1-1-2011

Protein-DNA Recognition Models for the Homeodomain and C2H2 Zinc Finger Transcription Factor Families

Ryan Christensen
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Computational Biology

Dissertation Examination Committee:
Gary D. Stormo, Chair
Michael R. Brent
Jeremy D. Buhler
James J. Havranek
Garland R. Marshall
Robi D. Mitra

PROTEIN-DNA RECOGNITION MODELS FOR THE HOMEODOMAIN AND C2H2 ZINC FINGER TRANSCRIPTION FACTOR FAMILIES

by

Ryan Goaslind Christensen

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2011

Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Protein-DNA Recognition Models for the Homeodomain and C2H2 Zinc Finger Transcription Factor Families

By

Ryan Goaslind Christensen

Doctor of Philosophy in Biology and Biomedical Sciences

Computational Biology

Washington University in St. Louis, 2011

Professor Gary D. Stormo, Chairperson

Transcription factors (TFs) play a central role in the gene regulatory network of each cell. They can stimulate or inhibit transcription of their target genes by binding to short, degenerate DNA sequence motifs. The goal of this research is to build improved models of TF binding site recognition. This can facilitate the determination of regulatory networks and also allow for the prediction of binding site motifs based only on the TF protein sequence. Recent technological advances have rapidly expanded the amount of quantitative TF binding data available. PBMs (Protein Binding Microarrays) have recently been implemented in a format that allows all 10mers to be assayed in parallel. There is now PBM data available for hundreds of transcription factors. Another fairly recent technique for determining the binding preference of a TF is an in vivo bacterial one-hybrid assay (B1H). In this approach a TF is expressed in *E. coli* where it can be used to select strong binding sites from a library of randomized sites located upstream of a weak promoter, driving expression of a selectable

gene. When coupled with high throughput sequencing and a newly developed analysis method, quantitative binding data can be obtained. In the last few years, the binding specificities of hundreds of TFs have been determined using B1H.

The two largest eukaryotic transcription factor families are the zf-C2H2 and homeodomain TF families. Newly available PBM and B1H specificity models were used to develop recognition models for these two families, with the goal of being able to predict the binding specific of a TF from its protein sequence. We developed a feature selection method based on adjusted mutual information that automatically recovers nearly all of the known key residues for the homeodomain and zf-C2H2 families. Using those features we find that, for both families, random forest (RF) and support vector machine (SVM) based recognition models outperform the nearest neighbor method, which has previously been considered the best method.

Acknowledgements

I want to thank my advisor, Gary Stormo for giving me the opportunity to join his lab and for everything he has taught. Not only is Gary an excellent scientist, he is a kind and patient mentor. I have appreciated his open door policy and his willingness to talk and give advice when ever I have questions.

I thank the funding organizations that have made this work possible. I was supported initially by National Institutes of Health training grant GM08802. I was also supported by National Institutes of Health grant GM078369.

I want to thank all of the members of the Stormo lab for their friendship and for creating a very collegial and supportive atmosphere. In particular, I wish to thank Yue Zhao for countless discussions and valuable insight and advice on every topic under the sun, as well as regarding science. I thank my lab mates, for help, encouragement and for their friendship. In particular, I wish to thank Gurmukh Sahota, Nnamdi Ihuegbu, Aaron Spivak, Yue Zhao and Barrett Foat.

My collaborators in the Wolfe lab at the University of Massachusetts have been a pleasure to work with and I thank them for the opportunity to work with them on a variety of exciting projects. In particular I thank Scot Wolfe, Marcus Noyes, and Ankit Gupta.

Finally, I need to thank my family. My parents have been encouraging and supportive and have given me so much. I especially thank my wife Susan for her love and continued support and for all of the sacrifices she has made on my behalf. I also thank my children Claire, William and Rose for the joy and purpose they bring to my life.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Transcription Factors

Transcription factors (TFs) are important components of gene regulatory networks. They bind to short degenerate DNA motifs, activating or inhibiting their target genes by either recruiting the transcriptional machinery or blocking it. The general principles of transcription regulation were first deduced from studies of the *lac* operon system in *E. coli* by Jacob and Monod (Jacob and Monod, 1961). They showed that the *lac* operon has cis-acting elements in a promoter region that are controlled by trans-acting proteins that bind to these regulator elements, either activating or repressing transcription.

TFs make up a significant percentage of the human genome, about 6 to 10% (Babu et al., 2004; Vaquerizas et al., 2009). However only a small percentage of the DNA motifs that these TFs bind have been determined. TFs also play an important role in disease. For instance, TFs are over represented among oncogenes (Furney et al., 2006; Vaquerizas et al., 2009). A third of the genes linked to birth defects in OMIM are transcription factors (Boyadjiev and Jabs, 2000).

C2H2 Zinc Finger and Homeodomain Families

There are many different types of transcription factors that utilize a wide variety of protein folds to bind DNA and recognize specific sites (Garvie and Wolberger, 2001; Luscombe et al., 2000; Luscombe and Thornton, 2002). The two largest transcription factor families in most metazoan genomes are the C2H2 zinc finger (hereafter referred to as ZF) family and the homeodomain (hereafter referred to as HD) family (Tupler et al., 2001).

Both of these families utilize a relatively small set of key residues to bind to DNA specifically (Ekker et al., 1994; Puppin et al., 2011; Wolfe et al., 2000), and they are capable of binding to a wide variety of different types of binding sites. It has been proposed that this ability to bind a wide array of possible DNA motifs is why these families have expanded so quickly (Itzkovitz et al., 2006; Luscombe and Thornton, 2002). The reasoning is that if only a few mutations at key residues can lead to a wide variety of different binding specificities, than a TF family should be able to expand through duplication successfully.

The HD and ZF are both relatively small domains, 30 and 60 amino acids long, respectively. Both families use an alpha helix, or recognition helix, to bind in the major groove. HD proteins also bind in the minor groove via an N-terminal arm (see chapters 2). ZF proteins generally contain tandem repeats of ZF domains that bind to overlapping

subsites (see chapter 5). The HD family is a subfamily of the HTH class of proteins, which are very abundant in prokaryotic genomes, although ZFs are not.

Protein-DNA Recognition Code

It has been a long standing goal to determine the rules that govern how TFs specifically recognize their target binding sites (Pabo and Sauer, 1984; Seeman et al., 1976). Initially, it was hoped that there would be a simple deterministic recognition code that once decoded would allow the regulatory network of the cell to be deciphered, just as learning the genetic code allowed protein sequences to be decoded. In 1976, before there were any known structures of protein-DNA complexes, Seeman et al. (Seeman et al., 1976) proposed a simple recognition code in which Arg would specifically bind to G-C base-pairs and Asn and Gln would specifically bind to A-T base-pairs. As the first crystal structures of TFs bound to DNA became available, however, it became apparent that there was no simple, universal recognition code (Mandel-Gutfreund et al., 1995; Matthews, 1988). The history of the idea of a recognition code is reviewed in Benos et al. (Benos et al., 2002b). Benos et al. point out that although there is no simple universal code, it may be possible to construct probabilistic models for individual families. A major obstacle for parameterizing these types of models is the size of the possible parameter space (Benos et al., 2001). Even for a motif

like the ZF domain that only has about three key residues that contact DNA, there are 8000 possible different ZF domains. However, if it can be assumed that all interactions are additive, then this reduces to only 60 different measurements. The extent to which interactions between key residues are additive is a matter of debate (Liu and Stormo, 2008). Until recently, not much effort had been spent on developing an HD recognition code. A recent paper (Alleyne et al., 2009) proposes that the a simple nearest neighbor model may have the best performance that can be hoped for, given the current data. Further more, the authors propose that the nearest neighbor approach may be essentially as good as or better more sophisticated machine learning methods for other TF families as well.

Methods For Determining TF Binding Preferences

Many different techniques have been developed to determine the affinity of TFs for specific DNA sites. Two of the most utilized methods are the DNase footprinting assay (Galas and Schmitz, 1978) and the Electrophoretic mobility shift assay (EMSA)(Fried and Crothers, 1981; Garner and Revzin, 1981). They can be used to determine the equilibrium disassociation constant (K_d) of a TF for a particular site by varying the concentration of the TF. The K_d is a measure of the absolute affinity of a TF for a site. These methods are low throughput and laborious. In the end, it is not very helpful for elucidating regulator

networks to know the K_d of just a few sites. What we really want to know is the relative binding affinity to any given site in the genome.

Biophysical Model of Binding

TFs do not bind to DNA in a binary fashion. The concentration of DNA is so high with in a prokaryotic cell or a eukaryotic nucleus that a TF is effectively always bound to DNA. Roughly speaking, there are 10^7 potential binding sites in a prokaryotic genome. With a cell volume of about 10^{-15} L, that is a concentration of sites in the mM range, which is well above the non-specific K_d for any TF, the affinity it has for DNA in general.

A given TF will bind to any site in the genome with some probability. However, it will generally have high affinity for only a small number of sites, but this is also a function of the TF's concentration (Bussemaker et al., 2007; Zhao et al., 2009). A TF binding to the DNA sequence S_i is a thermodynamic processes governed by two rate constants, k_{on} and k_{off}:

$$TF + S_i \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} TF \cdot S_i$$

These rate constants define the affinity of the TF for the site, S_i, expressed in terms of the equilibrium disassociation constant:

$$K_d(S_i) = \frac{[TF][S_i]}{[TF \cdot S_i]} = \frac{k_{off}}{k_{on}}$$

The occupancy of site S_i by the TF, or the probability that site S_i will be bound by the TF is a function of the concentration of the TF and of the disassociation constant:

$$P(S_i) = \frac{[TF \cdot S_i]}{[TF \cdot S_i] + [S_i]} = \frac{1}{1 + \frac{1}{K_d[TF]}} = \frac{1}{1 + e^{E_i - \mu}}$$

where $P(S_i)$ is the probability that site S_i will be bound by the TF, $E_i = -\ln(K_i)$ is the Gibbs free energy of binding in units of the RT (the gas constant times the temperature in degrees Kelvin) and $\mu = \ln[TF]$ is the chemical potential (Zhao et al., 2009).

Models of TF Specificity

Models of binding are desirable, since they can provide insight into the mechanism of binding. Also, models can be parameterized using multiple different types of data. This can help to filter out noise and biases inherent to different experimental techniques. On the one extreme, the model of DNA specificity with the greatest number of parameters is just a list where every possible site of length L, of which there are 4^L, has an energy value assigned to it. This provides no real insight in to the mechanism of binding, however. On the other hand, the simplest possible model is just the TFs consensus sequence (Schneider, 2002).

In 1982, Stormo et al. first introduced the concept of a PWM (Position Weight Matrix) to represent RNA sites that function as translation initiation sites in *E. coli* (Stormo

et al., 1982). PWM models assume that binding to one position of a DNA site is independent of binding to any other position of that site (Stormo, 2000). Because the energy of binding to each position of the motif is assumed to be independent, these energies can be added together to obtain the overall energy of binding to any site of length L. The extent to which additivity is a valid approximation is still an open question (Alleyne et al., 2009; Badis et al., 2009; Benos et al., 2002a; Maerkl and Quake, 2007b; O'Flanagan et al., 2005). Careful analysis has resolved many of the concerns that have been raised, showing that additivity tends to be a reasonable assumption in most cases (Benos et al., 2002a; Stormo and Zhao, 2007; Zhao and Stormo, 2011).

Additivity is certainly a very useful assumption. If the partial free energies of binding to each position of a site are independent, than it is not necessary to measure the affinity to every site. To completely parameterize a PWM model then, it is not necessary to measure affinities for all 4^L sites, but only for 4L different sites, greatly shrinking the parameter space. PWMs also have many other useful properties. PWMs can be aligned (Mahony et al., 2007) and easily used to score all potential binding sites in a genome. PWMs can score all possible L long sequences, even if they are only trained on a handful of sequences, so they can help impute missing data. When necessary, additional higher order interactions terms, such as terms describing interactions between adjacent nucleotides, can be add to augment a PWM

model (Stormo, 2011). At the very least, they can serve as null models to help detect non-additive interactions between motif positions.

Recent Technological Advances

In order to parameterize protein-DNA recognition models, it is important to be able to measure the affinities for many different TFs, since the parameter space of possible TFs is very large. Recently, technological advances have greatly increased the rate at which new TFs can be analyzed (Stormo and Zhao, 2010). These new technologies include MITOMI (Fordyce et al., 2010; Maerkl and Quake, 2007a), a method that utilize microfluidics to measure affinities for all 8mers; CSI (Cognate Site Identification), a similar technique to PBMs that uses micro arrays (Hauschild et al., 2009; Puckett et al., 2007; Warren et al., 2006). This dissertation will focus on three main methods: PBMs (Protein Binding Microarrays), HT-SELEX (high throughput Systematic Evolution of Ligands by Exponential Enrichment), and the bacterial one-hybrid assay (B1H) (Meng et al., 2005; Meng and Wolfe, 2006) (See figure 1.1).

PBMs (Protein Binding Microarrays) use double stranded DNA micro arrays to obtain values proportional to affinity values for all 10mers (Berger et al., 2006; Bulyk et al., 2001; Mukherjee et al., 2004). SELEX (Systematic Evolution of Ligands by Exponential Enrichment) has long been used to determine the specificity of TFs. Before the advent of next generation sequencing, initially SELEX was used obtain a few high affinity sites

(Blackwell and Weintraub, 1990; Oliphant et al., 1989; Tuerk and Gold, 1990; Wright et al., 1991). Coupled with massively parallel Illumina sequencing, high throughput or HT-SELEX can now determine highly accurate specificities for TFs (Zhao et al., 2009).

The B1H assay is an *in vivo* method that uses a reporter gene to select for randomized binding sites. In this approach a TF is expressed in *E. coli* fused to the ω subunit of RNA Polymerase, turning any DNA binding protein into a transcriptional activator. A library of randomized binding sites located upstream of a weak promoter, driving expression of a selectable gene. Only sites with high affinity for the TF will survive selection. In the last few years, the binding specificities of hundreds of TFs have been determined using B1H. One of the main advantages of B1H is that unlike PBMs or SELEX, the transcription factor of interest does not need to be purified. We show in chapter 4 that when coupled with high throughput sequencing and a new analysis method which we introduce, that the B1H assay yields accurate models able to predict HT-SELEX results.

Recently, new *in vivo* technologies such as ChIP-Chip (Ren et al., 2000) and more recently ChIP-Seq (Johnson et al., 2007) have been applied to determine the binding sites of TFs on a genomic scale. This begs, the question, if *in vivo* methods can determine binding sites directly, are in vitro technologies still relevant? There are many factors at work that together determine which sequences are bound. Chromatin structure, for instance, can determine which regions of the genome are accessible to TFs. It is possible that otherwise

high affinity sites are not available to a TF due to competition with other factors. At the very least, *in vitro* models can provide a valuable null model. When *in vivo* binding profiles deviate from in vitro results, it means that additional factors are at work. Also, methods like B1H are simple to perform and do not require the purification or pull-down of the TFs.

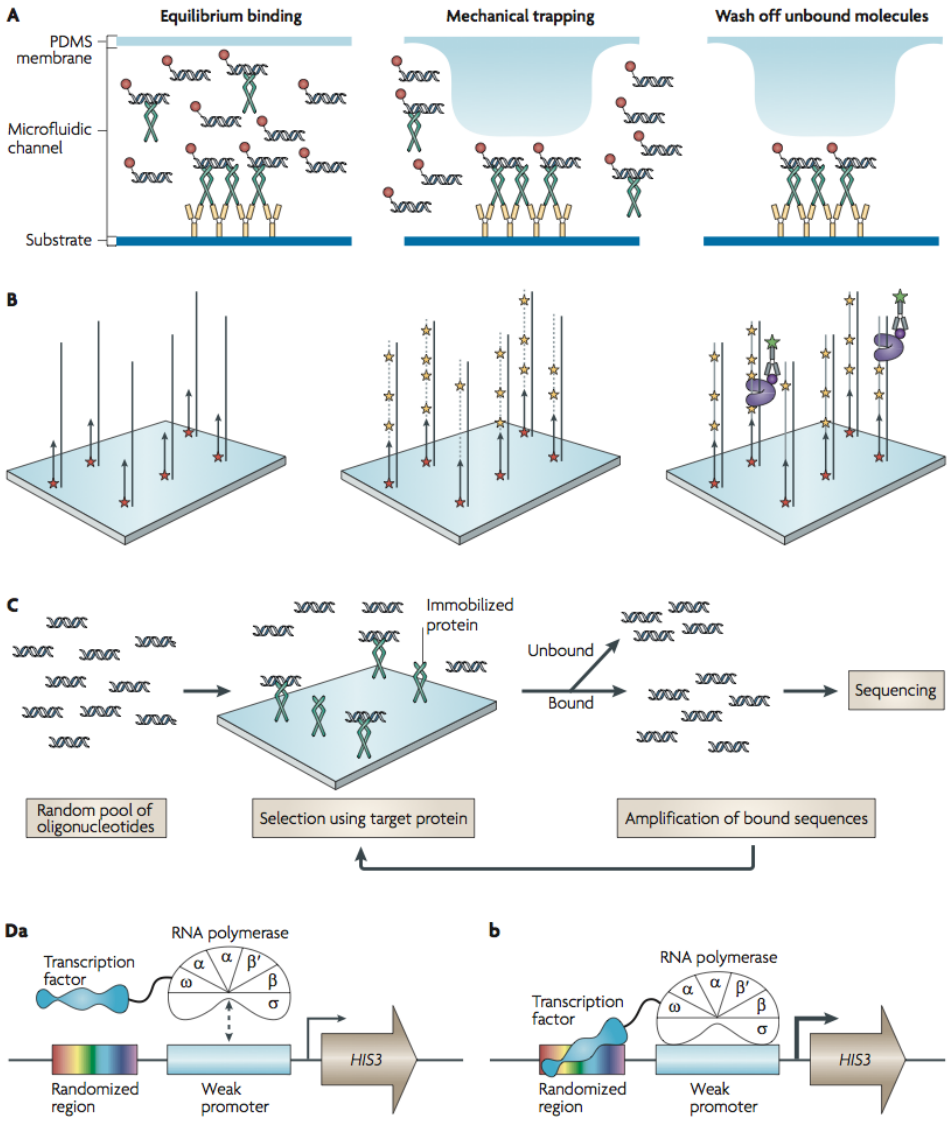


Figure 1.1 Overview of recent *in vitro* high throughput, quantitative methods for determining TF specificity data:

(A) MITOMI (B) PBM (C) High-throughput SELEX (D) B1H method. Taken from (Stormo and Zhao, 2010)

Comparative Genomics Methods

As genome sequences became available for multiple related species, comparative approaches were developed, such as PhyloNet (Wang and Stormo, 2005) that use only genomic sequence data to determine a library of conserved motifs. The comparative genomics approach has been fairly successful, but there are limitations. The greatest draw back is that without further experiments, it is not yet possible to determine which TFs are paired with each motif. Also, it is only possible to study TFs that are held in common between a group of species. Because gene regulation is thought to play a major role in evolution, some of the most interesting motifs may not be possible to find with this method. Having a better recognition models for the largest TF families may help bridge the gap and link motifs discovered with purely computational programs such as PhyloNet with putative TFs.

Overview

This chapter addresses the question raised early on by Seeman about whether a recognition code exists and if so how complex it is (Seeman et al., 1976). While ZF proteins have been studied the most, and there are reasonably good predictive models available (Benos et al., 2001; Benos et al., 2002b, c), there is still ample room for improvement. New

data from high throughput technologies such as PBM and B1H allow us to explore the question again. In chapter 5, new models with substantially better accuracy are described. For HD proteins, not nearly as much effort has been spent trying to find suitable recognition models. In fact a recent paper (Alleyne et al., 2009) proposes that one cannot not construct recognition models that perform much better than a nearest neighbor approach, which is not very informative about mechanism and is not very useful for HD design. We re-examine the question, again relying on newer and larger datasets. We first find that the motifs can be aligned, which the previous paper suggested was not feasible. We find quite good predictive models, significantly better than nearest neighbor approach.

Chapter 2 discusses a large scale characterization of all HD proteins in the Drosophila genome using B1H. This work was expanded with the aim to determine the specificity of all Drosophila TFs. In chapter 3, a database to house this data and an analysis pipeline are introduced. High throughput Illumina sequencing was coupled with a new variant of B1H called constrained variation B1H and a new biophysical based model is introduced to analyze this data in chapter 4. In chapter 5, new ZF and HD models are introduced. They are shown to perform better than the nearest neighbor method and other existing models. Finally, chapter 6 summarizes our findings and discusses future directions.

References

Alleyne, T.M., Pena-Castillo, L., Badis, G., Talukder, S., Berger, M.F., Gehrke, A.R., Philippakis, A.A., Bulyk, M.L., Morris, Q.D., and Hughes, T.R. (2009). Predicting the binding preference of transcription factors to individual DNA k-mers. Bioinformatics *25*, 1012-1018.

Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., and Teichmann, S.A. (2004). Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol *14*, 283-291.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X.*, et al.* (2009). Diversity and complexity in DNA recognition by transcription factors. Science *324*, 1720-1723.

Benos, P.V., Bulyk, M.L., and Stormo, G.D. (2002a). Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res *30*, 4442-4451.

Benos, P.V., Lapedes, A.S., Fields, D.S., and Stormo, G.D. (2001). SAMIE: statistical algorithm for modeling interaction energies. Pac Symp Biocomput, 115-126.

Benos, P.V., Lapedes, A.S., and Stormo, G.D. (2002b). Is there a code for protein-DNA recognition? Probab(ilistical)ly. Bioessays *24*, 466-475.

Benos, P.V., Lapedes, A.S., and Stormo, G.D. (2002c). Probabilistic code for DNA recognition by proteins of the EGR family. J Mol Biol *323*, 701-727.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol *24*, 1429-1435.

Blackwell, T.K., and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science *250*, 1104-1110.

Boyadjiev, S.A., and Jabs, E.W. (2000). Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. Clin Genet *57*, 253-266.

Bulyk, M.L., Huang, X., Choo, Y., and Church, G.M. (2001). Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc Natl Acad Sci U S A *98*, 7158-7163.

Bussemaker, H.J., Foat, B.C., and Ward, L.D. (2007). Predictive modeling of genome-wide mRNA expression: from modules to molecules. Annual review of biophysics and biomolecular structure *36*, 329-347.

Ekker, S.C., Jackson, D.G., von Kessler, D.P., Sun, B.I., Young, K.E., and Beachy, P.A. (1994). The degree of variation in DNA sequence recognition among four Drosophila homeotic proteins. EMBO J *13*, 3551-3560.

Fordyce, P.M., Gerber, D., Tran, D., Zheng, J., Li, H., DeRisi, J.L., and Quake, S.R. (2010). De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nat Biotechnol *28*, 970-975.

Fried, M., and Crothers, D.M. (1981). Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. Nucleic acids research *9*, 6505-6525.

Furney, S.J., Higgins, D.G., Ouzounis, C.A., and Lopez-Bigas, N. (2006). Structural and functional properties of genes involved in human cancer. BMC Genomics *7*, 3.

Galas, D.J., and Schmitz, A. (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic acids research *5*, 3157-3170.

Garner, M.M., and Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic acids research *9*, 3047-3060.

Garvie, C.W., and Wolberger, C. (2001). Recognition of specific DNA sequences. Molecular cell *8*, 937-946.

Hauschild, K.E., Stover, J.S., Boger, D.L., and Ansari, A.Z. (2009). CSI-FID: high throughput label-free detection of DNA binding molecules. Bioorg Med Chem Lett *19*, 3779-3782.

Itzkovitz, S., Tlusty, T., and Alon, U. (2006). Coding limits on the number of transcription factors. BMC Genomics *7*, 239.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol *3*, 318-356.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Liu, J., and Stormo, G.D. (2008). Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. Bioinformatics.

Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. (2000). An overview of the structures of protein-DNA complexes. Genome Biol *1*, REVIEWS001.

Luscombe, N.M., and Thornton, J.M. (2002). Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol *320*, 991-1009.

Maerkl, S.J., and Quake, S.R. (2007a). A systems approach to measuring the binding energy landscapes of transcription factors. Science *315*, 233-237.

Maerkl, S.J., and Quake, S.R. (2007b). A systems approach to measuring the binding energy landscapes of transcription factors. Science *315*, 233-237.

Mahony, S., Auron, P.E., and Benos, P.V. (2007). DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies. PLoS Comput Biol *3*, e61.

Mandel-Gutfreund, Y., Schueler, O., and Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. J Mol Biol *253*, 370-382.

Matthews, B.W. (1988). Protein-DNA interaction. No code for recognition. Nature *335*, 294-295.

Meng, X., Brodsky, M.H., and Wolfe, S.A. (2005). A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol *23*, 988-994.

Meng, X., and Wolfe, S.A. (2006). Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. Nat Protoc *1*, 30-45.

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A., and Bulyk, M.L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat Genet *36*, 1331-1339.

O'Flanagan, R.A., Paillard, G., Lavery, R., and Sengupta, A.M. (2005). Non-additivity in protein-DNA binding. Bioinformatics *21*, 2254-2263.

Oliphant, A.R., Brandl, C.J., and Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Mol Cell Biol *9*, 2944-2949.

Pabo, C.O., and Sauer, R.T. (1984). Protein-DNA recognition. Annu Rev Biochem *53*, 293-321.

Puckett, J.W., Muzikar, K.A., Tietjen, J., Warren, C.L., Ansari, A.Z., and Dervan, P.B. (2007). Quantitative microarray profiling of DNA-binding molecules. J Am Chem Soc *129*, 12310-12319.

Puppin, C., Fabbro, D., Pellizzari, L., and Damante, G. (2011). Using the recognition code to swap homeodomain target specificity in cell culture. Mol Biol Rep.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E.*, et al.* (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-2309.

Schneider, T.D. (2002). Consensus sequence Zen. Appl Bioinformatics *1*, 111-119.

Seeman, N.C., Rosenberg, J.M., and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci U S A *73*, 804-808.

Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*, 16-23.

Stormo, G.D. (2011). Maximally efficient modeling of DNA sequence motifs at all levels of complexity. Genetics *187*, 1219-1224.

Stormo, G.D., Schneider, T.D., and Gold, L.M. (1982). Characterization of translational initiation sites in E. coli. Nucleic acids research *10*, 2971-2996.

Stormo, G.D., and Zhao, Y. (2007). Putting numbers on the network connections. Bioessays *29*, 717-721.

Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. Nat Rev Genet *11*, 751-760.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science *249*, 505-510.

Tupler, R., Perini, G., and Green, M.R. (2001). Expressing the human genome. Nature *409*, 832-833.

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.

Wang, T., and Stormo, G.D. (2005). Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. Proc Natl Acad Sci U S A *102*, 17400-17405.

Warren, C.L., Kratochvil, N.C., Hauschild, K.E., Foister, S., Brezinski, M.L., Dervan, P.B., Phillips, G.N., Jr., and Ansari, A.Z. (2006). Defining the sequence-recognition profile of DNA-binding molecules. Proc Natl Acad Sci U S A *103*, 867-872.

Wolfe, S.A., Nekludova, L., and Pabo, C.O. (2000). DNA recognition by Cys2His2 zinc finger proteins. Annual review of biophysics and biomolecular structure *29*, 183-212.

Wright, W.E., Binder, M., and Funk, W. (1991). Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. Mol Cell Biol *11*, 4104-4110.

Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring binding energies from selected binding sites. PLoS Comput Biol *5*, e1000590.

Zhao, Y., and Stormo, G.D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat Biotechnol *29*, 480-483.

Chapter 2: Analysis of Homeodomain Specificities Allows the Family-Wide Prediction of Preferred Recognition Sites[1]

[1] This chapter was adapted from: Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell *133*, 1277-1289. I performed all of the computational work. This includes analyzing all of the B1H data sets, clustering the motifs, constructing trees, the adjusted mutual information analysis, and development of the recognition model and motif prediction web site

Abstract

We describe the comprehensive characterization of homeodomain DNA-binding specificities from a metazoan genome. The analysis of all 84 independent homeodomains from *D. melanogaster* reveals the breadth of DNA sequences that can be specified by this recognition motif. The majority of these factors can be organized into 11 different specificity groups, where the preferred recognition sequence between these groups can differ at up to 4 of the 6 core recognition positions. Analysis of the recognition motifs within these groups led to a catalog of common specificity determinants that may cooperate or compete to define the binding site preference. Using these recognition principles, a homeodomain can be reengineered to create factors where its specificity is altered at the majority of recognition positions. This resource also allows prediction of homeodomain specificities from other organisms, which is demonstrated by the prediction and analysis of human homeodomain specificities.

Introduction

In humans, as well as many other metazoans, homeodomains comprise the second largest class of sequence-specific transcription factors (TFs) (Tupler et al., 2001). Homeotic

genes were first identified in *D. melanogaster* because their altered activity resulted in dramatic phenotypes such as the formation of an additional pair of wings (Lewis, 1978). Cloning of these genes led to the landmark observation that they contain a common sequence motif that encodes a DNA-binding domain (Gehring et al., 1994a). Subsequent studies have identified a large number of additional homeodomain proteins in *Drosophila* that regulate diverse developmental processes. A remarkable number of these genes have mammalian homologs with conserved developmental functions and biochemical properties (Banerjee-Basu and Baxevanis, 2001; Mukherjee and Burglin, 2007).

Insights into the mechanisms of sequence-specific DNA binding by homeodomains have been provided by the three-dimensional structures of individual protein-DNA complexes coupled with directed mutagenesis and biochemical analysis (Ades and Sauer, 1995; Gehring et al., 1994b; Wolberger, 1996). The homeodomain consists of approximately 60 amino acids that fold into a stable 3-helix bundle preceded by a flexible N-terminal arm. Interactions with a 5 to 7 base pair DNA binding site are formed by positioning a single "recognition" helix in the major groove and the N-terminal arm in the minor groove (Figure 1A and B). Despite a common DNA-binding architecture, there is significant variation in the sequence composition within the homeodomain family; for example the two superclasses of homeodomains, denoted as typical and atypical (Banerjee-Basu and Baxevanis, 2001;

Mukherjee and Burglin, 2007), share low sequence identity and recognize substantially different DNA sequences, yet their docking with the DNA is nearly identical (Kissinger et al., 1990; Wolberger et al., 1991). This conserved binding geometry allows differences in amino acid sequence and DNA-binding specificity for various homeodomains to be interpreted within a common structural framework. Residues at positions 2, 3 and 5-8 on the N-terminal arm, as well as residues at positions 47, 50, 51, 54 and 55 on the recognition helix, can all contribute to DNA-binding specificity (Ades and Sauer, 1995; Damante et al., 1996; Ekker et al., 1994; Fraenkel et al., 1998; Passner et al., 1999; Piper et al., 1999; Wolberger et al., 1991) (Figure 1B and C).

How specific sequence variations between homeodomains lead to different recognition preferences has been defined in several cases. Seminal experiments demonstrated that Lys50 promotes recognition of TAATCC by the Bicoid class of homeodomains instead of the TAAT(T/G)(A/G) recognized by the Gln50-containing Antp and En classes (Hanes and Brent, 1989; Percival-Smith et al., 1990; Treisman et al., 1989). Beachy and colleagues mapped differences in binding site position 2 specificity for the posterior HOX protein Abd-B (TTATGG) and more anterior HOX family members (TAATGG) to amino acids at positions 3, 6 and 7 in the N-terminal arm (Ekker et al., 1994). Interestingly, substitutions at amino acids that overlap with these positions (6-8) are sufficient to switch the specificity of

an NK-2 type homeodomain (CAAGTG) to the specificity of an Antp-type homeodomain (TAAGTG) at the *neighboring* base, binding site position 1 (Damante et al., 1996). This complexity is not limited to the N-terminal arm, as residues at different amino acid positions, such as 47 and 54, can potentially contact the same base pair (Fraenkel et al., 1998; Gruschus et al., 1997; Wolberger et al., 1991). This diversity in potential recognition contacts has hindered efforts to globally reengineer homeodomain specificity (Mathias et al., 2001). Consequently, a comprehensive description of the determinants of homeodomain DNA-binding specificity remains an important goal.

A complete survey of DNA-binding specificity on a large family of DNA-binding domains has not been previously attempted. We have recently described a bacterial one-hybrid (B1H) system that allows the specificities of a DNA-binding domain to be rapidly characterized with sufficient ease that multiple factors can be assayed in parallel (Meng et al., 2005; Meng and Wolfe, 2006). Using this system, we analyze the DNA-binding specificities for all 84 homeodomains in *D. melanogaster* that are not associated with an additional DNA-binding domain as well as 16 mutant homeodomains with changes in residues that contribute to DNA recognition. Our analysis reveals a diverse array of DNA-binding specificities with a minimum of seventeen unique specificities in *D. melanogaster, of* which the majority of homeodomains can be clustered into 11 specificity groups. Members of a given

specificity group typically share common recognition residues. Combining this data with previous structural and biochemical work on the homeodomain family, we propose and evaluate a detailed set of recognition determinants for homeodomains and use this information to broadly and accurately predict the specificities of homeodomains in the human genome.

