Predicting specificity in bZIP coiled-coil protein interactions

Citation: Fong, Jessica H, Amy E Keating, and Mona Singh. "Predicting specificity in bZIP coiled-coil protein interactions." Genome Biology 5, no. 2 (2004): R11.

As Published: http://dx.doi.org/10.1186/gb-2004-5-2-r11

Publisher: BioMed Central

Persistent URL: http://hdl.handle.net/1721.1/100901

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Method

Predicting specificity in bZIP coiled-coil protein interactions

Jessica H Fong*, Amy E Keating† and Mona Singh*

Addresses: *Computer Science Department and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Olden Street, Princeton, NJ 08544, USA. †Department of Biology, Massachusetts Institute of Technology, Massachusetts Avenue, Cambridge, MA 02139, USA.

Correspondence: Amy E Keating. E-mail: keating@mit.edu. Mona Singh. E-mail: msingh@princeton.edu

Published: 15 January 2004

Genome ***Biology*** 2004, **5:**R11

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/2/R11

Received: 22 September 2003
Revised: 21 November 2003
Accepted: 12 December 2003

Abstract

We present a method for predicting protein-protein interactions mediated by the coiled-coil motif. When tested on interactions between nearly all human and yeast bZIP proteins, our method identifies 70% of strong interactions while maintaining that 92% of predictions are correct. Furthermore, cross-validation testing shows that including the bZIP experimental data significantly improves performance. Our method can be used to predict bZIP interactions in other genomes and is a promising approach for predicting coiled-coil interactions more generally.

Background

High-throughput experimental techniques have recently begun to uncover protein-protein interactions at the proteomic scale [1-6]. As the number of fully-sequenced organisms grows, however, it becomes increasingly necessary to develop computational methods for predicting these interactions. The difficulty of computationally predicting protein structures suggests a strategy of concentrating first on interactions mediated by specific interfaces of known geometry.

In this article, we focus on a common and well-characterized protein interaction interface - the parallel two-stranded coiled coil. Coiled coils are found in proteins that participate in many diverse processes, including transcription, oncogenesis and membrane fusion; predicting protein-protein interactions mediated by this motif will have important biological ramifications. Coiled coils consist of two or more α-helices that wind around one another with a slight left-handed superhelical twist. A characteristic heptad repeat $(\mathbf{abcdefg})_n$ defines the placement of residues in each helix relative to the interaction interface (Figure 1). The buried positions **a** and **d** usually contain hydrophobic amino acids, and the more exposed positions **g** and **e** often contain charged and polar amino acids. This simple structure and periodicity permit recognition of potential coiled-coil sequences through statistical methods (for example, [7-12]), as well as detailed predictions of the structure and energetics of their hydrophobic interfaces through molecular modeling [13-15].

Much of what is known about the structure and specificity of parallel, two-stranded coiled coils has been ascertained through biophysical studies of peptides derived from bZIP transcription factors (for example, [16-22]). The coiled-coil regions of these proteins are also known as leucine zippers because the core **d** positions are dominated by leucine residues. bZIPs homo- and hetero-dimerize with each other via their coiled-coil regions. Despite readily apparent sequence homology, they exhibit a high degree of partnering selectivity that allows them to function in diverse pathways. Recently, protein array technology was used to determine coiled-coil interactions within a near-complete set of human bZIP transcription factors [23]. The interactions uncovered showed high reproducibility and excellent consistency with previously published studies, giving us high confidence in the data.

In this work, we apply a method for predicting coiled-coil interactions [24] to the human bZIP proteins. The array data provide an excellent opportunity to test our method and to

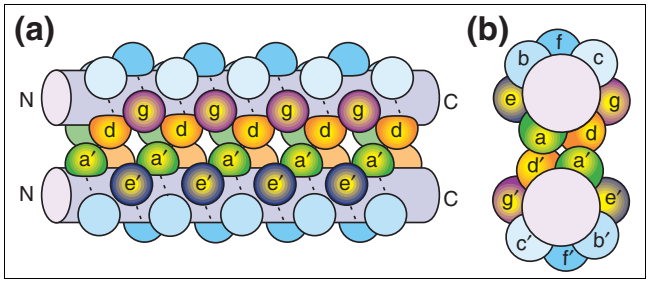

Figure 1
Cartoon of a parallel two-stranded coiled coil. **(a)** Side view and **(b)** top view. The interface between the α-helices in a coiled-coil structure is formed by residues at the core positions **a**, **d**, **e** and **g**. Positions in the two helices are distinguished by the prime notation; for example, **a** and **a'** are analogous positions in the two helices. N, amino terminus; C, carboxy terminus.

assess its utility for the general coiled-coil prediction problem. Our method represents coiled coils in terms of their interhelical interactions and it derives, from a base dataset of sequence and experimental data, a 'weight' that indicates how favorable each residue-residue interaction is. Our method is able to predict interaction partners with high confidence, identifying a significant fraction (70%) of strong bZIP pairings while maintaining that the majority (92%) of predicted interactions are correct. Further cross-validation testing demonstrates the extent to which the human bZIP data refines our method, and suggests levels of confidence, based on shared sequence similarity, for predicting bZIP interactions within new genomes.

Prior to this work, there has been only modest success in predicting the partnering specificity of naturally occurring coiled-coil proteins. An earlier version of our method was tested on fibrous coiled coils. It was able to eliminate a large fraction of non-interacting partners for a given coiled-coil sequence but not to find the actual partner [24]. Several other groups have counted the number of favorable and unfavorable electrostatic interactions to make some specific predictions about the nature of particular coiled-coil interactions [18,25,26]. Recently, simple rules incorporating both **ge'** electrostatic and **aa'** polar interactions have been used to evaluate potential bZIP dimerization in the *Drosophila* genome [27], but as yet most of these predictions have not been experimentally corroborated. On human bZIP data, such simple rules are able to identify only a small fraction of known strong interactions at a high level of precision. For example, when they are defined so as to identify at least one third of the strong interactions, they give rise to as many false positives (FP) as true positives (TP).

Whole- and cross-genomic approaches to predicting protein partners have had some success [28-34]. Our work, however, is the first to demonstrate large-scale, high-confidence computational predictions for any protein interaction motif.

