Towards Predicting Coiled-Coil Protein Interactions

Mona Singh
Department of Computer Science
Princeton University
msingh@cs.princeton.edu

Peter S. Kim
Howard Hughes Medical Institute
Whitehead Institute for Biomedical Research
Department of Biology
Massachusetts Institute of Technology
kimadmin@wi.mit.edu

ABSTRACT

Protein-protein interactions play a central role in many cellular functions, and as whole-genome data accumulates, computational methods for predicting these interactions become increasingly important. Computational methods have already proven to be a useful first step for rapid genome-wide identification of putative protein structure and function, but research on the problem of computationally determining biologically relevant partners for given protein sequences is just beginning. In this paper, we approach the problem of predicting protein-protein interactions by focusing on the 2-stranded coiled-coil motif. We introduce a computational method for predicting coiled-coil protein interactions, and give a novel framework that is able to use both genomic sequence data and experimental data in making these predictions. Cross-validation tests show that the method is able to predict many aspects of protein-protein interactions mediated by the coiled-coil motif, and suggest that this methodology can be used as the basis for genome-wide prediction of coiled-coil protein interactions.

RECOMB 2001, Montreal, Canada

1. INTRODUCTION

Protein-protein interactions play a central role in many cellular functions, including DNA replication, transcription and translation, signaling cascades, metabolic pathways, and protein trafficking and secretion. Since a genome contains a complete "parts list" of an organism, whole-genome data allows one to begin to address exhaustively the problem of determining and predicting which proteins can interact with each other. Traditionally, protein-protein interactions have been determined using biochemical and genetic experiments; however, as whole-genome data accumulates, it becomes increasingly important to develop computational methods for predicting these interactions. Computational methods have already proven to be a useful first step for rapid genome-wide identification of putative protein structure and function, but research on the problem of computationally determining biologically relevant partners for a given protein sequence is just beginning.

The difficulty of the general protein structure prediction problem precludes structure-based prediction of all protein-protein interactions. Our approach to this problem focuses on a specific, well-characterized structural motif that mediates protein-protein interactions: the parallel, 2-stranded coiled coil. Predicting protein-protein interactions mediated by the coiled-coil motif is an important problem, as coiled coils are found in proteins involved in transcription, in cell-cell and viral-cell fusion events, and in maintaining the structural identity of cells. Coiled-coil interactions are known to be specific [19, 32, 41], and coiled coils are also quite common–they are predicted to comprise 3%–5% of sequence databases [25, 42]. More generally, the methods we develop for this problem may influence computational methods for prediction of protein-protein interactions mediated by other structural motifs. Protein-protein interactions can sometimes be predicted without structural information by exploiting information gleaned from multiple fully sequenced genomes [35, 26, 8]; the structural approach outlined here can be augmented by these non-structural, cross-genomic approaches.

The coiled-coil motif consists of two or more right-handed α-helices wrapped around each other with a slight left-handed superhelical twist. These helices may associate with each other in a parallel or anti-parallel orientation, and the sequences making up the helices may either be the same (homo-oligomers) or different (hetero-oligomers). Coiled coils show

a characteristic heptad repeat $(\mathbf{abcdefg})_\mathbf{n}$ spread out along two turns of the helix, with positions **a** and **d** containing generally hydrophobic residues, and positions **e** and **g** containing generally charged residues (see Figure 1). Their simple, repeating structures make them particularly amenable to computational methods, and several methods have been developed that are effective in identifying potential coiled-coil strands within *single* protein sequences [34, 25, 2, 42, 1, 39, 38].

These methods cast the coiled-coil recognition problem within a probabilistic framework, and use databases of known coiled-coil sequences to tabulate frequencies of amino acids appearing in particular heptad repeat positions. Several of these methods use *intrahelical* pairwise correlations between residues in coiled coils, but up to this point none has explicitly used *interhelical* correlations within coiled-coil structures. All of these methods predict the likelihood that a single protein sequence is part of a coiled-coil structure; none of these methods predict whether a particular protein pair (or triplet, etc.) is likely to form a coiled-coil structure. Since the coiled coil is one of the few structural motifs for which effective prediction methods exist, it is a natural motif for which to begin development of computational methods for predicting protein-protein interactions. In the past, several groups have counted the number of favorable and unfavorable electrostatic interactions to make some specific predictions about the nature of particular coiled-coil protein-protein interactions [33, 27, 41]; however, it is known that many other factors play a role in coiled-coil specificity (e.g., [32, 24, 14]) and thus such simple approaches are limited in their applicability.