Results

Analysis of homeodomains using a modified bacterial one-hybrid (B1H) system

We have modified our B1H system to rapidly characterize the DNA-binding specificity of a homeodomain (Meng et al., 2005; Meng and Wolfe, 2006). Homedomains are expressed as fusions to both the omega subunit of RNA-polymerase (Dove and Hochschild, 1998), which provides better dynamic range than fusions to alpha (data not shown), and to zinc fingers 1 and 2 of the protein Zif268 (Zif12; Figure 1D). Because zinc finger-homeodomain chimeras exhibit increased affinity and specificity (Pomerantz et al., 1995), even homeodomains with relatively low DNA binding activity can be readily characterized, A library with 10 randomized base pairs adjacent to a Zif12 binding site (ZF10) was used to

isolate recognition sequences that are complementary to the homeodomain in this selection system (Figure 1D and Supplementary Figure 1).

This system was used to determine DNA-binding specificities for all 84 of the homeodomains in the *D. melanogaster* genome that are not associated with an auxiliary DNA-binding domain (Supplementary Figure 2 and Supplementary Table 1). These homeodomains cluster into previously described families (Banerjee-Basu and Baxevanis, 2001; Mukherjee and Burglin, 2007) based on their amino acid similarity (Supplementary Figure 3), where approximately 85% of these homeodomains are in the "typical" superclass. Present in the collection of Drosophila homeodomains are diverse sets of amino acids at DNA-recognition positions, which suggests that a range of DNA-binding specificities is possible (Figure 1C). One notable exception is Asn at position 51 of the recognition helix, which is present in all but one of these homeodomains.

Comparisons to earlier studies confirm that the motifs obtained by the B1H method accurately reflect the DNA-binding specificities of homeodomains. For example, all of our specificities for the homeotic (HOX) gene family share a common consensus – T(A/T)AT(T/G)(A/G) (Supplementary Figure 4), consistent with previous studies (Pearson et al., 2005). Furthermore, subtle differences in the specificity of Ubx, Dfd and Abd-B that

were previously observed in biochemical assays (Ekker et al., 1994; Ekker et al., 1992) are also present in our data, such as the preference of Abd-B for Thy over Ade at binding site position 2. Thus, even subtle differences in homeodomain specificity can be captured by the B1H analysis. The accuracy of our B1H-generated data was further validated by competition gel mobility shift assays performed for 9 factors that display different specificities (Supplementary Figure 5).

Global alignment and clustering of homeodomain binding sites

Remarkable diversity exists in the B1H-determined DNA-binding specificities for the entire set of homeodomains (Supplementary Figure 2). The conservation of Asn51, which specifies Ade at binding site position 3 (Fraenkel et al., 1998; Wolberger et al., 1991), in combination with our ability to infer the orientation of each homeodomain on its binding site (Supplementary Figure 6 and Supplementary Table 2) provides a basis for aligning all of these recognition sequences. Using this master alignment (Supplementary Table 3), hierarchical clustering of the *D. melanogaster* homeodomains was performed based on the similarity of their DNA-binding specificities (Figure 2A). The majority of these factors can be organized into eleven different specificity groups and the average specificity of these groups was determined for the purposes of comparison (Figure 2). In this analysis, we used

only the core 6 base pair element recognized by these factors. Consistent with the idea that many homeodomain proteins prefer similar TAAT-related motifs, slightly more than half (43) of the homeodomains fall into the Antp or En specificity groups. There are also a number of specificity groups, such as the Abd-B and NK-1 group, which differ in sequence preference from the Antp or En groups at only one or two positions. However, other groups, such as the TGIF-Exd group, differ at four positions relative to the Antp or En groups. Outside of these specificity groupings are six factors that exhibit unique specificities. The observed diversity of specificities reveals the adaptability of the homeodomain architecture for the recognition of a variety of DNA sequences.

Clustering the *D. melanogaster* homeodomains by specificity has revealed that homeodomains that share strong amino acid sequence similarity are not always found in the same specificity group (Figure 2C). In 10 examples, two factors share strong sequence similarity, but fall into different specificity groups. In eight of these comparisons, this difference can be explained by the presence of a different residue at one or more of the key DNA-recognition positions (5, 47, 50, 51, 54 and 55, see below). Pairs of factors with high overall sequence similarity, but different specificities, may represent recently diverged gene duplications where one factor has acquired new target genes.

Distinguishing features of homeodomain specificity groups

The contribution of specific residues toward binding site preference for one or more group members has been demonstrated in previous studies. Below, we use correlations between the average group recognition motifs and the amino acid distributions at key DNA recognition positions (Figure 2B) to systematically describe the characteristics of each group that lead to differences in binding specificity.

Typical superclass

Antp and En groups: The largest groups of homeodomains provide a reference point to describe how differences in amino acid sequence correlate with DNA-binding specificity. The Antp and En groups share similar recognition motifs and amino acid distributions at the key recognition positions. However, at binding site position 5, the En group prefers Thy, whereas the Antp group tolerates either Gua or Thy. There is a corresponding difference at amino acid position 54: Ala for the En group and Met for the Antp group. In the Antp-DNA structure, the side chain of Met54 is neighboring this base pair (Fraenkel and Pabo, 1998).

Bcd group: Typical homeodomains utilize Lys50 to specify Cyt at binding site positions 5 and 6 through the interaction of Lys50 with the complementary Gua at these positions (Tucker-Kellogg et al., 1997).

NK-1, Bar and Ladybird groups: Many of these homeodomains are members of the NK or DL homeodomain classes (Banerjee-Basu and Baxevanis, 2001) and generally have Thr at position 47 or 54. Compared to the Antp and En groups, the homeodomains with Thr47 have reduced specificity at binding site positions 4 and/or 5 (Supplementary Figure 7).

NK-2 group: The members of this group prefer Gua at position 4, due to an interaction between Tyr54 and the complementary Cyt (Gruschus et al., 1997). Their specificities vary at binding site position 1, which correlates with differences at residues 6 and 7 of the N-terminal arm (Damante et al., 1996) (Supplementary Figure 8).

Abd-B group: These factors prefer Thy over Ade at position 2. In Abd-B, this preference has been mapped to amino acid positions 3, 6 and 7 of the N-terminal arm (Ekker et al., 1994); however, the variability within the N-terminal arm precludes a simple correlation of binding preference and amino acid sequence.

Atypical homeodomains

The atypical groups generally prefer Gua at binding site position 2, and Cyt and Ade at positions 4 and 5 (Figures 2B and 3A). In CG11617, the Iroquois group and the TGIF group, the preference for Cyt and Ade at positions 4 and 5 correlates with the presence of Arg54, consistent with the structure of MATa2 (Wolberger et al., 1991) (Figure 3B). The single exception to this correlation, Onecut, contains a unique residue (Met50), which may contribute to its distinct binding preference. Likewise, with the exception of the Iroquois group, homeodomains that contain Arg55 prefer Gua at position 2, consistent with the Exd and Pbx structures (Passner et al., 1999; Piper et al., 1999).

TGIF-Exd group: Our data are consistent with previously described specificities for individual members of the TGIF - Exd group (TGA(C/t)A) (Bertolino et al., 1995; Chang et al., 1996).

Six group: All members of this group (So, Six4 and Optix) display a specificity that overlaps with the recognition motif TGATAC and share identical residues at the key DNA-recognition positions (47, 50, 51, 54 and 55). Our data are consistent with a known So motif ((T/C)GATAC) (Hazbun et al., 1997). A discrepancy between our data and a motif (TAAT) reported for an Optix homolog, Six3 (Zhu et al., 2002), is investigated in the analysis of human homeodomains described below.

Iroquois group: Our monomeric motif (ACA) reflects part of the palindromic, homodimer binding site (ACANNTGT) for a full-length Mirr protein (Bilioni et al., 2005). Homeodomains in this group have weak preferences at binding site positions 1 and 2, despite containing notable specificity determinants (Arg5 and Arg55). One striking feature of the Iroquois group is Ala at position 8 (Supplementary Figure 3). In other homeodomains, a large hydrophobic residue at this position binds in a cleft formed by the homeodomain helices and appears to position the N-terminal arm over the 5' end of the binding site (Figure 4). To examine the effect of residue 8 on Iroquois specificity, an Ala8Phe mutation was introduced into Caup (Figure 4D). This mutation restores, albeit incompletely, the anticipated specificity at positions 1 and 2. The incomplete transformation suggests that additional determinants also contribute to specificity at the 5' end of the binding site (Supplementary Figure 9).

Our assessment of the typical and atypical superclasses suggests two overlapping, but distinct sets of protein-DNA interactions (Figure 2B and 3B). Both classes generally share Arg5 and Asn51, which typically specify Thy and Ade at binding site positions 1 and 3, as well as common set of phosphate contacting residues (Supplementary Figure 3), which should result in a similar docking arrangement of all of these homeodomains with the DNA.

Thus, specificity differences between these homeodomains primarily arise from distinct combinations of residues that directly interact with DNA or that influence these contact residues, rather than changes in the overall conformation of the homeodomain-DNA complex.

Common specificity determinants for homeodomain proteins

Computational and qualitative approaches were used to decipher how variations in homeodomain amino acid sequences across all specificity groups lead to differences in the preferred bases at each binding site position. Mutual information (MI) analysis was used to identify potential specificity determinants by evaluating homeodomain residues that co-vary with changes in binding site preferences (Gutell et al., 1992; Mahony et al., 2007). A simple MI analysis identified some expected correlations at the protein/DNA interface (Supplementary Table 4), but was complicated by the limited variability at some individual positions (Supplementary Figure 10A). To compensate for differences in variability, the MI matrix was transformed into a joint rank product matrix (Supplementary Figure 10B). This plot identifies many known homeodomain-DNA interactions; for example, strong MI is observed between recognition helix positions 50 and 54 and binding site positions 6 and 4, respectively. However, a strong correlation between residue 47 and binding site position 2 is

likely due to evolutionary linkage; the residue present at position 47 correlates to the superclass of the homeodomain (atypical or typical) and each superclass typically prefers different bases at this position. Although evolutionary history complicates MI analysis, novel positions are identified that may be new hallmarks for predicting binding specificity.

To identify which amino acids lead to different binding site preferences, we examined the correlations between amino acid sequence and recognition preference in the context of homeodomain structures and existing or new mutagenesis experiments. The keystone for this analysis is recognition of Ade at position 3 by Asn51. Inferences about specificity determinants may not be valid in the absence of this interaction. Below, residues that most frequently contribute to specificity are summarized for each position in the binding site (Figure 5) and a more detailed analysis is available in the supplementary discussion.

Binding Site (BS) Position 1: 89% of the aligned recognition sequences have Thy at this position. Consistent with this preference, the majority of homeodomains (94%) have Arg5 in the N-terminal arm, which specifies Thy (Ades and Sauer, 1995).

BS Position 2: Preferences for Ade, Gua or Thy are observed among the different homeodomains. 83% of the aligned recognition sequences have Ade at this position. Most

typical homeodomains contain Arg2 or Arg3, which help specify Ade (Ades and Sauer, 1995; Hovde et al., 2001). Most atypical homeodomains contain Arg55, which can specify Gua.

BS Position 3: Asn51 specifies Ade at this position.

BS Position 4: Any base can be specified at this position. Thy is the most common base (80%) and is strongly correlated with the presence of Ile or Val at position 47.

BS Position 5: Preferences for Ade, Thy and Cyt are observed among different homeodomains. For many specificity groups, correlations exist between combinations of residues at positions 47, 50 and 54 and certain base preferences.

BS Position 6: Preferences for Ade, Gua and Cyt are observed among the different homeodomains. Like binding site position 5, residues at positions 47, 50 and 54 appear to be the primary determinants of specificity.

These results imply that there is rarely a simple one-to-one correlation between a specific residue and the preferred base at a binding site position. This complexity precludes the construction of a basic "recognition code" that defines specificity based on a subset of

residues at key recognition positions; however, this analysis reveals some general principles regarding how certain combinations of residues influence specificity. Multiple homeodomain positions can contact a single base pair (e.g. residues 47 and 54 at base position 4 and residues 3 and 55 at base position 2), and when more than one determinant is present for a single base pair, these residues can be in competition (see next section). In addition, other residues can indirectly contribute to specificity by influencing the conformation of potential contact residues. For example Ala8 affects specificity in the N-terminal arm (Figure 4). Similarly, Lys50 displays distinct base preferences in the Bcd and Six groups, likely due to different neighboring residues at positions 47 and 54. These examples support the general conclusion that the contribution of individual specificity determinants to DNA recognition is modulated by additional residues at the protein-DNA interface.

Bcd uses competing contact residues

We have used Bcd to explore the role of competition in determining specificity, as it contains Ile47 and Arg54, which can specify Thy and Cyt, respectively, at binding site position 4. At this position, Bcd displays a strong preference for Thy, a weak preference for Gua and no evidence of tolerance for Cyt (Figure 6A and Supplementary Figure 11). The weak preference for Gua at position 4 has been previously demonstrated (Dave et al., 2000),

and is likely due to Lys50, as this residue can interact simultaneously with the carbonyls of the base at position 4 on the primary strand and position 5 on the complementary strand in the context of the consensus binding site, TAATCC (Tucker-Kellogg et al., 1997).

The absence of Cyt in the recognition motif at position 4 suggests that Ile47 or Lys50 may prevent Arg54 from contributing to the base preference. When Ile47 is mutated to Asn, a residue commonly found in atypical homeodomains that contain Arg54, a slight tolerance for Cyt is observed, indicating the influence of Arg54 (Figure 6A). When Lys50 is mutated to Ala, a complete shift to an En-like specificity (TAATTA) is observed. In the double mutant Ile47Asn and Lys50Ala, a preference for Cyt at position 4 - the base specified by Arg54 in most atypical homeodomains - is revealed. Thus, three different potential specificities are embedded within Bcd. Lys50 and Arg54 are less influential, likely because they are more flexible and are able to make other favorable interactions: Lys50 with bases at positions 5 and 6, and Arg54 with the phosphodiester backbone.

Engineering the DNA-binding specificity of En

We used our catalog of specificity determinants to shift the specificity of En from a typical homeodomain (TAATTA) to a TGIF-type atypical homeodomain (TGACA). En

and TGIF differ in binding site preference at four out of six positions (Figure 6B) and share only 28% amino acid sequence identity overall. While homeodomain specificities have been previously altered at one or two binding site positions, attempts to produce more dramatic changes have failed (Mathias et al., 2001).

Two partial conversions were performed in parallel to assess the flexibility of the En-scaffold for each end of the binding site (Figure 6B): two mutations (R3K and K55R) were sufficient to alter specificity at position 2 (T**G**ATTA) and two other mutations (I47N and A54R) altered specificity at positions 4-6 (TAA**CA**). The combination of both pairs of mutations (R3K, I47N, A54R and K55R) resulted in the desired 5' specificity, but an intermediate 3' specificity (TGA(T/C)(T/A)(G/A); Figure 6B), which suggests additional competing specificity determinants. Gln50, although passive in the I47N, A54R mutant, might influence specificity in the quadruple mutant context. Indeed, addition of the Q50A mutation creates an almost complete conversion to the desired TGACA specificity, as demonstrated by motif clustering analysis (Supplementary Figure 12). The intermediate and final transformations of binding specificity demonstrate that En is a robust scaffold for engineering novel DNA-binding specificities (Supplementary Figure 13). In addition, these results highlight how the impact of an individual specificity determinant (i.e. Gln50) can be influenced by its context at the homeodomain-DNA interface.

Predicting the specificity of the human homeodomains

We used our analysis of *Drosophila* homeodomain specificities to predict the specificity of most human homeodomain proteins. Pairs of homeodomains with the highest overall sequence similarity can have different specificities, likely due to differences at their key recognition positions (Figure 2C). Therefore, three criteria were employed in making predictions for the independent human homeodomains: 1) the presence of Asn51, 2) the overall sequence similarity of each human homeodomain to each fly homeodomain, and 3) the number of identical residues at five recognition positions (5, 47, 50, 54 and 55). The recognition motifs for 153 of 193 human homeodomains (79%) were constructed from the selected binding sites of up to three fly factors that share the highest overall sequence homology and the most similar recognition residues (Supplementary Figure 14). A cross-validation test with the fly homeodomain set was used to assess the accuracy of these predictions (Supplementary Table 5). The human predictions were binned into four confidence levels based on the cross-validation analysis (Supplementary Table 6) from highest (1) to lowest (4). 113 (74%) of the predictions fall in the top two confidence levels. These predictions were confirmed for six human homeodomains (BarHL1, Nkx3-2, PitX2, Six3, TGIF2, Vsx1) by determining their specificities using the B1H system (Figure 7). The

determined and predicted specificities are very similar (all p-values < $2x10^{-6}$), indicating that this approach should be applicable to homeodomains from a broad range of species. This conclusion is supported by an independent comparison with the specificities for non-fly homeodomains in TRANSFAC (Matys et al., 2003) with our predicted specificities for these factors (Supplementary Table 7). Predictions of homeodomain specificities from other species can be made through our web-page where a user enters the homeodomain amino acid sequence and a recognition motif is generated if homeodomains are present in our dataset that meet the user-defined criteria (Supplementary Figure 15). Our specificity predictions for the human homeodomain set, their corresponding PWMs, and the interactive prediction tool are available at http://ural.wustl.edu/flyhd.

Discussion

A major limitation for understanding transcriptional regulation in animal cells is the paucity of defined specificities for the majority of encoded transcription factors. The B1H system offers many potential advantages for the analysis of transcription factor specificity. First, selected binding sites are assayed for the ability to activate a biological response in the context of competition from a pool of potential sites in the *E. coli* genome. More importantly, the ability to determine the orientation of the homeodomain on each selected

binding site allows even partially symmetric sites to be properly aligned when constructing recognition motifs (Supplementary Figure 6). Correct alignment of selected sites is not only important for ranking predicted recognition sequences in genomic DNA sequences, but it is also required to understand the structural basis for variations in DNA binding specificity.

This study provides a complete analysis of homeodomain specificities in a metazoan and it dramatically increases the number of characterized homeodomains in this organism, as only 18 of 84 had any binding site information in the FlyREG database (Bergman et al., 2005). We find that the homeodomain family displays an extensive range of specificities in which a wide variety of bases can be preferred at most positions within the core 6 bp binding site. Overall, the majority of homeodomains (93%) in our dataset can be clustered into 11 different specificity groups with an additional 6 homeodomains that display unique specificities. This clustering strategy allowed us to describe how common variations in residues at a given position in the homeodomain contribute to differences in specificity. However, even within these groups there are homeodomains that display differences in binding site preference. For example, members of the NK-2 group differ in their base preference at the 5'-most position and Exd specificity clearly differs from other members of the TGIF group (Supplementary Figure 8, Figure 3A). In addition, differences outside the core 6 base pair binding site motifs lead to further diversity among homeodomain

specificities (Supplementary Figure 2). Thus, the 17 specificities described by the 11 groups and 6 unique homeodomains represent the minimum number of different specificities recognized by Drosophila homeodomains.

Our analysis demonstrates that the overall sequence similarity between two homeodomains is a useful, but sometimes misleading indicator of the degree of similarity in their DNA-binding specificities. Once factors are clustered into specificity groups, it is possible to compare binding specificity with their degree of sequence homology (Figure 2C). As expected, a substantial correlation between sequence similarity and preferred recognition motif is observed. However, we find multiple examples where pairs of closely related homeodomains cluster into different specificity groups. In both naturally-occurring and engineered homeodomains, single amino acid changes at putative DNA recognition positions are sufficient to alter specificity. These observations illustrate the importance of defining the amino acid positions that contribute to variations in binding site specificity in order to make accurate specificity predictions.

In addition to providing a better understanding of DNA-recognition for this family, this dataset provides a resource for the prediction and interpretation of homeodomain binding sites in regulatory targets within the *D. melanogaster* genome. The specificity of individual

homeodomains has proven instrumental in the identification of functional regulatory sites utilized by these factors *in vivo* (a subset of examples in *D. melanogaster* are listed in Supplementary Table 8) and in the computational identification of target genes with evolutionarily conserved binding sites (Berman et al., 2004; Kheradpour et al., 2007; Schroeder et al., 2004; Sinha et al., 2003). Comparisons with chromatin immunoprecipitation (ChIP) data confirm that Bicoid monomer binding sites are enriched at sites that are occupied *in vivo* (Li et al., 2008) and that the combination of ChIP data and analysis of conserved transcription factor binding sites generally provides significant improvement in the prediction of functional targets over either method alone (Kheradpour et al., 2007). The complete analysis of *D. melanogaster* specificities also highlights the importance of identifying factors with overlapping specificities, as conserved binding sites may reflect recognition sequences for a number of potential factors.

Homeodomains can bind DNA as monomers, homodimers, heterodimers or higher order complexes; in several examples, the preferred recognition sequence of monomers in these complexes may even be modified (Pearson et al., 2005; Ryoo and Mann, 1999; Wilson and Desplan, 1999). Both structural data and our analysis suggest that a likely site for modified specificities is in the flexible N-terminal arm (Figures 1, 2 and 4). The recently described structures of Scr-Exd heterodimers bound to DNA reveal how complex

formation can alter the interaction of residues within and beyond the N-terminal arm with DNA (Joshi et al., 2007). Thus, while the primary sequence determinants within the N-terminal arm help define sequence preferences, intramolecular (e.g. Ala8 in Caup; Figure 4) or intermolecular (e.g. Scr-Exd) interactions can also influence recognition. It is currently unclear how frequently monomeric specificities are modified by protein-protein interactions, but our systematic characterization of monomeric specificities provides a foundation to explore this question.

The analysis of homeodomain specificities in *D. melanogaster* also provides the basis to predict most homeodomains specificities in other organisms. We predicted the DNA-binding specificities of 79% of the independent homeodomains in the human genome with moderate to high confidence (Supplementary Figure 14). This prediction scheme can be applied to homeodomains from any species, providing a resource to help identify binding sites in cis-regulatory regions. In the future, incorporation of a probabilistic recognition code to approximate the specificities of factors that do not have good homologs in our database should allow more comprehensive specificity predictions based on homeodomain amino acid sequence (Benos et al., 2002; Liu and Stormo, 2005).

Continued analysis of homeodomain specificity will lead to more detailed understanding of recognition by this family. Our current experiments have led to a catalogue of specificity determinants that can be used to rationally engineer the DNA-binding specificity of homeodomains. The throughput of the B1H system will facilitate the synthesis of a more comprehensive recognition model as more naturally-occurring and mutant homeodomains are characterized. The B1H system can also be used to perform selections on pools of mutagenized homeodomains to assess the range of residues that are compatible with recognition of a given motif. Given the high success rate of the B1H method, a systematic characterization of other classes of DNA-binding domains can be used to produce a complete map of transcription factor specificities in a genome.

Experimental Procedures:

Homeodomain binding site selections: A detailed description of the general B1H selection protocol has been previously described (Meng et al., 2005; Meng and Wolfe, 2006), modifications to this procedure and a detailed description of the construction of the ZF10 randomized library are presented in the Supplementary Methods. The 84 independent *D. melanogaster* homeodomains were identified as described in Supplementary Methods. The sequences of the homeodomains used in the B1H selection and the raw selected binding sites are found in Supplementary Table 1.

Construction of the master alignment of sites for clustering and MI analysis:

The master alignment contains 1860 binding sites for 83 of the 84 Drosophila homeodomain proteins as well as Oct1 (Lag1 was excluded because it lacks Asn51). These alignments were constructed from overrepresented motifs identified for each factor using CONSENSUS (Hertz and Stormo, 1999). Details on the alignment construction, motif clustering and MI analysis can be found in the Supplementary Methods. All Sequence logos (Schneider and Stephens, 1990) for these factors were generated using WebLogo (Crooks et al., 2004). Because the number of selected binding sites that comprise a particular logo is modest (22 on average), the significance of bases that are absent or occur infrequently in a motif cannot be fully assessed.

Specificity Predictions for the human homeodomain set

193 homeodomains containing proteins were annotated in the SMART human genome database and 175 of these were independent homeodomains containing Asn51. To predict the DNA-binding specificity of this set we used the DNA-binding specificity of up to 3 of the fly homeodomains with the highest BLOSUM45 similarity scores (calculated from a sequence-to-profile multiple sequence alignment (Edgar, 2004) between the query sequence and the 84 fly homeodomain profiles) provided that: 1) they contained Asn51; 2) they contained identical residues at the other 5 key recognition positions (5, 47, 50, 54 and 55); and 3) they passed a BLOSUM45 similarity score threshold. The similarity score threshold was set to 200, based on a cross validation analysis of the fly homeodomain set (data not shown). Additionally, once a reference protein passed all of our filters, additional reference proteins were only added to the predictive set if their similarity score was within 40 similarity score units of the most similar reference protein. If no reference homeodomain passed these three criteria, we considered up to 3 homeodomains within the set that contained identical residues at 4 of the 5 key recognition positions, as long as they also passed the similarity score threshold. Specificity predictions comprise all of the selected binding sites for all of the reference homeodomains that passed the filters. In some cases no fly homeodomains met these criteria and consequently no prediction was made.

Cross-validation analysis and comparison of predicted and determined motifs

To assess the accuracy of the specificity predictions we performed a cross-validation analysis where the binding specificity of each fly homeodomain was predicted based on the information of all of the other homeodomain proteins. All TRANSFAC 10.2 datasets associated with proteins classified as homeodomains (TRANSFAC classes C0006, C0027, C0047, C0053) and that contain at least 20 binding sites were extracted from the database (Matys et al., 2006). The 47 groups of binding sites that met these requirements were reanalyzed with CONSENSUS to generate new motifs. 27 of these 47 transcription factors were sufficiently similar to a *D. melanogaster* homeodomain to make a prediction based on our criteria (described in the text). In some cases (8), multiple homeodomains were associated with one dataset in TRANSFAC and vice versa (5). In these cases, we compared the predicted matrix for a factor to each of the CONSENSUS matrices associated with it. We used the Average Log Likelihood Ratio (ALLR) score to determine the best local alignment (Matalign-v2a, Wang, T & Stormo, G. D. *unpublished*) between the predicted and CONSENSUS matrices. Based on these alignments, we assessed the degree of similarity using the ALLR similarity score, the ALLR based distance and the e-value computed by Matalign.

Figure Legends

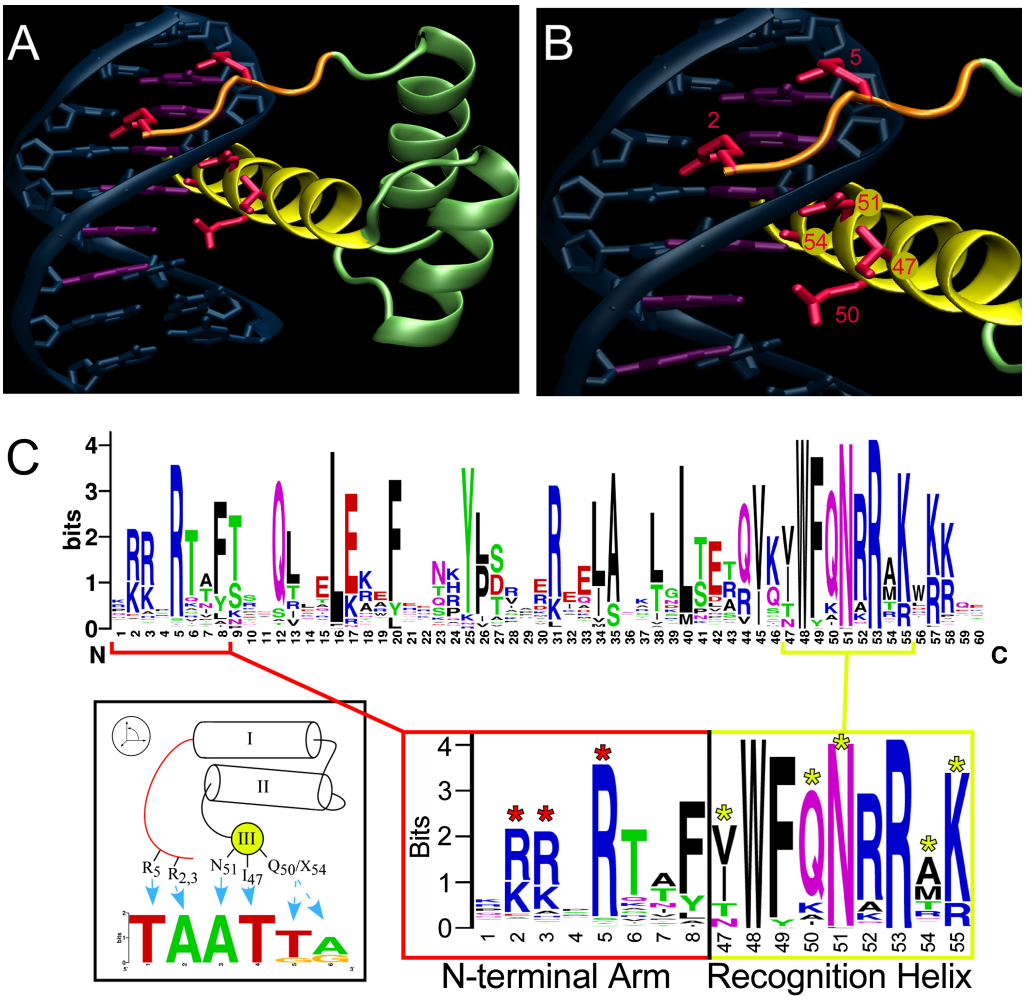

A
B
5
2
51
54
47
50
C
bits
N
C
I
II
III
R5 R2,3 N51 I47 Q50/X54
TAATTA
Bits
N-terminal Arm
Recognition Helix
D
RNA
Polymerase
HD
Zif12
2 1
ω α α β' β σ
?
21 bp
10bp
Randomized
region
Zif12
site
-35
-10
Weak
promoter
HIS3
URA3

Figure 2.1 DNA recognition by the homeodomain family.

A) The structure of Msx-1 bound to DNA is representative of homeodomain-DNA interactions (Hovde et al., 2001). B) Detailed view of the recognition contacts (red), where residues at positions 2 and 5 of the N-terminal arm (orange) interact with bases in the minor groove and residues at positions 47, 50, 51 and 54 of the recognition helix (yellow) are positioned to make contacts in the major groove. C) (Top) Sequence logo representation of the diversity in our set of 84 homeodomains. (Bottom) Windows highlighting the diversity in the DNA-recognition regions - the N-terminal arm (red) and recognition helix (yellow). The key recognition positions are indicated with asterisks. D) Cartoon depicting recruitment of omega-Zif12-HD (homeodomain) fusions to the weak promoter driving the *HIS3* and *URA3* reporters used in the B1H system (Meng et al., 2005; Meng and Wolfe, 2006).

