ABSTRACT

MCFERRIN, LISA GAIL. Modeling the Molecular Evolution of Protein Domains and Networks. (Under the direction of William R. Atchley and Eric Stone.)

Protein sequences are subject to a mosaic of constraints. Such constraints often manifest in patterns of conservation and can reveal important structural or functional features among homologous sequences.

Regions of constraint on the tertiary structure of a protein result in loose segmentation of its primary structure into stretches of slowly- and rapidly-evolving amino acids. We demonstrate that the regional nature of structural and functional constraints asserts a positive autocorrelation on the evolutionary rates of neighboring sites. Using a simple dispersion statistic that quantifies the degree of non-synonymous clustering for genome-wide interspecific comparisons of orthologous protein pairs, we show non-synonymous clustering intensifies with increasing purifying selection, revealing a strong log-linear relationship between the degree of clustering and the intensity of constraint. This relationship is also preserved in yeast, even after accounting for other selection correlates such as protein abundance and dispensability. While we do not claim that the dispersion ratio is an optimal statistic, we propose that it may help uncover conflicting signals of constraint that would otherwise be lost in a historical switch in selective regime. In general, the correlation between clustering and evolutionary rate supports the use of *de novo* annotation methods that have implicitly assumed these selective constraints.

Regional constraint also forms conserved domains such as the basic-helix-loop-helix-zipper (bHLHZ) domain responsible for DNA binding and dimerization. The bHLHZ domain defines the interactions between Max and Mlx transcription factor network members (Max, Mlx, Mondo, Mga, Mnt, Mxd, and Myc), which are integral for differentiation, proliferation, and energy metabolism. We identified the bHLHZ domain of Max and Mlx network members in all sampled animal lineages, dating these networks to ≈1 billion years ago. Throughout evolution, only four major network configurations emerged 1) most species contain core network members (Myc, Max, Mlx, Mondo, Mnt, and Mxd), 2) vertebrates experienced additional radiation of protein families (c-, N-, and L-Myc, Mxd1-4, MondoA and MondoB), 3) Mxd was lost in the Diptera lineage, and 4) a major network reconstruction

in nematodes formed a Mxd-like (MDL-1), a Myc and Mondo-like (MML-1), a Mlx-like (Mxl-2), and two Max-like (Mxl-1, Mxl-2) proteins. Phylogenetic reconstruction of the bHHZ domain shows distinct conservation among orthologous proteins, while multivariate discriminant analysis reveals particular residues that classify proteins, families, and network configurations. Such differences likely affect the recognition of gene target sequences, and hence alter the function of these transcription factors.

In vertebrates, MondoA and MondoB paralogs regulate energy homeostasis by transactivating genes involved in glucose metabolism. In addition to the bHLHZ domain, five unique N-terminus domains named Mondo Conserved Regions (MCRI-V) regulate their transactivational activity in response to glucose-6-phosphate (G6P). Through sequence conservation, we identified an additional constrained region (MCR6) that is located in a region capable of transactivation. MCR6 has a SxSxxT motif that is similar to known G6P binding regions, suggesting G6P may bind to MCR6 allosterically. MCR6 also matches the sequence signature of a 9aa transactivation domain known to interact with histone acetyltransferases such as MondoB interaction partner CBP/p300. Structural predictions of the N-terminal region agree with existing experimental evidence, which indicate that intramolecular interactions between MCRI-IV, MCR6 and MCRV may affect MondoA and MondoB conformation in different glucose conditions. Hence, MCR6 may provide the necessary link in determining how MondoA and MondoB are able to transactivate genes in response to glucose.

Modeling the Molecular Evolution of Protein Domains and Networks

by
Lisa Gail McFerrin

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2010

APPROVED BY:

Eric Stone Co-chair	William Atchley Co-chair
Jeffrey Thorne	Spencer Muse

Ignazio Carbone

Dedication

To my family and friends. Without you, nothing would be possible.

Rachel and Ashley, you have been there for me even when you didn't know it. Thank you for being my rock and keeping me optimistically honest.

My Hokies, you are my spirit. Thank you for keeping my passion for life strong and never letting me take myself too seriously.

My friends both new and old, you always know how to make me smile. Thank you for filling my memories with laughter.

My family, you continually give me strength. Thank you for teaching and encouraging me to be the best I can be.

I love you all very much. Thank you for making my life so wonderful.

Biography

Lisa McFerrin graduated in 2005 from Virginia Polytechnic Institute and State University with two Commonwealth Scholar Bachelors of Science in both Mathematics and Computer Science. During her time at Virginia Tech, she participated in two Research Experiences for Undergraduates, one at Mississippi State University and the other through the Institute of Pure and Applied Mathematics at University of California, Los Angeles. Through this latter experience she worked with a company called BioDiscovery and developed algorithms for identifying key genes in brain cancer. This prompted her to pursue a degree in the field of bioinformatics, which couples her skills in math and computer science with the fascinating field of biology. In the fall of 2005, she enrolled in the Bioinformatics program at North Carolina State University and has since focused on the application of statistical methods to develop protein models in systems associated with metabolic disease and cancer.

Acknowledgements

Thank you to my advisors Bill Atchley and Eric Stone. You have been there for me as both friend and mentor. I will never forget our trips to New York, China, and Applebees. I couldn't have asked for better people to share in this learning endeavor.

Thank you to my committee members Jeff Thorne, Spencer Muse, and Ignazio Carbone for your time, comments, and contribution.

Thank you to the faculty and staff of the Bioinformatics Program and Department of Genetics at the North Carolina State University for your friendliness and unabated willingness to help.

A special thanks to Benjamin Dorshorst, Jonathan Keebler, Rachel Myers, and Shengdar Tsai. You receive my utmost gratitude, respect, and admiration. It has been a pleasure getting to know you these past several years and I am honored to be considered your colleague.

Table of Contents

List of Tables

List of Figures

Forward

This dissertation on protein evolution and structure focuses on the development and application of statistical methods to identify functional variation in protein domains and networks. Chapter 1 provides a broad introduction to sequence analysis and methods of detecting selection, while a technical review and comparison of methods used to handle high dimensional molecular data is presented in Appendix A. The body of this research is covered in chapters 2, 3, and 4 and is composed of two related parts. Chapter 2 describes the first project, which investigates the effect of protein constraint on the gene sequence by quantifying the correspondence between selection and dispersion of amino acid altering changes. Chapter 3 and 4 cover the second project, which concentrates on Max and Mlx network proteins. Chapter 3 focuses on the identification and ramification of amino acid changes within the dimerization and DNA binding basic helix-loop-helix leucine zipper (bHLHZ) domain within the C-terminus, while Chapter 4 proposes a model that explains the glucose responsive transactivation activity governed by the N-terminus of the Mlx network proteins MondoA and ChREBP. Chapter 5 provides a general discussion. Figures and references are located within each chapter.

Chapter 1

Introduction and Background

A protein is a linear chain of amino acids connected by peptide bonds that can fold into a three-dimensional tertiary structure. Amino acid physicochemical properties such as hydrophobicity, polarity, volume, Van der Waals forces, and hydrogen bonds confer physicochemical constraints that affect protein structure and function (Kidera, Konishi et al. 2009). Often amino acids with similar, overlapping traits can function interchangeably within the protein context, e.g. leucine to isoleucine, while other replacements would likely abrogate protein function, e.g. leucine to proline (Majewski and Ott 2003). One method of inferring the extent of these constraints is to compare orthologous proteins that originate from a common ancestor yet retain similar functions in independently evolving species. Sites with particular amino acid restrictions are expected to be conserved over large evolutionary times while sites exhibiting variability are likely to indicate random or adaptive changes. In general, buried residues and active sites are highly conserved while surface residues are more variable in a folded protein (Ma, Elkayam et al. 2003; Tuncbag, Gursoy et al. 2009). Herein we focus on the development and application of methods to accurately quantify and distinguish factors constraining protein variation in the pursuit of understanding protein function and evolution.

Multiple Sequence Alignments

According to the NCBI website, 564 Eukaryotic and 1956 Prokaryotic genomes have been or are being sequenced as of October 2010 (Sayers, Barrett et al. 2010). One can query such a genome database to find significantly conserved sequences from several diverse species and infer their evolutionary relationship or functional similarities.

The predominant query method is Basic Local Alignment Search Tool (BLAST), which uses a heuristic algorithm to first locate short, highly conserved matches that anchor the alignment, then contingently extends each match to include less conserved regions (Altschul, Gish et al. 1990). Significance of a BLAST alignment score is determined by the

Gumbel Extreme Value Distribution, which depends on the query sequence length, database sequence length, substitution matrix, gap penalties, sequence composition, and the number of sequences in the database. The expected value (E-value) of a particular BLAST alignment score S is the expected number of sequences that would score at least as high as S by chance within the database. Alignments with extremely low E-values may indicate a common function or origin between sequences, although it is important to note that similarity does not necessarily imply homology. Since many genomes are not fully annotated, these genomic resources are useful for identifying and annotating potential orthologs in diverse organisms and supplying a comprehensive sample for comparative analysis.

Analyses of sequence evolution rely heavily upon adequate sampling of sequence variability as well as quality of the sequence alignment. Correct alignment is crucial for testing the magnitude and direction of evolutionary pressure on a site. Unfortunately, aligning multiple orthologous sequences is not a trivial task (Carrillo and Lipman 1988; Wang and Jiang 1994). Dynamic programming methods like the Needleman-Wunsch global alignment and Smith-Waterman local alignment algorithms provide optimal solutions, but require too much computational time and memory to be feasible for many sequences (Feng and Doolittle 1987; Taylor 1988; Lipman, Altschul et al. 1989; Elias 2006; Kumar and Filipski 2007). Ultimately, the trade-off in alignment algorithms is speed versus accuracy, with the former generally prevailing. Hence visual inspection of the aligned sequences is necessary to ensure alignment columns accurately represent homologous sites.

ClustalW is one of the most widely used alignment programs and employs a two stage progressive algorithm (Feng and Doolittle 1987). First ClustalW constructs a guide tree from a distance matrix of pairwise alignment scores. The guide tree is typically formed using a Neighbor Joining (NJ) algorithm (Saitou and Nei 1987), described in the next section. ClustalW aligns the most closely related sequences and merges their nodes on the guide tree by consolidating the distance matrix according to an objective function. The most similar sequence or group of sequences is progressively added until all nodes are merged and a single alignment is formed. While the progressive method used in ClustalW and T-Coffee (Notredame, Higgins et al. 2000) is fast and fairly reliable, it is sensitive to initial alignment

errors and to sequences of different lengths (Fitch and Smith 1983; Thompson, Plewniak et al. 1999; Pollard, Bergman et al. 2004; Kumar and Filipski 2007).

The global Muscle (Edgar 2004) and local Dialign (Subramanian, Kaufmann et al. 2008) iterative methods avoid initial alignment traps by realigning subsets of sequences and show mild improvements for some alignments. Other methods include the likelihood and optimization algorithms used in Hidden Markov Models (HMMs), simulated annealing, and genetic algorithms, although they are used less commonly due to setbacks in speed and parameter requirements (Kim and Pramanik 1994; Notredame and Higgins 1996; Lassmann and Sonnhammer 2005; Song, Liu et al. 2010). Mafft, which uses a fast Fourier transform to identify homologous regions, has been shown to outperform most of these methods in both speed and accuracy (Katoh, Misawa et al. 2002).

A good operational scheme for attaining accurate alignments is to use an algorithm such as Altavist (Morgenstern, Goel et al. 2003), which computes both a local (Dialign) and global alignment on the data and allows for alignment comparison. Regions where the two algorithms agree are accepted as "correct", while regions where the two algorithms differ can be manually adjusted. Once an alignment is established, evolutionary constraints can be inferred through several methods including selection, phylogenetic and covariance models.

Models of Selection

Protein constraint is the limitation in (physicochemical, structural, binding, etc.) properties a functional polypeptide can tolerate within a certain environment. From sequence alignments, protein constraint can be observed by the non-stochastic pattern of changes in nucleotides, codons, or amino acids.

Population genetics theory states the probability of fixation for a mutation depends on the population size and fitness conferred (Fisher 1930; Wright 1931; Kimura 1962). A mutation that affects the reproductive ability of an organism is likely to be selectively removed or sustained within subsequent generations depending on its deleterious or advantageous impact, respectively. However, Kimura's neutral theory of evolution states that most interspecies changes are caused by random genetic drift rather than selection (Kimura 1968), while Fisher claims that drift accounts for only a minor portion of genetic variability

especially for large populations (Fisher 1930). A more recent theory of "near neutrality" looks for a balance between these concepts and considers the interaction and importance of drift and selection at various levels, e.g. silent and amino acid altering substitutions and sequence turnover of regulatory elements (Ohta 2002). Consequently, the lack of a definitive evolutionary model lead many to agree with the well-known statistician George Box, "All models are wrong, but some are useful" (E. P. Box and Richard Draper 1987).

Still, models of evolution can determine the relative intensity and direction of selection among orthologous proteins by comparing the number and type of observed changes. The expected number of changes per site, or evolutionary distance, between orthologs provides a measure of sequence conservation via a specified model of substitution (Jukes and Cantor 1969; Tavaré 1986). Let q_{ij} be the instantaneous rate of a site changing from state i to j and Q be the matrix of all possible transitions between states. For DNA models, states consist of the four nucleotides, i, j ={A, C, G, T}, while protein models have 20 amino acid states and codon models have 61 states of nucleotide triplets (excluding stop codons). To retain the probabilistic framework, each site is required to be within a defined state and each row of Q must sum to zero. The probability a site in state i will change to state j after time t is defined as $P(t) = e^{Qt}$. Assuming site independence in a pairwise sequence comparison, the likelihood of observing the alignment of sites k=1...N is thus $L = \prod_k \pi_{x_k^a} P\left(x_k^d \mid x_k^a, Qt\right)$ where π represents the state stationary frequency for ancestral sequence x^a and descendant sequence x^d separated by time t (Figure 1a). The goal is to optimize the estimated substitution rate parameters associated with Q that maximize the probability of observed data.

Note that rate and time cannot be disentangled, but must be jointly estimated. The expected number of changes between sequences after time t is empirically estimated as distance $\hat{d} = t\sum_{i \neq j} \pi_j \hat{q}_{ij}$ where substitution rate $\hat{q}_{ij}$ is estimated from the data. This distance estimates the divergence between sequences and is used to construct the branch lengths in phylogenetic trees, as discussed in the following section.

The simplest model of sequence evolution is the Jukes Cantor (JC69) model of nucleotide substitution (Jukes and Cantor 1969). The JC69 model assumes every nucleotide changes at the same rate λ, so $q_{ij}=\lambda$ when $i \neq j$ and $q_{ij}=-3\lambda$ otherwise (Figure 1b). Thus, after time t, a site has probability $p_r(t) = \frac{1}{4}\left(1+3e^{-4d/3}\right)$ of remaining in the same state and probability $p_r(t) = \frac{1}{4}\left(1-e^{-4d/3}\right)$ of being in a different state where distance $d=3\lambda t$. The estimated distance between two aligned sequences is then $\hat{d} = -\frac{3}{4}\ln\left(1-4\hat{p}/3\right)$, where $\hat{p}$ is the observed proportion of nucleotide changes. Other increasingly complex models account for different substitution rates among transitions and transversions (Kimura 1980), unequal base compositions (Felsenstein 1981; Tamura and Nei 1993), and combinations thereof (Hasegawa, Kishino et al. 1985).

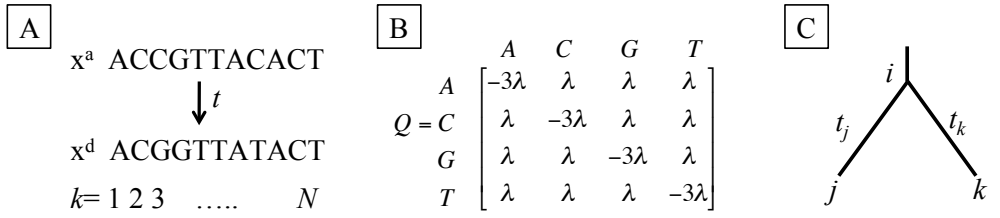


Figure 1: Probabilistic Models of Evolution

A) Ancestral and Descendant Sequences. The likelihood of observing an alignment of ancestral x^a and descendant x^d sequences separated by time t depends on the probability of states for each of the N sites. **B**) JC69 Substitution Matrix Q. **C**) A Simple Tree. Ancestral sequences are inferred at node i using descendant sequences at children nodes j and k connected by branch lengths t_j and t_k, respectively.

Codon models further incorporate the effects of non-synonymous (amino acid altering) and synonymous (silent) changes on protein structure and function. The non-synonymous/synonymous substitution rate ratio (ω) is predominantly used to assess the selective pressure for a protein, where $\omega>1$ and $\omega<1$ indicate diversifying (positive) and purifying (negative) selection, respectively. A neutral rate of codon evolution occurs when non-synonymous changes accumulate at an equal rate as synonymous changes, or $\omega=1$. Methods using counting techniques estimate ω for individual sites and optionally incorporate variation in the transition/transversion ratio (Li, Wu et al. 1985; Nei and Gojobori 1986; Comeron 1995; Ina 1995) and base/codon frequency bias (Yang and Nielsen 2000). These

methods conservatively estimate ω by assuming site independence and averaging ω over the entire sequence. Alternatively, sophisticated likelihood methods can be coupled with probabilistic models of phylogenetic relationships to incorporate the evolutionary history of a protein. Numerous codon models exist with variable specifications on substitution rates, including MG94 (Muse and Gaut 1994) and GY94 (Goldman and Yang 1994), which allow for position specific nucleotide substitution rates within the codon.

Several recent studies have found that the synonymous substitution rate is not neutral and can be influenced by factors such as exonic splicing enhancers, codon usage bias, isochores, translational rate or mRNA structure (reviewed in Chamary, Parmley et al. 2006). Kosakovsky Pond and Muse (2005) found that allowing separate non-synonymous and synonymous rates for each site more accurately represented their datasets 9 of 10 times. In a similar manner, Mayrose et al. (2007) allow for separate non-synonymous and synonymous rates and propose the instantaneous rate of changing from codon *i* to codon *j* be defined by:

$$Q_{ij} = \begin{cases} \lambda_s \cdot \kappa \cdot \pi_j & i \text{ and } j \text{ differ by one synonymous transition} \\ \lambda_s \cdot \pi_j & i \text{ and } j \text{ differ by one synonymous transversion} \\ \lambda_a \cdot \kappa \cdot \pi_j & i \text{ and } j \text{ differ by one non-synonymous transition} \\ \lambda_a \cdot \pi_j & i \text{ and } j \text{ differ by one non-synonymous transversion} \\ 0 & \text{otherwise} \end{cases}$$

where π_j is the target codon frequency calculated using the F3x4 model (Yang, Nielsen et al. 2000), κ is the transition/transversion ratio, and λ_s and λ_a are the synonymous and non-synonymous substitution rates. Substitution rate parameters can be optimized to maximize the likelihood of observed data and estimate the evolutionary distance between sequences.

Models can also allow for heterogeneous substitution rates among sites, which can be estimated from given a distribution, e.g. Poisson, Beta, Normal, or Gamma. While rate heterogeneity approximates a more biologically realistic model (Yang, Nielsen et al. 2000), it is computationally intensive and often requires discretization of a continuous distribution. Due to the pliability of the gamma distribution based on the shape ($\alpha > 0$) and scale ($\beta > 0$) parameters, it is frequently used to model substitution rate variation among sites (Yang 1994). However, the continuous gamma distribution does not have a closed form solution for

the likelihood. To circumvent this issue, a discrete gamma model can be approximated by *K* rate categories, each weighted by their probability of occurrence and represented by substitution rate r_k, a summary statistic for rate category *k* (Yang 1994; Stern and Pupko 2006). When $\alpha < 1$ the gamma distribution becomes skewed and has an L-shape, $\alpha \geq 1$ creates a $\cap$-shape, while $\alpha \rightarrow \infty$ reduces the model to a single rate across all sites. By estimating parameters α and β, the mean α/β and variance α/β^2 of the gamma distribution can be used to describe the distribution of synonymous λ_s and non-synonymous λ_a substitution rates. For example, when $\hat{\beta}_s = \hat{\alpha}_s$ the synonymous substitution rate $\lambda_s \sim \Gamma(\hat{\alpha}_s, \hat{\alpha}_s)$ will be neutrally evolving with mean 1 and variance $1/\hat{\alpha}_s$ while a non-synonymous substitution rate $\lambda_a \sim \Gamma(\hat{\alpha}_a, \hat{\beta}_a)$ will have a variable level of selection (Mayrose, Doron-Faigenboim et al. 2007). After maximizing the likelihood, rigorous statistical methods can test for significant trends in selection by comparing the estimated non-synonymous and synonymous substitution rates.

For large evolutionary distances, the use of nucleotide or codon sequence alignments may not be appropriate due to the possible saturation of substitutions. Instead heuristic amino acid models better reflect the probability of changes affecting protein structure and function (Halpern and Bruno 1998). Originally Dayhoff et al. empirically estimated the substitution rates among amino acids using observed changes in an alignment of orthologous proteins with at least 85% identity (Dayhoff, Schwartz et al. 1978). The symmetric 20x20 matrix was then standardized to create PAM1, which has one accepted point mutation per 100 amino acids. Assuming changes occur independently, repeatedly multiplying PAM1 creates matrices that reflect more divergent sequences such as PAM250 with an average of 2.5 changes per residue. Since only a few protein families were available during the initial calculation, a similar method has since been reapplied to create the updated JTT matrix based on a large set of protein databases (Jones, Taylor et al. 1992).

Henikoff and Henikoff created a similar series of BLOSUM (BLOcks of amino acid SUbstitution Matrix) matrices using the BLOCKS database of very conserved ungapped protein sequence alignments (Henikoff and Henikoff 1992). BLOSUM matrix values are

log-odds scores comparing the log ratio of likelihoods for residues in a biological alignment against those paired by random chance. The BLOSUM80 matrix was calculated by comparing sequences with at least 80% identity while BLOSUM45 used sequences with at least 45% identity. Hence lower valued BLOSUM matrices are more appropriate for estimating relationships for more divergent sequences.

Phylogenetics

The rate of sequence evolution varies extensively among genes and DNA segments. Representing the estimated evolutionary relationships of homologous sequences in a phylogenetic tree provides a useful tool for visualizing the patterns and processes of evolution (Felsenstein 1981). For example, cladogenesis can indicate speciation or duplication events, while branch lengths can signify the extent of divergence. However, a matrix of pairwise distance estimates does not necessarily form an exact or unique bifurcating tree due to violations of the triangle inequality (Beyer, Stein et al. 1974) and inaccuracies arising from the summarization of the phylogenetic information into a single distinct value. Exhaustively comparing all possible trees is infeasible since an alignment of just m=10 sequences will have $(2m-3)!!=(2m-3)!/2^{m-2}(m-2)!=34{,}459{,}425$ possible bifurcating, rooted trees (Felsenstein 1978). Stepwise algorithms and heuristic search methods are often applied to reduce computational cost at the expense of rigor. Hence determining proper tree reconstruction and inference methods is not straightforward because of the statistical and biological assumptions used to circumvent these issues.

Two main classes of methods are used for phylogenetic inference: distance and character based (Felsenstein 2004). Distance based methods like Unweighted Pair Group Method with Arithmetic mean (UPGMA) and Neighbor Joining (NJ) use pairwise comparisons of molecular sequences to form a distance matrix applied within a rule-based clustering algorithm and arrive at a single best tree. UPGMA (Sokal and Michener 1958) is based on the molecular clock hypothesis, which assumes all sequences evolve at a constant rate, and generally forms a poor tree reconstruction (Hollich, Milchert et al. 2005). In contrast, extensive tests found that the greedy, bottom up clustering method used in NJ frequently arrives at tree similar to the optimal solution (Huelsenbeck 1995). However, NJ

uses total tree length as the criterion for tree selection, and may poorly estimate large distances or trees based on alignments containing numerous gaps. A related method called BIONJ addresses this shortcoming in NJ trees by incorporating the approximate variance and covariance of distance estimates (Gascuel 1997). Likewise, WEIGHNJ uses an approximate likelihood criterion for joining nodes to make the method more robust and produce trees similar to the maximum likelihood method described below (Bruno, Socci et al. 2000).

Parsimony, Maximum Likelihood (ML) and Bayesian methods are all character based tree estimation procedures. These methods search the tree space using an objective function, fit each observed character (nucleotide, codon, or amino acid) at every site to a tree, and measure the fit of data to that tree. This procedure is very laborious and time consuming for a large number of sequences, as each iteration through the tree space requires a measure of fit calculation (Felsenstein 2004). Nonetheless, rich, statistical theory has been established to allow for comparison of estimates and likelihoods among trees.

ML and Bayesian methods easily incorporate complex models with realistic biological assumptions to provide a powerful and flexible framework for phylogenetic inference and hypothesis testing. The ML method fits data from a given alignment (D) to a tree topology (T) using a particular model of evolution by maximizing the likelihood function $L=P(D|\theta,T)$, where θ is the set of model parameters to be estimated. These parameters include the frequency of character states, branch lengths, substitution rates, and rate heterogeneity among sites (Durbin, Eddy et al. 1998). Assuming a general time reversible model where the rate of substitution from state x_i to x_j is the same as x_j to x_i and $\pi_i Q_{ij} = \pi_j Q_{ji}$, we can apply Felsenstein's pruning algorithm to quickly calculate the likelihood at each site (Felsenstein 1981). For node i with children nodes j and k connected by branch lengths t_j and t_k respectively (Figure 1c), define $L_i(x_i) = \sum_{x_j} p_{x_i x_j}(t_j) L_j(x_j) \times \sum_{x_k} p_{x_i x_k}(t_k) L_k(x_k)$ as the probability of observing the data at node i and its descendants, given its character is in state x_i. The total likelihood for site s is then $L^{(s)} = \sum_{x_i} \pi_{x_i} L_0(x_i)$ where L_0 is the likelihood at the root and π_{x_i} is the probability of state x_i at the root. When sites are treated independently, the

likelihood of the tree is simply the product of site likelihoods. Since this value is typically small and difficult to store computationally, the log likelihood is calculated instead where $\ell = \log(L) = \sum_s \log\left(L^{(s)}\right)$. A deficiency with the ML algorithm is that the common and biologically more realistic case of intercorrelated or dependent sites complicates the likelihood model (Yang 1995; Felsenstein and Churchill 1996; Mayrose, Doron-Faigenboim et al. 2007; Fernandes and Atchley 2008). This condition exponentially increases the size for substitution matrix Q and is rarely implemented due to the extremely cumbersome computation.

The Bayesian method of phylogenetic reconstruction similarly uses an explicit model of evolution to optimize a likelihood function. This procedure optimizes the posterior probability $P(\theta|D)$ using Bayes Theorem,

$$P(\theta \mid D) = P(D \mid \theta)P(\theta) \Big/ P(D) = P(D \mid \theta)P(\theta) \Big/ \int P(D \mid \theta)P(\theta)d\theta$$

where T is incorporated in the set of model parameters θ, D is as described above, $P(D|\theta)$ is the likelihood, $P(D)$ is the marginal probability of the data given a tree, and $P(\theta)$ is the prior distribution describing the tree parameters. A prior distribution allows for the incorporation of additional biological information during estimation and defines a confidence interval for the model parameters rather than a point estimate as in ML. When no information is available *a priori*, a noninformative prior is assumed and a uniform distribution is used. Arguably, a uniform distribution does not imply an uninformed prior and there is significant disagreement over the use and interpretation of prior distributions (Syversveen 1998). To complicate matters further, the marginal probability of the data in the denominator is difficult if not impossible to calculate. The Markov Chain Monte Carlo (MCMC) algorithm circumvents this issue by optimizing the posterior distribution (Metropolis, Rosenbluth et al. 1953; Hastings 1970).

In a non-periodic Markov Chain, each state represents a set of θ values with a stationary distribution of $P(\theta|D)$, assuming each state can reach any other state within a finite number of steps. The MCMC method approximates $P(\theta|D)$ by stochastically sampling a

large number of θ values from $P(\theta|D)$ under the expectation that θ values with a high posterior probability are more likely to be sampled than those with a low posterior probability. Thus running the Markov Chain for a sufficient period of time can approximate $P(\theta|D)$ by the probability of being at a particular state. Starting at time t=0 with parameter values $\theta^{(t)}$, the algorithm steps through the Markov Chain and randomly proposes a new set of parameter values θ^*. The Hastings algorithm accepts these new θ^* values with probability $min(1, r)$, where $r = \frac{P(D \mid \theta^*)P(\theta^*)J(\theta^{(t)} \mid \theta^*)}{P(D \mid \theta)P(\theta)J(\theta^* \mid \theta^{(t)})}$ and $J(\theta^{(t)} \mid \theta^*)$ is the probability of "jumping" from $\theta^{(t)}$ to θ^* (Hastings 1970). Current limitations to the MCMC Bayesian method arise from difficulties in assessing independence from initial parameter specifications, correlations in consecutive parameter values, and ensuring the process does not get trapped in local maxima. In spite of these difficulties, the MCMC Bayesian method provides a straightforward interpretation of probabilities and easily lends itself to biological models (Marjoram and Tavaré 2006).

Likelihoods produced by Bayesian and ML methods quantify the fit of data for a given tree and model of evolution. Statistical methods can then compare multiple likelihoods to identify the model of better fit. The likelihood ratio statistic, $LRT = -2\left(L_1 \big/ L_2\right)$, compares likelihoods for two nested models to test if the likelihood of a model with more parameters, L_2, is significantly better than that with one or more fixed parameters, L_1 (Casella and Berger 2002). The Chi-squared distribution can then be used with df_1-df_2 degrees of freedom to compare the fit of models. However, by definition $L_2 \geq L_1$ and the increase in parameters may cause overfitting. The Akaike information criterion (AIC) and Bayes information criterion (BIC) measure the goodness of fit for model selection while penalizing overparameterization (Akaike 1974; Schwarz 1978). Lower AIC or BIC values indicate the preferred model. In general, $AIC = 2k - 2\ln(L)$ with k estimated parameters and maximized likelihood L for the given model. BIC imposes a stricter penalty than AIC for including additional parameters and is defined by $BIC = -2\ln\left(p(x \mid k)\right) = -2\ln(L) + k\ln(n)$, where data x is in the exponential family with n observations. An overview of some software (PAML, Mr

Bayes, MEGA, Phylip, PhyML, BioNJ, Weighbor) for constructing and testing phylogenies is available in Chapter 3.

High Dimensional Molecular Data

Some methods of identifying selection focus on statistically quantifying protein property variability. Many such tests require extensive replication for sufficient power to determine significance, and most exploratory techniques like genome sequencing and microarrays inundate biologists with massive amounts of high dimensional molecular data (HDMD). These huge data sets lead to major statistical issues since HDMD typically have many more variables than observations. This condition, often called "the curse of dimensionality" (Bellman 1961), causes various multivariate statistical analyses to fail, become intractable, or produce misleading results (Donoho 2000; Ransohoff 2005; Clarke, Ressom et al. 2008). Further, there are serious problems with procedures used to effectively reduce the dimensionality of HDMD while retaining relevant biological information. Widely used multivariate techniques like cluster analysis (MacQueen 1966; Overbeek, Fonstein et al. 1999; Tan, Steinbach et al. 2006), principal component analysis (PCA) (Jolliffe 2002; Johnson and Wichern 2007), factor analysis (FA) (L. Gorsuch 1983), and discriminant analysis (DA) (J. Huberty 1994) have underlying assumptions that are often not properly addressed.

In Chapter 3 we employ FA and DA to identify variability due to selection in the DNA binding and dimerization domain of Max and Mlx proteins. While that chapter includes a brief summary and purpose in using these methods, we provide Appendix A to further discuss a number of problems pertaining to HDMD and the relative merits of these and various multivariate statistical procedures.

Max and Mlx Networks

To identify functional variation from multiple sequence alignments, in Chapters 3 and 4 we focus on two related protein-protein interaction networks that are involved with cell maintenance, growth, and metabolism. For simplicity, we will refer to the networks as the Max and Mlx networks due to the central and pivotal role of these proteins.

The Max network of interacting transcription factors regulates cell cycle progression via coordinated activation or repression of genes involved in ribosome biogenesis, growth, proliferation, differentiation, and apoptosis (reviewed in Lüscher 2001). Antagonistic dimerization of Myc and Mxd/Mnt to obligate partner Max governs transcriptional activation and cell proliferation or transcriptional repression and cell differentiation, respectively. Similarly, repressors Mxd/Mnt and transcriptional activator Mondo competitively and antagonistically dimerize with Max-like protein Mlx in the parallel Mlx network, which is implicated in cellular proliferation, growth, energy homeostasis, and metabolism (Billin and Ayer 2006).

Implications in Disease

Deregulation of the Max and Mlx networks leads to improper cell growth, unbalanced energy homeostasis, and possibly death (Lüscher 2001; Billin and Ayer 2006). Myc family member c-Myc has long been known as a potent proto-oncogene that is necessary for cell growth in mammals, yet often deregulated in actively growing tumors (Eilers and Eisenman 2008). Although not properly identified as a tumor suppressor protein, Mnt antagonizes c-Myc function and can also promote tumor growth when conditionally deleted (Hooker and Hurlin 2006). Since Mnt, c-Myc and Max have complex function and are essential for mammalian viability, using them as drug targets in cancer treatments has proved challenging (Freie and Eisenman 2008). Similarly, MondoB is implicated in cell growth and particularly in glucose metabolism (Yamashita, Takenoshita et al. 2001). As a major factor in *de novo* lipogenesis (Ma, Robinson et al. 2006), MondoB has been linked to metabolic syndrome and type II diabetes (Dentin, Benhamed et al. 2006; Denechaud, Dentin et al. 2008; Iizuka and Horikawa 2008). While MondoB knockouts in rodents are viable, their promising attenuation of insulin resistance is at the expense of increased blood glycogen levels (Towle 2005; da Silva Xavier, Rutter et al. 2006; Uyeda and Repa 2006). Considering the prevalence of cancer and type II diabetes, understanding the general function of these networks can provide great insight to the penetrability and treatment of these diseases.

Known Max and Mlx Network Configurations

Less complex organisms can often act as adequate surrogates for annotating proteins as well as protein networks. To date, three main Max and Mlx network configurations have been documented in vertebrates, flies, and nematodes (Figure 2).

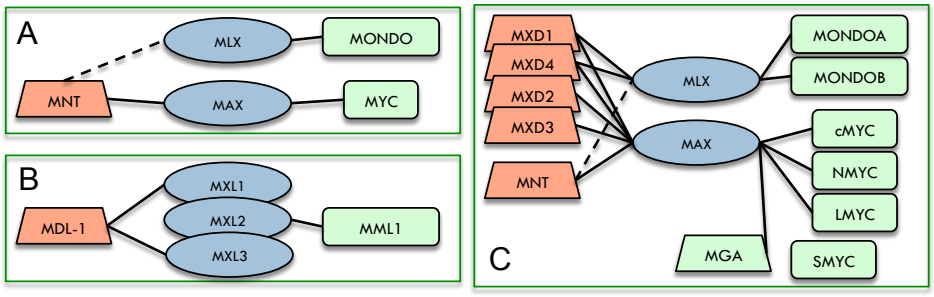


Figure 2: Known Max and Mlx Network Configurations
A) *Drosophila melanogaster*, B) *Caenorhabditis elegans*, C) *Mus musculus*. Red trapezoids represent repressor proteins, green rectangles activating proteins, and blue ovals central interaction partners. Solid lines indicate known bHLHZ interactions, while dotted lines have not been firmly established.

The vertebrate network has paralogous families for Myc (c-, L-, and N-Myc), Mxd (Mxd1-4, formerly Mad1, Mxi1, Mad3, and Mad4) and Mondo (MondoA and MondoB), along with single copies of Max, Mlx, Mnt, and Mga genes. In addition, rodents have experienced lineage specific duplications to create S-Myc and b-Myc, while the human network contains a separate duplication denoted L-Myc2 (Lüscher 2001). For the purposes of this discussion, b-Myc in rodents is not included within the Max network since it lacks a bHLHZ domain (Burton, Mattila et al. 2006). Mediated by the highly conserved bHLHZ domain, Max can homodimerize with itself and heterodimerize with Mnt, Mga, and members of the Myc and Mxd families (Hurlin and Huang 2006). The Mlx bHLHZ domain exhibits a more restrictive binding pattern and interacts with only Mxd1 and Mxd4 of the Mxd family in addition to itself, MondoA, and MondoB (Billin and Ayer 2006).

Drosophila melanogaster contains a simpler network with single genes coding for dMyc, dMax, dMlx, dMnt, and dMondo (Gallant 2006). Vertebrate c-Myc is considered a direct ortholog of dMyc due to their similar cellular expression and reciprocal ability to rescue dMyc$^{-/-}$ *Drosophila* mutants and c-Myc null murine embryo fibroblasts (MEFs) (Gallant, Shiio et al. 1996; Delacova and Johnston 2006). As before, dMax can dimerize

with dMax, dMyc and dMnt, while dMlx binds to dMondo. Interestingly, the loss of Mxd in flies makes dMnt the only repressor within this network.

Caenorhabditis elegans has a markedly different yet clearly orthologous network, presumably due to massive gene reduction and rearrangement in nematodes (Witherspoon and Robertson 2003; Denver, Morris et al. 2004; Coghlan 2005). Notably, *C. elegans* lacks both Mnt and Myc orthologs. Instead, Myc and Mondo-like protein MML-1 and Mad-like ortholog MDL-1 alone comprise the activator and repressor components of the network, respectively, while two Max orthologs (Mxl-1 and Mxl-3) and a single Mlx ortholog (Mxl-2) act as central dimerization partners (Yuan, Tirabassi et al. 1998). Mxl-2 dimerizes to transcriptional activator MML-1, while MDL-1 dimerizes with either Mxl-1 or Mxl-3 to repress transcription. Surprisingly, Mxl-1 cannot homodimerize and does not interact with mouse c-Myc.

As the defining factor for Max and Mlx network members, the bHLHZ domain is integral for their protein-protein and protein-DNA interactions (Nair and Burley 2006; Maerkl and Quake 2009). Dimerization of obligate partners through the bHLHZ domain creates a four helix amphipathic bundle for which several structures have been determined and are available in the protein databank (PDB): Max:Max heterodimer (1R05), Max:Max heterodimer recognizing DNA (1AN2, 1HLO), Max:Mxd1 heterodimer recognizing DNA (1NLW), Max:Myc heterodimer/heterotetramer recognizing DNA (1NKP), and Max:c-Myc leucine zipper (1A93, 2A93) (Ferré-D'Amaré, Prendergast et al. 1993; Brownlie, Ceska et al. 1997; Lavigne, Crump et al. 1998; Nair and Burley 2003; Sauvé, Tremblay et al. 2004).