In this paper, we present a computational method for predicting whether a particular pair of protein sequences can form a coiled-coil structure. Unlike previous methods that cast coiled-coil recognition within a probabilistic framework, here coiled coils are explicitly represented in terms of the interhelical pairwise interactions that make them up. This simple switch in representation captures many of the essential structural features of coiled coils, and allows coiled-coil sequence data to be easily augmented with experimentally derived information about coiled coils. We give an optimization method based on both sequence and experimental data that computes a weight for each possible interhelical interaction. Intuitively, for each possible interhelical interaction, the method computes the corresponding "weight" that represents how favorable the interhelical interaction is. The score for any coiled-coil structure is then the sum of the computed weights corresponding to each interhelical interaction making up the coiled coil.

A novel feature of this method is the ability to use both coiled-coil sequence data and coiled-coil experimental data. Since the coiled-coil motif is well-studied and much has been determined about coiled-coil specificity in experimental work (e.g., see [32, 24, 14, 41, 11, 10]), it is advantageous to have a method that incorporates this knowledge. On the other hand, not all pairwise interactions of interest have been studied experimentally, and taking advantage of statistical features evident in sequence databases is also necessary. Additionally, the method can tolerate some amount of error, either in the experimental data or in the way in which sequence data is used.

Since potential coiled-coil strands can be identified efficiently at the genome level [25, 2, 42], and in particular, 2-stranded coiled-coil regions can be distinguished [42], ultimately we would like to use our method to make genome-wide predictions of coiled-coil protein interactions. As a first step, we show that the method is able to predict many aspects of protein-protein interactions mediated by the coiled-coil motif. In particular, the general problem of coiled-coil partner prediction can be broken down into the following easier subproblems:

- **Predicting helix-alignment:** given two coiled-coil sequences that are known to partner, predict how the helices align with each other. That is, two coiled-coil regions may interact with each other in several shifts, and we would like to predict, for example, which **a** position in one helix is across the coiled-coil interface from a given **a** position in the other helix.
- **Heterodimeric preferences:** given two coiled-coil sequences, predict whether the heterodimer formed by these sequences is preferred over the two corresponding homodimers. Heterodimer preferences are measured by comparing the heterodimer species with the two corresponding homodimer species; in some cases, heterodimer preferences exist because one of the two homodimer species is very unfavorable (e.g., for the fos-jun heterodimer, the instability of the fos homodimer drives heterodimer formation [32]).
- **Partner elimination:** given a coiled-coil sequence, eliminate the vast majority of possible coiled-coil partners. The ultimate goal is to get rid of all incorrect possibilities; for this work, we try to quantify how well we do in this regards.

Cross-validation tests show that our method performs well on the above problems; we know of no other method that has the type of performance we obtain. Since solutions to the above subproblems are clearly necessary for solving the general coiled-coil partner prediction problem, our methodology makes substantial progress on the problem of predicting coiled-coil protein interactions.

2. METHODS

In this section, we describe a framework for making predictions about coiled-coil protein interactions. First, we show how structural features important for coiled-coil specificity are used to motivate a representation of coiled coils based on interhelical interactions. Second, we show how this representation allows incorporation of experimental information about coiled-coil partnering specificity. Third, we show how to include coiled-coil sequence data into this framework. Finally, we show how this framework can handle error, either in the experimental data or in the way in which sequence data is used.

The methodology presented below is significantly different from previous computational methods for coiled-coil recognition [34, 25, 2, 42, 1, 39, 38]. Unlike the methodology

presented in this paper, these previous methods have all viewed coiled-coil recognition within a probabilistic framework. Additionally, none of these methods has attempted to predict coiled-coil protein-protein interactions, and none provides an obvious way to incorporate experimental information.

Representing coiled coils. The interface between α-helices in a parallel, 2-stranded coiled coil is formed by residues at the core positions **a**, **d**, **e** and **g** (see Figure 1). In particular, crystal structures of parallel 2-stranded coiled coils [31, 9] have revealed that the side chains in the interaction interface display knobs-into-holes packing [7]. That is, the side chain of i-th heptad **a** position residue of one helix packs into a "hole" surrounded by the i-th heptad **a** position residue, the $i-1$-st heptad **g** position residue, and the $i-1$-st and i-th **d** position residues of the other helix; and the side chain of i-th heptad **d** position residue of one helix packs into a "hole" surrounded by the i-th heptad **d** position residue, the i-th heptad **e** position residue, and the i-th and $i+1$-st **a** position residues of the other helix. Additionally, the solvent-exposed portion of the interaction interface consists of side chains in the **g** and **e** positions, with the side chain in the i-th heptad **g** position of one helix possibly interacting with the side chain in the $i+1$-st heptad **e** position of the other helix.

