Quantitative Biology 2015, 3(3): 115–123
DOI 10.1007/s40484-015-0045-y

RESEARCH ARTICLE

# Determination of specificity influencing residues for key transcription factor families

**Ronak Y. Patel[1,†,*], Christian Garde[2,†] and Gary D. Stormo[1,*]**

[1] Department of Genetics, School of Medicine, Washington University, St. Louis, MO 63108, USA
[2] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kgs. Lyngby, DK 2800, Denmark
* Correspondence: ronak.patel@bcm.edu, stormo@genetics.wustl.edu

Received March 1, 2015; Revised May 18, 2015; Accepted May 21, 2015

**Transcription factors (TFs) are major modulators of transcription and subsequent cellular processes. The binding of TFs to specific regulatory elements is governed by their specificity. Considering the gap between known TFs sequence and specificity, specificity prediction frameworks are highly desired. Key inputs to such frameworks are protein residues that modulate the specificity of TF under consideration. Simple measures like mutual information (MI) to delineate specificity influencing residues (SIRs) from alignment fail due to structural constraints imposed by the three-dimensional structure of protein. Structural restraints on the evolution of the amino-acid sequence lead to identification of false SIRs. In this manuscript we extended three methods (direct information, PSICOV and adjusted mutual information) that have been used to disentangle spurious indirect protein residue-residue contacts from direct contacts, to identify SIRs from joint alignments of amino-acids and specificity. We predicted SIRs for homeodomain (HD), helix-loop-helix, LacI and GntR families of TFs using these methods and compared to MI. Using various measures, we show that the performance of these three methods is comparable but better than MI. Implication of these methods in specificity prediction framework is discussed. The methods are implemented as an R package and available along with the alignments at http://stormo.wustl.edu/SpecPred.**

**Keywords:** protein-DNA interactions; residue co-variance; motifs; co-evolution; feature selection; direct information; specificity determinants

## INTRODUCTION

Transcription factors (TFs) are important components of cellular regulatory networks. Knowledge of TF specificity is essential for understanding regulatory networks of physiological pathways [1], annotating non-coding or disease causing variants [2], design of TF-nucleases/TF-lysine demethylase for site specific modifications of genetic [3] and epigenetic features [4], respectively, and modulation of metabolic pathways for commercial purposes [5]. TF specificities also serve as input for several prediction frameworks and global models of cellular regulation [6,7]. A recent estimate of the number of TFs encoded by the human genome is in the range of 1700–1900 [8]. Although large scale efforts have been undertaken by various groups, there are less than 500 human TFs for which the specificity is known. TF specificities of model and pathological organisms are also far from complete; even the well studied *Escherichia coli* is far from complete. To close this gap, TF specificity prediction models are urgently needed.

Although a simple, deterministic recognition code has been disproven [9], there are several reports of successful TF family-specific probabilistic recognition codes [10]. Specificity prediction models have been developed for zinc-fingers [10–16] and homeodomain (HD) [17], and have been reported to perform well on test data sets using various measures.

Current TF specificity prediction methods usually refer to prediction of specificity based on position weight matrices (PWMs). Most eukaryotic sequence specific TFs bind to 8–11 base pairs and hence their specificity is

† These authors contributed equally to this work

described by PWMs of equivalent width. On the other hand, the number of amino acids in the primary structure of the DNA-binding domains of TFs are much larger (e.g., 23 for zinc fingers, 58 for HDs). Most amino acids are required to maintain the three dimensional (3D) structure of TFs while a few are involved in determining specificity. Providing the entire amino acid sequence for predicting specificity at a given position, which is influenced by only a couple of residues, can result in overfitted models. Hence, identifying residues that influence specificity for a TF family under consideration is important. In previous studies such specificity influencing residues (SIRs) were determined either from structural information of the interacting positions in the protein and DNA or using variable selection from multiple alignments of proteins and their binding sites (or motifs). Although inferring SIRs from structural information is straightforward, rearrangement of side-chains at the protein-DNA interface do occur [16,18] making any one-to-one correspondence incomplete. Instead of relying on structural information, covariance based measures can be used to infer interacting positions. This approach works well for predicting base pairs in RNA structures because the interactions are mainly one-to-one. However, residue variations in a given structural family of functional proteins is constrained by its 3D structure with many-to-many contacts that can result in a chain of correlations and even superadditive correlations [19]. Lapedes and colleagues pointed out the problem and outlined a solution utilizing maximum entropy estimates of interaction parameters [20], and in 2002 showed that this could be an effective means of identifying the directly interacting positions in protein sequences [21]. Since then, several methods have been developed to disentangle directly and indirectly co-varying positions and shown to reliably predict protein structures from deep alignments [22–27] and even to demonstrate the ability to identify interacting residues between proteins in multi-protein complexes [27–29].