Results

High-confidence predictions using base dataset

Each pair of bZIP coiled coils was scored using simple electrostatic weights, coupling energy weights and the 'base-optimized weights', that is, the weights optimized using our base dataset (see Methods). Using any of the methods, TPs are the correctly identified strong interactions, true negatives (TN) are the correctly identified non-interactions, FPs are the non-interactions incorrectly identified as interactions, and false negatives (FN) are the strong interactions incorrectly identified as non-interactions.

The distribution of scores computed using the base-optimized weights is depicted in Figure 2. Despite some overlap, the lowest scoring pairs correspond to non-interactions and the highest scoring pairs correspond to strong interactions, suggesting that it is possible to make high-confidence predictions. This can be quantified using the receiver-operator characteristic (ROC) curves in Figure 3a, which plot TP as a function of FP. All three methods identify some interactions, but the base-optimized weights consistently identify at least twice as many interactions as either of the other two sets of weights, over a wide range of allowed FP. Using the base-optimized weights, it is possible to predict a significant fraction of bZIP coiled-coil interactions with high confidence. In particular, the top scores include 56 strong coiled-coil interactions (70% of the experimental strong interactions) and five false positives; so at this threshold 92% of the predictions are true interactions. Simple electrostatic weights perform particularly poorly in comparison. For example, to identify a third (27) of the strong interactions, simple weights would misclassify 31 non-interactions, for a precision of 46.6%.

Non-interactions are plotted correspondingly in Figure 3b. In this case, the objective is to disqualify a large fraction of

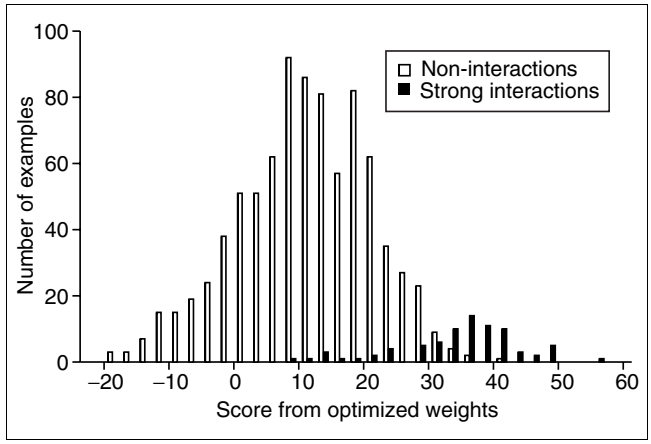

Figure 2
Histogram of scores using base-optimized weights. Non-interactions are shown in white and strong interactions are shown in black. Bins are of size 2.5.

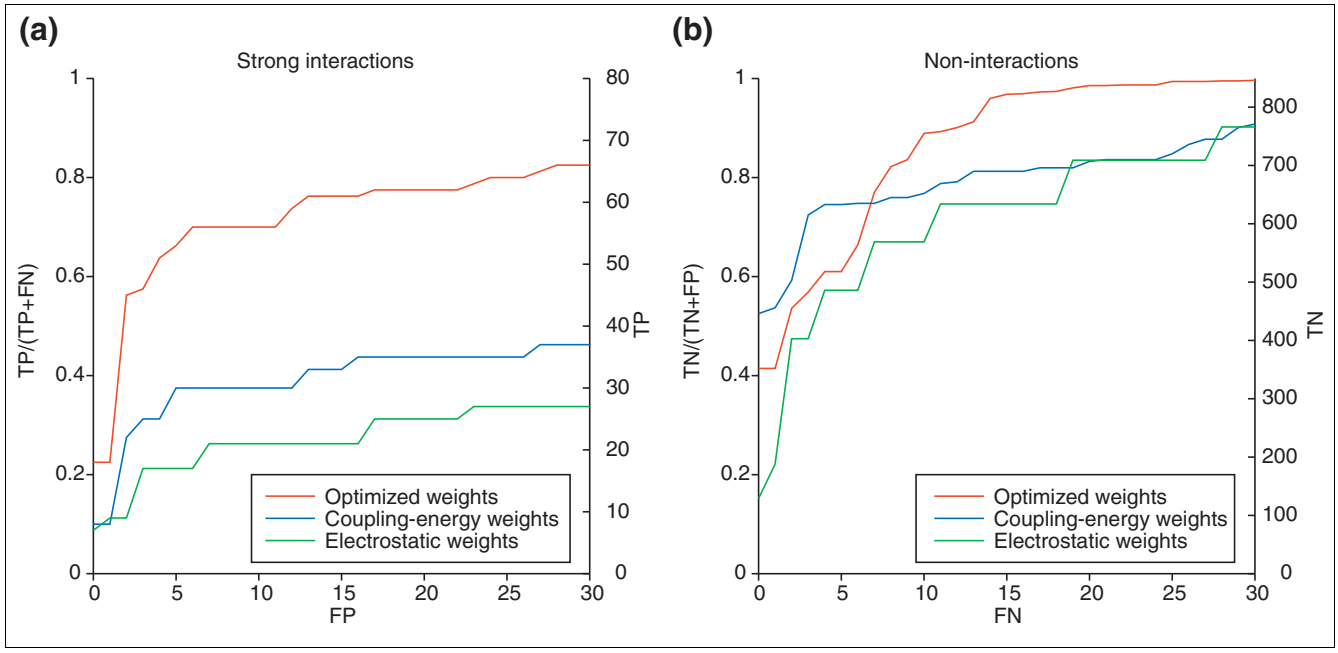

Figure 3
Prediction of strong interactions and non-interactions. Predictions are shown in red for the base-optimized weights, in blue for the coupling-energy weights, and in green for the simple electrostatic weights. **(a)** The fraction of strong coiled-coil interactions correctly identified as interactions (TP/(TP+FN)) as a function of the number of non-interactions incorrectly identified as interactions (FP). The second y-axis shows the number of strong coiled-coil interactions (TP). **(b)** The fraction of non-interactions correctly identified (TN/(TN+FP)), as a function of the number of strong interactions incorrectly identified as non-interactions (FN). The second y-axis shows the number of non-interactions (TN).

non-interactions from the pool of candidates while limiting the number of false negatives. All three sets of weights are effective in throwing out non-interacting coiled-coil pairs. For example, using the base-optimized weights, it is possible to throw out the bottom 89% of the scores, and still leave more than 83% of the strong interactions; at this cut-off, the negative predictive value (TN/(TN+FN)) is 98.4%. The base-optimized weights consistently outperform the simple electrostatic weights; relative performance of the optimized weights and coupling energy weights depends on the number of FN allowed, with the base-optimized weights performing better when allowing at least seven FN.

