated with the secondary antibodies. Reactions were performed in buffer containing 50 mM Hepes at pH 7.8, 150 mM NaCl, 5 mM $MgCl_2$, 20% (vol/vol) glycerol, and 0.1% Triton X-100 (immunoprecipitation buffer) and washed in the same buffer containing in addition 0.5% Triton, 0.5% Igepal, and 0.5% sodium deoxycholate, or a buffer containing 0.5% Igepal and NaCl from 150 to 900 mM. Immunoprecipitated proteins were eluted by boiling in SDS sample buffer, resolved by electrophoresis in SDS/4–22% polyacrylamide gradient gels, and detected by immunoblotting.

For hydroxylation of endogenous Rpb1, 150 μg of total protein extract was incubated in the presence of prolyl hydroxylase cofactors or inhibitors for 15–30 min at 30°C. The extracts were then processed for the immunoprecipitation reaction with anti-HA antibody. For elutions of Rpb1 from pVHL-associated complexes, proteins coimmunoprecipitated with anti-HA antibody from PC12VHL(WT) nuclear extracts were incubated in 40 mM Tris buffer in the presence of the respective peptides for 1.5 h. The eluates were analyzed by SDS/PAGE. To dephosphorylate Rpb1, extracts were incubated with 25 units of alkaline phosphatase (Roche Molecular Biochemicals) with or without 10 mM NaF for 30 or 60 min at room temperature. Dephosphorylated extracts were used for immunoprecipitations with anti-HA antibodies.

***In Vitro* Ubiquitination.** Four-hundred micrograms of total cellular extract was immunoprecipitated with anti-HA or H14 antibody, followed by four washes with high detergent immunoprecipitation buffer and two washes in buffer containing 50 mM Tris at pH 8 and 3 mM DTT. The immunoprecipitated complexes were resuspended in a final reaction volume of 50 μl containing ATP-regenerating buffer, 5 $\mu g \cdot \mu l^{-1}$ ubiquitin, 100 $ng \cdot \mu l^{-1}$ ubiquitin aldehyde, CbzLLn, and 16 μl of purifed rabbit reticulocyte fraction II, incubated for 2 h at 37°C, washed in 50 mM Tris, pH 8.0/3 mM DTT buffer, and analyzed by SDS/PAGE.

Computational Analysis. The PROSITE server (29) was used to identify proteins containing proline hydroxylation motifs. HIF-1α secondary structures were predicted by using the PsiPRED (30) server. The sequence of the human Rpb1 subunit (GenBank accession no. NP_000928) was first aligned optimally with the yeast Rpb1 structure (PDB ID code 1I50, chain A, 60% identical with mostly conservative substitutions), and then used to build the optimal structurally biased sequence alignment with the sequences of the human HIF-α factor (GenBank accession no. BAB70608). HIF-1α residues 530–577 were aligned with the yeast Rpb1 (structure 1I50, chain A) by using the LOOPP program and structurally biased sequence alignment (refs. 31–33; LOOPP is available at www.tc.cornell.edu/CBIO/loopp). Residues 571–679 of HIF-1α were aligned with the structure of the human Rpb6 subunit (structure 1qkl, chain A) by using the 3D-PSSM server (34).

Results

Similarity of Rpb1/Rpb6 Subunits and HIF-1α Oxygen-Dependent Degradation Domain (ODDD). The human, murine, and yeast Rpb1 subunits contain an analogous, L(XY)LAP, motif located amino-terminal to the binding site for Rpb6 and to the beginning of the unstructured CTD (Fig. 1; ref. 35). Comparing the sequence of the HIF-1α ODDD with representative libraries of protein structures identified a region with similarities to a fragment of Rpb1 and the adjacent Rpb6 subunit. The $\approx$50-aa Rpb1 counterpart is 30% identical and contains the L(XY)LAP motif, including P1465 as a counterpart of the HIF-1α P564 residue. The HIF-1α secondary structures, as predicted by the PSI-PRED method (30), are also consistent with those of Rpb1 and Rpb6. The plausible pVHL binding pocket between Rpb1 and Rpb6 with the critical pVHL-binding motif (Fig. 1*B*) is located on the surface of the RNA polymerase II complex (see the legend to Fig. 7, which is published as supporting information on the PNAS web site, www.pnas.org). The estimated statistical significance of individual alignments of the HIF-1α sequence into the Rpb1 and Rpb6 structures is low (Fig. 7). However, two weak matches into adjacent structural domains of the RNA polymerase II complex make the overall prediction stronger than suggested by the individual estimates of significance. This prediction is further strengthened by the recently published partial structure of the ODDD peptide and pVHL complex (36, 37), which suggests that ODDD exists in an extended conformation and reveals that the adjacent DLQL motif stabilizes the pVHL binding. A closely related motif (DLLL) is found in the predicted Rpb1 counterpart of ODDD.

pVHL Binds Hyperphosphorylated Rpb1 in a Proline Hydroxylation-Dependent Manner. The immobilized Rpb1 peptide (amino acids 1440–1475) containing the hydroxylated proline binds to WT

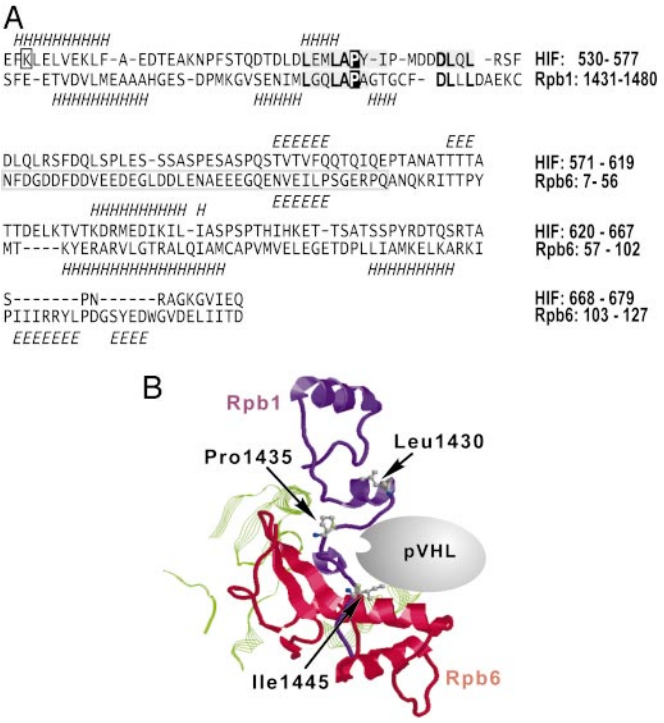


Fig. 1. Computational prediction of the pVHL-binding pocket in the RNA polymerase II complex. (*A*) Sequence-to-structure alignments of the HIF-1α ODDD fragment into the carboxyl-terminal fragments of human Rpb1 and Rpb6 subunits. The Rpb1 and Rpb6 secondary structures are indicated below, and the predicted HIF-1α secondary structures are shown above their sequences. *H*, α (and other)-helices; *E*, extended β-strands. HIF-1α motifs that make contact with the pVHL complex (including the Pro-564 residue) are shaded and, if conserved in Rpb1, bold. The critical HIF-1α residues (L559, L562, P564, and D571) are conserved in the Rpb1 structure. The K532 residue, ubiquitinated on HIF-1α, is boxed. The human and yeast Rpb6 structures (PDB ID codes 1QKL and 1I50, chain F, respectively) are different by an additional β-strand occurring only on the human Rpb6 structure (boxed fragment). (*B*) The predicted pVHL-binding pocket (Rpb1, purple; Rpb6, red; other fragments in contact with the binding pocket are green). The critical proline residue and the flanking amino acids are indicated by using ball and stick models of their side chains. The numbering of residues is according to the yeast Rpb1 structure with the yeast Leu-1430, Pro-1435, and Ile-1445 residues corresponding to Leu-1460, Pro-1465, and Leu-1475 of the human Rpb1, respectively.

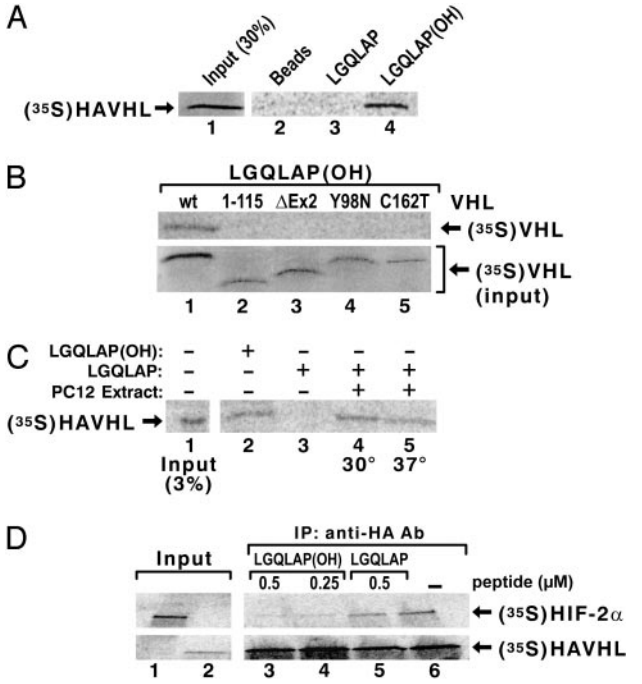


Fig. 2. pVHL binds to the Rpb1 synthetic 36-aa peptide with hydroxylated P1465. (*A*) Binding of [^{35}S]pVHL to the Rpb1 peptide hydroxylated (lane 4), or nonhydroxylated (lane 3), on P1465. (*B*) Binding of the *in vitro*-translated mutated forms of pVHL to the hydroxylated peptide. To ensure that the amounts of the labeled mutant proteins used in the peptide-binding reactions were the same as for the WT pVHL, the amounts of the lysate with radioactively labeled mutant proteins used in the binding reactions were normalized accordingly by using the PhosphorImager quantification. (*C*) Hydroxylation of the Rpb1 peptide in extract from PC12 cells. HAVHL, HA-tagged pVHL. (*D*) Competition experiment of [^{35}S]HIF-2α–[^{35}S]pVHL binding by hydroxylated (lanes 3 and 4) or nonhydroxylated (lane 5) peptide.

[^{35}S]pVHL (Fig. 2*A*), but not to pVHL with deletions of exon 3 or exon 2 or with point mutations within the β (C162T) or α (Y98N) domain (Fig. 2*B*). The nonhydroxylated peptide does not bind pVHL, but it acquires pVHL-binding properties after incubation with PC12 cell extracts in the presence of Fe(II), ascorbic acid, and 2-oxoglutarate (Fig. 2*C*). The hydroxylated, but not the nonhydroxylated, peptide competes with full-length [^{35}S]HIF-2α for [^{35}S]pVHL binding (Fig. 2*D*). [^{35}S]pVHL and Rpb6 do not interact under these conditions.

Coimmunoprecipitation experiments using anti-pVHL antibodies in nuclear extracts from PC12 cells overexpressing HA-tagged human pVHL, or in control vector-transfected PC12 cells (18), reveal that both anti-VHL and anti-HA antibodies are able to coimmunoprecipitate hyperphosphorylated Rpb1 as detected by the H14 antibody, which is specific for phosphoserine-5 within the CTD repeats (refs. 38–40; Fig. 3*A*). In contrast, pVHL fails to coimmunoprecipitate the nonphosphorylated Rpb1 as detected by the C21 antibody (Fig. 3 *A* and *B*), which is specific for the nonphosphorylated peptide sequence from the CTD. The pVHL–Rpb1 complex is stable in high salt (up to 900 mM NaCl washes), consistent with other pVHL-binding proteins (1–3) (Fig. 3*B*). Dephosphorylation of extracts with alkaline phosphatase greatly attenuates binding of pVHL to the hyperphosphorylated Rpb1, and does not induce binding of the hypophosphorylated Rpb1 (Fig. 3*C*). pVHL–Rpb1 complex is also formed in extracts derived from RCC cells, either expressing endogenous truncated and nonfunctional pVHL or stably transfected with HA-tagged pVHL (ref. 1; Fig. 3*D*).