Here we apply a similar method to identify the SIRs in protein-DNA complexes. We extended three methods to infer direct from mixed correlations to infer SIRs from alignment of proteins and corresponding binding site motifs. The methods are compared with each other and a simple measure, mutual information (MI). We assessed the accuracy of the methods by mapping the identified SIRs to crystal structures.

## RESULTS

The protein domains of the four families used in this study are in the range of 46–64 amino-acids, and their specificity spans 5–9 degenerate bases. Only a few amino-acids in the protein domains (SIRs) determine the specificity. To identify SIRs from composite-alignments, four quantities, MI, adjusted mutual information (MIp), direct information (DI) and PSICOV (PC), were computed. Heat-maps representing MI, MIp, DI and PC for inter-molecular pairs are shown in Figure 1 for HD family. Heat-maps for other families are given in the Supplementary Materials (Figures S1, S2 and S3). Top ranked amino-acid residues that influence specificity, identified using DI, are exact or immediate neighbors of contact bases (as determined by crystal structures). Similar trends were observed when PC and MIp were used to rank the SIRs. However, when MI was used, the top ranked residues are not necessarily close in crystal structure.

Receiver operating characteristic (ROC) curves were generated to assess the compromise between true and false positive contact pairs identified using the four measures (Figure 2). The true positives and false positive pairs were identified using crystal structures as reference with different distance cut-offs. Generally, the ROCs are very steep until sensitivities of 0.4 to 0.5. When ROCs of different methods are compared for pairs identified using 4 Å cutoffs, it is evident that DI performs better than other methods especially in the left bottom quadrant. This is a critical area for ROCs and recommended as a good measure of classification performance. When the different methods are compared, for all families, DI is somewhat better consistently in identifying ~50%–60% of true contacts at a lower false positive rate. Performances of MIp and PC are not very different from DI, while performance of MI is worse than random.

The ROC computed using 4 and 5 Å cut-offs has a larger area under the curve as compared to 6 and 7 Å (Figure 2 and Figures S4–S6). This indicates that DI and PC avoid false positives to some extent and are better at identifying the closely interacting contacts. Except for GntR, 40% and 60% of the true contacts are identified at a false positive rate of around 10% and 25%, respectively. It should be noted that the ROCs are computed using crystal structure of one of the members as a reference. The protein-DNA interface is known to rearrange depending upon the crystal structure of the family member and mutation in protein that alters the specificity [16,18].

Next, we mapped top interacting amino-acid-specificity pairs identified by DI on the solved crystal structures. Contacts made by DNA bases to amino-acid were first identified from crystal structure (less than 5 Å) and used as reference (Figure 3). The top ranked pairs identified by DI, when mapped on crystal structures, show that most of the top ranked SIRs-specificity pairs are in physical proximity of each-other. A few pairs that are not physically close to each other were also identified in the top list, some probable reasons are given in discussions. The same maps using the other methods are included in Supplemental Figure S7.