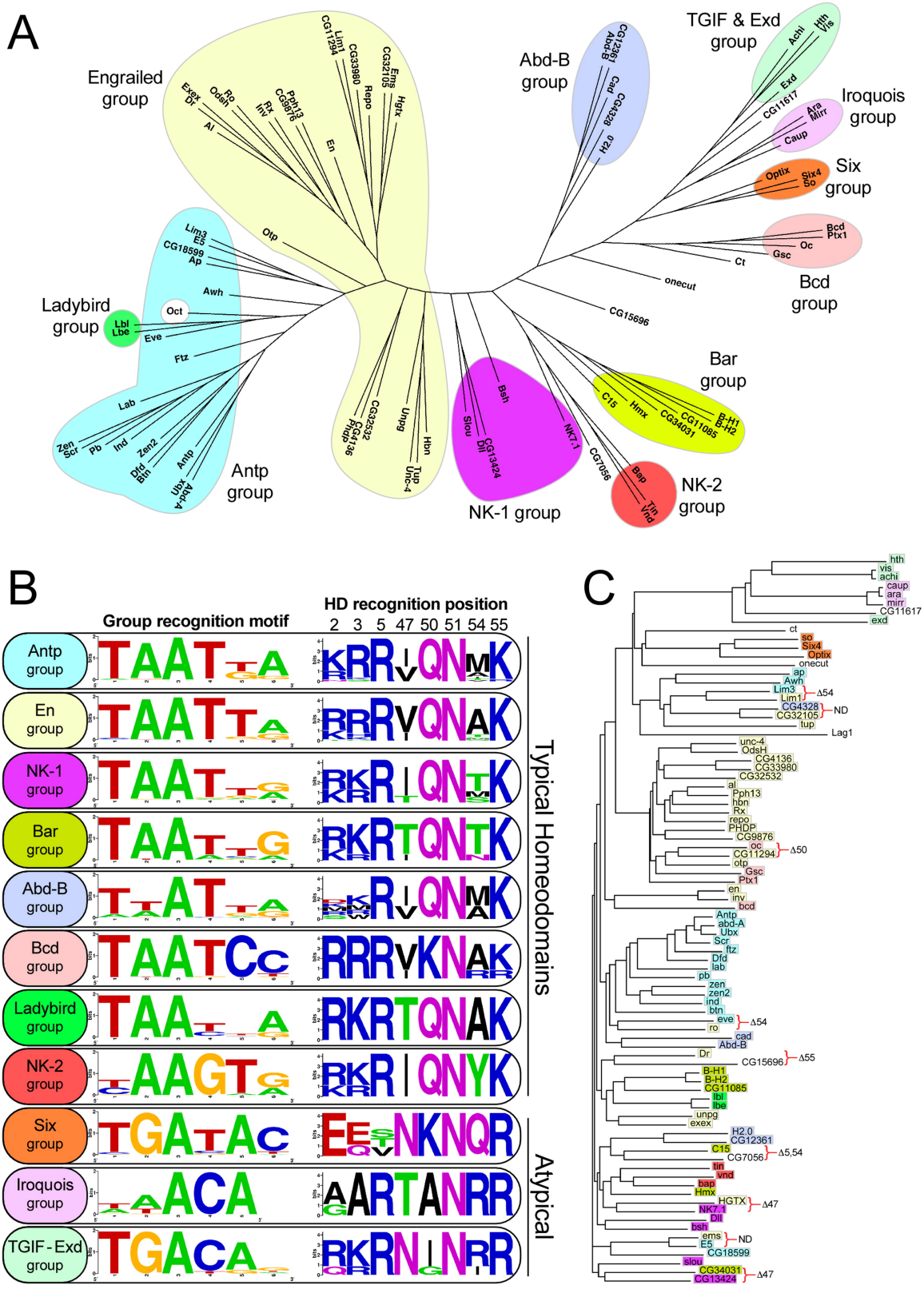

A
Engrailed group
Abd-B group
TGIF & Exd group
Iroquois group
Six group
Bcd group
Ladybird group
Antp group
NK-1 group
Bar group
NK-2 group
onecut
CG15696
Otp
B
Group recognition motif
HD recognition position
2 3 5 47 50 51 54 55
Antp group
En group
NK-1 group
Bar group
Abd-B group
Bcd group
Ladybird group
NK-2 group
Six group
Iroquois group
TGIF - Exd group
Typical Homeodomains
Atypical
C
Δ54
ND
Lag1
Δ50
Δ55
Δ5,54
Δ47
ND
Δ47

Figure 2.2. Clustering of the 84 Drosophila homeodomains.

(A) Clustering based on the similarity between the recognition motifs of these factors, which we have organized into eleven different specificity groups. (B) The typical and atypical homeodomains are distributed into separate groups. The average specificity of each group is indicated under the Group recognition motif, and to the right is the Sequence logo of the key recognition positions. (C) The specificity groups (colored rectangles) are mapped onto the homeodomain amino acid sequence similarity tree. In instances where neighbors have been assigned to different specificity groups (indicated by red brackets) any difference in residue type at a key recognition position (5, 47, 50, 54 or 55) is noted (ND = No difference).

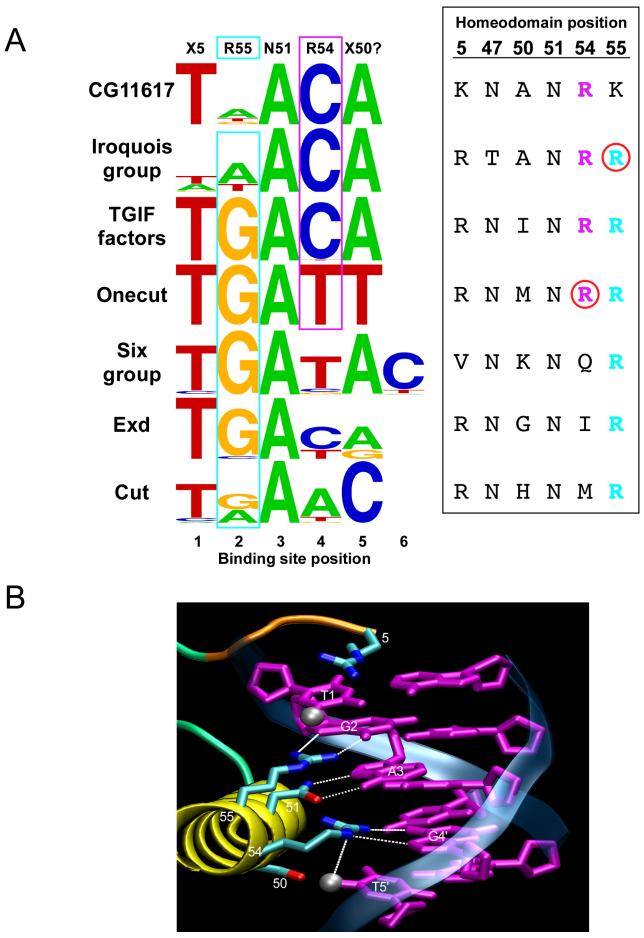


Figure 2.3 Atypical homeodomain specificity and correlations with positions 54 and 55.

A) (Left) Sequence logos for types of atypical homeodomains (either groups or outliers). (Right) The corresponding amino acid sequences at the key DNA contact positions. Arg at position 54 (magenta) correlates with a preference for Cyt at binding site position 4. Arg at position 55 (cyan) correlates with a preference for Gua at binding site position 2. Notable exceptions are indicated by red circles. B) Structural model of DNA recognition for atypical family members constructed from a superposition of the contacts observed in the MATα2-DNA (Wolberger et al., 1991) and Exd-Ubx-DNA structures (Passner et al., 1999). The arginines potentially specify the contacted Gua and the 5' Thy due to the favorable van der Waals interaction (~4 Å) with the T-methyl group (silver sphere).

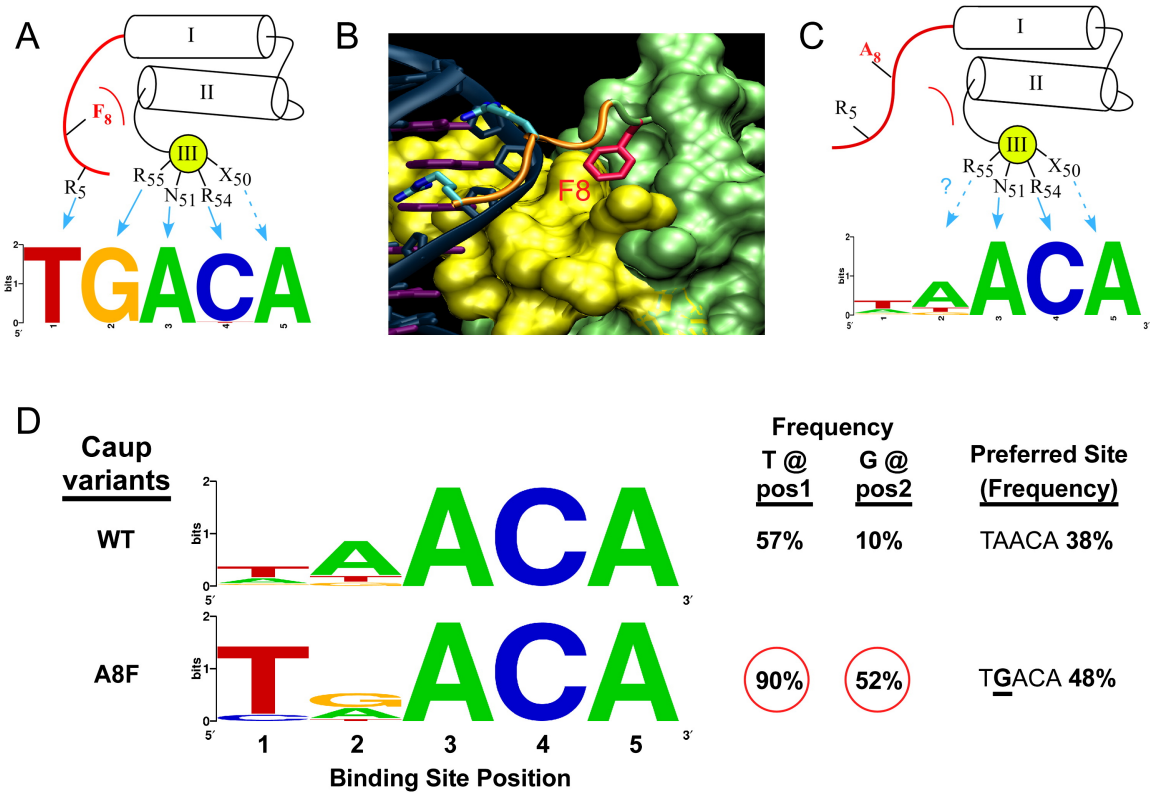


Figure 2.4 The role of position 8 in organizing the N-terminal arm.

A) A large hydrophobic residue at position 8 docks into a pocket formed by the three-helix bundle of the homeodomain fold anchoring the N-terminal arm over the minor groove. B) Surface rendering of the homeodomain (residues 9-60, recognition helix shown in yellow; Msx-1 structure (Hovde et al., 2001)). Phe8 (red) sits in a structural pocket. C) Iroquois family members contain Ala at position 8, allowing the N-terminal arm to sample other conformations that reduce the specificity of the factor. D) Reintroduction of the Phe at position 8 in Caup (A8F) dramatically alters the specificity of the protein at positions 1 and 2 of the binding site.

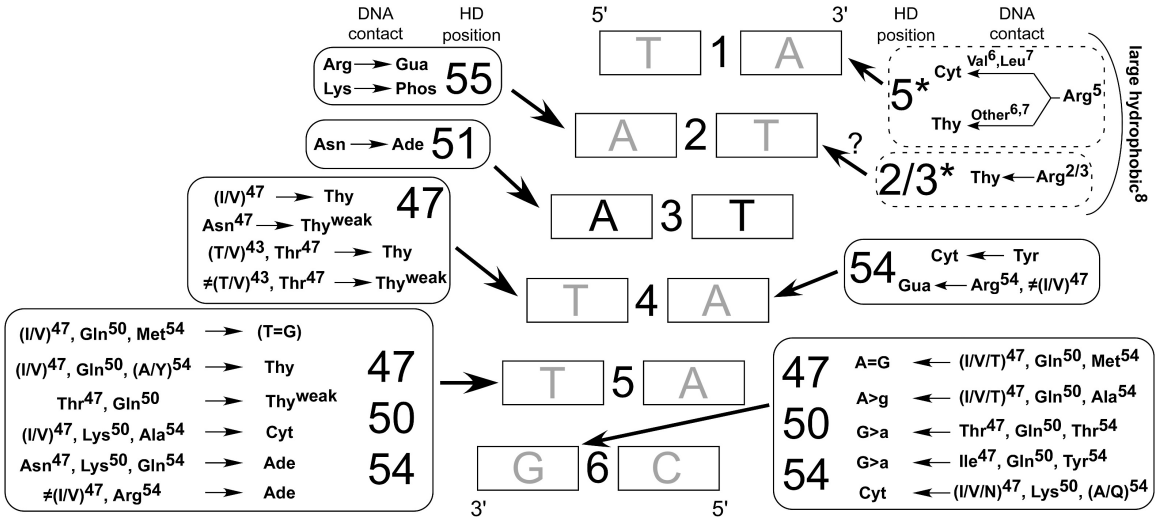

DNA contact
HD position
Arg → Gua
Lys → Phos
55
Asn → Ade
51
(I/V)47 → Thy
Asn47 → Thyweak
(T/V)43, Thr47 → Thy
≠(T/V)43, Thr47 → Thyweak
47
(I/V)47, Gln50, Met54 → (T=G)
(I/V)47, Gln50, (A/Y)54 → Thy
Thr47, Gln50 → Thyweak
(I/V)47, Lys50, Ala54 → Cyt
Asn47, Lys50, Gln54 → Ade
≠(I/V)47, Arg54 → Ade
47
50
54
5'
3'
T 1 A
A 2 T
A 3 T
T 4 A
T 5 A
G 6 C
3'
5'
HD position
DNA contact
5* Cyt Val6,Leu7
Thy Other6,7
Arg5
?
2/3* Thy Arg2/3
large hydrophobic8
54 Cyt Tyr
Gua Arg54, ≠(I/V)47
47 A=G (I/V/T)47, Gln50, Met54
50 A>g (I/V/T)47, Gln50, Ala54
G>a Thr47, Gln50, Thr54
54 G>a Ile47, Gln50, Tyr54
Cyt (I/V/N)47, Lys50, (A/Q)54

Figure 2.5 Catalog of common specificity determinants for Asn51-containing homeodomains. Amino acid positions that are most likely to influence the sequence preference at a particular position are indicated in boxes (solid line – major groove, dotted line – minor groove) surrounding the core 6 bp binding element. An arrow points from the box of potential interactions to the base within each base pair that it describes. For simplicity some interactions, such as Lys50 with binding site positions 5 and 6, are described as influencing specificity on the primary strand of the DNA when in reality direct contacts are made to the complementary strand. DNA recognition by residues in the N-terminal arm is also dependent on the type of residue at position 8 as observed for the Iroquois group.

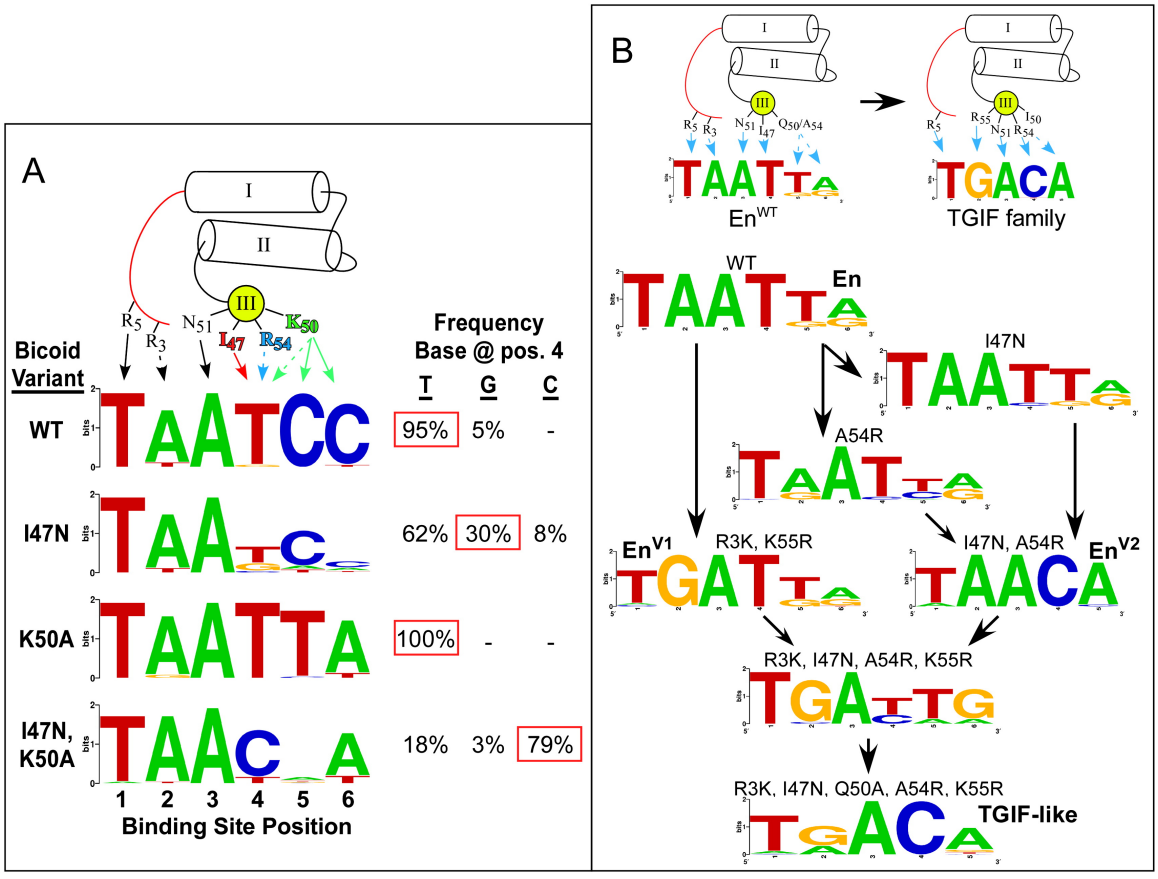


Figure 2.6 Exploring DNA-binding specificity through mutagenesis.

A) Mutational analysis of binding site position 4 in Bcd. Three different mutants (I47N, K50A and I47N with K50A) were characterized to determine the alteration in base preference at this position. The frequency that each base was recovered at position 4 is indicated to the right of the Sequence logo for each factor. B) Conversion of Engrailed (En) into a homeodomain with TGIF-like specificity. (Top) Schematic representation of the critical base contacts responsible for specificity in En and TGIF family members. (Bottom) Flow diagram of the mutations required to complete the specificity conversion. Two intermediate specificity conversions (EnV1 and EnV2) were obtained first, and these mutations were combined along with Q50A to produce TGIF-like specificity.

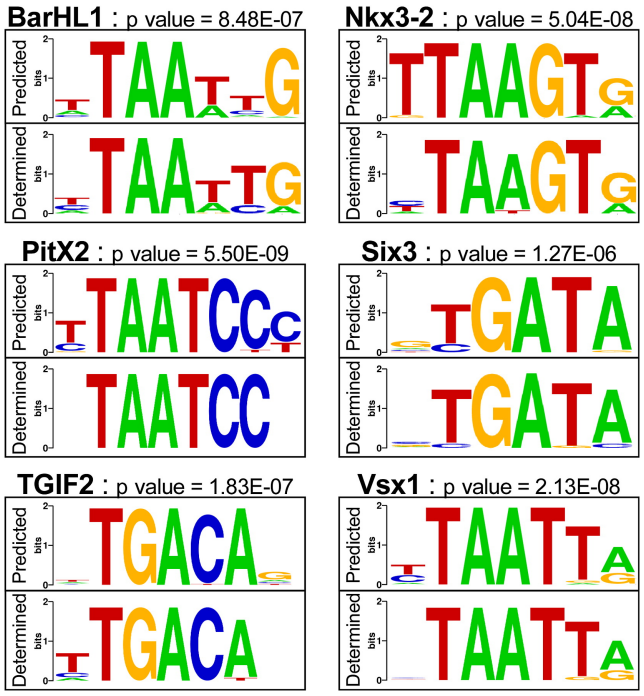


Figure 2.7 Comparison of the predicted and determined recognition motifs for 6 human homeodomains.

The specificities of the human factors were determined using the B1H system. In each case the "Determined" compares favorably with the "Predicted" motif generated using our algorithm. The p-value for each comparison was calculated from the weight matrices for each motif as described in the Methods with additional metrics of these comparisons in Supplementary Table 9. Of particular note, the specificity of Six3 is consistent with other Six family members; it does not specify TAAT as previously described (Zhu et al., 2002).

Supplemental Material

Due to the large quantity of supplemental material, including all of the aligned selected sites, please refer to the supplemental material available on Cell's web site:

http://www.sciencedirect.com/science/article/pii/S009286740800682X

doi:10.1016/j.cell.2008.05.023

Acknowledgments

We would like to thank Xiangdong Meng for his valuable advice and technical support. We would like to thank the Berkeley Drosophila Genome Project (BDGP) for producing the cDNA clones used in this study, the Drosophila Genomics Resource Center (DGRC) for distributing the clones, and Mark Stapleton and Susan Celniker for sharing unpublished results. Some of these ORFs were obtained from clones produced by BDGP under National Institutes of Health grant (HG002673 to S. E. Celniker). We would like to thank Adam Richards for technical support. S.A.W. and M.B.N. were supported by NIH grant 1R21HG003721 from NHGRI. A.W. was supported in part by NIH grant 1R21HG003721 from NHGRI. M.H.B. and A.W. were supported in part by a New Scholar in Aging Award from the Ellison Medical Foundation and American Cancer Society grant RSG-05-026-01-CCG. R.G.C. was supported by training grant T32 GM08802. G.D.S. was supported by NIH grant HG00249 from NHGRI.

References

Ades, S.E., and Sauer, R.T. (1995). Specificity of minor-groove and major-groove interactions in a homeodomain-DNA complex. Biochemistry *34*, 14601-14608.

Banerjee-Basu, S., and Baxevanis, A.D. (2001). Molecular evolution of the homeodomain family of transcription factors. Nucl Acids Res *29*, 3258-3269.

Benos, P.V., Lapedes, A.S., and Stormo, G.D. (2002). Probabilistic code for DNA recognition by proteins of the EGR family. Journal of molecular biology *323*, 701-727.

Bergman, C.M., Carlson, J.W., and Celniker, S.E. (2005). Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics (Oxford, England) *21*, 1747-1749.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., and Celniker, S.E. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. Genome Biol *5*, R61.

Bertolino, E., Reimund, B., Wildt-Perinic, D., and Clerc, R.G. (1995). A novel homeobox protein which recognizes a TGT core and functionally interferes with a retinoid-responsive motif. J Biol Chem *270*, 31178-31188.

Bilioni, A., Craig, G., Hill, C., and McNeill, H. (2005). Iroquois transcription factors recognize a unique motif to mediate transcriptional repression in vivo. Proceedings of the National Academy of Sciences of the United States of America *102*, 14671-14676.

Chang, C.P., Brocchieri, L., Shen, W.F., Largman, C., and Cleary, M.L. (1996). Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. Mol Cell Biol *16*, 1734-1745.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome research *14*, 1188-1190.

Damante, G., Pellizzari, L., Esposito, G., Fogolari, F., Viglino, P., Fabbro, D., Tell, G., Formisano, S., and Di Lauro, R. (1996). A molecular code dictates sequence-specific DNA recognition by homeodomains. Embo J *15*, 4992-5000.

Dave, V., Zhao, C., Yang, F., Tung, C.S., and Ma, J. (2000). Reprogrammable recognition codes in bicoid homeodomain-DNA interaction. Mol Cell Biol *20*, 7673-7684.

Dove, S.L., and Hochschild, A. (1998). Conversion of the omega subunit of *Escherichia coli* RNA polymerase into a transcriptional activator or an activation target. Genes Dev *12*, 745-754.

Edgar, R. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Ekker, S.C., Jackson, D.G., von Kessler, D.P., Sun, B.I., Young, K.E., and Beachy, P.A. (1994). The degree of variation in DNA sequence recognition among four Drosophila homeotic proteins. EMBO J *13*, 3551-3560.

Ekker, S.C., von Kessler, D.P., and Beachy, P.A. (1992). Differential DNA sequence recognition is a determinant of specificity in homeotic gene action. EMBO J *11*, 4059-4072.

Fraenkel, E., and Pabo, C.O. (1998). Comparison of X-ray and NMR structures for the Antennapedia homeodomain-DNA complex. Nat Struct Biol *5*, 692-697.

Fraenkel, E., Rould, M.A., Chambers, K.A., and Pabo, C.O. (1998). Engrailed homeodomain-DNA complex at 2.2 A resolution: a detailed view of the interface and comparison with other engrailed structures. Journal of molecular biology *284*, 351-361.

Gehring, W.J., Affolter, M., and Burglin, T. (1994a). Homeodomain proteins. Annu Rev Biochem *63*, 487-526.

Gehring, W.J., Qian, Y.Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A.F., Resendez-Perez, D., Affolter, M., Otting, G., and Wuthrich, K. (1994b). Homeodomain-DNA recognition. Cell *78*, 211-223.

Gruschus, J.M., Tsao, D.H., Wang, L.H., Nirenberg, M., and Ferretti, J.A. (1997). Interactions of the vnd/NK-2 homeodomain with DNA by nuclear magnetic resonance spectroscopy: basis of binding specificity. Biochemistry *36*, 5372-5380.

Gutell, R.R., Power, A., Hertz, G.Z., Putz, E.J., and Stormo, G.D. (1992). Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucl Acids Res *20*, 5785-5795.

Hanes, S.D., and Brent, R. (1989). DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. Cell *57*, 1275-1283.

Hazbun, T.R., Stahura, F.L., and Mossing, M.C. (1997). Site-specific recognition by an isolated DNA-binding domain of the sine oculis protein. Biochemistry *36*, 3680-3686.

Hertz, G.Z., and Stormo, G.D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics *15*, 563-577.

Hovde, S., Abate-Shen, C., and Geiger, J.H. (2001). Crystal structure of the Msx-1 homeodomain/DNA complex. Biochemistry *40*, 12013-12021.

Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B., and Mann, R.S. (2007). Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell *131*, 530-543.

Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 Drosophila genomes. Genome research *17*, 1919-1931.

Kissinger, C.R., Liu, B., Martin-Blanco, E., Kornberg, T.B., and Pabo, C.O. (1990). Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: A framework for understanding homeodomain-DNA interactions. Cell *63*, 579-590.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565-570.

Li, X.Y., Macarthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C.L.*, et al.* (2008). Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm. PLoS biology *6*, e27.

Liu, J., and Stormo, G.D. (2005). Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions. Nucleic acids research *33*, e141.

Mahony, S., Auron, P.E., and Benos, P.V. (2007). Inferring protein-DNA dependencies using motif alignments and mutual information. Bioinformatics (Oxford, England) *23*, i297-304.

Mathias, J.R., Zhong, H., Jin, Y., and Vershon, A.K. (2001). Altering the DNA-binding specificity of the yeast Matalpha 2 homeodomain protein. J Biol Chem *276*, 32696-32703.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V.*, et al.* (2003). TRANSFAC(R): transcriptional regulation, from patterns to profiles. Nucl Acids Res *31*, 374-378.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K.*, et al.* (2006). TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. Nucl Acids Res *34*, D108-110.

Meng, X., Brodsky, M.H., and Wolfe, S.A. (2005). A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol *23*, 988-994.

Meng, X., and Wolfe, S.A. (2006). Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. Nat protocols *1*, 30-45.

Mukherjee, K., and Burglin, T.R. (2007). Comprehensive Analysis of Animal TALE Homeobox Genes: New Conserved Motifs and Cases of Accelerated Evolution. J Mol Evol *65*, 137-153.

Passner, J.M., Ryoo, H.D., Shen, L., Mann, R.S., and Aggarwal, A.K. (1999). Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. Nature *397*, 714-719.

Pearson, J.C., Lemons, D., and McGinnis, W. (2005). Modulating Hox gene functions during animal body patterning. Nat Rev Genet *6*, 893-904.

Percival-Smith, A., Müller, M., Affolter, M., and Gehring, W.J. (1990). The interaction with DNA of wild-type and mutant fushi tarazu homeodomains. EMBO J *9*, 3967-3974.

Piper, D.E., Batchelor, A.H., Chang, C.P., Cleary, M.L., and Wolberger, C. (1999). Structure of a HoxB1-Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. Cell *96*, 587-597.

Pomerantz, J.L., Sharp, P.A., and Pabo, C.O. (1995). Structure-based design of transcription factors. Science *267*, 93-96.

Ryoo, H.D., and Mann, R.S. (1999). The control of trunk Hox specificity and activity by Extradenticle. Genes Dev *13*, 1704-1716.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucl Acids Res *18*, 6097-6100.

Schroeder, M.D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E.D., and Gaul, U. (2004). Transcriptional control in the segmentation gene network of Drosophila. PLoS biology *2*, E271.

Sinha, S., van Nimwegen, E., and Siggia, E.D. (2003). A probabilistic method to detect regulatory modules. Bioinformatics (Oxford, England) *19 Suppl 1*, i292-301.

Treisman, J., Gönczy, P., Vashishtha, M., Harris, E., and Desplan, C. (1989). A single amino acid can determine the DNA binding specificity of homeodomain proteins. Cell *59*, 553-562.

Tucker-Kellogg, L., Rould, M.A., Chambers, K.A., Ades, S.E., Sauer, R.T., and Pabo, C.O. (1997). Engrailed (Gln50-->Lys) homeodomain-DNA complex at 1.9 A resolution: structural basis for enhanced affinity and altered specificity. Structure *5*, 1047-1054.

Tupler, R., Perini, G., and Green, M.R. (2001). Expressing the human genome. Nature *409*, 832-833.

Wilson, D.S., and Desplan, C. (1999). Structural basis of Hox specificity. Nat Struct Biol *6*, 297-300.

Wolberger, C. (1996). Homeodomain interactions. Curr Opin Struct Biol *6*, 62-68.

Wolberger, C., Vershon, A.K., Liu, B., Johnson, A.D., and Pabo, C.O. (1991). Crystal structure of a MATalpha2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. Cell *67*, 517-536.

Zhu, C.C., Dyer, M.A., Uchikawa, M., Kondoh, H., Lagutin, O.V., and Oliver, G. (2002). Six3-mediated auto repression and eye development requires its interaction with

members of the Groucho-related family of co-repressors. Development (Cambridge, England) *129*, 2835-2849.

Chapter 3: FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system[1]

[1] This chapter was adapted from: 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. I developed the analysis pipeline and methods for analysing all of the B1H data sets. I genenerated all of the logos, alignments and matricies that were subsequently added to the database

Abstract

FlyFactorSurvey (http://pgfe.umassmed.edu/TFDBS/) is a database of DNA binding specificities for *Drosophila* transcription factors (TFs) primarily determined using the bacterial one-hybrid system. The database provides community access to over 400 recognition motifs and position weight matrices (PMWs) for over 200 TFs, including many unpublished motifs. Search tools and flat file downloads are provided to retrieve binding site information (as sequences, matrices and Sequence logos) for individual TFs, groups of TFs or for all TFs with characterized binding specificities. Linked analysis tools allow users to identify motifs within our database that share similarity to a query matrix or to view the distribution of occurrences of an individual motif throughout the *Drosophila* genome. Together, this database and its associated tools provide computational and experimental biologists with resources to predict interactions between *Drosophila* TFs and target *cis*-regulatory sequences.

Introduction

The first critical step in converting genomic sequence into temporally and spatially patterned gene expression is the regulation of transcription. This process is typically controlled by cis-regulatory modules (CRMs), discrete sequences that contain groups of binding sites for sequence specific transcription factors (TFs). Experimental methods such as chromatin immunoprecipitation allow

direct genome-wide analysis of TF binding in a specific cell type and experimental condition (Johnson et al., 2007; Li et al., 2008; Ren et al., 2000; Zeitlinger et al., 2007). Work in *Drosophila* and other organisms has shown that matrix representations of the recognition motif of a TF can be used to computationally map enrichment of its binding sites across the genome (Berman et al., 2004; Schroeder et al., 2004; Sinha et al., 2004). By analyzing homo- and heterotypic clusters of TF binding sites, conservation of these sites across species, and the spatial and temporal expression of TFs and their potential target genes, it is possible to computationally construct transcription regulatory networks (Janssens et al., 2006; Kheradpour et al., 2007; Schroeder et al., 2004; Segal et al., 2008). In the case of the *Drosophila* anterior-posterior patterning network, we have shown that the accuracy of networks predicted based on a nearly complete set of TF binding site motifs can be similar to that obtained using chromatin immunoprecipitation data for these TFs (Kazemian, 2010). In principal, such computational approaches could be applied in any cell type where sufficiently complete gene expression and TF binding specificity data is available. Currently, one major limitation for this type of analysis is the incomplete description of recognition motifs for the majority of sequence-specific TFs.

The first studies of TF DNA binding specificity used biochemical methods such as DNAse I footprinting to identify individual binding sites in known target regulatory sequences. Compilations of these sites (Matys et al., 2006; Portales-Casamar et al., 2010) have provided a rich but crude source of descriptions of binding site preferences. In *Drosophila*, motifs constructed from these compiled sites (Bergman et al., 2005), provided a basis for many early studies of TF-CRM regulatory interactions. Subsequently, a variety of additional methods have been developed to study binding

specificities more systematically, including systematic evolution of ligands by exponential enrichment (SELEX) (Kinzler and Vogelstein, 1990; Tuerk and Gold, 1990), SELEX with deep sequencing (Jolma et al., 2010; Zhao et al., 2009; Zykovich et al., 2009), and protein binding microarrays (PBMs) (Berger and Bulyk, 2009; Jolma et al.; Roulet et al., 2002). As an alternative to these purely *in vitro* methods, we have developed the bacterial one-hybrid system (B1H) that allows TF specificities to be determined without purification and in the context of competition from the bacterial genome (Meng et al., 2005). The relative efficiency of this system has allowed for the systematic characterization of large numbers of TFs in *Drosophila*, including all members of the large family of homeodomain proteins and all of the known core components of the embryonic A-P patterning network (Noyes et al., 2008a; Noyes et al., 2008b). This method allows relatively large libraries of randomized binding sites (~10^8) to be rapidly interrogated for potential recognition sequences, where hundreds to thousands of binding sites can be recovered and characterized using high-throughput SOLEXA sequencing.