Incubation of cellular lysates in the presence of Fe(II), 2-oxoglutarate, and ascorbic acid substantially increases the formation of the pVHL–Rpb1 complex as determined by coimmunoprecipitation reactions (Fig. 4 *A* and *B*), whereas this complex is inhibited in the presence of iron chelators, desferrioxamine, and 2,2′-dipyridyl (41), or by the addition of $ZnCl_2$, a potent divalent inhibitor of collagen prolyl hydroxylases (ref. 42; Fig. 4*A Right*). These treatments do not affect the total amount of hyperphosphorylated Rpb1 in the extracts (Fig. 4*A Left*). Pretreatment of lysates with cofactors of prolyl hydroxylases augments both the association of pVHL with Rpb1 and the formation of the pVHL–HIF-2α complex (Fig. 4*B*). However, these treatments do not affect the formation of the pVHL–elongin BC, cullin-2, and Rbx-1 complex (Fig. 4*B*), and they do not induce binding of pVHL to the hypophosphorylated Rpb1 (data not shown). The synthetic 36-aa P1465-hydroxylated Rpb1 peptide elutes Rpb1 and HIF-2α from the anti-HA immunoprecipitated complex, whereas the nonhydroxylated peptide is only marginally effective (Fig. 4*C*). These data indicate that, similar to HIF-α, hyperphosphorylated Rpb1 binds pVHL in a proline hydroxylation-dependent manner.

pVHL Regulates Ubiquitination and Accumulation of Hyperphosphorylated Rpb1. The amount of hyperphosphorylated, but not of hypophosphorylated, Rpb1 in PC12 cells correlates inversely

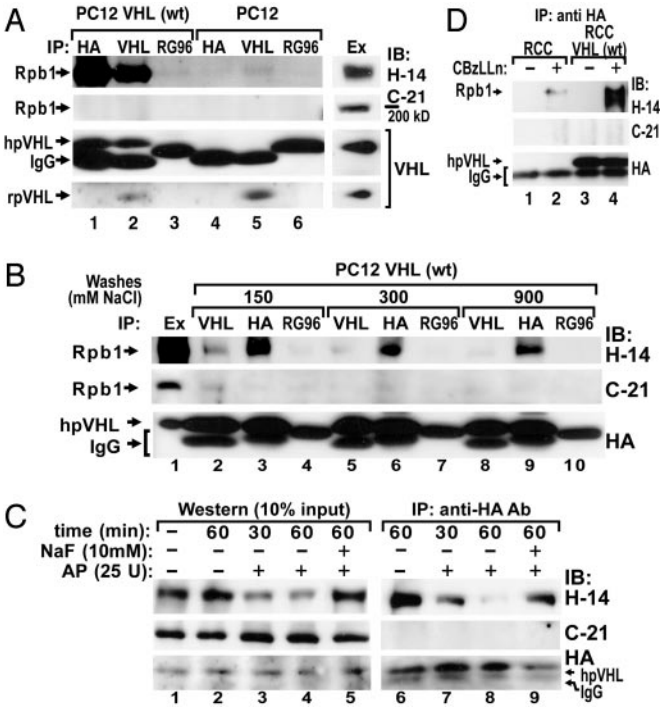


Fig. 3. pVHL specifically interacts with the hyperphosphorylated Rpb1 in nuclear extracts from PC12 and RCC cells. Coimmunoprecipitations (IP) using monoclonal antibodies against HA or pVHL or mouse anti-rabbit IgG (RG-96) in nuclear extracts from PC12 cells overexpressing human HA-tagged pVHL [PC12 VHL (WT)] (*A* and *B*), or control PC12 cells stably transfected with an empty vector (*A*). (*C*) Dephosphorylation of Rpb1 in PC12 cellular extracts for the indicated times by treating the extracts with alkaline phosphatase (AP) in the absence (lanes 3, 4, 7, and 8) or presence (lanes 6 and 9) of NaF. Treated extracts were subjected to immunoprecipitations using anti-HA antibodies. (*D*) Anti-HA immunoprecipitations in nuclear extracts from RCC 786-O cells lacking pVHL function (lanes 1 and 2) or from cells stably transfected with HA-pVHL (1) (lanes 3 and 4). PC12 and indicated RCC cells were pretreated with 10 μM CbzLLn for 6 h to increase accumulation of the hyperphosphorylated Rpb1. The immunoprecipitates were washed with high-detergent immunoprecipitation buffer (*A*, *C*, and *D*), or immunoprecipitation buffers containing up to 900 mM NaCl and 0.5% Igepal (*B*). Blots were probed with the indicated antibodies, human (h)pVHL and rat (r)pVHL, respectively.

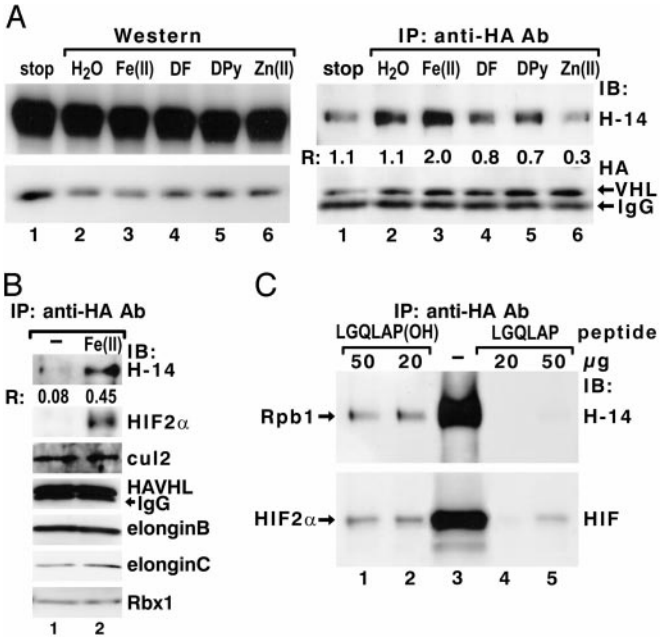


Fig. 4. pVHL binds Rpb1 in a proline hydroxylation-dependent manner. (*A*) Preincubation of PC12 cellular extracts with $FeCl_2$, ascorbic acid, and 2-oxoglutarate (100 μM each) (lane 3), or with 100 μM each of iron chelators: desferrioxamine (DF, lane 4) and 2,2′-dipyridyl (DPy, lane 5), or $ZnCl_2$ (lane 6), followed by Western blot analysis (*Left*) or coimmunoprecipitations (IP) with anti-HA antibodies (*Right*). IB, immunoblotting antibody. (*B*) Coimmunoprecipitation of the components of pVHL-associated complex by using anti-HA antibody in cellular lysates (lane 1) or lysates treated under hydroxylating conditions with Fe(II), ascorbate, and 2-oxoglutarate, as in *A*. Immunoblots were probed with the indicated antibodies. (*C*) Elution of hyperphosphorylated Rpb1 and HIF-2α with a hydroxylated 36-aa Rpb1 peptide. R describes the ratio of the signal detected with H14 antibody to the signal detected with anti-HA antibody, as quantified by using optical density measurements.

with the levels of pVHL (Fig. 5*A*). Cells overexpressing pVHL exhibit low levels of constitutively accumulated hyperphosphorylated Rpb1, whereas cells expressing reduced levels of pVHL (19) exhibit high levels of Rpb1 detected with H14 (Fig. 5 *A* and *C*). After treatment with the proteasomal inhibitor CbzLLn, cells expressing high levels of pVHL accumulate the more slowly migrating forms of Rpb1, whereas cells expressing low levels of pVHL do not. In contrast, steady-state levels of the hypophosphorylated form of Rpb1 are not affected by CbzLLn treatment, and are independent of pVHL levels (Fig. 5*A*). Formation of the pVHL-Rpb1 complex is proportional to the concentration of pVHL and increases in cells treated with CbzLLn (Fig. 5*B*). Consistent with these data, *in vivo* ubiquitination of the hyperphosphorylated Rpb1 correlates with the levels of pVHL in a CbzLLn-dependent manner (Fig. 5*C*).

To further investigate whether the pVHL complex directly ubiquitinates hyperphosphorylated Rpb1, protein complexes were coimmunoprecipitated with either anti-HA or H14 antibody, washed stringently, and subjected to *in vitro* ubiquitination (Fig. 5*D*). The more slowly migrating forms of Rpb1 are detected only with the full ubiquitination-reaction-containing complexes coimmunoprecipitated with anti-HA, but not with H14 (Fig. 5*D*, lanes 2 and 5). The ubiquitinated forms of Rpb1 are not detected if ATP-regenerating buffer or ubiquitin is omitted (Fig. 5*D*, lanes 3 and 4). These data show that pVHL targets hyperphosphorylated Rpb1 for ubiquitination.

UV Irradiation Induces pVHL–Rpb1 Interaction. In cells overexpressing pVHL, UV irradiation induces an early and transient increase in the accumulation of hyperphosphorylated Rpb1 detected with H14, which decreases during 8 h of recovery (Fig. 6*A*). In contrast, in cells with reduced levels of pVHL, hyperphosphorylated Rpb1 does not increase after UV irradiation, but declines with a delay beginning after 6 h of recovery (Fig. 6*A*). The disappearance of Rpb1 in pVHL-overexpressing cells is prevented by proteasomal inhibitors, indicating that the loss of Rpb1 results from proteasomal degradation (Fig. 6*A*). Levels of the hypophosphorylated Rpb1 are not affected by UV exposure and do not depend on the concentration of pVHL. UV irradiation increases the amount of Rpb1 coimmunoprecipitated with pVHL, in the presence and absence of proteasomal inhibitors (Fig. 6*B*). The UV stimulus clearly increases ubiquitination of hyperphosphorylated Rpb1 in cells overexpressing pVHL, but fails to induce its ubiquitination in cells expressing low levels of pVHL (Fig. 6*C*). These data indicate that pVHL contributes to the processing of the RNA polymerase complex in response to UV stress.

Discussion

Our findings extend the role of the pVHL complex. They show that, in addition to its role as an E3 ubiquitin ligase, which regulates the accumulation of HIF-α protein (9–12), and thereby

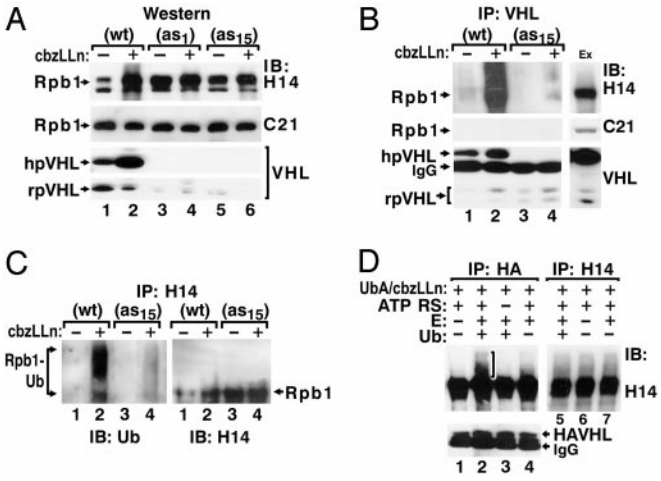


Fig. 5. Accumulation and ubiquitination of hyperphosphorylated Rpb1 in cells with different levels of pVHL. (*A*) Western blot analysis of hyperphosphorylated (H14) and hypophosphorylated (C21) Rpb1 in nuclear extracts from PC12VHL(WT) or two different clones of PC12VHL antisense (as) cells. (*B*) Coimmunoprecipitations using anti-pVHL antibody from nuclear extracts of WT and antisense cells. Ex, extract. (*C*) Immunoprecipitation of ubiquitinated forms of hyperphosphorylated Rpb1 from denatured cellular lysates by using H14 antibody. (*D*) *In vitro* ubiquitination reactions on protein complexes coimmunoprecipitated by using anti-HA (lanes 1–4) or H14 (lanes 5–7) antibodies from cellular extracts from PC12VHL(WT) cells. H14 antibody does not coimmunoprecipitate pVHL. UbA, ubiquitin aldehyde; E, purified enzymatic fraction II from reticulocyte lysate; ATP RS, ATP-regenerating solution. The bracket marks ubiquitinated forms of Rpb1.

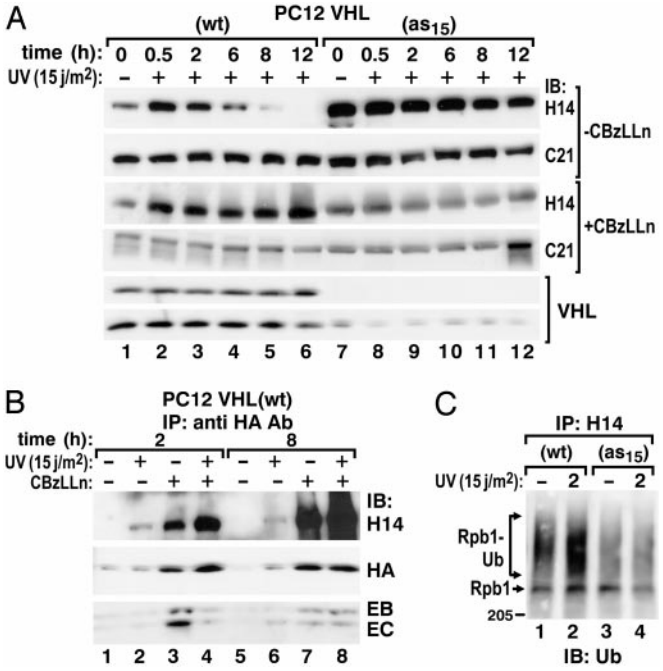


Fig. 6. Rpb1–pVHL interactions in response to the UV treatment. (*A*) Western blot analysis of Rpb1 and pVHL in nuclear extracts from cells in the absence (*Upper*) or presence (*Lower*) of CbzLLn. (*B*) Coimmunoprecipitation of hyperphosphorylated Rpb1 with anti-HA antibody in nuclear extracts from cells treated with UV for the indicated times. Blots were probed with indicated antibodies. (*C*) Ubiquitination of the hyperphosphorylated Rpb1 in response to the UV treatment in PC12VHL(WT) and antisense cells. Two hours after UV irradiations denatured cellular lysates were immunoprecipitated with H14 antibodies and the immunoblots were probed with anti-ubiquitin antibody.

expression of hypoxia-inducible genes, the pVHL complex can function as an E3 ligase, which targets the hyperphosphorylated Rpb1 for ubiquitination and degradation. Importantly, binding of pVHL to the full-size Rpb1 requires hydroxylation of proline-1465 within Rpb1 and phosphorylation of the CTD. To date, the proteins that pVHL targets for ubiquitination include, in addition to HIF-αs, a subfamily of deubiquitinating enzymes (43, 44) and activated atypical protein kinase C (45).