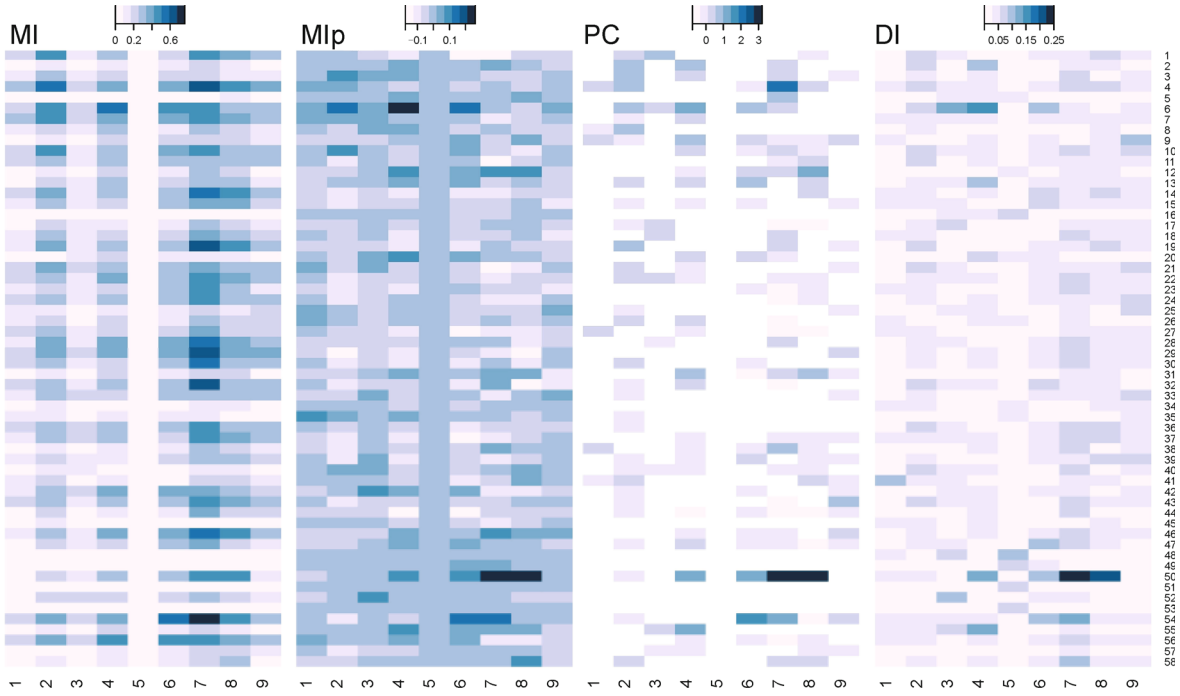


**Figure 1. Heat maps showing MI, MIp, PC and DI for intermolecular contact pairs for the HD family of TFs.** Color keys are given at the top of each plot. The *X*-axes correspond to specificity positions and *Y*-axes show the position of amino-acid for a given family.

## DISCUSSION

Simple approaches to predict SIRs, like MI and correlations, are prevalent in the literature. However, due to constraints imposed by the 3D structure of the protein, such measures fail to deduce correct SIRs given amino-acid sequence and motif alignments. Constraints imposed by 3D structures on evolution of protein results in false, super-additive or chained correlations. Such chained correlations are also observed in gene regulatory networks [30] and successful attempts to deduce direct correlation from indirect correlations have been proposed [22,24–26]. Here we used such approaches to deduce SIRs from alignments of transcription factor proteins and DNA binding site motifs.

Two methods based on inverse covariance matrix (direct information and graphical lasso (glasso)) and a method based on adjusting mutual information (MIp) depending upon background were used to identify SIRs. The results are compared with simple MI. Heatmaps (Figure 1), ROCs (Figure 2) and rendering of contacts on 3D structures (Figure 3) showed that the performance of DI, MIp and PC in identifying SIRs are mostly comparable and much superior to MI, with DI being somewhat better. DI is capable of extracting contacts even if the residues are fully conserved and specificity is invariant across the family, although fully conserved positions do not carry any information [24]. The performance gained from using the global methods (DI, PC and MIp) over a local method (MI) is achieved through the attempt to eliminate indirect couplings which is inherent to the MI. Differences in performance between the global methods are determined by how well they achieve to do this with the available data (e.g., depth of the alignments). Defining analytically (in the mathematical sense) what makes one global method superior to another is not straight forward as there are numerical approximations to different theories for solving the chaining problem.

Although the ROCs shown in Figure 2 are significantly better than random for three methods, there is room for improvements. Likely reasons for not capturing all the correct contacts might be a limited number of proteins in a family with known specificity. For the successful protein contact prediction using similar approaches, the number of sequences in the alignment should be 5 times the length of protein sequence [24]. However, unlike the number of known protein orthologs, the number of TFs for which the specificity is known is very limited. To meet this challenge, we limited the protein sequences to their domains, thereby reducing the width of the alignment

 