For each sequence, Figure 4 illustrates the distribution of scores obtained using the base-optimized weights for all strong interactions and non-interactions. The high-confidence predictions described above correspond to examples that score above 32.8 (predicted interactions) and below 27.8 (predicted non-interactions); these thresholds were chosen empirically based on trade-offs between the number and accuracy of predictions (as shown in Figure 3). Because interaction scores depend to some extent on the residue composition of the sequences, for some sequences all pairings fall below the threshold for predicting interactions. Sequences in the smMaf, lgMaf and CNC families produce the 'dip' in scores in Figure 4 (see Discussion). Regardless of this, the highest scoring pairings for most sequences represent coiled-coil interactions. Thus, a simple method for identifying the most likely partners for any sequence is to select the highest scoring pairings for that sequence.

The grid in Figure 5 presents our predictions of strong interactions and non-interactions for all of the bZIP proteins studied; similar sequences are grouped into families. Many strong interactions are intrafamily [23], so a large number of TPs lie along the diagonal. But we also identify TN intrafamily interactions and TP interfamily interactions (off-diagonal boxes), indicating the sensitivity of the method and its ability to predict both homo- and heterodimeric pairings.

The raw fluorescence data from the array experiment are subject to a range of different interpretations [23]. In the results described above, we made predictions for 54.3% of the data, consisting of the strongest and weakest interactions. However, by requiring less internal consistency in the experimental data (for example, fewer observations of an interaction), we can relax the definitions so that 81.1% of possible pairwise combinations are classified as interactions or non-interactions. (Note that we are less certain of the reliability of these classifications.) In this case, using the same thresholds as earlier, we correctly identify 54% of interactions with 82% positive predictive value (TP/(TP+FP)), and 95% of the

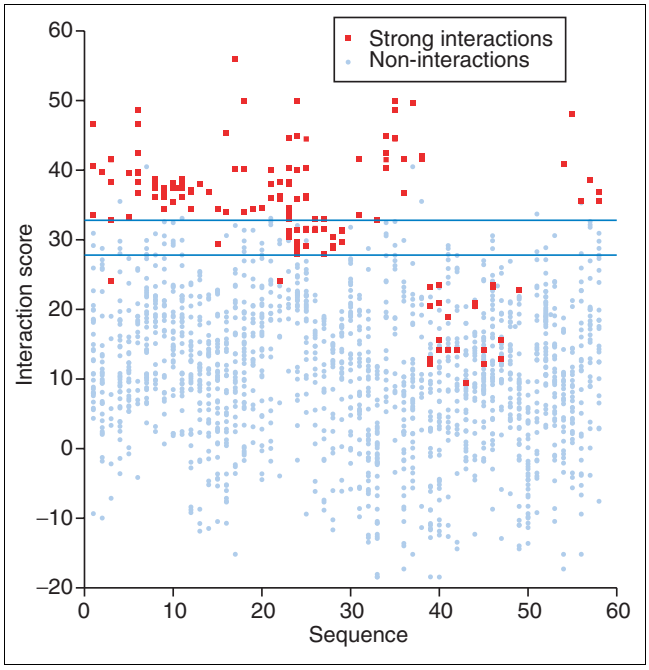

Figure 4
Interaction scores for each protein. Each column shows the interaction scores using the base-optimized weights for one sequence's strong interactions, shown in red, and non-interactions, shown in blue. The 58 sequences are grouped by similarity, and ordered as in Table 1 of Additional data file 1. The blue horizontal lines mark off high-confidence predictions of interactions (more than 32.8) and non-interactions (less than 27.8). Note that all heterodimer interactions appear twice on the graph.

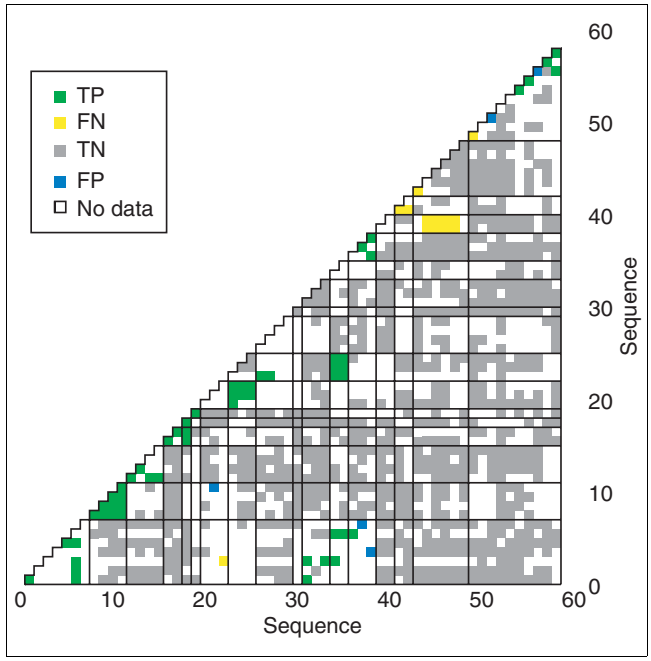

Figure 5
Full grid depiction of high-confidence predictions. Correct predictions are colored green (TP) or grey (TN) whereas incorrect predictions are colored yellow (FN) or blue (FP). White boxes represent pairs of sequences that are not classified as strong interactions or non-interactions. A point is 'positive' if its score using the base-optimized weights is greater than 32.8. A point is 'negative' if its score is lower than 27.8. Strong interactions and non-interactions are as defined in Methods. The sequences are grouped by similarity, and are numbered as in Table 1 of Additional data file 1. The families, separated by the gridlines, are ordered as follows: C/EBP, sequences 1-7; CREB, 8-11; OASIS, 12-15; ATF-6, 16-17; XBP, 18; E4BP4, 19; ATF-2, 20-22; JUN, 23-25; FOS, 26-29; ATF-3, 30; ATF-4, 31-33; B-ATF, 34-35; PAR, 36-38; smMAF, 39-40; lgMAF, 41-42; CNC, 43-48; and YEAST, 49-58.

non-interactions with 96% negative predictive value (TN/(TN+FN)) (data not shown).