These structural features motivate the simplifying assumption that all coiled-coil interhelical interactions can be captured by interactions between these four core positions. Considering just these core position residues is also a reasonable assumption given what is known experimentally about dimeric coiled-coil specificity; for example, see [32]. The further assumption is made that considering pairwise interactions is sufficient.[1] With these two assumptions, each coiled coil can thus be represented by the pairwise interhelical interactions that make it up. That is, each coiled coil can be represented by a vector $\vec{x}$, where $x_{(p,q),i,j}$, is the number of times residues i and j appear across the helical interface in positions p and q respectively.

Note that the interhelical pairwise positions (p and q) to be considered can be chosen based on what interactions are thought to be important for coiled coils. In particular, based on structural features of the coiled-coil interhelical interface [31, 9] as well as experiments on determinants of coiled-coil specificity [32, 24, 41], the following pairwise positions are used: (i)**g**-$(i+1)$**e'** (that is, the interaction between the **g** position in the i-th heptad of one helix with the **e** position in the $i+1$-st heptad of the other helix), (i)**g'**-$(i+1)$**e**, (i)**g**-$(i+1)$**a'**, (i)**g'**-$(i+1)$**a**, (i)**d**-(i)**e'**, (i)**d'**-(i)**e**, (i)**d**-$(i+1)$**a'**, (i)**d'**-$(i+1)$**a**, (i)**a**-(i)**d'**, and (i)**a'**-(i)**d**.[2] Note that pairs of these interactions are symmetric (e.g., (i)**g**-$(i+1)$**e'** and (i)**g'**-$(i+1)$**e** represent the same interhelical interaction). Thus, five different pairwise interactions are used, and since 20 amino acids are possible for each position, coiled coils are represented by vectors of dimensionality $5 \cdot 20 \cdot 20 = 2000$.

Suppose that for each possible interaction $x_{(p,q),i,j}$, the corresponding "weight" $w_{(p,q),i,j}$ that represents how favorable the interhelical interaction is known. Then for a particular coiled coil represented by $\vec{x}$, its *score* is given by $\vec{w} \cdot \vec{x}$. Of course, initially this weight vector $\vec{w}$ is not known, and we show below constraints that this vector $\vec{w}$ should satisfy.

Constraints from experimental information. This representation of coiled coils allows us to state rigorously certain types of experimental constraints. For example, if coiled coil $\vec{x}$ is more "favorable" (e.g., more stable) than coiled coil $\vec{y}$, then we would like weight vector $\vec{w}$ to satisfy:

$$\vec{w} \cdot \vec{x} \quad > \quad \vec{w} \cdot \vec{y} \tag{1}$$

Furthermore, this representation allows knowledge about specific weight elements to be incorporated. As an example, say that it is more favorable to have a Lysine in a **g** position in one helix with a Glutamic Acid in the following position **e** in the other helix than it is to have Glutamic Acid in both these positions (i.e., **g-e** K E is "better than" **g-e** E E). Then the following should be true:

$$w_{(g,e),K,E} \quad > \quad w_{(g,e),E,E} \tag{2}$$

Constraints from sequence databases. Given databases of sequences that form dimeric coiled coils and sequences that do not, we can further constrain the weight vector $\vec{w}$. In particular, we would like coiled coils to have a higher score than non-coiled-coils. One way to constrain the weight vector $\vec{w}$ to satisfy this is to require that:

$$\vec{w} \cdot \vec{x} \quad > \quad 0, \text{ for all coiled coils } \vec{x} \tag{3}$$

$$\vec{w} \cdot \vec{y} \quad < \quad 0, \text{ for all non-coiled-coils } \vec{y} \tag{4}$$

Allowing errors and determining the weight vector. Note that without loss of generality, each of the above constraints i (equations 1, 2, 3 and 4) can be rewritten in vector notation $\vec{z}^{(i)}$ such that $\vec{w}$ is constrained to satisfy $\vec{w} \cdot \vec{z}^{(i)} > 0$ for each $\vec{z}^{(i)}$.

Since there may be errors in either the sequence data or the experimental data, we would like to relax these constraints on $\vec{w}$. Errors can be allowed by adding additional variables for each $\vec{z}^{(i)}$ as follows:

$$\vec{w} \cdot \vec{z}^{(i)} \quad \geq \quad -\epsilon_i \tag{5}$$

$$\epsilon_i \quad \geq \quad 0 \tag{6}$$

The goal now is to find $\vec{w}, \vec{\epsilon}$ to minimize $\sum \epsilon_i$ such that each constraint (given in equations 5 and 6) is satisfied. Thus,

[1] Pairwise interactions were sufficient in earlier work on coiled-coil recognition [2, 1, 42]. Note also that theoretically we can consider more than pairs of amino acids at a time; the major difficulty is in the size of our coiled-coil databases.