Figure 3: Myc:Max and Mxd:Max bHLHZ Structures
bHLHZ structure images taken from (Nair and Burley 2003). PDB IDs are in parentheses. **A)** Human cMyc:Max heterodimer recognizing DNA (1NKP). **B)** Human Mxd1:Max heterodimer recognizing DNA (1NLW). **C)** Human cMyc:Max heterotetramer recognizing DNA (1NKP).

Max network proteins are known to bind to the E-box motif 5'-CACGTG-3' (Blackwood and Eisenman 1991; Blackwell, Huang et al. 1993), while MondoA, MondoB, and Mlx bind a ChORE motif characterized by two E-boxes separated by exactly five nucleotides, CAYGNGN$_5$CNCRTG (Ma, Robinson et al. 2006). Due to the importance of the bHLHZ domain in these essential and potentially proto-oncogenic proteins, we examined the sequence conservation and functional characteristics of the bHLHZ domain among network members, and present our results in Chapter 3.

Sequence variability in regions surrounding the conserved bHLHZ domain and presence of other functionally disparate domains suggests these proteins arose through ancient domain shuffling events, Figure 4 (Morgenstern and Atchley 1999; Jones 2004). The N-terminus of activators Myc, Mondo, and MML-1 contain a transactivation domain (TAD), while repressors Mnt, Mxd, and MDL-1 have a Sin3 interaction domain (SID) (Lüscher 2001; Hurlin and Huang 2006). Myc's TAD associates with coactivator complex TRRAP/GCN5 and the subsequent Myc/Max/TRRAP/GCN5 active complex binds to the DNA E-box motif, acetylates surrounding histones, and permits transcriptional activation of gene targets. In contrast, Mnt, Mxd, and MDL-1 recruit the Sin3/HDAC/N-Cor/Ski/Sno complex to its SID, which in turn deacetylates histones and represses transcription. Max and Mlx lack any other known functional domain and therefore simply serve as necessary coactivators.

As transcription factors, these proteins are capable of regulating multiple genes within several signaling pathways. Dynamic expression patterns, competitive dimerization, and overlapping gene targets of Max network members create a complex and essential system controlling cell fate. In particular, coordinated expression of Myc and Mxd given background levels of Max and Mnt mediate the transition between states in the cell cycle (Walker, Zhou et al. 2005).

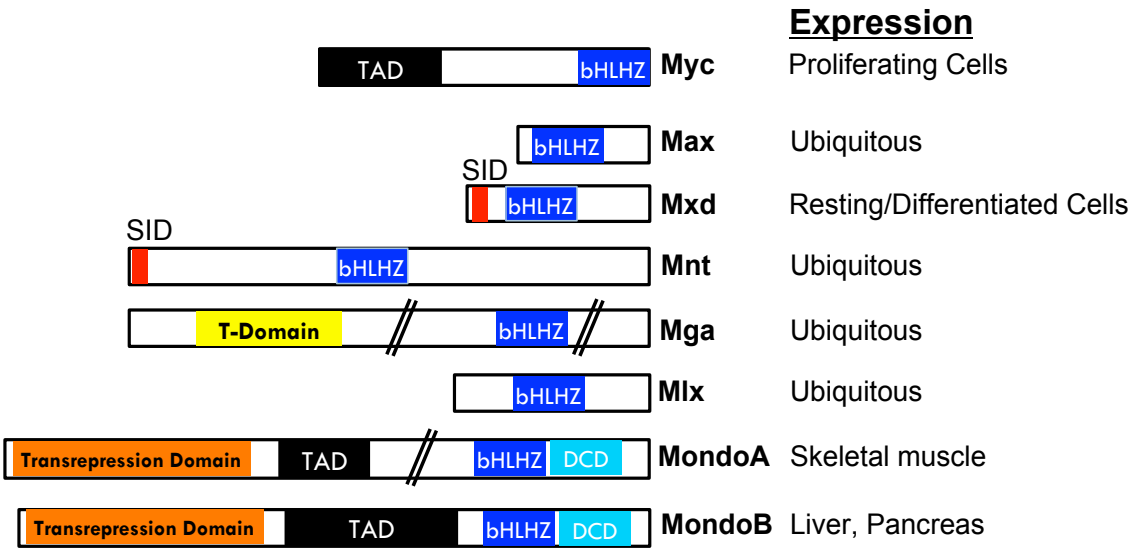


Figure 4: Max and Mlx Network Member Domains
Modified from (Lüscher 2001). TAD: transactivation domain, bHLHZ: basic-Helix-Loop-Helix-Leucine Zipper domain, Transrepression Domain: region involved in repression, SID: mSin3-interaction domain, T-Domain: T-box DNA binding domain, DCD: dimerization and cytoplasmic localization domain.

Myc and Mxd affect Cell Cycle Progression

The cell cycle is responsible for cell duplication and division and is defined by the resting phase (G0), two gap or growth phases (G1, G2), a DNA synthesis phase (S), and mitosis (M). Quiescent or differentiated cells are in resting (G0) phase of the cell cycle, while proliferating cells cycle through G1, S, G2, and M.

In normal cells, mitogen-induced expression by c-Myc is sufficient to initiate cell cycle entry in otherwise quiescent or differentiated cells endogenously expressing Mxd1, Mxd2, and Mxd4 (Eilers, Schirm et al. 1991; Walker, Zhou et al. 2005). c-Myc accumulates quickly during G1 phase and displaces Mnt as the primary Max dimer to activate key mitosis checkpoint proteins like cdk4, CycD, CycE, and Rbf (Hooker and Hurlin 2006). Ubiquitin-mediated proteolysis degrades existing c-Myc rapidly, while a negative feedback loop prevents further protein production. This subsequent degradation of c-Myc is important to protect against tumorigenesis and increased sensitivity to apoptosis as high levels of c-Myc during S phase are linked to an increased susceptibility to mutation (Felsher and Bishop

1999). c-Myc levels return to a low, steady state by S phase, whereby Mxd3 accumulates, peaks and diminishes again.

The coordinated oscillation of c-Myc and Mxd expression during cell cycle transitions of proliferative/quiescent states and G1/S phases suggests these proteins act as an on/off switch governing transcriptional activation or repression of genes specific to cell cycle progression. Such transitioning between Max heterodimerization partners depends on the relative expression levels, complex stability, subcellular location, and protein degradation of Myc, Mxd, and Mnt proteins. Max levels are in excess compared to its binding partners and competition may only occur during peak phases of expression for its interacting proteins (Lüscher 2001).

The following addresses the individual and coordinated functions for Max network components, with regard to their role in cell cycle regulation and transformation.

Max is a central dimerization partner

Max is a small essential protein with ubiquitous and cell autonomous expression and is maintained at constant, stable levels with a half-life of >24hours (Blackwood et al, 1992). An 'RKKLR' motif located near the bHLHZ domain serves as a nuclear localization signal (NLS) that transports Max to the nucleoplasm where it is uniformly distributed (Grinberg, Hu et al. 2004). In the absence of other interacting proteins, Max readily forms Max:Max homodimers that bind to E-box motifs and weakly decrease transcription by ≈2-fold (Billin, Eilers et al. 1999; Mcduff, Naud et al. 2009). The function of DNA-bound Max:Max is unknown, since Max has no discernable phenotype and contains no other known domain. Instead Max may mark genes poised for transcriptional regulation, possibly binding to DNA as a monomer prior to dimerization. As such, Max function depends on the conditional expression, cellular location, and relative binding affinity of the Myc, Mnt, Mxd, and Mga interacting proteins described below.

Myc is a potent proto-oncogene

Myc (identified as *v-myc*) was first isolated as the transforming factor in chicken retroviruses over 30 years ago and has since been identified as a potent proto-oncogene (Sheiness, Fanshier et al. 1978; Dang 1999).

Myc abundance is tightly regulated due to its essential role in growth and is typically expressed only at low levels in proliferating cells (Hooker and Hurlin 2006). That said, Myc deregulation is associated with most if not all cancers, attributing to around 70,000 deaths a year (myccancergene.org). Translocation and amplification, activation of growth signals, and increased protein stability contribute to its deregulation (Spencer and Groudine 1991). The transformation behavior stems from Myc's ability to induce proliferation through cell cycle entry, block cell cycle exit, and sensitize cells to apoptosis in a manner dependent upon cell type and physiological status (Grandori, Cowley et al. 2000; Nasi, Ciarapica et al. 2001). Although aberrant overexpression of Myc does not accelerate cell division, it can prevent cell cycle exit with a correspondingly marked increase in cell and consequently body size (Johnston, Prober et al. 1999).

In a normal cell, overabundance of Myc triggers a Myc induced apoptotic pathway in both a p53-dependent and -independent fashion as a failsafe mechanism to limit excessive growth and control tumor progression (Hoffman and Liebermann 2008). Frequently, cells with overexpressed Myc have mutations that disrupt this pathway, allowing survival, exacerbating proliferation, and further promoting transformation (Grandori, Cowley et al. 2000; Hooker and Hurlin 2006). Ectopic upregulation of Myc can also disrupt normal cell function and result in tumor formation, e.g. just a 1.47 fold increase of c-Myc was observed in certain cases of Burkitt's Lymphoma (Sáez, Artiga et al. 2003). Moreover, Myc can induce cell competition, where neighboring cells with significantly lower levels of Myc are signaled for apoptosis (Johnston, Prober et al. 1999; Secombe, Pierce et al. 2004). Thus cells experiencing Myc deregulation can increase cell growth and demand for more nutrients, survive longer when the apoptotic pathway is disrupted, and deplete surrounding healthy cells.

Artificially regulating Myc is problematic since the upstream and downstream signaling cascades surrounding Myc expression are complex and undefined (Orian, van Steensel et al. 2003; Cole and Nikiforov 2006; Lee and Dang 2006; Knoepfler 2007). Myc null mutants are also difficult to examine since Myc deletion causes early embryonic lethality. *Drosophila* dMyc$^{-/-}$ mutants die within the second instar (Pierce, Yost et al. 2004),

mouse c-Myc$^{-/-}$ mutants arrest development at day 9.5 postcoitum (Davis, Wims et al. 1993), and mice with N-Myc$^{-/-}$ mutants fail to develop heart, lungs, and a nervous system (Charron, Malynn et al. 1992). Curiously, L-Myc$^{-/-}$ mutant mice are viable and lack any major defects (Hatton, Mahon et al. 1996). In comparison, c-Myc hypomorphs have a normal cell size, although the body size is smaller, development is delayed, and females are sterile due to a defect in oogenesis (Gallant, Shiio et al. 1996; Gallant 2006).

Supporting Myc's role in proliferation, differentiation, and growth, Myc is shown to target genes involved in metabolism (CAD, ornithine decarboxylase, Lactate dehydrogenase A, dihydrofolate reductase), cell cycle progression (cyclinA, cyclinD2, cyclinE, cdc25A, telomerase genes), and ribosome biogenesis (tRNAs and 5S rRNA) (Hooker and Hurlin 2006). However, recent genome scans show Myc can bind globally, spanning ~15% of both the *Drosophila* and human genome in both inter- and intra-genic regions (Orian, van Steensel et al. 2003). Thus Myc can affect hundreds to thousands of genes, albeit weakly.

To bind DNA and transactivate genes, as well as interact with Inr elements such as Miz-1, Myc must first dimerize with Max (Seoane, Pouponnot et al. 2001; Staller, Peukert et al. 2001; Herold, Wanzel et al. 2002). However, recent evidence in *Drosophila* indicates Myc can also function in a Max independent manner. Experiments with *Drosophila* dMax$^{-/-}$ mutants found dMyc can induce biological processes such as cell autonomous death, endoreplication of polyploid larval cells, cell competition, and regulation of cell growth (Steiger, Furrer et al. 2008). Specifically, activation of RNA polymerase II and Myc induced apoptosis, by and large, do not require Max dimerization. Nonetheless, increasing Max expression in response to Myc overexpression can reduce cell transformation, possibly preventing Myc from binding to other interaction partners such as those involved in RNA polymerase formation (Mäkelä, Koskinen et al. 1992).

In the Max network, Mnt and Mxd family members antagonize Myc transactivation ability by competitively dimerizing with Max and binding to overlapping gene targets. Although Myc is clearly an important and interesting protein, understanding its role in cell cycle progression, proliferation, and cell growth ultimately relies on its relationship within the network.

Mnt represses cell growth

Myc and Mnt exhibit opposing roles in regulating cell growth, with similar implications in tumor formation when deregulated. Mnt overexpression results in markedly reduced cell and body size, while conditional deletion of Mnt leads to increased cell size, tumor formation, disruption in T cell development, and in general phenocopies c-Myc overexpression (Hurlin, Zhou et al. 2004; Nilsson and Cleveland 2004; Loo, Secombe et al. 2005; Walker, Zhou et al. 2005; Wahlström and Henriksson 2007). However, loss of Mnt function is not equivalent to aberrant Myc expression, since Myc abundance is still under several other regulatory mechanisms (Hooker and Hurlin 2006). Instead, Mnt depletion may increase the expression level and sensitivity of target genes to Myc activation, as seen during cell cycle progression (Walker, Zhou et al. 2005). Correspondingly, tumor formation resulting from conditional Mnt deletion is less penetrant and takes longer than ectopic overexpression of Myc (Hurlin, Zhou et al. 2003). Despite these tumor suppressor properties, cancer cells rarely have mutations in Mnt (Sommer, Waha et al. 1999).

Mnt and Myc's reciprocal roles in controlling cell growth suggest they share overlapping gene targets. As expected, dMnt and dMyc have both unique and overlapping binding sites (Orian, van Steensel et al. 2003). Since Mnt is ubiquitously expressed in proliferating, quiescent, and differentiating cells, while Myc expression is limited to proliferating cells, Mnt inhibits Myc transactivation as well as suppresses Myc-dependent transformation (Zhou and Hurlin 2001; Hurlin, Zhou et al. 2003). Hence, Mnt can act as a general repressor of Myc to control cell progression, even though Mnt expression is independent of the cell cycle.

The balance between Mnt and Myc function has been investigated through combinatorial deletions in *Drosophila* and mouse models, with surprising results. Mnt null mice die within a day after birth due to poor lung development and impaired lung function (Toyo-Oka, Hirotsune et al. 2004; Hurlin and Huang 2006), while *Drosophila* dMnt$^{-/-}$ mutants have increased cell size, a shortened lifespan and are viable (Delacova and Johnston 2006). Moreover mouse c-Myc$^{-/-}$/Mnt$^{-/-}$ mutants are not viable, although MEFs with simultaneous deletion of both c-Myc and Mnt continue to proliferate at a reduced rate

(Walker, Zhou et al. 2005; Hooker and Hurlin 2006). This is in stark contrast to *Drosophila* dMyc$^{-/-}$/dMnt$^{-/-}$ mutants, which grow significantly larger and with increased viability over single mutant dMyc$^{-/-}$ as a result of increased endoreplication and growth of larval tissues and imaginal discs (Pierce, Yost et al. 2008). Since Myc relieves Mnt repression in both mouse and *Drosophila* models, the discordance of these findings suggests a divergence in network dynamics between species.

As previously mentioned, dMnt is the only repressor in the *Drosophila* Max network, while Mnt and Mxd1-4 repress Myc family proteins in vertebrates. Also, in contrast to the ubiquitous expression of vertebrate Mnt, dMnt has a dynamic expression pattern temporally dependent upon cell type, where it is present in actively replicating mitotic and endoreplicating tissue as well as differentiating cells (Loo, Secombe et al. 2005). Hence *Drosophila* has adopted a distinct balance among Max network members that does not necessarily require Mnt or Mxd proteins. Similarly, *C. elegans* does not have a Mnt ortholog, but rather contains a Mxd-like ortholog, MDL-1, involved in post-embryonic development and suppression of transformation (Yuan, Tirabassi et al. 1998).

Mxd proteins have dynamic patterns of repression in vertebrates

Like Mnt, Mxd proteins repress Myc and can suppress c-Myc-dependent cell transformation (Lahoz, Xu et al. 1994; Koskinen, Ayer et al. 1995; Västrik, Kaipainen et al. 1995; Zhou and Hurlin 2001). However, the Mxd family of proteins in vertebrates exhibits a dynamic, yet distinct expression pattern that may mediate specialized roles in Myc repression (Hooker and Hurlin 2006).

Mxd1-4 expression levels correlate with cell cycle transitions, where their upregulation corresponds to downregulation of Myc levels. Although there is extensive overlap in expression of all Mxd proteins during early mouse embryogenesis, Mxd1 and Mxd4 are most prevalent in quiescent or terminally differentiated cells (Hurlin, Quéva et al. 1995; Västrik, Kaipainen et al. 1995; Quéva, Hurlin et al. 1998). Mxd2 and Mxd3, unlike the other Mxd proteins, are also expressed early during differentiation or in proliferating adult cells (Zervos, Gyuris et al. 1993; Larsson, Pettersson et al. 1994), while the varying

expression of Mxd3 peaks during S phase of the cell cycle (Hurlin, Quéva et al. 1995; Quéva, Hurlin et al. 1998; Fox and Wright 2001; Quéva, McArthur et al. 2001).

When bound to Max, Mxd proteins repress transcription of gene targets by ≈4 fold and mediate cell cycle checkpoints (Billin, Eilers et al. 1999). In particular, Mxd1 represses Myc targets cyclin D2 and human telomerase reverse transcriptase (hTERT), which are important for the transition into and progression of S phase (Günes, Lichtsteiner et al. 2000; Bouchard, Dittrich et al. 2001; Xu, Popov et al. 2001). Mxd3 amplification and c-Myc depletion during S phase further suggests Mxd proteins affect cell cycle progression through phase-specific regulation (Hooker and Hurlin 2006). The reciprocal relationship between Mxd and Myc reflects a switch from activation to repression of gene targets, associated with a transition from cell proliferation to differentiation (Ayer and Eisenman 1993). Due to their short (15-20 minute) half-life, Myc and Mxd families yield a steady turnover rate that enables the network to readily respond to external stimuli and transition between cell states (Blackwood, Lüscher et al. 1992; Amati and Land 1994; Sears 2004; Adhikary and Eilers 2005).

Regulation of cell cycle progression and Mxd mediated antagonism of Myc suggested this family could contain potential tumor suppressor proteins. Accordingly, deletion or Loss of heterozygosity (LOH) in a region containing Mxd2 is associated with tumors including astrocytoma, glioblastoma, lymphocytic leukemia, prostate adenocarcinoma, malignant melanoma, small cell and squamous cell carcinomas of the lung (Edelhoff, Ayer et al. 1994; Chen, Willingham et al. 1995; Eagle, Yin et al. 1995; Foley and Eisenman 1999; Engstrom, Youkilis et al. 2004). However, alterations of Mxd2 were found in only a subpopulation of cells for each tumor and a nearby tumor suppressor gene PTEN/MMAC1 may be responsible (Li, Yen et al. 1997). Similarly, there is no clear indication that Mxd1, Mxd3, or Mxd4 are consistently altered in tumors (Schreiber-Agus and Depinho 1998; Baudino and Cleveland 2001).

Still, Mxd knockouts in mice exhibit a modest increase in sensitivity to tumorigenesis. Mice with Mxd1, or Mxd3 deleted had minimal abnormalities, while $Mxd2^{-/-}$ mice had a slight predisposition to tumorigenesis although all were viable, fertile and

outwardly healthy (Schreiber-Agus and Depinho 1998; Foley and Eisenman 1999; Rottmann and Lüscher 2006). Mxd3$^{-/-}$ mice are healthy, with no apparent defect in cell cycle entry or exit, although thymocytes and neuronal cells are sensitive to radiation-induced apoptosis (Quéva, McArthur et al. 2001). Mxd1$^{-/-}$ and Mxd3$^{-/-}$ mice were also more sensitive to granulocyte differentiation (Foley, McArthur et al. 1998). Surprisingly, simultaneous deletion of Mxd1, Mxd2, and Mxd3 produced mice that were fertile, viable, and 15-20% larger than controls, with a hyperproliferative phenotype in thymocytes, and splenic B and T cells (Rottmann and Lüscher 2006). Since Mxd4 knockouts have not been investigated, it is unknown if Mxd4 is compensating for the triple mutant knockout or if Mxd proteins are dispensable as a whole.

Evidence of protein compensation among Mxd family members suggests a restrictive homeostatic feedback mechanism of upregulation in response to individual gene loss (Rottmann and Lüscher 2006). While other Mxd levels do not change in Mxd2 or Mxd3 knockout mice, Mxd2 and Mxd3 expression is upregulated when Mxd1 is removed (Ayer, Kretzner et al. 1993; Quéva, McArthur et al. 2001). There may also be functional compensation/complementation from proteins outside the Max network, as seen by the synergism between Mxd2 and cki p27 proteins in promoting cell cycle exit during differentiation (McArthur, Foley et al. 2002).

Max and Mlx Networks Overlap

Mxd1 and Mxd4 have also been shown to dimerize to Mlx, thus linking the Max and Mlx networks in vertebrates (Billin, Eilers et al. 1999; Meroni, Cairo et al. 2000). Interestingly, Mlx is a predominantly cytoplasmic protein, and the subcellular location of these interactions is unknown. While Mxd1 is nuclear, Mxd4 contains a necessary and sufficient N-terminus nuclear export signal (NES) that relocates it to the cytoplasm when expressed alone (Grinberg, Hu et al. 2004). However, mutations in the C-terminus of Max and N-terminus of Mxd4 verified that the NLS of Max overcomes the NES of Mxd4 and mediates their relocation to the nucleus. Mxd2 and Mxd3 relocation to nuclear foci is independent of Max, suggesting they may contain a NLS or CES that limits their interaction with Mlx and prevents heterodimerization.

Immunoprecipitation from cell lysates showed that Mxd1 preferentially dimerizes with Max over Mlx, while two hybrid and gel shift assays using recombinant proteins show similar binding affinities (Billin and Ayer 2006). This indicates post-translational modification moderates the cellular location of these proteins and the preferential binding partner of Mxd1. The preferential binding of Mxd4 to Max or Mlx is unknown.

Mnt has also been found to dimerize with Mlx, but this interaction has not been firmly established. Mnt:Mlx and Mlx:Mlx heterodimers have been isolated (Meroni, Cairo et al. 2000) and Mnt was identified as a Mlx interactor when isolating MondoB:Mlx complexes (Cairo, Merla et al. 2001), yet no further experiments have validated the presence of Mnt:Mlx. In opposition, Mnt:Mlx or Mlx:Mlx interactions were not found (Billin, Eilers et al. 1999) and Mnt:Mlx interaction *in vitro* or *in vivo* has not been observed (Hurlin and Huang 2006). Although Mnt:Mlx repression on E-box genes was established (Meroni, Cairo et al. 2000), the level of repression was not significantly different upon addition of Mlx. This suggests the observed repression occurs only under particular cellular conditions and could be mediated by either endogenous Max or Mlx levels or Mlx:Mnt dimerization. Moreover, combinatorial knockouts of *Drosophila* dMyc (dm^4), dMnt (Mnt^1) and dMax (Max^1), show dm^4Mnt^1, dm^4Max^1, and $dm^4Mnt^1Max^1$ mutants have strong similarity in phenotype (Pierce, Yost et al. 2008). This implies that both dMax and dMnt do not interact with other partners including dMlx or dMondo, at least for the cell type and condition assayed.

Mga function and origin is unknown

Along with Mnt and Mxd family of repressors and Myc family of activators, a novel Max-interacting protein Mga has recently been discovered in vertebrates (Hurlin, Steingrìmsson et al. 1999). Appropriately named, Mga was isolated within mouse as a 14 kb, 3006 amino acid protein with 25 exons and two dimerization domains. In addition to its bHLHZ domain 39% similar to c-Myc in humans, Mga also has a N-terminal T-box domain, which lacks the canonical exon structure conserved among T-box family proteins (Lardelli 2003; Minguillon and Logan 2003). The intronless T-box domain of Mga suggests it was inserted via a reverse transcription event, coupling it with a Max interacting bHLHZ domain.

The unusual coupling of multiple DNA binding domains complicates the functional role Mga.

Mga acts as either a repressor or activator of reporter genes in a Max dependent manner (Hurlin, Steingrìmsson et al. 1999). Mga:Max activates transcription when the canonical E-box motif, T-domain, or both are present. Mga:MaxΔBR also activates T-domain reporters, where MaxΔBR is a dominant-negative repressor that lacks the basic region and cannot bind to the E-box. In the absence of endogenous Max, Mga binds and subsequently represses a T-domain containing reporter. This implies that Mga:Max dimerization through the bHLHZ domain not only permits binding to the E-box motif, but may also block a bHLHZ dependent co-repressor for T-domain targets. Thus Mga acts as a dual specific transcriptional regulator, where activation is dependent upon Max dosage.

From yeast two hybrid assays, Mga was identified in mouse embryonic cDNA libraries at days E9.5 and E10.5 as well as murine kidney cells at E14.5 with highest levels found in limb buds, branchial arches and the tail region (Hurlin, Steingrìmsson et al. 1999). Mga was not isolated in other cell types or adult tissues using Northern blot assay, except in rat pheochromocytoma cell line PC12 and C2C12 myoblasts. The overlapping expression of Mga with other T-box proteins such as Brachyury, Tbx2, 3, 4, and 5 during embryonic development suggests that it also regulates regions involved in mesoderm and mesodermal-epithelial interactions during development. However, little is still known about this large protein including the possible identification of other domains, role of alternative splicing, subcellular location, and cross talk between the bHLHZ and T-box domains.

The Mlx Network

In general, Max network members control cell growth by regulating target genes involved in proliferation, differentiation and cell cycle progression. Similarly, the parallel Mlx network of transcription factors affects cell growth by targeting genes involved in glucose metabolism.

Balancing glucose use and storage is critical for proper energy homeostasis. Mammals store sugars such as glucose in the form of glycogen or triglycerides for immediate or compact energy storage, respectively. When food is scarce, glycogen and triglycerides

can be decomposed into glucose and processed through glycolysis and the citric acid cycle (TCA) for energy use. Upon food intake, increased glycolitic flux stimulates insulin secretion from pancreatic islets (β-cells) and activates the glycogen synthesis or *de novo* lipogenesis pathways in the liver to promote energy storage (Postic, Dentin et al. 2007).

In mammals, MondoB expression in the liver promotes triglyceride and lipid formation in response to excess carbohydrates by activating genes involved in *de novo* lipogenesis (Ma, Robinson et al. 2006). Vertebrate paralog MondoA also affects energy homeostasis in skeletal muscle, where its expression is negatively correlated with glucose uptake (Billin, Eilers et al. 2000; Sans, Satterwhite et al. 2006; Stoltzman, Peterson et al. 2008). As biosensors to intracellular glucose levels, unraveling the role of MondoA:Mlx and MondoB:Mlx heterodimers in glycolysis will enhance our understanding of glucose associated diseases such as type II diabetes and metabolic syndrome (Burgess, Iizuka et al. 2008; Denechaud, Dentin et al. 2008; Iizuka and Horikawa 2008; Stoltzman, Peterson et al. 2008; Sears, Hsiao et al. 2009).

Defects and Disease

Diabetes is a disease characterized by high blood glucose levels due to either a failure to produce insulin (Type I) or insulin resistance (Type II), where cells do not properly respond to the insulin produced. A distinctive feature for both forms of diabetes is loss or dysfunction of pancreatic islets, which are responsible for insulin secretion and signaling (Noordeen, Khera et al. 2010). Consequently, glucose is not properly absorbed or stored in muscle or liver tissues, respectively, resulting in higher glucose levels in the bloodstream.

Alterations in mitochondrial activity within skeletal muscle also play a decisive role in the physiopathology of insulin resistance (Postic, Dentin et al. 2007; Zorzano, Liesa et al. 2009). Impaired fat oxidation, designated by the reduced ability of muscle to properly oxidize glucose and lipids during fasting and fasting-to-fed conditions is a main feature of insulin resistance (Kelley, Goodpaster et al. 1999). Obese and Type II diabetics have a higher capacity for lipid oxidation than switching to glucose oxidation, exhibiting "metabolic inflexibility" (Felber, Ferrannini et al. 1987; Kelley and Mandarino 2000; Storlien, Oakes et al. 2004; Zorzano, Liesa et al. 2009). Hence glucose is not fully processed, leading to a

reduction in both TCA activity and respiratory chain (Kelley, He et al. 2002), with increased lactate levels (Consoli, Nurjhan et al. 1990; Cusi, Consoli et al. 1996).

Obesity, hypertension, and glucose or insulin intolerance are risk factors also tightly linked to the development of metabolic syndrome. Metabolic syndrome is a combination of disorders that increase the susceptibility of cardiovascular disease and diabetes and is estimated to affect almost 25% of the US population (Ford, Giles et al. 2002). While the etiology is still not known, aberrant storage of glucose is a major component of metabolic syndrome, leading to hepatic steatosis or fatty liver, characterized by excessive accumulation of triglycerides (Marchesini, Brizi et al. 2001; Postic, Dentin et al. 2007).

An increase in glucose metabolism is also observed in cancer cells, which are highly proliferative and energy dependent (Medina, Sánchez-Jiménez et al. 1992; Kim and Dang 2006). Most tumor cells and lines exhibit the Warburg effect, where glucose uptake and ATP production via glycolysis is high, even in aerobic growth conditions (Warburg 1956; Tong, Zhao et al. 2009). A significant fraction of these cells also direct glucose into *de novo* lipogenesis, nucleotide biosynthesis, and lactic acid production (Matés, Segura et al. 2009; Tong, Zhao et al. 2009).

MondoA and MondoB Regulate Glucose Metabolism

Initially, sterol regulatory element binding protein (SREBP1) was thought to be the main factor in glucose metabolism and insulin response (Horton, Bashmakov et al. 1998; Foretz, Guichard et al. 1999; da Silva Xavier, Rutter et al. 2006). However, glycolitic and lipogenic gene expression cannot be fully explained by insulin mediated SREBP1 activity (Postic, Dentin et al. 2007). Most of these genes require both insulin and glucose to be fully induced (Koo, Miyashita et al. 2009).

MondoB, also called carbohydrate response element binding protein (ChREBP), predominantly regulates glucose responsive genes in the liver by binding to the promoter region of genes involved in gluconeogenesis, glycolysis, and *de novo* lipogenesis, i.e. acetyl-CoA carboxylase (ACC), L-type pyruvate kinase (L-PK), fatty acid synthase (FAS), thyroid hormone-inducible hepatic protein (S14), glucose kinase regulatory protein (GKRP), and (GLUT4) (Shih, Liu et al. 1995; O'Callaghan, Koo et al. 2001; Yamashita, Takenoshita et al.

2001; Ma, Robinson et al. 2006). Its vertebrate paralog MondoA binds to a distinct set of glycolitic genes involved in energy conversion and blocking glucose uptake, i.e. thioredoxin-interacting protein (TXNIP), lactate dehydrogenase A (LDH-A), 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase (PFKFB3), and hexokinase II (HKII) (Minn, Hafele et al. 2005; Sans, Satterwhite et al. 2006; Stoltzman, Peterson et al. 2008).

Clearly, MondoA and MondoB are major factors in glucose regulation. Although their complex phenotype currently prevents their use as therapeutic drug targets, they may prove useful for treating diabetes and metabolic syndrome in the future.

Individual knockouts of MondoA and MondoB in mice display a complex phenotype of glucose metabolism (Towle 2005). Inhibition of MondoB in the liver decreased the rate of lipogenesis and glycolysis resulting in higher liver glycogen content, lower free fatty acid levels and reduced adipose tissue (da Silva Xavier, Rutter et al. 2006; Dentin, Benhamed et al. 2006; Postic, Dentin et al. 2007). Despite the clear defect in glucose utilization, MondoB$^{-/-}$ mutants are viable, have a normal life span, and improve steatosis and insulin resistance. However, when MondoB$^{-/-}$ mice are fed high fructose or sucrose diets, they rapidly lose weight, exhibit hypothermia and become moribund within days (Uyeda and Repa 2006). Similarly, loss of MondoA increases glucose uptake through reduced TXNIP expression, reduces glycolysis, and stimulates cell proliferation (Sans, Satterwhite et al. 2006; Kaadige, Looper et al. 2009).

In addition, loss of heterozygosity of a genomic 1Mb around MondoB is associated with the Williams Beuren syndrome (WBS). Patients with this condition exhibit an array of abnormalities such as supravalvular aortic stenosis, impaired visual-spatial constructive cognition, mental retardation, and infantile hypercalcemia (Merla, Howald et al. 2004). Around 75% of WBS patients also have elevated glucose levels indicating impaired glucose tolerance or silent diabetes (Pober, Wang et al. 2010). Hemizygosity of MondoB, also called William Beuren Syndrome containing region 14 (WBSCR14), is predicted to contribute to this defect (de Luis, Valero et al. 2000).

Expression and Regulation of Mlx, MondoA, and MondoB

Since MondoA and MondoB have no known cytoplasmic function, their activity as transcription factors is dependent upon interaction with obligate partner Mlx through the bHLHZ dimerization and DNA binding domain (Cairo, Merla et al. 2001; Stoeckman, Ma et al. 2004; Peterson, Stoltzman et al. 2010). Akin to Max, Mlx is a stable protein that is ubiquitously expressed and has a half-life of at least 6-8 hours (Billin, Eilers et al. 1999). Although Mlx can homodimerize and weakly activate transcription by ≈2-fold, Mlx expression alone is not associated with any change in phenotype. Instead, dimerization of Mlx with transcriptional repressors Mxd1 and Mxd4 and activators MondoA and MondoB characterize the Mlx network.

In contrast to the restricted expression of Mxd1 and Mxd4, MondoA and MondoB are constitutively expressed in both embryonic and adult tissues (Billin and Ayer 2006). However, MondoA transcripts are elevated in the developing CNS and skeletal muscle (Billin, Eilers et al. 2000), while MondoB is most abundant in cells with active *de novo* lipogenic pathways such as the liver, pancreas, small intestine, and white and brown adipose tissues (Towle 2005). Still, MondoA:Mlx and MondoB:Mlx complexes do not constitutively transactivate target genes despite their widely distributed expression; additional mechanisms control their transcriptional activity (Kawaguchi, Takenoshita et al. 2001; Eilers, Sundwall et al. 2002; Li, Chang et al. 2006).

One aspect of Mlx, MondoA, and MondoB regulation depends upon their unique dimerization and cytoplasmic localization domain (DCD) located directly after the bHLHZ domain (Eilers, Sundwall et al. 2002; Billin and Ayer 2006). A strong cytoplasmic retention signal (CRS) within the DCD retains Mlx, MondoA, and MondoB monomers in the cytoplasm until dimerization blocks the signal and allows the complex to be transported into the nucleus. Surprisingly, the specific mechanism mediating the CRS has not been investigated and the process of Mlx, MondoA and MondoB nuclear transport is still incomplete.

Some evidence suggests Mlx homodimers can relocate to the nucleus. Although a specific nuclear localization signal (NLS) in Mlx has not yet been determined, comparisons

among the three Mlx isoforms (Mlx-α, -β, -γ) identified a basic residue rich region in the alternative first exon of Mlx-γ is necessary for it nuclear transport (Meroni, Cairo et al. 2000). In HeLa cells, full length Mlx-γ is predominantly nuclear, while shorter Mlx-α and Mlx-β isoforms remain in the cytoplasm. Mlx-α also relocates to the nucleus when coexpressed with Mlx-γ, possibly piggy-backing through dimerization. However, fusion proteins show the Mlx-γ basic region is not a sufficient NLS (Meroni, Cairo et al. 2000). Moreover, conflicting reports show all Mlx isoforms are cytoplasmic in NIH3T3 cells (Billin and Ayer 2006) and Mlx is primarily restricted to the cytoplasm in HL60, K562, and PC12 cells, although the isoform was not specified (Billin, Eilers et al. 2000).

MondoA and MondoB mutants possessing only their C-terminus can bind Mlx, block the CRS, actively transport to the nucleus, and constitutively transactivate genes (Eilers, Sundwall et al. 2002; Billin and Ayer 2006; Li, Chang et al. 2006; Li, Chen et al. 2008; Davies, O'callaghan et al. 2010; Peterson, Stoltzman et al. 2010). However, the subcellular localization of their full length constructs is further regulated by N-terminus domains, comprised of five unique sequence motifs designated Mondo Conserved Region (MCRI-V). MCRI-IV comprise a low glucose inhibitory domain (LID), while MCRV is contained within a glucose-response activation conserved element (GRACE) sufficient for transactivation (Li, Chang et al. 2006). The coordinated activity of these regions displays a complex mechanism controlling both the subcellular localization and transactivation ability of MondoA and MondoB in response to glucose.

Mondo Conserved Regions (MCRI-V)

In low glucose MondoA and MondoB are mainly cytoplasmic, yet continually cycle between the cytoplasm and nucleus with little to no transactivation ability (Billin, Eilers et al. 2000; Sans, Satterwhite et al. 2006; Davies, O'Callaghan et al. 2008; Stoltzman, Peterson et al. 2008; Peterson, Stoltzman et al. 2010). Upon nuclear entry, MondoA and MondoB are actively exported by a potent CRM1 dependent nuclear export signal (NES) in MCRII (Eilers, Sundwall et al. 2002; Li, Chang et al. 2006; Davies, O'Callaghan et al. 2008). Blocking the NES, and thus trapping MondoA or MondoB in the nucleus, does not result in constitutive activation, showing an additional level of activation is necessary (Davies,

O'Callaghan et al. 2008; Li, Chen et al. 2008; Sakiyama, Wynn et al. 2008; Fukasawa, Ge et al. 2010). Moreover, site-specific mutations independent of the NES reveal MCRII is also required for transactivation (Davies, O'Callaghan et al. 2008; Peterson, Stoltzman et al. 2010).

The mechanisms controlling MondoA and MondoB nuclear localization are not as clear. A bipartite NLS has been identified in MondoB MCRIV, although mutations may have compromised the integrity of the MCRIV NLS in MondoA (Billin and Ayer 2006). As expected, NLS mutants were unable to relocate MondoBΔNLS to the nucleus (Kawaguchi, Takenoshita et al. 2001; Fukasawa, Ge et al. 2010). However, MondoA and MondoB mutants lacking MCRI-IV are constitutively active (Eilers, Sundwall et al. 2002; Li, Chang et al. 2006; Li, Chen et al. 2008; Davies, O'callaghan et al. 2010; Peterson, Stoltzman et al. 2010) and the dependency of MCRIV for interaction with nuclear transport proteins importin α and β has not been demonstrated (Sakiyama, Wynn et al. 2008; Davies, O'callaghan et al. 2010). Thus MCRIV may not be required for nuclear import. Instead, conservation of MCRIV may be related to its role in suppression of MondoA and MondoB transactivation. Mutations to MCRIV in MondoA results in increased nuclear accumulation as well as transactivation, while MCRIV was shown to be necessary and sufficient for inhibiting MondoB transactivation in low glucose (Li, Chen et al. 2008; Peterson, Stoltzman et al. 2010).