Several existing databases house collections of TF DNA binding information. The commercial database Transfac (Matys et al., 2006) and the publically accessible database JASPAR (Portales-Casamar et al., 2010) both include matrix descriptions of recognition motifs for TFs across multiple species generated from a variety of methodologies, including compiled sequences, SELEX, PBMs and B1H (Portales-Casamar et al.; Wingender, 2008; Wingender et al., 2000). The Redfly database, which is specific to *Drosophila*, provides an extensive compilation of published experimental data identifying CRMs and individual TF binding sites within these CRMs (Halfon et al., 2008). The

Uniprobe database provides specificity information for TFs derived from a single technique, PBMs, providing access to the underlying raw data, which allows investigators to directly employ the binding site preferences determined by the data producer or to develop alternative representations of these data to describe recognition (Newburger and Bulyk, 2009).

FlyFactorSurvey provides an important complement to these existing databases. The current version focuses exclusively on the description of DNA binding specificities for *Drosophila* TFs determined by B1H and other methods. This database provides a repository for an ongoing project to determine specificities for all TFs in *Drosophila*, which is one of the primary model organisms for the analysis of transcriptional regulatory networks in metazoa. This database houses more than 400 recognition motifs for over 200 factors often generated from hundreds to thousands of selected binding sites. In keeping with the spirit of other genome-wide analysis projects, a large number of these binding specificities have been released prior to publication to facilitate the use of this information for the analysis of transcriptional regulation at the level of individual transcription factors, genes, or regulatory pathways.

Database Content

The primary source of recognition motifs within FlyFactorSurvey is TF binding site selections performed using the B1H method (Meng et al., 2005; Noyes et al., 2008a; Noyes et al., 2008b). An

outline of the important selection parameters captured in the database, as well as the data processing pipeline is shown in Figure 1. In brief, the predicted DNA binding domain of each *Drosophila* TF is expressed as a fusion to a component of *E. coli* RNA polymerase and transformed into cells with a library of reporter plasmids containing a randomized DNA sequence upstream of a weak promoter driving expression of the *His3* gene. When plated on media lacking histidine and containing a His3 inhibitor, plasmids with a complementary binding site within the randomized region are required for colony growth. TF binding sites are recovered by sequencing the randomized region of the reporter plasmid recovered from visible colonies. The MEME algorithm (Bailey et al., 2009; Bailey et al., 2006) is used to identify enriched sequence motifs within these sequences, which should represent the DNA binding specificity of the assayed TF. These motifs can be represented in a variety of formats, including alignments of the sequences containing the motifs, a counts matrix depicting the number of sequences with a given base at each position and a position weight matrix (PWM) depicting the log-odds score at each position (Stormo, 2000). FlyFactorSurvey houses three classes of information regarding each selection. First, some general information about each transcription factor is provided, including direct links to other wide-ranging databases describing *Drosophila* genes and proteins, such as FlyBase and FlyMine (Drysdale, 2008; Lyne et al., 2007). Second, parameters of B1H selections and sequence analysis are described. Third, recovered sequences and the sequence motif output from MEME and images depicting the information content in these motifs are stored.

A strength of the B1H system is the variety of parameters that can be adjusted to ensure a successful selection or change the complexity of the recovered recognition motif. Many of these important

parameters are captured with the database (Figure 1). For each TF characterized using the B1H system, the type of the DNA-binding domain(s) and the amino acid sequence of the portion of the protein characterized is provided. In some cases this is the entire coding sequence, but in others it is the region spanning the DNA-binding domain as well as conserved flanking elements that may play accessory roles in recognition. For heterodimeric TF complexes characterized by the B1H method, the amino acid sequence of both protein fragments is included. The TF expression vector employed dictates the TF fusion partner (either the alpha or omega subunit of *E. coli* RNA polymerase) and the strength of the promoter (lppC > UV5 > UV2) driving expression of the hybrid protein. In addition, some vectors contain two fingers of the zif268 zinc finger protein ("zif12") as an accessory DNA-binding domain to increase the activity of the hybrid protein (Noyes et al., 2008a). Finally, a subset of these vectors allow expression of a second monomer as an untethered protein for studies of heterodimeric complexes. The vector name provided within the database captures each of these parameters. Additional selection parameters captured in the database include the type of binding site library, which can vary in the length of the randomized region and contain a fixed DNA binding site for the zif12 protein fusion partner. The stringency of the selection is primarily influenced by the concentration of the inducer (IPTG) regulating TF expression and concentration of the His3 inhibitor, (3AT), where higher concentrations of inhibitor require higher affinity interactions between the hybrid protein and DNA binding site for colony growth.

TF DNA binding sites are characterized from bacterial colonies by two methods. First, the randomized region is amplified and sequenced (Sanger method) from 24 to 48 individual colonies.

MEME is used to identify an overrepresented sequence motif within this population of sequences that should represent the recognition motif of the TF. If a significant motif is identified, the randomized region is amplified from a pool of all colonies on the selection plate, characterized by SOLEXA sequencing and the unique sequences are analyzed using MEME. The Sanger-generated motif is compared with the SOLEXA-generated motif as a quality control step. Aligned sequences comprising the MEME-identified motif are input into WebLogo(Crooks et al., 2004) to generate a Sequence logo (Schneider and Stephens, 1990) describing the information content at each position in the recovered motif. For each recognition motif the associated unique sequences, aligned sequences, counts matrix, and WebLogo image are captured within the database. For published B1H motifs, the associated Pubmed ID number is also provided.

FlyFactorSurvey can also host DNA binding specificity information derived from other experimental methods. In these cases, the construct and selection information may be incomplete, but counts matrices and WebLogo images describing the motif can still be generated and associated with a given TF and publication. The database currently contains this information for DNase I footprint derived motifs described by Bergman and colleagues (Bergman et al., 2005).

Database Structure and User Interface

Database schema and website

The TFDBS database application was developed in house using a MySQL relational database hosted in a database server as the back-end, with the business/presentation layer written using PHP, and client access through a standard web client (browser) hosted by a Apache server. The database consists of several tables. The TF table stores detailed information about the transcription factors. The PWM table stores information about the position weight matrices. The TF_PWM table links TFs to PWMs allowing a many to many relationship. The DNA_BindingSites table contains the raw selected sequences as well as the aligned sequences used to derive the PWM. The Selection table contains the detailed selection conditions used to obtain the sequences that ultimately generated the PWMs. The DNAbindingDomain table contains detailed information about the DNA binding domains associated with the TFs. Lastly, the Users table stores user information and associated roles for access control and monitoring. The database contains constraints, indices, and keys to ensure data integrity and high performance. In addition, each editable record contains meta-data to monitor recent edits with regards to the user, date, time and location for any changes. The detailed relationship among these tables can be viewed as the Entity Relation Diagram (ERD) at http://pgfe.umassmed.edu/TFDBS/Documentation/FFS_schema.pdf.

Data access

The website home page provides several paths to navigate to TF binding site data of interest. The header contains a link to a "browse" page that lists all TFs with associated DNA binding specificity

data within the database (currently 250 factors). Links to individual TF summary pages - one per factor - are on the left of each table in the "view" column. The home page also provides two search windows that allow users to either search for TFs of interest based on gene name or other identification information, or to search for TFs based on the presence of the type of DNA binding domain motif contained within the protein (e.g. homeodomain or C_2H_2 zinc finger). Each of these searches can be restricted to comprise only TFs with associated binding specificity information within the database or only TFs characterized using the B1H method. These searches return matching TFs in the same format provided in the browse page.

The TF summary page for each factor is split into two sections (Figure 2). The top section provides key descriptive information for the TF, including links to the FlyBase (Tweedie et al., 2009) and FlyMine (Lyne et al., 2007) databases. Below the TF summary are individual panels for each associated recognition motif. In most cases, multiple motifs are available for each TF. These may be derived from different methods to identify binding sites (B1H and DNAse I), different sequencing methods, different selection conditions or different expression constructs. In addition, TFs that recognize their bindings sites as dimers may have different motifs associated with different hetero- or homo-dimeric binding partners. For each recognition motif, the WebLogo image is shown and download buttons allow retrieval of the aligned binding sites, all unique raw sequences analyzed to discover the motif, or matrices representing the motif in formats compatible with different sequence analysis programs. Information describing the parameters of the B1H selection that generated the motif is provided on the right side of each motif. In the case of selections performed with

heterodimers, a link to the TF summary page for the partner and its amino acid sequence are also displayed.

FlyFactorSurvey also provides an option for obtaining binding site information and matrices for all stored motifs as a flat file download. The home page header contains a link to a Downloads and Resources page. As on each TF summary page, the motif sequences and matrices for the entire dataset are accessible in a variety of formats to maximize the ability of biologists to employ data for large numbers of factors with existing computational tools.

Analysis tools

The sequence and motif formats provided for download are compatible with many commonly used computational tools to analyze TF binding specificities and to map potential binding sites for a TF or set of TFs within a genomic sequence. Two analysis tools of use to the *Drosophila* research community have been implemented in conjunction with this database. The first is an implementation of the TOMTOM motif comparison tool from the MEME suite (Bailey et al., 2009; Gupta et al., 2007). A version of the tool populated with all FlyFactorSurvey motifs is accessible via a link from the home page. This program allows a user to input a query motif and identify similar motifs within the database. For example, an investigator might identify an enriched motif in promoters of genes expressed in a given cell type and then query whether any of the TFs in FlyFactorSurvey have a similar DNA binding specificity. Each resulting match provides an image of the query and subject

motifs and a link to the TF summary page containing the matching motif within the database. An example of such a search and the resulting TOMTOM output is described in Supplementary Figure 1.

The second tool is an implementation of GenomeSurveyor that allows investigators to examine the distribution of matches to a given DNA binding site motif within regions of the *Drosophila* genome. GenomeSurveyor uses a Hidden Markov Model to score DNA binding site motif matches for single or multiple TFs in 500 base pair regions of the *Drosophila* genome and then display them as Z score tracks on a genome browser (Gbrowse) (Donlin, 2009; Noyes et al., 2008b; Sinha et al., 2004). Within the TF summary page for a TF in FlyFactorSurvey, each listed motif has a link to a GenomeSurveyor page where the default settings display three Z score tracks for the motif calculated over the *Drosophila melanogaster* genome and as averages across the genomes of two or eleven *Drosophila* species (Figure 3). An additional track displays the position of individual high scoring matches to the motif. When viewed in smaller genomic regions, the matching genomic sequence can be directly observed. The default genomic region (20 kB surrounding the *eve* gene) can be shifted to any desired region of the genome.

Availability

All data is freely available for distribution at the website. The authors request that they be contacted if multiple unpublished motifs from the database are being used prior to formal publication by the authors. Database and website code are available on request.

Acknowldegements

We would like to thank Marcus Noyes and Adam Richards for their contributions to the initial stages of this project.

Funding

This work was supported by the National Human Genome Research Institute of the National Institutes of Health, [grant R01 number HG004744-01 to MHB and SAW].

Figure Legends

Figure 3.1. Schematic of data flow into FFS database. The majority of motifs present in the database originate from B1H binding site selections. Information on the factor constructs and selection conditions is captured within the database for each motif generated. Clones from each selection are sequenced either individually via Sanger sequencing or as a pooled population via SOLEXA sequencing, and binding site motifs are identified from these sequences using MEME (Bailey and Elkan, 1994). The FFS database provides users with access to published and unpublished motif information on our characterized *Drosophila* TFs as well as links to tools to mine our database of motifs as well as utilize these for searches within the *Drosophila* genome.

Figure 3.2. Screen shot of *bicoid* summary page within FlyFactorSurvey. (top) Header information for each factor contains identification information, the type(s) of DNA binding domains found within the gene, and links to factor information in Flybase (Tweedie et al., 2009), Unipro (The UniProt Consortium, 2010) and FlyMine (Lyne et al., 2007). (bottom) Recognition motifs determined for the factor through different methods, under different conditions, or assayed via different sequencing methods are displayed in independent panels. Each panel displays the recognition motif as a Sequence logo (Schneider and Stephens, 1990) and download buttons to obtain count, position-specific probability (PSPM) or position-specific scoring (PSSM) matrices. The other information panel summarizes the methods used to select and sequence the factor binding sites. For motifs determined using the B1H system, the expression vector and the selection conditions are indicated,

where different stringencies can result in motifs with different complexity. In this example, only two of the four *bicoid* Motif panels within the database are shown; these panels illustrate that the increased inhibitor concentration used in the selection to generate the lower panel resulted in a more stringent motif. For each individual motif, a direct link is provided to view the relative frequency of the motif in the *Drosophila* genome using Genome Surveyor (see Figure 3).

Figure 3.3. Screen shot of Genome Surveyor interface directly linked from the Bcd_SOLEXA motif within the FFS database. The relative enrichment (profile) of this motif in 500 bp windows surrounding the *eve* locus is represented as a Z score relative to the genome-wide average. Three different motif profiles are shown: Single species (*D. melanogaster*), Multi species (*D. melanogaster* & *D. pseudoobscura*) and BMPhylo (11 Drosophila species). Individual high scoring sequence motif matches are also shown at bottom. This tool provides a rapid assessment of the overrepresentation of any motif in the database within the *Drosophila* genome and additional functions such as combined motif searches (Noyes et al., 2008b).

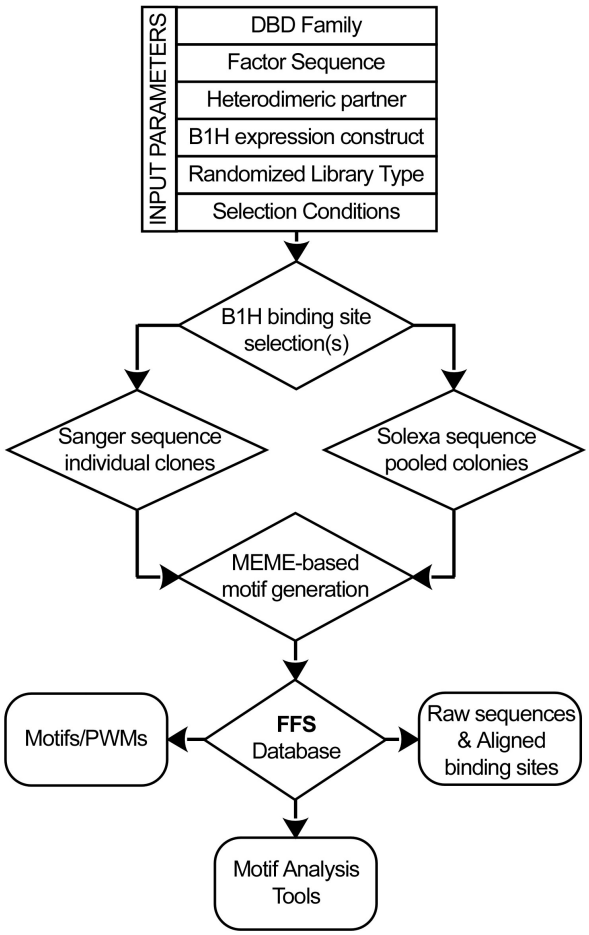


Figure 3.1 Schematic of data flow into FFS database

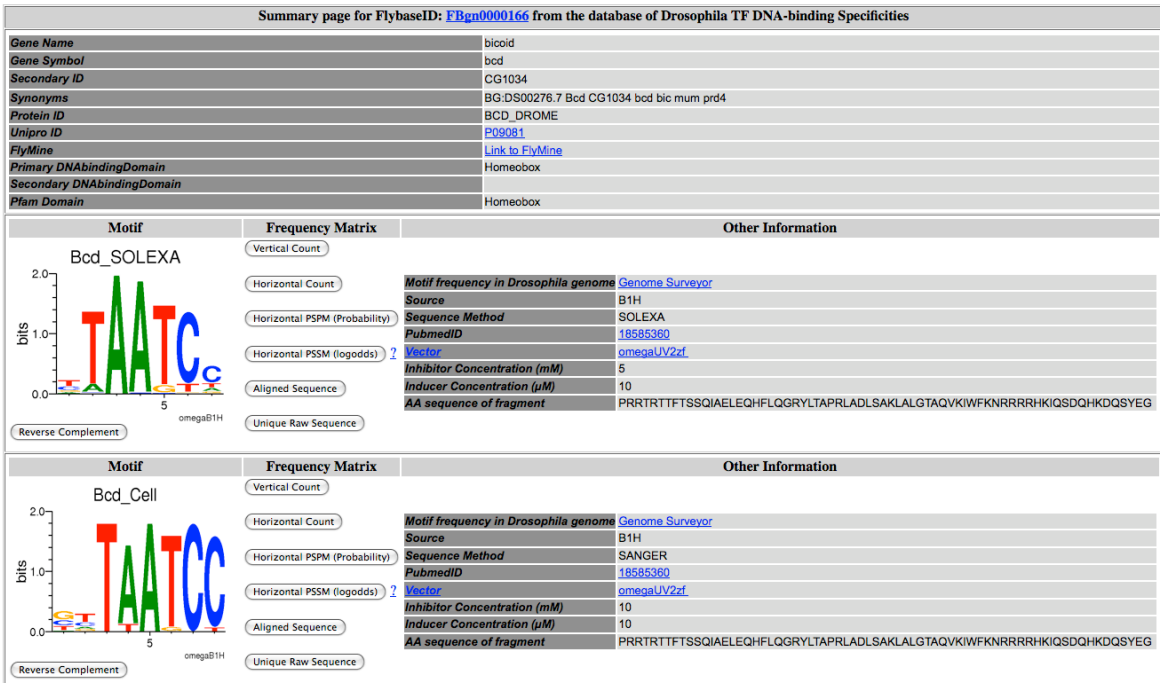
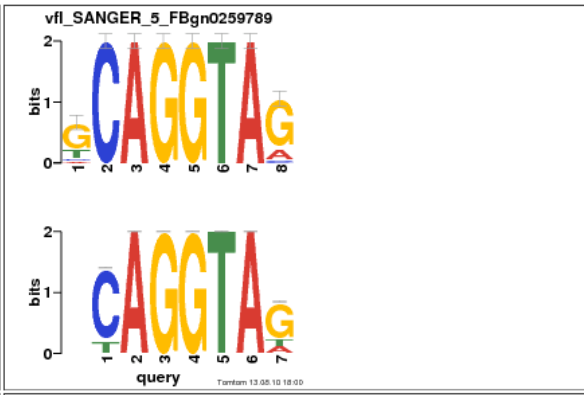
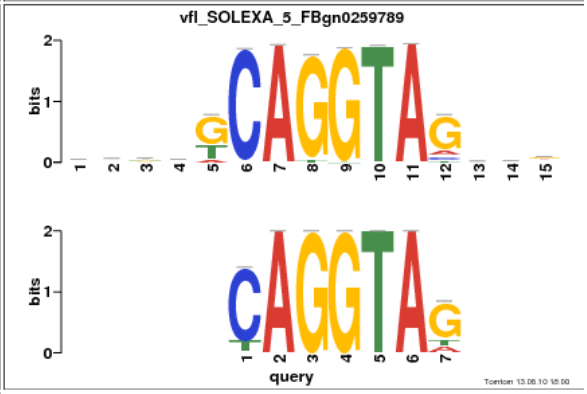


Figure 3.2 Screen shot of *bicoid* summary page within FlyFactorSurvey

Figure 3.3 Screen shot of Genome Surveyor interface directly linked from the Bcd_SOLEXA motif within the FFS database

Target Motif: vfl_SANGER_5_FBgn0259789 **Target Description:** **Query Motif:** query **Query Description:** ***p*-value:** **1.9e-06** ***E*-value:** **0.0009** ***q*-value:** **0.0013** **Overlap:** 7 **Query Offset:** 1 **Orientation:** + **Figures:** [EPS][PNG]	
Target Motif: vfl_SOLEXA_5_FBgn0259789 **Target Description:** **Query Motif:** query **Query Description:** ***p*-value:** **2.9e-06** ***E*-value:** **0.0014** ***q*-value:** **0.0013** **Overlap:** 7 **Query Offset:** 5 **Orientation:** + **Figures:** [EPS][PNG]	
Total matches with *q*-value ≤ 0.5: 2	

Supplemental Figure 3.1 Example of TOMTOM analysis identifying a TF within our database that recognizes a motif similar to an enriched promoter sequence motif.

The "TAGteam" motif was identified as a sequence motif related to the consensus sequence CAGGTAG that is overrepresented in the upstream regions of genes expressed in the *Drosophila* pre-cellular blastoderm embryo (see Sup Fig. 1 in (Bosch et al., 2006)). We constructed a position weight matrix from the four most enriched individual sequences in these promoters and used the TOMTOM program to search within FlyFactorSurvey for factors with similar DNA binding site motifs. The only significant hit found within our database is *vfl* (also known as *zelda*), which was recently shown to encode a TF that regulates gene expression through this motif in the early zygotic genome (Liang et al., 2008). The screenshot shown illustrates the output of the TOMTOM search where the query is the motif constructed from promoter-derived sequences and the subjects are two *vfl* motifs obtained from either Sanger or SOLEXA sequencing.

References

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res *37*, W202-208.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol *2*, 28-36.

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res *34*, W369-373.

Berger, M.F., and Bulyk, M.L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc *4*, 393-411.

Bergman, C.M., Carlson, J.W., and Celniker, S.E. (2005). Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics *21*, 1747-1749.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., and Celniker, S.E. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol *5*, R61.

Bosch, J.R.t., Benavides, J.A., and Cline, T.W. (2006). The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription. Development *133*, 1967-1977.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

Donlin, M.J. (2009). Using the Generic Genome Browser (GBrowse). Curr Protoc Bioinformatics *Chapter 9*, Unit 9 9.

Drysdale, R. (2008). FlyBase : a database for the Drosophila research community. Methods Mol Biol *420*, 45-59.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol *8*, R24.

Halfon, M.S., Gallo, S.M., and Bergman, C.M. (2008). REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. Nucleic Acids Res *36*, D594-598.

Janssens, H., Hou, S., Jaeger, J., Kim, A.R., Myasnikova, E., Sharp, D., and Reinitz, J. (2006). Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. Nat Genet *38*, 1159-1165.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science *316*, 1497-1502.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpaa, M.J.*, et al.* (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res *20*, 861-873.

Kazemian, M. (2010). Quantitative analysis of the Drosophila segmentation regulatory network using pattern generating potentials. PLoS Biology *in press*.

Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res *17*, 1919-1931.

Kinzler, K.W., and Vogelstein, B. (1990). The GLI gene encodes a nuclear protein which binds specific sequences in the human genome. Mol Cell Biol *10*, 634-642.

Li, X.Y., Macarthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C.L.*, et al.* (2008). Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm. PLoS Biol *6*, e27.

Liang, H.-L., Nien, C.-Y., Liu, H.-Y., Metzstein, M.M., Kirov, N., and Rushlow, C. (2008). The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. Nature *456*, 400-403.

Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., Janssens, H., Ji, W., McLaren, P., North, P.*, et al.* (2007). FlyMine: an integrated database for Drosophila and Anopheles genomics. Genome Biol *8*, R129.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K.*, et al.* (2006). TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. Nucl Acids Res *34*, D108-110.

Meng, X., Brodsky, M.H., and Wolfe, S.A. (2005). A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol *23*, 988-994.

Newburger, D.E., and Bulyk, M.L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res *37*, D77-82.

Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008a). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell *133*, 1277-1289.

Noyes, M.B., Meng, X., Wakabayashi, A., Sinha, S., Brodsky, M.H., and Wolfe, S.A. (2008b). A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucleic Acids Res *36*, 2547-2560.

Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucl Acids Res *38*, D105-110.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E.*, et al.* (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-2309.

Roulet, E., Busso, S., Camargo, A.A., Simpson, A.J., Mermod, N., and Bucher, P. (2002). High-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites. Nat Biotechnol *20*, 831-835.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res *18*, 6097-6100.

Schroeder, M.D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E.D., and Gaul, U. (2004). Transcriptional control in the segmentation gene network of Drosophila. PLoS Biol *2*, E271.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gaul, U. (2008). Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature *451*, 535-540.

Sinha, S., Schroeder, M.D., Unnerstall, U., Gaul, U., and Siggia, E.D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. BMC Bioinformatics *5*, 129.

Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*, 16-23.

The UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. Nucl Acids Res *38*, D142-148.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science *249*, 505-510.

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R.*, et al.* (2009). FlyBase: enhancing Drosophila Gene Ontology annotations. Nucl Acids Res *37*, D555-559.

Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief Bioinform *9*, 326-332.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res *28*, 316-319.

Zeitlinger, J., Zinzen, R.P., Stark, A., Kellis, M., Zhang, H., Young, R.A., and Levine, M. (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes Dev *21*, 385-390.

Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring binding energies from selected binding sites. PLoS Comput Biol *5*, e1000590.

Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. Nucleic Acids Res *37*, e151.

Chapter 4: A Modified Bacterial One-Hybrid System Yields Improved Quantitative Models of Transcription Factor Specificity[1]

[1] This chapter was adapted from: Christensen, R.G., Gupta, A., Zuo, Z., Schriefer, L.A., Wolfe, S.A., and Stormo, G.D. (2011). A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. Nucleic acids research. Epub April 20. I performed all of the computational work, produced all of the figures and wrote the bulyk of the paper. The experiments were designed by Gary and I. Most of the experimental work was done in the Wolfe lab by Ankit Gupta. In the Stormo lab, Larry Schriefer performed some of the 28 bp selections and Zheng Zuo did all of the B1H selections in liquid media.

Abstract

We examine the use of high-throughput sequencing on binding sites recovered using a bacterial one-hybrid (B1H) system and find that improved models of transcription factor (TF) binding specificity can be obtained compared to standard methods of sequencing a small subset of the selected clones. We can obtain even more accurate binding models using a modified version of B1H selection method with constrained variation (CV-B1H). However, achieving these improved models using CV-B1H data required the development of a new method of analysis - GRaMS (Growth Rate Modeling of Specificity) - that estimates bacterial growth rates as a function of the quality of the recognition sequence. We benchmark these different methods of motif discovery using Zif268, a well characterized C_2H_2 zinc finger transcription factor on both a 28bp randomized library for the standard B1H method and on 6bp randomized library for the CV-B1H method for which forty-five different experimental conditions were tested: 5 time points and three different IPTG and 3-AT concentrations. We find that GRaMS analysis is robust to the different experimental parameters whereas other analysis methods give widely varying results depending on the conditions of the experiment. Finally, we demonstrate that the CV-B1H assay can be performed in liquid media, which produces recognition models that are similar in quality to sequences recovered from selection on solid media.

Introduction

Determining the specificity of transcription factors (TFs) is an important step in elucidating regulatory networks. It is also an essential step in developing rules describing the relationship between the protein sequence of a TF and its preferred binding sites, which can be used to predict the specificities of uncharacterized TFs and to design TFs with novel specificities. Traditionally, determining the specificity of a TF was a slow and laborious process. Recent technological advances have greatly increased the rate at which new TFs can be analyzed (Stormo and Zhao, 2010). One new method, MITOMI (Fordyce et al., 2010; Maerkl and Quake, 2007), provides good estimates of binding affinities to different DNA sequences in a moderately high-throughput format, including a recent advance that allows affinity measurements for all possible 8-long (8mer) binding sites. PBMs (Protein Binding Microarrays) were first described about ten years ago, and recently have been implemented in a format that allows all 10mers to be included in the analysis (Berger et al., 2006; Bulyk et al., 2001; Mukherjee et al., 2004). CSI (Cognate Site Identification) is a related technique with similar capabilities (Hauschild et al., 2009; Puckett et al., 2007; Warren et al., 2006). SELEX (Systematic Evolution of Ligands by Exponential Enrichment) has long been used to determine the specificity of TFs, but initially it was used in a low-throughput manner that only returned the consensus sequence and some measure of the variability tolerated at different positions (Blackwell and Weintraub, 1990; Oliphant et al., 1989; Tuerk and Gold, 1990; Wright et al., 1991). Several years ago it was coupled with a serial analysis of gene expression (SAGE) method to create a moderate throughput method that greatly increased the accuracy of specificity determination (Roulet et al.,

2002). In the last year SELEX has been scaled up to utilize next generation sequencing methods and is now capable of determining highly accurate specificities for TFs (Jolma et al., 2010; Zhao et al., 2009; Zykovich et al., 2009). One advantage of SELEX over the other methods is that it is capable of analyzing binding sites of essentially any length; the only limitation is that the library of potential binding sites is limited to about 10^{12} and the number of sites that can be sequenced is about 10^{8}, both of which are much greater than all possible 10mers (10^{6} different sequences), the limit of methods such as PBM.

Another method to determine the binding specificities of TFs is a bacterial one-hybrid (B1H) system (Meng et al., 2005; Meng et al., 2006; Noyes et al., 2008b). In this approach a TF is expressed in *E. coli* fused to the ω subunit of RNA Polymerase. This turns any DNA binding protein into an activator of transcription. A library of randomized binding sites is located upstream of a weak promoter driving expression of a selectable gene. Under appropriate growth conditions only sites with high affinity for the TF will survive selection. As with SELEX, this approach has the advantage that binding sites of any size can be studied, the only limitation being that the library size is constrained by the transformation efficiency of bacteria, which is about 10^{8} individual sequences. Another advantage of this approach is that the TF does not have to be purified, or expressed *in vitro*; any TF that can be functionally expressed in *E. coli* can be assayed with this method making it rapid and easy to use. It can also be used with TFs that have very low specificity by fusing them to two fingers of a zinc finger protein to create a chimeric protein with sufficient specificity and affinity for function within the B1H system (Noyes et al., 2008b). Previously binding sites were sequenced from a small number of surviving colonies, typically 20-40, and a model of the specificity of the TF would

be inferred using a motif finding program (Noyes et al., 2008a; Noyes et al., 2008b), such as Consensus (Hertz and Stormo, 1999) or MEME (Bailey and Elkan, 1994).

Regardless of the method employed, the goal is to obtain an accurate quantitative model of the specificity of the TF. In this paper we test several different variations of the B1H method, including different approaches to analyzing the data, using the well-characterized Zif268 zinc-finger protein as the standard for comparison. We find that the standard B1H method, which employs a large randomized library and determines the binding sites from a few selected colonies, has reasonable accuracy which can be further improved by the application of high-throughput sequencing methods that determine the frequencies of selected binding sites across the distribution from high affinity to low affinity. We also show that using a library with limited variability, in which part of the binding site is fixed and the other part randomized, combined with high-throughput sequencing allows us to measure the growth rate of colonies containing each site in the library. An algorithm that models the relationship between binding energy and growth rate can then further increase the accuracy of the quantitative specificity model. We call this experimental approach CV-B1H for "constrained variation B1H" and the analysis method GRaMS for "Growth Rate Modeling of Specificity" and we compare its performance under many different experimental protocols to other experimental and analysis methods. Overall we show that CV-B1H is an inexpensive, fast and easy method for accurately determining the specificity of a TF, where optimal results are obtained when the data are analyzed using an appropriate model that accounts for the dynamic growth of cells.

Methods

Zif268 B1H selections

All of the B1H binding site selections were performed as described previously using an ω-Zif268 fusion protein expressed from a UV2 promoter in the plasmid pB1H2ω (Noyes et al., 2008b). Zif268 was used for these experiments because it has been thoroughly characterized by a number of other methods allowing comparison of the recognition models we obtain to its previously defined specificity.

Randomized 28bp binding site library

Four independent B1H binding site selections for Zif268 were performed using a 28-bp randomized library in pH3U3 reporter vector as previously described (Noyes et al., 2008b). Approximately $2x10^7$ cotransformed cells containing the library and the ω-Zif268 expression plasmid were plated under each selection condition on selective media plate containing 0, 2(replicated) or 5 mM 3-AT and 10 uM IPTG. These selections were incubated at 37 °C for 36-48 hours following which surviving cells were washed off the plate as a pool. The plasmid DNA from the pooled colonies was isolated. Library regions from the recovered reporter plasmids were PCR-amplified, adaptor ligated with barcodes identifying each selection, and the library for Illumina-sequencing was prepared as previously described (Gupta et al.). The initial 28bp library was also Illumina-sequenced where about 10^7 reads were obtained to provide a background model for subsequent motif analysis.