Comparison accuracy using base dataset

An important step in understanding coiled-coil specificity is to predict relative interaction binding strengths. The bZIP arrays are able to determine the relative strengths of different interactions involving a common probe [23]. Here, we assess the ability of the base-optimized weights to predict these relative strengths, defining two interactions as well-ordered if the stronger interaction is given a higher score than the other. This analysis relies on raw data and does not require classifying the possible bZIP pairings into different groups (for example, strongly interacting or non-interacting), and thus can use more of the experimental data.

We consider accuracy in ordering interaction strengths as a function of the difference, d, between raw fluorescence signals for the two interactions under comparison. To assign a consistent experimental ordering with separation d to two interactions, we require that there be at least four possible data comparisons, and that in at least 90% of these comparisons the signal for the stronger interaction is at least d greater than the signal for the other interaction. (The experimental data for each interaction include four replicate measurements from each of five experiments. Measurements from the same probe and experiment are comparable, permitting a maximum of 80 direct data comparisons per pair of interactions, as some of these measurements may be missing.) Our method is robust (data not shown) for a range of criteria, reflecting the internal consistency of the data. The chosen cutoffs allow 33,186 experimental orderings. In theory, 95,874 comparisons can be made (1,653 pairs of interactions for each probe), but the ranges of signals for two interactions may overlap when neither is significantly stronger than the other. For example, relatively few comparisons between two non-interactions are clearly ordered.

Figure 6 plots the comparison accuracy as a function of the difference d. At a separation of at least $d = 500$, comparison accuracy exceeds 85%; most of these comparisons are between strong interactions and non-interactions. Separation near zero is the most difficult case, as it includes comparisons of weak interactions.

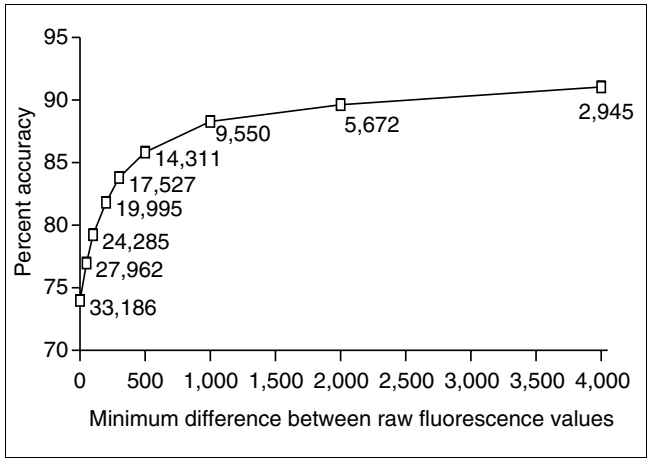

Figure 6
Accuracy in predicting relative strengths of interactions. Percent of comparisons correct using base-optimized weights, as a function of separation of raw fluorescence values. Labels on points show the number of comparisons with consistent experimental data and the given level of separation.

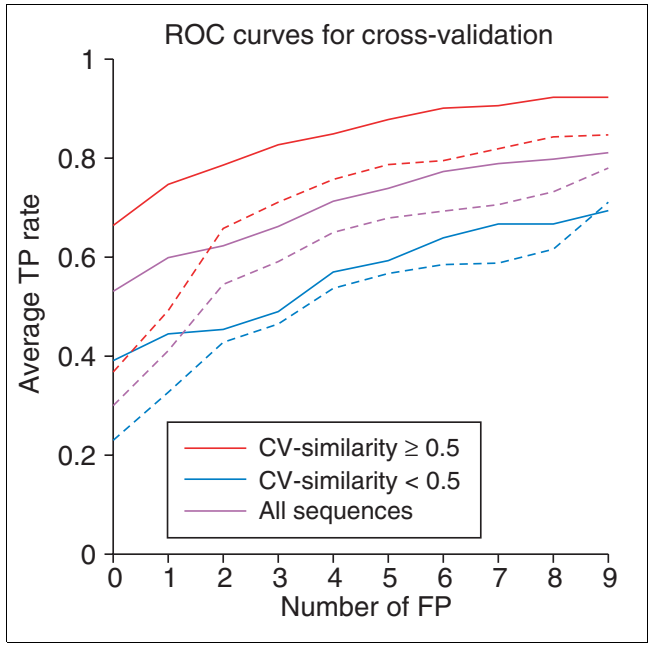

Figure 7
Cross-validation testing. The average fraction of correctly identified coiled-coil interactions as a function of the number of FP. Solid lines give averages computed using the appropriate cross-validation weights, and dotted lines give averages computed using the base-optimized weights. The TP rates shown were averaged over all human bZIP sequences, shown in purple, over human bZIP sequences with CV-similarity equal to or greater than 50%, shown in red, and over human bZIP sequences with CV-similarity less than 50%, shown in blue.

Cross-validation testing incorporating bZIP data

We perform cross-validation testing to show the contributions of the human bZIP data to the optimized weights. We then determine how performance on new bZIP sequences is expected to vary as a function of sequence similarity to the human bZIP data.

Our general cross-validation setup is as follows. For each human bZIP family *F* (as given in Table 1 of Additional data file 1), we optimize a weight vector with our base dataset plus human bZIP data that does not involve any sequence in family *F*. Specifically, we incorporate likely bZIP interactions and non-interactions in the form of Equations 3 and 4, and comparisons between bZIP pairings in the form of Equation 2. Likely interactions and non-interactions follow the relaxed definitions given in Methods, and TP, TN, FN, and FP are defined with respect to these; comparisons are made between pairings that share a probe and differ in raw intensity by at least 500. The resulting weights are used to score all possible interactions of each sequence in family *F* and generate a TP rate (TP/(TP+FN)) for each sequence as a function of the number of FP. This process is repeated for all families. An overall TP rate, corresponding to the average over individual TP rates for each sequence, is shown in Figure 7. For comparison, the average TP rate using the base-optimized weights is also shown; the use of human bZIP data clearly improves performance over a wide range of allowed false positives.

To judge cross-validation performance as a function of sequence similarity, for each sequence *x* in family *F*, we compute its 'CV-similarity' as the maximum percent identity over the **a**, **d**, **g** and **e** positions between *x* and any human bZIP not in family *F*. The average CV-similarity for human bZIP sequences is 48.7%, with values ranging from 29.1% to 62.5% for individual sequences. The sequences are grouped into those with CV-similarity of 50% or more (24 sequences) and those with CV-similarity less than or equal to 50% (23 sequences). (One of the human bZIP sequences, ATF4-L1, has no experimentally-determined interactions.) The average TP rate in each group is computed (Figure 7). Incorporating the human bZIP data improves performance on both groups. However, sequences with CV-similarity greater than or equal to 50% benefit particularly from cross-validation training.