MCRIII also affects nuclear transport by interacting with protein 14-3-3 at a non-consensus binding site, which slows nuclear import and enhances nuclear export of MondoA and MondoB (Eilers, Sundwall et al. 2002; Billin and Ayer 2006; Li, Chen et al. 2008; Sakiyama, Wynn et al. 2008; Fukasawa, Ge et al. 2010). Human and rat MondoB sequences contain a RRKpSP motif that is similar to a consensus 14-3-3 binding target RXXXpSP, where X is any amino acid and pS is phosphoserine (Bridges and Moorhead 2004). However, MondoA lacks such a sequence. Instead, an α-helix in MCRIII was proven essential for 14-3-3 binding in both MondoA and MondoB (Eilers, Sundwall et al. 2002; Merla, Howald et al. 2004; Li, Chen et al. 2008; Sakiyama, Wynn et al. 2008), leaving the necessity of S140 phosphorylation in the MondoB recognition motif under contention (Merla, Howald et al. 2004; Li, Chen et al. 2008; Sakiyama, Wynn et al. 2008). Still,

MondoB double mutant S140D/S196D, which replaces serine with aspartic acid, mimics a phosphorylated status, increases MondoB binding affinity with 14-3-3, and reduces nuclear accumulation whereas alanine mutation S140A/S196A mimics a dephosphorylated status and has increased nuclear accumulation (Sakiyama, Wynn et al. 2008).

MCRI and MCRV also influence subcellular transport, but their functions are still vague and broadly associated with glucose dependent activation and LID dependent repression, respectively (Davies, O'Callaghan et al. 2008; Tsatsos, Davies et al. 2008; Peterson, Stoltzman et al. 2010). Mutations abrogating MCRI impede MondoA and MondoB transactivation in high glucose, while mimicking phosphorylation at S56 in MondoB increases its transactivation potential (Tsatsos, Davies et al. 2008). Complete removal of the MCRI-IV LID region indicates that MCRV has no repressive affects on transactivation for either MondoA or MondoB (Eilers, Sundwall et al. 2002). However, mutations to MCRV in the presence of the LID region reveal its involvement in synergistic repression (Davies, O'Callaghan et al. 2008). Knockouts of MCRI, II, III, or IV do not alleviate the low glucose inhibition and often abrogate high glucose activation (Li, Chen et al. 2008; Davies, O'callaghan et al. 2010; Peterson, Stoltzman et al. 2010), emphasizing the complex coordination of MCRs in regulating MondoA and MondoB glucose response.

Current Models of Mondo Glucose Response

As more information is gleaned, more sophisticated models have been developed to explain MondoA and MondoB glucose response. Initially, dephosphorylation of sites S196, S626, and T666 in MondoB was believed to be sufficient for its nuclear localization and activation (Kawaguchi, Takenoshita et al. 2001; Sakiyama, Wynn et al. 2008). The dephosphorylation of S196 and T666 directly enhances the nuclear transport and DNA binding activity, respectively, of MondoB in hepatocytes (Dentin, Benhamed et al. 2005). These are both targets of cAMP activated protein kinase A (PKA) and expected to be phosphorylated in low glucose (Kawaguchi, Takenoshita et al. 2001). When glucose levels increase, pentose shunt intermediate xylulose 5-phosphate (X5P) levels rise and activate protein phosphatase PP2A, which dephosphorylates S196, S626, and T666 in MondoB (Kawaguchi, Takenoshita et al. 2001; Sakiyama, Wynn et al. 2008). However, cAMP levels

did not significantly change in low and high glucose conditions and overall MondoB phosphorylation did not decrease in response to glucose (Tsatsos and Towle 2006). Although MondoB is undoubtedly a phosphoprotein, further investigation revealed that the triple mutant S196A/S626A/T666A mutant was not constitutively active and additional mechanisms exist that mediate glucose responsiveness (Li, Chang et al. 2006; Tsatsos and Towle 2006; Tsatsos, Davies et al. 2008). Furthermore, MondoA is also glucose responsive and lacks the phosphorylation potential at these sites (Li, Chang et al. 2006).

Recent evidence demonstrates that nuclear accumulation and activation of both MondoA and MondoB is directly dependent upon glucose, and in particular is correlated with the formation of glucose-6-phosphate (G6P) through glucose phosphorylation (Stoltzman, Peterson et al. 2008). Current models propose G6P affects MondoA and MondoB nuclear accumulation and activation of transcriptional machinery in distinct steps:

MondoB: In low glucose, the LID region represses MondoB activity though intramolecular contacts with MCRV. As glucose levels rise, G6P causes MondoB to adopt an activated conformation either through direct interaction or an allosterically-regulated intermediate, such as a protein kinase. This relieves the repression of LID on MCRV and allows the interaction with additional cofactors involved in transactivation or chromatin remodeling (Davies, O'callaghan et al. 2010).

MondoA: In low glucose, CRM1 binds to MondoA:Mlx and actively transports it from the nucleus. In high glucose conditions, G6P binding dissociates CRM1 from MondoA and links the MCRs and basic region. This change in conformation leads to increased nuclear accumulation and release of repression. Subsequently, MCRs attract a histone H3 acetyltransferase to gene targets enhancing MondoA transactivation activity. (Peterson, Stoltzman et al. 2010).

Based on the similarity between MondoA and MondoB annotation and protein sequence homology, we build upon these models to present a cohesive, yet generalized model of Mondo glucose regulation within Chapter 4.

References

"The Myc Target Gene Database." from myccancergene.org.

Adhikary, S. and M. Eilers (2005). "Transcriptional regulation and transformation by Myc proteins." Nat Rev Mol Cell Biol **6**(8): 635-45.

Akaike, H. (1974). "A new look at the statistical model identification." IEEE Transactions on Automatic Control **19**(6): 716-723.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." Journal of Molecular Biology **215**(3): 403-10.

Amati, B. and H. Land (1994). "Myc-Max-Mad: a transcription factor network controlling cell cycle progression, differentiation and death." Curr Opin Genet Dev **4**(1): 102-8.

Ayer, D. E. and R. N. Eisenman (1993). "A switch from Myc:Max to Mad:Max heterocomplexes accompanies monocyte/macrophage differentiation." Genes & Development **7**(11): 2110-9.

Ayer, D. E., L. Kretzner, et al. (1993). "Mad: a heterodimeric partner for Max that antagonizes Myc transcriptional activity." Cell **72**(2): 211-22.

Bellman, E. R. (1961). "Adaptive control processes: a guided tour?" 255.

Beyer, W. A., M. L. Stein, et al. (1974). "A Molecular Sequence Metric and Evolutionary Trees." Mathematical Biosciences **19**: 9-25.

Billin, A. N. and D. E. Ayer (2006). "The Mlx network: evidence for a parallel Max-like transcriptional network that regulates energy metabolism." Curr Top Microbiol Immunol **302**: 255-78.

Billin, A. N., A. L. Eilers, et al. (1999). "Mlx, a novel Max-like BHLHZip protein that interacts with the Max network of transcription factors." J Biol Chem **274**(51): 36344-50.

Blackwell, T. K., J. Huang, et al. (1993). "Binding of myc proteins to canonical and noncanonical DNA sequences." Mol Cell Biol **13**(9): 5216-24.

Blackwood, E. M. and R. N. Eisenman (1991). "Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc." Science **251**(4998): 1211-7.

Blackwood, E. M., B. Lüscher, et al. (1992). "Myc and Max associate in vivo." Genes & Development **6**(1): 71-80.

Bouchard, C., O. Dittrich, et al. (2001). "Regulation of cyclin D2 gene expression by the Myc/Max/Mad network: Myc-dependent TRRAP recruitment and histone acetylation at the cyclin D2 promoter." Genes & Development **15**(16): 2042-7.

Bridges, D. and G. B. Moorhead (2004). "14-3-3 proteins: a number of functions for a numbered protein." Sci STKE **2004**(242): re10.

Brownlie, P., T. Ceska, et al. (1997). "The crystal structure of an intact human Max-DNA complex: new insights into mechanisms of transcriptional control." Structure **5**(4): 509-20.

Bruno, W. J., N. D. Socci, et al. (2000). "Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction." Molecular Biology and Evolution **17**(1): 189-97.

Burton, R. A., S. Mattila, et al. (2006). "B-myc: N-terminal recognition of myc binding proteins." Biochemistry **45**(32): 9857-65.

Carrillo, H. and D. Lipman (1988). "The Multiple Sequence Alignment Problem in Biology." SIAM Journal of Applied Mathematics **48**(5): 11.

Casella, G. and R. L. Berger (2002). "Statistical inference." 660.

Chamary, J. V., J. L. Parmley, et al. (2006). "Hearing silence: non-neutral evolution at synonymous sites in mammals." Nat Rev Genet **7**(2): 98-108.

Clarke, R., H. W. Ressom, et al. (2008). "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data." Nat Rev Cancer **8**(1): 37-49.

Cole, M. D. and M. A. Nikiforov (2006). "Transcriptional activation by the Myc oncoprotein." Curr Top Microbiol Immunol **302**: 33-50.

Consoli, A., N. Nurjhan, et al. (1990). "Mechanism of increased gluconeogenesis in noninsulin-dependent diabetes mellitus. Role of alterations in systemic, hepatic, and muscle lactate and alanine metabolism." J Clin Invest **86**(6): 2038-45.

Cusi, K., A. Consoli, et al. (1996). "Metabolic effects of metformin on glucose and lactate metabolism in noninsulin-dependent diabetes mellitus." J Clin Endocrinol Metab **81**(11): 4059-67.

Dang, C. V. (1999). "c-Myc target genes involved in cell growth, apoptosis, and metabolism." Molecular and Cellular Biology **19**(1): 1-11.

Davies, M., B. O'callaghan, et al. (2010). "Activation and Repression of Glucose-stimulated ChREBP Requires the Concerted Action of Multiple Domains within the MondoA Conserved Region." AJP: Endocrinology and Metabolism: 38.

Davies, M. N., B. L. O'Callaghan, et al. (2008). "Glucose activates ChREBP by increasing its rate of nuclear entry and relieving repression of its transcriptional activity." J Biol Chem **283**(35): 24029-38.

Dayhoff, M. O., R. M. Schwartz, et al. (1978). A model of evolutionary change in proteins. Washington, DC, Natl. Biomed. Res. Found.

Delacova, C. and L. Johnston (2006). "Myc in model organisms: A view from the flyroom." Seminars in Cancer Biology **16**(4): 303-312.

Dentin, R., F. Benhamed, et al. (2005). "Polyunsaturated fatty acids suppress glycolytic and lipogenic genes through the inhibition of ChREBP nuclear protein translocation." J Clin Invest **115**(10): 2843-54.

Donoho, D. (2000). "Mathematical Challenges of the 21st Century - High-Dimensional Data Analysis: The Blessings and Curses of Dimensionality." from http://www.stat.stanford.edu/~donoho/Lectures/AMS2000/MathChallengeSlides2*2.pdf.

Durbin, R., S. Eddy, et al. (1998). "Biological sequence analysis: probabilistic models of proteins and nucleic acids, ." Cambridge University Press.

E. P. Box, G. and N. Richard Draper (1987). "Empirical model-building and response surfaces?" 669.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Research **32**(5): 1792-7.

Eilers, A. L., E. Sundwall, et al. (2002). "A novel heterodimerization domain, CRM1, and 14-3-3 control subcellular localization of the MondoA-Mlx heterocomplex." Molecular and Cellular Biology **22**(24): 8514-26.

Eilers, M. and R. Eisenman (2008). "Myc's broad reach." Genes & Development **22**(20): 2755-2766.

Eilers, M., S. Schirm, et al. (1991). "The MYC protein activates transcription of the alpha-prothymosin gene." EMBO J **10**(1): 133-41.

Elias, I. (2006). "Settling the intractability of multiple alignment." J Comput Biol **13**(7): 1323-39.

Felber, J. P., E. Ferrannini, et al. (1987). "Role of lipid oxidation in pathogenesis of insulin resistance of obesity and type II diabetes." Diabetes **36**(11): 1341-50.

Felsenstein, J. (1978). "The Number of Evolutionary Trees." Systematic Zoology **27**(1): 27-33.

Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." J Mol Evol **17**(6): 368-76.

Felsenstein, J. (2004). "Inferring phylogenies." 664.

Felsenstein, J. and G. A. Churchill (1996). "A Hidden Markov Model approach to variation among sites in rate of evolution." Molecular Biology and Evolution **13**(1): 93-104.

Felsher, D. W. and J. M. Bishop (1999). "Transient excess of MYC activity can elicit genomic instability and tumorigenesis." Proc Natl Acad Sci USA **96**(7): 3940-4.

Feng, D. F. and R. F. Doolittle (1987). "Progressive sequence alignment as a prerequisite to correct phylogenetic trees." J Mol Evol **25**(4): 351-60.

Fernandes, A. D. and W. R. Atchley (2008). "Site-specific evolutionary rates in proteins are better modeled as non-independent and strictly relative." Bioinformatics **24**(19): 2177-83.

Ferré-D'Amaré, A. R., G. C. Prendergast, et al. (1993). "Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain." Nature **363**(6424): 38-45.

Fisher, A. R. (1930). "The genetical theory of natural selection." 291.

Fisher, R. (1930). "The Distribution of Gene Ratios for Rare Mutations." Proceedings of the Royal Society of Edinburgh **50**: 205-220.

Fitch, W. M. and T. F. Smith (1983). "Optimal sequence alignments." Proc Natl Acad Sci USA **80**(5): 1382-6.

Foley, K. P., G. A. McArthur, et al. (1998). "Targeted disruption of the MYC antagonist MAD1 inhibits cell cycle exit during granulocyte differentiation." EMBO J **17**(3): 774-85.

Ford, E. S., W. H. Giles, et al. (2002). "Prevalence of the metabolic syndrome among US adults: findings from the third National Health and Nutrition Examination Survey." JAMA **287**(3): 356-9.

Fox, E. J. and S. C. Wright (2001). "S-phase-specific expression of the Mad3 gene in proliferating and differentiating cells." Biochem J **359**(Pt 2): 361-7.

Freie, B. W. and R. N. Eisenman (2008). "Ratcheting Myc." Cancer Cell **14**(6): 425-6.

Gallant, P. (2006). "Myc/Max/Mad in invertebrates: the evolution of the Max network." Curr Top Microbiol Immunol **302**: 235-53.

Gascuel, O. (1997). "BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data." Molecular Biology and Evolution **14**(7): 685-95.

Goldman, N. and Z. Yang (1994). "A codon-based model of nucleotide substitution for protein-coding DNA sequences." Mol Biol Evol **11**(5): 725-36.

Günes, C., S. Lichtsteiner, et al. (2000). "Expression of the hTERT gene is regulated at the level of transcriptional initiation and repressed by Mad1." Cancer Res **60**(8): 2116-21.

Halpern, A. L. and W. J. Bruno (1998). "Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies." Molecular Biology and Evolution **15**(7): 910-7.

Hasegawa, M., H. Kishino, et al. (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA." J Mol Evol **22**(2): 160-74.

Hastings, W. K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." Biometrika **57**(1): 97-109.

Hatton, K., K. Mahon, et al. (1996). "Expression and activity of L-Myc in normal mouse development." Molecular and Cellular Biology **16**(4): 1794-804.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci USA **89**(22): 10915-9.

Hoffman, B. and D. Liebermann (2008). "Apoptotic signaling by c-MYC." Oncogene **27**(50): 6462-6472.

Hollich, V., L. Milchert, et al. (2005). "Assessment of protein distance measures and tree-building methods for phylogenetic tree reconstruction." Molecular Biology and Evolution **22**(11): 2257-64.

Hooker, C. W. and P. Hurlin (2006). "Of Myc and Mnt." Journal of Cell Science **119**(Pt 2): 208-16.

Huelsenbeck, J. P. (1995). "The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining." Molecular Biology and Evolution **12**(5): 843-9.

Hurlin, P. and J. Huang (2006). "The MAX-interacting transcription factor network." Seminars in Cancer Biology **16**(4): 265-274.

Hurlin, P., Z. Q. Zhou, et al. (2003). "Deletion of Mnt leads to disrupted cell cycle control and tumorigenesis." EMBO J **22**(18): 4584-96.

Hurlin, P. J., C. Quéva, et al. (1995). "Mad3 and Mad4: novel Max-interacting transcriptional repressors that suppress c-myc dependent transformation and are expressed during neural and epidermal differentiation." EMBO J **14**(22): 5646-59.

J. Huberty, C. (1994). "Applied discriminant analysis." 466.

Johnson, A. R. and W. D. Wichern (2007). "Applied multivariate statistical analysis." 773.

Johnston, L. A., D. A. Prober, et al. (1999). "Drosophila myc regulates cellular growth during development." Cell **98**(6): 779-90.

Jolliffe, T. I. (2002). "Principal component analysis." 487.

Jones, D. T., W. R. Taylor, et al. (1992). "The rapid generation of mutation data matrices from protein sequences." Comput Appl Biosci **8**(3): 275-82.

Jukes, T. and C. Cantor (1969). Evolution of protein molecules. New York, Academic Press.

Katoh, K., K. Misawa, et al. (2002). "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." Nucleic Acids Research **30**(14): 3059-66.

Kawaguchi, T., M. Takenoshita, et al. (2001). "Glucose and cAMP regulate the L-type pyruvate kinase gene by phosphorylation/dephosphorylation of the carbohydrate response element binding protein." Proc Natl Acad Sci USA **98**(24): 13710-5.

Kelley, D. E., B. Goodpaster, et al. (1999). "Skeletal muscle fatty acid metabolism in association with insulin resistance, obesity, and weight loss." Am J Physiol **277**(6 Pt 1): E1130-41.

Kelley, D. E., J. He, et al. (2002). "Dysfunction of mitochondria in human skeletal muscle in type 2 diabetes." Diabetes **51**(10): 2944-50.

Kelley, D. E. and L. J. Mandarino (2000). "Fuel selection in human skeletal muscle in insulin resistance: a reexamination." Diabetes **49**(5): 677-83.

Kidera, A., Konishi, et al. (2009). "Statistical Analysis of the Physical Properties of the 20 Naturally Occuring Amino Acids." 33.

Kim, J. and S. Pramanik (1994). "An efficient method for multiple sequence alignment." Proc Int Conf Intell Syst Mol Biol **2**: 212-8.

Kimura, M. (1962). "On the probability of fixation of mutant genes in a population." Genetics **47**: 713-9.

Kimura, M. (1968). "Evolutionary rate at the molecular level." Nature **217**(5129): 624-6.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." J Mol Evol **16**(2): 111-20.

Knoepfler, P. (2007). "Myc Goes Global: New Tricks for an Old Oncogene." Cancer Research **67**(11): 5061-5063.

Koo, H., M. Miyashita, et al. (2009). "Replacing dietary glucose with fructose increases ChREBP activity and SREBP-1 protein in rat liver nucleus." Biochem Biophys Res Commun **390**(2): 285-289.

Koskinen, P. J., D. E. Ayer, et al. (1995). "Repression of Myc-Ras cotransformation by Mad is mediated by multiple protein-protein interactions." Cell Growth Differ **6**(6): 623-9.

Kumar, S. and A. Filipski (2007). "Multiple sequence alignment: In pursuit of homologous DNA positions." Genome Research **17**(2): 127-135.

L. Gorsuch, R. (1983). "Factor analysis." 425.

Lahoz, E. G., L. Xu, et al. (1994). "Suppression of Myc, but not E1a, transformation activity by Max-associated proteins, Mad and Mxi1." Proc Natl Acad Sci USA **91**(12): 5503-7.

Larsson, L. G., M. Pettersson, et al. (1994). "Expression of mad, mxi1, max and c-myc during induced differentiation of hematopoietic cells: opposite regulation of mad and c-myc." Oncogene **9**(4): 1247-52.

Lassmann, T. and E. L. Sonnhammer (2005). "Kalign--an accurate and fast multiple sequence alignment algorithm." BMC Bioinformatics **6**: 298.

Lavigne, P., M. P. Crump, et al. (1998). "Insights into the mechanism of heterodimerization from the 1H-NMR solution structure of the c-Myc-Max heterodimeric leucine zipper." Journal of Molecular Biology **281**(1): 165-81.

Lee, L. A. and C. V. Dang (2006). "Myc target transcriptomes." Curr Top Microbiol Immunol **302**: 145-67.

Li, M., B. Chang, et al. (2006). "Glucose-dependent transcriptional regulation by an evolutionarily conserved glucose-sensing module." Diabetes **55**(5): 1179-89.

Li, M., W. Chen, et al. (2008). "Glucose-Mediated Transactivation of Carbohydrate Response Element-Binding Protein Requires Cooperative Actions from Mondo Conserved Regions and Essential Trans-Acting Factor 14-3-3." Mol Endocrinol **22**(7): 1658-1672.

Lipman, D. J., S. F. Altschul, et al. (1989). "A tool for multiple sequence alignment." Proc Natl Acad Sci USA **86**(12): 4412-5.

Lüscher, B. (2001). "Function and regulation of the transcription factors of the Myc/Max/Mad network." Gene **277**(1-2): 1-14.

Ma, L., L. N. Robinson, et al. (2006). "ChREBP*Mlx is the principal mediator of glucose-induced gene expression in the liver." J Biol Chem **281**(39): 28721-30.

Majewski, J. and J. Ott (2003). "Amino acid substitutions in the human genome: evolutionary implications of single nucleotide polymorphisms." Gene **305**(2): 167-73.

Marjoram, P. and S. Tavaré (2006). "Modern computational approaches for analysing molecular genetic variation data." Nat Rev Genet **7**(10): 759-770.

Mayrose, I., A. Doron-Faigenboim, et al. (2007). "Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates." Bioinformatics **23**(13): i319-i327.

McArthur, G. A., K. P. Foley, et al. (2002). "MAD1 and p27(KIP1) cooperate to promote terminal differentiation of granulocytes and to inhibit Myc expression and cyclin E-CDK2 activity." Mol Cell Biol **22**(9): 3014-23.

Metropolis, N., A. W. Rosenbluth, et al. (1953). "Equations of State Calculations by Fast Computing Machines." Journal of Chemical Physics **21**(6): 1087–1092.

Morgenstern, B., S. Goel, et al. (2003). "AltAVisT: comparing alternative multiple sequence alignments." Bioinformatics **19**(3): 425-6.

Muse, S. V. and B. S. Gaut (1994). "A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome." Molecular Biology and Evolution **11**(5): 715-24.

Nair, S. K. and S. K. Burley (2003). "X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors." Cell **112**(2): 193-205.

Nielsen, R. and Z. Yang (1998). "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene." Genetics **148**(3): 929-36.

Noordeen, N. A., T. K. Khera, et al. (2010). "Carbohydrate-responsive element-binding protein (ChREBP) is a negative regulator of ARNT/HIF-1beta gene expression in pancreatic islet beta-cells." Diabetes **59**(1): 153-60.

Notredame, C. and D. G. Higgins (1996). "SAGA: sequence alignment by genetic algorithm." Nucleic Acids Res **24**(8): 1515-24.

Notredame, C., D. G. Higgins, et al. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." Journal of Molecular Biology **302**(1): 205-17.

Ohta, T. (2002). "Near-neutrality in evolution of genes and gene regulation." Proc Natl Acad Sci USA **99**(25): 16134-7.

Orian, A., B. van Steensel, et al. (2003). "Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network." Genes & Development **17**(9): 1101-14.

Peterson, C., C. Stoltzman, et al. (2010). "Glucose Controls Nuclear Accumulation, Promoter Binding, and Transcriptional Activity of the MondoA-Mlx Heterodimer." Molecular and Cellular Biology **30**(12): 2887-2895.

Pober, B. R., E. Wang, et al. (2010). "High prevalence of diabetes and pre-diabetes in adults with Williams syndrome." Am J Med Genet C Semin Med Genet **154C**(2): 291-8.

Pollard, D. A., C. M. Bergman, et al. (2004). "Benchmarking tools for the alignment of functional noncoding DNA." BMC Bioinformatics **5**: 6.

Pond, S. K. and S. V. Muse (2005). "Site-to-site variation of synonymous substitution rates." Mol Biol Evol **22**(12): 2375-85.

Postic, C., R. Dentin, et al. (2007). "ChREBP, a transcriptional regulator of glucose and lipid metabolism." Annu Rev Nutr **27**: 179-92.

Quéva, C., P. J. Hurlin, et al. (1998). "Sequential expression of the MAD family of transcriptional repressors during differentiation and development." Oncogene **16**(8): 967-77.

Quéva, C., G. A. McArthur, et al. (2001). "Targeted deletion of the S-phase-specific Myc antagonist Mad3 sensitizes neuronal and lymphoid cells to radiation-induced apoptosis." Mol Cell Biol **21**(3): 703-12.

Ransohoff, D. F. (2005). "Bias as a threat to the validity of cancer molecular-marker research." Nat Rev Cancer **5**(2): 142-9.

Rottmann, S. and B. Lüscher (2006). "The Mad side of the Max network: antagonizing the function of Myc and more." Curr Top Microbiol Immunol **302**: 63-122.

Sáez, A. I., M. J. Artiga, et al. (2003). "Development of a real-time reverse transcription polymerase chain reaction assay for c-myc expression that allows the identification of a subset of c-myc+ diffuse large B-cell lymphoma." Lab Invest **83**(2): 143-52.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Molecular Biology and Evolution **4**(4): 406-25.

Sakiyama, H., R. M. Wynn, et al. (2008). "Regulation of nuclear import/export of carbohydrate response element-binding protein (ChREBP): interaction of an alpha-helix of ChREBP with the 14-3-3 proteins and regulation by phosphorylation." J Biol Chem **283**(36): 24899-908.

Sauvé, S., L. Tremblay, et al. (2004). "The NMR solution structure of a mutant of the Max b/HLH/LZ free of DNA: insights into the specific and reversible DNA binding mechanism of dimeric transcription factors." Journal of Molecular Biology **342**(3): 813-32.

Schwarz, G. E. (1978). "Estimating the dimension of a model." Annals of Statistics **6**(2): 461-464.

Sears, R. C. (2004). "The life cycle of C-myc: from synthesis to degradation." Cell Cycle **3**(9): 1133-7.

Sheiness, D., L. Fanshier, et al. (1978). " Identification of nucleotide sequences which may encode the oncogenic capacity of avian retrovirus MC29." J. Virol. **28**: 600-610.

Sokal, R. and C. Michener (1958). "A statistical method for evaluating systematic relationships." University of Kansas Science Bulletin **38**: 1409-1438.

Song, J., C. Liu, et al. (2010). "Alignment of multiple proteins with an ensemble of hidden Markov models." Int J Data Min Bioinform **4**(1): 60-71.

Spencer, C. A. and M. Groudine (1991). "Control of c-myc regulation in normal and neoplastic cells." Adv Cancer Res **56**: 1-48.

Steiger, D., M. Furrer, et al. (2008). "Max-independent functions of Myc in Drosophila melanogaster." Nat Genet **40**(9): 1084-1091.

Stoltzman, C. A., C. W. Peterson, et al. (2008). "Glucose sensing by MondoA:Mlx complexes: a role for hexokinases and direct regulation of thioredoxin-interacting protein expression." Proc Natl Acad Sci USA **105**(19): 6912-7.

Storlien, L., N. D. Oakes, et al. (2004). "Metabolic flexibility." Proc Nutr Soc **63**(2): 363-8.

Subramanian, A. R., M. Kaufmann, et al. (2008). "DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment." Algorithms for molecular biology : AMB **3**: 6.

Syversveen, A. R. (1998). "Noninformative Bayesian priors: interpretation and problems with ...‡." 11.

Tamura, K. and M. Nei (1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees." Molecular Biology and Evolution **10**(3): 512-26.

Tavaré, S. (1986). "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences." Lectures on Mathematics in the Life Sciences **17**: 57-86.

Taylor, W. R. (1988). "A flexible method to align large numbers of biological sequences." J Mol Evol **28**(1-2): 161-9.

Thompson, J. D., F. Plewniak, et al. (1999). "A comprehensive comparison of multiple sequence alignment programs." Nucleic Acids Res **27**(13): 2682-90.

Towle, H. C. (2005). "Glucose as a regulator of eukaryotic gene transcription." Trends Endocrinol Metab **16**(10): 489-94.

Tsatsos, N. G., M. N. Davies, et al. (2008). "Identification and function of phosphorylation in the glucose-regulated transcription factor ChREBP." Biochem J **411**(2): 261-70.

Tsatsos, N. G. and H. C. Towle (2006). "Glucose activation of ChREBP in hepatocytes occurs via a two-step mechanism." Biochem Biophys Res Commun **340**(2): 449-56.

Uyeda, K. and J. J. Repa (2006). "Carbohydrate response element binding protein, ChREBP, a transcription factor coupling hepatic glucose utilization and lipid synthesis." Cell Metab **4**(2): 107-10.

Västrik, I., A. Kaipainen, et al. (1995). "Expression of the mad gene during cell differentiation in vivo and its inhibition of cell growth in vitro." The Journal of Cell Biology **128**(6): 1197-208.

Walker, W., Z. Q. Zhou, et al. (2005). "Mnt-Max to Myc-Max complex switching regulates cell cycle entry." The Journal of Cell Biology **169**(3): 405-13.

Wang, L. and T. Jiang (1994). "On the complexity of multiple sequence alignment." J Comput Biol **1**(4): 337-48.

Wright, S. (1931). "Evolution in Mendelian Populations." Genetics **16**(2): 97-159.

Xu, D., N. Popov, et al. (2001). "Switch from Myc/Max to Mad1/Max binding and decrease in histone acetylation at the telomerase reverse transcriptase promoter during differentiation of HL60 cells." Proc Natl Acad Sci USA **98**(7): 3826-31.

Yang, Z. (1994). "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods." J Mol Evol **39**(3): 306-14.

Yang, Z. (1995). "A space-time process model for the evolution of DNA sequences." Genetics **139**(2): 993-1005.

Yang, Z. (2002). "Inference of selection from multiple species alignments." Curr Opin Genet Dev **12**(6): 688-94.

Yang, Z. and R. Nielsen (2000). "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models." Molecular Biology and Evolution **17**(1): 32-43.

Yang, Z., R. Nielsen, et al. (2000). "Codon-substitution models for heterogeneous selection pressure at amino acid sites." Genetics **155**(1): 431-49.

Zervos, A. S., J. Gyuris, et al. (1993). "Mxi1, a protein that specifically interacts with Max to bind Myc-Max recognition sites." Cell **72**(2): 223-32.

Zhou, Z. Q. and P. J. Hurlin (2001). "The interplay between Mad and Myc in proliferation and differentiation." Trends Cell Biol **11**(11): S10-4.

Zorzano, A., M. Liesa, et al. (2009). "Role of mitochondrial dynamics proteins in the pathophysiology of obesity and type 2 diabetes." Int J Biochem Cell Biol **41**(10): 1846-54.

Chapter 2

The Non-Random Clustering of Non-Synonymous Substitutions and its Relationship to Evolutionary Rate

Abstract

Protein sequences are subject to a mosaic of constraint. Changes to functional domains and buried residues, for example, are more apt to disrupt protein structure and function than are changes to residues participating in loops or exposed to solvent. Regions of constraint on the tertiary structure of a protein often result in loose segmentation of its primary structure into stretches of slowly- and rapidly-evolving amino acids. This clustering can be exploited, and existing methods have done so by relying on local sequence conservation as a signature of selection to help identify functionally important regions within proteins. Here we invert this paradigm leveraging the regional nature of protein structure and function to both illuminate and make use of genome-wide patterns of local sequence conservation. Our hypothesis is that the regional nature of structural and functional constraints will assert a positive autocorrelation on the evolutionary rates of neighboring sites, which, in a pairwise comparison of orthologous proteins, will manifest itself as the clustering of non-synonymous changes across the amino acid sequence. Using genome-wide interspecific comparisons of orthologous protein pairs, we show this to be the case and reveal a strong log-linear relationship between the degree of clustering and the intensity of constraint. We further demonstrate how this relationship varies with the evolutionary distance between the species being compared. Because there is evidence that both purifying and positive selection will promote the clustering of non-synonymous changes, we examine whether proteins with a history of positive selection deviate from the established genome-wide trend. We make the case that conflicting signals of clustering and constraint may be indicative of a historical period of relaxed selection.

Introduction

For functional biological sequences, and for proteins in particular, similarity in sequence is often predictive of similarity in structure and function. This has great utility, because while it is challenging to glean knowledge of structure and function, sequence information is comparatively easy to obtain. For this reason, and because comparing two sequences in an alignment can often be straightforward, pairwise alignments are often the first step toward annotating a sequence whose folded structure and biological function are unknown. When two sequences show extensive similarity and one of the two has been annotated, transferring that annotation provides an effortless functional prediction; however, even in the complete absence of annotation, alignments can be used to ascribe functional importance to sites and regions in a sequence (Hardison 2000). Consider, for example, two distantly-related sequences, say a pair of orthologous genes in human and chicken. Both the coding sequences of these genes and the amino acid sequences that they encode may be very different, yet particular stretches of residues may be well conserved (Takata et al. 2002). While such surprising similarity can arise by random chance, it may also be the footprint of purifying selection, indicating a region of the sequence that is functionally important and resistant to evolutionary change.

In proteins, functionally and structurally important residues are often organized into domains that as units are themselves structurally and functionally important. Thus, in a comparison of related sequences, domains may be apparent as regions of surprising similarity. This style of *de novo* annotation is exploited routinely and underlies a number of web-accessible methods including but not limited to the Evolutionary Trace (ET) (Lichtarge, Bourne, and Cohen 1996; Joachimiak and Cohen 2002; Mihalek, Res, and Lichtarge 2006) and Evolution-Structure-Function analysis (ESF) (Simon, Stone, and Sidow 2002; Binkley et al. 2010). The success of these methods relies upon two general characteristics of protein sequences, namely (1) that there exists heterogeneity among the rates at which sites in a protein evolve and (2) that the rates are spatially autocorrelated (see Figure 1). Consequently, more sophisticated *de novo* annotation schemes gain resolution through a combination of improved evolutionary models, accounting for site autocorrelation, and

respecting spatial proximities, induced by tertiary structure, e.g. (Doron-Faigenboim et al. 2005; Glaser et al. 2005; Landau et al. 2005).

Just as surprising regional similarity in a pairwise comparison may be of biological interest, interesting biology may be responsible for regions that are surprisingly distinct. For example, in a comparison of closely-related species, say human and chimpanzee, one expects a great deal of sequence similarity. Regions of surprising dissimilarity may encode positively selected adaptations that have helped to distinguish us from our primate cousins, e.g. (Pollard et al. 2006). Within a protein-coding gene, there is evidence that sites undergoing diversifying positive selection, that is, those evolving more rapidly than the rate of neutral evolution would predict, cluster non-randomly along the primary sequence (Wagner 2007; Zhou, Enyeart, and Wilke 2008). The web-accessible tool SWAKK, which is similar in spirit to ET and ESF, exploits this non-random distribution to identify positively-selected regions within a protein (Liang, Zhou, and Landweber 2006).

Synthesizing the above, there is evidence that both negative purifying selection and positive diversifying selection promote the clustering of amino acid differences in a pairwise comparison of protein sequences. In contrast, in the absence of selection at the protein level (e.g. for a pseudogene or fully redundant duplicate), clustering is not expected, unless for example the mutation process is biased or there is selection on the encoding DNA. In a snapshot of evolutionary time, most proteins are under purifying selection, whereby non-synonymous mutations that change the encoded protein are more likely to fix if they affect regions of the sequence with less functional importance. This raises the possibility that for proteins under stronger purifying selection the clustering of amino acid differences in a pairwise comparison is more intense. To explore this and other possibilities, we introduce a simple statistic that quantifies the degree to which non-synonymous changes are clustered in a pairwise alignment.

In this manuscript, we consider aligned pairs of putatively orthologous protein-coding sequences across a variety of species. Within that focus, we hypothesize that: (1) there exists a genome-wide trend relating the intensity with which purifying selection acts on a protein sequence to the intensity with which non-synonymous changes are clustered in a pairwise

alignment; (2) gene pairs which have undergone periods of relaxed or reversed constraint, such as might occur subsequent to gene duplication, appear as deviations from the genome-wide trend; and (3) the intensity with which non-synonymous changes are clustered in a pairwise alignment is a strong non-redundant predictor of the relationship of evolutionary rates. Using our new "dispersion ratio" statistic, we provide evidence in support of each hypothesis as well as show that the hypotheses are robust to the choice of genomes compared.

Results

The dispersion ratio as a simple measure of clustering

In this section we introduce the dispersion ratio, a statistic designed to detect the clustering of non-synonymous changes in a pairwise alignment. To illustrate the approach, in Figure 2 we present a hypothetical 27aa protein sequence that is composed of alternating rapidly- and slowly-evolving segments. As shown in the figure, to construct the dispersion ratio from a pair of aligned protein-coding sequences, we begin by identifying the aligned codon positions at which the encoded amino acids disagree. At each position where the encoded amino acids disagree, we label the adjacent sites in the primary sequence with an "**a**". We then partition the alignment into two subalignments: one composed exclusively of the sites labeled "**a**", and one composed of the remaining sites, which we label "**i**" for isolated. Within each of these subalignments, we compute the ratio of the rate of non-synonymous substitutions to the rate of synonymous substitutions (ω_I and ω_A for the isolated and adjacent subalignments, respectively). The dispersion ratio ρ is the ratio of ratios ω_I / ω_A .

The dispersion ratio measures the degree to which non-synonymous changes are clustered along a protein's primary sequence. It specifically quantifies the propensity for non-synonymous changes to neighbor one another in a comparison of homologous proteins. The philosophy of ρ can be conveyed through Figure 2 by simply tallying where the non-synonymous changes fall; there 2 of 11 isolated sites (18%) harbor a non-synonymous change, as compared to 7 of 16 adjacent ones (44%), suggesting a dispersion ratio smaller than one. As the name implies, larger values of ρ indicate that non-synonymous changes are

more dispersed, whereas smaller values indicate a greater degree of clustering. Supplied with this definition of ρ, we can rephrase our first hypothesis as follows: if ω is the ratio of the rate of non-synonymous substitutions to the rate of synonymous substitutions for the entire protein, then we hypothesize a genome-wide trend that relates ω to ρ.