[2] Note that (i)**a**-(i)**a'** and (i)**d**-(i)**d'** are very important components of the interhelical interface (e.g., [14, 24]). However, since our database is biased towards homodimeric coiled coils, and homodimeric coiled coils have identical resides in the analogous **a** positions across the interhelical interface (as well as identical residues in the analogous **d** positions across the interhelical interface), the described methodology would not be able to handle these positions effectively.

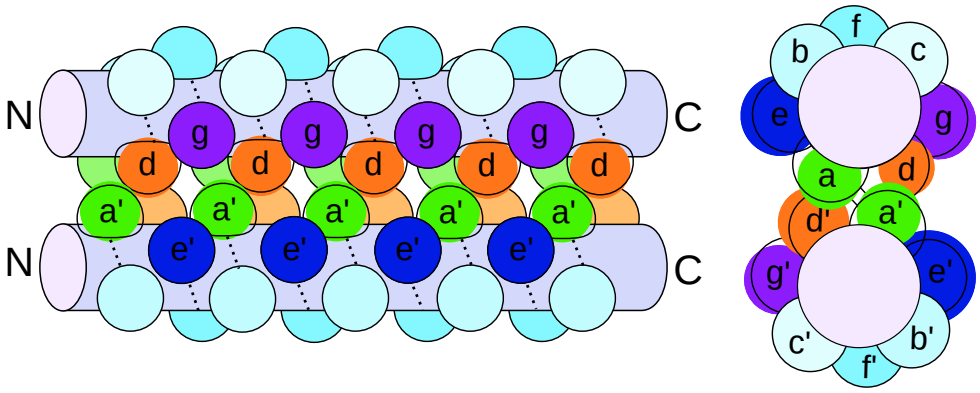

Figure 1: (a) Side view of a parallel 2-stranded coiled coil. (b) Top view of a parallel 2-stranded coiled coil. The interface between the α-helices in a coiled-coil structure is formed by residues at the core positions a, d, e and g. For notational convenience, positions in the two helices are distinguished by the prime notation (e.g., a and a' are analogous positions in the two helices).

by searching for such a $\vec{w}$, we obtain a set of weights corresponding to each relevant interhelical interaction in a coiled coil structure.

Tradeoffs between generalization and error led us to the theory of SVMs (Support Vector Machines) [40, 4], and thus for a chosen value of C, we actually minimized:

$$\frac{1}{2} \parallel \vec{w} \parallel^2 + C(\sum \epsilon_i) \tag{7}$$

subject to

$$\vec{w} \cdot \vec{z}^{(i)} \geq 1 - \epsilon_i \quad \forall i \tag{8}$$

$$\epsilon_i \geq 0 \quad \forall i \tag{9}$$

In SVMs, the first term ($\frac{1}{2} \parallel \vec{w} \parallel^2$) avoids overfitting, the second term ($\sum \epsilon_i$) is the error term, and the value of C determines the tradeoff between generalization and error [40, 4]. The SVM-lite package of [21] was used to implement portions of this work, with $C = 1$ for the testing described below.

Note that SVMs allow one to work efficiently in higher dimensional spaces using kernel functions, but in this work we assume that all coiled-coil interhelical interactions can be captured by certain pairwise interactions, and projection into a higher dimensional space violates this assumption. Nevertheless, in future work, it may be worthwhile to relax this assumption and to experiment with higher dimensional spaces.

The framework described is independent of the optimization method chosen, and it may be possible to improve performance by experimenting with different methods of optimization. For example, note that instead of optimizing a quadratic program as given in equation 7, it is possible to restate the problem as a linear program; for example, see [3].

Sequence and experimental data. For the work described here, we have built two sequence databases. The first is a homodimeric database consisting of myosin, tropomyosin, and types III and V intermediate filament (IF) proteins; this database was extracted from the approximately 58,000 residue 2-stranded database of [2]. The second is a heterodimeric database (6650 residues) built of keratin pairs that are known to interact with each other [36, 29, 15, 37, 6]. The two databases consist of approximately 250 protein sequences. For all these sequences, the coiled-coil strands are in exact axial register (review, [5]), so the correct structural alignment of the helices with respect to each other is known (i.e., for each coiled coil, it is known which **a** position in one helix is across from each **a** position in the other helix). Each coiled-coil region in these databases adds a constraint of the type in equation 3.