It is important to note that our method does not predict trivially the interactions and non-interactions for a sequence to be those of its closest homolog included in the optimization. Such a sequence-similarity based approach on average correctly identifies 30.5% of likely interactions for a particular sequence while allowing 3.1 FPs. In contrast, our cross-validation optimization on average correctly identifies 53.1% of interactions when allowing no FPs, or 66.2% of interactions when allowing three FPs (Figure 7).

Discussion

We have applied a computational method for predicting coiled-coil partnering specificity to a near-complete set of human and yeast bZIP proteins. The ability to make high-confidence predictions of coiled-coil interactions and non-interactions illustrates the effectiveness of our method. Strong interactions for each sequence score higher than non-interactions in almost all cases (Figure 4); this suggests an approach to predicting interacting partners for all bZIP sequences in a genome.

Our method performs significantly better than simple rules, based on **ge'** electrostatic interactions and **aa'** polar interactions. Similar rules have been used to predict potential bZIP dimerization in the *Drosophila* genome [27], and we note that slight variations in these rules do not change performance dramatically on the human bZIP data. Our method also performs better than rules based on experimentally determined coupling energies. In fact, because the base-optimized weights are constrained using both simple electrostatic rules and coupling energy measurements, the 56 high-confidence predictions made by the optimized weights largely encompass the highest scoring pairs using the other methods. For example, they include the 27 highest scoring strong interactions using the electrostatic weights, and 36 out of the 37 highest scoring strong interactions using the coupling energy weights. To identify these smaller numbers of strong interactions, both the electrostatic weights and coupling energy weights would identify approximately as many false positives as false negatives. Similarly, the 13 FN given by the base-optimized weights (Figure 5) fall among the lowest scoring strong interactions using the electrostatic weights, and among the bottom half of scores for strong interactions using the coupling energy weights. This suggests that the strong interactions with low scores using the base-optimized weights do not have particularly favorable **ge'** electrostatic interactions or **aa'** hydrophobic interactions.

Our results provide several insights into how specificity is encoded in coiled-coil sequences. First, we note that simple electrostatic weights are more effective in identifying non-interactions than interactions (Figure 3). This suggests that electrostatic repulsion and unpaired buried asparagine residues may play key roles in preventing the formation of some coiled coils [18,35]. Second, both the coupling energy weights and the cruder simple electrostatic weights rely on **ge'** and **aa'** interactions. Our method considers **dd'**, **da'**, **ad'**, **ga'** and **de'** interactions as well. The increased sensitivity we obtain indicates that these pairwise interactions play an important role in partnering specificity.

High-quality experimental data are important for the success of our method. This is demonstrated by the cross-validation results (Figure 7), and also by the fact that including experimental constraints in the base dataset significantly enhanced the predictive power of the method (data not shown). Almost all of the experimental constraints in the base dataset focus on residues in the **ge'** and **aa'** positions, however. Data for the other interactions are derived primarily from sequence databases. As more experimental information is gathered about interactions between residues in other positions, it can be readily incorporated in our methodology and should contribute to improvements in performance.

In general, residue interactions that are not observed frequently in the base dataset are likely to be problematic for the optimization method, as the corresponding weights are under-constrained; this phenomenon almost certainly plays a role in the observed FP and FN. Moreover, interaction scores vary depending on residue composition (Figure 4), and thus the same cut-off scores are not ideal for all bZIP sequences. For example, the top scoring pairs for some sequences in the smMaf, lgMaf and CNC families (which produce the dip in interaction scores in Figure 4 and the block of yellow FNs in the upper right of Figure 5) form strong interactions but score below the cut-off of 32.8. The composition of **a** positions in these sequences partially explains their lower interaction scores. These sequences contain two to three **a** positions with basic amino acids. Coupling energy measurements for **aa'** residue positions demonstrate that interactions involving lysine have, on average, unfavorable energetic contributions as compared to those involving strictly hydrophobic amino acids [19]. Accordingly, the optimized weights for interactions with **a** position basic amino acids are not large and contribute little to overall interaction scores. Furthermore, many of the strong interactions for these sequences are between the (very similar) smMaf and CNC families and within the lgMaf family, exaggerating the influence of these lower weights.

The dependence of interaction scores on residue composition is not specific to the use of base-optimized weights. For example, a coiled-coil sequence with few charged amino acids in the **g** and **e** positions is unlikely to have any high scoring interactions using the simple electrostatic weights. Although using the same scoring cut-offs for all pairings allows the base-optimized weights to make high-confidence predictions for more than 95% of experimentally classified strong interactions and non-interactions (Figure 4), future improvements to our method will include correcting the interaction score for two sequences based on their residue compositions.

As an immediate next step, we plan to make novel predictions about bZIP interactions in other eukaryotic genomes, and to test experimentally the high-confidence predictions. Our cross-validation testing was designed to mimic the situation where bZIP proteins in other genomes do not have direct orthologs in human. This testing shows that incorporating human bZIP interaction data in our optimization procedure will be helpful in making predictions for novel bZIP proteins in other genomes, and that we may judge confidence in predictions as a function of both interaction scores and sequence identity.

Conclusions

The problem of predicting protein interactions mediated by the coiled coil is far from being solved. Our work demonstrates, however, that it is possible to reliably identify from sequence a significant fraction of bZIP coiled-coil protein partners. We also note that our methodology is directly applicable to parallel, two-stranded coiled-coil interactions that are not bZIP interactions [24]. Practical limitations include the paucity of experimental and sequence data on heterotypic **dd'** interactions; biophysical studies on these interactions will be especially useful for improving performance. Further directions include incorporating molecular modeling approaches, either as a way of constraining the weight vector, or as a way of predicting pairings whose scores fall between the thresholds for predicting either high-confidence non-interactions or high-confidence interactions. Finally, we note that an approach similar to the one outlined here may be useful in studying other structurally well-defined protein interfaces that are also well-studied experimentally.