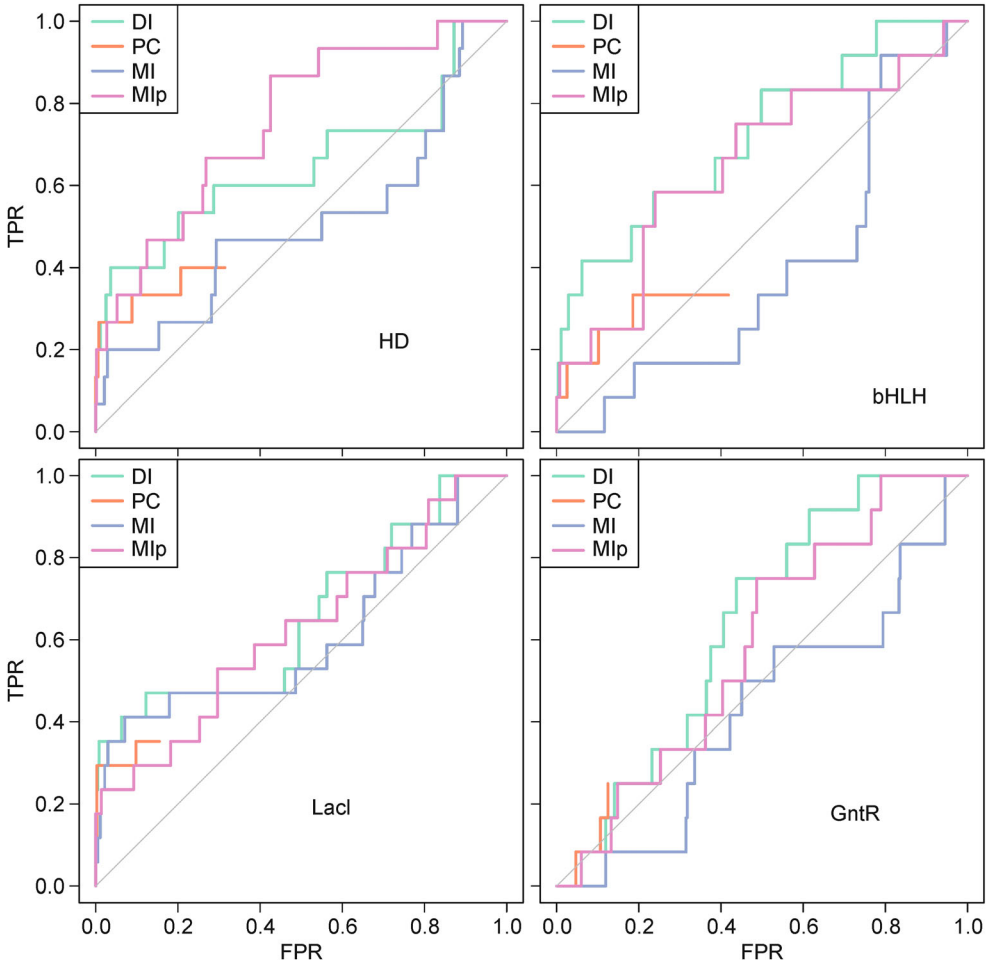


**Figure 2. ROC showing the trade-off between true and false positives.** True or false positive contacts were assigned based on crystal structures. The contacts were calculated using 5Å distance cut-offs. As PC only outputs a subset of the contacts, the ROCs are truncated. ROCs computed using 4, 6 and 7Å are given in Supplementary Materials.

considerably, which in turn reduces the demand on the alignment depth. As is stated in Table 1, there is a different number of samples for each TF and as low as 87 sequences for bHLH still show an improvement over MI. It should be noted that in the current approach the contacts present in crystal structure were considered as a ruler to compare performance, however, reports of sidechain rearrangement of SIRs depending upon different DNA partners is well known.

The major usefulness of identifying correct SIRs is in specificity prediction methods [17]. We attempted to develop a random forest based specificity predictor using SIRs identified by DI and PC for the HD family using a similar approach as Christensen et al. [17], however, our performance is comparable to that of this previous study. This is expected as the Christensen et al [17] uses MIp to identify SIRs, which here shows comparable performance to DI and PC, at least with the given depth of the alignments.

## CONCLUSIONS

In the present article we used four methods to identify SIRs from transcription factor amino-acid sequence and corresponding specificity (given by PWMs). While MI fails to correctly identify SIRs from alignments, other methods (MIp, PC and DI) are comparable and are much better than MI. The alignments of amino-acid and motifs along with software to calculate DI and PC are available at http://stormo.wustl.edu/SpecPred.

## METHODS

### Database

Protein and specificity collection

We used variable selection and specificity prediction methods for two eukaryotic families (HD and bHLH) and

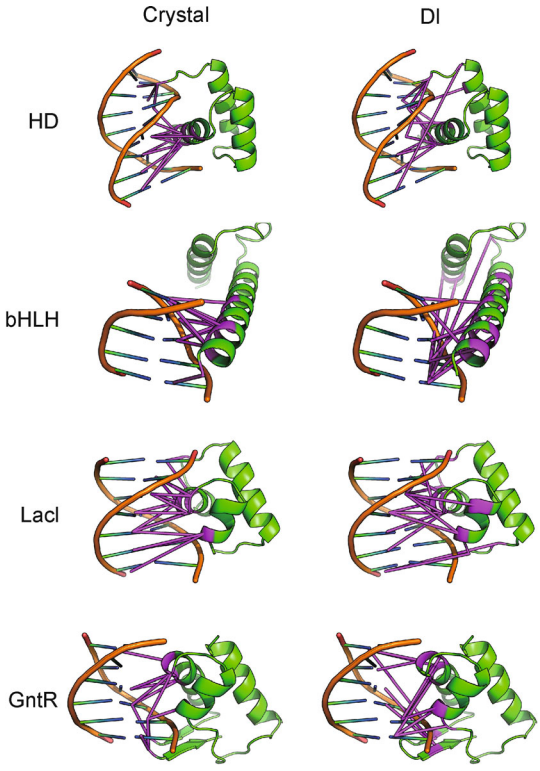


**Figure 3. Contacts from the crystal structure and predicted by DI.** Actual (within 5Å) and DI predicted SIRs-specificity contacts (Only top N contacts with highest DI are shown, where N is the number of contacts identified within 5Å in the crystal structure) traced on the solved TF-DNA complex. TF and DNA are rendered in green and orange backbone, respectively. The interacting or top ranked pairs are shown by a connection using magenta lines. Such lines in the left panel shows amino-acid-base contacts inferred from crystal structure and right panels show pairs from the top of the sorted list of DIs. The crystal structure used are 9ANT, 1NK4, 1EFA and 4EGY for HD, bHLH, LacI and GntR family, respectively. SDRs identified by other methods are rendered in Supplementary Materials.

two prokaryotic families (LacI and GntR). Sources from which specificity or preferred DNA sequences are derived are given in Table 1. Most specificities of eukaryotic TFs were derived from FlyFactorSurvey [31], UniProbe [32] and [33]. Specificities of prokaryotic TFs were derived from RegPrecise [34]. RegPrecise does not include experimentally determined specificity but manually curated, inferred binding sites. The corresponding amino-acid sequences of TFs were collected using the accession numbers given in the respective source of specificities or preferred sequences mostly from UniProt [35] and Microbes Online [36].