The non-coiled-coils necessary for our developed methodology (equation 4) are constructed in two ways. First, strands that are known to interact with each other are paired together in an incorrect way; that is, a non-coiled-coil is constructed such that the relative shift of the two helices with respect to each other is not seen in the actual coiled-coil structure. Second, strands from different protein families are paired together. While both these methods of constructing non-coiled-coils may result in some pairs that could form partners *in vitro*, coiled-coil interactions are known to be quite specific (e.g., see [19, 41]), and our methodology allows for some errors.

In addition to sequence databases, we use information from various biophysical studies that took a coiled-coil host system, mutated amino acids, and determined melting temperatures. Although the melting temperatures from different studies are often incomparable (e.g., due to different experimental conditions or constructs), each study provides a rank ordering of the stabilities of the coiled coils considered, and thus constraints of the type in equation 1 can be introduced. Data from the following studies were incorporated. O'Shea *et al.* determined melting temperatures for peptides consisting of the coiled-coil regions of GCN4, c-fos, and c-jun transcription factors, as well as several core position mutants [32]. Using the bZIP homodimeric protein chicken VBP as a host molecule, Krylov *et al.* measured melting temperatures for **g** and **e** position mutants, and Moitra *et al.* measured melting temperatures for **d** position mutants [23, 22, 28]. Jelesarov and Bosshard [20] determined melting

temperatures for designed peptides, with Alanine mutations in the **d** positions.

Experimental data from some genetic approaches were also used. Here, GCN4 leucine zipper sequences were randomized in core positions, and the resulting mutants were assayed using λ repressor-zipper fusions in which repressor function depends on oligomerization mediated by the leucine zipper [18, 17, 43]. Thus, this assay determines whether a particular mutant GCN4 coiled coil forms or not. In [18, 17], functional mutants were not assayed for oligomerization state, so only the non-functional mutant data is used.

3. RESULTS

In this section, we show that our methodology is able to predict many aspects of protein-protein interactions mediated by the coiled-coil motif. We begin by explaining the cross-validation set up, and then describe the performance of our method on three subproblems of the general problem of coiled-coil partner prediction. Solutions to these subproblems are clearly necessary for solving the general coiled-coil partner prediction problem, and we know of no other method that has the type of performance we obtain.

Cross-validation set up. In order to do cross-validation experiments, the dimeric coiled-coil sequence database was split into 11 different groups, based on the fact that our database consists of 11 different homologous coiled-coil regions: keratins, type V IFs segment 1B, type III IFs segment 1B, type III IFs segment 2B, two tropomyosin groups (corresponding to rabbit skeletal muscle tropomyosin residues 15–182, and 196–269) and five myosin groups (corresponding to nematode myosin residues 864–1160, 1212–1386, 1409–1583, 1606–1808, 1831–1929). The regions were chosen as described in [2]. In this manner, we ensure that there is no significant sequence similarity between sequences in different groups. Thus, to test the method, we applied the optimization to all the data resulting from 10 out of the 11 groups, as well as all the experimental weight constraint information described above, and then tested on the group left out.

Once the weights have been determined, scoring and comparing all possible coiled coils is straightforward. Namely, given two coiled-coil strands, each alignment of the two strands with each other specifies a vector of relevant pairwise interactions, and the weights corresponding to these interactions are simply summed up to give the interaction score.

Predicting helix-alignment. Given two coiled-coil regions that are known to partner with each other, is it possible to predict how the helices align with each other? That is, two coiled-coil regions may potentially interact with each other in several registers, and we would like to predict, for example, which **a** position in one helix is across the coiled-coil interface from a particular **a** position in the other helix. A solution to this problem is clearly necessary for the general partner prediction problem. Additionally, a solution to this problem can provide structural information about coiled coil protein-protein interactions determined in other ways (e.g., either obtained experimentally, as in [30], or computationally from various non-structural whole- and cross-genome methods [35, 26, 8]).

The weights were optimized in the manner just described, with 11 optimizations for the different cross-validation sets. We tested performance in the following way. Since the entire coiled-coil region can play a role in specificity, for each coiled-coil region in the database, performance was measured on progressively smaller subregions. In particular, given two coiled-coil regions A and B of length l in our database that are known to partner with each other, we began by considering two coiled-coil subregions of A: A_1 that is missing the first heptad of A, and A_2 that is missing the last heptad of A. Both A_1 and A_2 have length $l-7$. We now use the optimized weights to score A_1 and A_2 against subregions of B of length $l-7$. Note that since in coiled coils an **a** position residue in one helix is across from an **a** position residue in the other helix, there are only two appropriate subregions of B of length $l-7$: B_1 that is missing the first heptad of B, and B_2 that is missing the last heptad of B. (We required that all the $l-7$ residues of A_1 and A_2 be used for a technical reason: we wanted to compare scores of coiled-coil regions of the same length in order to avoid possible biases in our testing if the weights either favor lengthening or shortening of the coiled-coil interface.) Then, the predicted helix-alignment for A_1 is the subregion of B (B_1 or B_2) with which it has the higher interaction score. Thus, in this first simple case of testing where we are just removing one heptad, guessing gives a performance of 50% correct. To increase difficulty of the testing, we then progressively remove more heptads from sequence A (down to a minimum length of 35 residues, or 5 heptads), and see how these subregions align using the optimized weights with appropriate length subregions of B (and vice-versa).