The exact role that ubiquitination of hyperphosphorylated Rpb1 by pVHL plays in the function of the RNA polymerase complex remains to be determined. Because ubiquitination of Rpb1 occurs in a transcription-dependent manner (22, 23), and because our earlier observations indicate that pVHL levels affect *in vivo* elongation of TH transcripts (18, 19), we anticipate that ubiquitination of Rpb1 by pVHL complex is likely to regulate efficient transcript elongation through elongation-pause and -arrest sites of specific genes. In particular, it may be involved in the regulation of TH transcript elongation (18–20). Such potential role of the pVHL–Rpb1 interaction is supported by the fact that pVHL binds in the pocket between Rpb1 and Rpb6, and that Rpb6 promotes elongation through arrest sites by binding to the elongation factor TFIIS (46). The pVHL-binding site is located on the surface of the elongating RNA polymerase II complex, and thus is accessible for pVHL binding during transcription. Interaction of pVHL with the RNA polymerase II complex may also locally titrate elongins B and C from elongation factor SIII (elongin ABC), thereby providing another mechanism by which pVHL could inhibit transcription elongation, as proposed based on *in vitro* studies (47).

We also anticipate that the pVHL–Rpb1 interaction has a more universal role and may regulate genes other than TH. The pVHL–Rpb1 interaction is regulated by UV stress, thus pVHL may play a role in the regulation of transcription complexes (transcription-coupled repair) under conditions of DNA damage, such as UV irradiation. In this respect, pVHL-negative cells undergo apoptosis in response to UV treatment, whereas the pVHL-positive cells do not (48). Our data also suggest a molecular mechanism by which the loss of pVHL function in von Hippel–Lindau disease may result in tumorigenesis.

Our results demonstrate that antisense cells having decreased levels of pVHL accumulate hyper- but not hypophosphorylated Rpb1, resulting in decreased ubiquitination of Rpb1. The most consistent explanation for this finding is that the antisense cells have decreased pVHL-associated E3 ligase activity toward the hyperphosphorylated Rpb1. However, it cannot be excluded that pVHL may affect the activity and/or expression of some kinases or phosphatases involved in the phosphorylation of CTD. In view of the role of CTD phosphorylation in pVHL binding, this last possibility might be an attractive regulatory mechanism increasing the pVHL–Rpb1 interaction under conditions of reduced amounts of pVHL.

These data provide biochemical evidence that Rpb1 can be modified by proline hydroxylation. It is unclear whether proline hydroxylation requires CTD phosphorylation. Two major groups of proline hydroxylases have been identified to date: the endoplasmic reticulum collagen proline hydroxylases (42) and the Egl-9-like group of proline hydroxylases involved in O_2-dependent regulation of HIF-α (7, 8). Both groups hydroxylate prolines in an O_2-, Fe(II)-, and 2-oxoglutarate-dependent manner; howeve, hydroxylases involved in collagen maturation are less sensitive to O_2 levels and are functional even under hypoxic conditions (42). In contrast, the HIF prolyl hydroxylases appear to be strictly O_2-sensitive, and their activities are inhibited by a decrease in pO_2 (7). At this time, it has not been possible to measure the O_2-sensitivity of prolyl hydroxylation of Rpb1 and pVHL binding because hyperphosphorylated Rpb1 disappears

rapidly from the hypoxic extracts by an as yet unknown mechanism (M.F.C.-K., unpublished results).

Computational models and experimental observations demonstrate significant structural similarity between the ODDD of HIF-1α and Rpb1/Rpb6. However, the HIF-1α pVHL-binding peptide needs to be rotated along a single bond in the central "bulge" to bring it into a good agreement with its predicted yeast Rpb1 counterpart. These different conformations of the pVHL-binding domains, as well as variations in the structure between human and yeast Rpb6, suggest the existence of some differences in the pVHL-binding mechanism between HIF-1α and Rpb1/Rpb6.

In summary, our work shows that the pVHL complex binds the hyperphosphorylated large subunit of the RNA polymerase II complex, in a proline hydroxylation- and CTD phosphorylation-dependent manner, targeting it for ubiquitination. These results indicate that pVHL plays a role in the regulation of gene expression and cellular function.

We thank Glenn Doerman for preparing the figures. The HA-VHL expression construct and RCC clones stably expressing HA-VHL were a gift from W. G. Kaelin, Jr. This work was supported by the following grants to M.F.C.-K.: National Institutes of Health Grants HL58687 and HL66312, American Cancer Society Research Scholar Grant GMC-101430, and a von Hippel–Lindau Family Alliance Research Grant. J.M. acknowledges support from the Children's Hospital Research Foundation Trustee Grant.

1. Kibel, A., Iliopoulos, O., DeCaprio, J. A. & Kaelin, W. G., Jr. (1995) *Science* **269,** 1444–1446.
2. Pause, A., Lee, S., Worrell, R. A., Chen, D. Y., Burgess, W. H., Linehan, W. M. & Klausner, R. D. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 2156–2161.
3. Kamura, T., Koepp, D. M., Conrad, M. N., Skowyra, D., Moreland, R. J., Iliopoulos, O., Lane, W. S., Kaelin, W. G., Jr., Elledge, S. J., Conaway, R. C., *et al.* (1999) *Science*, **284,** 657–661.
4. Cockman, M. E., Masson, N., Mole, D. R., Jaakola, P., Chang, G.-W., Clifford, S. C., Maher, E. R., Pugh, C. W., Ratcliffe, P. J. & Maxwell, P. H. (2000) *J. Biol. Chem.* **275,** 25733–25741.
5. Kamura, T., Sato, S., Iwai, K., Czyzyk-Krzeska, M. F., Conaway, R. C. & Conaway, J. W. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10430–10435.
6. Ohh, M., Park, C. W., Ivan, M., Hoffman, M. A., Kim, T. Y., Huang, L. E., Pavletich, N., Chau, V. & Kaelin W. G., Jr. (2000) *Nat. Cell Biol.* **2,** 423–427.
7. Epstein, A. C. R., Gleadle, J. M., McNeill, L. A., Hewiston, K. S., O'Rourke, J., Mole, D. R., Mukherji, M., Metzen, E., Wilson, M. I., Dhanda, A., *et al.* (2001) *Cell* **107,** 43–54.
8. Taylor, S. (2001) *Gene* **275,** 125–132.
9. Ivan, M., Kondo, K., Yang, H., Kim, W., Valiando, J., Ohh, M., Salic, A., Asara, J. M., Lane, W. S. & Kaelin, W. G., Jr. (2001) *Science* **292,** 464–468.
10. Jaakkola, P., Mole, D. R., Tian, Y. M., Wilson, M. I., Gielbert, J., Gaskell, S. J., Kriegsheim, A. V., Hebestreit, H. F., Mukherji, M., Schofield, C. J., *et al.* (2001) *Science,* **292,** 468–472.
11. Yu, F., White, S. B., Zhao, Q. & Lee, F. S. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 9630–9635.
12. Bruick, R. K. & McKnight, S. L. (2001) *Science* **294,** 1337–1340.
13. Iliopoulos, O., Levy, A. P., Jiang, C., Kaelin, W. G., Jr., & Goldberg, M. A. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 10595–10599.
14. Wizigmann-Voos, S., Breier, G., Risau, W. & Plate, K. H. (1995) *Cancer Res.* **55,** 1358–1364.
15. Yang, H. & Kaelin, W. G., Jr. (2001) *Cell Growth Differ.* **12,** 447–455.
16. Walther, M. M., Reiter, R., Keiser, H. R., Choyke, P. L., Venzon, D., Hurley, K., Gnarra, J. R., Reynolds, J. C., Glenn, G. M., Zbar, B. & Linehan, W. M. (1999) *J. Urol.* **162,** 659–665.
17. Eisenhofer, G., Walther, M. M., Huynh, T.-T., Li, S.-T., Bornstein, S. R., Vortmeyer, A., Mannelli, M., Goldstein, D. S., Linehan, W. M., Lenders, J. W. M. & Pacak, K. (2001) *J. Cell. Endocrinol. Metab.* **86,** 1999–2008.
18. Kroll, S. L., Paulding, W. R., Schnell, P. O., Barton, M. C., Conaway, J. W., Conaway, R. C. & Czyzyk-Krzeska, M. F. (1999) *J. Biol. Chem.* **274,** 30109–30114.
19. Bauer, A. L., Paulding, W. R., Striet, J. B., Schnell, P. O. & Czyzyk-Krzeska, M. F. (2002) *Cancer Res.* **62,** 1682–1687.
20. Hawryluk, P. & Luse, D. (2002) M.S. thesis (Case Western Reserve Univ., Cleveland).
21. Dahmus, M. (1996) *J. Biol. Chem.* **271,** 19009–19012.
22. Mitsui, A. & Sharp, P. A. (1999) *Proc. Natl. Acad. Sci. USA.* **96,** 6054–6059.
23. Lee, K. B., Wang, D., Lippard, S. J. & Sharp, P. A. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 4239–4244.
24. Ratner, J. N., Balasubramanian, B., Corden, J., Warren, S. L. & Bergman, D. B. (1998) *J. Biol. Chem.* **273,** 5184–5189.
25. Bregman, D. B., Halaban, R., Van Gool, A. J., Henning, K. A., Friedberg, E. C. & Warren, S. L. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 11586–11590.
26. Rockx, D. A., Mason, R., van Hoffen, A., Barton, M. C., Citterio, E., Bregman, D. B., van Zeeland, A. A., Vrieling, H. & Mullenders, L. H. F. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10503–10508.
27. Svejstrup, J. Q. (2002) *Nat. Rev.* **3,** 21–29.
28. Beaudenon, S. L., Huacani, M. R., Wang, G., McDonnell, D. P. & Huibregtse, J. M. (1999) *Mol. Cell. Biol.* **19,** 6972–6979.
29. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K. & Bairoch, A. (2002) *Nucleic Acids Res.* **30,** 235–238.
30. Jones, D. T. (1999) *J. Mol. Biol.* **292,** 195–202.
31. Meller, J. & Elber, R. (2001) *Proteins Struct. Funct. Genet.* **45,** 241–261.
32. Meller, J. & Elber, R. (2000) Learning, Observing, and Outputting Protein Patterns (LOOPP), a Program for Protein Recognition and Design of Folding Potentials (Cornell Univ., Ithaca, NY), Version 2.0.
33. Frary, A., Nesbitt, T. C., Grandillo, S., Knaap, E., Cong, B., Liu, J., Meller, J., Elber, R., Alpert, K. B. & Tanksley, S. D. (2000) *Science* **289,** 85–88.
34. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. E. (2000) *J. Mol. Biol.* **299,** 501–520.
35. Cramer, P., Bushnell, D. A. & Kornberg, R. D. (2001) *Science* **292,** 1863–1876.
36. Min, J. H., Yang, H., Ivan, M., Gertler, F., Kaelin, W. G., Jr., & Pavletich, N. P. (2002) *Science* **296**, 1886–1889.
37. Hon, W. C., Wilson, M. I., Harlos, K., Claridge, T. D., Schofield, C. J., Pugh, C. W., Maxwell, P. H., Ratcliffe, P. J., Stuart, D. I. & Jones, E. Y. (2002) *Nature* **417,** 975–978.
38. Kim, E., Du, L., Bregman D. B. & Warren S. L. (1997) *J. Cell Biol.* **136,** 19–28.
39. Bregman, D. B., Du, L., Van der Zee, S. & Warren, S. L. (1995) *J. Cell Biol.* **129,** 287–298.
40. Patturajan, M., Schulte, R. J., Sefton, B. M., Berezney, R., Vincent, M., Bensaude, O., Warren, S. L. & Corden, J. L. (1998) *J. Biol. Chem.* **273,** 4689–4694.
41. Franklin, T. J., Morris, W. P., Edwards, P. N., Large, M. S. & Stephenson, R. (2001) *Biochem. J.* **353,** 333–338.
42. Kivirikko, K. I. & Pihlajaniemi, T. (1998) *Adv. Enzymol. Relat. Areas Mol. Biol.* **72,** 325–398.
43. Li, Z., Na, X., Wang, D., Schoen, S. R., Messing, E. M. & Wu, G. (2002) *J. Biol. Chem.* **277,** 4656–4662.
44. Li, Z., Wang, D., Na, X., Schoen, S. R., Messing, E. M. & Wu, G. (2002) *Biochem. Biophys. Res. Commun.* **294,** 700–709.
45. Okuda, H., Saitoh, K., Hirai, S., Iwai, K., Takaki, Y., Baba, M., Minato, N., Ohmo, S. & Shuin, T. (2001) *J. Biol. Chem.* **276,** 43611–43617.
46. Ishiguro, A., Nogi, Y., Hisatake, K., Muramatsu, M. & Ishihama, A. (2000) *Mol. Cell. Biol.* **20,** 1263–1270.
47. Duan, D. R., Pause, A., Burgess, W. H., Aso, T., Chen, D. Y. T., Garrett, K. P., Conaway, R. C., Conaway, J. W., Linehan, W. M. & Klausner, R. D. (1995) *Science* **269,** 1402–1406.
48. Schoenfeld, A. R., Parris, T., Eisenberger, A., Davidowitz, E. J., De Leon, M., Talasazan, F., Devarajan, P. & Burk, R. D. (2000) *Oncogene* **19,** 5851–5857.