A significant log-linear relationship between selection and dispersion

To test hypothesis (1), we conducted a genome-wide comparison between each human protein-coding gene and its ortholog, when present and unambiguous, across eight vertebrate species (Figure 3). We restricted ourselves to unique orthologs as designated by Ensembl (see Methods) and used their previously computed alignments. For each alignment, we used the model of Yang and Nielsen (2000) as implemented in PAML to estimate ω, ω_I, and ω_A as described in the previous section. Each aligned pair of orthologs thus provides a (ω, ρ) coordinate pair that can be entered into a species-specific scatterplot of genes. These eight scatterplots – one for each non-human species in the phylogeny of Figure 3 – show a consistent, non-linear monotonic trend; as ω decreases, so too does ρ, indicating that the degree to which non-synonymous changes cluster increases with the strength of purifying selection (data not shown). When the two axes are log-transformed, so that log(ρ) is plotted against log(ω), the relationship becomes linear and highly significant. In Figure 4, log(ρ) is plotted against log(ω) in blue for 11894 aligned pairs of orthologous genes identified in human and mouse (see Methods for inclusion criteria). The linear trend depicted in black is highly significant (r = 0.3877; p-value < 2.2e-16) and is not limited to the comparison of human and mouse. Indeed, as Table 1 shows, each of the eight comparisons provides strong evidence of a significant log-linear trend relating our chosen measures of selection and dispersion.

To emphasize the significance of our findings, the scatterplot of Figure 4 in red presents a control. Our control, constructed in the spirit of the dispersion ratio, follows the construction illustrated in Figure 2 for synonymous rather than for non-synonymous changes. Thus, whereas ρ is created by first partitioning sites in the alignment according to the location of non-synonymous changes, the synonymous dispersion ratio ρ_S is created by first partitioning sites according to where synonymous changes are observed. Figure 4 plots

log(ρ_S) against log(ω) in red for the human/mouse comparison. As the figure shows, the relationship is not significant (r = -0.0156; p-value = 0.087), suggesting that in strong contrast to non-synonymous changes, the clustering of synonymous changes does not depend on the intensity of purifying selection on the protein sequence.

As a final validation, we turned to a permutation-based approach whereby the order of sites in each alignment was shuffled. The effect of this, for any one aligned pair of orthologs, is to hold ω fixed while varying ρ in a random, non-biological way. Permuting each aligned human/mouse pair creates an alternative version of the blue scatterplot in Figure 4; the observed correlation can be thought of as a sample from a null distribution under which selection and dispersion are not biologically related. Owing to edge effects and the discrete nature of the data, the mean correlation under the null hypothesis is biased away from zero; nevertheless, the correlation observed in our original data is uniformly and substantially larger than any of the permuted realizations, and this persists regardless of the comparison. This, once again, supports the existence of a genome-wide trend relating the intensity with which purifying selection acts on a protein and the intensity with which non-synonymous changes are clustered.

Genes under recent positive selection deviate from the trend

In a pairwise comparison of protein-coding sequences, it is difficult to disentangle the mode and tempo of the evolutionary process. For example, genes under recent positive selection in the human lineage may not appear as such in a pairwise comparison if purifying selection is acting upon the gene in the sister lineage. Put another way, the pairwise comparison reflects the aggregated effects of two evolutionary regimes, one in which the protein evolves at a rate faster than expected under neutrality, and one in which the protein evolves at a rate slower than expected under neutrality. As a consequence of this aggregation, the individual regimes that compose such a mixed regime may be obscured, unless of course additional information is incorporated in the analysis. We hypothesize that the dispersion ratio provides useful information toward disentangling mixed evolutionary regimes. Evidence of this comes from the observation that both purifying selection and positive selection appear to promote the clustering of non-synonymous changes: if both

regimes promote clustering, than the degree of clustering observed under a mixed regime may be surprisingly large given the apparent intensity of selection. The stable relationship between log(ρ) and log(ω) presented in the previous section suggests that log(ω) can be predicted from log(ρ); in a pairwise comparison that spans a mixed regime, log(ω) may be appear too large when compared to a prediction based on the value of log(ρ) that was observed. In other words, we hypothesized that a mixed regime might lead to evolutionary rates that are "too fast" for the degree of clustering observed.

As a test of this hypothesis, we turned to a set of protein-coding genes implicated as being under positive selection in the human lineage after the human/chimpanzee split (Vallender and Lahn 2004; Sabeti et al. 2006). Reversing the axes from Figure 4, in Figure 5 we identified these genes in a human/chimpanzee scatterplot of log(ω) vs. log(ρ) (see Methods). Qualitatively, the positively-selected genes (in orange) appear to have larger-than-average values of ω for any given ρ; quantitatively, we assessed this using a linear model that includes an indicator variable. Letting X_i and Y_i denote the log(ρ) and log(ω) values for gene i, respectively, and defining the indicator P_i be equal to one if gene i was deemed to be under recent positive selection and equal to zero otherwise, we tested whether $\gamma = 0$ in the linear model $Y_i = \alpha + \beta X_i + \gamma P_i + \varepsilon_i$. We were able to reject the null hypothesis $\gamma = 0$ when tested against the one-sided alternative $\gamma > 0$ (p-value < 0.00467), concluding that as compared to the overall clustering trend the rates of "mixed-regime" genes appear to be elevated.

The dispersion ratio is a useful predictor of evolutionary rate

Recall that, as depicted in Figure 2a, one interpretation of the dispersion ratio is that it captures the latent segmentation of rate classes within a protein sequence. This segmentation, in turn, may be due to constraints on a protein's structure and function. Viewed in this way, it is not unreasonable to consider the dispersion ratio as a crude but informative surrogate of the structural constraints acting upon a protein. We have provided evidence that this structural surrogate is predictive of the ratio of rates at which a protein evolves (i.e. ω), and we have shown that the clustering measured by ρ is independent of ω when the sequences have been permuted (i.e. in the absence of structuring). In this section,

we investigate how ρ compares with other established correlates of evolutionary rate.

We have structured this comparison to bring it in accord with the literature. The manuscripts we sought to parallel collectively introduce a diverse set of potential correlates of a protein's evolutionary rate. The measures we consider span a wide range of protein-related attributes, including mRNA expression level (Holstege et al. 1998), protein abundance (Ghaemmaghami et al. 2003), translational efficiency (as measured by the codon adaptation index) (Fraser et al. 2004; Wall et al. 2005), dispensability (i.e. fitness when deleted) (Warringer et al. 2003; Deutschbauer et al. 2005), sequence length (Drummond, Raval, and Wilke 2006), the number of protein-protein interaction partners (Han et al. 2004), the protein's contact density (Bloom et al. 2006), the fraction of residues in the protein that are at least 25% buried, and the fraction of residues involved in various secondary structure elements (helix, strand, turn, coil) (Bloom et al. 2006). In addition to correlating these attributes both to $\log(\rho)$ and $\log(\omega)$, we considered each as a controlling variable to test the persistence of a significant log-linear relationship between ρ and ω in yeast.

The yeast dataset we employ comes from (Kellis, Birren, and Lander 2004) and includes annotated protein-coding genes from four *Saccharomyces* species: *S. cerevisiae*, *S. paradoxus*, *S. mikitae*, and *S. bayanus*. We again focused on groups of unique orthologs, and because here for each protein-coding gene we have four sequences instead of two, we were forced to extend the dispersion ratio beyond pairwise comparisons. Our approach was to treat the unrooted phylogeny from (Kellis, Birren, and Lander 2004), ((*S. cerevisiae*, *S. paradoxus*), (*S. mikitae*, *S. bayanus*)), as representing five separate pairwise comparisons to be aggregated (though see Discussion for alternatives). To accomplish this required us to infer the sequences at the internal nodes of the tree, and we did so under a probabilistic model from (Yang, Kumar, and Nei 1995), using the algorithm of (Pupko et al. 2000). For each pair of sequences spanning a branch on the tree, we partitioned their alignment as in Figure 2 to obtain four values: (1) Ka_{A}, the adjacent rate of non-synonymous changes, (2) Ks_{A}, the adjacent rate of synonymous changes, (3) Ka_{I}, the isolated rate of non-synonymous changes, and (4) Ks_{I}, the isolated rate of synonymous changes. Note that whereas before we combined these to compute ρ, here we have kept them separate so that each can be summed

across the tree. In this way, we computed the dispersion ratio for each yeast protein-coding gene as $\left(\sum Ka_I / \sum Ks_I\right) / \left(\sum Ka_A / \sum Ks_A\right)$.

As before, we find a highly significant log-linear relationship between the dispersion ratio and ratio of evolutionary rates. To test whether or not that relationship persists after controlling for the aforementioned protein-related attributes, we used the method of partial correlation. The partial correlation between two variables X and Y, after controlling for a variable (or a set of variables) $\boldsymbol{Z}$, can be found by regressing each of X and Y on $\boldsymbol{Z}$ and then computing the correlation between residuals. We took $\log(\rho)$ and $\log(\omega)$ as X and Y and let $\boldsymbol{Z}$ range across each of the protein-related attributes in Table 2. The results show that the log-linear relationship between selection and dispersion remains highly significant even after controlling for a variety of established evolutionary correlates. The strength of that relationship, in comparison to those observed for other attributes, is remarkable (see Table 2) and suggests that the dispersion ratio is capturing an important determinant of evolutionary rate.

Methods

Genome-wide pairwise comparisons of selection and dispersion

We obtained from Ensembl 46 pairwise codon alignments of all one-to-one orthologous protein coding sequences between human and eight other species: *Pan troglodytes, Macaca mulatta, Mus musculus, Rattus norvegicus, Canis familiaris, Monodelphis domestica, Gallus gallus, and Danio rerio*. As illustrated in Figure 2, we identified the sites in each alignment at which the encoded amino acids were distinct; these comprise the visible subset of all sites where a non-synonymous change has taken place. We labeled as "adjacent" all sites adjacent to, but not necessarily including, any site identified as non-synonymous by amino acid comparison; the remaining sites were labeled "isolated". The complete alignment was then partitioned into its adjacent and isolated components, yielding two disjoint subalignments. Within each genome-wide comparison, individual proteins were excluded from consideration unless the two subalignments each contained both a transition and a transversion event.

We used the method of Yang and Nielsen (2000), as implemented in PAML (yn00, version 3.15), to estimate *Ka* and *Ks* for each complete alignment (no partitioning) and its two subalignments. The subalignment *Ka* and *Ks* estimates were denoted Ka_A and Ks_A, for the Adjacent alignment, and Ka_I and Ks_I, for the Isolated alignment. We obtained from PAML the standard errors for each estimate as well. We computed $\omega = Ka/Ks$ for the complete alignment, $\omega_A = Ka_A/Ks_A$ for the adjacent alignment, and $\omega_I = Ka_I/Ks_I$ for the isolated alignment. The dispersion ratio was calculated as $\rho = \omega_I/\omega_A$. Within each genome-wide comparison, individual proteins were again excluded when either log(ω) or log(ρ) (i.e. the values plotted and compared in Figure 4 and Table 1) had a standard error greater than one. Standard errors were approximated using the delta method as

$$SE_\rho = \sqrt{\left(\frac{SE_{Ka_I}}{Ka_I}\right)^2 + \left(\frac{SE_{Ks_I}}{Ks_I}\right)^2 + \left(\frac{SE_{Ka_A}}{Ka_A}\right)^2 + \left(\frac{SE_{Ks_A}}{Ks_A}\right)^2} \quad \text{and} \quad SE_\omega = \sqrt{\left(\frac{SE_{Ka}}{Ka}\right)^2 + \left(\frac{SE_{Ks}}{Ks}\right)^2}.$$

Saccharomyces data and analysis

We obtained from Kellis et al. (2004) the protein-coding genes and ortholog assignments (grouped by ORFs with unambiguous correspondence) for four *Saccharomyces* species: *S. cerevisiae*, *S. paradoxus*, *S. mikitae*, and *S. bayanus*. We considered only those proteins for which all four sequences were present, and these were aligned using ClustalW and subjected to phylogenetic analysis assuming the fixed unrooted topology ((*S. cerevisiae*, *S. paradoxus*), (*S. mikitae*, *S. bayanus*)). The method of Yang et al. (1994), as implemented in PAML (codeml; version 3.15), was used to jointly infer "ancestral" sequences at the coalescence of *cerevisiae*/*paradoxus* and of *mikitae*/*bayanus*. This facilitated five pairwise comparisons that collectively span the tree: (1) *cerevisiae* vs. *cerevisiae*/*paradoxus*, (2) *paradoxus* vs. *cerevisiae*/*paradoxus*, (3) *mikitae* vs. *mikitae*/*bayanus*, (4) *bayanus* vs. *mikitae*/*bayanus*, and (5) *cerevisiae*/*paradoxus* vs. *mikitae*/*bayanus*. Subsequently, Ka_A, Ks_A, Ka_I, and Ks_I were calculated for each. To compute a dispersion ratio for the tree, we first summed each of these measures across the five branches comprising ((*S. cerevisiae*, *S. paradoxus*), (*S. mikitae*, *S. bayanus*)). The dispersion ratio for each gene was thus given by

$\left(\sum Ka_I / \sum Ks_I\right) / \left(\sum Ka_A / \sum Ks_A\right)$ where each sum ranges over the five aforementioned pairwise comparisons.

Comparing selection and dispersion for genes under recent positive selection

Within the human/chimpanzee dataset gathered from Ensembl, we identified those genes implicated as being under positive selection in the human lineage (Vallender and Lahn 2004; Sabeti et al. 2006). We then fit the model $Y_i = \alpha + \beta X_i + \gamma P_i + \varepsilon_i$, where the response variable Y_i is the log(ω) value for gene *i*, the continuous predictor variable X_i is the log(ω) value for gene *i*, and

$$P_i = \begin{cases} 1 & \text{if gene } i \text{ was under recent positive selection} \\ 0 & \text{otherwise} \end{cases}$$

Comparing the dispersion ratio to established correlates of evolutionary rate

Measures of protein-related attributes in *Saccharomyces cerevisae* were collected from various sources (see Table 2). Careful attention was paid to ensure that we chose exclusion criteria and data transformations consistent with published studies. After exclusion and transformation, each of the protein-related attributes described above was investigated for correlation to both log(ω) and log(ρ) (Table 2, $r_{\log(\omega),X}$ and $r_{\log(\rho),X}$ respectively). Partial correlations were computed between log(ω) and log(ρ) after controlling for each of the protein-related attributes individually (Table 2, $r_{\log(\omega),\ \log(\rho)|X}$).

Discussion and Conclusions

As a protein-coding gene evolves, non-synonymous substitutions do not accumulate uniformly along its sequence. There is heterogeneity among the rates at which individual sites within a protein evolve, and part of that heterogeneity is induced by structural and functional constraints. Though the structural and functional domains that comprise proteins are contingent upon tertiary folding, there is enrichment within domains for residues that are contiguous along the primary sequence. As such, within proteins there exists rate autocorrelation that can be, and has been, exploited to annotate regions of putative importance.

In a pairwise comparison of protein-coding genes, rate heterogeneity manifests in the non-random placement of non-synonymous changes. One expects a dearth of such changes in regions of structural and functional importance and a relative excess where the intensity of selection is less. The aggregation of changes outside of important regions may lead to the appearance that non-synonymous changes are clustering. We speculated that the appearance of clustering would increase with an increasing intensity of selection, and we developed the dispersion ratio to test that hypothesis. Confirming our speculation, we found a highly significant log-linear relationship between the dispersion ratio and ratio of evolutionary rates. This relationship was observed to be robust to both choice of species and degree of evolutionary divergence.

Just as purifying selection acts to cluster substitutions along the sequence of a protein, there is evidence that diversifying selection leads to clustering as well. This led us to consider the case of genes whose modes of evolution differ on sister lineages. In cases when the evolutionary trajectory spanned by a pairwise comparison contains a mixture of purifying and diversifying selection, we hypothesized an effect on the relationship between the dispersion ratio and ratio of evolutionary rates. Having already observed that the degree to which non-synonymous changes cluster is predictive of the ratio of rates at which a protein is evolving, we reasoned that for mixed regimes such predictions would be biased downward. At least for the data we examined, this turned out to be the case: for genes under positive selection in the human lineage, the evolutionary rate ratio estimated from a human/chimpanzee comparison was greater than what the degree of clustering would predict.

To place in perspective the contribution of the dispersion ratio as a predictor of the evolutionary rate ratio, we compared its explanatory power to those of a diverse set of protein-related attributes. In doing so, we found $\log(\rho)$ to be a highly significant and non-redundant correlate of the logarithmic ratio of rates, $\log(\omega)$. The correlation between $\log(\rho)$ and $\log(\omega)$, and its persistence after conditioning on other correlates of evolutionary rate, speaks to either a determinant of evolutionary rate that has not yet been characterized or a deficiency in the way evolutionary rate has been quantified in this particular set of studies. Whatever the case, it appears that non-synonymous clustering is a reliable, non-redundant,

sequence-based predictor of ω.

Because the dispersion ratio behaves differently under neutrality and under purifying selection, and because permutations can be used to populate a sensible null distribution, one can envision using the dispersion ratio in a test of selection. Nevertheless, we did not devise ρ as a statistic to test the behavior of individual genes, and such tests, though conceivable, would likely be underpowered and inferior to existing methods (e.g. (Wagner 2007; Zhou, Enyeart, and Wilke 2008)). Instead, we were motivated by simplicity and proposed the dispersion ratio as an intuitive means of testing the existence of genome-wide evolutionary trends. Other measures of clustering are likely to perform similarly, and indeed we observe similar results to those presented when ρ is replaced by a model-based measure of autocorrelation (taken from (Mayrose, Friedman, and Pupko 2005); data not shown).

The intuition behind our statistic and its relationship to evolutionary rate is grounded in dependencies induced by protein tertiary structure. Though ρ is a function of sequence and not structure the dispersion ratio, like the methods from which it was inspired (e.g. ET, ESF, SWAKK), leverages the fact that adjacent residues in the sequence are structurally proximal. It seems reasonable that a structurally-informed analog of the dispersion ratio would be superior to ρ in validating the hypotheses of this manuscript, but we did not find this to be the case (data not shown). This may be due to, among other possibilities, the limited number of structures available or the manner in which we extended our statistic.

In summary, we proposed a simple statistic that quantifies the degree of non-synonymous clustering in a pairwise comparison, and we did so to test hypotheses about how clustering varies with the ratio of evolutionary rates. We found ample evidence of a strong log-linear relationship, and we tested the robustness and validity of our observations in a number of ways. To investigate the generality, we considered eight vertebrate pairwise comparisons spanning a wide rang of evolutionary divergence, as well as a comparison of four *Saccharomyces* yeast. To investigate potential artifacts, we used as controls both a permutation approach and a synonymous dispersion statistic. To investigate methodological dependence, we considered alternatives to the dispersion ratio, including the idea of simply "counting" synonymous and non-synonymous changes as suggested by Nei and Gojobori

(1986) or Li (1993) (data not shown). In every case, for every comparison, we find that non-synonymous clustering intensifies with increasing purifying selection. The ubiquity of this relationship supports the concept of a loose segmentation model for protein sequences as well as the use of *de novo* annotation methods that have implicitly capitalized upon it.

Figures

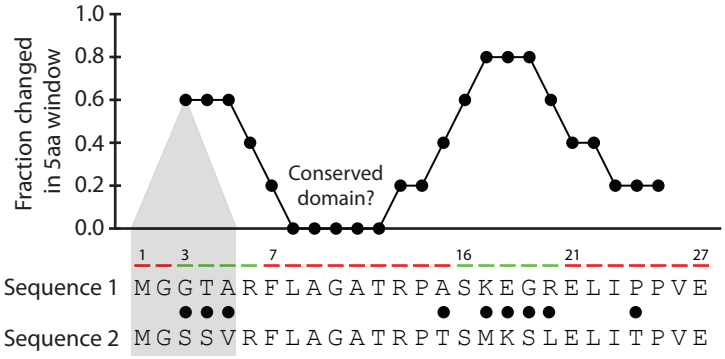


Figure 1: Illustration of simple *de novo* annotation.

Shown is a comparison of two aligned protein orthologs, each of which is 27 amino acids in length. Filled circles between the sequences indicate sites at which the amino acids are distinct. The sequence has been segmented into red (more slowly evolving) and green (more rapidly evolving) regions to illustrate the biological motivation. The figure above the alignment shows, for each of positions 3 through 25, the fraction of mismatched amino acids among positions *i*-2 through *i*+2 plotted as a function of *i* (highlighted for *i* = 3 in gray). The region from positions 8 to 12 shows a deficit of changes, suggesting the possible presence of a conserved domain.

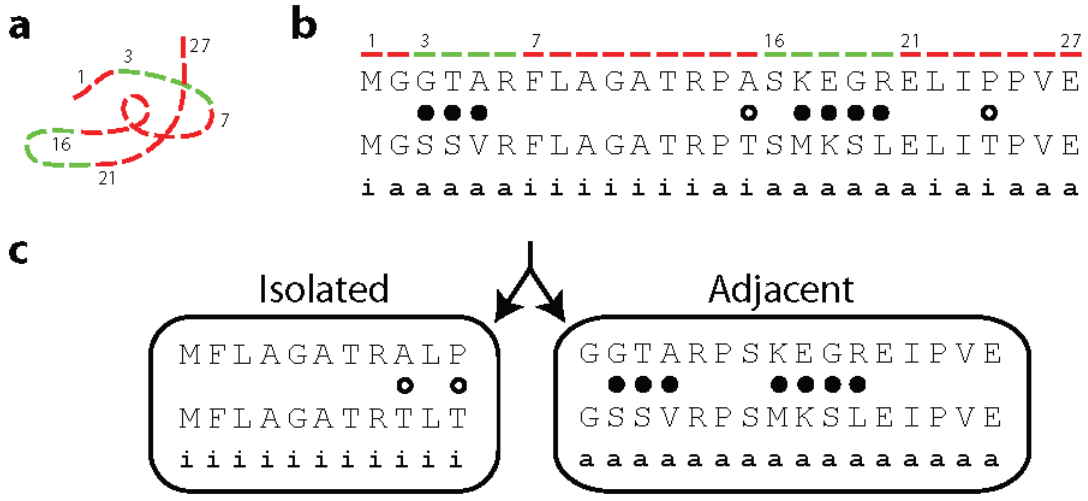


Figure 2: Construction of the Dispersion Ratio

(**a**) A protein sequence of 27 amino acids in length. The sequence has been segmented into red (more slowly evolving) and green (more rapidly evolving) regions to illustrate the biological motivation. (**b**) A pairwise alignment of two protein orthologs of (a). Amino acids are shown rather than their encoding trinucleotides for clarity. Circles indicate sites at which the amino acids are distinct. Sites adjacent to sites at which the amino acids are distinct are labeled with an "**a**"; the remaining sites are labeled "**i**" for isolated. As shown, filled circles denote amino acid differences at adjacent sites, whereas the circles indicating amino acid differences at isolated sites are hollow. (**c**) The alignment is partitioned into its isolated and adjacent constituents, and the selection parameter ω is estimated for each (as ω_I and ω_A, respectively). The dispersion ratio ρ is computed as ω_I / ω_A .

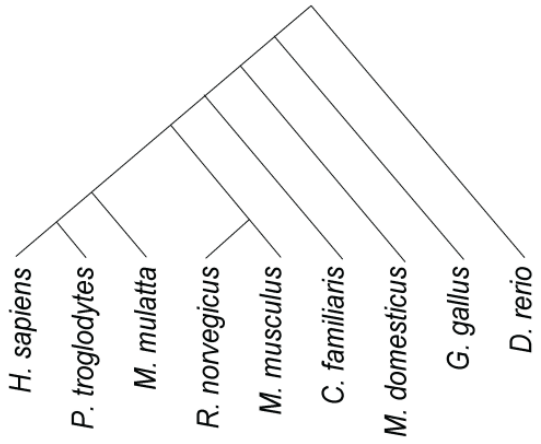


Figure 3: Phylogeny of the eight species considered in pairwise comparisons

From left: human, chimpanzee, Macaca, rat, mouse, dog, opossum, chicken, and zebrafish.

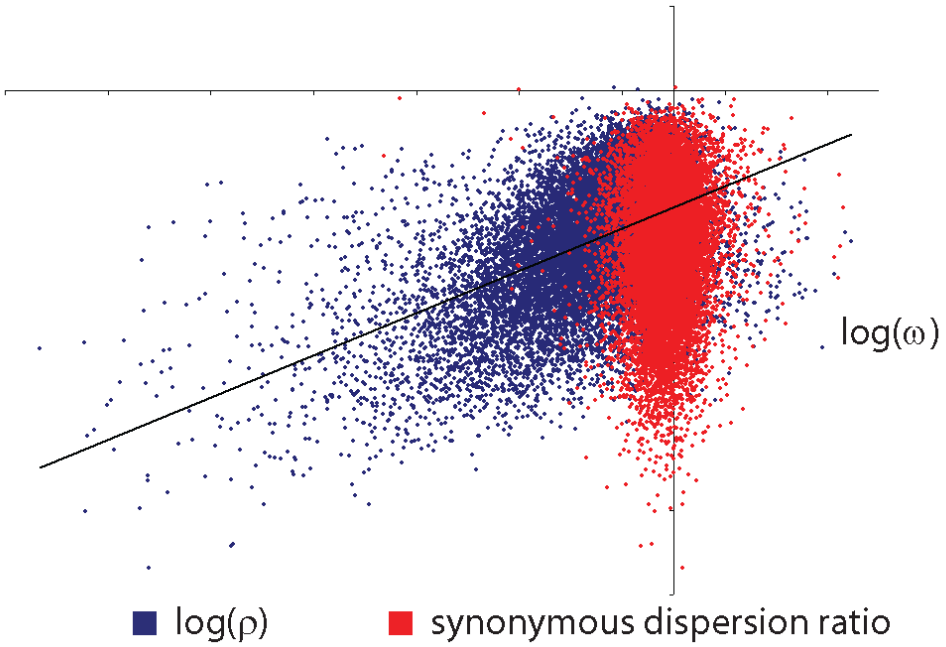


Figure 4: Relationship between measures of selection and dispersion

In blue, log(ω) is plotted against log(ρ) for aligned pairs of orthologous proteins shared between human and mouse (N = 11,894). The plot in red features the same *y*-axis but shows the synonymous dispersion ratio (see text) on the *x*-axis. The linear trend for log(ρ) vs. log(ω) is highly significant (black line; $p < 2.2e\text{-}16$) whereas the plot in red shows no significant linear trend.

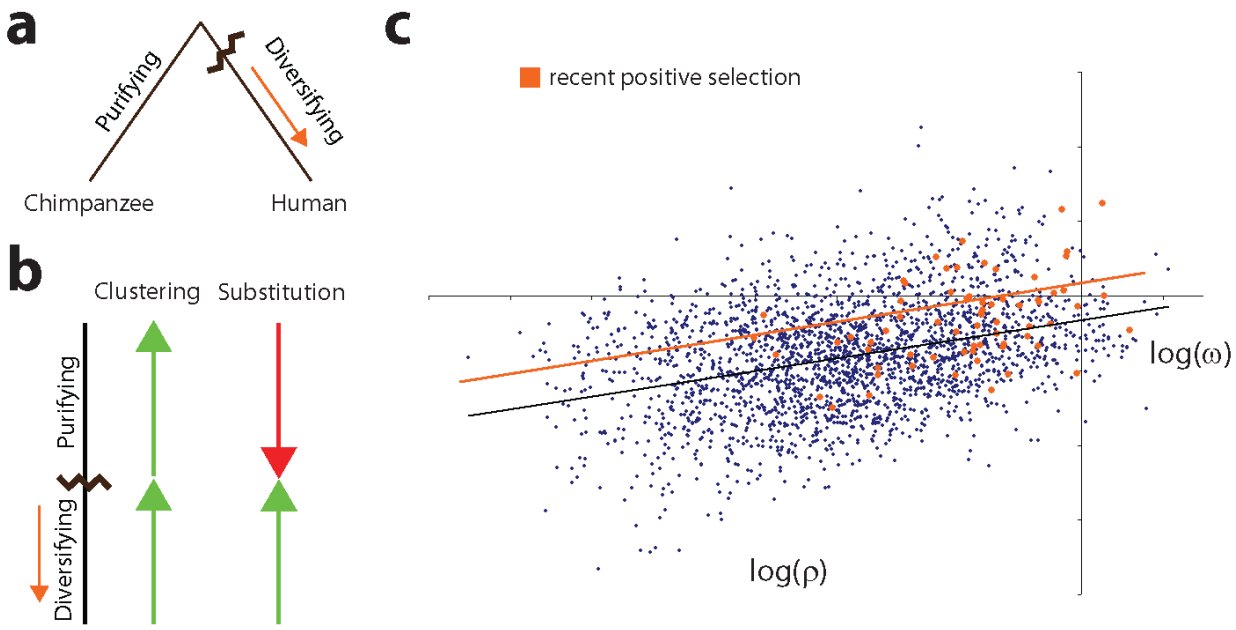


Figure 5: Deviation of genes under recent positive selection in humans

(a) Illustrated mode of evolution for a gene shared by human and chimpanzee that is under positive selection in the human lineage. **(b)** Putative effect of evolutionary mode on intensity of clustering and intensity of non-synonymous change. The rate of substitution is different under diversifying and purifying selection, however both may promote the clustering of changes along the sequence. **(c)** Plot of log(ω) vs. log(ρ) for human/chimpanzee orthologs. Genes annotated by Sabeti et al. (2006) as being under selection in the human lineage are highlighted in orange. In black is the fitted line $y = \hat{\alpha} + \hat{\beta}x$; in orange is the fitted line $y = \hat{\alpha} + \hat{\gamma} + \hat{\beta}x$. Both γ and β were found for be significantly larger than zero.

Tables

Table 1. Genome-wide relationships between log(ω) and log(ρ) for eight pairwise comparisons

Species compared to human	Number of ortholog pairs	Number after exclusion	Correlation Coefficient	Squared Correlation Coefficient
Chimpanzee	16,496	796	0.3094	0.09461
Macaque	16,412	6,255	0.3856	0.1486
Mouse	14,757	11,894	0.3877	0.1503
Rat	14,146	11,401	0.3895	0.1517
Dog	14,794	11,414	0.3991	0.1592
Opossum	13,272	9,381	0.3766	0.1418
Chicken	10,936	5,927	0.3944	0.1554
Zebrafish	7,348	1,419	0.3359	0.1122

Table 2. Correlation and partial correlation between log(ω) and various protein attributes

Attribute (X)	N	$r_{\log(\omega), X}$ (p-value)	$r_{\log(\rho), X}$ (p-value)	$r_{\log(\omega), \log(\rho)\|X}$ (p-value)	Reference
log(ρ)	2,897	0.40077295 (0)	-	-	
mRNA expression	2,701	-0.3807253 (6.7e-94)	-0.2072764 (1.34e-27)	0.3558986 (0)	(Holstege et al. 1998)
Protein abundance	1,930	-0.3878717 (2.572e-70)	-0.1586315 (2.4e-12)	0.3727785 (0)	(Ghaemmaghami et al. 2003)
Codon adaptation index[1]	2,895	-0.3741477 (7.23e-97)	-0.1898758 (6.63e-25)	0.3621437 (0)	(Fraser et al. 2004)
Codon adaptation index[2]	2,643	-0.4055142 (3.568e-105)	-0.2027786 (6.31e-26)	0.3558753 (0)	(Wall et al. 2005)
Sequence Length	2,895	-0.01921694 (0.301313)	-0.01095773 (0.5556)	0.4006604 (0)	(Drummond, Raval, and Wilke 2006)
Dispensibility[1]	1,562	0.1832102 (2.94e-13)	0.09173406 (0.000283)	0.3922312 (0)	(Deutschbauer et al. 2005)
Dispensibility[2]	49	-0.2296285 (0.1124)	0.01099192 (0.94025)	0.4143947 (0.00201)	(Warringer et al. 2003)
Degree	674	-0.1502817 (8.98e-5)	-0.0850535 (0.02724)	0.3938752 (0)	(Han et al. 2004)
Centrality	674	-0.0193294 (0.616415)	-0.03150676 (0.414129)	0.4004375 (0)	(Han et al. 2004)
Contact density	84	0.1411473 (0.2003)	0.05072567 (0.646781)	0.3981061 (9.39e-5)	(Bloom et al. 2006)
Fraction buried 25%	84	0.2146396 (0.04992)	0.184923 (0.09218)	0.3761856 (0.000258)	(Bloom et al. 2006)
SS (helix)	84	-0.1465735 (0.18337)	0.01745651 (0.8748)	0.4077974 (5.8299e-5)	(Bloom et al. 2006)
SS (strand)	84	0.05027238 (0.64973)	-0.05152868 (0.6416)	0.4044114 (6.90152e-5)	(Bloom et al. 2006)
SS (turn)	84	0.07785531 (0.48147)	-0.05314373 (0.6311)	0.406718 (6.1537e-5)	(Bloom et al. 2006)
SS (coil)	84	-0.2148053 (0.04973)	-0.02548217 (0.818)	0.4048788 (6.743566e-5)	(Bloom et al. 2006)

References

Binkley, J., K. Karra, A. Kirby, M. Hosobuchi, E. A. Stone, and A. Sidow. 2010. ProPhylER: a curated online resource for protein function and structure based on evolutionary constraint analyses. Genome Res **20**:142-154.

Bloom, J. D., D. A. Drummond, F. H. Arnold, and C. O. Wilke. 2006. Structural determinants of the rate of protein evolution in yeast. Molecular Biology and Evolution **23**:1751-1761.

Deutschbauer, A. M., D. F. Jaramillo, M. Proctor, J. Kumm, M. E. Hillenmeyer, R. W. Davis, C. Nislow, and G. Giaever. 2005. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics **169**:1915-1925.

Doron-Faigenboim, A., A. Stern, I. Mayrose, E. Bacharach, and T. Pupko. 2005. Selecton: a server for detecting evolutionary forces at a single amino-acid site. Bioinformatics **21**:2101-2103.

Drummond, D. A., A. Raval, and C. O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. Molecular Biology and Evolution **23**:327-337.

Fraser, H. B., A. E. Hirsh, D. P. Wall, and M. B. Eisen. 2004. Coevolution of gene expression among interacting proteins. Proc Natl Acad Sci USA **101**:9033-9038.

Ghaemmaghami, S., W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman. 2003. Global analysis of protein expression in yeast. Nature **425**:737-741.

Glaser, F., Y. Rosenberg, A. Kessel, T. Pupko, and N. Ben-Tal. 2005. The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. Proteins **58**:610-617.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11**:725-736.

Han, J. D., N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature **430**:88-93.

Hardison, R. C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet **16**:369-372.

Holstege, F. C., E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. Cell **95**:717-728.

Joachimiak, M. P., and F. E. Cohen. 2002. JEvTrace: refinement and variations of the evolutionary trace in JAVA. Genome Biol **3**:RESEARCH0077.

Kellis, M., B. W. Birren, and E. S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature **428**:617-624.

Landau, M., I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz, T. Pupko, and N. Ben-Tal. 2005. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res **33**:W299-302.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol **36**:96-99.

Liang, H., W. Zhou, and L. F. Landweber. 2006. SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis. Nucleic Acids Res **34**:W382-384.

Lichtarge, O., H. R. Bourne, and F. E. Cohen. 1996. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol **257**:342-358.

Mayrose, I., N. Friedman, and T. Pupko. 2005. A Gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics **21 Suppl 2**:ii151-ii158.

Mihalek, I., I. Res, and O. Lichtarge. 2006. Evolutionary trace report_maker: a new type of service for comparative analysis of proteins. Bioinformatics **22**:1656-1657.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3**:418-426.

Pollard, K. S., S. R. Salama, N. Lambert, M. A. Lambot, S. Coppens, J. S. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel, A. D. Kern, C. Dehay, H. Igel, M. Ares, Jr., P. Vanderhaeghen, and D. Haussler. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. Nature **443**:167-172.

Pupko, T., I. Pe'er, R. Shamir, and D. Graur. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Molecular Biology and Evolution **17**:890-896.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. Positive natural selection in the human lineage. Science **312**:1614-1620.

Simon, A. L., E. A. Stone, and A. Sidow. 2002. Inference of functional regions in proteins by quantification of evolutionary constraints. Proc Natl Acad Sci U S A **99**:2912-2917.

Takata, M., S. Tachiiri, A. Fujimori, L. H. Thompson, Y. Miki, M. Hiraoka, S. Takeda, and M. Yamazoe. 2002. Conserved domains in the chicken homologue of BRCA2. Oncogene **21**:1130-1134.

Vallender, E. J., and B. T. Lahn. 2004. Positive selection on the human genome. Human Molecular Genetics **13 Spec No 2**:R245-254.

Wagner, A. 2007. Rapid detection of positive selection in genes and genomes through variation clusters. Genetics **176**:2451-2463.

Wall, D. P., A. E. Hirsh, H. B. Fraser, J. Kumm, G. Giaever, M. B. Eisen, and M. W. Feldman. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci USA **102**:5483-5488.

Warringer, J., E. Ericson, L. Fernandez, O. Nerman, and A. Blomberg. 2003. High-resolution yeast phenomics resolves different physiological features in the saline response. Proc Natl Acad Sci USA **100**:15724-15729.

Yang, Z., S. Kumar, and M. Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641-1650.

Yang, Z., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution **17**:32-43.

Zhou, T., P. J. Enyeart, and C. O. Wilke. 2008. Detecting clusters of mutations. PLoS One **3**:e3765.

Chapter 3

Evolution of the Max and Mlx Networks in Animals

Abstract

Transcription factors are essential for the regulation of gene expression and often form emergent complexes to perform vital roles in cellular processes. In this paper, we focus on the parallel Max and Mlx networks of transcription factors because of their critical involvement in cell cycle regulation, proliferation, growth, metabolism, and apoptosis. A basic-helix-loop-helix-zipper (bHLHZ) domain mediates the competitive protein dimerization and DNA binding among Max and Mlx network members to form a complex system of cell regulation. To understand the importance of these network interactions, we identified the bHLHZ domain of Max and Mlx network proteins across the animal kingdom and carried out several multivariate statistical analyses. The presence and conservation of Max and Mlx network proteins in animal lineages stemming from the divergence of Metazoa (circa 1 billion years ago) indicates these networks have ancient and essential functions. Phylogenetic analysis of the bHLHZ domain identified clear relationships among protein families with distinct points of radiation and divergence. Multivariate discriminant analysis further isolated specific amino acid changes within the bHLHZ domain that classify proteins, families, and network configurations. These analyses on the Max and Mlx networks provide a model for characterizing the evolution of transcription factors involved in essential networks.

Introduction

Organism development requires the coordination of complex biological processes involving gene-regulatory, protein interaction, and metabolic networks (Barabási and Oltvai 2004; Siegal, Promislow et al. 2006). Transcription factors (TFs) form important links in such networks by responding to cellular signals, recruiting cofactors to promoter regions, and regulating the transcription of target genes that determine cell function and fate. Hence, the protein and DNA interactions that comprise transcription factor networks are fundamental for proper cellular regulation.