Methods

Representing coiled coils

In dimeric coiled coils, residues at the **a**, **d**, **e** and **g** positions form the protein-protein interface [36,37] (Figure 1). Experimental studies show that specificity is largely driven by interactions between residues at these core positions (for example, see [35]). Our method includes the assumption that considering interhelical interactions among these residues in a pairwise manner is sufficient. (We can consider three or more amino acids at a time but this would require a larger coiled-coil database.) Based on structural features of the interhelical interface as well as experiments on determinants of specificity, the following seven interhelical interactions are assumed to govern partnering in coiled coils:

$$\mathbf{a}_i\mathbf{d'}_i,\ \mathbf{d}_i\mathbf{a'}_{i+1},\ \mathbf{d}_i\mathbf{e'}_i,\ \mathbf{g}_i\mathbf{a'}_{i+1},\ \mathbf{g}_i\mathbf{e'}_{i+1},\ \mathbf{a}_i\mathbf{a'}_i,\ \mathbf{d}_i\mathbf{d'}_i. \quad (1)$$

The prime differentiates the two strands and the subscript denotes the relative heptad number (for example, the first interaction, $\mathbf{a}_i\mathbf{d'}_i$, is between the **a** position in the *i*-th heptad of one helix and the **d** position in the same heptad of the other helix). Interactions between strands are symmetric (for example, both $\mathbf{a}_i\mathbf{d'}_i$ and $\mathbf{a'}_i\mathbf{d}_i$ are included).

Consequently, each coiled coil is represented as a 2,800-dimension vector $\boldsymbol{x}$, the entries of which tabulate the occurrences of amino-acid pairs in the above interactions. Specifically, entry $\boldsymbol{x}_{(p,q),i,j}$ indicates the number of times amino acids *i* and *j* appear across the helical interface in positions *p* and *q*, respectively.

Scoring framework

For each possible interhelical interaction, we seek a weight $\boldsymbol{w}_{(p,q),i,j}$ that denotes how favorable the interaction is between amino acid *i* in position *p* and amino acid *j* in position *q*. A potential coiled coil represented by $\boldsymbol{x}$ is then scored by computing $\boldsymbol{w} \bullet \boldsymbol{x}$ where $\boldsymbol{w}$ is a vector of such weights. Although the ideal weight vector is unknown, it may be approximated in several different ways. We compare our framework, which is based on computationally optimizing the weights, with two previously proposed methods.

In the first alternate method, simple rules count favorable and unfavorable electrostatic interactions [18,25,26]. We consider an extension of these rules given by the following weight vector [23] that additionally rewards buried asparagines [20]: $\boldsymbol{w}_{(g,e),E,+} = \boldsymbol{w}_{(g,e),+,E} = 1.0$; $\boldsymbol{w}_{(g,e),D,+} = \boldsymbol{w}_{(g,e),+,D} = \boldsymbol{w}_{(g,e),Q,+} = \boldsymbol{w}_{(g,e),+,Q} = 0.5$; $\boldsymbol{w}_{(g,e),-,-} = \boldsymbol{w}_{(g,e),+,+} = -1.0$; $\boldsymbol{w}_{(a,a),N,N} = 1.0$. Here, '+' denotes amino acids lysine and arginine and '-' denotes aspartic acid and glutamic acid. All other weight elements are equal to zero. We refer to this set of weights as 'simple electrostatic weights'.

In the second alternate method, weights come from experimental coupling energies [19,35]. Such coupling energies have been measured for some **aa'** and **ge'** residue interactions. We refer to this set of weights as 'coupling energy weights'.

On the other hand, we propose a method that optimizes these weights to satisfy constraints derived from experimental data or sequence information [24].

In the optimization framework, experimental information on relative coiled-coil stability (for example, the observation that coiled coil $\boldsymbol{x}$ is more stable than coiled coil $\boldsymbol{y}$) is used to constrain the weight vector $\boldsymbol{w}$ by requiring that

$$\boldsymbol{w} \bullet \boldsymbol{x} > \boldsymbol{w} \bullet \boldsymbol{y} \quad (2)$$

This allows, for example, experimental information used to derive coupling energies [19,38,39] to constrain the weight vector.

Additionally, sequences known to form coiled coils should score higher than those that do not (see section on base dataset, below):

$$\boldsymbol{w} \bullet \boldsymbol{x} > 0, \text{ for all coiled coils } \boldsymbol{x}, \quad (3)$$

$$\boldsymbol{w} \bullet \boldsymbol{y} < 0, \text{ for all non-coiled coils } \boldsymbol{y}. \quad (4)$$

Finally, knowledge about specific weight elements can be directly incorporated. For example, we require

$$\boldsymbol{w}_{(g,e),K,E} > 0,\ \boldsymbol{w}_{(g,e),E,E} < 0. \quad (5)$$

These types of constraints are used to capture some of the features of the simple electrostatic weights.

Indexing each constraint with i, the above constraints (equations 2-5) can be rewritten using vectors $\boldsymbol{z}^{(i)}$, such that $\boldsymbol{w}$ is constrained to satisfy $\boldsymbol{w} \bullet \boldsymbol{z}^{(i)} > 0$. Including non-negative slack variables ε_i to allow for errors in sequence or experimental data, each constraint can then be relaxed as $\boldsymbol{w} \bullet \boldsymbol{z}^{(i)} > -\varepsilon_i$. The goal is to find $\boldsymbol{w}$ and ε_i such that each constraint is satisfied and $\sum \varepsilon_i$ is minimized. Tradeoffs between training and generalization error suggest the approach of support vector machines (SVMs) [40,41], in which the following quadratic objective function is minimized, subject to a variation of the previously described set of linear constraints:

Minimize $(1/2)||\boldsymbol{w}||^2 + \sum \varepsilon_i$

Subject to $\boldsymbol{w} \bullet \boldsymbol{z}^{(i)} \geq 1 - \varepsilon_i$

$\varepsilon_i \geq 0$.

The SVM-lite package of [42] was used to implement the optimization framework, with all vectors $\boldsymbol{z}^{(i)}$ normalized using the L_2 norm.

Base dataset

The optimization framework relies on a base dataset comprising known coiled coils, hypothesized non-interactions and experimentally-determined relative stabilities to determine constraints on $\boldsymbol{w}$. Coiled-coil sequences are taken from a database of non-redundant interacting coiled coils containing approximately 29,000 residues from one of two classes: homodimeric coiled coils in myosin, tropomyosin, cortexillin and types III and V intermediate filament proteins; and heterodimeric coiled coils in keratin proteins [8,24].

A set of non-coiled coils is created by misaligning known partnering strands (that is, strands are paired but with a shift of one or more heptads from the correct alignment). Although this may result in some pairs that could form *in vitro*, coiled-coil interactions are known to be quite specific (for example, see [18,23,43]), and our methodology allows for some errors.