Randomized 6bp binding site library

The binding site library

(GCGGCCACTGGGCAGCTGGCCA*NNNN*AAAAAT**<u>NNNNNN</u>**<u>GCGG</u>TACCTAGGTTCTT
CGAATTC) cloned between the *Eco*RI and *Not*I sites in pH3U3 contains two different randomized regions: a 6bp element (bold underlined) that is associated with the four 3' bases of the Zif268 recognition sequence (GCGG, underlined), and a 4bp randomized region (italics) that serves as an internal control to identify sequences that may be enriched in the selections or preparation for sequencing sample due to jackpot effects. We did not observe any evidence of a jackpot effect. Auto-activating clones within this library were removed by 5-FOA counter-selection as previously described (Noyes et al., 2008b). Approximately 10^6 co-transformed cells containing the library and the ω-Zif268-expression plasmid were plated under each selection condition on selective media plates containing 0.5, 1 or 2 mM 3-AT and 0, 10 or 50 μM IPTG, where these selections were incubated at 37 °C for 4, 8, 12, 18 or 24 hours. This was a total of 45 independent selections. At the desired time-point surviving cells were washed off the plate as a pool. Isolated plasmid DNA from the pooled cells was prepared for Illumina sequencing as for the 28bp library. Using barcodes for each experiment, sequences from all 45 experiments were obtained from a single Illumina sequencing lane that contained over 15 million reads, leading to an average of about 300,000 binding sites per experiment. There are only 4096 different 6mer sites so this quantity of sequences is sufficient for good coverage of all possible binding sites. We also performed CV-B1H from the same initial library in liquid media with 5 mM 3-AT and 50 μM IPTG. After 4 hrs the cells were pelleted, plasmids

isolated and they were prepared for Illumina sequencing as with the experiments on plates. We independently sequenced the counter selected library, which was the input to each of the binding site selection experiments, to define the initial frequency of each 6mer. More than 16 million reads were obtained and every 6mer was observed at least 472 times. This allowed us to determine the enrichment of each site after selection. The sequences from each dataset are available at http://ural.wustl.edu/htb1h_zif68 and from the GEO database (GSE26767).

Binding site modeling using existing programs

We model the binding energy of Zif268 for any sequence using a position weight matrix (PWM) (Stormo, 2000). We used four different motif discovery methods on the different datasets. BioProspector (Liu et al., 2001) was used on both the 28bp datasets with a site size of 10bp, and on the 6bp datasets with a site size of 6bp. For the 6bp dataset the orientation was fixed whereas for the 28bp datasets sites could be discovered in either orientation. A 3rd-order Markov model, based on the sequences of the respective initial libraries, was used for the background model. MEME (Bailey and Elkan, 1994) was run only on the 28bp datasets because it requires sites longer than 6bp for motif analysis. The 3000 most abundant 28mers served as the input to MEME with each sequence being used only once in the input set. Sites were allowed to occur in either orientation. BEEML (Zhao et al., 2009) requires the alignment of the binding sites for motif analysis, so it was used only on the 6bp datasets with the background model derived from the 6mer counts in the initial library. On the 6bp datasets we also tested a simple Log-Odds method that determines the value of each PWM element from the ratio of the observed frequency of each base at each position in the aligned binding sites to

the observed frequency of each base at each position in the initial library (from the randomized region). We also tested the accuracy obtained from the consensus sequence, GCGTGGGCGG, where its energy is set to 0 and the optimal mismatch energy (over all possible integer values) is 2.

Binding site modeling based on growth rate analysis

We model protein-DNA binding using a biophysical model described previously (Zhao et al., 2009). Briefly, the probability that the sequence S_i is bound at equilibrium is:

$$P(S_i\ bound) = \frac{[TF \cdot S_i]}{[TF \cdot S_i] + [S_i]} = \frac{[TF]}{[TF] + K_d(S_i)} \quad (1)$$

where K_d is the dissociation constant and square brackets indicate concentrations. It is convenient to express the energy of binding, E_i, relative to the Gibbs free energy of binding to a reference sequence; we use the consensus sequence, in units of RT, with its energy defined as, $E_{ref} = 0$:

$$P(S_i\ bound) = \frac{1}{1 + e^{E_i - \mu}} \quad (2)$$

where

$$E_i = \Delta\Delta G_i^\circ / RT = (\Delta G_i^\circ - \Delta G_{ref}^\circ) / RT \quad (3)$$

and

$$\mu = \ln \frac{[TF]}{K_d(S_{ref})} \quad (4)$$

Binding sites with $E_i=\mu$ have a binding probability of one-half.

In order to grow and replicate, cells must express sufficient His3 enzyme to meet their histidine requirements. We define the growth rate of an allele as the number of doublings that a cell possessing it undergoes each hour during exponential growth phase. The equation

$$N_i(t) = N_i(0)2^{r_i t} \tag{5}$$

describes the exponential growth of a colony, where t is the number of hours, $N_i(t)$ is the final number of cells possessing site S_i present at time t, r_i is the growth rate for cells containing that site in doublings/hr, and $N_i(0)$ is the initial number of cells with that site at time 0.

Histidine is a rate limiting reagent, and we make the simplifying assumption that the amount of histidine is directly proportional to the occupancy of the His3 promoter by the TF (up to some saturating level) and that the growth rate, r_i, of cells possessing S_i is directly proportional to the amount of histidine produced, up to a level where it is no longer limiting. The relationship between binding energy of the TF for site S_i and the growth rate is then:

$$r_i = \log_2\left(\frac{N_i(t)}{N_i(0)}\right)/t = \frac{M}{1+e^{E_i-\mu}} \tag{6}$$

where M is the maximum growth rate for these cells under the same conditions but with histidine not being limiting. Supplemental figure 1A shows a simulated ideal experiment where the counts for each sequence depend on the binding energies as described in the biophysical model of the preceding equations. Data taken at different time points will fall on different curves, but when converted to

growth rates all of the data sets converge to a common curve describing the relationship between growth rate and binding energy (Supplemental figure 1B).

We are only able to determine the frequency of each allele from the Illumina reads. In order to convert these frequencies into numbers of cells, we need to know the initial number of cells plated, n_I, and the final number of cells on the plate, n_F, at time t. The growth rates determined by the frequencies at time t will be off by a constant

$$c = \log_2\left(\frac{n_F}{n_I}\right)/t \tag{7}$$

such that

$$r_i = \log_2\left(\frac{f_i(t)}{f_i(0)}\right)/t + c \tag{8}$$

where $f_i(t)$ is the frequency of site S_i at time t, and $f_i(0)$ is the initial frequency of S_i before selection. We refer to the quantity

$$\frac{f_i(t)}{f_i(0)} \tag{9}$$

as the enrichment of site S_i at time t. For a given experiment, every growth rate will be off by the same constant. If we assume that the minimum growth rate is 0 (cells may not divide but they do not disappear from the plate) we can determine the constant by assuming the plateau of high energy binding sites represents a growth rate of 0. For the remainder of the paper, including all of the

figures, the calculated growth rates for each site have been adjusted such that the median of the high energy plateau is defined as 0.

For a given PWM, the predicted growth rates, $\hat{r}_i$, depends on the energy model via:

$$\hat{r}_i = \frac{M}{1 + e^{\vec{S}_i \cdot \vec{W} - \mu}} \tag{10}$$

where $\vec{S}_i$ is the encoded sequence, S_i, and $\vec{W}$ is the PWM. In this analysis, M was fixed to the maximum growth rate for each data set. We use the Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963; Moré, 1978) in a program called GRaMS (Growth Rate Modeling of Specificity) to perform a least squares fit between the measured and predicted growth rates in order to find the optimum PWM.

Assessment of different protocols and analysis methods

For each experimental dataset and each analysis method we obtain a position weight matrix (PWM). We adjust the elements such that those corresponding to the reference sequence are assigned 0, and the other elements are estimates of the binding energy differences for each other base at each position in the binding site, as proposed by Berg and von Hippel (24). We determine the accuracy of each method by measuring, using the squared Pearson Correlation Coefficient (R^2), how well it predicts the binding data from a single-round SELEX experiment (Zhao et al., 2009). In that experiment a large library of random 10mers were bound to Zif268 and the bound fraction as well as the initial library were Illumina-sequenced. For each PWM the values of μ and a non-specific binding

energy, E_{ns}, are found that maximize the fit for that model so that the comparisons are strictly between how well the PWMs capture the energy differences for each base at each position. BEEML (Zhao et al., 2009) was developed specifically to model that SELEX data so we determined its R^2 value when trained on the SELEX data directly as the maximum that any other PWM could be expected to obtain. This was 0.93 and 0.96 for the 10bp PWMs and 6bp PWMs, respectively. The remaining variance is probably due to experimental noise as well as binding energy contributions not captured by the simple PWM which are known to exist but be small for Zif268 (Benos et al., 2002).

Results

Selections from 28 bp library

We first characterized how well the PWM obtained from the 28bp library can predict the zif268 SELEX data. T shows a simulated ideal he PWM from Meng et al. (2005), which was based on only 17 sequences from selected colonies, gives an R^2=0.67. This is nearly as high as that obtained from a Zif268 PWM obtained from PBM data (R^2=0.69)(Berger et al., 2006) and is much better than simply using a consensus sequence (GCGTGGGCGG) to predict quantitative binding affinities, which gives an R^2=0.27. We next tested whether using high-throughput sequencing methods when applied to the B1H selected clones would provide a motif with even higher accuracy. We collected all of the cells on the plate from a selection using Zif268, which includes the large colonies, small colonies and even individual cells that have not divided, purified the binding site plasmids and subjected the entire mixture to Illumina sequencing. We did this for four different growth conditions and obtained

between 117,475 and 928,304 sequences from each selection (after removing poor quality reads that did not match the fixed sequences flanking the library). We used BioProspector and MEME to obtain alignments and PWMs from each dataset, and then used those PWMs to predict the SELEX data. The results from both motif discovery methods were nearly identical, with median R^2 of 0.79 and 0.80 for BioProspector and MEME, respectively (Figure 1). This is a significant improvement over the PWM based on only 17 sequences. We also tested how many reads are required to obtain that accuracy by randomly selecting subsets of sequences of various sizes from our datasets. We found that the maximum accuracy was achieved by both BioProspector and MEME analyses with a population of about 3000 reads. Thus, a large number of different B1H binding site selections can be multiplexed together in a single Illumina lane to minimize sequencing costs, while still obtaining good recognition models for each experiment.

Selections from 6 bp library

While quite good, those R^2 values still leave considerable room for improvement. We next tested whether we could get further improvement by fixing part of the binding site, in this case 4bp, and only randomizing the remaining 6bp for our B1H binding site selections. This eliminates problems related to aligning the binding sites because the TF should always prefer the orientation and position that overlaps the fixed region; we found no exceptions to that expectation in the analysis of the data that was generated. Moreover, because there are only 4096 different 6bp sequences, we can obtain good frequency estimates for all binding sites in both the initial library and in the selected sites while at the same time multiplexing many different experiments in a single Illumina lane. We averaged

about 300,000 reads for the each of 45 different selections that explore a range of different selection conditions (3-AT, IPTG and incubation time). Motif comparisons for data generated from these experiments were made to the subset of Zif268 SELEX data that contained the GCGG sequence in the last four positions on the binding site, to be consistent with the constraints in our selections. The BEEML PWM predicts the SELEX data (that it was trained on) with an R^2=0.96, which is the maximum we would expect from any other method trained on alternative datasets. The simple consensus sequence, with optimal mismatch penalty, predicts the SELEX data with an R^2=0.50 and the 6bp segment of the previous B1H Zif268 PWM (Meng et al., 2005) fits with R^2=0.74 (Figure 2). We performed a traditional B1H experiment on this library, picking and sequencing just 22 colonies, and the resulting PWM had an R^2 of only 0.52, much worse than the previous PWM from a 28bp library. Equally unexpected was that BioProspector and Log-Odds analyses on the various 6bp experiments were highly variable and generally much poorer than for the 28bp library. The median values of R^2 were only 0.52 and 0.54 for BioProspector and Log-Odds, respectively, and the maximum values were only 0.71 and 0.74 (Figure 2). Motifs generated from selections with short incubation times displayed the worst performance, but none of the experimental conditions performed very well on this library.

We think these results are explained by the fact that the initial library, which has been counter-selected to remove autoactivating sequences, has a very low proportion of the consensus binding site and some other closely related sites. Their low initial frequencies ensure that even after the 24hr time selections they have not become the most abundant sites, therefore leading to PWMs with sub-optimal parameters. Although both the Log-Odds method and BioProspector take the initial library

into account through their background estimates, those are only based on the total composition, in the case of Log-Odds, or a 3rd-order Markov model, neither of which really captures the explicit deficiency of specific binding sites, some of which are high affinity sites. We therefore tested the BEEML program on the 45 datasets. It takes into account each specific binding site in both the initial and selected libraries and, based on a biophysical model for enrichment based on affinity, does a non-linear regression to find the optimal parameters for a PWM. While its performance is still quite poor on the earliest time points, its median R^2 is 0.86 and its best is 0.92, both significantly better than the other methods (Figure 2). This level of performance makes BEEML analysis of the 6bp CV-B1H data even better than the BioProspector and MEME performance on the 28bp high-throughput B1H datasets (Figure 1).

Since we expect the differences in binding energies for different sites to affect their relative growth rates, we developed GRaMS to obtain optimal PWMs for CV-B1H data according to the model described in Methods. Supplemental Figure 1 shows, for ideal simulated data, how the occurrences of difference binding sites at various time points fall on different lines, but when converted to growth rates they all converge to a single line that shows the relationship between growth rate and binding energy. Figure 3 shows the results from one experiment (8hrs, 50μM IPTG, 2mM 3-AT) where the binding energies are the predictions from the GRaMS model (Figure 4). Supplemental Figure 2 shows the same curve for all 45 datasets. While obviously noisier than the simulated data, the curves are all very similar and are consistent with our model. Supplemental Figure 3 shows the logos for all 45 datasets. Note that the models are very similar indicating that with GRaMS analysis the resulting models are relatively insensitive to the exact experimental protocol. Motifs from the 4hr time points

are still the least accurate and those from the late time points have slightly reduced accuracy probably due to the onset of colony saturation for the highest affinity binding sites. At the earlier times points increased stringency, using higher concentrations of 3-AT, improved the quality of the motifs somewhat but the results are not much affected by the concentration of IPTG. In contrast, Supplemental Figure 4 shows the Logos for the 45 different PWMs obtained by BioProspector. While overall they are not too bad, they are more variable between different conditions and none fit the SELEX data as well as the GRaMS models.

Using GRaMS we obtained R^2 values with a median of 0.92 (Figure 2) and with a maximum value of 0.94, nearly as good as the maximum expected. The entire range is from 0.84 to 0.94, again indicating that the models are relatively insensitive to the exact experimental protocol. Only one sample gave an R^2 as low as 0.84, with the next lowest value being 0.88. On average, the lowest R^2 values were obtained using GRaMS models trained on the 4 hour data sets. None of the methods performed particularly well on these data sets, but GRaMS performed the best (Supplemental Figure 5). On datasets from the later time points the R^2 values ranged from 0.89 to 0.94 with a median of 0.92 (Supplemental Table).

The largest difference between the logos from GRaMS (Supplemental Figure 2) and from BioProspector (Supplemental Figure 3) is at position 5, where BioProspector nearly always shows a slight preference for A over G, whereas GRaMS has a somewhat larger preference for G over A, which is consistent with the SELEX data. This difference can be attributed to bias in the initial library (probably due to the counter-selection), which contains many more sequences with A at

position 5 than with G. Even after selection up to 24 hours, A remains the most common base in each dataset, which causes the BioProspector PWM to prefer A. But the growth rate of sites with G in position 5 is, on average, greater than for those with A, so GRaMS infers the higher affinity for G. Presumably BioProspector would perform better if the initial library were less biased, as we observed in the 28bp library, but one advantage of GRaMS is that it takes the bias into account directly through its background model and so is not strongly influenced by it.

We also performed HT-B1H in liquid culture and found that similar models could be obtained. Supplemental Figure 6 shows the motif obtained through GRaMS analysis after 4 hrs of growth which attained an R^2=0.93 on the SELEX quantitative predictions. This may be the most straightforward method to employ for selections in practice but we have not examined its performance in detail as we have with the plate growth method. Performing the B1H assay in liquid media has the potential artifact that the cells with low affinity sites may be able to grow using histidine made in excess by other cells, although this should not be a major limitation at early time points when the cell densities are very low.

Given that GRaMS performed so well with CV-B1H data we wondered whether it would also improve the models obtained from the 28bp B1H experiments. One limitation is that GRaMS does not generate a native alignment of binding sites from a population of selected sequences, it requires an existing alignment to generate a binding motif. Consequently, we generated PWMs with GRaMS using the aligned sites generated by BioProspector and MEME, and we used for the background model the counts of 10mers from the input 28bp library. This provided almost no improvement over

the original MEME and BioProspector models, where the median R^2 values increased by only 0.01 for both sets (data not shown). We think this stems from several factors including: an incomplete sampling of the binding sites, especially the low affinity sites, in the BioProspector and MEME alignments; an incomplete sampling of the sequences in the initial library for the construction of the background model; and the influence on the activity of a sequence by its distance from the promoter, which is evident in the strong preferences for the recovery of binding sites at specific registers in the randomized sequences, that is not accounted for within the current GRaMS model. Therefore as currently implemented, GRaMS does not provide an improved analysis method for general B1H experiments, even with high-throughput sequencing data, but with the appropriate experimental design, as in the CV-B1H experiments, it can be used to obtain highly accurate, quantitative models of TF specificity.

Discussion

The B1H assay has proven to be a robust technique applicable to a wide variety of different transcription factor families. For instance, it has recently been successfully applied to more than 200 different Drosophila transcription factors from a variety of different families (i.e. pfam families: bZIP_1, bZIP_2, CBF_beta, Fork_head, HLH, HMG_box, PAX, RHD, Runt, zf-C2H2, zf-C4) (Zhu et al.) We find that applying massively parallel sequencing methods to all of the selected binding sites on an entire plate can lead to more accurate, quantitative models of TF specificity. The

new models are more accurate than those obtained from the same library when sequencing only a few selected colonies, as might be expected. These new models are also slightly better than those obtained from PBM experiments on Zif268. Despite the increased accuracy from the high-throughput sequencing there remains substantial room for improvement. We show that by constraining the variability in the library, which eliminates ambiguities in the alignment of the sites and allows for deep sampling of the population, very accurate models can be obtained. However, even when using the CV-B1H protocol, the accuracy of the resulting motifs depends on the data analysis method employed. By measuring growth rates of cells across the distribution from high affinity to low affinity sites and using a biophysical model for the relationship between growth rate and binding energy, GRaMS is able to obtain more accurate models from B1H data than any other approach we tested. This approach is fairly insensitive to the exact B1H protocol used and we obtained good models under all of the variations that we tested except for the very early time point (4 hours). From selections in liquid culture we were able to obtain a good model even after only 4 hours of growth. Increased 3-AT concentrations, which increase the stringency of the selection, increased the accuracy of the resulting model slightly on average. The IPTG concentration had little effect, although 10 and 50 μM were slightly better than 0, on average.

We have used some simplifying assumptions in our biophysical model, but the fact that we consistently get good PWMs suggests that the assumptions are reasonable. In particular we have not used a coupling factor, referred to as λ by Berg and von Hippel (Berg and von Hippel, 1987, 1988), that relates the binding energy to the functional activity of a binding site. In essence we assume $\lambda=1$, which is within the range of 0.5 to 1.5 that they found for several natural systems. If we empirically

determine a λ for each of our PWMs that converts them to the optimal PWM for Zif268, we find that it decreases at late times, but we think this is best explained by the saturation effects of the faster growing colonies beginning to reach their maximum size. For early time points, and for most conditions, assuming λ=1 appears to be a good approximation. It is unclear whether this relationship will hold true for other TFs, but the fact that the TFs analyzed by this approach use the ω-fusion as the means of coupling DNA-binding to transcriptional activation suggests that assuming λ=1 is likely to be reasonable in general.

AVAILABILITY

The sequences from each dataset are available at http://ural.wustl.edu/htb1h_zif68 and via GEO (accession GSE26767). GRaMS was implemented as a MATLAB program and is available for download from http://ural.wustl.edu/resources.html#Software.

FUNDING

This work was supported by the National Institutes of Health [grants R24GM078369 , R01HG004744 and R01HG00249].

ACKNOWLEDGEMENTS

We thank the member of the Stormo lab for helpful discussions and advice; in particular, we thank Yue Zhao for his insightful advice. We also thank the Washington University Genomic Technology Access Center and the Center for Genome Sciences for Illumina sequencing support.

Figure Legends

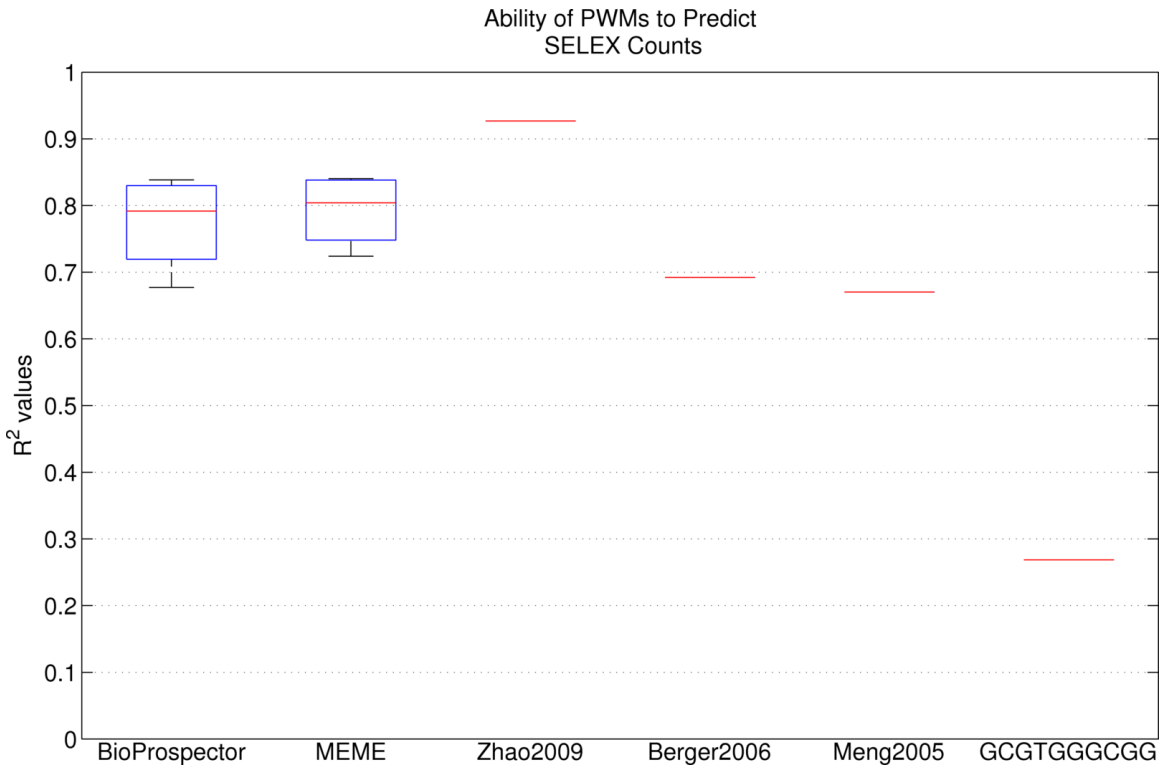


Figure 4.1 Boxplot showing the ability of the set of MEME and BioProspector motifs learned from the four 28bp B1H data sets to predict the SELEX data.

For each PWM, the squared Pearson correlation coefficient (R^2) was calculated to determine the correlation between the predicted and observed SELEX counts. The performance of three PWMs from the literature is also shown. Zhao2009, Berger2006 and Meng2005 were learned from SELEX, PBM and B1H data, respectively. The GCGTGGGCGG consensus sequence PWM was constructed using an optimal mismatch penalty term.

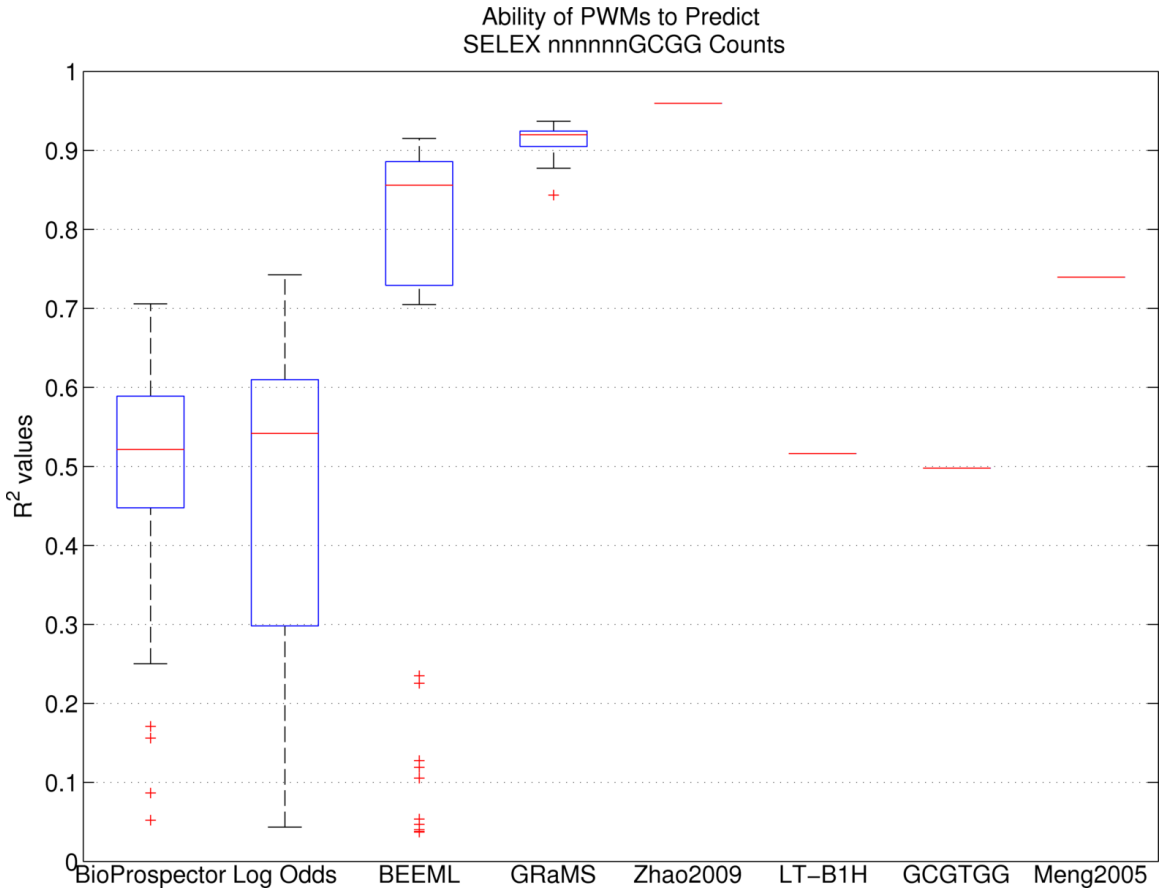


Figure 4.2 Boxplot showing the ability of the 45 PWMs produced by each analysis method using each B1H data set as training data to predict the SELEX nnnnnnGCGG data.

For each model, the squared Pearson correlation coefficient (R^2) was calculated to determine the correlation between the predicted and observed SELEX counts. The performance of four individual PWMs is also indicated. Two of these PWMs, Zhao2009 and Meng2005, were obtained from published SELEX and B1H studies respectively; the first six positions of these PWMs were used. The LT-B1H PWM was learned from 22 sequences obtained from a CV-B1H experiment. The GCGTGG consensus sequence PWM was constructed using an optimal mismatch penalty term.

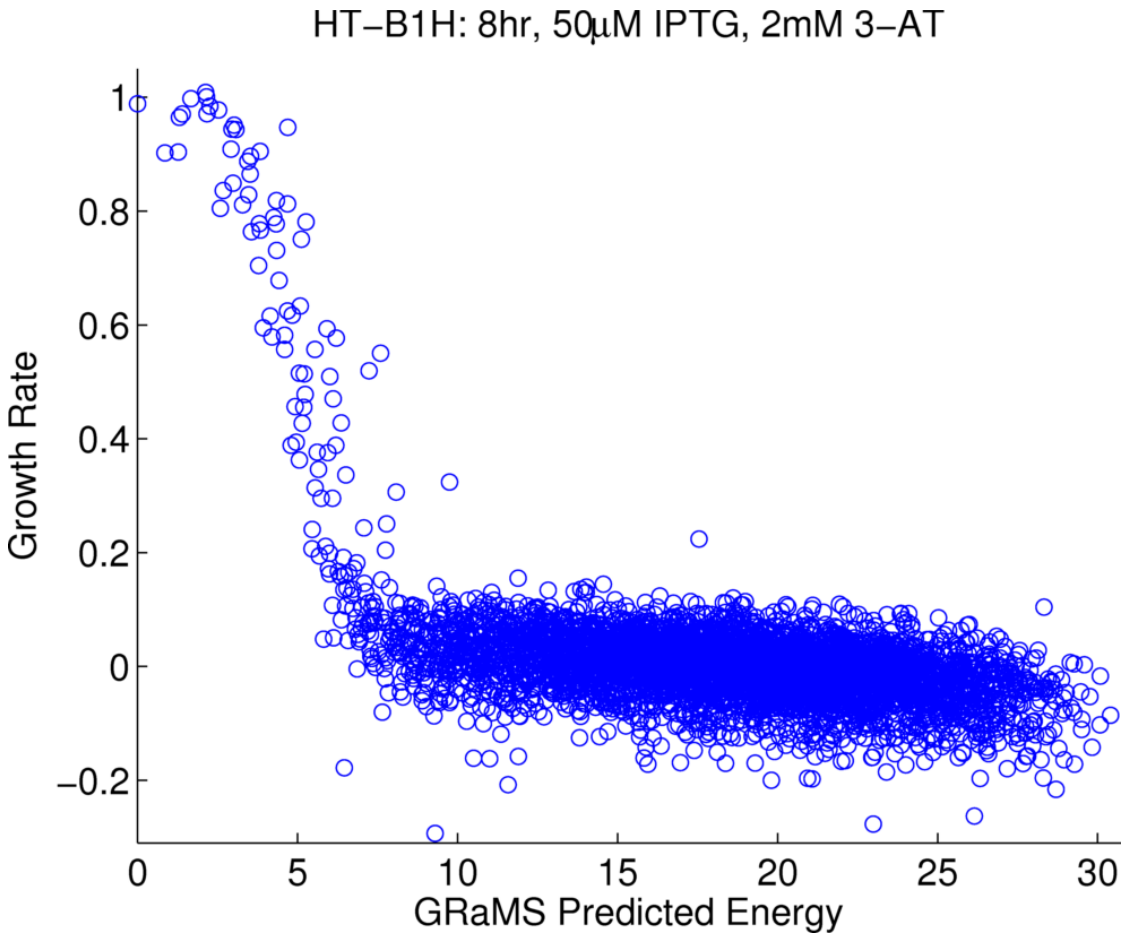


Figure 4.3 Plot of predicted energies versus growth rates per 6mer.

The GRaMS PWM (8 hour, 50µM IPTG, 2mM 3-AT), was used to predict the energies. The growth rates (shifted so that the median value is zero) are from the 8 hour, 50µM IPTG, 2mM 3-AT dataset used to estimated the GRaMS PWM.

Figure 4.4 Sequence logo for the GRaMS PWM obtained from the same dataset.

The y-axis indicates the information content of each position in bits. Sequence logos were produced using in-house software, svgSeqLogo, written by RGC.

Supplemental Material

Due to the large amount of supplemental material, we refer the reader to the NAR web site:

http://nar.oxfordjournals.org/content/early/2011/04/20/nar.gkr239.long

doi: 10.1093/nar/gkr239

References

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol *2*, 28-36.

Benos, P.V., Bulyk, M.L., and Stormo, G.D. (2002). Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res *30*, 4442-4451.

Berg, O.G., and von Hippel, P.H. (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol *193*, 723-750.

Berg, O.G., and von Hippel, P.H. (1988). Selection of DNA binding sites by regulatory proteins. Trends Biochem Sci *13*, 207-211.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol *24*, 1429-1435.

Blackwell, T.K., and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science *250*, 1104-1110.

Bulyk, M.L., Huang, X., Choo, Y., and Church, G.M. (2001). Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc Natl Acad Sci U S A *98*, 7158-7163.