Understanding the evolutionary dynamics of TF networks is critical for discerning the essential components regulating key pathways among organisms. Changes to TF networks are known to appreciably contribute to the morphological and developmental differences observed between related species (Fujimoto, Ishihara et al. 2008; Maerkl and Quake 2009). Such network evolution is characterized by natural selection acting on the individual members as well as their interacting partners. Consequently, different patterns of variability and conservation occur, which can alter network interactions and result in functional divergence. The ability for a TF network to withstand such perturbations over large evolutionary distances indicates the network is functionally robust and is likely vital for important cellular processes (Alberghina, Höfer et al. 2009).

One approach to understanding such complex, biological systems is to reduce them to their individual parts. This reductionist method has been largely successful in cataloging functional annotations. However, TFs often act in concert and their essentiality may depend on their interacting partners or positioning within a signaling cascade (Hlavacek, Faeder et al. 2003; Hahn, Conant et al. 2004; Siegal, Promislow et al. 2006). Such networks that rely on a dependency structure or interaction of multiple components instead lend to a systems biology approach (Ideker, Galitski et al. 2001; Kitano 2002; Barabási and Oltvai). By approaching from a network perspective, the essentiality and plasticity of TFs can be more readily integrated within a biological system for a greater contextual understanding.

One large superfamily of TFs characterized by the basic-helix-loop-helix (bHLH) DNA binding and dimerization domain is critical for development in almost all eukaryotes

(Jones 2004). Individual bHLH proteins form dimer complexes that recognize the 5'-CANNTG-3' E-box binding motif in promoter regions to regulate transcription of diverse gene targets. bHLH proteins are well known to contribute to neurogenesis, myogenesis, heart development, hematopoiesis, cell proliferation, and cell lineage determination (Atchley and Fitch 1997; Massari and Murre 2000; Robinson and Lopes 2000; Jones 2004; Kewley, Whitelaw et al. 2004).

Through modular evolution, multiple domain shuffling events coupled bHLH and other domains to create a functionally heterogeneous set of TFs (Morgenstern and Atchley 1999; Moore, Björklund et al. 2008). Furthermore, gene duplications, gene deletions and changes to the bHLH domain have modified bHLH TF network interactions and altered the complexity of transcriptional regulation (Levine and Tjian 2003; Van Dam, Snel et al. 2008). For example, some bHLH proteins have a leucine zipper region (Z) adjacent to the carboxyl end of the bHLH region that stabilizes dimerization and subsequently restricts interaction between bHLHZ proteins (Dang, McGuire et al. 1989; Orian, van Steensel et al. 2003), while others have a PAS domain to enhance binding specificity (Partch and Gardner 2010).

Herein, we focus on two parallel bHLHZ transcription factor networks that are critically involved in regulating cell growth, metabolism, apoptosis, proliferation, and differentiation, i.e. the Max and Mlx networks (Table 1) (Lüscher 2001). Extensive studies in model organisms such as *Mus musculus*, *Drosophila melanogaster*, and *Caenorhabditis elegans* demonstrate that the Max and Mlx networks have maintained functional similarity over extensive evolutionary time, although they have evolved in terms of their membership and complexity (Lüscher 2001).

Max and Mlx network members, including Max, Myc, Mnt, Mxd, Mlx, and Mondo proteins, are defined by a highly conserved C-terminus bHLHZ domain that specifies dimerization with either the Max or Mlx proteins. Their bHLHZ region is defined by a 13 residue basic region (b1-13), 2 α-helices each consisting of 15 residues (H101-115, H201-215), a 28 residue leucine zipper (Z1-Z28), and a variable length loop (L) (Atchley and Fernandes 2005). Each bHLHZ monomer forms two asymmetric alpha helices (bH and HZ) that can dimerize and fold into a globular, left-handed, four-helix bundle. However,

additional dimerization restrictions and DNA binding preferences exist for each bHLHZ protein.

The ancient condition, represented in the fruitfly *D. melanogaster* exhibits a minimal network consisting of single copies of dMax, dMlx, dMnt, dMyc, and dMondo genes (Table 1; Figure 1c2)(Peyrefitte, Kahn et al. 2001). Nematodes are distantly related to flies and other arthropods in the Ecdysozoa lineage (Budd and Telford 2009) and *C. elegans*, for example, has a markedly different yet clearly orthologous network. This is presumably due to massive gene reduction and rearrangement that occurred in nematodes (Witherspoon and Robertson 2003; Denver, Morris et al. 2004; Coghlan 2005). In *C. elegans*, two Max orthologs (Mxl-1 and Mxl-3) and a single Mlx ortholog (Mxl-2) act as central dimerization partners for the Mad-like ortholog MDL-1 and Myc and Mondo-like protein MML-1, respectively (Table 1; Figure 1c3)(Yuan, Tirabassi et al. 1998; Gallant 2006; Pickett, Breen et al. 2007).

In contrast, Max and Mlx networks in *H. sapiens* and *M. musculus* contain several members, with paralogous families for Myc (c-, L-, and N-Myc), Mxd (Mxd1-4, formerly Mad1, Mxi1, Mad3, and Mad4) and Mondo (MondoA and MondoB), along with single copies of Max, Mlx, Mnt, and Mga genes (Table 1; Figure 1c4) (Gallant 2006). Rodents and humans possess additional Max interacting proteins S-Myc and L-Myc2, respectively, indicating they sustained separate evolutionarily ancient duplication events (Depinho, Hatton et al. 1987; Doskocil 1996). Another c-Myc homolog, B-Myc, exists in the murine lineage and lacks the C-terminal bHLHZ sequence. Consequently, it cannot interact with Max or bind DNA (Burton, Mattila et al. 2006).

Despite differences in network structure, Max and Mlx network member domains and functions remain stable among species (Yuan, Tirabassi et al. 1998; Gallant 2006; Steiger, Furrer et al. 2008). Myc and Mondo family proteins promote gene transcription by interacting with Max and Mlx, respectively, and recruiting a histone acteylase complex to their N-terminus transactivation domain (TAD) (Facchini and Penn 1998; McMahon, Van Buskirk et al. 1998; Billin, Eilers et al. 1999; Dang 1999; Billin, Eilers et al. 2000; de Luis, Valero et al. 2000; Cairo, Merla et al. 2001). In an antagonistic fashion, Mnt and Mxd family

proteins competitively dimerize with Max and recruit a histone deacytelase complex through a N-terminus Sin3 interaction domain (SID) that represses transcription (Hurlin, Quéva et al. 1997). While there are contradicting results regarding Mnt and Mlx dimerization (Meroni, Reymond et al. 1997; Meroni, Cairo et al. 2000; Cairo, Merla et al. 2001), vertebrate Mxd1 and Mxd4 proteins can also heterodimerize with Mlx and potentially antagonize Mondo function (Billin and Ayer 2006). Since Max and Mlx have no intrinsic transcriptional activity, they ostensibly serve as obligate dimerization partners during transitions in transcriptional signaling. Hence Max and Mlx networks differentially regulate gene transcription according to competitive dimerization and reciprocal behavior of protein members (Grinberg, Hu et al. 2004).

Mnt and Mad antagonize Myc in a general and cell specific manner, respectively, by differentially regulating transcription for overlapping gene targets (Hurlin, Quéva et al. 1997; Orian, van Steensel et al. 2003). Such DNA binding specificity arises from protein specific residues that interact with flanking regions of the canonical 'CACGTG' motif. Myc shows a preference for 5'-GC, 5'-CG, or 5'-AG prior to the E-box (Lüscher and Larsson 1999), Mxd1:Max heterodimers prefer an extended 'CCACGTGG' E-box (Rottmann and Lüscher 2006), while MondoB recognizes the carbohydrate response element (ChORE) designated by two 'CACGTG' E-boxes separated by exactly 5 nucleotides (Shih and Towle 1992; Shih, Liu et al. 1995). Moreover, the synthetic lethal interaction of *D. melanogaster* orthologs dMondo and dMyc indicate both are necessary to regulate at least one essential gene involved in cell growth (Billin and Ayer 2006). This orchestration of Max and Mlx network members enables cells to refine the regulation of shared gene targets through a complex system of activation and repression.

The coordinated expression and dimerization of Max and Mlx network members is essential for normal development (Blackwood, Lüscher et al. 1992; Charron, Malynn et al. 1992; Amati, Littlewood et al. 1993; Grandori, Cowley et al. 2000; Shen-Li, O'Hagan et al. 2000; Nilsson, Maclean et al. 2004; Walker, Zhou et al. 2005; Hooker and Hurlin 2006). Mnt and Myc family proteins are essential for proper cell growth (Charron, Malynn et al. 1992; Pierce, Yost et al. 2004; Toyo-Oka, Hirotsune et al. 2004; Benassayag, Montero et al. 2005;

Loo, Secombe et al. 2005; Pierce, Yost et al. 2008), while Mnt, Myc and Mad family proteins are important for cell cycle progression (Amati and Land 1994; Hanson, Shichiri et al. 1994; Hurlin, Quéva et al. 1995; Zhou and Hurlin 2001). In parallel, Mlx and Mondo family proteins are important in growth and energy homeostasis (Billin, Eilers et al. 2000; Meroni, Cairo et al. 2000; Ma, Tsatsos et al. 2005; Ma, Robinson et al. 2006; Sans, Satterwhite et al. 2006; Iizuka and Horikawa 2008; Stoltzman, Peterson et al. 2008). Although not individually essential, MondoA and MondoB are important for proper glucose metabolism and formation of triglycerides (Ma, Robinson et al. 2006; Peterson, Stoltzman et al. 2010).

The relative abundance and activity of member proteins in these two networks is tightly controlled due to the substantial effects of even some minor perturbations (Grandori, Cowley et al. 2000; Hooker and Hurlin 2006). Most notably, deregulation of Myc is directly associated with oncogenesis and attributes to over 70,000 human deaths a year in the US (myccancergene.org). Loss of Mnt can also result in tumor formation (Hurlin, Zhou et al. 2004; Nilsson, Maclean et al. 2004; Hooker and Hurlin 2006), although no significant observations have been able to classify Mad or Mnt as tumor suppressors (Schreiber-Agus, Meng et al. 1998; Rottmann and Lüscher 2006). Moreover, the central role of MondoB in lipid synthesis and glucose response implicates it as a possible contributing factor in fatty liver, obesity, and Type II diabetes (Postic, Dentin et al. 2007)

Parallel and essential regulation of the Max and Mlx networks shows these TFs exhibit distinct characteristics necessary for proper cell development. The homologous bHLHZ domain is integral in distinguishing the preference of protein interactions, complex structure, and gene targets that direct the downstream effects of these transcription factors. Still, the importance and evolution of Max and Mlx interactions is relatively unknown. The function and origin of Max interacting protein Mga has not been formally addressed, distinctions in Mxd function and binding have yet to be determined, and ramifications of Max and Mlx network gene loss in *C. elegans* and *D. melanogaster* are uncertain.

Herein, we employ tools from computational biology to explore the evolution of two networks of important TFs. Using Max and Mlx networks as a model, we investigate how networks involving TFs essential for organism development change during organismal

diversification over extensive evolutionary time and distances. Using phylogenetic and multivariate statistical analyses, we characterize Max and Mlx network interactions in animals by comparing the bHLHZ domain of its members across diverse species. In particular, we address several questions regarding evolution of network structure and the bHLHZ interaction domain. Did network structure diverge in bursts of diversification or through several incremental evolutionary events? Is the DNA binding and protein-protein interaction bHLHZ domain conserved among orthologous members or in particular lineages? And what residues in the bHLHZ domain restrict and distinguish potential dimerization and DNA binding patterns?

Methods

Obtaining and Aligning Max and Mlx Network bHLHZ Sequences

Approximately 100 eukaryotic species were surveyed for Max and Mlx network members. Initial amino acid sequences were obtained from Ensembl (Flicek, Aken et al. 2010) and NCBI (Sayers, Barrett et al. 2010) annotations while sequences of unannotated species were gathered from eukaryotic genomic databases, i.e. JGI (JGI 2010), Baylor (Gibbs 2010), Dana Farber (Quackenbush, Cho et al. 2001), Metazome, Flybase (Tweedie, Ashburner et al. 2009), Vectorbase (Lawson, Arensburger et al. 2009), Sanger (Sanger 2010), Broad (McCarthy 2005), Washington University (2010), Wormbase (Harris, Antoshechkin et al. 2010), and Kegg (Kanehisa, Goto et al. 2010) databases (Table 2). When no known ortholog was available, we performed TBLASTN and BLASTP (Altschul, Gish et al. 1990) queries on relevant databases using known protein sequences of similar species. Validated EST and predicted transcripts were given priority, followed by blast hits on scaffolds and unassembled whole genome shotgun (WGS) reads. A protein was considered absent within a species if distinguishing features in the bHLHZ domain could not be identified manually (Atchley and Fernandes 2005). Note that absence in the database does not necessarily indicate absence within the organism. Rather it could reflect inadequate sampling or sequencing of the genome.

To adequately represent the distribution of species across the Metazoa, our analyses were restricted to a subset of 45 diverse species (19 Deuterostomes: 14 Chordates, 3

Urochordates, 2 Echinodermes; 21 Protostomes: 16 Ecdysozoans, 4 Lophotrochozoans, 1 Trematode; 2 Cnidarian, 1 Placozoan, 1 Porifera, and 1 Choanoflagellate). While the Choanoflagellida lineage is not part of the Metazoa, it is closely related and serves as an outgroup for the animal lineage. Using ClustalW (Larkin, Blackshields et al. 2007), Muscle (Edgar 2004), and Dialign (Subramanian, Kaufmann et al. 2008) algorithms provided similar amino acid alignments of the bHLHZ domain with small deviations in gap location within the loop region. Morgenstern and Atchley (1999) previously described issues with gaps during phylogenetic reconstruction of bHLH sequences and the inability to determine proper homology between proteins for the loop region. To circumvent these problems, we removed the loop and optimized over the bHHZ sequence when comparing different protein families.

Although most of the surveyed species conform to the aforementioned canonical bHLHZ structure (Figure 4), we also removed a few species-specific, systematic deviations. After position b11, one of the two Mxd2 genes in *Tetraodon nigroviridis* contains a Ser (S) insertion, while Mnt in platypus *Ornithorhyncus anatinus* contains a Thr-Leu-Leu (TLL) insertion. After position H5, Mondo in mollusk *Aplysia californica*, MondoB in opossum *Monodelphis domestica*, and one copy of Mxd2 in cow *Bos taurus* have a Thr (T), Trp (W), and Ile (I) insertion, respectively. Platypus Mnt has an additional Gly-Glu-Ala (GEA) insertion after H210, while moth *Bombyx mori* Mnt experienced an Arg-Gln-Val-Leu-Arg (RQVLR) insertion after Z1. Although most loop regions are in the range of 5-13 residues, the predicted transcript of *Hydra magnipapillata* Max has an extensive 55 residue loop. These autapomorphies have been removed from further analysis to focus on cladist information from homologous sites within the defined structure, since their inclusion would only contribute to branch length divergence with no cladistical relevance.

Phylogenetic Reconstruction using the bHHZ domain

A battery of phylogenetic algorithms was employed including Bayesian, maximum likelihood, and distance methods. This was done to expose algorithm bias when determining the phylogenetic relationships within the bHHZ domain among Max and Mlx network proteins. Max and Mlx network members belong to the DNA binding class B 5'-CACGTG-3' E-box binding group, which is suggested to represent the ancestral HLH sequence

(Atchley and Fitch 1997). Since the history of divergence among Max and Mlx members is uncertain, we included several additional class B bHHZ sequences as outgroup sequences for comparison in each phylogenetic analysis of the 352 taxa. These outgroup sequences included *H. sapiens*, *D. melanogaster* and *C. elegans* orthologs of SREBF1, USF2, TCF3, MYOD, and HES1.

Table 3 lists all parameter combinations and programs used for each method. However, the strengths and limitations of the methods vary depending on the model used. For example, distance based methods provide a single phylogenetic tree based on defined similarity measures and optimization criteria. Neighbor Joining (NJ) uses a greedy, bottom up clustering algorithm that optimizes over the total tree length (Saitou and Nei 1987). We estimated several NJ trees based on different models of selection using HyPhy (Pond, Frost et al. 2005). We also used BioNJ (Gascuel 1997), which iteratively reduces the variance of distance estimates for a minimum evolution tree by applying weighted averages. The initial distance matrix required for BioNJ was created using ProtDist of the Phylip package (Felsenstein 2005). Protpars, a parsimony method also developed by Felsenstein (2005) uses stepwise addition to calculate the minimum number of changes required for a given tree until all taxa have been included and the maximum parsimony determined. However, parsimony overlooks underlying biological information and is not a consistent estimation technique (Felsenstein 1981).

Further, we applied a Bayesian approach and several maximum likelihood (ML) methods for statistical comparison of phylogenies. Bayesian and ML methods optimally fit data from a given alignment to a tree topology using a particular model of evolution by maximizing the likelihood function (Wróbel 2008). Typically ML methods use a single stochastic model to represent the evolution of amino acid sequences (Yang 1996; Yang 2007). This is obviously violated in sequences that contain a highly conserved functional domain flanked by unconserved, variable residues. Bayesian analysis and some ML methods allow for mixed models and rate heterogeneity, although the distribution of rates must still be defined *a priori*. Moreover, the Bayesian posterior distribution cannot be calculated directly and must be methodically sampled instead from the tree space (Ronquist and Huelsenbeck

2003). To ensure we adequately sampled possible evolutionary models, we used multiple programs and parameterizations during tree reconstruction.

PAML provides a framework for complex models during ML phylogenetic reconstruction (Yang 2007). However, PAML was unable to build an initial tree due to gaps in the alignment and segmentation faults related to the number of taxa. To circumvent this issue, we provided an initial Neighbor Joining (NJ) tree and estimated the tree topology by nearest neighbor interchange (NNI) (Moore, Goodman et al. 1973) for tree optimization. Comparatively, ProML (Felsenstein 2005) and PhyML (Guindon and Gascuel 2003) use an internal BioNJ method to build an initial tree. PhyML couples stepwise addition with topology rearrangement to simultaneously optimize branch lengths and likelihood probabilities for each iteration of its hill-climbing algorithm. This method claims to reduce computational time while maintaining comparable accuracy levels with other ML approaches. Felsenstein's ML application ProML developed in the Phylip package also computes topologies and maximum likelihood estimates. We also used Mr. Bayes for a comparable Bayesian reconstruction of the phylogeny (Ronquist and Huelsenbeck 2003).

Entropy as a Conservation Score

In the context of protein sequence analysis, entropy measures the amount of information or conservation at a site by the observed distribution of amino acids (Shannon 1948). We calculated the Shannon Entropy for all sites, where $H_i = -\sum_j p_j \log_b p_j$ is the entropy for site i with probability p_j of being in state j. Entropy can be standardized so $H \in [0,1]$ by setting b equal to the number of possible states. Amino acid entropy assumes independence among amino acid states and standardizes by log base b=20. However, treating each amino acid independently does not reflect the similarity in physiochemical properties. To accentuate changes in physicochemical properties at a site, Atchley et al. (1999) developed a functional entropy measure that groups amino acids into eight functional categories then standardizes values by log base b=8. A site with a low functional entropy but high amino acid entropy suggests it is conserved for a particular physiochemical property but not a particular residue.

Transforming Amino Acid Sequences into Metric data using Factor Scores

Statistically rigorous analyses of amino acid variability procedures typically require a numeric representation of the alphabetic amino acid codes in protein sequence data. To interpret the structural and functional attributes of bHLHZ sites, we transformed each amino acid sequence into the 5 multivariate physiochemical metrics proposed by (Atchley, Zhao et al. 2005) that independently describe the multidimensional characters of the various amino acids. Atchley et al. (2005) used factor analysis to distinguish the common and unique variance of approximately 500 amino acid indices. They found five basically orthogonal factors adequately summarized the latent variable structure and denoted these vectors by polarity, accessibility, and hydrophobicity (PAH), propensity for secondary structure (PSS), molecular size (MS), codon composition (CC), and electrostatic charge (EC). Each column in the amino acid alignment is then represented by a five element vector of converted PAH, PSS, MS, CC, and EC values.

Discriminant Analysis of Proteins, Networks, and Binding Partners

To identify the structure of variation in physicochemical properties among proteins, we statistically ranked sites according to their ability to distinguish protein groups by using stepwise linear discriminant analysis (LDA) (Fisher 1936). LDA is a widely used, robust statistical method for discriminating variables among *a priori* defined groups. While canonical LDA considers all variables simultaneously when building the discriminatory model, the stepwise method discriminates groups by iteratively incorporating variables that maximize the between versus within group variance after conditioning on prior variables included.

In order to reveal the impact of natural selection among orthologs in different network configurations, we first grouped all orthologous species sequences by their network topology (Figure 1c; Table 2): 1) core, 2) nematode, 3) Diptera, or 4) vertebrate. Since paralogs may be under different selection pressures, they were considered distinct proteins, e.g. c-Myc, N-Myc, L-Myc, S-Myc, and L-Myc2 were all grouped separately. However, LDA performs poorly on discrete data (Dillon and Westin 1982). Hence we independently used each of the five factor score transformations of the amino acid sequences to annotate

sites according to these distinct physiochemical properties. Gaps in the alignment were replaced by zeros, although imputing missing residues gave comparable results (data not shown). Implemented in SAS, stepwise LDA produced an ordered list of sites along with the average square canonical correlation (ASCC), which explains the cumulative amount of among class variance explained by the included sites (Table 4).

Results and Discussion

Myc, Max, Mnt, Mad, Mlx, and Mondo proteins comprise the basic members of the Max and Mlx interaction networks found throughout the Metazoa. The presence of an identifiable bHLHZ sequence for Max and Mlx network members in all animals surveyed emphasizes the importance of these ancient transcription factors (Figure 1b). Using several multivariate statistical analyses (i.e. phylogenetics, entropy, and LDA), we identified protein specific residues within the bHLHZ domain that potentially restrict gene targets and influence patterns of transcriptional regulation.

Max and Mlx Network Protein Presence/Absence in Metazoa

Protein sequences from approximately 100 species were obtained from an array of genome databases using sequence annotations, predicted transcripts, and significant blast hits (see Methods). We used well-defined and highly conserved sequence of the bHLHZ domain of Max and Mlx network members to ascertain if the various proteins occurred in a given organism. As shown in Figure 1b, we identified core network members ('X') in almost all surveyed species and predict their existence (blue) even if a particular member was only partially found ('*') or unidentifiable (blank). Exceptions occur when a gene has been experimentally validated as missing ('0') (Yuan, Tirabassi et al. 1998; Gallant 2006), and we conjecture that consecutive absences are deletions (grey). However, absence in the database does not denote absence in the organism, since several of the queried genome assemblies are still in draft or assembly phase with low coverage (Table 2).

Lineage specific radiation and deletion of Max and Mlx network components resulted in four main network configurations in animals (Figure 1c). At the stem of Bilateria and Radiata divergence, six core proteins represent the ancestral Max and Mlx network topology. This core topology consists of Max, Mlx, Myc, Mxd, Mnt, and Mondo proteins, for which all

animals surveyed contain at least one identifiable bHLHZ sequence. However, nematodes, flies, and vertebrates have distinct topologies and derived configurations.

Organisms that diverged near the root of the Metazoa can provide significant insight into the origin and evolution of network members. *Trichoplax adhaerens* of the Placazoa lineage is the simplest known animal with the smallest known genome (Srivastava, Begovic et al. 2008), while the choanoflagellate *Monosia brevicollis* is one of the closest single celled organisms related to animals (King, Westbrook et al. 2008). The presence of Myc and Max in both *Trichoplax* and *Monosiga* strongly implies that these proteins have ancient roots and are important for basic cellular function. Max, Myc, Mxd, Mlx, and Mondo bHLHZ sequences were recovered in *Trichoplax*, while the first identifiable instance of Mnt occurs within the Cnidaria and Bilateria lineages. Hence, the origin of the Max and Mlx networks dates to over 500 mya and was established prior to or during the divergence of animals.

Flies and nematodes are the only known organisms to be missing a core network member (Gallant 2006). Previous reports of yeast two-hybrid assays, interaction screens, and genome searches indicate flies lack a Mxd gene, while nematodes are missing both Mnt and Myc orthologs (Yuan, Tirabassi et al. 1998; Gallant 2006). We observe that fruitfly *Drosophila melanogaster* and the mosquitoes *Aedes aegypti*, *Anopheles gambiae*, and *Culex pipens* lack an identifiable Mxd sequence, while the moth *Bombyx mori* possesses an orthologous bHLH sequence. This reinforces the idea that Mxd loss is specific to the Diptera lineage. We could not find Mxd in the ticks *Boophilus microplus* or *Ixodes scapularis*, indicating ticks, which are part of the Arachnida lineage, may have also independently lost Mxd. Similarly, nematodes *C. elegans*, *C. briggsae*, and *Brugia malayi* do not have Myc or Mnt orthologs. Instead these species along with the trematode *Schistosoma mansoni* contain similar yet divergent orthologs for Max, Mlx, Mxd, and Mondo.

In contrast, two whole genome duplication (WGD) events during vertebrate divergence (Dehal and Boore 2005) ostensibly resulted in the radiation of Myc, Mxd, and Mondo proteins. Only a singe copy of Max, Mlx, and Mnt exist in vertebrates despite multiple duplication events, suggesting the regulation of these proteins is highly controlled by natural selection. In contrast, Myc has experienced additional, independent duplication

events. Approximately 35-50 mya new and old world primates, but not prosimians, exhibit a duplication of L-Myc denoted L-Myc2 (Morton, Nussenzweig et al. 1989; Arnason, Gullberg et al. 1998). Since L-Myc2 is intronless, it presumably arose via a reverse transcriptase event. The murine lineage, including mouse and rat, also exhibit a duplication of N-myc, forming Myc family member S-Myc. The presence of both the 5' and 3' UTR and absence of conventional N-Myc introns suggests S-Myc was formed by an N-Myc cDNA sequence reintegrating into the genome. The murine lineage also contains a N-terminal homolog of c-Myc, named B-Myc, that lacks the bHLH domain and cannot interact with Max or bind DNA (Burton, Mattila et al. 2006).

Another Max network member, Mga, also arose during vertebrate divergence. Mga is predicted to be a Myc family member since its bHLHZ domain is most similar to c-Myc (Hurlin, Steingrìmsson et al. 1999). However, the origin of Mga is ambiguous due to issues with genome coverage and prediction for the 12,189 bp transcript.

Like other network members, Mga has a C-terminus bHLHZ domain with conserved sites in the basic region responsible for E box recognition. However, it also contains a second DNA recognition domain in its N-terminus that recognizes the DNA Brachyury T-box motif (Hurlin, Steingrìmsson et al. 1999). Unlike the characteristic exon structure in other T-box proteins, the T-domain in Mga lacks introns, implying it was inserted via reverse transcription.

Branchiostoma floridae (lancelet or amphioxus) of the cephalochordate lineage contains a sequence with 33.8% identity and 53.8% similarity to Mga in humans over its bHLHZ domain. However, the *B. floridae* sequence does not contain a T-domain. Instead, this 11,851 bp hypothetical transcript contains a second N-terminus bHLHZ domain. Since the divergence of *Branchiostoma* was prior to the vertebrate WGD events (Putnam, Butts et al. 2008; Kawashima, Kawashima et al. 2009), Mga may have arisen independently whereby the T-domain insertion into this ancestral duplicate altered the transcript 5' end. Alternatively, Mga truly arose during the radiation in vertebrates and is a divergent member of the Myc family.

While the Diptera and Nematoda lineages represent experimentally validated gene loss, other instances of member absence may simply be the result of missing data. For example, our criterion for protein identification reports the chicken *Gallus gallus* ortholog Mxd3 as absent, although it is likely to exist in the genome. We found the 5' UTR of Mxd3 overlaps the 3' UTR of the Prelid1 gene in all vertebrates sampled. Sequencing in this region in *Gallus* is of poor quality with non-overlapping contigs. By searching the HGTS database in NCBI, we identified fragments of Mxd3 in BAC clone AC195499 (Wilson 2007). Although we were unable to amplify Mxd3 based on primers of this sequence (data not shown), the conservation of identifiable sequence fragments provides strong evidence that Mxd3 exists and is functional in chicken.

Myc, Mxd, and Mondo Family Genes exhibit Synteny in Vertebrates

The syntenic region around paralogs gives evidence for the regional conservation of duplications and suggests an order of divergence. As shown in Figure 2, Mxd3 is genetically linked with mitochondrial precursor protein Prelid1 (Fox, Stubbs et al. 2004) and an unannotated protein similar to zinc finger protein ZFYVE28. Similarly, Mxd4 is associated with ZFYVE28 in *Monodelphis domesticus* (opossum), *Gallus gallus* (chicken) and *Xenopus tropicalis* (clawed frog), and also linked with the second copy of Prelid1 in the pufferfish *Tetraodon nigroviridis*. This synteny suggests that Mxd3 and Mxd4 are within similarly conserved and paralogous genetic regions. Mxd1, Mxd2, and Mxd4 paralogs are also genetically linked with the three member ADD family of cytoskeleton proteins (Anong, Franco et al. 2009). The relative orientation of these genes supports evidence that these families radiated during the two WGD events that occurred prior to vertebrate divergence.

The Myc family of proteins is syntenic with the FAM84(A,B) and FAM49(A, B, C) families associated with DNA repair and unknown functions, respectively (McDonald, Pavlova et al. 2003). Mga has been proposed to be a Myc family member (Hurlin, Steingrìmsson et al. 1999), although we found no paralogous families that corroborate this speculation. c-Myc and N-Myc are genetically linked with FAM84 and FAM49 homologs, while L-Myc is in proximity to only FAM49C. Since L-Myc is not essential for viability

(Hatton, Mahon et al. 1996), dispensable promoter elements may affect the selective pressure on surrounding genes.

Although knockout studies in mice indicate both c-Myc and N-Myc are essential for growth (Charron, Malynn et al. 1992; Davis, Wims et al. 1993; Moens, Stanton et al. 1993; Sawai, Shimono et al. 1993), we were unable to identify N-Myc in opossum. Chromosomal rearrangements show N-Myc is no longer flanked by FAM84A and FAM49A, which are located within 4Mb of the distal end of opossum Chromosome 1 and 20Mb upstream, respectively. Hence opossum N-Myc may have been lost during this translocation and N-Myc may be conditionally dispensable.

In contrast to Myc and Mxd protein families, the Mondo family contains only two paralogs despite their coincidental emergence during vertebrate divergence. The origin of MondoA and MondoB duplication can be extrapolated from their genetic linkage with BCL7(A, B, C), CLIP(1, 2, 3, 4), and VPS37(A, B, C, D) protein families. The most recent common ancestor of the four paralogs in CLIP and VPS37 dates to the origin of vertebrates (Flicek, Aken et al. 2010). However, no combination of VPS37A, VPS37C, CLIP3, CLIP4, and BCL7C are genetically linked in pufferfish, clawed frog, chicken, opossum, or human.

Max and Mlx Network bHHZ domains show clear Phylogenetic relationships

Variable selective pressures among homologs in different lineages may cause inferred evolutionary relationships to differ from the order of divergence. Phylogenetic trees display the association of multiple taxa by grouping sequences according to a measure of similarity (Hedges 2002). Using phylogenetic reconstructions, we infer the relationship and divergence of the homologous bHHZ domain to determine the relative importance of DNA binding and dimerization among Max and Mlx network proteins.

We used several Bayesian, maximum likelihood, and distance based phylogenetic methods to diversify reconstruction strategies and compare the resulting optimal phylogenetic trees (see Methods, Table 3). Since the root of the tree is unknown, we also included orthologous bHHZ sequences for SREBF1, USF2, TCF3, MYOD, and HES1 to compare the relationship to outgroup sequences.

For all tree methods, orthologous protein sequences were consistently grouped within a clade, although the divergence structure of individual taxa had slight differences (Figure 5). In most cases, trees only differed in the order of species divergence within a particular protein group as a result of forcing bifurcations among taxa with highly similar, if not identical sequences. However, the relationship between protein groups showed slight variability among methods, which we classify into three general types of tree topologies (Table 3, Figure 5).

The PhyML tree using a Gamma distribution with 4 rate categories had the highest maximum likelihood (Figure 5A), and several other reconstructions had similar topologies. This topology groups Mga as a distinct clade closely related to Mlx and Mondo. In comparison, a similar tree was constructed by the BioNJ method using the PMB rate matrix (Figure 5B), although this topology group associates Mga with Myc. Most tree reconstructions resemble these topologies, where the Mnt and Mxd clade and Mondo and Mlx clade are distantly related while Max, Myc, Mga, and outgroup sequences are at intermediate distances. In particular, Max is closer to Mondo and Mlx, Myc is closer to Mnt and Mxd, and the positioning of Mga and outgroups varies between trees.

The third topology type consists solely of the Bayesian reconstruction using a gamma rate distribution (Figure 5C). In this tree Mga and Myc share a clade, while all outgroup sequences are within a single clade with Mlx and Mondo. During Bayesian analysis, the chains representing different tree reconstructions failed to converge due to variability among tree topologies regardless of parameter combinations and starting trees. The smallest standard deviation (SD=0.101) occurred with 1M generations, default posterior probability for chains (T=0.2), and estimating the proportion of invariant sites starting from a given NJ tree. High conservation within and large distances between protein families likely caused these convergence issues due to variable, unresolved topologies within each chain. Hence we conclude type A and B topologies more likely represent the true tree.

Still, orthologous sequences for each protein formed distinguishable groups in all phylogenetic reconstructions. This indicates the bHHZ sequences of Max, Mlx, Myc, Mondo, Mnt, Mxd, and Mga have distinct sequence attributes that contribute to their

dimerization and DNA binding with similar patterns of conservation that have been retained over millions of years of evolution. Mondo and Mlx as well as Mnt and Mxd proteins form distinct sister clades in all tree reconstructions. Since these proteins have an equally ancestral origin, their relationship is likely due to functional and structural constraints. Moreover, outgroup sequences were consistently separate from Max and Mlx network member clades in type A, although USF2 and SREBF1 often grouped alongside Max in type B topologies. Rooting the tree with any of these outgroup sequences forms distinct lineages separating Mlx and Mondo from Myc, Mnt, and Mxd (Figure 6).

Consistent branching patterns also depict distinct groupings among paralogs for vertebrate protein families (Figure 6). Mxd1 and Mxd2 form sister clades as does Mxd3 and Mxd4. This finding supports the shared conservation and descent of these paralogs through two WGDs. MondoA and MondoB also form distinct sister clades suggesting these paralogs retain features similar to their ancestral sequence while accumulating distinguishing characteristics. L-Myc and N-Myc are closer paralogs than c-Myc, which agrees with previous findings (Atchley and Fitch 1995), and all are distinguishable from the invertebrate orthologs. This relationship is in contrast to the syntenic regions and attests to the congruence between c-Myc and its ancestral form. Mga bHHZ sequences are tightly grouped, although their relationship with other proteins varies among tree constructions and largely defines the distinction between type A and B topologies.

Sequences from each protein group also exhibit branching patterns largely analogous to speciation events (Figure 6). Branching of nematode sequences, however, do not correspond with the order of species divergence. MML-1, Mxl-1, Mxl-2, Mxl-3, and MDL-1 bHHZ sequences show large divergences from Max and Mlx network members despite their clearly orthologous proteins. We identified one Max ortholog in *Schistosoma* that is closely related to Mxl-3 and a Mlx ortholog similar to Mxl-2. Mxd in *Schistosoma mansoni* is an outgroup of both Mxd and Mnt clades. Thus Mnt may truly be lost in this lineage whereby Mxd contains binding functions attributable to both proteins. Nematode MDL-1 orthologs are more closely related to Mxd, which signifies a potential loss of Mnt function in this lineage. Moreover, the bHHZ domain of MML-1 is most similar to Mondo proteins while its

binding partner Mxl-2 is an outgroup for Mlx. Hence the Mlx network is conserved in nematodes and the antagonistic behavior of Myc and Mnt transcriptional regulation is presumably lost.

The bHLHZ domain exhibits site specific constraint

To quantify amino acid variability at sites, we compare Shannon Entropy values (Shannon 1948), where low entropy signifies site conservation and high values represent variation. This standardized amino acid (H_{AA}) entropy treats all changes equally to stress the conservation of a particular amino acid. However, some amino acids are functionally and structurally similar and confer comparable functional attributes, e.g. leucine, isoleucine, and valine. Hence we also use a functional (H_{FG}) entropy value developed by (Atchley, Terhalle et al. 1999) based on eight groups of amino acids, which accentuates similarity between amino acids and the variability in functional changes.

We find several highly conserved sites within the bHLHZ domain known to be responsible for DNA binding and stable dimer formation. As seen in Figure 3, sites b5, b9, b12, b13, H110, and H205 have H_{AA} entropy values close to zero and are thus highly conserved in all Max and Mlx network members. Sites b5, b9, and b13 of Max and Mlx network members make base contacts with DNA that restrict binding to the class B 5'-CACGTG-3' E-box motif (Ferré-D'Amaré, Prendergast et al. 1993; Nair and Burley 2003), while the helical structure creates a surface consisting of sites b1, b2, b6, b10, b12, and b13 that make phosphodiester backbone contacts (Lüscher and Larsson 1999; Nair and Burley 2003). Buried sites and specific amino acid interactions further direct structural conformation during dimerization. Site H110 is a buried site that interacts with H204 and H205, while H114 packs against sites H212 and H213 in Max (Atchley and Zhao 2007).

Low H_{FG} entropy values at sites b2, H103, H104, and H215 denote particular amino acid attributes are important for these sites, although a specific amino acid is not required. Hence the structural restrictions on buried site H103 and phosphate backbone contacts by H104 slightly varies between proteins and may distinguish binding abilities (Atchley and Zhao 2007). Crystal structures further show that H215 interacts with its symmetry mate in Max (Ferré-D'Amaré, Prendergast et al. 1993). Similarly, the conservation of leucine heptad

repeats necessary for stable dimerization is shown by the relative decrease in entropy for sites Z14 and Z21 within the zipper.