Information is incorporated from various biophysical studies that take a coiled-coil host system, mutate amino acids, and determine melting temperatures. Each of these studies [14,16,19,38,39,44-47] provides an ordering of the stabilities of the coiled coils considered, and thus constraints of the type in equation 2 are introduced. Additional constraints are derived from the experimental data of [48-54]. The bZIP proteins VBP, Fos, Jun and GCN4 are used as host systems in some of the experimental studies considered. No other bZIP proteins, and in particular no human bZIP interaction data from [23], are part of the base dataset.

In all, 6,379 constraints are included in the base dataset (2,612 from coiled-coil or non-coiled coil sequences, 3,575 from experiments and 192 simple constraints as in equation 2).

Experimental bZIP interaction dataset

The bZIP interaction data collected in [23] are used for testing. In that work, coiled-coil domains of 62 bZIP proteins were printed onto glass slides and then probed with a corresponding set of fluorescently-labeled peptides. For each probe sequence, its raw fluorescence signals from interactions with all surface peptides were normalized to obtain Z-scores. Of the 62 sequences, we removed three duplicate sequences and the sequence CREM-Ia(K324R), which has the same interactions as CREM-Ia; only data for the remaining 58 sequences (given in Table 1 of Additional data file 1) are considered. Note that the bZIP peptides used in the array studies have coiled-coil regions of varying lengths and contain varying amounts of non-coiled coil sequence. We used the coiled-coil regions defined in [23]. When considering whether two sequences partner, we assume that the coiled-coil interface has exactly the length of the shorter region (that is, the extra residues of the longer coiled coil are ignored). The alignment between the two strands used for scoring is shown in Table 1 of Additional data file 1. Strands can be unambiguously aligned using the conserved, basic DNA binding region that occurs immediately amino-terminal to the coiled-coil region (not shown).

Based on Z-scores, we classify potential bZIP coiled-coil pairings into 80 strong interactions and 849 non-interactions as follows. Note that for each pair of sequences, the normalization produces two Z-scores that should validate one another (one value is taken when the first sequence is the probe and the second is on the surface, and the other *vice versa*). A strong interaction is defined as having two Z-scores greater than 10 and a non-interaction has two Z-scores less than 1. These definitions correspond to the bZIP pairings whose classifications are most certain, given the experimental data, and our testing is done with respect to them. Strong interactions and non-interactions account for 54.3% of possible pairwise combinations.

Among the remaining bZIP pairings, those with two Z-scores greater than 2.5 are considered likely coiled-coil interactions and those with one Z-score less than 1 and another less than 1.5 are considered likely non-interactions. These relaxed definitions, which are used in cross-validation training and testing, lead to 186 likely interactions and 1,250 likely non-interactions, and account for 83.9% of possible pairwise combinations.

Program availability

The program used is available online at [55]. Note that it has been validated only for predicting interactions between bZIP proteins and not for coiled coils generally.

Additional data

The following additional data are included with the online version of this article: a table of the human and yeast bZIP

transcription factor coiled coils considered, a list of the experimental data used to derive constraints (Additional data file 1), and a list of the base-optimized weights (Additional data file 2).

Acknowledgements

J.H.F. thanks the NIH for a Predoctoral Fellowship Award. M.S. thanks the NSF for PECASE award MCB-0093399 and DARPA for grant MDA972-00-1-0031. The authors thank Chris Burge, John Newman, Sam Sia and members of the Keating and Singh labs for comments on the manuscript.