Fordyce, P.M., Gerber, D., Tran, D., Zheng, J., Li, H., DeRisi, J.L., and Quake, S.R. (2010). De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nat Biotechnol *28*, 970-975.

Gupta, A., Meng, X., Zhu, L.J., Lawson, N.D., and Wolfe, S.A. (2011). Zinc finger protein-dependent and -independent contributions to the in vivo off-target activity of zinc finger nucleases. Nucleic Acids Res *39*, 381-392.

Hauschild, K.E., Stover, J.S., Boger, D.L., and Ansari, A.Z. (2009). CSI-FID: high throughput label-free detection of DNA binding molecules. Bioorg Med Chem Lett *19*, 3779-3782.

Hertz, G.Z., and Stormo, G.D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics *15*, 563-577.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpaa, M.J.*, et al.* (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res *20*, 861-873.

Levenberg, K. (1944). A method for the solution of certain problems in least squares. . Quart Applied Math *2*, 164-168.

Liu, X., Brutlag, D.L., and Liu, J.S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput, 127-138.

Maerkl, S.J., and Quake, S.R. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. Science *315*, 233-237.

Marquardt, D.W. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. . SIAM Journal on Applied Mathematics *11*, 431-441.

Meng, X., Brodsky, M.H., and Wolfe, S.A. (2005). A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol *23*, 988-994.

Meng, X., Smith, R.M., Giesecke, A.V., Joung, J.K., and Wolfe, S.A. (2006). Counter-selectable marker for bacterial-based interaction trap systems. Biotechniques *40*, 179-184.

Moré, J. (1978). The Levenberg-Marquardt algorithm: Implementation and theory. (Berlin / Heidelberg, Springer).

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A., and Bulyk, M.L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat Genet *36*, 1331-1339.

Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008a). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell *133*, 1277-1289.

Noyes, M.B., Meng, X., Wakabayashi, A., Sinha, S., Brodsky, M.H., and Wolfe, S.A. (2008b). A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucleic Acids Res *36*, 2547-2560.

Oliphant, A.R., Brandl, C.J., and Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Mol Cell Biol *9*, 2944-2949.

Puckett, J.W., Muzikar, K.A., Tietjen, J., Warren, C.L., Ansari, A.Z., and Dervan, P.B. (2007). Quantitative microarray profiling of DNA-binding molecules. J Am Chem Soc *129*, 12310-12319.

Roulet, E., Busso, S., Camargo, A.A., Simpson, A.J., Mermod, N., and Bucher, P. (2002). High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. Nat Biotechnol *20*, 831-835.

Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*, 16-23.

Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. Nat Rev Genet *11*, 751-760.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science *249*, 505-510.

Warren, C.L., Kratochvil, N.C., Hauschild, K.E., Foister, S., Brezinski, M.L., Dervan, P.B., Phillips, G.N., Jr., and Ansari, A.Z. (2006). Defining the sequence-recognition profile of DNA-binding molecules. Proc Natl Acad Sci U S A *103*, 867-872.

Wright, W.E., Binder, M., and Funk, W. (1991). Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. Mol Cell Biol *11*, 4104-4110.

Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring binding energies from selected binding sites. PLoS Comput Biol *5*, e1000590.

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*, D111-117.

Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. Nucleic Acids Res *37*, e151.

Chapter 5: Protein-DNA Recognition Models for the C2H2 Zinc Finger and Homeodomain Transcription Factor Families[1]

[1] The work in this chapter has not yet been published. Our collaborators are completing additional B1H experiments, which will expand the zinc finger data set that I describe in this chapter. We anticipate that this data will be available by the end of August 2011, at which point the manuscript will be updated and submitted for publication. The analysis pipeline I have written is almost completely automated, so it will be a simple matter to perform the required analysis for the finalized C2H2 zinc finger data set.

I performed all of the computational work described in this chapter and wrote it with editorial assistance from Gary Stormo. Our collaborators in the Wolfe performed all the experimental work. Ankit Gupta, Marcus Noyes, and Metewo Enuameh in Scot Wolfe's lab at the University of Massachusetts performed all of the B1H analysis. Ankit performed all of the Zif268 CV-B1H selections and generated the designed Zif268 mutants. Marcus Noyes and Metewo Enuameh performed the homeodomain B1H selections. Marcus performed all of the initial homeodomain selections that were sequenced using the Illumina platform. Metewo performed additional follow up selections. Keith Joung's lab at Harvard performed all of the B2H experiments. All of the homeodomain data used in this study is publicly available. The Zinc finger data sets are not yet publicly available.

Abstract

It is long standing goal to a transcription factor's DNA-binding specificity directly from its protein sequence. While it was initially hoped that the protein-DNA recognition code would be simple, it is now apparent that interactions between transcription factors and DNA are more complicated. Different families of transcription factors utilize different folds to bind to DNA. Family based approaches to learn C2H2 Zinc Finger recognition models have met with some success. Here we describe random forest based recognition models for the C2H2 zinc finger and homeodomain transcription factor families and show that they perform better than all of the other methods tested.

Introduction

It is a long-standing goal to predict the DNA-binding specificity of a transcription factor (TF) based only on its protein sequence. That ability would allow for the inference of regulatory networks from genome sequences alone as well as for the design of TFs with desired binding sites. In 1976, before there were any known structures of protein-DNA complexes, Seeman et al. (Seeman et al., 1976) proposed a simple recognition code in which Arg would specifically bind to G-C base-pairs and Asn and Gln would specifically bind to A-

T base-pairs. But soon after the first few structures were determined using X-ray crystallography, the title of a famous Nature paper stated "Protein-DNA interaction. No code for recognition" (Matthews, 1988). However, it was clear that this referred to a simple, one-to-one deterministic code, in which there were only the types of interactions proposed by Seeman et al. As the structures of more DNA-protein complexes were determined it became apparent that there were definite preferences for interactions between particular amino acids and base-pairs, and that those preferences might vary depending on the class of TF. Zinc finger proteins, in particular, were heavily studied and a degenerate, qualitative recognition code was shown to be moderately successful at predicting their binding specificities (Choo and Klug, 1994a, b, 1997; Wolfe et al., 1999; Wolfe et al., 2000). This general idea of class-specific, degenerate recognition codes was further developed into a quantitative probabilistic code for zinc finger proteins that was more predictive than previous models although still much less accurate than one would like (Benos et al., 2001; Benos et al., 2002a, b; Kaplan et al., 2005).

This chapter pursues the goal of developing quantitative recognition models, focusing on the two most abundant TF families in mammalian genomes, the zinc finger and homeodomain families (Tupler et al., 2001). It takes advantage of new high-throughput methods for determining the specificity of individual TFs and the resulting large increase in the number of TFs from each family with well-characterized binding motifs (Stormo and

Zhao, 2010). It compares several different algorithms for predicting the specificities from protein sequences using 10-fold cross-validation assessment. The zinc finger models were also tested on additional data that were not included in any of the training sets. It demonstrates that a random forest machine learning method performs better than other methods tested and better than previously published methods.

Figure 5.1 Cartoon style image of PDB structure 1ZAA of Zif268-DNA complex produced using Jmol (Herraez, 2006).

C2H2 Zinc Finger Family

The C2H2 family of zinc finger TFs (Pfam name: zf-C2H2) constitutes the largest TF family in mammals and most other sequenced eukaryotic genomes (Tupler et al., 2001). Throughout the rest of this chapter we will refer to this family simply as ZF. The DNA binding domain consist of a compact set of about 30 amino acids which form a $\beta\beta\alpha$ motif that coordinates a zinc ion which helps stabilize the structure (Figure 5.1). Members of this family share a distinct $CX_nCX_nHX_nH$ motif (Figure 5.3). The two cysteine and histidine residues are critical for coordinating a zinc ion that allows the motif to maintain its structure. The alpha helix, or recognition helix, binds in the major groove where its amino acids bind to bases in the major groove and confer specificity. The recognition helix also makes multiple contacts with the backbone. Crystal structures indicate that there is nearly a one-to-one relationship between DNA contacting residues (key residues) in the well studied ZF protein Zif268 and its binding site. Figure 5.2 shows the set of key residues in Zif268 and the bases they contact. This is known as the canonical model of zinc finger binding. There are exceptions to this general mapping between key residues in the recognition helix and the binding site observed in some crystal structures for other proteins. See Pabo et al. for a review (Pabo et al., 2001).

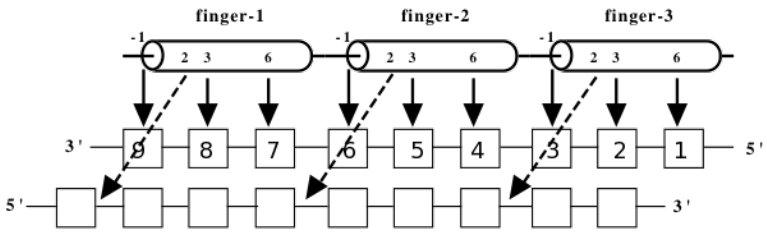


Figure 5.2 Canonical DNA binding model for Zif268 (Benos et al., 2001).

Recognition helix positions -1, +3, +6 contact bases at positions 3, 2, 1 respectively in each finger's sub site. The +2 position (overlapping base) binds to the complementary base at position 4. The key residues of each recognition helix are labeled. The bases in the DNA site are numbered from 5' to 3'.

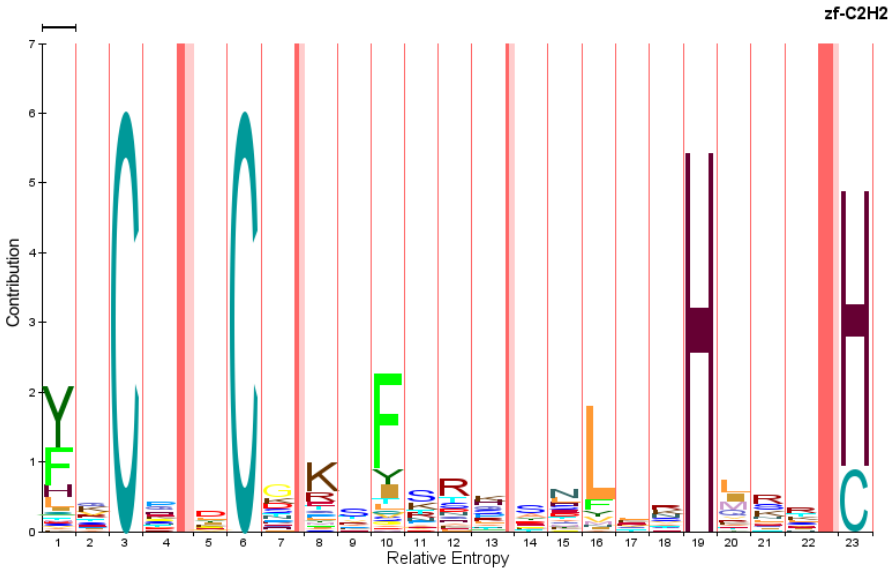


Figure 5.3 HMM logo for the ZF family from Pfam.

This domain follows a clear $CX_nCX_nHX_nH$ pattern, hence the name, ZF.

The ZF data set used for training the recognition models in this study consists of PWMs for over 250 different Zif268 mutants. The specificity of these mutants was determined using the CV-B1H assay (Christensen et al., 2011). Some of these mutants were selected by our collaborator Keith Joung's lab at Harvard to bind specific target sequences of the form GNNNNG*gcgg*. The first finger was fixed to the wild type sequence and portions of the recognition helix of fingers 2 and 3 were randomized. The consensus site for the wild type finger 1 is gcgg. Mutants were selected from a large randomized library using a bacterial two-hybrid assay (Joung et al., 2000; Serebriiskii et al., 2005). For the mutants selected by the B2H assay, only recognition helix positions +5 and +6 of finger 2 and -1, +1, +2 of finger 3 were randomized. The ZF data set also contains other mutants that were designed by our collaborator Scot Wolfe's lab at the University of Massachusetts. Existing B2H selected mutants were modified in an attempt to create variants with greater specificity for the desired target sites. Figure 5.4 shows that for the most part, the data set consists of proteins that were randomized at positions 5 and 6 of finger 2 and positions -1 to 3 of finger 3. Position 4 of fingers 2 and 3 was the only position that was never varied.

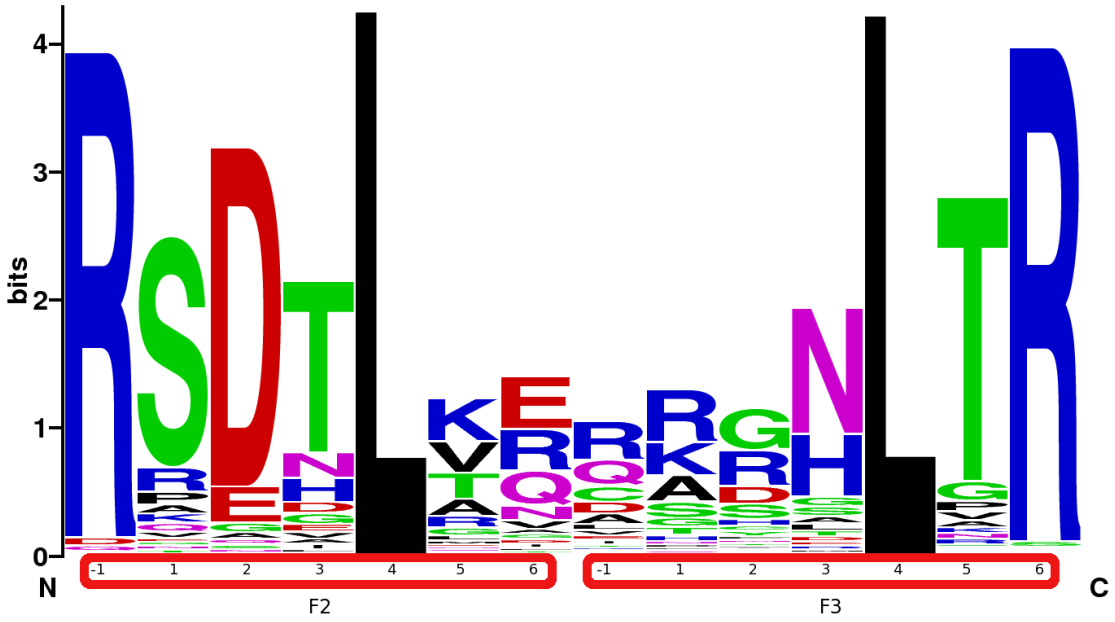


Figure 5.4 Sequence logo showing the variability at each position of the Zif268 mutants in the data set.

Homeodomain Data Set

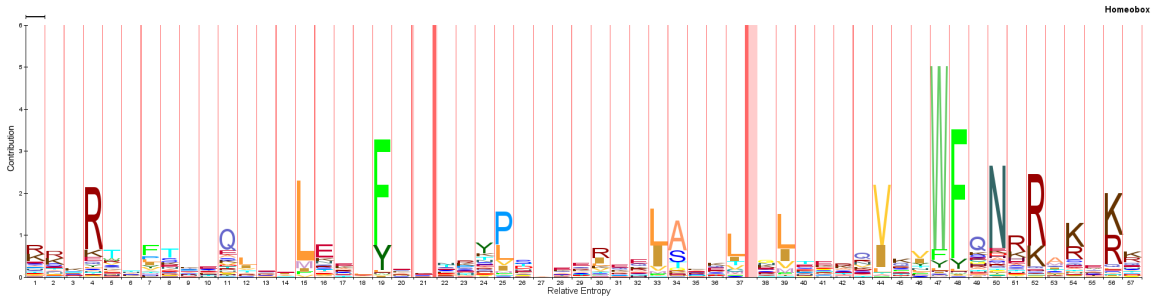


Figure 5.5 HMM logo for the Pfam homeobox family.

Notice that the first position of this logo corresponds to the second position in our MAFFT alignment

The homeodomain TF family (Pfam name: homeobox) (Figure 5.5) is the second largest TF family in humans and in most metazoans (Tupler et al., 2001). This DBD motif was first discovered in Drosophila. TFs possessing a homeodomain DBD generally have only a single DBD. The domain consists of about 60 residues that fold into a stable bundle of three alpha helices. The C-terminal helix, or recognition helix, binds in the major groove. There is an unstructured N-terminal arm that occurs at the very beginning of the domain. This N-terminal arm binds in the minor groove. See chapter 2 for more details about the homeodomain family. Hereafter we will refer to this family as 'HD.'

In Noyes et al., we described a web application available at http://ural.wustl.edu/flyhd called hd_pred that used all of the Drosophila B1H data available at that time to make DNA specificity predictions along with estimates of the quality of those predictions (Noyes et al., 2008). To the best of our knowledge, this was the first recognition model for the homeodomain family that predicted position specific frequency matrices (PFMs) based on the protein sequence of the query proteins. The hd_pred model was based on the observations that (1) HD factors with similar key residues tend to bind very similar motifs and (2) this tendency holds true as long as the overall similarity is

sufficiently high between two proteins. This model is similar to the k nearest neighbors method (KNN), except that we considered both the overall similarity as measured by the BLOSSUM similarity score as well as exact identity at a few selected key residue positions (residues 5, 47, 50, 51, 54, 55, indexed relative to the Drosophila engrailed protein).

Shortly after the hd_pred model was publish, an analysis paper was published claiming that of all the machine learning methods they tried (nearest neighbor, SVM, random forests, principle component regression) the simplest model, nearest neighbor, worked essentially as well as any of the more sophisticated methods (Alleyne et al., 2009). Alleyne et al. did not predict PWMs as our hd_pred method does. Instead, they predicted the PBM Z-scores for all 32,896 non-redundant 8mer probe intensities based essentially on the Hamming distance between the protein sequence of the query and reference proteins. Machine learning algorithms like SVM and random forests can take multiple dependent variables as input, but they can predict only a single dependent variable. For this reason, Alleyne learned 32,896 individual SVM and RF models, one for every non-redundant 8mer. Predicting PBM Z-scores precluded the inclusion of the Drosophila B1H HD data set (Noyes et al., 2008) in their analysis. They decided to predict 8mer probe intensities partly because they believed that many HD motifs could not reliably be aligned (Alleyne et al., 2009; Berger et al., 2006), and partly because they did not believe that PWMs are capable of adequately modeling PBM data.

We decided to revisit the HD recognition model problem using all of the available HD B1H and PBM data. We combined all of the B1H models for fly proteins with all of the PWM models for the mouse, yeast and worm proteins characterized using PBM. Since our previous study, all of the B1H experiments have been repeated for the wild type fly proteins and the selected sites were sequenced using Illumina sequencing in a multiplexed sequencing run. This yielded many more selected sites than were obtained in the previous study. We did not use the motifs produced by the Bulyk group's seed-and-wobble algorithm (Berger et al., 2006), which are available in JASPAR (Portales-Casamar et al., 2010) and in the online database UniPROBE (Newburger and Bulyk, 2008). Instead a new set of PWMs generated by an algorithm called BEEML-PBM (Zhao and Stormo, 2011) were utilized.

These newly available BEEML-PBM PWMs have been shown (Zhao and Stormo, 2011) to explain the original PBM data significantly better than the PWMs produced by the seed-and-wobble algorithm. Furthermore, Zhao and Stormo showed that PWMs are able in most every case examined to model the PBM data well.

Looking at the B1H and BEEM-PBM PWMs, it became clear that it should be possible to align the mouse HD motifs, contrary to the conclusions drawn by Alleyne et al. and Berger et al. (Alleyne et al., 2009; Berger et al., 2008). Some of the UniPROBE HD PFMs looked like they contained two half sites and not a single monomeric motif. The

BEEML-PBM motif set did not contain any motifs that looked like they had this issue. This new motif set made it clear that motif alignment ought to be possible.

Non-Additivity in Recognition Models

Position weight matrix (PWM) models assume that binding to one position of a TF site is independent of binding to any other position of that site (Stormo, 2000). Because the energy of binding to each position of the motif is assumed to be independent, these energies can be added to obtain the overall relative energy of binding. The extent to which additivity is a valid approximation for protein-DNA recognition models is still an open question. Many recognition models exist which make the simplifying assumption that all interactions are independent. These models make the same independence assumption that PWMs make, which we call the assumption of 'independence between bases'. They also assume that the energy of an amino acid, A, at position i in the protein binding to base b at position j in the motif is not affected by the identity of the other amino acids in the protein. In other words, it is assumed that amino acids at other key residue positions can be mutated with out affecting the binding at other positions. We call this assumption 'independence between residues.' These assumptions of independence between bases and between amino acids are almost certainly not true, but the question is one of degrees. To paraphrase George Box,

"all models are wrong, but some models are useful." While these assumptions of additivity are oversimplifications, the questions is how big of a sacrifice we need to make in accuracy in order to have a model that is as simple as possible. It would be ideal if a simple, easily interpretable model was all that was needed. We show however, that while simple recognition models like SAMIE perform reasonably well, more complicated models that do not make strict assumptions of independence have better performance.

Existing Family Recognition Models

SAMIE

SAMIE consists of several different matrices. The first matrix is a contact matrix that defines the mapping between positions in the recognition helix and positions in the triplet DNA motif. SAMIE models the interaction between each key residue and the base it binds as being independent from the other interactions. It also assumes that the identity of neighboring bases does not affect the energy of binding. Furthermore, it assumes that each finger binds independently of the other fingers. In addition to the contact matrix, SAMIE has three 20x4 energy matrices that model the energy of each key residue.

SAMIE ignores the contact made by the +2 position of the adjacent finger on the opposite strand. SAMIE was parameterized using the selected proteins or sites from phage display experiments and sites selected using SELEX taken from the literature.

The energy of each AA-DNA contact is modeled by a 20x4 energy matrix, E_{ij}, where i and j indicate the indexes of the positions. Once it has been populated, this matrix is used as a look-up table where each element indicates the energy of residue A_i to base B_j, where i indicates the position in the recognition helix and j indicates the position in the motif.

KFM05

There are several other matrix based ZF models in the literature. One of these, KFM05, has a contact matrix and four 20x4 AA-DNA energy matrices to model the partial binding energies contributions of positions -1,+3, +6 and also +2 of the adjacent C-terminal finger (Kaplan et al., 2005). Note that refer to all of the MBMs besides SAMIE and those learned in this study by their author's first initials and the last 2 digits of the year the model was published, after Persikov et al. (Persikov et al., 2009). KFM05 was parameterized using unaligned short sequences from TRANSFAC (Wingender et al., 2000). An EM algorithm was employed to both find the staring sites of all the binding sites and to parameterize the 4 energy models and an additional parameter, α, which weights the contribution of the +2 and +6 matrices, since both of the residues together determine the specificity of the first position in a single finger triplet motif (Wolfe et al., 2000).

ZifNet

We also tested a neural network based model called ZifNet which was trained on the same data set used to train SAMIE (Liu and Stormo, 2008). The goal of ZifNet was to take into account the amino acid context of the key residues -1, +3, +6. It takes as input only these three residues and ignores the adjacent C-terminal finger +2 overlap position. It outputs a vector of scores for all 64 DNA 3mers and converts this vector to a PFM.

Single Matrix Models

Other simpler models have been reported to perform fairly well at predicting ZF motifs. These models consist of a single 20x4 energy matrix that is used to model the energy of all residue (20 rows) and base (4 columns) binding contacts. A contact matrix can then be constructed for a particular structural family of transcription factors and used with this energy matrix. One such model, SBGY95, was constructed by Suzuki et al. after looking at crystal structures of transcription factors from different families bound to DNA (Suzuki et al., 1995). They assigned unitless, semi-arbitrary 'chemical merit points' to different possible nucleotide and amino acid binding interactions. We find that these 'chemical merit points' are roughly proportional to energy values, once they are scaled appropriately, and that this pseudo-quantitative model actually performs relatively well. We also test another single matrix model, MGM98, that was parameterized by looking at crystal structures containing protein-DNA complexes. The authors took the frequency of observed interactions and a

theoretical background model and generated a log-odds energy matrix (Mandel-Gutfreund and Margalit, 1998). Interestingly, the theoretical SBGY95 matrix performs better than the empirically parameterized MGM98 model (see results).

Methods

Analysis of ZF B1H Data

GRaMS$_c$

Because the CV-B1H version of the B1H assay was employed (see chapter 4), the alignment of all 6bp zinc finger motifs was trivial. For every protein but one, all of the selected sites were already aligned. In this one exception, some of the selected sites seemed to be shifted by 1 base pair relative to the majority of sites and relative to the expected motif. Thus, we did not need to use our multiple PFM alignment method, since the sites were already aligned due to the fixed finger 1 subsite adjacent to the 6bp randomized region. This enabled us to use GRaMS to learn PWMs from the ZF data sets.

In the original implementation of GRaMS, nonlinear regression was employed to parameterize a model consisting of a PWM and a parameter, μ, which describes the degree

of saturation due to the free concentration of the TF (see chapter 4). Many of the Zif268 mutants appear to be more specific then wild type Zif268 for their target sequences. This is not surprising since the mutant proteins in our collection were selected or designed to bind as specifically as possible to their target sequences. However, in order to get properly scaled PWMs, (i.e. PWMs with the correct information content), we find that it is critical to estimate μ accurately. Otherwise, when there are only a handful (i.e. less than 20) sites that are significantly enriched, the original version of GRaMS has a difficult time parameterizing μ correctly. This is because μ is equal to the $\Delta\Delta G$ of those sites that are bound 50% of the time. Those sites that are half saturated or are nearly half saturated occur in the linear region of the sigmoid energy VS occupancy curve. If none of the observed sites fall in this linear region on the 'true' energy versus occupancy curve (figure 4.3), then it is difficult or impossible to determine μ accurately from only a few saturated sites. We found in practice that for very specific proteins, it worked best to re-arrange the GRaMS objective function equation and fit directly to the observed counts per 6mer rather than to the relative growth rate of each 6mer allele. Otherwise, when there were very few appreciably enriched sites, the tendency was to over fit to the noise when the growth rates were fit directly. We found it also helped to set M, the maximum growth at which 100% saturation occurs to 1.02 the observed maximal growth rate. In the original version, M was set to the maximum observed relative growth rate. We call this version of the program $GRaMS_c$ and we used it to analyze

all of the B1H Zif268 mutant data. Figure 5.6 shows that our ZF training set contained a great deal of diversity in the set of motifs at positions 2 through 5. Positions 1 and 6 of the motifs were largely invariant because the key residues known to bind these positions were not varied for the most part.

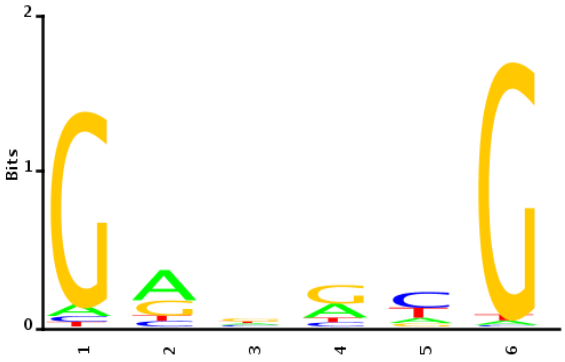


Figure 5.6 Average PFM for the mutant Zif268 data set produced by averaging all 251 GRaMS$_c$ PWMs, after converting them to PFMs.

Analysis of HD Data Set

Protein Alignment

Only the HD proteins needed to be aligned, since the ZF proteins contained no insertions or deletions, only point mutations. The hmmsearch program from the HMMER suite (Eddy, 2001; Eddy, 1998) was used to extract the DBD from every protein fragment

used for either the PBM or B1H analysis using the homeobox Pfam hmm model (Finn et al., 2010). All wild type protein sequences were obtained from the UniPROBE (Newburger and Bulyk, 2008) or FlyFactorSurvey (Zhu et al., 2011). Table 5.1 shows the source of all of the HD motifs and proteins used in this study.

Table 5.1 Source of HD motif matrices

Source	Species	Number
PBM	worm	1
PBM	yeast	4
B1H, Sanger	human	8
B1H, Sanger	fly directed mutants	13
B1H, SOLEXA	fly	84
PBM	mouse	154

The protein multiple alignment program MAFFT (Katoh et al., 2005), was used to align each set of extracted DBDs. We found that using MAFFT rather than other alignment methods such as clustalX or hmmalign from the HMMER suite worked best. The N-terminal domain in particular is difficult for many alignment programs to align correctly. Other programs tended to insert too many gaps in this region. We included one flanking residue on either side of the homeobox domain matches returned by hmmsearch, so the first position in the homeobox hmm logo corresponds to the second position in our alignment.

One of the best studied HD proteins is engrailed (Ades and Sauer, 1994; Clarke et al., 1994; Draganescu and Tullius, 1998; Fraenkel et al., 1998; Grant et al., 2000; Kissinger et al.,

1990; Liu et al., 1990; Sato et al., 2004; Tucker-Kellogg et al., 1997). The engrailed homologs in moue (en1 and en2) and fly (en) are all very similar. For consistency with previous studies that generally number residues with respect to engrailed, all columns in the multiple sequence alignment that contained insertions relative to engrailed were removed. Only a small minority of the proteins in the data set had insertions relative to engrailed. There were three main insertions relative to the engrailed proteins. One insertion of three residues that occurred between engrailed positions 21 and 22 was present in 26 out of 264 proteins. These 26 proteins were member of the atypical subgroup and are known to have this small insertion. The two other insertions only occurred in the mouse proteins Hdx, Hmbox1, Tcf1, and Tcf2.

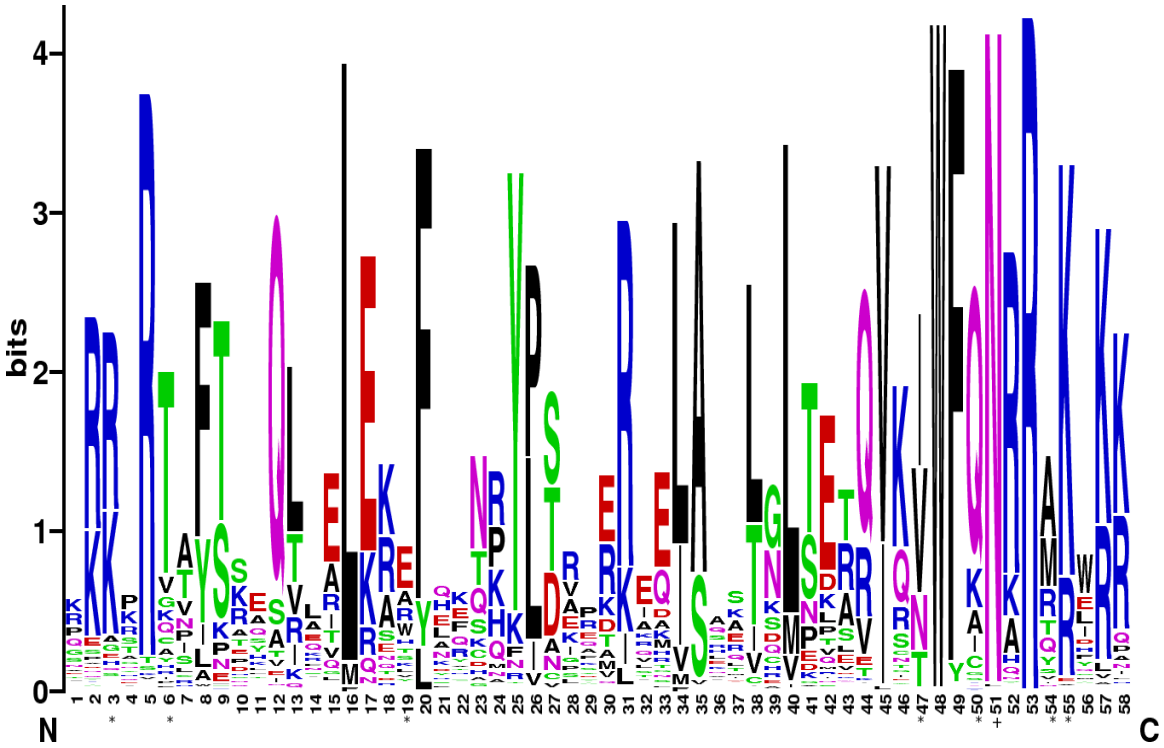


Figure 5.7 Sequence logo for the HD multiple sequence alignment used for training the recognition models.