Table 1 shows results with subregions of up to 6 heptads shorter. For each of the 11 cross-validation sets, the percent of correctly identified helix-alignments was calculated, and the average value over the cross-validation sets is reported. Overall performance starts with 97% and drops down to 90%. Since the entire coiled-coil region may play some role in specificity, some drop off in performance as progressively more heptads are removed is expected. In general, for genome-wide analysis, we will not know exact coiled-coil boundaries. Nevertheless, it is expected that current single-strand coiled-coil recognition methods [25, 2, 42] can detect dimeric coiled-coil boundaries within a heptad on each side, so this performance on helix-alignment is encouraging in terms of genome-wide application.

Heterodimer preferences. If we are given two coiled-coil sequences, can we predict whether the heterodimer formed by these sequences is preferred over the two corresponding homodimers? Heterodimer preferences are measured by comparing the heterodimer species with the two corresponding homodimer species; in some cases, heterodimer preferences exist because one of the two homodimer species is very unfavorable (e.g., for the fos-jun heterodimer, the instability of the fos homodimer drives heterodimer formation [32]). Again, a solution to this subproblem will be useful for the general partner prediction problem; in considering whether two strands are likely to partner with each other, it will be necessary to consider the two corresponding homodimers as well.

Heptads removed:	1	2	3	4	5	6
% alignments correct:	97	96	96	95	93	90
% performance if guessing:	50	33	25	20	17	14

Table 1: Helix-alignment performance of optimized weights. Percent of correctly identified helix-alignments is given as a function of the number of heptads removed. The percent of correctly identified helix-alignments given is the average performance over 11 cross-validation sets. Performance if guessing as a function of the number of heptads removed is also given for reference. In all cases, the method is performing significantly better than guessing.

Keratins are known to be preferential heterodimers [16]. For each pair of heterodimeric keratin coiled-coil partners, using the weights optimized with all constraints except those involving keratins, the score of the heterodimer pair was compared with the average of the scores of the two corresponding homodimers. In 25 out of 27 pairs considered, the score for the heterodimer is better. Additionally, for the fos-jun heterodimer, the optimized weights again predict preferential heterodimers. Thus, the weights resulting from our computational procedure have good performance in predicting heterodimer preferences.

Partner elimination. Given a coiled-coil sequence, can we eliminate the vast majority of possible coiled-coil partners? For each coiled-coil region A, we compared the score of A and its true partner with all of the scores of A and its possible partners. The possible partners were simply defined as the coiled-coiled subsequences of the same length as A, but in other cross-validation sets. We did not consider possible partners from the same cross-validation set since many of these sequences are close homologues that are likely to form coiled-coil partners *in vitro*. Note that a particular coiled-coil region B may contribute many potential coiled-coil partners for A, as all possible subsequences of the appropriate length were considered. As with the helix-alignment testing, we are only comparing interaction scores of coiled-coil regions of the same length. For all coiled-coil regions in our database, Table 2 shows the relative position of its true partner's score in comparison with the scores of all possible partners. In particular, we keep track of how often the true partner's score is in the top 5%, 10%, 25%, 30% and 50% of all possible partner scores.

Top 5%:	88.7%
Top 10%:	94.8%
Top 25%:	98.3%
Top 30%:	98.8%
Top 50%:	99.2%

Table 2: Partner elimination. The percent of coiled-coil regions in our database whose true partner's score is in the top 5%, 10%, 25%, 30% and 50% of all possible partner scores.

The ultimate goal in partner elimination is to get rid of all incorrect possibilities; here, we quantify how well we do in this regards. Our results here are very encouraging. For example, for 94.8% of the coiled-coil regions considered, the true partner's score was higher than the score of 90% of all possible partners. Pragmatically, this means that for 94.8% of coiled-coil regions in our database, if we eliminate the possible partners with interaction scores in the bottom 90%, the correct partner is still left. Thus, for most coiled-coil regions, the vast majority of its incorrect partners can be eliminated using the described methodology.