**Table 1. Sources and number of TFs for different families.**

| Family | Source | Number of instances |
|---|---|---|
| HD | FlyFactorSurvey | 85 |
| | UniProbe (BEEML-PBM) | 127 |
| | Jolma et al., 2013 | 84 |
| bHLH | FlyFactorSurvey | 31 |
| | UniProbe | 21 |
| | Jolma et al., 2013 | 35 |
| LacI | RegPrecise | 404 |
| GntR | RegPrecise | 977 |

## Data set curation

A number of proteins for LacI and GntR families available on RegPrecise were excluded from our model. Proteins were filtered based on the following criteria: i) The referred protein sequences does not contain TF family motif, ii) Position frequency matrices (PFMs) inferred from preferred binding sites were not aligned with most other PFMs of the given family. Some obvious outliers in the aligned motifs were also removed following visual inspection.

## Alignment of proteins

Amino acid sequences were aligned using *hmmalign* module of HMMER [37] using HMM-profiles from PFAM [38]. For HD, bHLH, LacI and GntR, profiles with accession IDs PF00046, PF00010, PF00356 and PF00392, respectively, were used. Sequence logos for the generated alignments are given in Figure 4.

## Encoding of motif specificity and alignment

Specificities available from UniProbe, FlyFactorSurvey and [33]are represented in the form of PWMs. The PWMs derived using BEEML-PBM for UniProbe experiments were used after applying a scaling factor as suggested by [17]. PWMs represented in the form of energy were converted to PFMs assuming they are Boltzmann distributed. The binding sites for IFs retrieved from RegPrecise were converted to PFMs. The PFMs were encoded to 15 letter IUPAC encoding [39]. PFMs were converted to 15 bits encoding using Euclidean distance as a measure to discretize PFMs represented in IUPAC space as described in Wang and Stormo [39,40]. We refer to such encoded PFMs as specificity-strings. The encoding of specificity is essential as the amino-acid sequences are sets of discrete alphabets and it is convenient to measure

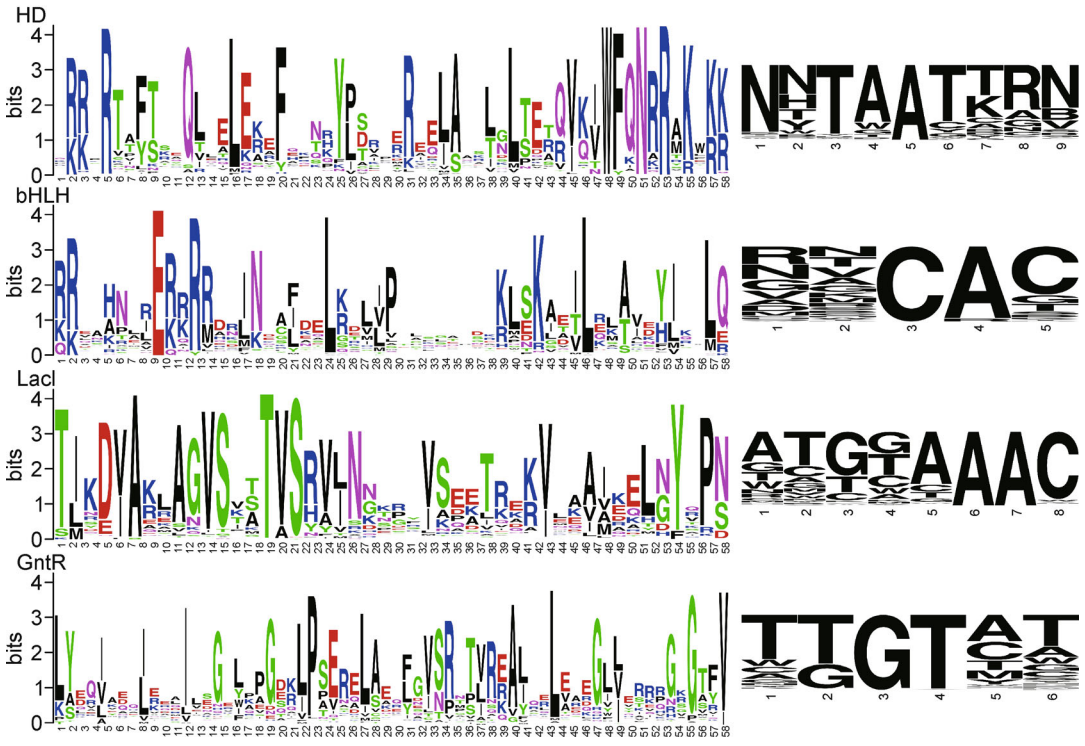


**Figure 4. Sequence conservation for proteins and binding sites.** Logos computed using amino-acid sequences for different families are given in the left panel. PFMs were discretized using the 15 digit standard IUPAC coding. Frequencies of alphabets at a given position is graphically shown in the right panel. The size of the letter is proportional to its frequency.

correlation between discrete variables by converting PFMs to discrete specificity-strings. The PFMs were aligned before encoding them into specificity-strings.