The '*' characters denote positions found by our feature selection method to be important. The '+' symbol by position 51 indicates that it is a critical residue. Most HD proteins have Asn51, which binds Adenine with high specificity. Those that do not, such as Lag1, tend to have very divergent motifs

Scaling BEEML-PBM Motifs

We noticed that the HD B1H position specific count matrices (PCMs) from Fly Factor Survey database (Zhu et al., 2011) and the PWMs from the BEEML-PBM set were

scaled differently when both were converted to PFMs. Although it was not clear which set was scaled correctly, they needed to be scaled similarly in order to combine both data sets for recognition model construction. A set of close mouse and fly homologs was assembled for comparison. We took this set of B1H and BEEML-PBM motif pairs and aligned each pair of motifs. All of the B1H PCMs were first converted to PFMs using a small pseudo count of 1e-10. The BEEML-PBM PWMs were also converted to PFMs. The elements of the PWMs were exponentiated and then each column of the resulting matrix was normalized to sum to one using the following formula:

$P_{i,b} = \frac{e^{-W_{i,b}}}{\sum_i e^{-W_{i,b}}}$ where W is the scaled PWM matrix, i is the index over the number of positions in the motif, and b is an index over the number of rows in the PWM, one for each of the four bases A, C, G, T. Initially, a factor of 2.5 was used to scale the BEEML-PBM PWMs before converting them to PFMs based on the observation that for a few mouse and fly homolog pairs, the scaling factor looked to be about 2.5. Each pair of fly and mouse PFMs was aligned using our MATLAB implementation of the SSD metric and the ungapped Smith-Waterman motif core alignment technique described by Mahony et al. (Mahony et al., 2007a). The best scaling factor was then found for each pair of aligned fly and mouse motifs by performing a linear fit between the models. A single global scaling factor of 2.2381 was obtained by averaging all of the scale factors together. This value was then used to scale all of the BEEML-PBM PWMs before converting them to PFMs.

Low Information Content PWMs Removed

A few of the BEEML-PBM PWMs had very low information content, even after being scaled. Very low information content was potentially due to a failed PBM experiments or a non-specific protein. BEEML-PBM PWMs with information contents lower than three standard deviations below the mean information content were removed. This resulted in seven PWMs being removed (Dbx1, Hoxb6, Hoxc6, Obox2, Pax6, Six6, Tlx2).

Multiple Motif Alignment

Mahoney et al. developed a multiple PFM alignment program called STAMP (Mahony et al., 2007a, b) which they demonstrated was able to produce motif alignments which it then used to determine some of the key residues via mutual information analysis for various TF families with some success. It was also shown to be an accurate motif database-searching tool. We implemented a similar multiple alignment program in MATLAB. We wanted to test additional scoring metrics and methods of guide tree construction, and we found that it was simpler to do this in a framework like MATLAB, which has a bioinformatics toolbox with many built in functions. STAMP, on the other hand, is written in C++ and is very fast, but the process for adding new motif column-column scoring

metrics is somewhat involved, since new empirical distributions must also be calculated in order to obtain p-values for the each new distance metric, which are necessary in STAMP for tree building. After developing metrics which incorporated the per column information content, we found that the best metric was SSD (sum of squared differences), also one of the two best metrics reported by Mahony et al. We also found that ungapped local alignments yielded the best alignments and that it was important to perform motif core alignment, as outlined by Mahoney et al. when aligning the relatively short HD motifs to generate a multiple motif alignment (MMA). Motif cores are define as consecutive positions in a motif having an information content above 0.3 bits, or if there were not at least 4 consecutive columns with IC > 0.3 bits, then the four consecutive columns with the highest total IC were used. If the motif was less than 4 base pairs long, then the entire motif was designated as the motif core.

In the end, the main difference between our progressive multiple motif alignment program and STAMP lies in how we calculated our distance matrix used to produce the guide tree. STAMP uses empirically derived p-values that are assigned to all alignments to generate the guide tree and to do database search analysis. It converts a pairwise matrix of p-values obtained for each alignment to a distance matrix and then generates a guide tree from these distance matrices. At least in this case, the result was not very different, but it does show that a simpler method for calculating the distance matrix for the guide trees was

sufficient. We did find that adjusting the mutual information scores tended to yield better results than calculating the mutual information between positions in the motif and protein multiple alignments as implemented in STAMP however.

Low Mean Information Content Positions Removed from MMA

Not all columns of the multiple motif alignment had similar information content (IC). Some of the flanking positions in the original multiple motif alignment consisted mostly of gaps, which had an IC of zero. Rather than attempt to model these positions that were not relevant for most motifs, we trimmed flanking positions from the MMA that had a mean IC of less than 0.05. The final trimmed multiple motif alignment consisted of 9 positions. Figure 5.8 shows the average PFM for the HD data set. All the PFMs in the final trimmed MMA were averaged together. The 'A' at position 9 in the average HD PFM is almost completely conserved. This is because almost all of the proteins in the data set have Asn at position 51, which binds strongly to adenine (see Figure 5.7).

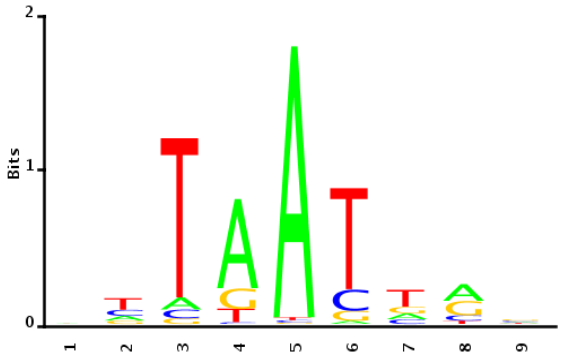


Figure 5.8 Average PFM for the trimmed HD multiple motif alignment

Feature Selection using Adjusted Mutual Information and Profile Alphabet

Machine learning methods like SVMs and random forests are relatively good at capturing both the non-linear and linear relationships between independent variables that give rise to the observed variable. In our case the independent variables are the dummy encoded amino acids in the recognition helices of each protein. The dependent variables are the elements of the energy matrix for each protein, 18 and 27 parameters for the ZF and HD data sets respectively. The SVM and RF methods are good at performing feature selection. However, we found that using an adjusted form of mutual information (MIp) to filter out the least informative features first resulted in even more accurate models.

The mutual information (MI) between two continuous variables is a measure of the degree of independence between those two variables. If the variables are completely independent, the MI will be zero. If the variables are completely correlated and the two categorical variables each have some degree of variability, then the MI will be maximal. Mutual information requires that the two variables being compared be discrete variables. In order to make each PFM matrix discrete, we used a profile alphabet (Wang and Stormo, 2005). Nineteen different multinomial probability distributions over the four bases A, C, G, and T were defined and assigned to labels in an alphabet. This allowed us to convert each column of a PFM to a single label. We began with an initial set of seed profiles chosen to cover the space of possible PFM columns sand then iteratively refined those profiles. The Euclidean distance between every column from every PFM in the multiple motif alignment (MMA) and all 19 profile vectors was calculated. Each PFM column was assigned to the nearest profile. The mean of the set of column vectors assigned to each profile then became the new profile vector. This process was repeated until convergence.

MIp

Adjusted mutual information (MIp) was used to sort the positions in the DBDs by their ability to determine the most predictive features. The adjustment to the mutual information proposed by Dunn et al. is based on the idea that the average entropy of each position in an alignments gives each position a particular propensity towards MI_b, or

background mutual information. Dunn et al. come up with a simple correction to the MI which they call the average product correction (APC). MIp(a,b) = MI(a,b) + APC(a,b). In other words, on average, the MI between positions a and b is due to the variability at those two positions and not due to direct interactions between these two positions. Subtracting APC(a,b) from MI(a,b) should yield the amount of MI due to functional interactions between positions a and b. A truly random data set of sufficient size should not need to be adjusted. The HD data set is much less random than the ZF data set, but even this data set is not completely random.

Every possible protein and motif position pair was ranked according to its MIp score. The maximum MIp score for each position in the protein alignment was calculated. The set of protein positions, or potential features, was then sorted according to max(MIp) resulting in a sorted set of features. The RF, SVM and KNN methods (described below) were each used to train a set of models, one model per element of the PFM. WYK encoding (Stormo, 2011) was used so that only free parameters were predicted by the models, since each position of a PFM only has three free parameters due to the constraint that each PFM column sums to one.

To avoid imposing an arbitrary MIp score cut off to determine the feature set to use for training the recognition models, the sorted set of features was used to construct progressively larger training sets. The first feature set contained just the single feature with

the highest MIp score. The last feature set in the series contained all of the residues in the DBD. Figure 5.10 and Figure 5.14 shows the mean squared error (MSE) for each model obtained by using progressively more features from the list of MIp ranked protein positions for the HD and ZF set, respectively. 10-fold cross validation was used to determine the MSE in every case. There is a clear plateau in the MSE scores for all three methods that indicates that adding features beyond a certain point did not appreciably increase the performance of the model.

STAMP on the other hand does not assign one single label to each PFM column. For each PFM column and protein residue pair, it assigns a total of one count to the appropriate joint and marginal count matrices, but it splits that count fractionally between 4 different discrete categories A,C,G, and T. This is equivalent to using the PFM to assign frequencies to all possible sites of length L, where L is the number of positions in the PFM. Then, calculating the MI using a multiple alignment of all sites for all proteins, where each protein is paired once with all 4^L sites. We believe that one of the reasons that our method of discretizing the PFMs improves performance is that it acts as a form of normalization and filters out noise. It tends to assign the same labels to PFM columns that are slightly different due to noise or differences in scaling due perhaps to differences in the B1H and PBM assays. PFM columns that are not equal may in reality be instances of the same multivariate distribution. In practice, using the profile alphabet to make the PFM columns discrete

helped us to detect the true key residue set more accurately for the HD data set and it did not hurt our accuracy in the zinc finger case.

Machine Learning Algorithms

K Nearest Neighbors

The k nearest neighbors (KNN) algorithm is a very simple method based on the principle that similar inputs generally yield similar outputs. The Euclidean distance between every dummy encoded pair of proteins was calculated. This is functionally equivalent to computing the Hamming distance between the proteins. For every query protein, the closest protein in the training set served as the reference protein. The reference protein's PFM was then used as the prediction for the query protein. If the user defined K parameter is set to 1, only the closest protein is included in the reference set. If it is set to 2, then the second closest protein is also included, etc. In the case of a tie or multiple reference proteins, all of the corresponding PFMs were averaged together. We used the knnflex R package for the KNN analysis. Surprisingly, in preliminary studies involving the HD data set, we found that tuning the K parameter did not increase performance, so we fixed the K parameter to 1 for all subsequent analysis.

Random Forest Regression

Random Forest (RF) is a fairly recent ensemble method that makes use of a collection of weak decision trees (Breiman, 2001). There are only two user specified parameters: the total number of trees in the ensemble (ntree) and the number of randomly selected features to use to parameterize every node in of each tree (mtry), so it is simple to tune. In practice, it works well even without any tuning, although we did tune the mtry parameter for every RF model. The R package that we used, randomForest, had its own tuning function, but we developed our own method. We iteratively tried increasing values of mtry, from 2 to 50. For all of these iterations, we set ntree to 50 for increased speed. Next, the mtry versus MSE curve was smoothed and the mtry parameter the yielded the best MSE value was determined. Then, we used the optimal mtry parameter, and we set ntree to 500, which is the default value.

State Vector Machine Regression

SVMs are a popular machine learning binary classifier that has been adapted to perform regression. SVMs use a kernel function to map a set of training vectors into a higher dimensional space. They find the linear separating hyperplane that maximizes the margin of separation in this higher dimensional space. A non-negative cost parameter, C, is set by the user and determines the weight of the error term in the minimization (Hsu et al.,

2010). The e1071 R package was used to train and tune the SVM models. This packages is based on the libsvm program (Chang and Lin, 2011). Of the available kernel functions, the authors of libsvm recommend the radial basis function as the simplest to tune and the most generally applicable. The radial basis function includes a gamma parameter, which we tuned along with the cost parameter using the grid search function implemented in the e1071 package. We searched over the range $2^{-15} \leq \gamma \leq 2^{3}$ and $2^{-5} \leq C \leq 2^{3}$ using a step size of 2^{2}. In preliminary studies involving the HD data set, we tried using all of the different kernel functions available in libsvm. However, we found that the radial basis function performed better or as well as the other kernel functions. Also, it was much less expensive to tune in most cases.

We also tried using R packages to construct neural networks and perform partial least squares regression. We found that these methods did not perform well compared to the random forest and SVM methods, so we did not use them in this study.

Normalization of Predicted PFMs

There is no guarantee that the columns of the raw predicted PFMs will sum to 1, although in practice the column total is very close to 1, once the predicted PFMs are converted from WYK to ACGT space. Sometimes the predicted frequency for a given base

is slightly negative. Since negative frequencies are not possible, all negative values were replaced with a very small pseudo count of 1e-10, and then all predicted PFMs were normalized.

Single Finger Models

The Zif268 mutant data consist of two finger motifs that are 6bp in length. In the ZF finger data set, key residue position -1 of finger 2 is almost always Arg and position +6 of finger 3 is almost invariably Arg as well. In order to parameterize a more broadly applicable single finger model, we combined specificity data for finger 2 and finger 3. We split up the 6bp motifs into finger 2 and 3 submotifs and paired them with the appropriate recognition helix protein sequence. In order to allow the models to capture the context effect that causes a finger to give rise to a different triplet motif depending on whether that finger occurs as the first, second or third finger (Liu et al., 2002), the sequence of the adjacent two fingers was included. In the case of finger 3, we encoded the fact that there is no adjacent C-terminal finger (ACF) by adding a series of seven gaps, one for each position in the recognition helix to the training data. Each of these gaps was dummy encoded to form a 20 long vector populated completely by zeros.

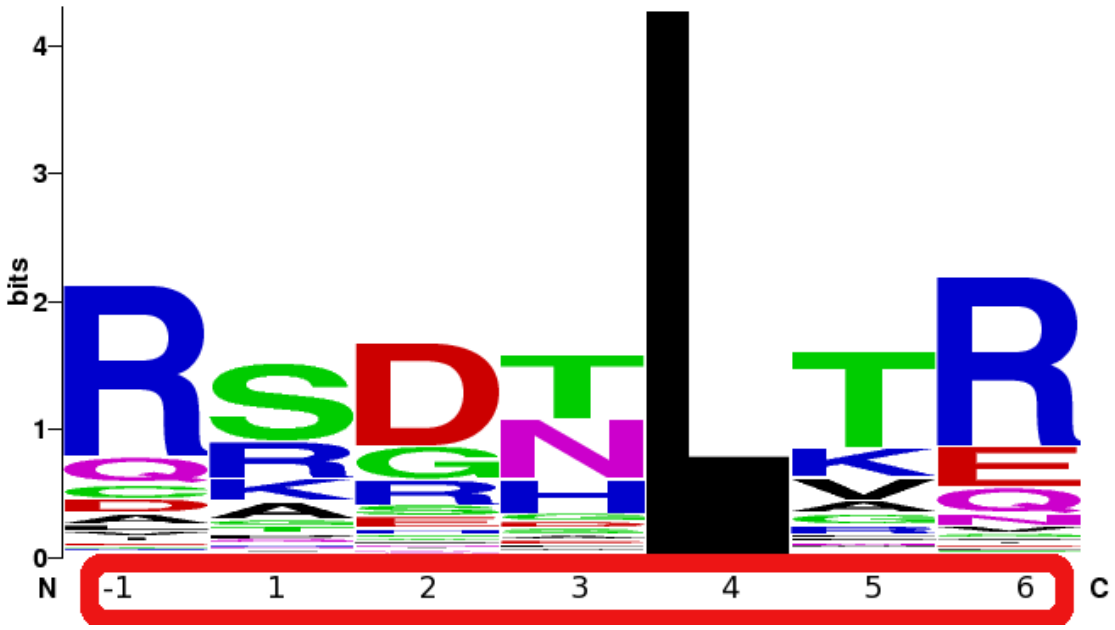


Figure 5.9 Sequence logo illustrating the diversity present in the combined finger 2 and 3 data set used to train the single finger models

To investigate the importance of interactions between amino acid residues in the recognition helix, we parameterized several matrix based models (MBMs). These models are similar to the SAMIE model (Benos et al., 2001). We developed four different types of MBMs, each using a different mapping or contact matrix, C. MBM1 has four different energy matrices, three for positions -1, +3, +6 of F_n and one for position +2 of F_{n+1}. Table 5.2 shows the contact matrix used for the MBM1 model. There is also a weighting parameter, alpha, which scales the contribution of the F_n:+6 and F_{n+1}:+2 matrices to position 1 of the triplet motif. (1-alpha) is used to weight F_n:+6 and alpha is used to weight the

F_{n+1}:+2 matrix. MBM2 is similar in design to MBM1, except that there is no F_(n+1):+2 matrix or alpha parameter.

MBM3 only has a single energy matrix that is used to model the amino acid and nucleotide interactions between positions -1, +3 and +6 in the recognition helix and positions 3, 2, and 1in the motif respectively. MBM4 also has only a single energy matrix, but it also uses its single matrix to model the ACF+2 position.

The reverse complement is taken to predict the F_{n+1}:+2 contribution, since F_{n+1}:+2 binds to the opposite strand.

Table 5.2 The contact matrix used by MBM1. Columns indicate the position in the recognition helix. Rows correspond to positions in the Nth fingers subsite.

	FN							F(N+1)						
	-1	1	2	3	4	5	6	-1	1	2	3	4	5	6
1	0	0	0	0	0	0	**1**	0	0	**1**	0	0	0	0
2	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0
3	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0

GNN Test Data

We used a series of 51 different single finger, 3bp long PCMs reported by Liu et al. to test our single finger ZF models in addition to the 10-fold cross-validation test (Liu et al.,

2002). The goal of Liu et al. was to design SP1 mutants to bind specifically to all 16 GNN triplets. SP1 is a three finger human ZF protein similar in structure to Zif268 (Berg, 1992). They used Sangamo Bioscience Inc.'s proprietary database of zinc fingers. They designed SP1 mutants to be specific for all 16 possible GNN sites. To investigate a possible finger context effect, they repeated this process for all three different finger positions for a total of 48 designed proteins, plus three additional mutants. Each protein was incubated with a library of randomized oligomers and a gel mobility shift assay was used iteratively to purify high affinity binding sites. The selected sites were then aligned to the consensus site and used to generate PCMs.

Results

HD Results

Our MIp analysis was able to successfully rank the most important key residues with only one false positive in the case of the HD data set. Position 19 had a high MIp score but to our knowledge, it has never been observed to contact DNA specifically. All of the other positions found by our MIp based feature selection analysis found key residues do interact with the bases directly in at least one structure. While other groups have used overlapping yet different sets of key residues, we find that including additional features beyond the set

selected using MIp actually hurt performance for the SVM and KNN models, and only helped performance slightly for RF.

Table 5.3 shows the set of MIp ranked features obtained from taking the maximum MIp score per position of the HD protein alignment (Figure 5.12). Those features that improved the MSE appreciably for all methods are highlighted in yellow. Only the top 30 features were considered. Figure 5.10 compares the 10-fold cross-validation performance of the KNN, SVM, and RF methods on the HD data set. The x-axis shows that after the seventh MIp ranked feature (Table 5.3) was included in the training set, adding additional parameters did not improve performance.

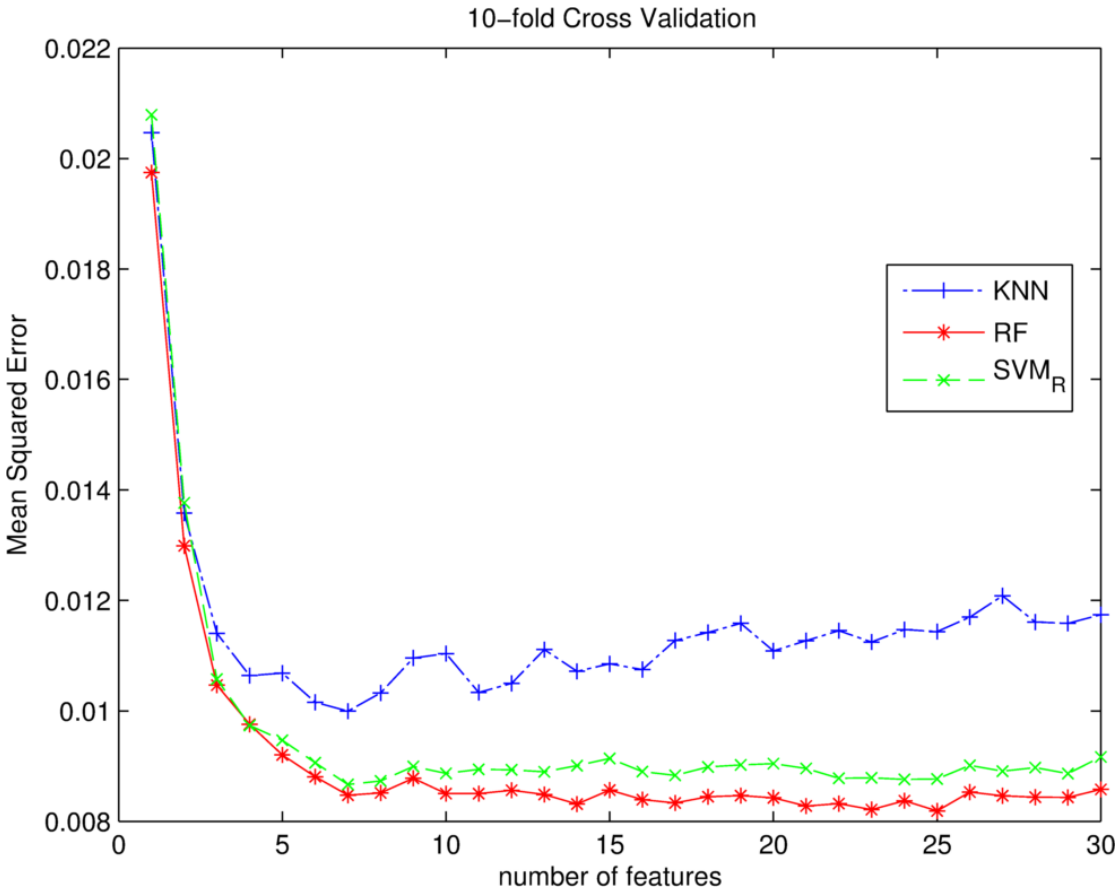


Figure 5.10 Plot of the number of features used to train the KNN, RF and SVM models versus the 10-fold cross validation MSE values.

After seven features are included the MSE stoped decreasing for KNN and SVM and did not decrease much for RF. Only the top 30 features were considered

Table 5.3 MIp Ranked HD Features

Features	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Residue Number	50	54	6	47	*19*	3	55	41	33	17	32	58	10	44	20

Features	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Residue Number	52	46	38	26	43	2	8	56	25	4	18	7	24	30	11

For the KNN and SVM methods, performance tended to decrease. The performance of the RF model did increase slightly in general as more features were added. Apparently the RF method does a slightly better job of feature selection here than even the SVM method. Based on the literature (Ekker et al., 1994; Kissinger et al., 1990; Noyes et al., 2008), all seven of these residues are thought to be important key residues in at least some contexts, except for position 19, which has not been reported to be a key residue. In chapter 2, based on our own findings and on a summary of the literatures, we report the following key residue set: 2, 3, 5, 6, 7, 43, 47, 50, 51, 54, 55. Alleyne et al. reviewed the literature and reported the following two key residues sets: 3, 5, 6, 25, 31, 44, 46, 47, 48, 50, 51, 53, 54, 55, 57 (Kissinger et al., 1990) and 3, 6, 7, 47, 50, 54 (Ekker et al., 1994).

To investigate how much our MIp based method for ranking features helped, we compared it to the MI analysis method implement in STAMP. The top seven residues with

the highest maximal STAMP MI values are, in order: 50, 54, 6, 19, 46, 56, 28 (Figure 5.11). Positions 46, 56 and 28 are not thought to be important for determining specificity. STAMP also finds residue 19, which was the one potential false positive found by our MIp based method. These, in order or rank, are the top ten features reported by Alleyne et al. (Alleyne et al., 2009): 50, 6, 46, 54, 7, 56, 14, 28, 4, 19. It is also interesting to note that there is a lot of overlap between the feature set reported by Alleyne et al. and the STAMP features set. It could be that some of the these positions have high entropy in the protein alignment and that this is why the MI is so high, not because of correlation due to coevolution (Dunn et al., 2008).

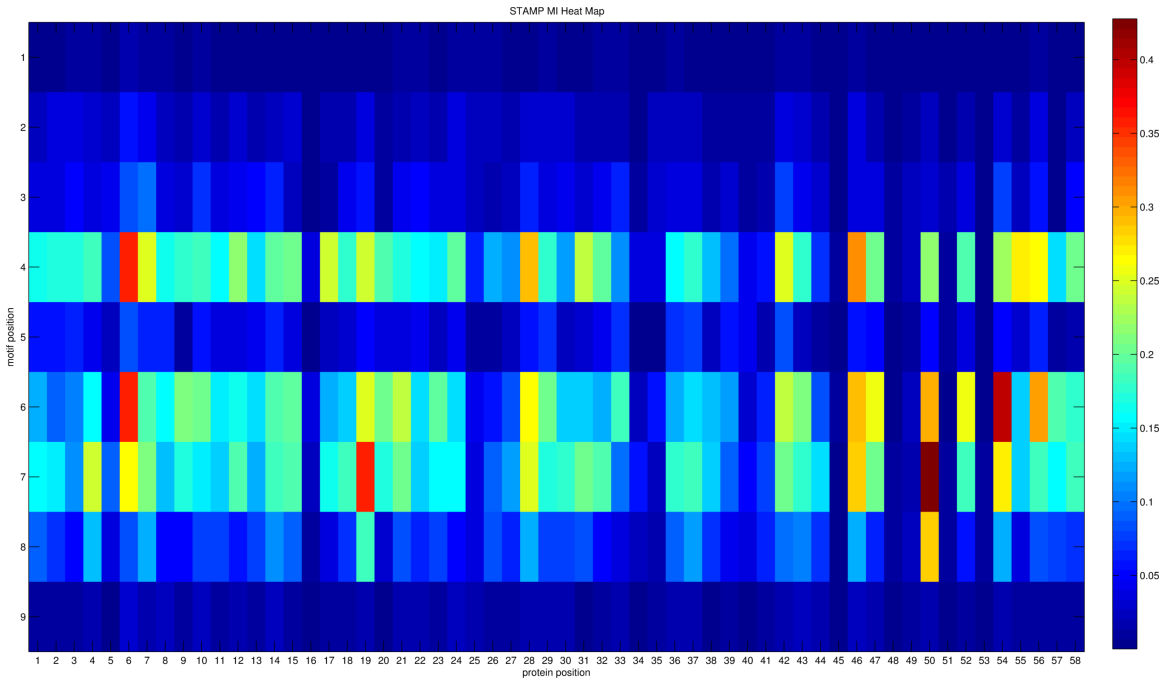


Figure 5.11 Heat map showing the MI matrix returned by STAMP for the HD PFM alignment produced by STAMP and our MAFFT protein alignment

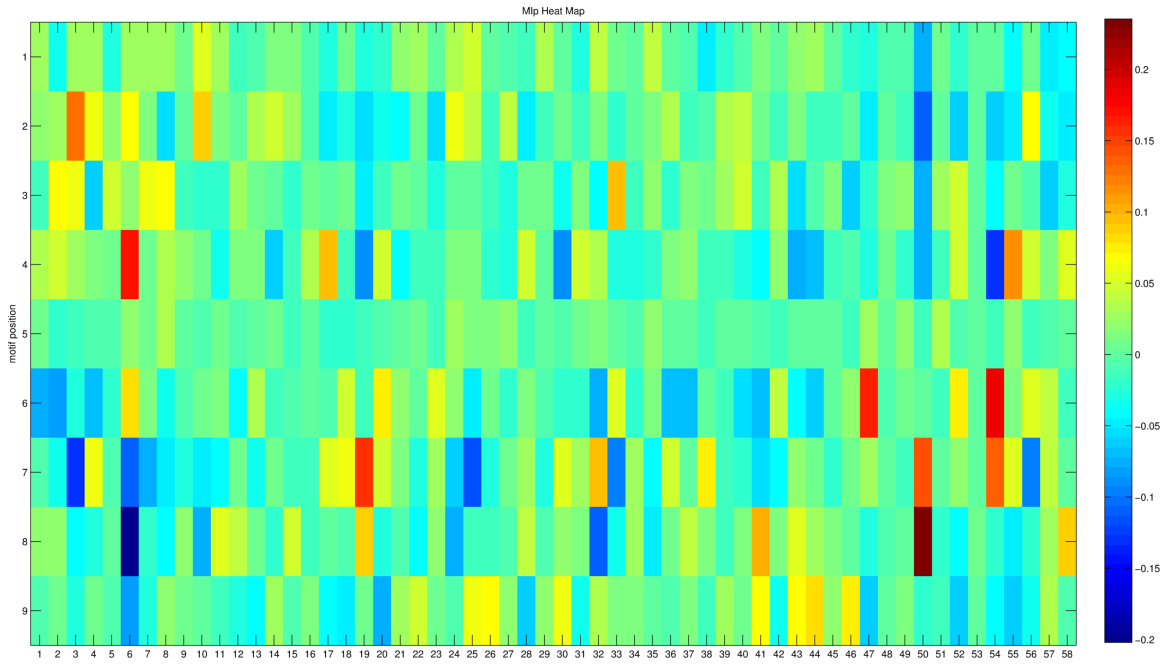


Figure 5.12 Heat map showing the protein alignment versus motif alignment MIp matrix. The PFM alignment was produced using our multiple motif alignment software. The PFMs where then made discrete using our profile alphabet. Position 5 in the MMA is the conserved A bound by Asn 51 in most cases.

The STAMP multiple motif alignment was very similar, yet not identical to our own. To control for the affect of the alignments being slightly different, we calculated the MI and MIp matrices using our own multiple PFM alignment without using the profile alphabet to make the PFMs discrete. Table 5.4 shows the top seven residues returned by the MI and MIp without using the profile alphabet. Using just the MI values gives the same set of

features, just in different order as were returned by STAMP. In this case, the profile alphabet helps find the correct set of key residues.

Table 5.4 Top Features found without using profile alphabet to make PFMs discrete

MI	54	50	19	6	46	56	28
MIp	54	50	17	19	55	52	20

ZF Results

For the zif268 mutant data, adding MIp ranked features improved the MSE until the sixth feature is included. After this, there is no improvement for any of the methods (Figure 5.14). All of the canonical features for these two fingers appear in this set of six features except for one. Finger 3 (F3), recognition helix position +6 was not found to be an important feature and it was not highly ranked in the MIp list (Table 5.5). This was doubtlessly due to the low variability in this position in the data set (see Figure 5.6 and Figure 5.9). For the most part, this position was never randomized or mutated. It is particularly interesting that F3:+2 was found to be important; including this feature noticeably lowered the MSE for all methods. This position is known to bind to the opposite strand of the DNA triplet bound by the adjacent finger 2. However, some models ignore the contribution of this position (i.e. SAMIE).