4. CONCLUSIONS

In this paper, we introduce a computational method for predicting coiled-coil protein interactions, and give a novel framework that is able to use both genomic sequence data and experimental data in making these predictions. Cross-validation tests show that the method is able to predict many aspects of protein-protein interactions mediated by the coiled-coil motif. In particular, for the coiled-coil regions in our database, almost 95% of the time, 90% of the incorrect partners can be eliminated. Ultimately, we would like to use our methodology to make genome-wide predictions of coiled-coil partnering interactions; that is, we would like to eliminate an even larger percent of incorrect partners. Obvious directions to pursue include looking at other optimization criterion, and trying to incorporate **a-a** and **d-d** interactions. The approach outlined here can also be augmented by other, non-structural, cross-genomic approaches [35, 26, 8], by more time-intensive coiled-coil molecular modeling approaches [13, 12], or by exploiting expression patterns of the protein sequences in consideration. While in some fraction of the cases it may be difficult to predict coiled-coil partners without molecular modeling calculations [13, 12], note that the sequence-based methods developed here get rid of the vast number of incorrect partner strands, and thereby allow time-intensive computations to be done on a more managable set. In conclusion, our methodology makes substantial progress on the problem of predicting coiled-coil protein interactions, and it is our hope that it will provide the basis of a system capable of predicting coiled-coil protein interactions at the genome level.

5. ACKNOWLEDGMENTS

We thank David Akey, Amy Keating, John Newman and Ethan Wolf for many helpful discussions.

6. REFERENCES

[1] B. Berger and M. Singh. An iterative method for improved protein structural motif recognition. *J. Computational Biol.*, 4(3):261–273, 1997.

[2] B. Berger, D. B. Wilson, E. Wolf, T. Tonchev, M. Milla, and P. S. Kim. Predicting coiled coils using pairwise residue correlations. *Proc. Natl. Acad. Sci. USA*, 92:8259–8263, 1995.

[3] B. Bradley and O. Mangasarian. Massive data discrimination via linear support vector machines. *Optimization Methods and Software*, 13(1):1–10, 2000.

[4] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[5] J. Conway and D. Parry. Structural features in the heptad substructure and longer range repeats of two-stranded α-fibrous proteins. *International Journal of Biological Macromolecules*, 12:328–333, October 1990.

[6] D. Cooper and T.-T. Sun. Monoclonal antibody analysis of bovine epithelial keratins. *Journal of Biological Chemistry*, 261:4646–4654, 1986.

[7] F. H. C. Crick. The packing of α-helices: simple coiled coils. *Acta Cryst.*, 6:689–697, 1953.

[8] A. Enright, I. Iliopoulos, N. Kyrpides, and C. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90, 1999.

[9] J. Glover and S. Harrison. Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature*, 373:257–261, 1995.

[10] L. Gonzalez, R. Brown, D. Richardson, and T. Alber. Crystal structures of a single coiled-coil peptide in two oligomeric states reveal the basis for structural polymorphism. *Nat. Struct. Biol.*, 3(12):1002–1009, 1996.

[11] L. Gonzalez, D. Woolfson, and T. Alber. Buried polar residues and structural specificity in the GCN4 leucine zipper. *Nat. Struct. Biol.*, 3(12):1011–1018, 1996.

[12] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim. High-resolution protein design with backbone freedom. *Science*, 282:1462–1467, 1998.

[13] P. B. Harbury, B. Tidor, and P. S. Kim. Predicting protein cores with backbone freedom: Structure prediction for coiled coils. *Proc. Natl. Acad. Sci. USA*, 92:8408–8412, 1995.

[14] P. B. Harbury, T. Zhang, P. S. Kim, and T. Alber. A switch between two-, three- and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science*, 262:1401–1407, November 1993.

[15] M. Hatzfeld and W. Franke. Pair formation and promiscuity of cytokeratins: Formation in vitro of heterotypic complexes and intermediate-sized filaments by homologous and heterologous recombinations of purified peptides. *Journal of Cell Biology*, 101:1826–1841, 1985.

[16] M. Hatzfeld and K. Weber. The coiled coil of in vitro assembled keratin filaments is a heterodimer of type I and type II keratins: Use of site-specific mutagenesis and recombinant protein expression. *J Cell Biol.*, 110:1199–1210, 1990.

[17] J. Hu, N. Newell, B. Tidor, and R. Sauer. Probing the roles of residues at the **e** and **g** positions of the GCN4 leucine zipper by combinatorial mutagenesis. *Protein Science*, 2:1072–1084, 1993.