Aligning motifs of high sequence identity can be accomplished by multiple motif alignment software like STAMP [41]. However, in the present study, the number of motifs to align is a few hundreds (for HD, LacI and GntR). In order to align motifs efficiently for large families, we first selected a few representative motifs (half sites for LacI and GntR due to symmetric binding site) and manually aligned them. After that, each of the motifs in the family was aligned to each selected representative motif using MatAlign [40]. The best alignment to each representative motif was identified using the *p*-value. Subsequently, all aligned motifs in a group were aligned to each other using the alignment of reference motifs. The aligned proteins and encoded specificities for a given family were then used for identification of SIRs with the methods described below. The alignments are available at http://stormolab.wustl.edu/SpecPred.

## Identification of SIRs using inverse covariance matrix

Correlations between aligned TFs amino-acid sequences and corresponding specificity-strings are a result of direct and indirect correlations. The direct correlations result from residues that influence the specificity. As these residues at the protein-DNA interface change, the corresponding preferred binding site and hence specificity changes. However, these amino-acids that determine the specificity are part of a protein and therefore co-evolved with one or few other amino-acids to maintain a three-dimensional structure characteristic to that family. This gives rise to confounding indirect correlations [20]. Several successful approaches have been suggested to disentangle direct from indirect interactions to interpret co-evolving protein residues (contact residue prediction) from multiple sequence alignments [20,22,23,25–27,30]. We extended two of these methods: i) PSICOV which uses graphical lasso (glasso) and average product correction (APC) [23], and ii) direct information [26,27] for SIRs prediction from amino-acid-specificity alignments.

Both methods are based on inversion of the co-variance matrix which has a unique property to disentangle direct and indirect information. The inverse covariance matrix has a direct relation to multiple linear regressions and computes a linear influence of a given residue on specificity in the presence of amino-acids at other positions. For $n$ continuous variables ($x_1$, $x_2$,…, $x_n$), the interpretation of the elements of the inverse covariance

matrix ($V^{-1}$) in terms of how the variable under consideration (e.g. $x_1$) is related to other variables (e.g. $x_2$, $x_3$, … , $x_n$) is described in Equation 1 [42].

$$V^{-1}=\begin{bmatrix} 1/[s_{11}(1-R_1^2)] & -\beta_{12}/[s_{11}(1-R_1^2)] & \cdots & -\beta_{1n}/[s_{11}(1-R_1^2)] \\ -\beta_{21}/[s_{22}(1-R_2^2)] & 1/[s_{22}(1-R_2^2)] & \cdots & -\beta_{2n}/[s_{22}(1-R_2^2)] \\ \vdots & \vdots & \cdots & \vdots \\ -\beta_{n1}/[s_{nn}(1-R_n^2)] & -\beta_{n2}/[s_{nn}(1-R_n^2)] & \cdots & 1/[s_{nn}(1-R_n^2)] \end{bmatrix} \quad (1)$$

where $\beta_{ij}$ is the regression coefficient for variable $j$ and $R_i^2$ is the coefficient of determination when variable $x_i$ is used as the response vector and the other variables as predictors. $s_{ii}$ is the variance of variable $x_i$.

### Direct information

The maximum entropy based frame-work to identify direct correlation for protein contact prediction [3,26] was adapted to identify the SIRs (see details after description of model). Briefly, the approach measures mutual information that results from direct interactions only. A two-site model is described as

$$P_{ij}^{dir}(A,B)=\frac{\exp\left(e_{ij}(A,B)+h_i(A)+h_j(B)\right)}{\sum_{a=1}^{q-1}\sum_{b=1}^{q-1}\exp\left(e_{ij}(a,b)+h_i(a)+h_j(b)\right)}. \quad (2)$$

The coupling term $e_{ij}$ is computed using inverse covariance matrix and $h_i$ and $h_j$ are computed iteratively by imposing marginal frequency restraints given the pair interaction term $e_{ij}$. The denominator is the partition function, which enters as a normalization factor.

For position $i$ and $j$ the above procedure results in a matrix of dimensions $(q-1)\times(q-1)$ which is converted into a metric, namely, the direct information ($DI_{ij}$), derived from mutual information formula using

$$DI_{ij}=\sum_{a=1}^{q-1}\sum_{b=1}^{q-1}P_{ij}^{dir}(a,b)\ln\frac{P_{ij}^{dir}(a,b)}{f_i(a)f_j(b)}. \quad (3)$$

Derivation and details of the model are given in the original papers [25–27] and are summarized in the Supplementary Materials.