Mutations within P2 domain of Norovirus Capsid Affect Binding to Human Histo-Blood Group Antigens: Evidence for a Binding Pocket

MING TAN[†], PENGWEI HUANG[†], JAROSLAW MELLER[†], WEIMING ZHONG,

TIBOR FARKAS AND XI JIANG[*]

Division of Infectious Diseases and Division of Pediatric Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

ABSTRACT

Noroviruses (NORs) are an important cause of acute gastroenteritis. Recent studies of NOR receptors showed that different NORs bind to different histo-blood group antigens (HBGAs) and at least four distinct binding patterns were observed. To determine the structure-function relationship for NORs and their receptors, two strains representing two of the four binding patterns were studied. Strain VA387 binds to HBGAs of A, B and O secretors, whereas strain MOH binds to HBGAs of A and B secretors only. Using multiple sequence alignments, homology modeling and structural analysis of NOR capsids, we identified a plausible "pocket" in the P2 domain that may be responsible for binding to HBGA receptors. This pocket consists of a conserved RGD/K motif surrounded by three strain-specific "hot spots" (N_{302}, T_{337}, and Q_{375} for VA387 and N_{302}, N_{338} and E_{378} for MOH). Subsequent mutagenesis experiments demonstrated that all four sites played important roles in binding. A single amino acid (aa) mutation at T_{337} (to A) in VA387 or a double aa mutation at RN_{338} (to TT) in MOH abolished binding completely. Change of the entire RGD motif to SAS abolished binding in case of VA387, whereas single aa mutations in that motif did not have an apparent effect on binding to A and B antigens but decreased binding to H antigen. Multiple mutations at the RGK motif of MOH (SIRGK to TFRGD) completely knocked out the binding. Mutation of N_{302} or Q_{375} in VA387 affected binding to type O HBGA only, while switch mutants with three aa changes at either site from MOH to VA387 resulted in a weak binding to type O HBGAs. A further switch mutant with three aa changes at E_{378} from MOH to VA387 diminished the binding to type A HBGA only. Taken together, our data indicated that the computationally identified putative pocket on the P2 domain of Norovirus capsid proteins is involved in receptor binding. Further studies of the binding specificity of individual strains and its structural determinants may help to elucidate the molecular pathogenesis of Norovirus infection in humans.

INTRODUCTION

Noroviruses (NORs) are the most important cause of non-bacterial epidemics of acute gastroenteritis, affecting individuals of all ages, in both developing and developed countries (7, 10). NORs are icosahedral, single-stranded, positive-sense RNA viruses whose capsids are composed of 180 copies of a single major structural protein (19, 23, 36). The viral capsid proteins expressed by baculovirus in insect cells self-assembly into virus-like particles (VLPs)(21, 22). These VLPs are morphologically and antigenically indistinguishable from authentic virions (21, 22), providing a useful tool for development of immunological assays and for study of receptor/virus interaction. Data from cryo-electron microscopy and X-ray crystallography showed that the

[*] Corresponding author.

[†]The first three authors contributed equally to this paper.

viral capsid protein folds into two major domains, the S and P domains (36, 37). The S domain forms the interior shell, while the P domain builds up arch-like structures that protrude from the shell. Morphogenesis studies showed that the S domain contains elements required for assembly of the capsid, whereas intermolecular contacts between dimeric subunits of the P domain increase the stability of the capsid (1). The P domain is further divided into P1 and P2 domains, with the latter located at the most exterior surface of the capsid. As opposed to the S and P1 domains, the P2 domain is characterized by relatively high sequence variability and therefore is believed to be critical in immune recognition and receptor binding.

NORs have been recently found to recognize human histo-blood group antigens (HBGAs) as receptors (14, 16-18, 27, 31). Moreover, the recognition of HBGAs by NORs was demonstrated to be strain specific. So far, four distinct binding patterns of NORs, which were defined by the ABO, Lewis and secretor types of human host (16), have been described. Human HBGAs are complex carbohydrates linked to glycoproteins or glycolipids that are present on the red blood cells and mucosal epithelial cells, or as free antigens in biological fluids, such as blood, saliva, intestinal content, and milk (30). These antigens are synthesized by sequential additions of monosaccharides to the antigen precursors by several glycosyltransferases that are genetically controlled and known as the ABO, Lewis and secretor gene families (30).

The prototype Norwalk virus (NV) represents one of the four binding patterns and it binds to HBGAs of types A and O secretors but not of non-secretors (16, 27). Human volunteer studies showed that saliva from volunteers with non-secretor status did not bind to NV and non-secretors were naturally resistant to NV infection following the challenge (27). In our studies, NV did not bind to saliva of type B secretors (16). A retrospective study of volunteers challenged with NV showed that type B individuals had a lower rate of infection to NV than other blood types following the challenge (17). The other three binding patterns recognize A, B and O secretors (strain VA387), A and B secretors (MOH) and Lewis positive secretors and non-secretors (strain VA207) (16). By analogy, we predict that each of the three binding patterns may have its own host ranges defined by human blood types, although direct evidence linking HBGAs with infection of these strains remains lacking.

In this study, we addressed the question of structural determinants of NOR capsids binding to HBGA receptors. Using computational approaches we identified a putative receptor binding site on the surface of the P2 domain. Mutagenesis data revealed that this putative binding pocket is indeed involved in specific binding to HBGAs. More importantly, single aa changes within this pocket knocked out the binding completely, whereas shifting mutations resulted in change of binding patterns, highlighting the importance of this newly identified site for the virus/host interaction.

MATERIALS AND METHODS

Protein sequence analysis and computer modeling. Multiple alignments of known NOR capsid sequences were carried out by computer software OMIGA 2.0 (Oxford Molecular Ltd.). The crystal structure of the prototype Norwalk virus capsid protein (PDB code 1I HM, (36)) was used to built homology models for other NOR strains. The initial sequence-to-structure alignments and the refined 3-dimensional models of the NOR capsids with minimized side chain conformations were obtained using the 3D-PSSM (6, 25) and MODELLER (33, 41) servers and programs as well as our own LOOPP (34) program.

Construction of mutant NOR capsids by site-directed mutagenesis. A series of mutant NOR capsids of strains VA387 (24) and MOH (11) were constructed using the QuikChange Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA). The capsid genes of VA387 and MOH were cloned into pGEM-T vector (Promega, Madison, WI). Primers for site-directed mutagenesis were designed according to manufacturer's instruction with at least 15 nucleotides at both ends to the nearest mutated nucleotides. For strain 387 the following primers were used: gttgtccaaccacaaagtgccagttgcacgactgatggc (NGR/SAA), ccaaggtgttttacggtccacgtgcagac (NGR/NAR), gtcaatatctgcaccttcagtgcggctgtcacccacattgcag (RGD/SAS), ctgcaccttcgcaggggatgtcacccac (RGD/AGD), ctgcaccttcagagcggatgtcacccac (RGD/RAG), ctgcaccttcagagggaaagtcacccacattgcag (RGD/RGK), gtcatgactatataatggctttggcatctcaaaattgg (N/A), gctcacccaaaccgcaagagaggatggc (T/A), gacacaaacaatgattttgcaactggccaaaacacg (Q/A). For strain MOH the following primers were used: gttgttcagccacagagtgctagcgtcacattagatggg (NGR/SAS), tgtaacatttgcaccttcgaggggacgtgacagggcag (SIRGK/TFRGD), cacatgtggaacatgaacctcacaaacctaaatggg (LEI/MNL), ggtgtgctcagccagacaaccagaggcgaaagcaac (RN/TT), aacaccaatgattttcaaacccaaccaacaaaattc (VEN/FQT). Chimera capsid of MOH with 11 aa replacement from VA387 at site II (chimera 1) or site IV (chimera 2) were made by overlapping PCRs. These were achieved by designing a pair of primers that can

anneal to the positions adjacent to the regions of replacement and have the overlapping sequences (mutated sequences) at their 5'-end (attttgagatgccaaattcattatatagtcatgattagggacctgccctgtcac and catgactatataatgaatttggcatctcaaaataacctaaatgggacgcaatttg for chimera 1; cgtgttttggccagtttgaaaatcattgtttgtccaagttccaatttgcactaag and acaaacaatgattttcaaactggccaaaacacgttcacccc aattggtttgaat for chimera 2). PCR products from these two primers, together with a primer at the beginning or end of the coding region, were gel purified and used together as templates for a second PCR using primers at the both ends of the coding region of the capsid gene. Chimera 3 that contains mutations of chimera 1 and 2 was prepared using the same method, with chimera 1 as the starting construct. The mutated sequences were further validated by sequencing.

Expression and purification of mutated capsid proteins. Mutated capsids were expressed in *Spodoptera frugiperda* (Sf9) cells using the Bac-to-Bac Baculovirus Expression System (Invitrogen, Carlsbad, CA) according to the manufacturer's manual. Briefly, the mutated capsid genes were subcloned into pFastBac1 donor plasmid and transpositioned into bacmid. Sf9 cells were then infected with wild-type baculovirus or the baculovirus recombinant bacmids containing the mutated capsid genes. Infected cells were harvested between the third and the sixth generation. After a few cycles of freeze-and-thaw, the cell lysates were centrifuged at 5,000 g for 15 min to separate the cell debris. The clear supernatants were centrifuged again at 10,000 g for 30 min to get rid of large protein complexes or baculovirus particles. VLPs in the supernatant were then purified by centrifugation at 100,000 g for 150 min. For further purification of the VLPs, the resuspended pellets were separated using sucrose step-gradient (5 to 45%), as described previously (21, 22). The recombinant proteins were stored in 1x PBS, pH 7.4, at -70ºC. In some cases, VLPs were visualized by electron microscope with a negative staining. To quantify the expressed proteins, a small aliquot of purified samples was separated on a 10 % SDS-PAGE gel. A series of dilutions of the quantitated recombinant wild type VLPs of the two strains (VA387 and MOH) were loaded on the same gel as standards. Proteins were then transferred to a nitrocellulose membrane for Western blot analysis with hyperimmune-antibodies raised in guinea pigs against the wild type capsids of VA387 and MOH, respectively. The protein concentrations in the samples were determined by comparing the signal intensities using the image analysis software Scion Image (4.0.2, Scion Corporation) after immuno-detection. Because we found background bands from wild type baculovirus in some Western blot experiments (Fig. 4), we did not use enzyme immune assays to determine the concentrations of mutated capsids.

Characterization of NOR capsid binding to HBGAs by saliva binding enzyme immune assays. Saliva binding enzyme immune assays were used to monitor recombinant VLP binding to HBGAs, as described previously (16). Briefly, saliva samples of known ABO, secretor and Lewis types were pretreated by boiling at 100 ºC and centrifugation at 10,000 g for 5 min. The supernatant was then coated on microtiter plates (Dynex Immulon, Dynatech, Franklin, MA) at a dilution of 1:1000 in 1x PBS (pH 7.4). After blocking with 5% dried milk (Blotto), known amounts of the wild type recombinant capsid proteins or their mutated forms were added with a serial dilution. The bound capsid proteins were detected by hyperimmune guinea pig anti-NORs antisera, and by adding horseradish peroxidase (HRP)-conjugated goat anti-guinea pig IgG (ICN, Aurora, OH). The HRP activity was detected by the TMB kit (Kirkegard & Perry Laboratories, Gaithersburg, MD) and the signal intensities (OD) were read by EIA spectra reader (Tecan, Durham, NC). In order to determine the binding affinity of the mutated capsids in relation with their wild type one, all capsids were assayed in the same dynamic conditions for their binding to HBGAs of A, B and O type saliva within a comparable range of protein concentrations.