Site conservation and distinguishing residues are clearly seen in the predicted HMMER sequences shown in Figure 4 (Durbin, Eddy et al. 1998). HMMER uses a profile hidden Markov model (HMM) to probabilistically infer the most likely residue at each site. Note that the majority of conservation (upper case, bold, blue) is within the basic region as well as sites that flank the loop. This is likely due to specific restrictions on dimerization, DNA binding, structural conformation, and stability.

bHLHZ sites can distinctly classify Max and Mlx Network Proteins

According to crystal structures, sites b3, b7, b10, and b11 point away from the DNA major groove and interact with regions outside the E-box (Nair and Burley 2006). These sites were found to be distinctly conserved among the Myc, Mxd, Mnt, and Max sequences in vertebrates and can differentially influence cellular transformation (O'Hagan, Schreiber-Agus et al. 2000). An arginine (R) at site b3 in human c-Myc abrogates DNA binding when there is a 5'-T immediately flanking the E-box because its molecular size prevents stable contacts in the major groove (Solomon, Amati et al. 1993). Moreover, an Arg-to-Ser (R3S) mutation at c-Myc b3 reduced c-Myc transformation capabilities, while a second mutation (V7E) at b7 partially restored the oncogenic potential (O'Hagan, Schreiber-Agus et al. 2000).

Interestingly, Myc b3 and b7 are variable in invertebrates with site b3 predominantly consisting of small and tiny amino acids (SNA). Site b10 also shows discriminatory power among Max and Mlx members; Mxd and Mnt have lysine (K), Mga, Max, and Myc have arginine (R), and Mondo and Mlx possess a glutamine (Q). These distinctly conserved sites potentially distinguish binding constraints among proteins and determine their overlapping or distinct gene targets.

Residues outside of the basic region and higher order conformations also affect DNA binding restrictions. Based on co-crystal formations and analysis of gene target promoters, Myc:Max heterodimers are expected to form a head to tail tetramer complex that recognizes two E-box motifs separated by approximately 100 basepairs and bends the DNA (Nair and Burley 2006). Stabilization of the Myc:Max heterotetramer structure results from extensive

hydrogen bonds and salt bridges formed by residues within the Myc zipper at sites Z11, Z15, Z17, Z18, Z19, Z22, Z23, and Z26 (Nair and Burley 2003). In human c-Myc, Z11:Glu forms polar contacts with Z15:Arg and Z18:Arg. While most species have polar residues at these sites, they are not highly conserved and human c-Myc is the only sequence to have a negatively charged residue at Z11.

It is similarly hypothesized that two tandemly arranged MondoB:Mlx heterodimers are required to stabilize binding with the ChORE element (Ma, Sham et al. 2007). Mutation experiments verified that the loop region of Mlx but not MondoB specify this interaction. Large hydrophobic residues L8:Phe (F) and L10:Ile (I) are predicted to create a favorable protein interaction interface, while basic residue L14:Lys (K) neutralizes electrostatic charges with the DNA backbone (Ma, Sham et al. 2007). Although L14:Lys (K) is highly conserved, only vertebrates have L8:Phe (F) in their extended 15-residue loop. Instead, arthropods have a 13-residue Mlx loop, the Mxl-2 loop has only 11 sites, and the Mlx loop is variable in other invertebrates.

While the zipper region also exhibits variability, multiple mutation studies have found it confers interaction preferences and is essential for dimerization (Reddy, Dasgupta et al. 1992; Arsura, Deshpande et al. 1995; Orian, van Steensel et al. 2003). Sites Z17 and Z18 form antiparallel contacts between monomers during Max dimerization and were found to deviate significantly in human Mxd1 and c-Myc (Nair and Burley 2003). The neutral charges of Z17:Gln-Z18:Asn (QN) in human Max allow homodimerization, yet cause flaring compared to the more stable interaction with positively charged residues Z17:Arg-Z18:Arg (RR) of c-Myc and complementary hydrogen bond interactions with Z17:Glu (E) of Mxd1. Hence Max more readily dimerizes with c-Myc or Mxd1 instead of homodimerizing (Nair and Burley 2003; Grinberg, Hu et al. 2004).

Sites Z17 and Z18 are invariant in Max, except for *Trichoplax* Max and nematode Mxl-1 and Mxl-3. Similarly, Mxd Z17:Glu-Z18:Gln (EQ) is largely conserved in all Mxd sequences, although Mxd4 Z18 is conserved for His (H) and Mxd3 Z17 varies between positively (KR) and negatively (ED) charged residues. Similarly, Myc Z17 is mainly composed of positively charged residues and Z18 is polar. In contrast, the Z10:Asp-Z15':Glu

(DE') and Z17:Lys-Z22':Arg (KR') repulsive forces in *C. elegans* Mxl-1 prevent homodimerization, where ' marks the opposing monomer (Yuan, Tirabassi et al. 1998). These patterns of conservation imply Myc, Max, and Mxd dimerization preferences are largely conserved among all species apart from deviations in nematode interactions.

Still, several residues within the Mxd bHLHZ have previously been documented as unique to the Mxd family (Yuan, Tirabassi et al. 1998). These distinctly conserved residues include H106:Cys (C), L16:Thr-H201:Thr-H202:Leu (TTL), H211:His-H212:Ile (HI), and Z17:Glu-Z18:Gln (EQ). However, site H6:Cys is not Mxd specific as Mnt is invariant for cysteine and Mxd4 contains a tyrosine (Y) in all sampled species. Additionally, conservation of H211:His-H212:Ile applies only to the Mxd duplicates in vertebrates, since Phe-Ile (FI) is conserved among arthropods and variable otherwise. Our results confirm the conservation of L16:Thr-H201:Thr-H202:Leu in all Mxd orthologs including MDL-1, with comparable conservation of L16:Ser-H201:Asn-H202:Leu (SNL) in Mnt. This differs from Myc variability between alanine and proline at L16 and invariability of lysine and valine at sites H201 and H202, respectively. Similarly, Mondo and Mlx are highly conserved at sites H107 (F/Y) and H202 (A/A). Strict amino acid conservation at these sites conveys their specific role in structure and function, such as the van der Waals contacts site H107 forms with H201 and H204 (Atchley and Fernandes 2005). Together sites H107, H201, and H202 discriminate the Max and Mlx protein groups and reveal their potential involvement in distinguishing protein structures.

Network Topologies have distinct bHLHZ sequences

Variations in network topology may also impose disparate restrictions on Max and Mlx network members. To infer potential structural or functional differences among the major species groups, we examine protein orthologs in 1) core, 2) nematode, 3) Diptera, and 4) vertebrate networks and identify discriminating sites among the network topologies. Since the alphabetic nature of amino acid sequences does not provide a basis for rigorous statistical procedures, we transformed each aligned protein sequence into five biologically relevant physiochemical metrics (Atchley, Zhao et al. 2005). This permits the residues within each amino acid sequence to be compared according to their multidimensional physicochemical

properties, i.e. polarity, accessibility and hydrophobicity (PAH), propensity for secondary structure (PSS), molecular size (MS), codon composition (CC) and electrostatic charge (EC). Stepwise discriminant analysis was performed on orthologous proteins using each metric separately to identify the best discriminating sites among networks (Table 4).

Nematode sequences showed the greatest amount of divergence for all orthologous proteins. Using the protein structure prediction program 3DJigsaw (Bates, Kelley et al. 2001), we predicted the structure for Mxl-1:MDL-1, Mxl-2:MML-1 and Mxl-3:Mxl-3 dimers based on PDB structures 1NLW, 1NKP, and 1HLO, respectively (Nair and Burley 2003) (Brownlie, Ceska et al. 1997). This allowed us to view the relative location of invariant residues and nematode specific sites within the dimer complex (Figure 7). The proximity of hydrophobic residues H106:His (H) and H203':Leu (L) in Mxl-3 may strengthen monomer interactions as compared to the polar H106:Ser (S) and H203:Gln (Q) residues conserved in Max sequences of other species. Discriminating sites in Mxl-3 appear to face away from the DNA and dimer interface, while distinct changes in Mxl-1 occur throughout the DNA and protein-binding region. This further suggests that Mxl-1 is divergent from Max and may confer variable transcriptional regulation for MDL-1. MDL-1 experienced only a few changes, which are also present in other vertebrate Mxd family members. Specifically, MDL-1 b4:Ala (A) and Mxd3 b4:Val (V) similarly changed to nonpolar residues while MDL-1 H102:Asn (N) and Mxd3 H102:Gln (Q) replaced positively charged residues.

Nematodes also exhibit distinctions in Mxl-2 and MML-1 interacting partners. Nematode Mlx-2 shows disparity at nearby sites H111:Lys (K), H201:Asn (N), and H206:Phe (F) compared to the otherwise conserved H111:Gln (Q) and H201:Lys (K) sites observed in other Mlx sequences. Meanwhile, MML-1 has a contrasting surface consisting of sites b11:Asn, H102:Ala, H105:Asp, and H109:Gln (N, A, D, Q) that faces away from the dimer complex. These differences in nematode orthologs accounts for the majority of variability among network members (Table 4).

In contrast, Max bHLHZ is highly conserved, with an expected 0.003 amino acid difference per million years, which is 16 times lower than that for Myc bHLHZ (Atchley and Fitch 1995). Interestingly, both Max and Myc bHLHZ domains required numerous sites to

explain at least 90% of the variability between network configurations. Since Max is a highly conserved sequence with minimal variation and Myc contains multiple changes that overlap network topologies, there was little structured variability for which DFA could easily distinguish classes. No sites were able to directly discriminate Max in the Diptera network, and only sites H108:His (H), H206:Asp (D), Z5:His (H), and Z19:Ala (A) showed any power in discriminating Max vertebrate sequences due to their changes in codon composition and charge. While these sites have not been previously annotated for conserved structure or function, the proximity of negatively charged H108:His (H) and positively charged H206':Glu (E) on opposing Max monomers may form stable contacts in vertebrates. In other species the charge of Max H108 is largely neutral while Max H206 is positive.

Myc also exhibited only minor differences between networks, with the Diptera lineage mostly discriminated by changes in hydrophobicity. *D. melanogaster* Myc b3:Asn (N) and H203:Asn (N) lost, while H111:Lys gained hydrophobic properties compared to almost all other species. Site H102 differed in both Diptera Myc and vertebrate c-Myc compared to the otherwise conserved Asp (D) residue, where cMyc H102:Glu (E) is bigger and dMyc H102:Gly (G) is smaller and not negatively charged. Sites b4:Thr (T), H108:Phe (F), and Z22:Lys (K) also discriminated c-Myc, while L-Myc displayed differences in aromatic b7 and neutrally charged H206. Interestingly, N-Myc showed overlapping similarities with either L-Myc or c-Myc at these residues and had no significantly discriminating sites of its own.

Primarily, residues within loop and zipper regions discriminated Mlx and Mondo orthologs among the core, Diptera, and vertebrate networks. Vertebrate paralogs MondoA and MondoB have H215:Ser (S) instead of proline that characteristically kinks and terminates the first α-helix in Max network members. MondoA also has a shorter loop consisting of only 7 residues, while MondoB resembles ancestral Mondo loop sequence with 11 residues and a proline at L6. As seen with Mlx:MondoB interactions, variability in the loop sequence is likely to have a prominent role in determining dimer and higher order conformations. However, vertebrates may have slightly different conformations due to the acquired charge at Mlx sites Z2:Lys (K) and Z3:Glu (E) and polar residues H204:Thr (T) and

H208:Thr (T) for both MondoA and MondoB. Other changes in the Diptera lineage include Mlx Z15 that is not positively charged, Mlx Z24 that is aliphatic, and distinct aliphatic residues at Z25 and Z28 in Mondo.

Do Mlx interacting proteins have distinct bHLHZ attributes?

Dimerization experiments have not been performed in an organism from the core network and must be inferred from orthologous network interactions. Mnt:Max, Myc:Max, and Mondo:Mlx heterodimers have been verified in both vertebrates and *Drosophila*, implicating their interactions are ancestral. The Mxd:Max interaction is also assumed to be ancestral since all Mxd family proteins can heterodimerize with Max and MDL-1 can interact with both Max orthologs (Baudino and Cleveland 2001). In addition, dimerization properties restricting Mlx interactions is currently unknown. Notably, the interaction between Mnt and Mlx is unresolved due to conflicting evidence (Meroni, Reymond et al. 1997; Meroni, Cairo et al. 2000; Cairo, Merla et al. 2001; Billin and Ayer 2006). If Mnt does not interact with Mlx, the Max and Mlx networks are decoupled in both fly and nematode lineages and Mondo lacks a known repressor counterpart within the Mlx network. In vertebrates, Mxd1 and Mxd4 can heterodimerize with Mlx, while Mxd2 and Mxd3 cannot. MDL-1 cannot interact with Mlx in nematodes (Cairo, Merla et al. 2001; Billin and Ayer 2006), suggesting that Mxd can dimerize only with Max in the core network and the interaction between Mxd1 and Mxd4 with Mlx is derived. Since the Mxd bHLHZ domain has a strictly defined loop consisting of 9 residues, these binding restrictions are likely the result of specific residue changes within homologous sites.

To predict if Mxd can heterodimerize with Mlx in species belonging to the core network, we used the Mxd protein family members to identify sites that discriminate Mlx binding properties. In vertebrates, 25 of the 80 Mxd bHLHZ sites are invariable. Of the remaining variable sites, the size of Z15, quantified by the factor score transformation, explains 90% of variability between Mxd and Mlx binding groups. Factor scores quantifying secondary structure, codon composition, and charge of Z15 also contribute to Max and Mlx binding discrimination. Mxd1 and Mxd4 Z15 are invariant for Gln (Q), while Mxd2 Glu (E) and Mxd3 Arg (R) are charged. Site Z8 PAH also shows discriminatory power, although it is

not conserved in Mxd1, Mxd2, or Mxd3 and has overlapping properties. Site Z15 and Z8 are variable among invertebrates with no clear pattern of size, charge, or hydrophobicity conservation.

Canonical discriminant analysis (DA) weights sites to standardize variability within groups and maximize among group variation. The resulting linear discriminant function gives the greatest separation among *a priori* defined groups. Using vertebrate Mxd sequences grouped according to Mlx binding ability, we applied DA to estimate discriminant coefficients that maximally discriminate between binding groups. We then predicted the binding ability of unclassified Mxd sequences by their posterior probability of membership to a particular Mlx binding group. That is, we let the discriminant function classify unknown data. While the linear discriminant function completely and correctly classifies known binding partners, the binding of non-vertebrate Mxd members is indeterminate. PAH (47.83%), PSS (50%), MS (30.43%), CC (30.43%), and EC (39.13%) metrics predict less than half of Mxd sequences within the core network can dimerize with Mlx. This indicates Mxd in invertebrates is unlikely to dimerize with Mlx, although it cannot be firmly established.

Differences within Mxd and Mnt sequences prevent adequate prediction of Mlx binding. However, Mnt is largely conserved among all species sampled, which indicates Mnt:Mlx binding is consistent among all species. Sites L1, L3, and Z23 differentiate the Diptera lineage, although *D. melanogaster* shows additional variability with no distinct conservation among amino acid attributes. Mnt H204:Val (V) discriminates vertebrates from the otherwise conserved isoleucine in other species, while vertebrate Z21:Thr (T) is larger than residues in sequences from the Diptera and core networks.

Summary and Conclusions

Max and Mlx network members are found in the earliest known precursor organisms to animals and throughout the animal kingdom. Retention of these proteins over a billion years of evolution in such a diverse array of organisms suggests the Max and Mlx networks have vital roles in cell regulation and organismal development. The presence of Myc and

Max in choanoflagellate *Monosiga brevicollis* further verifies their evolution is both ancient and highly constrained. Extensive studies in model organisms such as *D. melanogaster*, *C. elegans*, and *M. musculus* confirm their intimate involvement in basic cell processes such as cell differentiation, proliferation, growth, metabolism, and apoptosis (Lüscher 2001).

Clear points of radiation and deletion shape the four major network configurations found in animals. Most animals exhibit the ancestral six-member core network consisting of Max, Mlx, Mondo, Myc, Mnt and Mxd. The emergence of c-, N-, L-Myc, MondoA, MondoB, and Mxd1-4 along with syntenic paralogous families demonstrates that these protein families radiated during vertebrate divergence, presumably due to two WGD events.

In comparison, flies lost Mxd and nematodes experienced a major network reconfiguration creating Mxl-1-3, MML-1 and MDL-1. That said, some species exhibit losses for certain members of the Max and Mlx networks. This can be attributed to 1) lineage specific duplication or deletion, 2) gene pseudogenization, 3) low coverage or unassembled genomes, and 4) unidentifiable orthology due to gene divergence. Although we are unable to identify the bHLHZ domain of some network members, it is still plausible that they exist, even in ancient lineages such as *Trichoplax* and *Monosiga*. Other cases, including chromosomal translocations surrounding N-Myc of *Monodelphis domestica* and absence of Mxd in ticks, imply a lineage specific gene loss may have occurred.

Although the ancestral divergence of trematodes is uncertain (Carranza, Baguñà et al. 1997), we provide evidence that nematodes and trematodes shared a common ancestor prior to arthropod divergence. The absence of an identifiable Myc or Mnt ortholog in *Schistosoma* and similar patterns in divergence for Max-, Mlx-, and Mondo-like sequences suggests nematodes and trematodes both experienced a major reconfiguration of the Max and Mlx networks. Moreover, the absence of a second Max ortholog in both *Schistosoma* and *Brugia malayi* and similarity between Mxl-3 and Max suggests that Mxl-1 originated from a duplication of Mxl-3 in *Caenorhabditis*. Likewise, nematode Mxl-2 and trematode Mlx divergence occurs at similar sites, yet both still demonstrate clear sequence orthology to Mlx. Specific to nematodes, we predict further divergence at packed sites H111:Lys (K) and H201:Asn (N) in Mlx-2 exhibit a reciprocal interaction for the otherwise conserved

H111:Gln (Q) and H201:Lys (K) sites observed in other Mlx sequences. In addition, inconsistent changes in hydrophobicity, accessibility, and size suggest the region around b11, H102, H105 and H109 in nematode MML-1 either lost or altered its involvement with an interacting partner.

Phylogenetic reconstructions further indicate Max and Mlx network members have distinguishable bHHZ sequences that are likely to confer distinct and specific DNA binding and dimerization properties. We predict the similarity between Mondo and Mlx bHHZ domains results from dimerization constraints and unique gene targets within the parallel network. Moreover, outgroups split Mondo and Mlx clades from Max, Mnt, Myc and Mxd, suggesting they share a divergent yet related common ancestor. Since Mnt and Mxd bHHZ domains do not interact, we anticipate their similarity relates to their role in gene repression. In contrast, the dissociation of Mondo and Myc proteins with transactivation activity denotes independent dimerization and DNA binding attributes. The lack of a distinct Myc clade further highlights its diversity and insinuates Myc orthologs have different propensities in dimerization and transcriptional regulation. Mga is a "wandering taxon" that is phylogenetically unstable and not consistently grouped with any outgroup sequences. Thus, we predict Mga rapidly diverged after duplication of a Max or Mlx network member and was subsequently conserved. Conversely, nematode sequences showed clear similarity to their respective orthologs, although they consistently acted as near outgroups. Furthermore, paralogs in vertebrate protein families formed separate clades and sequences generally bifurcated in order of species divergence, demonstrating strong selective forces are acting on these sequences.

Several sites exhibit common and unique characteristics of the bHLHZ domain that depict the divergence of Max and Mlx network members in animals. Sites b5, b9, b12, b13, H110, and H205 are largely invariant among network members due to site-specific restrictions in E-box DNA binding and dimerization stability. Likewise, sites b2, H103, H104, and H215 have low functional entropy values presumably due to their role in contacting the DNA phosphate backbone and involvement in protein conformation. While

the zipper is required for stable dimerization, the relatively low entropy of Z14 and Z21 suggest these leucine repeats are important contact points between monomers.

Using discriminant analysis, we statistically identified specific sites that distinctly classify proteins, network topologies, and potential dimerization patterns. Sites H107, H201, H202 completely discriminate Max and Mlx network proteins. While site H202 is not annotated, site H201 forms van der Waals contacts with H107 and anchors the second helix to DNA (Atchley and Zhao 2007). Such variability in important residues likely alters both DNA and protein binding abilities that determine gene target recognition and protein function.

Similarly, changes among orthologs may display evolutionary adaptations. Specifically, site b3 in Myc is unconserved in invertebrate sequences, which can affect DNA recognition and transformation capabilities. Interestingly, N-myc had overlapping similarities with c-Myc and L-Myc discriminating sites with no distinct sites of its own, suggesting these changes have cumulative or compensatory effects among Myc family members. Protein dimerization may also differ among species due to variability between Max and Mlx network members at sites Z17 and Z18, which were found to attribute Max dimerization preferences. Similarity in loop length and conservation between MondoB and invertebrate Mondo sequences, suggests they have corresponding dimerization and DNA binding restrictions. However, heterotetramer conformation may differ in invertebrates due to the lack of L8:Phe (F). Instead this higher order structure may rely on the negatively charged L7:Asp (D) and hydrophobic L11:Gly (G) residues, which are highly conserved among animal species.

Dimerization properties among Max and Mlx network members have been investigated *in vivo* for *C. elegans*, *D. melanogaster*, and *M. musculus* (Blackwood, Lüscher et al. 1992; Amati and Land 1994; Arsura, Deshpande et al. 1995; Yuan, Tirabassi et al. 1998; Hurlin, Steingrìmsson et al. 1999; Billin, Eilers et al. 2000; Meroni, Cairo et al. 2000; Cairo, Merla et al. 2001; Orian, van Steensel et al. 2003). However, interactions between members in the core network are unknown. While our predictions for invertebrate Mxd binding using DA were indeterminant, we anticipate Mxd1 and Mxd4 binding with Mlx is

derived and results from independent changes within the bHLHZ domain. Furthermore, conflicting reports on Mnt and Mlx heterodimerization raise several questions concerning the extent of Mnt repression and Mondo regulation.

For example, do Mad or Mnt competitively dimerize with Mlx to regulate Mondo? How does the loss of Mxd2 and Mxd3 or gain of Mxd1 and Mxd4 binding with Mlx affect Mondo regulation in vertebrates? Does the loss of Mad in flies change dMnt function? Although Mxd is dispensable in flies and individual knockouts in mice have minor changes in phenotype, the persistence of Mxd in most other species including nematodes indicates it has a basic and important role in cell maintenance.

These evolutionary analyses provide a basis for understanding important aspects of Max and Mlx network interactions and function in animals. Although no direct ortholog of Myc or Max has been found in yeast (Brown, Cole et al. 2008), yeast contains interacting homologs Sin3 and GCN5 as well as E-boxes and may still be harboring unidentified Max and Mlx network orthologs. Using the protein distinctions we have described, it is now possible to distinguish Max and Mlx network member bHLHZ domains, search for unannotated sequences in highly divergent species, and attribute structural and functional differences among these proteins. Hence, these predictions will enable the refinement of protein annotation within an evolutionary context of network interactions and facilitate the functional analysis of important proteins such the Myc proto-oncogene.

Figures

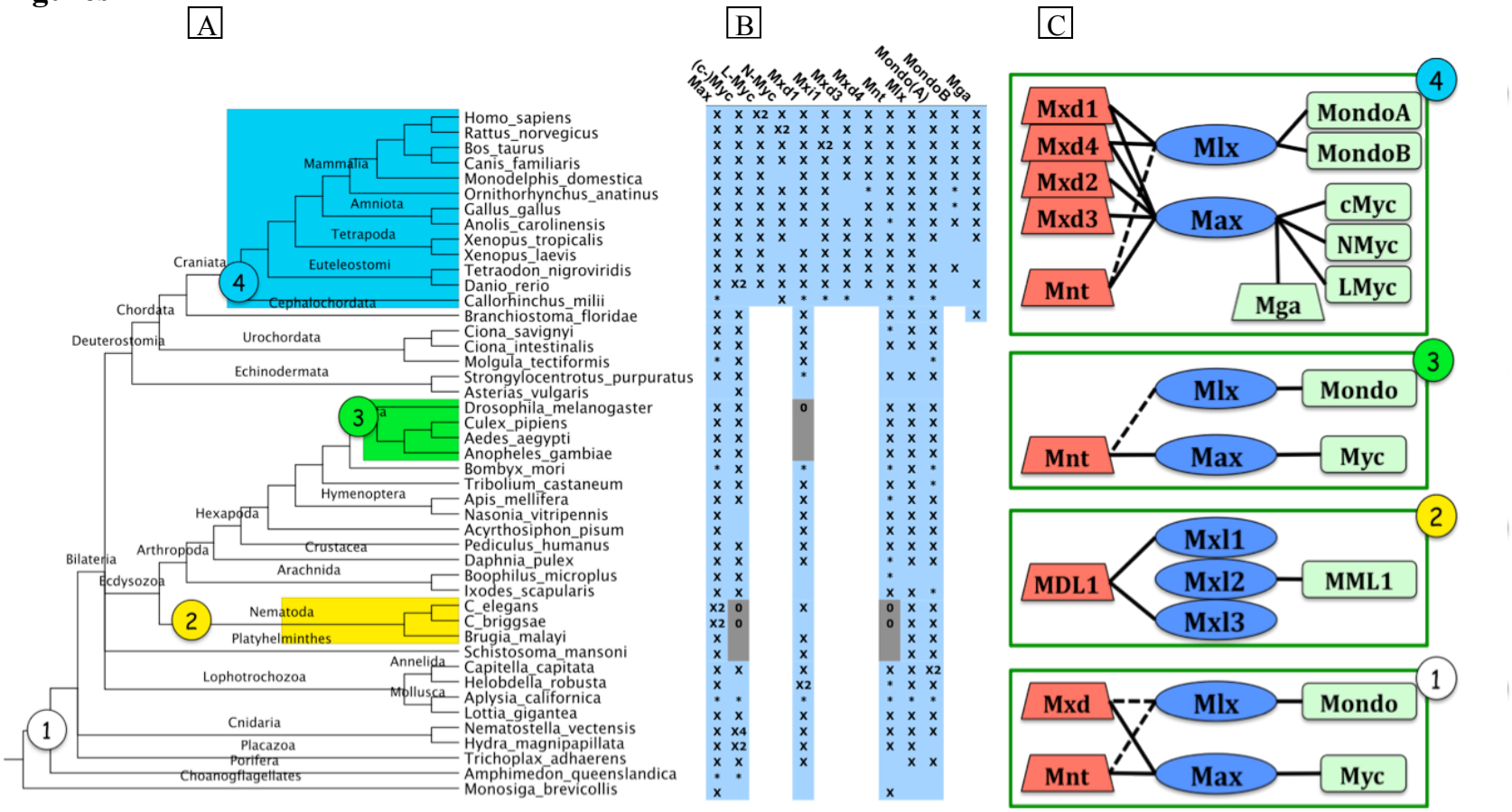


Figure 1:Max and Mlx Network Protein Distribution

A)Species tree determined by Flybase, Ensembl and Tree of Life resources. **B)** Light blue and Grey cells indicate the protein is expected to be present or absent, respectively, within the organism. 'X' means the bHLHZ was found within a protein or EST, '*' means part of the sequence was found or all was found within a genetic region, 0 means the protein is known to be absent. **C)** Circled numbers correspond to the emergence of the labeled network. Green squares activate and red trapezoids repress transcription; blue ovals are obligate dimers and lack an active domain. Solid lines indicate known dimerizations, while debated or unknown interactions are shown by a dotted line. Rodents also contain S-Myc and humans have L-Myc2, which interact with Max.

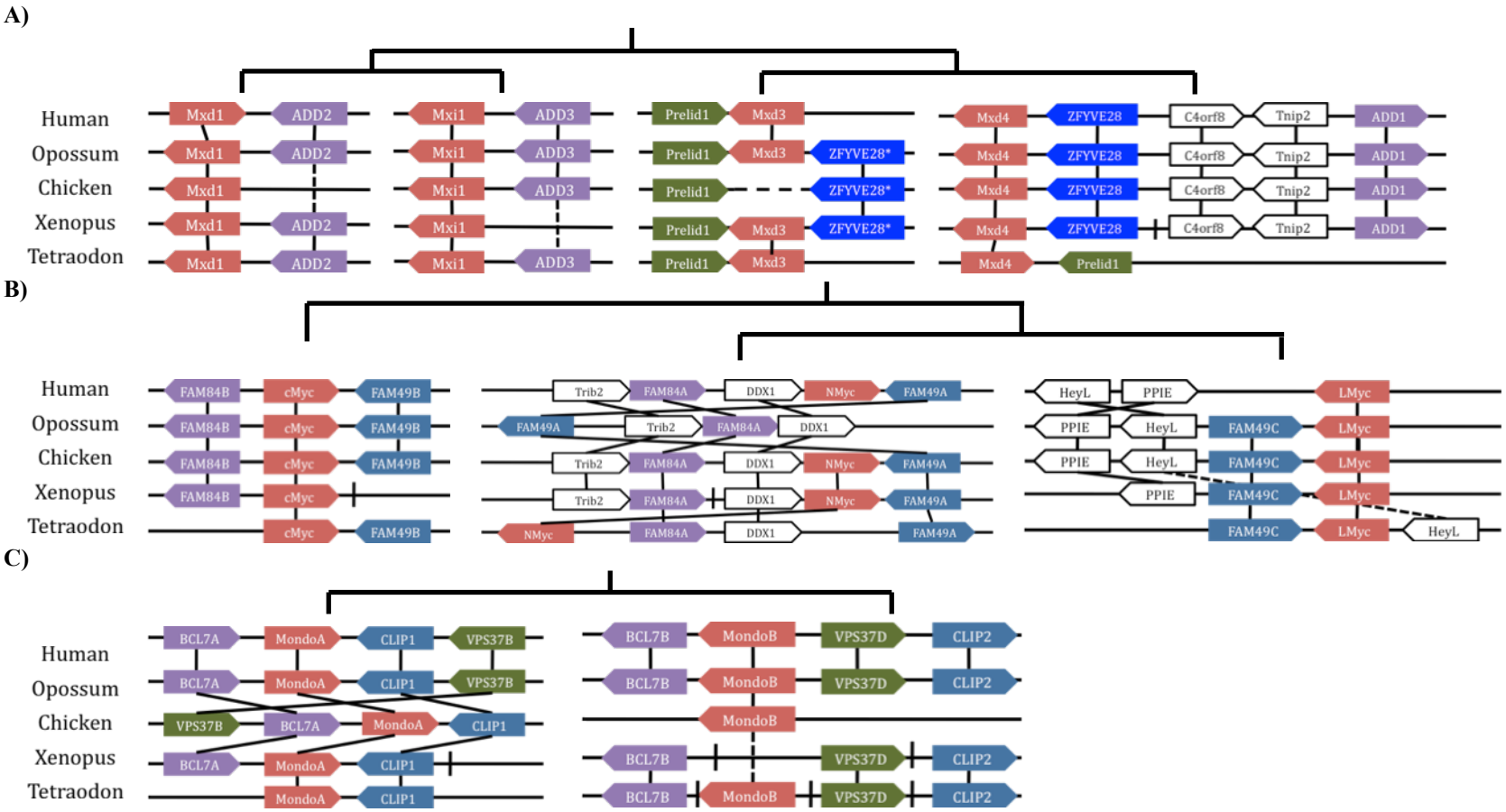


Figure 2: Mxd, Myc and Mondo Synteny

Cartoon depiction of genetically linked homologs. Synteny among paralogous gene families (colored boxes) suggest a common origin, while orthologs (white boxes) confirm orientation and structure. Tree structure displays proposed order of duplication prior to divergence. Solid lines between species indicate conserved orthology, dashed lines indicate intermediate species have a missing or unlinked ortholog, and hashes between genes show breaks in contig sequences. Gene sizes and distances are not to scale. **A**) The Mxd family is linked with ADD paralogs. Tetraodon carries two copies of Prelid1 and ZFYVE28* is an unnamed duplicate of ZFYVE28. A gap in the chicken genome coverage suggests Mxd3 is conserved yet unavailable. **B**) The Myc family is linked to Fam84 and Fam49 paralogs. Translocations surrounding NMyc in oppossum potentially resulted in its loss. **C**) The Mondo family is flanked by BCL7, Clip, and VPS37 paralogs. MondoB was unidentifiable in Xenopus. BCL7B, VPS37D, and Clip2 were all found on different contigs for Xenopus and Tetraodon.

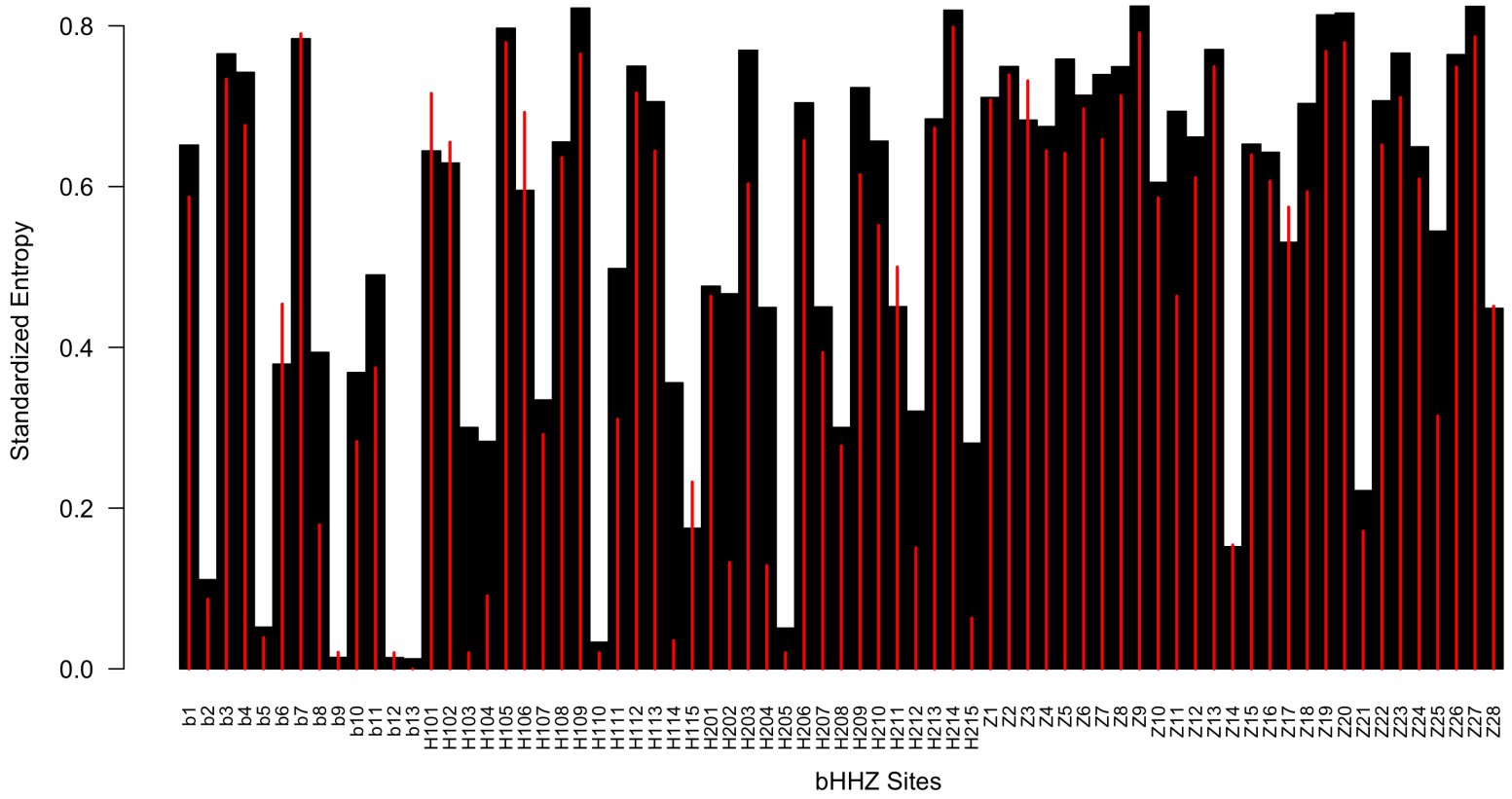


Figure 3: bHHZ Entropy for Max and Mlx Network Members

Black columns represent standardized AA Entropy. Red bars represent standardized Functional Entropy (Atchley et al 1999). All network proteins were included in calculation.

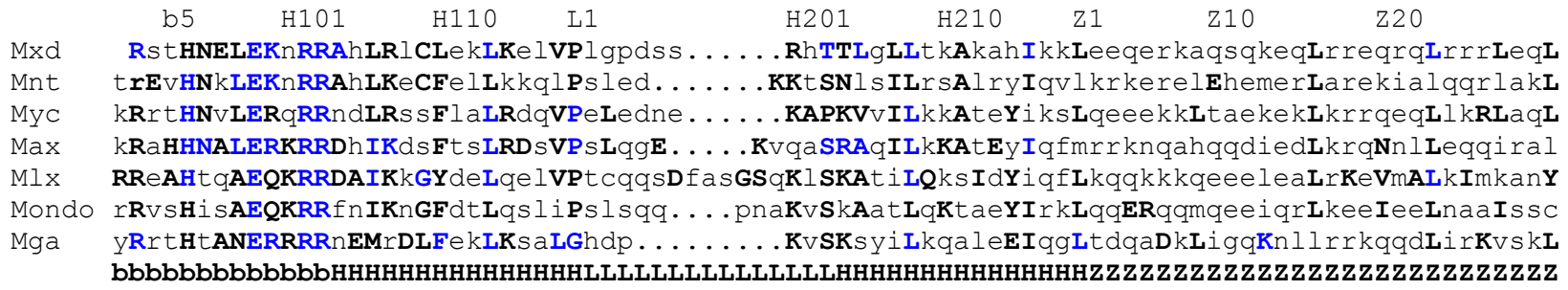


Figure 4: HMMER Sequence of bHLHZ Domain

Highly conserved residues that occur with over 90% probability are uppercase, bold letters, while other lower case letters show the most explanatory residue for that site. Invariant sites are colored blue to emphasize their importance. Dots are simply placeholders for the loop alignment and are not included in loop numbering for individual proteins.