To accommodate the DNA-protein contact inference we had to adapt the structure of the co-variance matrix from that presented in [26,27]. For the purpose of intra-molecular contact inference of the protein, as there is no encoded specificity involved, the covariance matrix is square and of order $n \times q$, where $n$ is the width of alignment (number of amino-acids per sequence) and $q$ is 21 (for 20 amino-acids and a gap state). For the inter-molecular contact inference, the covariance matrix is square and of order $(n \times q_a) + (m \times q_s)$, where $n$ and $m$ are the width of the amino-acid and specificity alignments, respectively. $q_a$ and $q_s$ are 21 (number of possible states of amino-acids) and 15 (number of possible states of encoded specificity), respectively.

### Graphical lasso (glasso) and APC

The approach and scoring is similar to a method published recently [23]. However, a few modifications are introduced in addition to the size of co-variance matrix. The observed frequencies of amino-acids were reweighted as in the computation of DI (Supplementary Materials). No pseudo-counts were added for adjusting the marginal and joint frequencies, as the glasso method is capable of inverting the matrix without pseudo-counts. In addition, the APC score is capable of adjusting for entropic bias that might be caused by not adding pseudo-counts. The inverse matrix was estimated using glasso with the regularization parameter (*rho*) set to 0.03. Once the inverse covariance matrix is available, the L1-norm is computed for converting the $q \times q$ matrix to a scalar value followed by computation of product correction (PC) score as described in original report and supplementary methods [23].

### Identification of SIRs using MI and MIp

Mutual information is a simple measure to compute correlation between two discrete variables. We used mutual information and an adjusted mutual information [43] in addition to the above methods and denote them as MI and MIp, respectively. The latter adjusts mutual information using APC.

## Validation

Heat maps were used to visualize MI, MIp, DI and PC for amino-acids and specificity contacts. Top ranked SIRs were mapped on the crystal structure of corresponding families in order to determine the physical proximity in solved structures. ROC curves were generated to gauge the performance against a validation set comprising contacts from DNA-protein co-crystal structures determined with different distance cutoffs.

## SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at DOI 10.1007/s40484-015-0045-y.

## ACKNOWLEDGEMENTS

This work was supported by NIH grant HG000249 to GDS. We also thank Chris Workman (DTU Systems Biology) and the Otto Monsted Foundation (J. NO. 13-70-1193) for the support of CG during his visit to Washington University.

## COMPLIANCE WITH ETHICS GUIDELINES

Ronak Y. Patel, Christian Garde and Gary D. Stormo declare they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