RESULTS

Identification of NOR capsid domains binding to HBGA receptors. Because of the surface location and high genetic variation of the P domain of NOR capsid, our initial sequence analyses were focused on this region. Multiple alignments of the capsid sequences of NORs representing the four distinct binding patterns to HBGA receptors revealed a conserved RGD-like motif in the P2 domain (aa 288-290 in VA387, Fig. 1). The RGD motif has been identified as a universal recognition site for cell-to-cell and cell-to-extracellular environment interactions, such as receptor-ligand, signal transduction, enzyme-substrate, and hormone-target interactions (39). Typical receptor-ligand interaction includes foot-and-mouth disease virus (32), Coxsachievirus and adenovirus receptor recognition (47). To determine whether the RGD-like motif also plays a role in NOR receptor binding, we used a combination of multiple alignments and structural analysis to search for sites conserved between strains of similar binding patterns and at the same time being in close spatial proximity. The results showed that the conserved RGD-like motif is surrounded by a cluster of strain-specific

residues, forming a pocket on the surface of the P2 domain (Fig. 2, based on NV). The RGD-like motif is located at the bottom, and the three hot spots N_{302}, T_{337}, and Q_{375} (VA387) surround the pocket. All these four sites are in a close spatial proximity. For convenience, the four sites are referred to as I, II, III, and IV, following their sequence order. The corresponding four sites in MOH are $RGK_{288-290}$, E_{302}, N_{338}, and E_{378}, respectively (Fig. 1).

Mutant construction and confirmation of VLP formation. To assess the sequence specificity of NOR capsid protein in binding to HBGAs, a series of mutants with mutations at each of the four sites have been constructed (Fig. 3) and the resulting recombinant capsid proteins were characterized for their binding to different HBGAs (Fig. 5- 8). To exclude the possible influence of a mutation on VLP formation, the recombinant capsid proteins were confirmed for VLP formation. This was achieved by purifying the mutated VLPs from sucrose gradients at the expected fractions (30-40%), and/or visualizing VLPs with electron microscopy (data not shown). In few cases VLPs were purified by a high-speed pelleting at 100,000 g for 150 min. All mutated capsids used in this study were confirmed to form VLPs. Fig. 4 shows the baculovirus expressed, mutated and wild type capsids after partial purification by high-speed centrifugation.

The RGD-like motif is important for NOR capsid binding to human HBGAs. The RGD motif has been shown to be responsible for receptor-ligand interaction in other viruses. To determine whether this motif also plays a role in NOR receptor binding, recombinant mutants with mutations in the RGD-like motif were constructed for VA387 and MOH. One VA387 mutant with the entire RGD motif mutated to SAA) abolished the binding completely (Fig. 5A). Mutants with only one aa change in the RGD motif (e.g. R_{288} to A, or G_{289} to A) did not affect binding to A or B type saliva (Fig. 5B). However, these mutants had reduced affinities to O type saliva comparing to the VA387 wild type. Modification of RGD to RAD led to a complete loss of binding to O type saliva. Our recent studies using synthetic oligosaccharides have shown that A, B, and H antigens (terminal carbohydrates) in corresponding saliva are responsible for NOR binding (16, Jiang, #15045) . MOH differs from VA387 only in the recognition of the H antigen. VA387 contains an RGD motif while MOH contains an RGK motif at site I. To test whether this difference plays a role in binding specificity, a switch mutant from RGD of VA387 to RGK of MOH was constructed. This mutant did not reveal significant change in biding to A, B and H antigens (Fig 5B). Another switch mutant involving larger sequence alteration in the vicinity of the RGD/K motif, shifting from MOH (SIRGK) to VA387 (TFRGD), did not gain binding to the H antigen either. Instead, it lost binding to the A and B antigens (Fig. 5C). In conclusion, the RGD/K motifs are directly involved in NORs binding to HBGA receptors. Whether they are also responsible for strain-specific binding remains unclear.

Site III is critical for NOR capsid binding to human HBGAs. Sites II, III and IV are located at the opening of the predicted binding pocket and they are not conserved among strains representing different receptor binding patterns. Thus, they may be responsible for the strain specificity to HBGAs. However, site-directed mutagenesis analysis did not support that site III is responsible for binding specificity. One VA387 mutant with a single aa change from T_{337} to A (Fig. 3) completely lost its binding to the HBGAs of A, B and O types (Fig. 6). A double aa switch mutant at this position from RN of MOH to TT of VA387 not only did not gain binding to the H antigen, but also lost binding to the A and B antigens (Fig. 7). Two more mutants with mutations at site III plus mutations at sites II and/or IV (Fig. 3) also led to the same results. Taken together, our data indicated that site III plays a critical role in NOR capsid binding to human HBGAs since even a single aa change can result in complete loss of binding.

Sites II and IV may play a role in binding specificity to HBGAs. Two mutant capsids with mutations at sites II (N_{302} to A) and IV (Q_{375} to A, Fig. 3) were constructed for VA387. Both mutant capsids did not reveal significant change in their binding to the A and B antigens, but did result in decreased binding to the H antigen in O type saliva, suggesting that sequences at these two sites might play a role in the binding specificity to human HBGAs. In addition, we constructed shift mutants from MOH to VA387 at these sites and found that shift mutants with three aa changes at either sites (LEI to MNL at site II or VEN to FQT at site IV) resulted in a weak binding to type O saliva (Fig. 7A and B). Moreover, the second mutant (VEN/FQT) also lost its binding to type A, but not to type B, antigens. In considering that both mutants gained only weak binding to O saliva, we constructed switch mutants with larger sequence changes to see if stronger binding to O saliva could be obtained. Among three chimeras with 11 aa shifting in the vicinity of sites II and/or IV from MOH to VA387 constructed (Fig. 3), none resulted in a gain of binding to type O saliva, instead, all lost binding completely to the types A and B antigens (data not shown). In conclusion, both sites II and IV are important for binding

specificity, although additional sites may also be involved. The loss of binding to HBGAs in the case of the three chimera mutants is most likely due to interruption of the required structure of the binding pocket.

The NGR motif is required for receptor binding. Sequence comparison of NOR capsid proteins revealed another highly conserved NGR motif that was found in all known human and animal enteric Noroviruses. Moreover, the Asparagine residue is conserved in all caliciviruses. This motif is located at 20 aa upstream of the RGD-like motif and at the interface between the P1-1 and P2 domains. Therefore, this motif is likely to play an important role in structure and function of NORs. The NGR motif has been found to be involved in the interaction with integrin for rotavirus (9). To determine if this motif is also involved in NOR receptor binding, the same knock-out mutants (NGR/SAA) were constructed for both VA387 and MOH. Both mutants resulted in low yields of the recombinant capsid proteins and VLPs in the insect cells, with less than one-quarter of that of other recombinant capsid proteins made in this study (Fig. 4), even with high titered recombinant viral stocks. Saliva binding assay of these mutants showed no detectable binding activity (Fig 8). To further dissect the motif for its role in receptor binding, another mutant (NGR/NAR) with a single aa modification was made. Again a low yield of the protein and no detectable binding to HBGAs were observed. Based on the above observations we speculated that the NGR motif is important for the maintenance of capsid structure. The loss of the binding to HBGAs in NGR-related mutants is probably due to a local or global conformation change(s) of the capsids, which directly or indirectly affects the conformation of the binding pocket.

DISCUSSION

This study characterized the structure of the NOR capsids and their binding to HBGAs. Using computational analyses, we identified a putative binding pocket in the P2 domain that could mediate binding to different HBGA receptors. Site-directed mutagenesis analyses provided further evidences supporting the binding pocket hypothesis. In particular, the following observations supported our conclusion. First, all four sites in the P2 domain that were predicted to be involved in the formation of the binding pocket by computational analysis were found to influence the ability of NORs to bind to different HBGAs based on site-directed mutagenesis analysis. Second, the two strains of NORs characterized in this study have distinct binding patterns, but the outcome of binding of their mutant capsids were very similar, suggesting that a common structure is formed for the two capsids despite differences in their primary sequences. Third, two of the four sites, the RGD-like motif and site III, were found to be critical for the binding. Changing sequences at either of the sites, particularly the single aa change at site III, resulted in a complete loss of binding. Finally, switch mutants with sequence shifting between the two strains at sites II and IV, as well as the RGD-like motif, revealed changing of binding patterns at certain levels, suggesting that these sites contribute to a certain extent to binding specificity. Thus, we believe that the binding pocket identified by computational analysis is likely to be the receptor binding site. At least four sites are involved in the pocket formation and receptor binding, although additional site(s) may also be important. The fact that the switch mutants with a three-aa shift from MOH to VA387 at site II and IV did not result in a significant binding to H antigen and that the switch mutants with mutations involving larger regions (11 and 22 aa) at these two sites resulted in a complete loss of binding suggested that the binding determinants are associated with a very strictly defined structure.

The two strains characterized in this study have multiple determinants for individual HBGAs. A previous study and our recent data indicated that MOH recognizes the A (glucose residue) and B (N-acetyl-glucosamine residue) antigens, while VA387 recognizes the H antigen (1,2-fucosyl residue) in addition to the A and B antigens (16, 20). One question is how variations in the binding pocket contribute to recognition of these substrates. As demonstrated by our results, site III is responsible for binding but not for strain specificity and single or double aa changes at this site resulted in loss of binding to all epitopes (A, B and/or H) for both strains. The other three sites showed certain levels of strain-specific binding but did not account fully for the binding specificity. Therefore, we hypothesize that the binding specificities of NOR capsids may not be determined by a linear eipitope or a single residue; instead, they may involve conformational epitopes and combination of several residues surrounding the pocket.

In our recent NOR receptor studies, we have observed that saliva samples from type A individuals could block VA387 and MOH binding to type B saliva and vice versa, suggesting that both A and B antigens share the same binding site. Once this site is occupied, it is no longer accessible to other molecules. Thus, we assume that each capsid protein has a single pocket. The overall shape and chemical characteristics (e.g. charge distribution) of the pocket is likely to be different for different strains due to different side chains of the residues surrounding the pocket, providing fitness to a specific group of HBGAs.

The identification of the RGD-like motif in the NOR capsids is an important finding. It led to the prediction of the binding pocket in the P2 domain. The location of the motif at the bottom of the pocket suggested that it interacts directly with HBGAs. NORs are genetically diverse but the RGD-like motif is conserved among all NORs and the Lagoviruses. The first aa of the motif of NORs is a basic and positively charged Arginine (R) or Lysine (K) residue, except for the bovine Newbury strain that contains a Valine (V). In Lagoviruses it is either Arginine or Serine. The second aa of the motif is highly conserved throughout the two genera and the third aa is more variable. The recent discovery of NORs in animals raised the questions of zoonotic transmission or animal reservoir for human NORs. However, direct evidence for interspecies transmission of NORs is still lacking. Recently, a large surveillance of bovine enteric caliciviruses (BoCVs) in the United Kingdom between 1976 and 2000 showed that the BoCVs represent a distinct genogroup III of NORs, but they did not pose a treat to human health (35). Human blood types are unique among most mammal species, and known NORs mainly infect humans. A recent study showed that the prototype Norwalk virus does not recognize blood antigens of many non-human mammals except chimpanzees (18). Most interestingly, the rabbit calicivirus recognizes the human H type 2 antigen (40). Thus, the RGD-like motif could be a genetic marker for host specificity between animal and human caliciviruses. Two exceptions, however, have been found so far. One is the Jena strain of bovine NORs (28) that contains an off-location RGD motif relative to those in human and rabbit strains (data not shown). The other exception is the swine NOR that belongs to Genogroup II and reveals a RGT motif (45). Whether these exceptional animal strains are truly zoonotic or of human origin and whether they pose a threat to human health remains unclear.

In contrast to the RGD-like motif, the role of the NGR motif may be significantly different. This motif is conserved in all animal and human NORs. Therefore, a common pressure not related with the host receptors must have selected it. The complete knock out of binding by mutations in this region suggests that this motif is indispensable. The location of this motif, near but not in the binding pocket, suggests that this motif is not directly involved in interaction with HBGAs. Thus, the lack of binding to receptors by VLPs containing NGR mutations indicates that the NGR motif may involve local conformation changes that perturb the structure of the binding pocket. The NGR motif is also likely to be required for capsid assembly, since mutations in this motif lead to low yields of the capsid proteins. Alternatively, the low yields are due to a low expression of the proteins. Low expression of mutated capsid proteins not due to conformational stability have been reported for mutations in other regions of the NV capsid (46).