References

1. Uetz P, Giot L, Cagney G, Mansfield T, Judson R, Knight J, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al.*: **A comprehensive analysis of protein-protein interactions in *S. cerevisiae*.** *Nature* 2000, **403:**623-627.
2. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA* 2001, **98:**4569-4574.
3. Newman JRS, Wolf E, Kim PS: **A computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2000, **97:**13203-13208.
4. Gavin A, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick J, Michon A, Cruciat C *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415:**141-147.
5. Ho Y, Gruhler A, Heilbut A, Bader G, Moore L, Adams S, Millar A, Taylor P, Bennett K, Boutilier K *et al.*: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry.** *Nature* 2002, **415:**180-183.
6. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T *et al.*: **Global analysis of protein activities using proteome chips.** *Science* 2001, **293:**2101-2105.
7. Lupas A, van Dyke M, Stock J: **Predicting coiled coils from protein sequences.** *Science* 1991, **252:**1162-1164.
8. Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, Kim PS: **Predicting coiled coils using pairwise residue correlations.** *Proc Natl Acad Sci USA* 1995, **92:**8259-8263.
9. Wolf E, Kim PS, Berger B: **Multicoil: a program for predicting two- and three-stranded coiled coils.** *Protein Sci* 1997, **6:**1179-1189.
10. Singh M, Berger B, Kim PS: **Learncoil-VMF: computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins.** *J Mol Biol* 1999, **290:**1031-1044.
11. Delorenzi M, Speed T: **An HMM model for coiled-coil domains and a comparison with PSSM-based predictions.** *Bioinformatics* 2002, **18:**617-625.
12. Woolfson DN, Alber T: **Predicting oligomerization state of coiled coils.** *Protein Sci* 1995, **4:**1596-1607.
13. Harbury PB, Tidor B, Kim PS: **Repacking protein cores with backbone freedom: structure prediction for coiled coils.** *Proc Natl Acad Sci USA* 1995, **92:**8408-8412.
14. Keating AE, Malashkevich V, Tidor B, Kim PS: **Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils.** *Proc Natl Acad Sci USA* 2001, **98:**14825-14830.
15. Havranek J, Harbury PB: **Automated design of specificity in molecular recognition.** *Nat Struct Biol* 2003, **10:**45-52.
16. O'Shea E, Rutkowski R, Kim PS: **Mechanism of specificity in the fos-jun oncoprotein heterodimer.** *Cell* 1992, **68:**699-708.
17. Harbury PB, Zhang T, Kim PS, Alber T: **A switch between two-, three- and four-stranded coiled coils in GCN4 leucine zipper mutants.** *Science* 1993, **262:**1401-1407.
18. Vinson C, Hai T, Boyd S: **Dimerization specificy of the leucine zipper-containing bZIP motif on DNA binding: prediction and rational design.** *Genes Dev* 1993, **7:**1047-1058.
19. Acharya A, Ruvinov S, Gal J, Moll JR, Vinson C: **A heterodimerizing leucine zipper coiled coil system for examining the specificity of a position interactions: amino acids I, V, L, N, A, and K.** *Biochemistry* 2002, **41:**14122-14131.
20. Lumb K, Kim PS: **A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil.** *Biochemistry* 1995, **34:**8642-8648.
21. Gonzalez L, Woolfson D, Alber T: **Buried polar residues and structural specificity in the GCN4 leucine zipper.** *Nat Struct Biol* 1996, **3:**1011-1018.
22. Gonzalez L, Brown R, Richardson D, Alber T: **Crystal structures of a single coiled-coil peptide in two oligomeric states reveal the basis for structural polymorphism.** *Nat Struct Biol* 1996, **3:**1002-1009.
23. Newman JRS, Keating AE: **Comprehensive identification of human bZIP interactions using coiled-coil arrays.** *Science* 2003, **300:**2097-2101.
24. Singh M, Kim PS: **Towards predicting coiled-coil protein interactions.** In *Proceedings of the 5th Annual International Conference on Computational Molecular Biology* ACM; 2001:279-286.
25. Parry DAD, Crewther WG, Fraser RD, MacRae TP: **Sequences of α-keratin: structural implication of the amino acid sequences of the type I and type II chain segments.** *J Mol Biol* 1977, **113:**449-454.
26. McLachlan A, Stewart M: **Tropomyosin coiled-coil interactions: evidence for an unstaggered structure.** *J Mol Biol* 1975, **98:**293-304.
27. Fassler J, Landsman D, Acharya A, Moll JR, Bonovich M, Vinson C: **bZIP proteins encoded by the *Drosophila* genome: evaluation of potential dimerization partners.** *Genome Res* 2002, **12:**1190-1200.
28. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23:**324-328.
29. Overbeek R, Fonstein M, D'Souza M, Pusch G, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.
30. Marcotte E, Pellegrini M, Ng H, Rice D, Yeates T, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285:**751-753.
31. Enright A, Iliopoulos I, Kyrpides N, Ouzounis C: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402:**86-90.
32. Goh C, Bogan A, Joachimiak M, Walther D, Cohen F: **Co-evolution of proteins with their interaction partners.** *J Mol Biol* 2000, **299:**283-293.
33. Ramani A, Marcotte E: **Exploiting the co-evolution of interacting proteins to discover interaction specificity.** *J Mol Biol* 2003, **327:**273-284.
34. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, **302:**449-453.
35. Vinson C, Myakishev M, Acharya A, Mir A, Moll JR, Bonovich M: **Classification of human bZIP proteins based on dimerization properties.** *Mol Cell Biol* 2002, **22:**6321-6335.
36. O'Shea E, Klemm J, Kim PS, Alber T: **X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil.** *Science* 1991, **254:**539-544.
37. Glover J, Harrison S: **Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA.** *Nature* 1995, **373:**257-261.
38. Krylov D, Mikhailenko I, Vinson C: **A thermodynamic scale for leucine zipper stability and dimerization specificity: e and g interhelical interactions.** *EMBO J* 1994, **13:**2849-2861.
39. Krylov D, Barchi J, Vinson C: **Inter-helical interactions in the leucine zipper coiled coil dimer: pH and salt dependence of coupling energy between charged amino acids.** *J Mol Biol* 1998, **279:**959-972.
40. Vapnik V: *Statistical Learning Theory* New York: Wiley; 1998.
41. Burges C: **A tutorial on support vector machines for pattern recognition.** *Data Mining and Knowledge Discovery* 1998, **2:**121-167.
42. Joachims T: **Making large-scale SVM learning practical.** In *Advances in Kernel Methods: Support Vector Machines* Edited by: Schölkopf B, Burges C, Smola A. Cambridge: MIT Press; 1999:169-185.
43. Hurst H: **Transcription factors 1: bZIP proteins.** *Protein Profile* 1995, **2:**101-168.
44. Moitra J, Szilak L, Krylov D, Vinson C: **Leucine is the most stabilizing aliphatic amino acid in the *d* position of a dimeric leucine zipper coiled coil.** *Biochemistry* 1997, **36:**12567-12573.
45. Jelesarov I, Bosshard HR: **Thermodynamic characterization of the coupled folding and association of heterodimeric coiled coils (leucine zippers).** *J Mol Biol* 1996, **263:**344-358.

46. Tripet B, Wagschal K, Lavigne P, Mant C, Hodges R: **Effects of side-chain characteristics on stability and oligomerization state of a de novo-designed model coiled-coil: 20 amino acid substitutions in position 'd'.** *J Mol Biol* 2000, **300:**377-402.
47. Akey DL, Malashkevich VN, Kim PS: **Buried polar residues in coiled-coil interfaces.** *Biochemistry* 2001, **40:**6352-6360.
48. Hu J, O'Shea E, Kim PS, Sauer R: **Sequence requirements for coiled coils: analysis with lambda repressor-GCN4 leucine zipper fusions.** *Science* 1990, **250:**1400-1403.
49. Hu J, Newell N, Tidor B, Sauer R: **Probing the roles of residues at the e and g positions of the GCN4 leucine zipper by combinatorial mutagenesis.** *Protein Sci* 1993, **2:**1072-1084.
50. Zeng X, Zhu H, Lashuel H, Hu J: **Oligomerization properties of GCN4 leucine zipper e and g mutants.** *Protein Sci* 1997, **6:**2218-2226.
51. Kammerer R, Frank S, Schulthess T, Landwehr R, Lustig A, Engel J: **Heterodimerization of a functional GABAB receptor is mediated by parallel coiled-coil alpha-helices.** *Biochemistry* 1999, **38:**13263-13269.
52. Porte D, Oertel-Buchheit P, John M, Granger-Schnarr M, Schnarr M: **DNA binding and transactivation properties of fos variants with homodimerization capacity.** *Nucleic Acids Res* 1997, **25:**3026-3033.
53. Smeal T, Angel P, Meek J, Karin M: **Different requirements for formation of Jun:Jun and Jun:Fos complexes.** *Genes Dev* 1989, **3:**2091-2100.
54. Amati B, Brooks M, Levy N, Littlewood T, Evan G, Land H: **Oncogenic activity of the c-Myc protein requires dimerization with Max.** *Cell* 1993, **72:**233-245.
55. **bZIP coiled-coil scoring form** [http://compbio.cs.princeton.edu/bzip]