# REFERENCES

1. Balwierz, P. J., Pachkov, M., Arnold, P., Gruber, A. J., Zavolan, M. and van Nimwegen, E. (2014) ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. Genome Res., 24, 869–884
2. Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. Science, 342, 1235587
3. Wright, D. A., Li, T., Yang, B. and Spalding, M. H. (2014) TALEN-mediated genome editing: prospects and perspectives. Biochem. J., 462, 15–24
4. Mendenhall, E. M., Williamson, K. E., Reyon, D., Zou, J. Y., Ram, O., Joung, J. K. and Bernstein, B. E. (2013) Locus-specific editing of histone modifications at endogenous enhancers. Nat. Biotechnol., 31, 1133–1136
5. Lin, Y., Chomvong, K., Acosta-Sampson, L., Estrela, R., Galazka, J. M., Kim, S. R., Jin, Y. S. and Cate, J. H. (2014) Leveraging transcription factors to speed cellobiose fermentation by *Saccharomyces cerevisiae*. Biotechnol. Biofuels, 7, 126
6. Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., Yan, K. K., Dong, X., Djebali, S., Ruan, Y., et al. (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res., 22, 1658–1667
7. Haynes, B. C., Maier, E. J., Kramer, M. H., Wang, P. I., Brown, H. and Brent, M. R. (2013) Mapping functional transcription factor networks from gene expression data. Genome Res., 23, 1319–1328
8. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. and Luscombe, N. M. (2009) A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet., 10, 252–263
9. Matthews, B. W. (1988) No code for recognition. Nature, 335, 294–295
10. Benos, P. V., Lapedes, A. S. and Stormo, G. D. (2002) Probabilistic code for DNA recognition by proteins of the EGR family. J. Mol. Biol., 323, 701–727
11. Gupta, A., Christensen, R. G., Bell, H. A., Goodwin, M., Patel, R. Y., Pandey, M., Enuameh, M. S., Rayla, A. L., Zhu, C., Thibodeau-Beganny, S., et al. (2014) An improved predictive recognition model for $Cys_2$-$His_2$ zinc finger proteins. Nucleic Acids Res., 42, 4800–4812
12. Kaplan, T., Friedman, N. and Margalit, H. (2005) Ab initio prediction of transcription factor targets using structural knowledge. PLoS Comput. Biol., 1, e1
13. Liu, J. and Stormo, G. D. (2008) Context-dependent DNA recognition code for $C_2H_2$ zinc-finger transcription factors. Bioinformatics, 24, 1850–1857
14. Persikov, A. V., Osada, R. and Singh, M. (2009) Predicting DNA recognition by $Cys_2His_2$ zinc finger proteins. Bioinformatics, 25, 22–29
15. Persikov, A. V. and Singh, M. (2014) *De novo* prediction of DNA-binding specificities for $Cys_2His_2$ zinc finger proteins. Nucleic Acids Res., 42, 97–108
16. Wolfe, S. A., Nekludova, L. and Pabo, C. O. (2000) DNA recognition by $Cys_2His_2$ zinc finger proteins. Annu. Rev. Biophys. Biomol. Struct., 29, 183–212
17. Christensen, R. G., Enuameh, M. S., Noyes, M. B., Brodsky, M. H., Wolfe, S. A. and Stormo, G. D. (2012) Recognition models to predict DNA-binding specificities of homeodomain proteins. Bioinformatics, 28, i84–i89
18. Stormo, G. D. (2013) Introduction to protein-DNA interactions: structure, thermodynamics, and bioinformatics. NewYork: Cold Spring Harbor Laboratory Press.
19. Giraud, B. G., Heumann, J. M. and Lapedes, A. S. (1999) Superadditive correlation. Phys. Rev. E, 59, 4983–4991
20. Lapedes, A. S., Giraud, B., Liu, L.C. and Stormo, G. D. (1999) Correlated mutations in models of protein sequences: phylogenetic and structural effects. The institute of mathematical statistics lecture notes-monograph series, 33, 236–256
21. Lapedes, A., Giraud, B. and Jarzynski, C. (2002) Using sequence alignments to predict protein structure and stability with high accuracy. q-bio. QM, arXiv, 1207.2484
22. Cocco, S., Monasson, R. and Weigt, M. (2013) From principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction. PLoS Comput. Biol., 9, e1003176
23. Jones, D. T., Buchan, D. W., Cozzetto, D. and Pontil, M. (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics, 28, 184–190
24. Kamisetty, H., Ovchinnikov, S. and Baker, D. (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc. Natl. Acad. Sci. USA, 110, 15674–15679
25. Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R. and Sander, C. (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS One, 6, e28766
26. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T. and Weigt, M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. USA, 108, E1293–E1301
27. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. and Hwa, T. (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc. Natl. Acad. Sci. USA, 106, 67–72
28. Burger, L. and van Nimwegen, E. (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol. Syst. Biol., 4, 165
29. Ovchinnikov, S., Kamisetty, H. and Baker, D. (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife, 3, e02030
30. Feizi, S., Marbach, D., Médard, M. and Kellis, M. (2013) Network

 

deconvolution as a general method to distinguish direct dependencies in networks. Nat. Biotechnol., 31, 726–733

31. Zhu, L. J., Christensen, R. G., Kazemian, M., Hull, C. J., Enuameh, M. S., Basciotta, M. D., Brasefield, J. A., Zhu, C., Asriyan, Y., Lapointe, D. S., et al. (2011) FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. Nucleic Acids Res., 39, D111–D117
32. Robasky, K. and Bulyk, M. L. (2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res., 39, D124–D128
33. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013) DNA-binding specificities of human transcription factors. Cell, 152, 327–339
34. Novichkov, P. S., Kazakov, A. E., Ravcheev, D. A., Leyn, S. A., Kovaleva, G. Y., Sutormin, R. A., Kazanov, M. D., Riehl, W., Arkin, A. P., Dubchak, I., et al. (2013) RegPrecise 3.0—a resource for genome-scale exploration of transcriptional regulation in bacteria. BMC Genomics, 14, 745
35. Magrane, M. and Consortium, U. (2011) UniProt Knowledgebase: a hub of integrated protein data. Database, 2011, bar009
36. Dehal, P. S., Joachimiak, M. P., Price, M. N., Bates, J. T., Baumohl, J. K., Chivian, D., Friedland, G. D., Huang, K. H., Keller, K., Novichkov, P. S., et al. (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res., 38, D396–D400
37. Eddy, S. R. (2011) Accelerated profile HMM searches. PLoS Comput. Biol., 7, e1002195
38. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014) Pfam: the protein families database. Nucleic Acids Res., 42, D222–D230
39. Wang, T. and Stormo, G. D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics, 19, 2369–2380
40. Wang, T. and Stormo, G. D. (2005) Identifying the conserved network of *cis*-regulatory sites of a eukaryotic genome. Proc. Natl. Acad. Sci. USA, 102, 17400–17405
41. Mahony, S. and Benos, P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res, 35 (Web Server issue), W253–W258
42. Kwan, C. (2014) A regression-based interpretation of the inverse of thesample covariance matrix. Spreadsheets in Education (eJSiE), 7, Article 3
43. Dunn, S. D., Wahl, L. M. and Gloor, G. B. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics, 24, 333–340

 