This is the first study to dissect the structure of NOR capsid in relationship with receptor binding. The attachment/entry of a virus to host cells could be the first step of viral infection. A precise map of binding domains and an elucidation of the structure of the interface between the receptor and viral capsid would facilitate understanding the virus/host interaction. This may lead to a discovery or design of specific compounds as antiviral drugs to block the virus infection. A growing number of viral and bacterial pathogens have been linked with histo-blood group antigens that serve as receptors for infection (2-5, 8, 12, 15, 26, 29, 38, 42-44). Although different pathogens cause different illness, they may share common mechanisms of interaction with histo-blood group antigen receptors. NOR capsids are formed by a single capsid protein, making this system simpler compared to many other viral and bacterial pathogens. Our recent studies have described four binding patterns of NORs. The target histo-blood group antigens within each pattern have been clearly defined (16) (unpublished data). Thus, NORs could provide a unique model of pathogen/host interaction for the human histo-blood group antigen system. Elucidation of this model promises to lead to new strategies for therapeutic control of emerging pathogens.

In this study we characterized only four sites within the P2 domain of NOR capsid. Whether additional sites within or adjacent to the predicted pocket are also involved in binding remains unclear. In addition, this study characterized two strains with close binding patterns and both strains belong to the same genogroup of NORs. So far, at least 20 genetic clusters within three genogroups of NORs have been identified (13). Strains within the same genogroup can target different HBGA receptors and strains in different genogroups can also have the same targets. Therefore, characterization of additional strains representing more genotypes and binding patterns is necessary. Finally, according to the biosynthetic pathways of HBGAs, the target antigens for individual binding patterns have been predicted. Thus, experiments using synthetic oligosaccharides representing these antigenic epitopes should be performed. Due to the unavailability of some of the synthetic oligosaccharides, our study used saliva-binding assays only. Future studies using defined oligosaccharides to confirm our results are necessary.

ACKNOWLEDGMENTS

The research described in this article was supported by the National Institute of Allergy and Infectious Diseases (RO1 AI37093-6).

FIGURES

Fig. 1. Sequence comparison of the P2 domain of NOR capsids. The four strains representing the four binding patterns to human HBGAs are VA387, NV, MOH and VA207. Four sites (I – IV) that are potentially responsible for building up a putative binding pocket are indicated in bold. The NGR motif upstream of the P2 domain is also indicated in bold. Strains VA387 (387) and Grimsby virus (GrV) bind to A, B, and O type saliva. The binding patterns of the Bristol (BV) and Lordsdale viruses (LV) are unknown but they share over 95% aa identity with VA387 and Grimsby virus. The prototype Norwalk virus (NV) is the only strain known to bind to A and O saliva. MOH and Mexico virus (MxV) bind to types A and B saliva. The binding patterns of their homologous strains Hillingdon virus (HIL) and Toronto virus (TV) remain to be determined. VA207 (207) binds to the Lewis epitope of secretors and non-secretors. According to our preliminary results, a new representative strain within genogroup I, Boxer (BX) also binds to the Lewis epitope (unpublished data) but additional characterization of this strain is necessary. The stars indicate conserved residues for all strains. Numbers on the right indicate the sequence position of capsid proteins.

Fig 1

Fig. 2. Computational prediction of a plausible binding pocket on the surface of the Norovirus capsid protein. The predicted pocket is located on top of the P2 domain and it is composed of a conserved RGD-like motif (R_{291}) and three strain-specific "hot spots" N_{300}, F_{335}, and N_{368} that are located in a close spatial proximity (see text for detail). "Ball and stick" models of the side chains are used to indicate the critical residues surrounding the putative binding pocket. The oval circle next to the P2 domain represents the P1 domain of the capsid protein. The S domain is not shown. The Rasmol visualization program was used to prepare the figure.

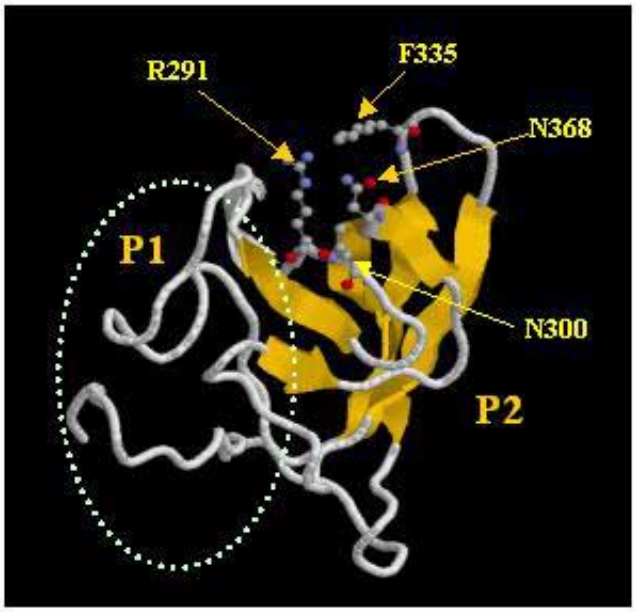
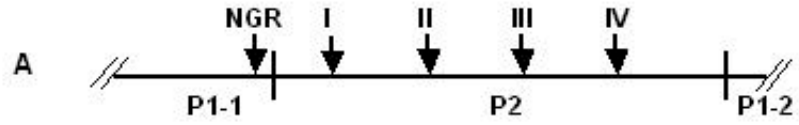


Fig 2

Fig. 3. Schematic representation of mutation constructs of NOR capsids used in this study. The graphic representation of the P domains with emphasis on P2 domain is shown on the top (A). Arrows indicate the positions of the four sites that are predicted to build up the binding pocket. The conserved NGR motif is also indicated. Panel B shows the sequences of different mutants in the NGR motif and the four sites of the VA387 and MOH capsid proteins.



B

VA387:

1.	NGR/SAA				
2.	NGR/NAR				
3.		RGD/SAS			
4.		RGD/RGK			
5.		RGD/RAD			
6.		RGD/AGD			
7.			N/A		
8.				T/A	
9.					Q/A

MOH:

10.	NGR/SAS				
11.		SIRGK/TFRGD			
12.			LEI/MNL		
13.				RN/TT	
14.					VEN/FQT
15.			LEI/MNL	RN/TT	
16.			LEI/MNL	RN/TT	VEN/FQT

17. MOH/VA387-Chimera-1
18. MOH/VA387-Chimera-2
19. MOH/VA387-Chimera-3

Fig. 3

Fig. 4. Western blot analysis of different mutant capsid proteins expressed in baculovirus infected Sf9 culture. Panel A shows mutants from VA387 and panel B shows mutants from MOH. Each sample contained partially purified VLPs corresponding to an equal amount of original insect culture. The proteins were detected by hyperimmune guinea pig antibodies against recombinant wild type VA387 (rVA387) and MOH (rMOH) capsids, respectively. In most cases, two major bands at ~58 and ~50 kDa were observed for each recombinant capsid. Arrows show the background bands from baculovirus.

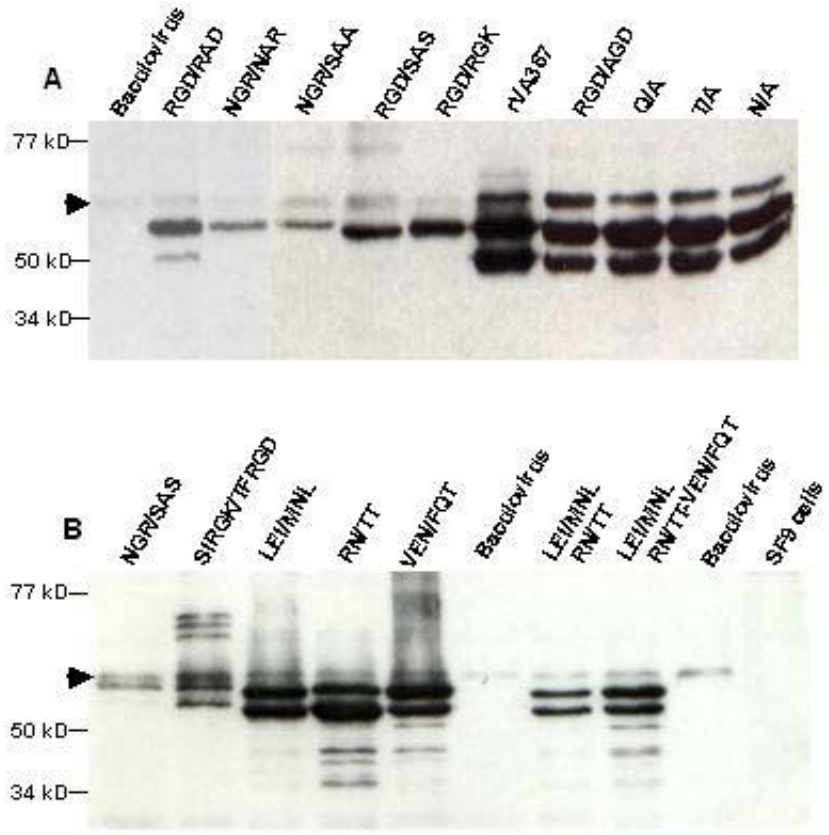


Fig 4

Fig. 5. Binding curves of mutants with mutations related to the RGD-like motif. The X axes indicate the concentration (pMol) of the capsid proteins and the Y axes indicate the OD values obtained from the saliva binding assay. Panel A shows mutants with amino acid changes of the entire RGD-like motif. Panel B shows mutants with single amino acid change in the RGD-like motif. Panel C shows mutant with longer sequences shifting from MOH to VA387. Data were averaged from at least two independent experiments. ●: A antigen; ▲: B antigen; ♦: H antigen (type O saliva).

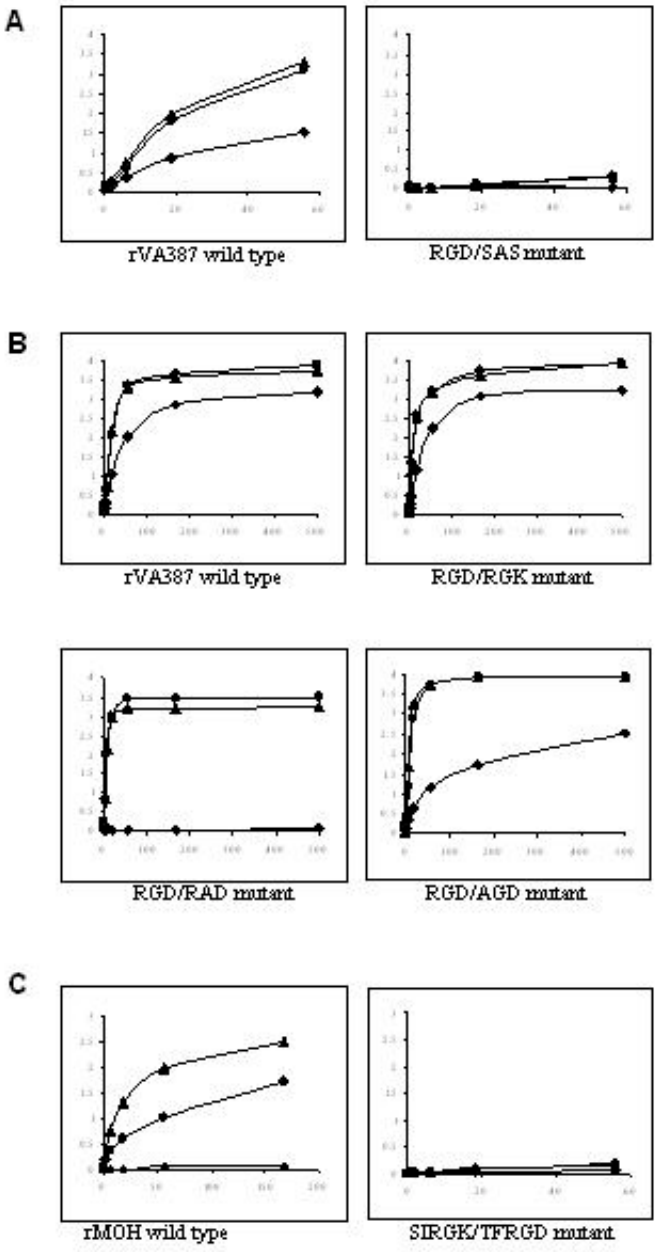


Fig 5

Fig. 6. Binding curves of mutants with single amino acid modification at sites II, III, and IV. The X and Y axes are the same as Fig. 5. Data were averaged from at least two independent experiments. ●: A antigen; ▲: B antigen; ♦: H antigen (type O saliva).