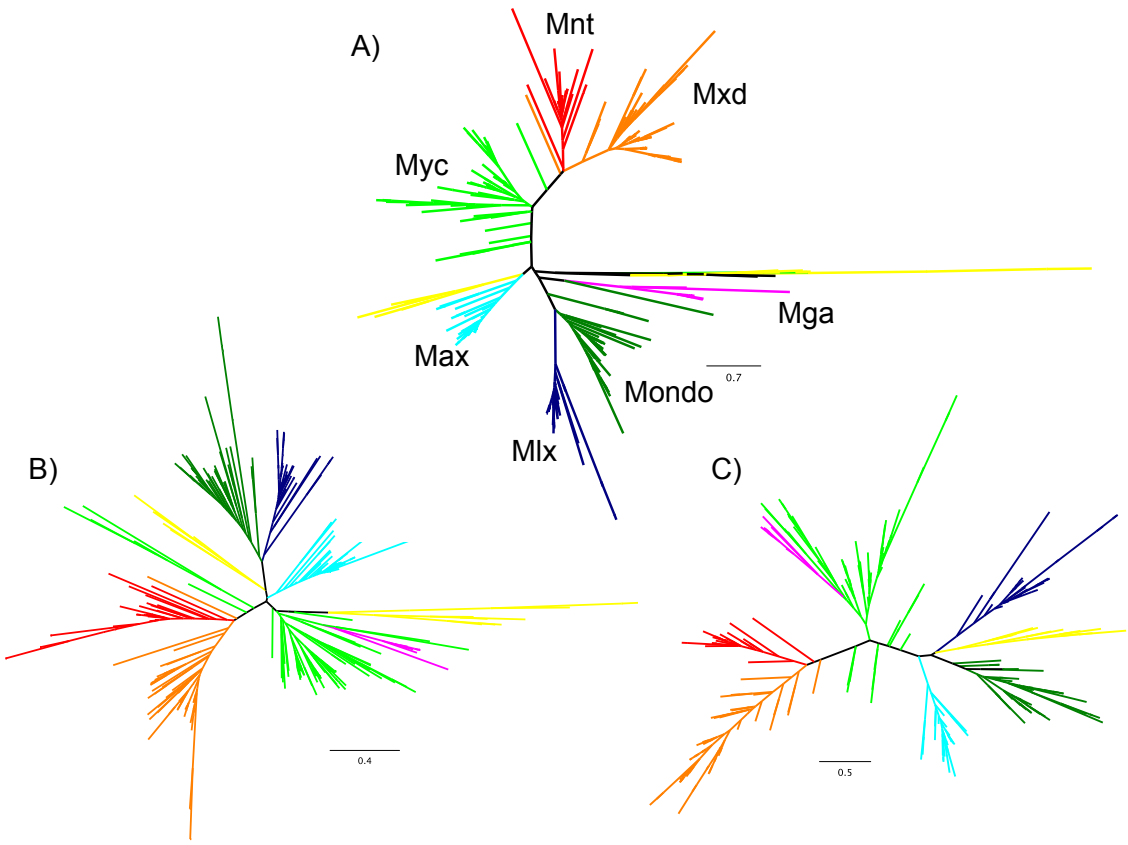


Figure 5: Phylogeny of bHHZ domain

Phylogenetic reconstruction of bHHZ domain for all Max and Mlx network members. **A)** PhyML algorithm using JTT rate matrix with 4 site rate categories estimated from a discretized Gamma distribution. **B)** BioNJ algorithm using PMB rate matrix and a single site rate. **C)** MrBayes algorithm using Gamma distribution of rate categories over 2 million generations. Mxd (orange), Mnt (red), Myc (light green), Max (light blue), Mlx (dark blue), Mondo (dark green), and Mga (magenta). Human, Drosophila, and C. elegans orthologs of SREBF, USF, TCF3, MYOD, and hes1 were used as outgroups (yellow).

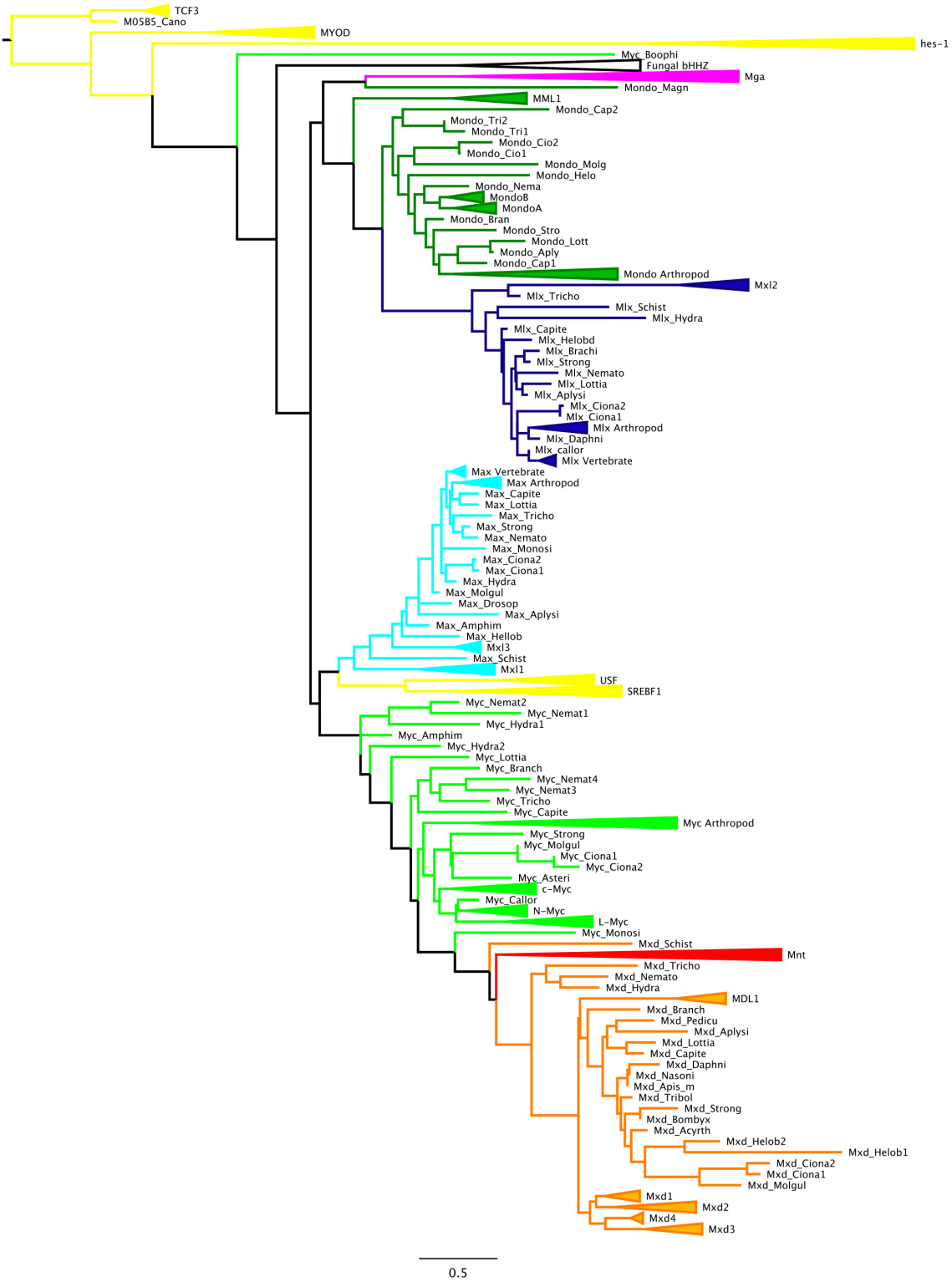


Figure 6: bHHZ PhyML Rooted Tree

PhyML tree using Gamma distribution for rates with 4 classes and rooted by the TCF3 outgroup.

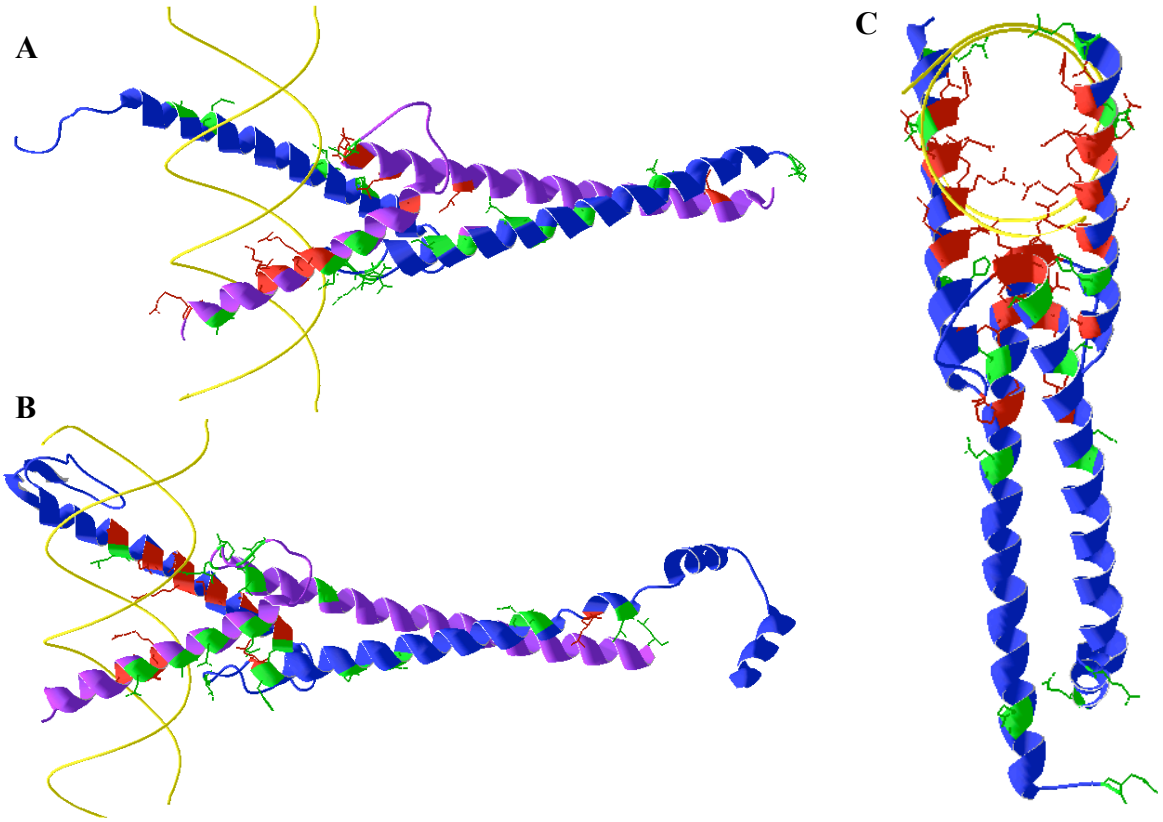


Figure 7: Nematode bHLHZ Structure

C. elegans dimers recognizing DNA (yellow). Sites distinguishing nematode orthologs are colored green, while identical sites for all orthologs are red. Backbone and side chain atoms for these sites are displayed. **A)** Mxl-1 (blue) and MDL-1 (purple) heterodimer. Identical Max sites are not shown for this structure. **B)** Full Mxl-2 sequence (blue) and MML-1 bHLHZ (purple) heterodimer **C)** Mxl-3 (blue) homodimer

Tables

Table 1: Max and Mlx Network Members

Max Network	*Overlap?*	*Mlx Network*
Core		
Myc	**Max** **Mxd** **Mnt**	**Mlx** **Mondo**
Diptera		
Myc (dMyc, dm)	**Max** (dMax) **Mnt** (dMnt)	**Mlx** (dMlx) **Mondo** (dMondo, Mio)
Nematode		
	Mxl-1 **MDL-1** **Mxl-2** **MML-1** (T20B12.6) **Mxl-3**	
Vertebrate*		
c-Myc (Myc2, Niard, Nird)	**Max** (Myn) **Mxd1** (Mad1) **Mnt** (Rox, Mad6, Mxd6) **Mlx** (BigMax) **MondoA** (bHLHe36, KIAA0867, MIR, MLXIP)	
N-Myc (N-Myc1,N-Myc2, MycN)	**Mxd2** (Mad2, Mxi1, Mxi) **MondoB** (ChREBP, WBSCR14, MLXIPL)	
L-Myc (MycL1, LMyc1)	**Mxd3** (Mad3, Myx)	
Mga (KIAA0518, Mad5, Mxd5)	**Mxd4** (Mad4, MSTP149, MST149)	

Network components are listed according to their presence in the four main animal networks. Columns represent orthologous proteins between networks and paralogous proteins within. Known aliases for each protein are provided in parentheses. *Rodents have an additional N-Myc duplicate termed S-Myc, while Primates have an L-Myc duplicate named L-Myc2. Mga has unknown origin within the vertebrate network.

Table 2: Sampled Genomes

	Genus species	Common Name	Source	Status	Published Genome
Vertebrate	*Homo sapiens*	Human		Complete	(Venter, Adams et al. 2001)
	Rattus norvegicus	Rat		Assembly	(Gibbs, Weinstock et al. 2004)
	Bos taurus	Cow		Assembly	(Consortium, Elsik et al. 2009)
	Canis familiaris	Dog	Broad	Assembly	(Lindblad-Toh, Wade et al. 2005)
	Monodelphis domestica	Opossum		Assembly	(Mikkelsen, Wakefield et al. 2007)
	Ornithorhynchus anatinus	Duckbill platypus	WashU	Assembly	(Warren, Hillier et al. 2008)
	Gallus gallus	Chicken		Assembly	(Consortium 2004)
	Anolis carolinensis	Green Anole Lizard		Assembly	
	Xenopus tropicalis	Western Clawed Frog	JGI	Assembly	
	Xenopus laevis	African Clawed Frog			
	Tetraodon nigroviridis	Green pufferfish	Broad	Assembly	
	Danio rerio	Zebrafish	Sanger	Assembly	
	Callorhinchus milii	Elephantfish		Assembly	(Venkatesh, Kirkness et al. 2007)
Core	*Branchiostoma floridae*	Florida lancet	JGI	Assembly	(Putnam, Butts et al. 2008)
	Ciona savignyi	Sea squirt	Broad	Assembly	
	Ciona intestinalis	Sea squirt	JGI	Assembly	(Dehal, Satou et al. 2002)
	Molgula tectiformis	Sea grapes			
	Strongylocentrotus purpuratus	Purple sea urchin	Baylor	Assembly	(Consortium, Sodergren et al. 2006)
	Asterias vulgaris	Sea star			
Diptera	*Drosophila melanogaster*	Fruitfly			(Adams, Celniker et al. 2000)
	Culex pipiens	Southern house Mosquito	Broad	Assembly	
	Aedes aegypti	Yellow fever mosquito	TIGR	Assembly	(Nene, Wortman et al. 2007)
	Anopheles gambiae	Malaria mosquito		Complete	(Sharakhova, Hammond et al. 2007)
Core	*Bombyx mori*	Silkworm moth		Assembly	(Consortium 2008)
	Tribolium castaneum	Red flour beetle	Baylor	Assembly	(Consortium, Richards et al. 2008)
	Apis mellifera	Honeybee	HGSC	Assembly	(Consortium 2006)
	Nasonia vitripennis	Jewel wasp	Baylor	Assembly	(Werren, Richards et al. 2010)
	Acyrthosiphon pisum	Pea aphid	Baylor	Assembly	(Consortium 2010)
	Pediculus humanus	Human louse			
	Daphnia pulex	Waterflea	JGI	Progress	
	Boophilus microplus	Southern cattle tick			
	Ixodes scapularis	Deer tick		Assembly	(Hill and Wikel 2005)
Nematode	*Caenorhabditis elegans*	Roundworm		Complete	(Hillier, Marth et al. 2008)
	Caenorhabditis briggsae	Roundworm	Sanger	Assembly	(Gupta and Sternberg 2003; Stein, Bao et al. 2003)
	Brugia malayi	Filariod worm	Sanger	Assembly	(Scott and Ghedin 2009)
Core	*Schistosoma mansoni*	Trematode		Assembly	
	Capitella capitata	Polycheate worm (Annelida)	JGI	Complete	
	Helobdella robusta	Leech (Annelida)			
	Aplysia californica	California sea hare	Broad	Assembly	
	Lottia gigantea	Owl limpet (sea snail)		Complete	
	Nematostella vectensis	Starlet sea anemone	JGI	Assembly	(Putnam, Srivastava et al. 2007)
	Hydra magnipapillata	Hydra	Venter	Assembly	(Chapman, Kirkness et al. 2010)
	Trichoplax adhaerens	Placazoa	JGI	Assembly	(Srivastava, Begovic et al. 2008)
	Amphimedon queenslandica	Sponge	JGI	Progress	
	*Monosiga brevicollis**	choanoflagellate	JGI	Complete	(King, Westbrook et al. 2008)

Table 3: Phylogenetic Reconstructions

Type	Method	Q	Site Rate	Log Lk	Tree
Bayesian	MrBayes	Mixed	fixed	-23893.743	**C***
Bayesian	MrBayes	Mixed	gamma, estimate pinvar	-23834.159	B
ML	PAML	JTT	pinvar	-21602.4493	B
ML	ProML	JTT	fixed	-21777.7348	B
ML	ProML	JTT	Γ:α=1.3, C=4	-20788.22426	B
ML	ProML	JTT	Γ:α=1.3, C=4, pairwise correlation	-20696.53095	B
ML	PhyML	JTT	fixed (pinvar=0)	-21614.3738	A
ML	PhyML	WAG	fixed (pinvar=0)	-21548.877	B
ML	PhyML	WAG	estimate pinvar	-21548.88396	B
ML	PhyML	JTT	estimate pinvar	-21612.50602	A
ML	PhyML	JTT	Γ:C=4, α, pinvar=0	-20550.54622	**A***
ML	PhyML	WAG	Γ:C=4, α, pinvar=0	-20675.15114	A
ML	PhyML	JTT	Γ:C=4, α=2, pinvar=0	-20582.61548	A
Distance	NJ (HyPhy)	PC	fixed		A
Distance	NJ (HyPhy)	PC_RV	fixed		A
Distance	NJ (HyPhy)	JTT	fixed		A
Distance	NJ (HyPhy)	JTT	Γ:C=4		B
Distance	NJ (HyPhy)	JTT+F	fixed		B
Distance	BioNJ	JTT	fixed		A
Distance	BioNJ	JTT	Γ:α=1		A
Distance	BioNJ	PMB	fixed		**B***
Distance	ProtPars		ordinary parsimony		B

Bayesian, Maximum Likelihood, and Distance methods for reconstructing the bHHZ tree. Trees fall under three main topologies (A, B, C) shown in Figure 5, where an asterisk indicates the tree shown. PC: Poisson correction, PC_RV: Poisson corrected with rate variation. WAG: Whelan Goldman model. JTT: Jones Taylor Thornton Model. PMB: Probability Matrix from Blocks. +F: with empirical character frequencies. Γ: Gamma rate distribution. C: number of rate categories. Pinvar: proportion of invariant sites, Mixed: Mixed Fixed Rate model explores rate matrices such as JTT and WAG, where each contributes to the rate in proportion to its posterior distribution of the converged model.

Table 4: Discriminant Analysis of Max and Mlx Network Proteins

A) Max

Identical	Step	PAH	(ASCC)	PSS		MS		CC		EC		Discriminate
b5	1	H203	(0.2757)	H106	(0.2489)	H206	(0.3012)	b1	(0.2726)	H206	(0.3089)	b1[1ab]
b6	2	b2	(0.3290)	H101	(0.3110)	H203	(0.5707)	b4	(0.3333)	L5	(0.3725)	b4[2a]
b8	3	H113	(0.3494)	b7	(0.3274)	H101	(0.5865)	H108	(0.5658)	H106	(0.6052)	b7[2b]
b9	4	L6	(0.3542)	b1	(0.3405)	L8	(0.5948)	Z19	(0.6206)	b7	(0.6358)	H102[2a]
b10	5	H210	(0.4665)	Z19	(0.5206)	H111	(0.6246)	Z5	(0.6358)	H102	(0.6589)	H106[2ab]
b12	6	H211	(0.5101)	Z22	(0.6409)	Z4	(0.6544)	Z16	(0.6646)	L6	(0.6654)	H108[4]
b13	7	b11	(0.5146)	b11	(0.6679)	Z25	(0.6793)	Z2	(0.6877)	Z16	(0.6945)	L6[2a]
H103	8	H209	(0.6085)	H209	(0.6810)	L13	(0.6840)	Z14	(0.6928)	L8	(0.7012)	H203[3]
H104	9	Z20	(0.6751)	H210	(0.7269)	H208	(0.7221)	Z7	(0.7016)	H209	(0.7025)	H206[4]
H110	10	Z22	(0.7069)	b3	(0.7500)	Z1	(0.7392)	H208	(0.7417)	Z27	(0.7903)	H209[2ab]
H115	11	L9	(0.7121)	H109	(0.7586)	Z27	(0.8069)	H213	(0.7518)	Z22	(0.7908)	H211[2a]
L14	12	Z8	(0.7293)	Z12	(0.7716)	Z3	(0.8249)	H211	(0.7624)	L10	(0.8230)	Z1[1]
H201	13	L3	(0.7467)	Z10	(0.7735)	L5	(0.8281)	Z8	(0.7828)	H211	(0.8499)	Z3[2a]
H202	14	Z11	(0.7537)	Z13	(0.7820)	H209	(0.8286)	L3	(0.7913)	L1	(0.8502)	Z5[1]
H205	15	L7	(0.7642)	H206	(0.8025)	b2	(0.8303)	L7	(0.8209)	Z18	(0.8578)	Z7[2a]
H212	16	Z1	(0.7853)	Z6	(0.8394)	b3	(0.8496)	Z17	(0.8256)	Z15	(0.8819)	Z15[2a]
	17	Z26	(0.7896)	L7	(0.8709)	Z17	(0.8648)	H210	(0.8778)	Z9	(0.8912)	Z19[2b4]
	18	Z18	(0.7934)	L3	(0.9002)	L11	(0.8716)	L10	(0.8908)	b2	(0.9217)	Z27[2ab]
	19	H213	(0.8160)			Z7	(0.8793)	Z15	(0.8922)			
	20	Z13	(0.8216)			H204	(0.8872)	L5	(0.9146)			
	21	Z23	(0.8480)			H106	(0.8909)					
	22	H109	(0.9061)			Z18	(0.8976)					
	23					L10	(0.8981)					
	24					L9	(0.8987)					
	25					Z15	(0.9297)					

B) Mlx

Identical	Step	PAH	(ASCC)	PSS		MS		CC		EC		Discriminate
b5	1	L8	(0.2968)	H206	(0.3186)	H206	(0.2861)	Z24	(0.2659)	H206	(0.3297)	b6[2]
b9	2	H206	(0.5835)	H201	(0.3333)	H201	(0.3333)	L16	(0.4298)	H202	(0.3333)	H111[2]
b12	3	H201	(0.6280)	Z3	(0.5391)	Z3	(0.5771)	H202	(0.5636)	Z2	(0.5868)	L8[2]
b13	4	b6	(0.6657)	H213	(0.6979)	Z15	(0.8027)	Z3	(0.7415)	Z15	(0.7880)	L16[2]
H103	5	Z24	(0.9016)	H209	(0.7591)	Z5	(0.8489)	H214	(0.7912)	Z4	(0.8431)	H201[2]
H106	6			H111	(0.8017)	Z4	(0.8745)	Z16	(0.8082)	H102	(0.8786)	H206[2]
H110	7			Z24	(0.8906)	Z16	(0.8930)	H111	(0.8885)	b6	(0.9062)	H213[3]
H205	8			Z16	(0.9650)	Z24	(0.9696)	H107	(0.8991)			H214[2]
Z21	9							Z14	(0.9259)			Z2[4]
												Z3[24]
												Z15[23]
												Z16[2]
												Z24[23]

C) Myc

Identical	Step	PAH	(ASCC)	PSS		MS		CC		EC		Discriminate
b5	1	H103	(0.1594)	H102	(0.1277)	b6	(0.1599)	b6	(0.1606)	b6	(0.1568)	b3[1]
b9	2	H206	(0.2740)	H206	(0.2419)	H102	(0.2812)	H108	(0.2771)	b10	(0.1752)	b4[4ae]
b12	3	H114	(0.2857)	H109	(0.3319)	b10	(0.2929)	L6	(0.3664)	H106	(0.3017)	b6[4c]
b13	4	b3	(0.4072)	H103	(0.4367)	H106	(0.3916)	H111	(0.4781)	H103	(0.3796)	b7[4cd]
H110	5	H107	(0.4315)	b10	(0.4716)	L7	(0.4955)	H103	(0.5584)	H107	(0.4078)	H102[34a]
H115	6	H102	(0.5527)	H107	(0.5034)	H103	(0.5716)	b10	(0.5953)	L7	(0.5122)	H103[4ae]
H205	7	Z22	(0.6511)	b6	(0.5372)	H107	(0.6123)	H107	(0.6111)	H104	(0.6101)	H108[4e]
	8	Z24	(0.7224)	L7	(0.6224)	b11	(0.6565)	H104	(0.6785)	b11	(0.6494)	H111[3]
	9	b11	(0.7552)	b4	(0.6518)	b8	(0.6823)	H206	(0.7173)	H202	(0.6779)	L7[4]
	10	H104	(0.7760)	H111	(0.7026)	H104	(0.7139)	L7	(0.7504)	H112	(0.7093)	H203[3]
	11	b7	(0.7943)	H203	(0.7291)	H111	(0.7831)	b1	(0.7830)	Z1	(0.7240)	H206[4c]
	12	L6	(0.8118)	L5	(0.7470)	Z1	(0.8047)	b8	(0.7999)	H111	(0.7825)	Z22[4a]
	13	Z5	(0.8166)	Z8	(0.7583)	H202	(0.8264)	H102	(0.8198)	b3	(0.8048)	Z24[3]
	14	H201	(0.8350)	H108	(0.7776)	H201	(0.8368)	H112	(0.8417)	H207	(0.8286)	
	15	L8	(0.8483)	Z9	(0.7883)	H108	(0.8563)	b11	(0.8528)	L5	(0.8406)	
	16	Z1	(0.8630)	L8	(0.8042)	H112	(0.8684)	H202	(0.8638)	Z13	(0.8653)	
	17	b4	(0.8696)	Z10	(0.8223)	b7	(0.8741)	H106	(0.8783)	H102	(0.8857)	
	18	H214	(0.8834)	H113	(0.8362)	Z15	(0.8877)	H203	(0.8910)	Z7	(0.8925)	
	19	Z26	(0.8881)	H114	(0.8395)	L5	(0.8974)	b7	(0.8972)	Z22	(0.9088)	
	20	H109	(0.8913)	H112	(0.8468)	Z16	(0.9076)	H213	(0.9049)			
	21	H202	(0.9003)	b2	(0.8552)							
	22			Z20	(0.8625)							
	23			Z19	(0.8656)							
	24			H207	(0.8748)							
	25			Z22	(0.8952)							
	26			Z14	(0.8995)							
	27			b11	(0.9033)							

D) Mondo

Identical	Step	PAH	(ASCC)	PSS		MS		CC		EC		Discriminate
b9	1	b11	(0.2418)	b11	(0.2401)	Z25	(0.1989)	Z25	(0.2499)	H201	(0.1851)	b11[2]
b12	2	H103	(0.2761)	H204	(0.4423)	Z14	(0.3041)	Z14	(0.2980)	Z25	(0.3608)	H102[2]
b13	3	H213	(0.4601)	L6	(0.6095)	L7	(0.4525)	H102	(0.4769)	Z14	(0.4945)	H105[2]
	4	H201	(0.5732)	b10	(0.6176)	H109	(0.5837)	H211	(0.4839)	L7	(0.6248)	H109[3]
	5	b2	(0.6395)	H202	(0.6252)	H113	(0.6241)	L2	(0.5915)	H113	(0.6724)	L5[2]
	6	Z6	(0.7290)	Z28	(0.7380)	H105	(0.7043)	L4	(0.6606)	H204	(0.7681)	L6[2]
	7	L8	(0.7614)	L7	(0.7899)	Z4	(0.7141)	Z15	(0.7483)	L4	(0.7820)	L8[4b]
	8	H209	(0.8022)	H205	(0.8084)	Z1	(0.7884)	H112	(0.7771)	L11	(0.8048)	L11[2]
	9	H208	(0.8208)	Z15	(0.8578)	Z27	(0.8486)	Z11	(0.8171)	H207	(0.8356)	H201[2]
	10	H203	(0.8583)	Z17	(0.8814)	L5	(0.8833)	b1	(0.8428)	L5	(0.8597)	H204[24]
	11	Z11	(0.8720)	L8	(0.9137)	Z22	(0.9030)	Z21	(0.8825)	L6	(0.8875)	H208[4]
	12	Z17	(0.9053)					b7	(0.8926)	H112	(0.9020)	H211[2]
	13							b5	(0.9097)	H201	(0.8922)	Z6[4a]
	14									Z22	(0.9098)	Z25[3]
												Z28[23]

E) Mnt

Identical	Step	PAH	(ASCC)	PSS		MS		CC		EC		Discriminate
b5	1	H115	(0.3237)	H113	(0.3861)	H204	(0.4244)	H204	(0.4889)	Z23	(0.3944)	L1[1]
b8	2	H212	(0.4596)	L3	(0.7095)	H201	(0.5000)	H208	(0.5000)	H204	(0.7556)	L3[1]
b9	3	Z27	(0.6955)	H212	(0.8186)	Z21	(0.7443)	H115	(0.8241)	H201	(0.8729)	H204[4]
b10	4	H112	(0.8102)	H206	(0.8483)	H212	(0.8572)	H212	(0.9674)	Z2	(0.9539)	Z21[34]
b12	5	L7	(0.8675)	H213	(0.8717)	Z4	(0.9056)					Z23[3]
b13	6	L10	(0.8923)	H204	(0.9430)							Z27[4]
	7	L1	(0.9112)									

F) Mxd

Identical	Step	PAH	(ASCC)	PSS		MS		CC		EC		Discriminate
b2	1	H102	(0.1885)	H106	(0.2000)	H106	(0.1874)	H106	(0.1874)	H106	(0.1874)	b4[24c]
b9*	2	H113	(0.3611)	H105	(0.3434)	b8	(0.3525)	b8	(0.3524)	b8	(0.3594)	b8[4a]
b10	3	Z11	(0.5064)	Z7	(0.4748)	H211	(0.4864)	Z11	(0.4812)	H102	(0.4932)	H102[24cd]
b12	4	H115	(0.5839)	L8	(0.5859)	Z16	(0.6016)	H211	(0.6034)	H114	(0.6135)	H105[24c]
b13	5	Z8	(0.6665)	H115	(0.6850)	H114	(0.6654)	H201	(0.6704)	Z16	(0.6911)	H106[4d]
H101*	6	H106	(0.6742)	Z23	(0.7355)	H102	(0.7570)	H214	(0.7120)	H115	(0.7354)	H113[4c]
H110*	7	Z22	(0.7462)	b11	(0.7724)	H115	(0.8081)	H113	(0.7443)	Z11	(0.7758)	H114[4b]
L9*	8	b4	(0.7721)	b8	(0.7949)	Z11	(0.8239)	Z7	(0.7634)	Z1	(0.8028)	L8[2]
H202*	9	L4	(0.7980)	Z5	(0.8043)	Z18	(0.8374)	H109	(0.7802)	H214	(0.8290)	H206[4b]
H205*	10	H210	(0.8167)	b5	(0.8316)	b11	(0.8642)	Z3	(0.7998)	H108	(0.8498)	H211[4]
H212	11	H201	(0.8284)	H103	(0.8463)	H203	(0.8729)	H208	(0.8287)	H209	(0.8553)	H214[4]
Z21	12	H213	(0.8458)	b3	(0.8646)	L2	(0.9011)	L4	(0.8470)	Z14	(0.8633)	Z7[4b]
	13	Z4	(0.8544)	L3	(0.8778)			H115	(0.8610)	H203	(0.8720)	Z8[4d]
	14	L2	(0.8594)	Z27	(0.8839)			Z9	(0.8711)	Z19	(0.8779)	Z11[4b]
	15	H206	(0.8896)	Z24	(0.8887)			L1	(0.9002)	b6	(0.8825)	Z13[4c]
	16	Z13	(0.8918)	Z19	(0.8928)					L7	(0.8854)	Z18[4d]
	17	H114	(0.8952)	H203	(0.9064)					H105	(0.9007)	
	18	H209	(0.9075)									

Stepwise discriminant analysis classifying each protein by network according to its bHLHZ sites. Conserved synapomorphies are highlighted and listed under the 'Discriminate' column where discriminating networks are designated by superscripts 1. Core 2. Nematode 3. Diptera 4. Vertebrate. Paralogs are treated as individual categories. **A)** Max.(1,2a: Mxl-1, 2b: Mxl-3,3,4) **B)** Mlx (1,2,3,4) **C)** Myc (1,3,4a: c-, 4b: N-, 4c: L-, 4d: S-, 4e: L-Myc2) **D)** Mondo (1,2,3,4a:MondoA, 4b: MondoB) **E)** Mnt (1,3,4) **F)** Mxd (1,2,4a:Mxd1, 4b:Mxi1, 4c:Mxd3, 4d:Mxd4). *The two Mxd2 sequences in Bos taurus contain multiple substitutions and were not considered for identifying identical or synapomorphic sites.

References

"The Myc Target Gene Database." from myccancergene.org.

(2010). "These data were produced by The Genome Center at Washington University School of Medicine in St. Louis and were obtained from http://genome.wustl.edu/tools/blast."

Adams, M. D., S. E. Celniker, et al. (2000). "The genome sequence of Drosophila melanogaster." Science **287**(5461): 2185-95.

Alberghina, L., T. Höfer, et al. (2009). "Molecular networks and system-level properties." Journal of Biotechnology: 10.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." Journal of Molecular Biology **215**(3): 403-10.

Anong, W., T. Franco, et al. (2009). "Adducin forms a bridge between the erythrocyte membrane and its cytoskeleton and regulates membrane cohesion." Blood **114**(9): 1904-1912.

Arnason, U., A. Gullberg, et al. (1998). "Molecular timing of primate divergences as estimated by two nonprimate calibration points." J Mol Evol **47**(6): 718-27.

Arsura, M., A. Deshpande, et al. (1995). "Variant Max protein, derived by alternative splicing, associates with c-Myc in vivo and inhibits transactivation." Molecular and Cellular Biology **15**(12): 6702-9.

Atchley, W. R. and A. D. Fernandes (2005). "Sequence signatures and the probabilistic identification of proteins in the Myc-Max-Mad network." Proc Natl Acad Sci USA **102**(18): 6401-6.

Atchley, W. R. and W. M. Fitch (1995). "Myc and Max: molecular evolution of a family of proto-oncogene products and their dimerization partner." Proc Natl Acad Sci USA **92**(22): 10217-21.

Atchley, W. R. and W. M. Fitch (1997). "A natural classification of the basic helix-loop-helix class of transcription factors." Proc Natl Acad Sci USA **94**(10): 5172-6.

Atchley, W. R., W. Terhalle, et al. (1999). "Positional dependence, cliques, and predictive motifs in the bHLH protein domain." J Mol Evol **48**(5): 501-16.

Atchley, W. R. and J. Zhao (2007). "Molecular architecture of the DNA-binding region and its relationship to classification of basic helix-loop-helix proteins." Molecular Biology and Evolution **24**(1): 192-202.

Atchley, W. R., J. Zhao, et al. (2005). "Solving the protein sequence metric problem." Proc Natl Acad Sci USA **102**(18): 6395-400.

Barabási, A. and Z. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nat Rev Genet **5**(2): 101-113.

Bates, P. A., L. A. Kelley, et al. (2001). "Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM." Proteins **Suppl 5**: 39-46.

Baudino, T. A. and J. L. Cleveland (2001). "The Max network gone mad." Molecular and Cellular Biology **21**(3): 691-702.

Benassayag, C., L. Montero, et al. (2005). "Human c-Myc isoforms differentially regulate cell growth and apoptosis in Drosophila melanogaster." Molecular and Cellular Biology **25**(22): 9897-909.

Billin, A. N. and D. E. Ayer (2006). "The Mlx network: evidence for a parallel Max-like transcriptional network that regulates energy metabolism." Curr Top Microbiol Immunol **302**: 255-78.

Billin, A. N., A. L. Eilers, et al. (2000). "MondoA, a novel basic helix-loop-helix-leucine zipper transcriptional activator that constitutes a positive branch of a max-like network." Molecular and Cellular Biology **20**(23): 8845-54.

Billin, A. N., A. L. Eilers, et al. (1999). "Mlx, a novel Max-like BHLHZip protein that interacts with the Max network of transcription factors." J Biol Chem **274**(51): 36344-50.

Brown, S., M. Cole, et al. (2008). "Evolution of the holozoan ribosome biogenesis regulon." BMC Genomics **9**(1): 442.

Brownlie, P., T. Ceska, et al. (1997). "The crystal structure of an intact human Max-DNA complex: new insights into mechanisms of transcriptional control." Structure **5**(4): 509-20.

Budd, G. and M. Telford (2009). "The origin and evolution of arthropods." Nature **457**(7231): 812-817.

Burton, R. A., S. Mattila, et al. (2006). "B-myc: N-terminal recognition of myc binding proteins." Biochemistry **45**(32): 9857-65.

Cairo, S., G. Merla, et al. (2001). "WBSCR14, a gene mapping to the Williams--Beuren syndrome deleted region, is a new member of the Mlx transcription factor network." Human Molecular Genetics **10**(6): 617-27.

Carranza, S., J. Baguñà, et al. (1997). "Are the Platyhelminthes a monophyletic primitive group? An assessment using 18S rDNA sequences." Molecular Biology and Evolution **14**(5): 485-97.

Chapman, J. A., E. F. Kirkness, et al. (2010). "The dynamic genome of Hydra." Nature **464**(7288): 592-6.

Charron, J., B. A. Malynn, et al. (1992). "Embryonic lethality in mice homozygous for a targeted disruption of the N-myc gene." Genes & Development **6**(12A): 2248-57.

Coghlan, A. (2005). "Nematode genome evolution." WormBook : the online review of C elegans biology: 1-15.

Consortium, B. G. S. a. A., C. Elsik, et al. (2009). "The genome sequence of taurine cattle: a window to ruminant biology and evolution." Science **324**(5926): 522-8.

Consortium, H. G. S. (2006). "Insights into social insects from the genome of the honeybee Apis mellifera." Nature **443**(7114): 931-49.

Consortium, I. A. G. (2010). "Genome sequence of the pea aphid Acyrthosiphon pisum." Plos Biol **8**(2): e1000313.

Consortium, I. C. G. S. (2004). "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution." Nature **432**(7018): 695-716.

Consortium, I. S. G. (2008). "The genome of a lepidopteran model insect, the silkworm Bombyx mori." Insect Biochem Mol Biol **38**(12): 1036-45.

Consortium, T. G. S., S. Richards, et al. (2008). "The genome of the model beetle and pest Tribolium castaneum." Nature **452**(7190): 949-55.

Dang, C. V. (1999). "c-Myc target genes involved in cell growth, apoptosis, and metabolism." Molecular and Cellular Biology **19**(1): 1-11.

Dang, C. V., M. McGuire, et al. (1989). "Involvement of the 'leucine zipper' region in the oligomerization and transforming activity of human c-myc protein." Nature **337**(6208): 664-6.

Davis, A. C., M. Wims, et al. (1993). "A null c-myc mutation causes lethality before 10.5 days of gestation in homozygotes and reduced fertility in heterozygous female mice." Genes & Development **7**(4): 671-82.

de Luis, O., M. C. Valero, et al. (2000). "WBSCR14, a putative transcription factor gene deleted in Williams-Beuren syndrome: complete characterisation of the human gene and the mouse ortholog." Eur J Hum Genet **8**(3): 215-22.