[18] J. Hu, E. O'Shea, P. S. Kim, and R. Sauer. Sequence requirements for coiled coils: Analysis with lambda repressor-GCN4 leucine zipper fusions. *Science*, 250:1400–1403, 1990.

[19] H. Hurst. Transcription factors 1: bzip proteins. *Protein Profile*, 2(2):101–168, 1995.

[20] I. Jelesarov and H. R. Bosshard. Thermodynamic characterization of the coupled folding and association of heterodimeric coiled coils (leucine zippers). *J. Mol. Biol.*, 263:344–358, 1996.

[21] T. Joachims. Making large-scale SVM learning practical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods–Support Vector Learning*. MIT-Press, 1999.

[22] D. Krylov, J. Barchi, and C. Vinson. Inter-helical interactions in the leucine zipper coiled coil dimer: pH and salt dependence of coupling energy between charged amino acids. *J. Mol. Bio.*, 279:959–972, 1998.

[23] D. Krylov, I. Mikhailenko, and C. Vinson. A thermodynamic scale for leucine zipper stability and dimerization specificity: *e* and *g* interhelical interactions. *The EMBO journal*, 13(12):2849–2861, 1994.

[24] K. Lumb and P. S. Kim. A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry*, 34:8642–8648, 1995.

[25] A. Lupas, M. van Dyke, and J. Stock. Predicting coiled coils from protein sequences. *Science*, 252:1162–1164, 1991.

[26] E. Marcotte, M. Pellegrini, H. Ng, D. Rice, T. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285:751–753, 1999.

[27] A. McLachlan and M. Stewart. Tropomyosin coiled-coil interactions: Evidence for an unstaggered structure. *J. Mol. Biol.*, 98:293–304, 1975.

[28] J. Moitra, L. Szilak, D. Krylov, and C. Vinson. Leucine is the most stabilizing aliphatic amino acid in the *d* position of a dimeric leucine zipper coiled coil. *Biochemistry*, 36:12567–12573, 1997.

[29] R. Moll, W. Franke, D. Schiller, B. Geiger, and R. Krepler. The catalog of human cytokeratins: Patterns of expression in normal epithelia, tumors and cultured cell. *Cell*, 31:11–24, November 1982.

[30] J. R. Newman, E. Wolf, and P. S. Kim. A computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA*, 97(24):13203–13208, 2000.

[31] E. O'Shea, J. Klemm, P. S. Kim, and T. Alber. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science*, 254:539–544, October 1991.

[32] E. O'Shea, R. Rutkowski, and P. S. Kim. Mechanism of specificity in the fos-jun oncoprotein heterodimer. *Cell*, 68:699–708, 1992.

[33] D. A. D. Parry. Sequences of α-keratin: Structural implication of the amino acid sequences of the type I and type II chain segments. *J. Mol. Biol*, 113:449–454, 1977.

[34] D. A. D. Parry. Coiled coils in alpha-helix-containing proteins: analysis of residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins. *Bioscience Rep.*, 2:1017–1024, 1982.

[35] M. Pellegrini, E. Marcotte, M. Thompson, D. Eisenberg, and T. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, 96(8):4285–4288, 1999.

[36] R. Quinlan, C. Hutchison, and B. Lane. Intermediate filament proteins. *Protein Profile*, 2(8):795–952, 1995.

[37] D. Schiller, W. Franke, and B. Geiger. A subfamily of relatively large and basic cytokeratin polypeptides as defined by peptide mapping is represented by one or several polypeptides in epithelial cells. *The EMBO Journal*, 6:761–769, 1982.

[38] M. Singh, B. Berger, and P. S. Kim. Learncoil-VMF: Computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins. *J. Mol. Biol*, 290:1031–1044, 1999.

[39] M. Singh, B. Berger, P. S. Kim, J. Berger, and A. Cochran. Computational learning reveals coiled coil-like motifs in histidine kinase linker domains. *Proc. Natl. Acad. Sci. USA*, 95:2738–2743, March 1998.

[40] V Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[41] C. Vinson, T. Hai, and S. Boyd. Dimerization specificy of the leucine zipper-containing bzip motif on DNA binding: prediction and rational design. *Genes Dev.*, 7(6):1047–1058, 1993.

[42] E. Wolf, P. S. Kim, and B. Berger. Multicoil: A program for predicting two- and three-stranded coiled coils. *Protein Sci.*, 6:1179–1189, 1997.

[43] X. Zeng, H. Zhu, H. Lashuel, and J. Hu. Oligomerization properties of GCN4 leucine zipper **e** and **g** mutants. *Protein Science*, 6:2218–2226, 1997.