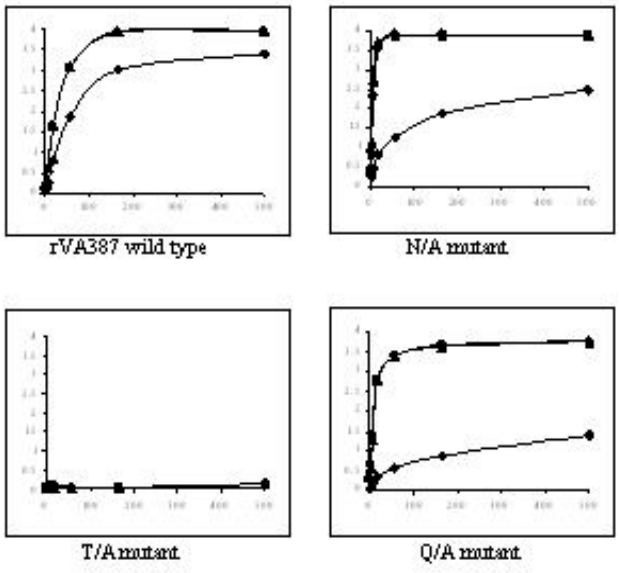


Fig 6

Fig. 7. Binding curves of shift mutants with sequence modifications from MOH to VA387 at sites II, III, and IV (A) and their comparison of binding to H antigen with the wild type (B). The X and Y axes in panel A are the same as Fig. 5. Data were averaged from at least two independent experiments. ●: A antigen; ▲: B antigen; ♦: H antigen (type O saliva). The Y axe in panel B indicates the OD value of the saliva-binding assay. The X axe indicates three mutants: LEI/MNL mutant (1), VEN/FQT mutant (2), and RN/TT mutant (3), and followed by the MOH wild type capsid (4).

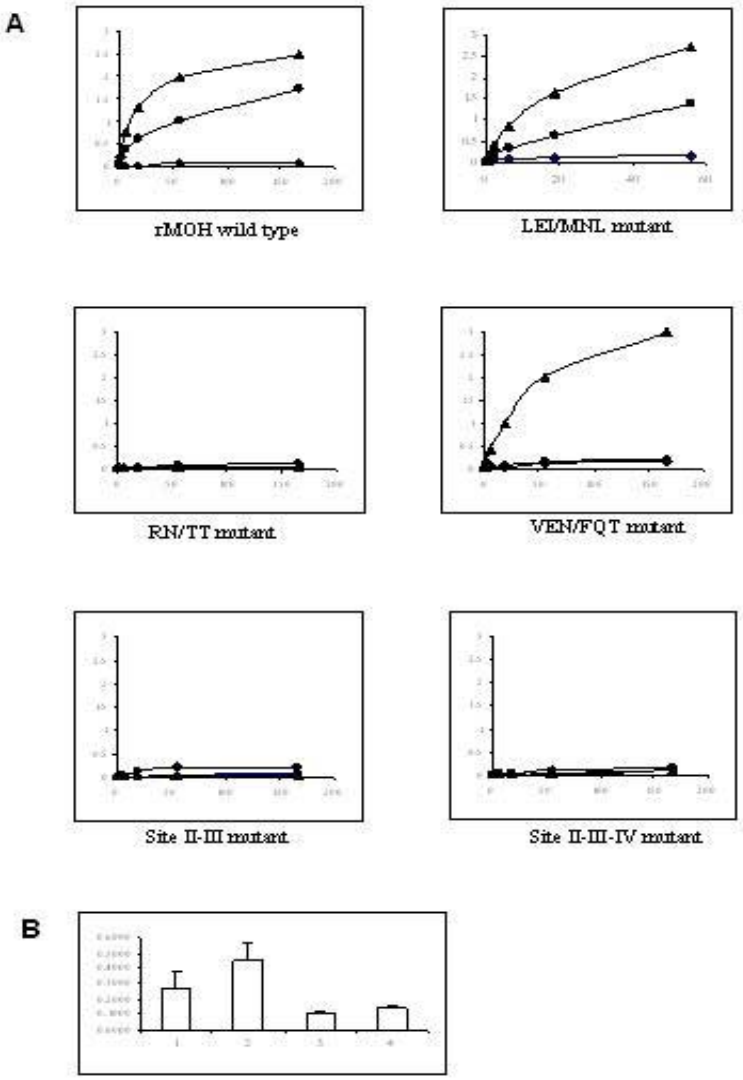


Fig 7

Fig. 8. Binding curves of mutants from VA387 (A) and MOH (B) with mutations related to NGR motif. The X and Y axes are the same as Fig. 5. Data were averaged from at least two independent experiments. ●: A antigen; ▲: B antigen; ♦: H antigen (type O saliva).

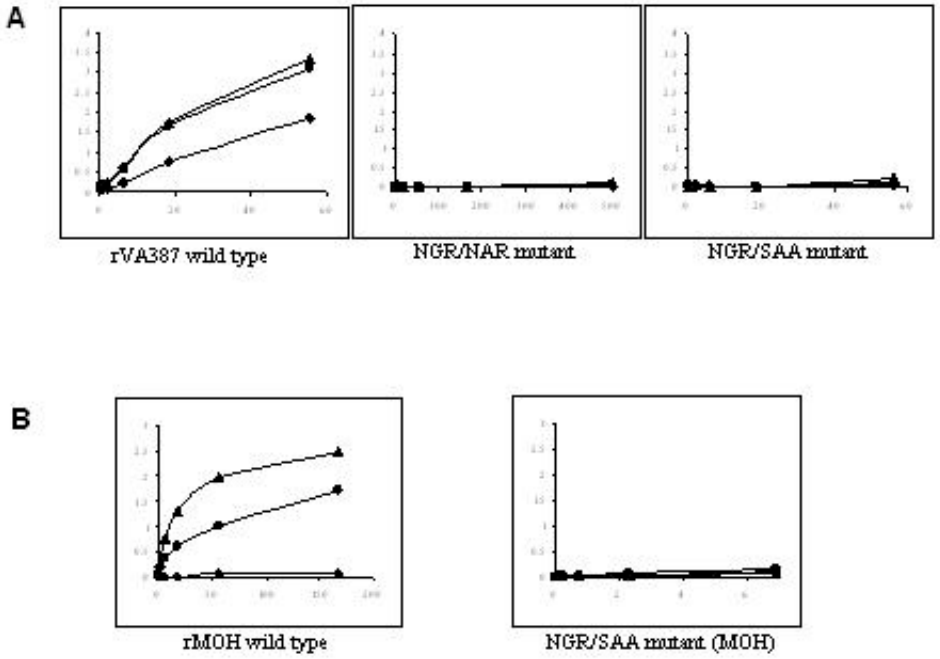


Fig. 8

REFERENCES

1. **Bertolotti-Ciarlet, A., L. J. White, R. Chen, B. V. Prasad, and M. K. Estes.** 2002. Structural requirements for the assembly of Norwalk virus-like particles. J Virol **76:**4044-55.
2. **Blackwell, C. C., F. Z. Aly, V. S. James, D. M. Weir, A. Collier, A. W. Patrick, C. G. Cumming, D. Wray, and B. F. Clarke.** 1989. Blood group, secretor status and oral carriage of yeasts among patients with diabetes mellitus. Diabetes Res **12:**101-4.
3. **Blackwell, C. C., S. J. May, R. P. Brettle, C. J. MacCallum, and D. M. Weir.** 1987. Secretor state and immunoglobulin levels among women with recurrent urinary tract infections. J Clin Lab Immunol **22:**133-7.
4. **Boren, T., P. Falk, K. A. Roth, G. Larson, and S. Normark.** 1993. Attachment of Helicobacter pylori to human gastric epithelium mediated by blood group antigens. Science **262:**1892-5.
5. **Boren, T., S. Normark, and P. Falk.** 1994. Helicobacter pylori: molecular basis for host recognition and bacterial adherence. Trends Microbiol **2:**221-8.
6. **Bower, M. J., F. E. Cohen, and R. L. Dunbrack, Jr.** 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol **267:**1268-82.
7. **Bresee, J., M. Widdowson, S. Monroe, and R. Glass.** 2002. Foodborne viral gastroenteritis: Challenges and opportunities. Clin Infect Dis **35:**748-53.
8. **Correa, P., and B. A. Schmidt.** 1995. The relationship between gastric cancer frequency and the ratio of gastric to duodenal ulcer. Aliment Pharmacol Ther **9:**13-9.
9. **Coulson, B. S., S. L. Londrigan, and D. J. Lee.** 1997. Rotavirus contains integrin ligand sequences and a disintegrin-like domain that are implicated in virus entry into cells. Proc Natl Acad Sci U S A **94:**5389-94.
10. **Fankhauser, R. L., S. S. Monroe, J. S. Noel, C. D. Humphrey, J. S. Bresee, U. D. Parashar, T. Ando, and R. I. Glass.** 2002. Epidemiologic and molecular trends of "Norwalk-like viruses" associated with outbreaks of gastroenteritis in the United States. J Infect Dis **186:**1-7.
11. **Farkas, T., T. Berke, G. Reuter, G. Szucs, D. O. Matson, and X. Jiang.** 2002. Molecular detection and sequence analysis of human caliciviruses from acute gastroenteritis outbreaks in Hungary. J Med Virol **67:**567-73.
12. **Glass, R. I., J. Holmgren, C. E. Haley, M. R. Khan, A. M. Svennerholm, B. J. Stoll, K. M. Belayet Hossain, R. E. Black, M. Yunus, and D. Barua.** 1985. Predisposition for cholera of individuals with O blood group, possible evolutionary significance. American Journal of Epidemiology **121:**791-796.
13. **Green, J., J. Vinje, C. I. Gallimore, M. Koopmans, A. Hale, and D. W. Brown.** 2000. Capsid protein diversity among Norwalk-like viruses. Virus Genes **20:**227-36.
14. **Harrington, P. R., L. Lindesmith, B. Yount, C. L. Moe, and R. S. Baric.** 2002. Binding of Norwalk virus-like particles to ABH histo-blood group antigens is blocked by antisera from infected human volunteers or experimentally vaccinated mice. J Virol **76:**12335-43.
15. **Hooton, T. M.** 2000. Pathogenesis of urinary tract infections: an update. J Antimicrob Chemother **46 Suppl 1:**1-7; discussion 63-5.
16. **Huang, P. W., T. Farkas, S. Marionneau, W. M. Zhong, N. Ruvoën-Clouet, L. K. Pickering, A. Morrow, M. Altaye, D. Newburg, J. LePendu, and X. Jiang.** 2003. Noroviruses bind to human ABO, Lewis and secretor histo-blood group antigens: Identification of four distinct strain-specific patterns. Jof Inf Dis **188: 19-31.**
17. **Hutson, A. M., R. L. Atmar, D. Y. Graham, and M. Estes.** 2002. Norwalk virus infection and disease is associated with ABO histo-blood group type. J Infect Dis **185:**1335-7.
18. **Hutson, A. M., R. L. Atmar, D. M. Marcus, and M. K. Estes.** 2003. Norwalk virus-like particle hemagglutination by binding to h histo-blood group antigens. Journal of Virology **77:**405-15.
19. **Jiang, X., D. Graham, K. Wang, and M. Estes.** 1990. Norwalk virus genome cloning and characterization. Science **250:**1580-1583.
20. **Jiang, X., P. W. Huang, W. M. Zhong, M. Tan, and T. Farkas.** Oligosaccharides with human histo-blood group antigen epitopes are involved in Norovirus binding. American Society for Virology Annual Meeting, University of California at Davis, July 12-16, 2003.
21. **Jiang, X., D. O. Matson, G. M. Ruiz-Palacios, J. Hu, J. Treanor, and L. K. Pickering.** 1995. Expression, self-assembly, and antigenicity of a Snow Mountain agent-like calicivirus capsid protein. J Clin Microbiol **33:**1452-5.
22. **Jiang, X., M. Wang, D. Y. Graham, and M. K. Estes.** 1992. Expression, self-assembly, and antigenicity of the Norwalk virus capsid protein. J Virol **66:**6527-32.
23. **Jiang, X., M. Wang, K. Wang, and M. K. Estes.** 1993. Sequence and genomic organization of Norwalk virus. Virology **195:**51-61.
24. **Jiang, X., W. M. Zhong, N. Wilton, T. Farkas, E. Barrett, D. Fulton, R. Morrow, and D. M. Matson.** 2002. Baculovirus expression and antigenic characterization of the capsid proteins of three Norwalk-like viruses. Arch Virol **147:**119-130.
25. **Kelley, L. A., R. M. MacCallum, and M. J. Sternberg.** 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol **299:**499-520.