It is also interesting to note that not only did our method selected the known key residues, but the mapping of key residues to positions in the motif was also perfectly recovered by MIp, except for the F3:+6 and motif position 6 interaction (Figure 5.13)

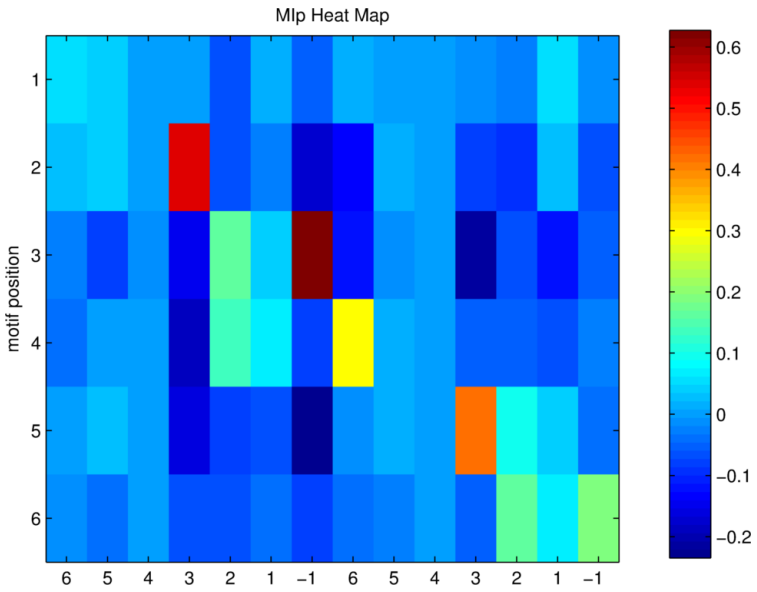


Figure 5.13 Heat map showing the protein alignment versus motif alignment MIp matrix. The PFMs were made discrete using our profile alphabet

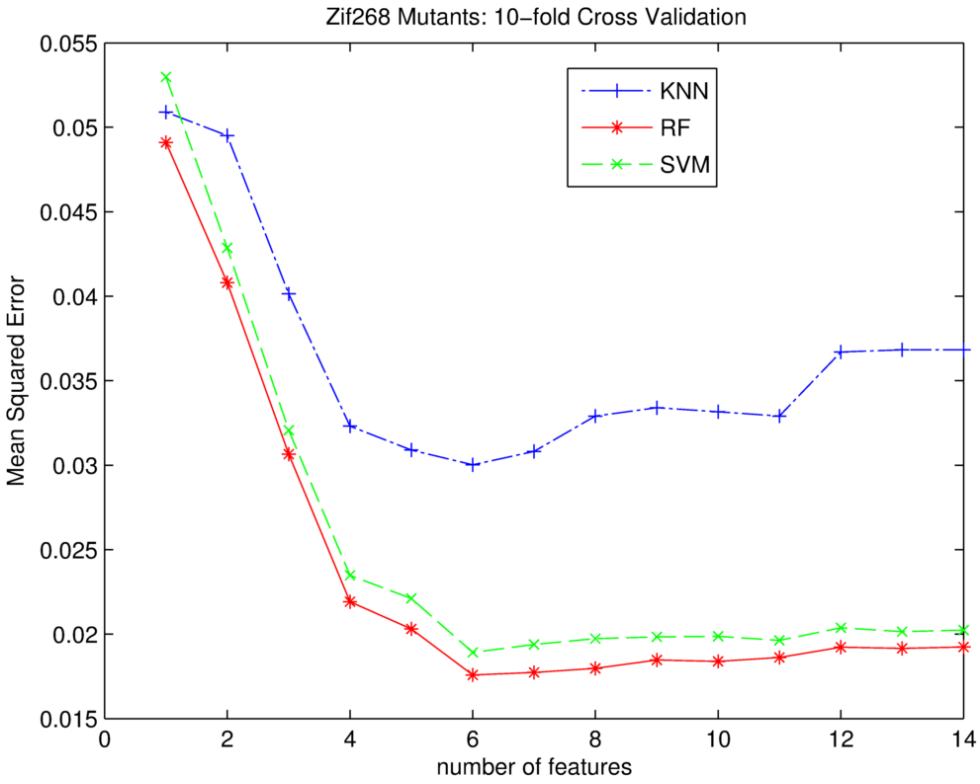


Figure 5.14 Plot of the number of features used to train the KNN, RF and SVM models versus the 10-fold cross validation MSE values for the ZF data set.

After six features were included the MSE stoped decreasing for all. These results are for the two finger models which predicted six position PFMs given the sequence of the finger 2 and finger 3 randomized recognition helices.

Table 5.5 Features ranked by MIp for Zif268 mutant data set

Feature Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Finger	3	3	2	2	2	3	2	2	3	3	3	2	3	2
Recognition Helix Position	-1	3	3	6	-1	2	2	1	1	6	5	5	4	4

We tested models from the literature (SAMIE, KFM05, ZifNet, MGM98, SBGY95) and our single finger RF, KNN, SVM and MBM models trained using recognition helix positions -1, +2, +3, +6 from fingers 2 and 3 of the mutant Zif268 training sets. All of the predicted PFMs using the models from the literature were scaled, except for the SAMIE PFMs, which were nearly optimally scaled. We tested them on the GNN data set, which was not used for training. Figure 5.15 shows the MSE values for these tests. The RF model had the lowest MSE, followed by SAMIE, which is a simple matrix based model. SAMIE performed even better than SVM on this test, which is surprising, while KNN performed relatively poorly. We also performed 10-fold cross-validation to test the models using the Zif268 mutants training set. Again, the RF model performed the best (Figure 5.16). Not surprisingly, all of the models trained on this data set (albeit via 10-fold cross-validation) performed better than all of the models from the literature, since they were not trained on completely different data sets. KNN did not perform very well here either. It only did better than two of our MBM models. It should be noted that these two models are single matrix models that utilize only

one residue-base energy matrix to predict specificities at every position in the motif, so it is not very impressive that KNN beat these two methods.

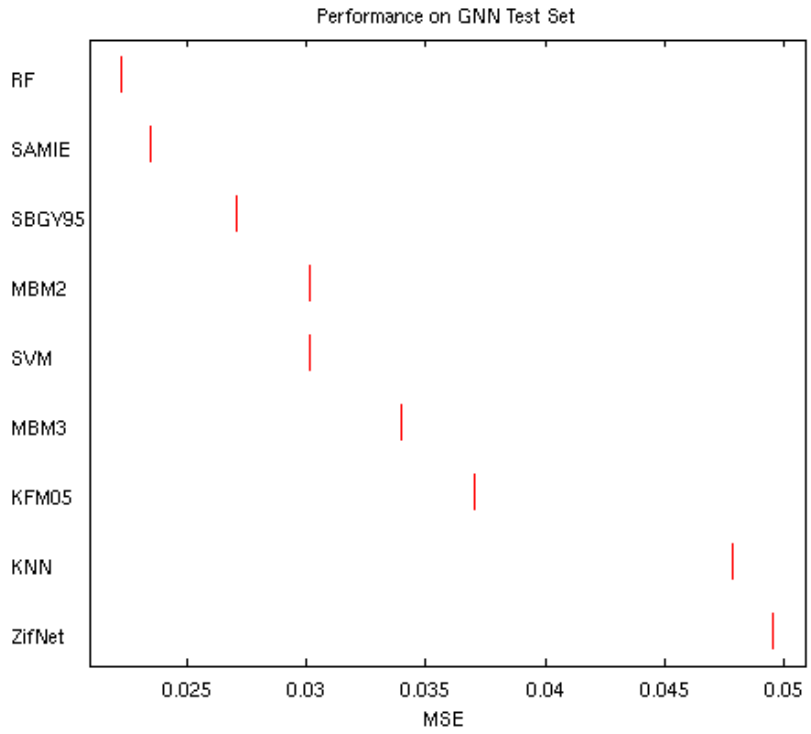


Figure 5.15 Performance on GNN test of single finger models.

The RF, SVM, KNN, MBM3 and MBM2 models are all single finger models trained using all -1,+2,+3,+6 positions from all 3 fingers in the ZF training set. The other models are all from the literature.

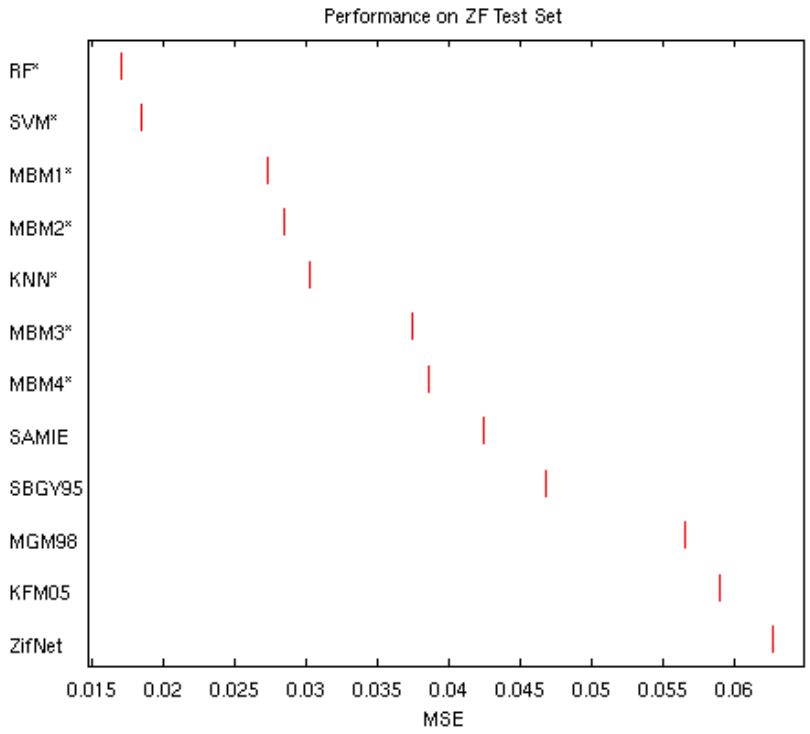


Figure 5.16 Performance on the ZF set.

The MSE scores are from 10-fold cross validation when there is an '*' next to model name. Other models are from the literature. All models not trained on ZF data set were scaled, except for SAMIE

It is interesting to note that our RF recognition model trained using GRaMS PWMs did very well on the ZF test set. Two completely different experimental methods were used to obtain the GNN and Zif268 mutant data sets, yet we did not need to re-scale our PWMs.

Discussion

The goal of this study was to find improved recognition models for the two most abundant TF families, ZF and HD. While ZF proteins have been studied extensively and reasonably good models exist that can predict PWMs, we show that our new models have even better performance. In particular, all of our RF based models out performed the models from the literature and the other methods we tried in every test. These new models performed well partly because of the new, large high quality data set we used for training and partly because we did not impose strict non-additivity imposed by all of the previous models we tested except for ZifNet, which actually performed relatively poorly in our tests. This may be because the method ZifNet uses for generating PWMs may not be optimal.

The HD family on the other hand has been much less studied in terms of a recognition model. Previous work suggested that more sophisticated models did not perform better than a simple nearest neighbor extrapolation. We show that it is possible to obtain good motif alignments that allow us to recover most of the most important key residues reported in the literature. Adding additional features beyond these key residues to the training set does not help any of the models much and actually hurts the performance of KNN.

Acknowledgements

I gratefully acknowledge Yue Zhao for providing me with the HD BEEML-PBM PWM models prior to the publication of his BEEML-PBM program and for helpful advice and discussions. I also thank the other members of the Stormo lab for insightful comments and discussions, in particular Gurmukh Sahota, and Ronak Patel.

References

Ades, S.E., and Sauer, R.T. (1994). Differential DNA-binding specificity of the engrailed homeodomain: the role of residue 50. Biochemistry *33*, 9187-9194.

Alleyne, T.M., Pena-Castillo, L., Badis, G., Talukder, S., Berger, M.F., Gehrke, A.R., Philippakis, A.A., Bulyk, M.L., Morris, Q.D., and Hughes, T.R. (2009). Predicting the binding preference of transcription factors to individual DNA k-mers. Bioinformatics *25*, 1012-1018.

Benos, P.V., Lapedes, A.S., Fields, D.S., and Stormo, G.D. (2001). SAMIE: statistical algorithm for modeling interaction energies. Pac Symp Biocomput, 115-126.

Benos, P.V., Lapedes, A.S., and Stormo, G.D. (2002a). Is there a code for protein-DNA recognition? Probab(ilistical)ly. Bioessays *24*, 466-475.

Benos, P.V., Lapedes, A.S., and Stormo, G.D. (2002b). Probabilistic code for DNA recognition by proteins of the EGR family. J Mol Biol *323*, 701-727.

Berg, J.M. (1992). Sp1 and the subfamily of zinc finger proteins with guanine-rich binding sites. Proceedings of the National Academy of Sciences of the United States of America *89*, 11109-11110.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T.*, et al.* (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266-1276.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol *24*, 1429-1435.

Breiman, L. (2001). Radom Forests. Machine Learning *45*, 5-32.

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology *2*, 1-27.

Choo, Y., and Klug, A. (1994a). Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. Proc Natl Acad Sci U S A *91*, 11168-11172.

Choo, Y., and Klug, A. (1994b). Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. Proceedings of the National Academy of Sciences of the United States of America *91*, 11163-11167.

Choo, Y., and Klug, A. (1997). Physical basis of a protein-DNA recognition code. Curr Opin Struct Biol *7*, 117-125.

Christensen, R.G., Gupta, A., Zuo, Z., Schriefer, L.A., Wolfe, S.A., and Stormo, G.D. (2011). A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. Nucleic acids research.

Clarke, N.D., Kissinger, C.R., Desjarlais, J., Gilliland, G.L., and Pabo, C.O. (1994). Structural studies of the engrailed homeodomain. Protein Sci *3*, 1779-1787.

Draganescu, A., and Tullius, T.D. (1998). The DNA binding specificity of engrailed homeodomain. J Mol Biol *276*, 529-536.

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.

Eddy, S. (2001). HMMER: Profile hidden Markov models for biological sequence analysis.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755-763.

Ekker, S.C., Jackson, D.G., von Kessler, D.P., Sun, B.I., Young, K.E., and Beachy, P.A. (1994). The degree of variation in DNA sequence recognition among four Drosophila homeotic proteins. EMBO J *13*, 3551-3560.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K.*, et al.* (2010). The Pfam protein families database. Nucleic acids research *38*, D211-222.

Fraenkel, E., Rould, M.A., Chambers, K.A., and Pabo, C.O. (1998). Engrailed homeodomain-DNA complex at 2.2 A resolution: a detailed view of the interface and comparison with other engrailed structures. J Mol Biol *284*, 351-361.

Grant, R.A., Rould, M.A., Klemm, J.D., and Pabo, C.O. (2000). Exploring the role of glutamine 50 in the homeodomain-DNA interface: crystal structure of engrailed (Gln50 --> ala) complex at 2.0 A. Biochemistry *39*, 8187-8192.

Herraez, A. (2006). Biomolecules in the computer: Jmol to the rescue. Biochem Mol Biol Educ *34*, 255-261.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2010). A practical guide to support vector classification (National Taiwan University).

Joung, J.K., Ramm, E.I., and Pabo, C.O. (2000). A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. Proc Natl Acad Sci U S A *97*, 7382-7387.

Kaplan, T., Friedman, N., and Margalit, H. (2005). Ab initio prediction of transcription factor targets using structural knowledge. PLoS Comput Biol *1*, e1.

Katoh, K., ichi Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res *33*, 511-518.

Kissinger, C.R., Liu, B.S., Martin-Blanco, E., Kornberg, T.B., and Pabo, C.O. (1990). Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: a framework for understanding homeodomain-DNA interactions. Cell *63*, 579-590.

Liu, B., Kissinger, C.R., and Pabo, C.O. (1990). Crystallization and preliminary X-ray diffraction studies of the engrailed homeodomain and of an engrailed homeodomain/DNA complex. Biochemical and biophysical research communications *171*, 257-259.

Liu, J., and Stormo, G.D. (2008). Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. Bioinformatics.

Liu, Q., Xia, Z., Zhong, X., and Case, C.C. (2002). Validated zinc finger protein designs for all 16 GNN DNA triplet targets. J Biol Chem *277*, 3850-3856.

Mahony, S., Auron, P.E., and Benos, P.V. (2007a). DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies. PLoS Comput Biol *3*, e61.

Mahony, S., Auron, P.E., and Benos, P.V. (2007b). Inferring protein DNA dependencies using motif alignments and mutual information. Bioinformatics *23*, i297-i304.

Mandel-Gutfreund, Y., and Margalit, H. (1998). Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. Nucleic Acids Res *26*, 2306-2312.

Matthews, B.W. (1988). Protein-DNA interaction. No code for recognition. Nature *335*, 294-295.

Newburger, D.E., and Bulyk, M.L. (2008). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res.

Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell *133*, 1277-1289.

Pabo, C.O., Peisach, E., and Grant, R.A. (2001). Design and selection of novel Cys2His2 zinc finger proteins. Annu Rev Biochem *70*, 313-340.

Persikov, A.V., Osada, R., and Singh, M. (2009). Predicting DNA recognition by Cys2His2 zinc finger proteins. Bioinformatics *25*, 22-29.

Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2010). JASPAR 2010: the

greatly expanded open-access database of transcription factor binding profiles. Nucleic acids research *38*, D105-110.

Sato, K., Simon, M.D., Levin, A.M., Shokat, K.M., and Weiss, G.A. (2004). Dissecting the Engrailed homeodomain-DNA interaction by phage-displayed shotgun scanning. Chem Biol *11*, 1017-1023.

Seeman, N.C., Rosenberg, J.M., and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci U S A *73*, 804-808.

Serebriiskii, I.G., Fang, R., Latypova, E., Hopkins, R., Vinson, C., Joung, J.K., and Golemis, E.A. (2005). A combined yeast/bacteria two-hybrid system: development and evaluation. Mol Cell Proteomics *4*, 819-826.

Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*, 16-23.

Stormo, G.D. (2011). Maximally efficient modeling of DNA sequence motifs at all levels of complexity. Genetics *187*, 1219-1224.

Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. Nat Rev Genet *11*, 751-760.

Suzuki, M., Brenner, S.E., Gerstein, M., and Yagi, N. (1995). DNA recognition code of transcription factors. Protein Eng *8*, 319-328.

Tucker-Kellogg, L., Rould, M.A., Chambers, K.A., Ades, S.E., Sauer, R.T., and Pabo, C.O. (1997). Engrailed (Gln50-->Lys) homeodomain-DNA complex at 1.9 A resolution: structural basis for enhanced affinity and altered specificity. Structure *5*, 1047-1054.

Tupler, R., Perini, G., and Green, M.R. (2001). Expressing the human genome. Nature *409*, 832-833.

Wang, T., and Stormo, G.D. (2005). Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. Proc Natl Acad Sci U S A *102*, 17400-17405.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. Nucleic acids research *28*, 316-319.

Wolfe, S.A., Greisman, H.A., Ramm, E.I., and Pabo, C.O. (1999). Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. J Mol Biol *285*, 1917-1934.

Wolfe, S.A., Nekludova, L., and Pabo, C.O. (2000). DNA recognition by Cys2His2 zinc finger proteins. Annual review of biophysics and biomolecular structure *29*, 183-212.

Zhao, Y., and Stormo, G.D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat Biotechnol *29*, 480-483.

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.

Chapter 6: Conclusions and Future Directions

Overview and Concluding Remarks

The goal of this dissertation was to develop improved recognition models for the two largest classes of transcription factors (TFs), the C2H2 zinc finger (ZF) and homeodomain (HD) families. More accurate recognition models have many different potential uses. They can be applied to predict the binding sites of uncharacterized TF. They can be used to engineer new proteins with novel specificities and new functions, such as zinc finger nucleases (Mani et al., 2005).

The HD family has not previously received as much attention as the ZF family with regards to learning recognition models. This is partly due to a lack of data. In the same issue of the journal Cell, two articles appeared that reported the entire catalog of HD DNA specificities for the mouse (Berger et al., 2008) and fly (see chapter 2) genomes. In chapter 2, a simple nearest neighbor based recognition model was introduced, that to our knowledge represented the first semi-quantitative recognition model for the HD family able to predict PFMs based on protein sequence alone. Analysis of the fly HD set highlighted the importance of a limited set of key residues in determining DNA binding preferences. Site

directed mutagenesis showed that changing only a few of these key residues could result in very different DNA binding preferences and that changing non-key residues did not generally have the same magnitude of an effect, giving further credence to the idea of an HD recognition code.

To facilitate the development of a new ZF recognition model, a new analysis method, GRaMS, was developed specifically to model Illumina sequenced B1H data. We introduced this method in chapter 4 and demonstrated that it was robust to different experimental conditions such as the 3-AT and IPTG concentrations as well as the duration of the experiment. We also introduced a B1H variant called CV-B1H, in which a portion of the binding site is fixed to the consensus, so that only some positions are randomized. We show that analyzing CV-B1H data using GRaMS yields models capable of predicting high throughput SELEX data (Zhao et al., 2009) with high accuracy. We also showed however that for 'unconstrained' B1H data produced using larger randomized regions, the popular program MEME (Bailey et al., 2006) could be run on the set with reasonably good results and that surprisingly GRaMS did not improve performance for this type of data. Chapter 3 describes the online database, FlyFactorSurvey, which is a repository for Drosophila B1H data sets. An analysis pipeline was developed based in part on our findings described in chapter 4 and this analysis pipeline was used to analyze B1H data sets for more than 200 different TFs from at least 11 different Pfam families.

In chapter 5, we presented a new random forest (RF) based recognition model which out performed all of the other methods and models that we examined, including K Nearest Neighbors (KNN). We showed that our new method has the potential to be generally applicable to multiple TF families, in that it was successfully applied to both the ZF and HD families. In the future, we wish to apply this method to other TF families as UniPROBE (Newburger and Bulyk, 2008), FlyFactorSurvey (Zhu et al., 2011) and other TF specificity data bases continue to expand.

For the HD analysis in chapter 5, we showed that it was possible to successfully combine PMB and B1H motif matrices produced from different types of experimental assays and using different analysis methods. It did prove necessary to scale the HD BEEML-PBM (Zhao and Stormo, 2011) and B1H motifs from FlyFactorSurvey, however. Why these two sets of motifs are scaled differently remains unknown and warrants further investigation.

As part of the training process for constructing the recognition models, we found that it was beneficial remove unnecessary protein residues, or features, from the training set. Our adjusted mutual information based feature selection method was able to recover known sets of key residues for both the ZF and HD families with high accuracy. We only recovered a single residues that we believe to be a false positive, residue number 19 in the HD protein alignment. It could be that this position is phylogenetically informative and that our

adjusted mutual information method is unable to account for this. In other words, perhaps position 19 is only serving as a marker for membership in a clade, and that members of this clade tend to have similar DNA binding preference. Residue 19 was picked out by a variety of different methods as well, so it potentially warrants further study. It would be relatively simple to investigate this position as a potential key residue via mutational studies using the B1H system. If mutating this residue has no significant effect on specificity, then it is probably only a marker for other traits. This highlights the benefits of data sets like the Zif268 mutant set described in chapter 5. Randomizing the recognition helix virtually eliminates any potential for phylogentically informative residues to be misidentified as key residues. It would be beneficial to carry out similar studies of the HD family. It would especially be interesting to do this since there are multiple key residues sets utilized by different subsets of this family, as discussed in chapter 2.

Contrary to the claims of a recent paper (Alleyne et al., 2009), we did not find that the simple nearest neighbor method to be as affective as random forest or SVM based recognition models. Alleyne et al. also claim that it is not feasible to align HD motifs. PFM alignment is necessary in order to construct a recognition model that predicts PWM or PFM matrices. In stead, they predicted the PBM Z-score transformed signal intensities for all non-redundant 8mers, requiring over 32,896 models. This must required a lot of CPU time to train, which doubtless contributed to their adoption of the KNN method, which only

requires a matrix of pairwise protein Hamming distances. In order to learn recognition models that predict a PFM, we only need to learn 18 or 9 different RF models for the ZF data, depending on if a single or two finger PFM is predicted, and 27 models to predict the 9 position trimmed HD PFMs (we learn a single model per free parameter in the PFM). This is a much simpler task computationally and is more satisfying since many fewer parameters are required.

In the future, a comparison should be made to determine the ability of the HD PFMs predicted by our RF recognition model to predict raw PBM data. A recent study shows that contrary to claims made in Alleyne et al. and elsewhere (Berger et al., 2006), PWMs, when trained with a biophysical based model such as BEEML-PBM, can explain PBM 8mer intensities quite well, often as well as replicate data sets (Zhao and Stormo, 2011). It may be that since the HD RF recognition model was trained using BEEML-BPM that the motifs that it predicts will also be able to explain raw PBM data well.

Future Directions

Recently Sahota et al. introduced a novel comparative genomics method for discovering motifs using hundreds of bacterial genomes (Sahota and Stormo, 2010). The novel aspect of their approach is that instead of searching for motifs in orthologous sets of

promoters, they rely on the observations that (1) TFs with identical key residues from the same family tend to bind similar motifs and (2) up to 80% of operons may be auto regulatory, if neighboring operons are considered. Based on these observations, they assembled set of promoters that corresponded to TFs that were identical at the key residue sets learned from crystal structures, but which were not otherwise required to be similar. Sahota et al. concentrated on two HTH subfamilies LacI and TetR and they were able to discover many motifs that matched experimentally determine motifs using only genomic data and a knowledge of the LacI and TetR key residue sets.

For my postdoctoral work, I plan to extend my research developing recognition models for the ZF and HD families to the HTH TF family using primarily genomic data as input in order to extend the work of Sahota et al. One of the issues that Sahota et al. faced is that many of the clusters of TFs with identical key residues are too small for effective motif discovery using methods like MEME (Bailey et al., 2006). It seems that it would be ideal to find a way to use all of the data at once to learn a family recognition model rather than to divide the data up into data sets that are often too small. Currently, the information contained in one group of TF's promoters cannot be applied to another group during the motif learning process. Combining information learned from different key residue clusters can currently only occur through some type of post processing set. I hope to develop a method similar in spirit to the method developed by Kaplan et al. that used the EM

algorithm to identify binding sites and learn a recognition model for the ZF family simultaneously (Kaplan et al., 2005) using only sequences known to be enriched for binding sites but without knowing their location (the model they learned, KFM05, is discussed in chapter 5).

Conclusion

In summary, we developed perhaps the first homeodomain recognition model that predicts PFMs based on identity at key residues in the protein and overall homology. We subsequently constructed a new HD recognition model using newly available BEEML-PBM models and B1H data sets available from the FlyFactorSurvey database. We developed methods to analyze B1H data, such as GRaMS that performed better than all other tested methods, using an HT-SELEX Zif268 data set as a benchmark. GRaMS was used to analyze more than 250 new B1H Zif268 mutant data sets. These PWMs were used to learn a new random forest based recognition model that we found to perform better than all methods that we tested. We show that contrary to recent claims, the nearest neighbor method does not compare favorably to random forest based recognition models. We also show that it is feasible to align most homeodomain PFMs and that these alignments can be used together with a novel feature selection method to determine key residue sets with high accuracy. We

believe the our recognition models have improved performance because they were trained with higher quality data and because the random forest method is able to model higher order interactions between key residues.

References

Alleyne, T.M., Pena-Castillo, L., Badis, G., Talukder, S., Berger, M.F., Gehrke, A.R., Philippakis, A.A., Bulyk, M.L., Morris, Q.D., and Hughes, T.R. (2009). Predicting the binding preference of transcription factors to individual DNA k-mers. Bioinformatics *25*, 1012-1018.

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res *34*, W369-W373.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Peña-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T.*, et al.* (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266-1276.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol *24*, 1429-1435.

Kaplan, T., Friedman, N., and Margalit, H. (2005). Ab initio prediction of transcription factor targets using structural knowledge. PLoS Comput Biol *1*, e1.

Mani, M., Kandavelou, K., Dy, F.J., Durai, S., and Chandrasegaran, S. (2005). Design, engineering, and characterization of zinc finger nucleases. Biochem Biophys Res Commun *335*, 447-457.

Newburger, D.E., and Bulyk, M.L. (2008). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res.

Sahota, G., and Stormo, G.D. (2010). Novel sequence-based method for identifying transcription factor binding sites in prokaryotic genomes. Bioinformatics.

Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring binding energies from selected binding sites. PLoS Comput Biol *5*, e1000590.

Zhao, Y., and Stormo, G.D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat Biotechnol *29*, 480-483.

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.

CURRICULUM VITAE

Ryan G. Christensen

Department of Genetics
Washington University School of Medicine
St. Louis, MO 63108

Lab: (314) 747-5535
Cell: (214) 797-4477
Email: christensen@wustl.edu

EDUCATION

Washington University in St. Louis 2004 – present

Ph.D. candidate in Computational Biology (anticipated August 2011)
Adviser: Dr. Gary Stormo
Thesis: Protein-DNA Recognition Models for the Homeodomain and C2H2 Zinc Finger Transcription Factor Families

Brigham Young University 2001 – 2004

B.S. in Bioinformatics and in Neuroscience

RESEARCH EXPERIENCE

Washington University in St. Louis St. Louis, MO

Pre Doctoral Research Fellow, Dr. Gary Stormo Lab 2004 – present

- Developed computational method to learn quantitative models of transcription factor DNA specificity from high throughput bacterial one hybrid (B1H) experiments
- Constructed predictive models of transcription factor DNA specificity for the Homeodomain and C2H2 Zinc Finger transcription factor families

Brigham Young University Provo, UT

Research Associate, David A. McClellan Lab 2002 – 2004

- Investigated the co-evolution of polyomaviruses and their hosts using a wide range of phylogenetic programs and bioinformatics tools
- Managed the lab undergraduate researchers program
- Helped develop the new Bioinformatics course and served as its first teaching assistant

TEACHING EXPERIENCE

Washington University in St. Louis Fall Semester, 2005
Graduate Teaching Assistant, Bio5495, Computational Molecular Biology
Course Master: Professor Sean Eddy

Brigham Young University Fall Semester, 2003
Teaching Assistant for new Bioinformatics course
Course Master: Professor David McClellan

HONORS AND AWARDS

- National Science Foundation Graduate Research Fellowship honorable mention 2006
- BYU Cancer Research Center Fellowship 2003
- BYU Office of Research and Creative Activities grant 2002

SERVICE

- Young Scientist Program educational outreach volunteer 2004-present
- Scout Master (Boy Scouts of America) 2007-2008
- Taught free English lessons, volunteered at the Red Cross in Hungary 1998-2000

PUBLICATIONS

1. **Christensen, R.G.**, A. Gupta, M.B. Noyes, M.H., S.A. Wolfe, G.D. Stormo. "Protein-DNA Recognition Models for the Homeodomain and C2H2 Zinc Finger Transcription Factor Families." (*in preparation*).
2. **Christensen, R.G.**, A. Gupta, Z. Zuo, L.A. Schriefer, S.A. Wolfe, G.D. Stormo. "A Modified Bacterial One-Hybrid System Yields Improved Quantitative Models of Transcription Factor Specificity." *Nucleic Acids Res. Epub (2011).*
3. Zhu, L.J., **R.G. Christensen**, M. Kazemian, C.J. Hull, M.S. Enuameh, M.D. Basciotta, J.A. Brasefield, C. Zhu, Y. Asriyan, D.S. Lapointe, S. Sinha, S.A. Wolfe, M.H. Brodsky. "FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system." *Nucleic Acids Res. 39:D111-7. Epub (2010).*
4. Noyes, M.B., **R.G. Christensen**, A. Wakabayashi, G.D. Stormo, M.H. Brodsky, S.A. Wolfe. "Analysis of Homeodomain Specificities Allows the Family-Wide Prediction of Preferred Recognition Sites." *Cell.* 133(7):1277-89 (2008).

5. Zeng, J., J. Yan, T. Wang, D. Mosbrook-Davis, K.T. Dolan, **R. Christensen**, G.D. Stormo, D. Haussler, R.H. Lathrop, R.K. Brachmann, S.M. Burgess. "Genome Wide Screens in Yeast to Identify Potential Binding Sites and Target Genes of DNA-Binding Proteins." *Nucleic Acids Res.* 36(1):e8 Epub (2007).
6. Stanley, S.L., S.E. Frey, P. Taillon-Miller, J. Guo, R.D. Miller, D.C. Koboldt, M. Elashoff, **R. Christensen**, N.L. Saccone, R.B. Belshe. "The Immunogenetics of Smallpox Vaccination." *J Infect Dis.* 196(2):212-219 (2007).
7. Perez-Losada, M., **R.G. Christensen**, D.A. McClellan, B.J. Adams, R.P. Viscidi, J.C. Demma, K.A. Crandall. "Comparing Phylogenetic Codivergence between Polyomaviruses and Their Hosts." *J Virol.* 80: 5663-9 (2006).
8. McClellan, D.A., E.J. Palfreyman, M.J. Smith, J.L. Moss, **R.G. Christensen**. "Physicochemical Evolution and Molecular Adaptation of the Cetacean and Artiodactyl Cytochrome B Proteins." *Mol Biol Evol.* 22: 437-55 (2005).
9. McClellan, D.A., D.F. Whiting, **R.G. Christensen**, J. Sailsbery. "Genetic Codes as Evolutionary Filters: Subtle Differences in the Structure of Genetic Codes Result in Significant Differences in Patterns of Nucleotide Substitution." *J Theor Biol.* 226: 393-400 (2004).
10. Crandall, K.A., M. Pérez-Losada, **R.G. Christensen**, D.A. McClellan, R.P. Viscidi. "Phylogenomics and Molecular Evolution of Polyomaviruses." In: Polyomaviruses and Human Diseases. Nasimul Ahsan, ed. Landes Bioscience, Georgetown, Texas (2004).

POSTERS

1. **Christensen, R.G.** and G.D. Stormo. *Systems Biology: Global Regulation of Gene Expression.* Cold Spring Harbor, NY (2010).
2. **Christensen, R.G.** and G.D. Stormo. *Intelligent Systems for Molecular Biology.* Toronto, CA (2008).
3. Vernon, H.J., **R.G. Christensen**, S. Woolley, M.D. Dyer, P.A Ringstrom, D. Almond, D.A. McClellan. *Society of Systematic Biology & Society for the Study of Evolution*, Fort Collins, CO (2004).
4. Sailsbery, J.K., **R.G. Christensen**, D.A. McClellan. (1) *Society of Systematic Biology & Society for the Study of Evolution*, Chico, CA; (2) *Society for Molecular Biology and Evolution*, Newport Beach, CA (2003).