26. **Klaamas, K., O. Kurtenkov, M. Ellamaa, and T. Wadstrom.** 1997. The Helicobacter pylori seroprevalence in blood donors related to Lewis (a,b) histo-blood group phenotype. Eur J Gastroenterol Hepatol **9:**367-70.
27. **Lindesmith, L., C. Moe, J. LePendu, X. Jiang, and R. Baric.** 2003. Determinants of Susceptibility and Protective Immunity to Norwalk Virus Infection. Nature Medicine **in press.**
28. **Liu, B. L., P. R. Lambden, H. Gunther, P. Otto, M. Elschner, and I. N. Clarke.** 1999. Molecular characterization of a bovine enteric calicivirus: relationship to the Norwalk-like viruses. J Virol **73:**819-25.
29. **Lomberg, H., U. Jodal, H. Leffler, P. De Man, and C. Svanborg.** 1992. Blood group non-secretors have an increased inflammatory response to urinary tract infection. Scand J Infect Dis **24:**77-83.
30. **Marionneau, S., A. Cailleau-Thomas, J. Rocher, B. Le Moullac-Vaidye, N. Ruvoen, M. Clement, and J. Le Pendu.** 2001. ABH and Lewis histo-blood group antigens, a model for the meaning of oligosaccharide diversity in the face of a changing world. Biochimie **83:**565-73.
31. **Marionneau, S., N. Ruvoen, B. Le Moullac-Vaidye, M. Clement, A. Cailleau-Thomas, G. Ruiz-Palacios, P. W. Huang, X. Jiang, and J. Le Pendu.** 2002. Norwalk virus binds to H typeS 1/3 histo-blood group antigens present on gastro-duodenal epithelial cells of "secretor" individuals. Gastroenterology **122.:**1967-1977.
32. **Mason, P. W., E. Rieder, and B. Baxt.** 1994. RGD sequence of foot-and-mouth disease virus is essential for infecting cells via the natural receptor but can be bypassed by an antibody-dependent enhancement pathway. Proc Natl Acad Sci U S A **91:**1932-6.
33. **Meller, J., and R. Elber.** 1998. Computer simulations of carbon monoxide photodissociation in myoglobin: structural interpretation of the B states. Biophys J **74:**789-802.
34. **Meller, J., and R. Elber.** 2001. Linear programming optimization and a double statistical filter for protein threading protocols. Proteins **45:**241-61.
35. **Oliver, S. T., A. M. Dastjerdi, S. Wong, L. El-Attar, C. Gallimore, D. W. Brown, J. Green, and J. C. Bridger.** 2003. Molecular characterization of bovine enteric caliciviruses: a distinct third genogroup of noroviruses (Norwalk-like viruses) unlikely to be of risk of humans. J. Virol. **77:**2789-98.
36. **Prasad, B. V., M. E. Hardy, T. Dokland, J. Bella, M. G. Rossmann, and M. K. Estes.** 1999. X-ray crystallographic structure of the Norwalk virus capsid. Science **286:**287-90.
37. **Prasad, B. V., R. Rothnagel, X. Jiang, and M. K. Estes.** 1994. Three-dimensional structure of baculovirus-expressed Norwalk virus capsids. J Virol **68:**5117-25.
38. **Raza, M. W., C. C. Blackwell, P. Molyneaux, V. S. James, M. M. Ogilvie, J. M. Inglis, and D. M. Weir.** 1991. Association between secretor status and respiratory viral illness. Bmj **303:**815-8.
39. **Ruoslahti, E., and M. D. Pierschbacher.** 1986. Arg-Gly-Asp: a versatile cell recognition signal. Cell **44:**517-8.
40. **Ruvoen-Clouet, N., J. P. Ganiere, G. Andre-Fontaine, D. Blanchard, and J. Le Pendu.** 2000. Binding of rabbit hemorrhagic disease virus to antigens of the ABH histo-blood group family. J Virol **74:**11950-4.
41. **Sali, A., and J. P. Overington.** 1994. Derivation of rules for comparative protein modeling from a database of protein structure alignments. Protein Sci **3:**1582-96.
42. **Sidebotham, R. L., J. H. Baron, J. Schrager, J. Spencer, J. R. Clamp, and L. Hough.** 1995. Influence of blood group and secretor status on carbohydrate structures in human gastric mucins: implications for peptic ulcer. Clin Sci (Colch) **89:**405-15.
43. **Stapleton, A., E. Nudelman, H. Clausen, S. Hakomori, and W. E. Stamm.** 1992. Binding of uropathogenic Escherichia coli R45 to glycolipids extracted from vaginal epithelial cells is dependent on histo-blood group secretor status. J Clin Invest **90:**965-72.
44. **Stroud, M. R., A. E. Stapleton, and S. B. Levery.** 1998. The P histo-blood group-related glycosphingolipid sialosyl galactosyl globoside as a preferred binding receptor for uropathogenic Escherichia coli: isolation and structural characterization from human kidney. Biochemistry **37:**17420-8.
45. **Sugieda, M., and S. Nakajima.** 2002. Viruses detected in the caecum contents of healthy pigs representing a new genetic cluster in genogroup II of the genus "Norwalk-like viruses". Virus Research **87:**165-72.
46. **White, L. J., M. E. Hardy, and M. K. Estes.** 1997. Biochemical characterization of a smaller form of recombinant Norwalk virus capsids assembled in insect cells. J Virol **71:**8066-72.
47. **Wickham, T. J., M. E. Carrion, and I. Kovesdi.** 1995. Targeting of adenovirus penton base to new receptors through replacement of its RGD motif with other receptor-specific peptide motifs. Gene Ther **2:**750-6.

PART IV:

Appendices

Selected acronyms and definitions:

Ab initio methods

A class of computational protocols for finding the overall three-dimensional structure of a protein from its sequence. *Ab initio* or *de novo* simulations attempt to reproduce the actual folding process, without using similarity to known protein structures.

Docking

Binding of a ligand to a specific binding site of a receptor molecule.

Dynamic Programming (DP)

A classical computer science technique for solving (in polynomial time) certain class of combinatorial optimization problems that are characterized by an exponentially scaling search space. Dynamic programming is widely used in bioinformatics as a tool to find optimal sequence alignments.

Fold recognition

A similarity-based approach to protein structure prediction, which assigns new proteins to previously characterized protein families by using sequence-to-structure matching.

Force field

A certain functional form of the potential energy of a system of interacting atoms with the parameters derived from *ab initio* calculations and experimental data.

Interior Point (IP) methods

A class of algorithms and techniques to solve Linear Programming (and in general convex optimization) problems. As opposed to the simplex algorithm, the IP methods are guaranteed to find the optimal solution in polynomial number of steps.

Linear Programming (LP)

A procedure for finding the maximum or minimum of a linear function where the arguments are subject to linear constraints. The simplex method is one well known algorithm for solving LP problems.

Molecular Dynamics (MD)

A technique for atomistic simulations of complex systems in which the time evolution of the system is followed using numerical integration of the equations of motion.

Monte Carlo (MC)

A simulation technique for conformational sampling and optimization based on a random search for energetically favourable conformations.

Threading Onion Model (THOM)

A class of scoring functions for sequence-to-structure matching (also known as threading potentials) considered in this dissertation.

Recent peer reviewed articles by the author of the thesis:

J. Meller and R. Elber; *Computer Simulations of Carbon Monoxide Photodissociation in Myoglobin: Structural Interpretation of the B states*, **Biophysical Journal**, 74, 789-802 (1998)

M. Turowski, N. Yamakawa, **J.Meller**, K. Kimata, T. Ikegami, K. Hosoya and N. Tanaka; *Deuterium isotope effect in chromatography as examined by various HPLC systems. Comments on the retention and H/D differentiation mechanism*, **Chromatography**, 19 (1998)

R. Elber, **J. Meller** and R. Olender; *Stochastic Path Approach to Compute Atomically Detailed Trajectories: Application to the Folding of C Peptide*, **Journal of Physical Chemistry B**, 103, 899-911 (1999)

J. Meller and W. Duch; *SGA derivation of matrix elements between spin-adapted perturbative wavefunctions*, **International Journal of Quantum Chemistry**, 74, 123-133 (1999)

A. Frary, T. C. Nesbitt, A. Frary, S. Grandillo, E. van der Knaap, B. Cong, J. Liu, **J. Meller**, R. Elber, K. B. Alpert, S. D. Tanksley; *fw2.2: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size*, **Science**, 289: 85-88 (2000)

Wan J, **Meller J**, Hada M, Ehara M, Nakatsuji H; *Electronic excitation spectra of furan and pyrrole: Revisited by the symmetry adapted cluster-configuration interaction method*, **Journal of Chemical Physics**, 113: 7853-7866 (2000)

Meller J, Elber R.; *Linear Optimization and a Double Statistical Filter for Protein Threading Protocols*, **Proteins, Structure, Function, and Genetics**, 45: 241-261 (2001)

J. Meller, Elber R; *Protein Recognition by Sequence-to-Structure Fitness: Bridging Efficiency and Capacity of Threading Models*, in Computational Methods for Protein Folding: A Special Volume of **Advances in Chemical Physics**, ed. R. A. Friesner, John Wiley & Sons 2002

J. Meller, Wagner M, Elber R; *Maximum Feasibility Guideline to the Design and Analysis of Protein Folding Potentials*, **Journal of Computational Chemistry**, 23: 111-118 (2002)

M. Wagner, **J. Meller** and R. Elber; *Large-Scale Linear Programming Techniques for the Design of Protein Folding Potentials*, **Mathematical Programming**, to appear (2003)

A. V. Kuznetsova, **J. Meller**, P. O. Schnell, J. A. Nash, Y. Sanchez, J. W. Conaway, R. C. Conaway and M. F. Czyzyk-Krzeska; *VHL binds hyperphosphorylated large subunit of RNA Polymerase II through a proline hydroxylation motif and targets it for ubiquitination*, **PNAS** vol. 100 (5), 2706-2711 (2003)

J. Meller, J. P. Malrieu and J. L. Heully; *Size-consistent multireference CI through the dressing of the norm of determinants*, **Molecular Physics** vol. 101 (13), 2029-2041 (2003)

J. Meller; *Molecular Dynamics*, in **Encyclopedia of the Human Genome**
Nature Publishing Group, Macmillan Publishers Ltd 2003

M. Tan, P. Huang, **J. Meller**, W. Zhong, T. Farkas and X. Jiang; *Mutations within P2 Domain of Norovirus Capsid Affect Binding to Human Histo-Blood Group Antigens: Evidence for a Binding Pocket*, **Journal of Virology**, to appear (2003)

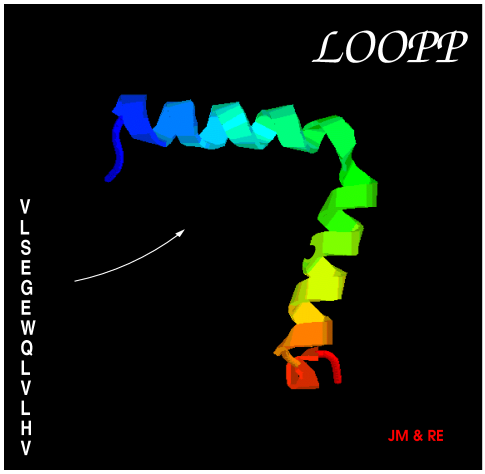
M. Turowski, N. Yamakawa, **J. Meller**, K. Kimata, T. Ikegami, K. Hosoya, N. Tanaka and E.R. Thornton; *Deuterium Isotope Effects on Hydrophobic Interactions. The Importance of Dispersion Interactions in the Hydrophobic Phase*, **Journal of American Chemical Society**, to appear (2003)

A. Porollo, R. Adamczak, M. Wagner and **J. Meller**; *Maximum Feasibility Approach for Consensus Classifiers: Applications to Protein Structure Prediction*, CIRAS 2003, accepted

S.-E. Olufemi, P. M. Snyder, K. L. Smith, Y. R. Su, M. C. Reif, R. Adamczak, **J. Meller** and A. G. Menon; *Polymorphic Variants Alter Function of the Human Epithelial Sodium Channel α-subunit: Evidence for a Role in Hypertension,* submitted

R. Adamczak and **J. Meller**; *On the Transferability of Folding and Threading Potentials and Sequence-Independent Filters for Protein Folding Simulations*, submitted

M.-D. Filippi, C. E. Harris, **J. Meller,** Y. Zheng and D. A. Williams; *Sequence Determinants of Rho GTPase Specification of Superoxide Generation and Directed Migration in Mammalian Cells*, submitted



Learning, Observing and Outputting Protein Patterns (LOOPP)

LOOPP is a program for **PROTEIN RECOGNITION** and design of **PROTEIN FOLDING** potentials.

LOOPP performs **sequence-to-sequence**, **sequence-to-structure** (threading), and **structure-to-structure** alignments. It further enables the optimization of potentials and scoring functions for the above applications. One may also use LOOPP to generate non-redundant libraries of folds, both for the training and recognition. A number of contact models is implemented in LOOPP, including continuous and step-wise pairwise potentials and several profile models, such as THOM1 and THOM2. LOOPP can be used to optimize parameters for these models using Linear Programming based protocols (however an external LP solver must be used). Gapless threading may be used in order to generate large samples of decoy structures for training. Alternatively, knowledge-based, statistical potentials may be computed. LOOPP was extensively used to generate many of the results discussed in this dissertation. LOOPP can be downloaded at http://www.tc.cornell.edu/CBIO/loopp.