Dehal, P. and J. Boore (2005). "Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate." Plos Biol **3**(10): e314.

Dehal, P., Y. Satou, et al. (2002). "The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins." Science **298**(5601): 2157-67.

Denver, D. R., K. Morris, et al. (2004). "High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome." Nature **430**(7000): 679-82.

Depinho, R., K. Hatton, et al. (1987). "The human myc gene family: structure and activity of L-myc and an L-myc pseudogene." Genes & Development **1**(10): 1311-1326.

Dillon, W. R. and S. Westin (1982). "Scoring Frequency Data for Discriminant Analysis: Perhaps Discrete Procedures Can Be Avoided." Journal of Marketing Research **19**(1): 44-56.

Doskocil, J. (1996). "The amplification of oligonucleotide themes in the evolution of the myc protooncogene family." J Mol Evol **42**(5): 512-24.

Durbin, R., S. Eddy, et al. (1998). "Biological sequence analysis: probabilistic models of proteins and nucleic acids, ." Cambridge University Press.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Research **32**(5): 1792-7.

Facchini, L. M. and L. Z. Penn (1998). "The molecular role of Myc in growth and transformation: recent discoveries lead to new insights." FASEB J **12**(9): 633-51.

Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." J Mol Evol **17**(6): 368-76.

Felsenstein, J. (2005). "PHYLIP (Phylogeny Inference Package) version 3.6. ." Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Ferré-D'Amaré, A. R., G. C. Prendergast, et al. (1993). "Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain." Nature **363**(6424): 38-45.

Fisher, R. (1936). "The Use of Multiple Measurements in Taxonomic Problems." Annals of Eugenics **7**: 179-188.

Fox, E. J., S. A. Stubbs, et al. (2004). "PRELI (protein of relevant evolutionary and lymphoid interest) is located within an evolutionarily conserved gene cluster on chromosome 5q34-q35 and encodes a novel mitochondrial protein." Biochem J **378**(Pt 3): 817-25.

Fujimoto, K., S. Ishihara, et al. (2008). "Network Evolution of Body Plans." PLoS ONE **3**(7): e2772.

Gallant, P. (2006). "Myc/Max/Mad in invertebrates: the evolution of the Max network." Curr Top Microbiol Immunol **302**: 235-53.

Gascuel, O. (1997). "BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data." Molecular Biology and Evolution **14**(7): 685-95.

Gibbs, R. (2010). "We thank the Honey Bee Genome Sequencing Consortium for making their data publicly available, and the BCM-HGSC for providing the genome and BLAST service. ."

Gibbs, R., G. Weinstock, et al. (2004). "Genome sequence of the Brown Norway rat yields insights into mammalian evolution." Nature **428**(6982): 493-521.

Grandori, C., S. M. Cowley, et al. (2000). "The Myc/Max/Mad network and the transcriptional control of cell behavior." Annu Rev Cell Dev Biol **16**: 653-99.

Grinberg, A. V., C. D. Hu, et al. (2004). "Visualization of Myc/Max/Mad family dimers and the competition for dimerization in living cells." Molecular and Cellular Biology **24**(10): 4294-308.

Guindon, S. and O. Gascuel (2003). "A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood." Systematic Biology **52**(5): 696-704.

Gupta, B. P. and P. W. Sternberg (2003). "The draft genome sequence of the nematode Caenorhabditis briggsae, a companion to C. elegans." Genome Biol **4**(12): 238.

Hahn, M. W., G. C. Conant, et al. (2004). "Molecular evolution in large genetic networks: does connectivity equal constraint?" J Mol Evol **58**(2): 203-11.

Hatton, K., K. Mahon, et al. (1996). "Expression and activity of L-Myc in normal mouse development." Molecular and Cellular Biology **16**(4): 1794-804.

Hedges, S. (2002). "The origin and evolution of model organisms." Nat Rev Genet **3**(11): 838-849.

Hill, C. A. and S. K. Wikel (2005). "The Ixodes scapularis Genome Project: an opportunity for advancing tick research." Trends Parasitol **21**(4): 151-3.

Hillier, L., G. T. Marth, et al. (2008). "Whole-genome sequencing and variant discovery in C. elegans." Nat Methods **5**(2): 183-8.

Hlavacek, W. S., J. R. Faeder, et al. (2003). "The complexity of complexes in signal transduction." Biotechnol Bioeng **84**(7): 783-94.

Hooker, C. W. and P. Hurlin (2006). "Of Myc and Mnt." Journal of Cell Science **119**(Pt 2): 208-16.

Hurlin, P., Z. Q. Zhou, et al. (2004). "Evidence of mnt-myc antagonism revealed by mnt gene deletion." Cell Cycle **3**(2): 97-9.

Hurlin, P. J., C. Quéva, et al. (1997). "Mnt, a novel Max-interacting protein is coexpressed with Myc in proliferating cells and mediates repression at Myc binding sites." Genes & Development **11**(1): 44-58.

Ideker, T., T. Galitski, et al. (2001). "A new approach to decoding life: systems biology." Annu Rev Genomics Hum Genet **2**: 343-72.

Iizuka, K. and Y. Horikawa (2008). "ChREBP: a glucose-activated transcription factor involved in the development of metabolic syndrome." Endocr J **55**(4): 617-24.

JGI (2010). "These sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community.".

Jones, S. (2004). "An overview of the basic helix-loop-helix proteins." Genome Biol **5**(6): 226.

Kanehisa, M., S. Goto, et al. (2010). "KEGG for representation and analysis of molecular networks involving diseases and drugs." Nucleic Acids Research **38**(Database issue): D355-60.

Kewley, R. J., M. L. Whitelaw, et al. (2004). "The mammalian basic helix-loop-helix/PAS family of transcriptional regulators." Int J Biochem Cell Biol **36**(2): 189-204.

King, N., M. J. Westbrook, et al. (2008). "The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans." Nature **451**(7180): 783-8.

Kitano, H. (2002). "Systems biology: a brief overview." Science **295**(5560): 1662-4.

Larkin, M. A., G. Blackshields, et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-8.

Lawson, D., P. Arensburger, et al. (2009). "VectorBase: a data resource for invertebrate vector genomics." Nucleic Acids Research **37**(Database issue): D583-7.

Levine, M. and R. Tjian (2003). "Transcription regulation and animal diversity." Nature **424**(6945): 147-51.

Lindblad-Toh, K., C. M. Wade, et al. (2005). "Genome sequence, comparative analysis and haplotype structure of the domestic dog." Nature **438**(7069): 803-19.

Loo, L. W., J. Secombe, et al. (2005). "The transcriptional repressor dMnt is a regulator of growth in Drosophila melanogaster." Molecular and Cellular Biology **25**(16): 7078-91.

Lüscher, B. (2001). "Function and regulation of the transcription factors of the Myc/Max/Mad network." Gene **277**(1-2): 1-14.

Lüscher, B. and L. G. Larsson (1999). "The basic region/helix-loop-helix/leucine zipper domain of Myc proto-oncoproteins: function and regulation." Oncogene **18**(19): 2955-66.

Ma, L., L. N. Robinson, et al. (2006). "ChREBP*Mlx is the principal mediator of glucose-induced gene expression in the liver." J Biol Chem **281**(39): 28721-30.

Ma, L., Y. Y. Sham, et al. (2007). "A critical role for the loop region of the basic helix-loop-helix/leucine zipper protein Mlx in DNA binding and glucose-regulated transcription." Nucleic Acids Research **35**(1): 35-44.

Ma, L., N. G. Tsatsos, et al. (2005). "Direct role of ChREBP.Mlx in regulating hepatic glucose-responsive genes." J Biol Chem **280**(12): 12019-27.

Maerkl, S. J. and S. R. Quake (2009). "Experimental determination of the evolvability of a transcription factor." Proc Natl Acad Sci USA **106**(44): 18650-5.

Massari, M. E. and C. Murre (2000). "Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms." Molecular and Cellular Biology **20**(2): 429-40.

McCarthy, A. A. (2005). "Broad institute: bringing genomics to real-world medicine." Chem Biol **12**(7): 717-8.

McDonald, W. H., Y. Pavlova, et al. (2003). "Novel essential DNA repair proteins Nse1 and Nse2 are subunits of the fission yeast Smc5-Smc6 complex." J Biol Chem **278**(46): 45460-7.

McMahon, S. B., H. A. Van Buskirk, et al. (1998). "The novel ATM-related protein TRRAP is an essential cofactor for the c-Myc and E2F oncoproteins." Cell **94**(3): 363-74.

Meroni, G., S. Cairo, et al. (2000). "Mlx, a new Max-like bHLHZip family member: the center stage of a novel transcription factors regulatory pathway?" Oncogene **19**(29): 3266-77.

Meroni, G., A. Reymond, et al. (1997). "Rox, a novel bHLHZip protein expressed in quiescent cells that heterodimerizes with Max, binds a non-canonical E box and acts as a transcriptional repressor." EMBO J **16**(10): 2892-906.

Mikkelsen, T., M. J. Wakefield, et al. (2007). "Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences." Nature **447**(7141): 167-77.

Moens, C. B., B. R. Stanton, et al. (1993). "Defects in heart and lung development in compound heterozygotes for two different targeted mutations at the N-myc locus." Development **119**(2): 485-99.

Moore, A., Å. Björklund, et al. (2008). "Arrangements in the modular evolution of proteins." Trends in Biochemical Sciences **33**(9): 444-451.

Moore, G. W., M. Goodman, et al. (1973). "An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets." Journal of Theoretical Biology **38**(3): 423-57.

Morgenstern, B. and W. R. Atchley (1999). "Evolution of bHLH transcription factors: modular evolution by domain shuffling?" Molecular Biology and Evolution **16**(12): 1654-63.

Morton, C. C., M. C. Nussenzweig, et al. (1989). "Mapping and characterization of an X-linked processed gene related to MYCL1." Genomics **4**(3): 367-75.

Nair, S. K. and S. K. Burley (2003). "X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors." Cell **112**(2): 193-205.

Nair, S. K. and S. K. Burley (2006). "Structural aspects of interactions within the Myc/Max/Mad network." Curr Top Microbiol Immunol **302**: 123-43.

Nene, V., J. R. Wortman, et al. (2007). "Genome sequence of Aedes aegypti, a major arbovirus vector." Science **316**(5832): 1718-23.

Nilsson, J. A., K. H. Maclean, et al. (2004). "Mnt loss triggers Myc transcription targets, proliferation, apoptosis, and transformation." Molecular and Cellular Biology **24**(4): 1560-9.

O'Hagan, R. C., N. Schreiber-Agus, et al. (2000). "Gene-target recognition among members of the myc superfamily and implications for oncogenesis." Nat Genet **24**(2): 113-9.

Orian, A., B. van Steensel, et al. (2003). "Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network." Genes & Development **17**(9): 1101-14.

Partch, C. and K. H. Gardner (2010). "Coactivator recruitment: A new role for PAS domains in transcriptional regulation by the bHLH-PAS family." J. Cell. Physiol.: n/a-n/a.

Peterson, C., C. Stoltzman, et al. (2010). "Glucose Controls Nuclear Accumulation, Promoter Binding, and Transcriptional Activity of the MondoA-Mlx Heterodimer." Molecular and Cellular Biology **30**(12): 2887-2895.

Peyrefitte, S., D. Kahn, et al. (2001). "New members of the Drosophila Myc transcription factor subfamily revealed by a genome-wide examination for basic helix-loop-helix genes." Mech Dev **104**(1-2): 99-104.

Pickett, C., K. Breen, et al. (2007). "A C. elegans Myc-like network cooperates with semaphorin and Wnt signaling pathways to control cell migration." Developmental Biology **310**(2): 226-239.

Pierce, S. B., C. Yost, et al. (2008). "Drosophila growth and development in the absence of dMyc and dMnt." Developmental Biology **315**(2): 303-16.

Pierce, S. B., C. Yost, et al. (2004). "dMyc is required for larval growth and endoreplication in Drosophila." Development **131**(10): 2317-27.

Pond, S. L., S. D. Frost, et al. (2005). "HyPhy: hypothesis testing using phylogenies." Bioinformatics **21**(5): 676-9.

Postic, C., R. Dentin, et al. (2007). "ChREBP, a transcriptional regulator of glucose and lipid metabolism." Annu Rev Nutr **27**: 179-92.

Putnam, N., M. Srivastava, et al. (2007). "Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization." Science **317**(5834): 86-94.

Quackenbush, J., J. Cho, et al. (2001). "The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species." Nucleic Acids Research **29**(1): 159-64.

Reddy, C. D., P. Dasgupta, et al. (1992). "Mutational analysis of Max: role of basic, helix-loop-helix/leucine zipper domains in DNA binding, dimerization and regulation of Myc-mediated transcriptional activation." Oncogene **7**(10): 2085-92.

Robinson, K. A. and J. M. Lopes (2000). "SURVEY AND SUMMARY: Saccharomyces cerevisiae basic helix-loop-helix proteins regulate diverse biological processes." Nucleic Acids Research **28**(7): 1499-505.

Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics **19**(12): 1572-4.

Rottmann, S. and B. Lüscher (2006). "The Mad side of the Max network: antagonizing the function of Myc and more." Curr Top Microbiol Immunol **302**: 63-122.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Molecular Biology and Evolution **4**(4): 406-25.

Sanger (2010). "These data were provided by the Zebrafish and C.elegans group at the Wellcome Trust Sanger Institute and were obtained from www.sanger.ac.uk/DataSearch/blast.shtml."

Sans, C., D. Satterwhite, et al. (2006). "MondoA-Mlx Heterodimers Are Candidate Sensors of Cellular Energy Status: Mitochondrial Localization and Direct Regulation of Glycolysis." Molecular and Cellular Biology **26**(13): 4863-4871.

Sawai, S., A. Shimono, et al. (1993). "Defects of embryonic organogenesis resulting from targeted disruption of the N-myc gene in the mouse." Development **117**(4): 1445-55.

Sayers, E. W., T. Barrett, et al. (2010). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Research **38**(Database issue): D5-16.

Schreiber-Agus, N., Y. Meng, et al. (1998). "Role of Mxi1 in ageing organ systems and the regulation of normal and neoplastic growth." Nature **393**(6684): 483-7.

Scott, A. L. and E. Ghedin (2009). "The genome of Brugia malayi - all worms are not created equal." Parasitol Int **58**(1): 6-11.

Shannon, C. E. (1948). "A Mathematical Theory of Communication." Bell System Technical Journal **27**: 379–423

623–656.

Sharakhova, M. V., M. P. Hammond, et al. (2007). "Update of the Anopheles gambiae PEST genome assembly." Genome Biol **8**(1): R5.

Shih, H. M., Z. Liu, et al. (1995). "Two CACGTG motifs with proper spacing dictate the carbohydrate regulation of hepatic gene transcription." J Biol Chem **270**(37): 21991-7.

Shih, H. M. and H. C. Towle (1992). "Definition of the carbohydrate response element of the rat S14 gene. Evidence for a common factor required for carbohydrate regulation of hepatic genes." J Biol Chem **267**(19): 13222-8.

Siegal, M., D. Promislow, et al. (2006). "Functional and evolutionary inference in gene networks: does topology matter?" Genetica **129**(1): 83-103.

Solomon, D. L., B. Amati, et al. (1993). "Distinct DNA binding preferences for the c-Myc/Max and Max/Max dimers." Nucleic Acids Research **21**(23): 5372-6.

Srivastava, M., E. Begovic, et al. (2008). "The Trichoplax genome and the nature of placozoans." Nature **454**(7207): 955-60.

Stein, L. D., Z. Bao, et al. (2003). "The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics." Plos Biol **1**(2): E45.

Stoltzman, C. A., C. W. Peterson, et al. (2008). "Glucose sensing by MondoA:Mlx complexes: a role for hexokinases and direct regulation of thioredoxin-interacting protein expression." Proc Natl Acad Sci USA **105**(19): 6912-7.

Subramanian, A. R., M. Kaufmann, et al. (2008). "DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment." Algorithms for molecular biology : AMB **3**: 6.

Toyo-Oka, K., S. Hirotsune, et al. (2004). "Loss of the Max-interacting protein Mnt in mice results in decreased viability, defective embryonic growth and craniofacial defects: relevance to Miller-Dieker syndrome." Human Molecular Genetics **13**(10): 1057-67.

Tweedie, S., M. Ashburner, et al. (2009). "FlyBase: enhancing Drosophila Gene Ontology annotations." Nucleic Acids Res **37**(Database issue): D555-9.

Van Dam, T., B. Snel, et al. (2008). "Protein Complex Evolution Does Not Involve Extensive Network Rewiring." PLoS Computational Biology **4**(7): e1000132.

Venkatesh, B., E. F. Kirkness, et al. (2007). "Survey sequencing and comparative analysis of the elephant shark (Callorhinchus milii) genome." Plos Biol **5**(4): e101.

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Werren, J. H., S. Richards, et al. (2010). "Functional and evolutionary insights from the genomes of three parasitoid Nasonia species." Science **327**(5963): 343-8.

Wilson, R. K. (2007). AC195499 Direct submission. Genetics, Genome Sequencing Center.

Witherspoon, D. J. and H. Robertson (2003). "Neutral evolution of ten types of mariner transposons in the genomes of Caenorhabditis elegans and Caenorhabditis briggsae." J Mol Evol **56**(6): 751-69.

Wróbel, B. (2008). "Statistical measures of uncertainty for branches in phylogenetic trees inferred from molecular sequences by using model-based methods." J Appl Genet **49**(1): 49-67.

Yang, Z. (1996). "Among-site rate variation and its impact on phylogenetic analyses." Tree **11**(9): 367-392.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Molecular Biology and Evolution **24**(8): 1586-91.

Yuan, J., R. S. Tirabassi, et al. (1998). "The C. elegans MDL-1 and MXL-1 proteins can functionally substitute for vertebrate MAD and MAX." Oncogene **17**(9): 1109-18.

Chapter 4

A Bioinformatics approach towards annotating the glucose response of MondoA and ChREBP

Abstract

Glucose is a fundamental energy source for both prokaryotes and eukaryotes. The balance between glucose utilization and storage is integral for proper energy homeostasis, and defects are associated with several diseases, e.g. metabolic syndrome and type II diabetes. In vertebrates, the transcription factor ChREBP is a major component in glucose metabolism, while its ortholog MondoA is involved in glucose uptake. Both MondoA and ChREBP contain five Mondo conserved regions (*MCRI-V*) that affect their cellular localization and transactivation ability. While phosphorylation has been shown to affect ChREBP function, the mechanisms controlling glucose response of both ChREBP and MondoA remain elusive. By incorporating sequence analysis techniques, structure predictions, and functional annotations, we propose a model involving the *MCR*s and two additional domains that determine ChREBP and MondoA glucose response. Paramount, we identified a conserved motif within the previously reported Myc box II-like region and propose that this region interacts with the phosphorylated form of glucose. In addition, we discovered a putative nuclear receptor box in invertebrate Mondo and vertebrate ChREBP sequences that reveals a potentially novel interaction with nuclear receptors. These interactions are likely involved in altering ChREBP and MondoA conformation to form an active complex and induce transcription of genes involved in glucose metabolism and lipogenesis.

Introduction

Glucose is a carbohydrate in the form of a simple sugar that is an important source of energy for both eukaryotes and prokaryotes (reviewed in Girard, Ferré et al. 1997; Towle 2005). In mammals, the liver is the primary organ that controls energy homeostasis by processing glucose for energy or storage. In fasting conditions, the liver produces glucose via *de novo* synthesis (gluconeogenesis) or decomposition of glycogen (glycogeneolysis). Glucose can then be converted to pyruvate through glycolysis and subsequently enter the citric acid (TCA) cycle within mitochondria to produce energy. In contrast, when excess carbohydrates are consumed, glucose can be stored according to two major pathways. Insulin induced enzymes trigger the glycogen synthase pathway to store glucose as glycogen. Alternatively, glucose can be converted to triglycerides through the *de novo* lipogenesis pathway for a more compact form of storage. Triglycerides within the liver can be further packaged into lipoproteins (i.e. VLDL, LDL, HDL) and transported into the blood stream and other tissues.

Initially, transcription factor SREBP1 was identified as the major factor involved in glucose metabolism and insulin response (Postic, Dentin et al. 2007). However, knockout experiments revealed an additional factor was necessary for the full glucose dependent transactivation of certain lipogenic genes, e.g. acetyl-CoA carboxylase (ACC) and fatty acid synthase (FAS) . The discovery of a conserved carbohydrate response element (ChORE) consisting of two E-boxes separated by exactly 5 residues ($CACGTGN_5CACGTG$) within the promoters of such genes facilitated the identification of this glucose responsive element ; ChORE binding protein ChREBP is mainly expressed in the liver and pancreatic beta cells within mammals (Cairo, Merla et al. 2001; Rufo, Teran-Garcia et al. 2001; Iizuka, Bruick et al. 2004) and has subsequently been implicated in transactivation of several genes that regulate the *de novo* lipogenesis pathway, e.g. liver pyruvate kinase (L-PK), malic enzyme (ME), glucose phosphoisomerase (GPI), ACC, and FAS (Yamashita, Takenoshita et al. 2001).

ChREBP protein, also named WBSCR14, MondoB and MLXIPL, has a paralog in vertebrates named MondoA. Like ChREBP, MondoA is ubiquitously expressed in

embryonic cells, although its expression in adult cells is limited to skeletal muscle (Ma, Robinson et al. 2006). The distinct expression profiles of MondoA and ChREBP underlie their downstream effects and separate roles in regulating genes involved in glucose metabolism: MondoA restricts glucose uptake and influences energy utilization while ChREBP signals energy storage through *de novo* lipogenesis (Billin, Eilers et al. 2000; Sans, Satterwhite et al. 2006).

Only a single mondo gene has been identified in invertebrate animals, including *Drosophila melanogaster* (*dmondo*) and *Caenorhabditis elegans* (*mml-1*), which are also named *mio* and *T20B12.6*, respectively (Ma, Tsatsos et al. 2005). Consistent with both MondoA and ChREBP expression, MML-1 is expressed in glucose dependent epidermal and intestinal cells during embryogenesis through all larval stages and adulthood (Billin and Ayer 2006). Likewise, dMondo is expressed in the amnioserosa during larval, pupal, and adult stages and has mixed expression throughout the head, testis, fat body, gonad, and embryonic tissue (Pickett, Breen et al. 2007).

Both ChREBP and MondoA are glucose responsive, whereby they are mainly located in the cytoplasm under low glucose conditions and have increased nuclear accumulation and transactivation of target genes in high glucose medium (Peyrefitte, Kahn et al. 2001; Ma, Tsatsos et al. 2005; Sans, Satterwhite et al. 2006). This nuclear translocation and DNA binding is dependent upon the dimerization to obligate partner Mlx, which is ubiquitously expressed. Mlx and Mondo proteins contain a C-terminal basic Helix-Loop-Helix-Leucine Zipper (*bHLHZ*) domain responsible for DNA binding and dimerization as well as a dimerization and cytoplasmic localization (*DCD*) domain that must be masked prior to nuclear entry (Yamashita, Takenoshita et al. 2001; Eilers, Sundwall et al. 2002). Dimerization through either the *bHLHZ* or *DCD* region is sufficient to block this cytoplasmic retention signal (CRS), but not sufficient for nuclear translocation (Eilers, Sundwall et al. 2002; Billin and Ayer 2006; Peterson, Stoltzman et al. 2010). While domain names are not generally italicized, we adopt this naming convention to avoid confusion with protein references throughout this manuscript.

Since MondoA and ChREBP are mainly cytoplasmic proteins, it was surprising to find that trapping them within the nucleus in low glucose conditions was not sufficient to replicate the transactivation potential (Davies, O'Callaghan et al. 2008; Peterson, Stoltzman et al. 2010). Consistent with this, both MondoA and ChREBP are known to shuttle between the cytoplasm and nucleus in both low and high glucose conditions, yet have increased transactivation only under high glucose. In contrast, proteins lacking the N-terminus are able to constitutively transactivate genes in both glucose mediums (Eilers, Sundwall et al. 2002; Li, Chang et al. 2006; Li, Chen et al. 2008; Stoltzman, Peterson et al. 2008), indicating additional N-terminal domains within MondoA and ChREBP contribute to their nuclear accumulation and transactivation in response to glucose (Li, Chang et al. 2006; Davies, O'callaghan et al. 2010).

MondoA and ChREBP proteins have five Mondo Conserved Regions (*MCRI-V*) in their N-terminus. These have previously been reported as *PADRE1*, *PADRE2*, and *MADRE* (Eilers, Sundwall et al. 2002) as well as a low glucose inhibitory domain (*LID*) and glucose responsive activation conserved element (*GRACE*) (Cairo, Merla et al. 2001). The *LID* spans *MCRI-IV*, *PADRE1* encompasses *MCRII-IV*, *PADRE2* and *GRACE* contain *MCRV*, and *MADRE* is a large central region that is variable. The distances between *MCRII*, *MCRIII*, and *MCRIV* are also conserved, implying they act as a functional module, while the regions linking *MCRI* and *MCRV* vary between MondoA and ChREBP (Li, Chang et al. 2006). *MCRII* contains a strong CRM1 dependent nuclear export signal (NES), almost identical to the high affinity LxxLFxxLSV motif (Billin and Ayer 2006). In contrast, *MCRIV* in ChREBP contains a bipartite nuclear localization signal (NLS) that mediates its nuclear entry (Kawaguchi, Takenoshita et al. 2001; Kutay and Güttinger 2005). Between these two regions *MCRIII* contains a binding motif recognized by the 14-3-3 protein that is involved in ChREBP and MondoA cytoplasmic retention, transactivation, and nuclear export (Yamashita, Takenoshita et al. 2001; Eilers, Sundwall et al. 2002; Merla, Howald et al. 2004). The functions of *MCRI* and *MCRV* are not as clear, although *MCRI* is necessary for glucose dependent transactivation in ChREBP (Li, Chen et al. 2008) and *MCRV* is within the *GRACE* region responsible for transactivation (Tsatsos, Davies et al. 2008).

Each *MCR* seems to have multiple and often opposing function. *MCRI* is necessary for glucose response, since alterations to *MCRI* (ChREBP: Δ1-71, Δ1-58; MondoA: Δ1-100, H78A/H81A/H88A) block transactivation in high glucose, yet mimicking phosphorylation (ChREBP: S56D) enhances it (Li, Chang et al. 2006; Davies, O'Callaghan et al. 2008; Li, Chen et al. 2008; Tsatsos, Davies et al. 2008). Likewise, altering the NES in *MCRII* (ChREBP: L89A, F90A; MondoA: F130A, M133A, Δ125-137) mildly enhances transactivation, while other mutations in *MCRII* (ChREBP: L86A/L93A, T85A, L95A, Δ72-99; MondoA: L129A) completely block it (Eilers, Sundwall et al. 2002; Fukasawa, Ge et al. 2010; Peterson, Stoltzman et al. 2010). In *MCRIII*, abrogating 14-3-3 protein binding sites (ChREBP: R128A, W130A; MondoA: I166A/W167A/R168A) inhibit transactivation, but so do mutations (ChREBP: N123A, I126A, Δ100-115) that are still capable of interacting with 14-3-3 (Davies, O'Callaghan et al. 2008; Li, Chen et al. 2008; Sakiyama, Wynn et al. 2008).

Intriguingly, changes within *MCRIV* have even more diverse effects. Some changes (ChREBP: Δ141-197, Δ158-181) likely block the NLS and thus prevent transactivation (Kawaguchi, Takenoshita et al. 2001; Peterson, Stoltzman et al. 2010), one change (ChREBP: Δ144-196) reduces transactivation function yet also removes glucose dependent inhibition (Davies, O'callaghan et al. 2010), while another change (MondoA: Y210D/W211D/K212) increases nuclear accumulation and transactivation (Li, Chen et al. 2008). While *MCRV* shows no repressive effects in the absence of *MCRI-IV*, changes to it (ChREBP: Y275A/V276A/G277A, L289A/Q290A/P291A; MondoA: Δ282-324) within the full-length sequence lead to an increase in nuclear accumulation and transactivation (Davies, O'Callaghan et al. 2008; Peterson, Stoltzman et al. 2010). Although the cellular conditions, site mutations, and reporter assays in these studies greatly vary, they individually and in combination suggest that the *MCRs* cooperatively repress and activate MondoA and ChREBP.

The N-terminal *LID* possesses a robust repressive mechanism that regulates the strong transactivation region within the *GRACE*. Contrary to prediction, individually deleting or mutating *MCRI, II, III,* or *IV* also abolishes MondoA or ChREBP transactivation in response to glucose (Li, Chen et al. 2008; Peterson, Stoltzman et al. 2010). Hence the *LID* participates

in repression in low glucose and activation in high glucose, where no individual *MCR* can sufficiently replicate the glucose response. Moreover, reversing the order of *LID* and *GRACE* regions results in a constitutively active ChREBP protein, indicating its structure and intramolecular contacts are major factors in regulating Mondo function (Davies, O'callaghan et al. 2010).

Recent evidence indicates phosphorylation of glucose by hexokinase to form G6P has a direct impact on the activation of MondoA and ChREBP, although the mechanism is still uncertain (Li, Chang et al. 2006). How G6P is able to promote transactivation within the *GRACE* and override the N-terminal repression imposed by the *LID* region is an important, yet unanswered question. In addition, low glucose repression seems to be independent of a cofactor and is likely a result of protein conformation (Peterson, Stoltzman et al. 2010). Determining the function and interactions of *MCR*s within the N-terminus is of great import to understanding MondoA and ChREBP glucose response and transactivation of genes involved in glucose metabolism.

To balance energy storage and usage, extracellular signals, including insulin and glucagon, instigate the expression and phosphorylation of proteins involved in the lipogenic pathway. ChREBP contains several such functional phosphorylation sites (Davies, O'callaghan et al. 2010). A ChREBP based phosphorylation model postulates that during starvation glucagon increases the concentration of cAMP in hepatocytes, which triggers the phosphorylation of ChREBP by cAMP dependent protein kinase A (PKA) (Li, Chang et al. 2006). Phosphorylation of ChREBP site Ser196 causes an adjacent bipartite nuclear localization signal (NLS) in *MCRIV* to be blocked and ChREBP to be sequestered in the cytosol (Yamashita, Takenoshita et al. 2001). Conversely, dephosphorylation events mediate a conversion to energy storage rather than usage after a high carbohydrate meal. Increased glucose and thus accelerated glycolytic flux increases the concentration of intermediate enzyme Xylulose-5-phosphate (X5P) within the pentose phosphate shunt, which stimulates an isoform of protein phosphatase 2A (PP2A) (Kawaguchi, Takenoshita et al. 2001). Cytosolic PP2A mediated dephosphorylation of S196 in ChREBP results in its nuclear localization, while ChREBP DNA binding and transactivation is enhanced by further

dephosphorylation of sites S626 and T666 via X5P activated PP2A in the nucleus (Figure 1) (Kabashima, Kawaguchi et al. 2003).

While this simple model is attractive, it is not complete and several questions still remain. Foremost, MondoA is glucose responsive although it does not contain many of the phosphorylation sites found in ChREBP, and mimicking the phosphorylation status in ChREBP is not sufficient to activate transcriptional machinery in low glucose. Herein, we address the following issues. First, can we expect MondoA and ChREBP domains to function similarly? Further, what does their overall conservation imply in terms of their structure and function? Second, how do these proteins sense changing glucose levels and how does this alter their transactivation potential? Finally, can we form a cohesive model based on the current information that explains MondoA and ChREBP subcellular localization and transactivation in response to glucose?

Results

MCRI-V, bHLHZ, and DCD domains are conserved among Mondo Sequences

Using sequences from species sampled across the animal kingdom, we identify and quantify the similarity among ChREBP, MondoA, and non-vertebrate Mondo proteins. As previously reported (Sakiyama, Wynn et al. 2008), the similarity within Mondo protein sequences is largely contained within the *MCRI-V*, *bHLHZ*, and *DCD* domains. This can be directly observed through the Jenson-Shannon Divergence (JS) score (Figure 2), which rates each site by an autocorrelated conservation value (Billin and Ayer 2006). Since conservation is a powerful predictor for detecting functional sites (Capra and Singh 2007), sites within more conserved regions have higher JS values and are thus more likely to affect protein function (Figure 2a). Similarly, entropy (H) measures the amount of information or variability within an alignment column where conserved sites have low entropy values. As expected, sites within the *MCR*, *bHLHZ*, or *DCD* regions are highly conserved and have correspondingly high JS and low H values (Figure 2).

However, the relationship between JS and H is nonlinear due to several autapomorphies within the full sequence alignment (Figure 2b). In these cases, sequence specific insertions or poor prediction of exon boundaries for unannotated sequences create

alignment columns with just a single or few residues. While the entropy values for these sites can largely vary according to the amino acid distribution, the JS score will be weighted according to the conservation of neighboring sites and should remain relatively stable. By removing alignment positions with less than ten residues, we were able to recover the correlation between entropy and JS scores (r^2=0.55) (Figure 2b), as well as reveal a bimodal distribution of entropy values (Figure 2c). From this reduced dataset, 127 (11.6%) sites are considered highly conserved with H<2.0. Since JS values are scored using an adjacency window, the JS distribution is smoothed to form a single peak and there is no clear delineation of conserved and variable sites (data not shown). Still, in accordance with entropy values, setting an arbitrary 90% threshold (JS>0.5597) shows the most conserved 10% of sites are within the *MCR* and *bHLHZ* regions (Figure 2a).

High JS scores were also observed for two new potentially important regions. The first region, which we name Mondo Conserved Region 6 (*MCR6*), was previously reported as a MBII-like region located between *MCRIV* and *MCRV* (Petrova and Wu 2006). However, the MB-II like region did not contain the highly conserved [ST]DTLF[ST] motif, where [ST] indicates either a serine or threonine. The conservation of *MCR6* residues, as well as *MCRI-V*, are depicted by the weblogos in Figure 3, where larger letters indicate more conserved sites. Based on the distribution of amino acids, we propose *MCR6* be defined by the 12 residue sequence signature [MLD][SNED][EDML][FIM][ST]DTLF[ST][STM][LTI].

Mondo and MondoB proteins contain a Nuclear Receptor Box

JS scores also revealed a LxQLLT motif located within the central region of MondoB and non-vertebrate Mondo protein sequences, but not MondoA (Figure 4). This sequence conforms to the LxxLL nuclear receptor box (*NRB*) signature that participates in the ligand dependent activation of nuclear receptors. *NRB*s are found within nuclear receptor coactivators such as the SRC-1 family of proteins (PF08832), which typically have multiple repeats of this motif, each sufficient for ligand interaction with several nuclear receptors (Billin and Ayer 2006). Mondo and MondoB proteins only contain one *NRB*.

Interestingly, ChREBP and nuclear receptor HNF4α have adjacent recognition sequences in the promoter sequence of liver pyruvate kinase (L-PK) (Nolte, Wisely et al.

1998; Odom, Zizlsperger et al. 2004; Ma, Robinson et al. 2006; Xu, Christian et al. 2006). Full activation of the L-PK gene requires both ChREBP and HNF4α (Zhang, Metukuri et al. 2010), and ChREBP:HNF4α:CBP is recruited as a complex to the L-PK promoter region in a glucose dependent manner (Xu, Christian et al. 2006). Taking this into consideration, it is reasonable to assume that the ChREBP *NRB* is capable of activating HNF4α.

The importance of MCR and DCD invariant positions

By isolating columns with zero entropy and hence no variation, we identify 24 invariant sites within the alignment, all of which are contained within the *MCR* and *DCD* regions. *MCRIII* contains two groupings of invariant residues P104/W106/F109 and R121/L122/N123/N124/W127/R128 (human ChREBP numbering). Interestingly, mutation to paralogous MondoA sites **P144A**/K145A/**W146A** (invariant sites in bold) did not affect 14-3-3 binding and no other phenotypic variations were reported (Burke, Collier et al. 2009). However, a serine or threonine immediately precedes P104 in all sequences, indicating this may be an important phosphorylation site for Mondo proteins. Likewise, the alpha helix spanning ChREBP sites 116-135 is essential for 14-3-3 binding as is R128A (Eilers, Sundwall et al. 2002), suggesting the RLNN motif is involved in 14-3-3 interactions. However an N123A mutation demonstrates it not necessary for 14-3-3 binding, but is essential for transactivation (Li, Chen et al. 2008). Hence, it is currently unclear how these invariant sites contribute to Mondo function.

Sites F145 and P148 are also invariant, yet have not been previously included in a specific *MCR* sequence. These residues (bold) are within a conserved [KR]x[KRN][NSTP][PLIV][VFI][CIV]x**F**[AVI][STV]**P**[LIV] motif that is located directly downstream of *MCRIII* (underlined). With the exception of upstream insertions within tunicate *Molgula tectiformis* (KILRRYGY), and nematodes *C. elegans* (KKQP) and *Brugia malayi* (RPDKD), this conserved region is contiguous with the remainder of MCRIII and thus we include these additional sites within MCRIII (Figure 3). As before, the bias for serine and threonine before P148 suggests a putative phosphorylation site in Mondo, but not MML-1 proteins, which have a valine instead.

MCRIV sites W170/Y181/W184/R185 are also invariant, along with P291 of *MCRV*. Alanine mutations of MondoA sites Y211/**W212**/K213 resulted in nuclear accumulation in low and high glucose as well as three-fold induction of TXNIP reporter gene in L6 myoblasts (Davies, O'Callaghan et al. 2008). Similar results were observed for L289A/Q290A/**P291A** mutation in ChREBP with two-fold ACC gene reporter expression in 832/13 cells (Peterson, Stoltzman et al. 2010). Hence these sites appear to be involved in Mondo repression. The remaining eight invariant positions are within the *DCD* region, represented by ChREBP sites L735, P736, W801, R812, P813, L819, L822, and P832. While their function is unknown, sites L735/P736 are located directly after the *bHLHZ* and may be important for correctly orienting the *DCD* domain.

Surprisingly, *MCRI*, *MCRII*, and the *bHLHZ* region lack invariant residues. However, high JS scores indicate these regions as well as others within *MCRIII* and *MCRIV* are still functionally conserved among species. For example, divergence of the predicted protein sequence in beetle *Tribolium castaneum* (XP_973749.2) prevents the identification of otherwise invariant residues **HSG**x**FM**x**S** within *MCRI*, where bold letters are conserved and x represents a variable site. *MCRII* in *Tribolium* is also not conserved, suggesting its N-terminal region is divergent or incorrectly identified. Regardless, most *MCRII* site variability arises from divergence in nematodes and other more distantly related species, which may indicate changes in selective pressure in Arthropoda and Deuterostoma lineages. In contrast, no single sequence is responsible for *bHLHZ* variability, although it appears that nematode, ghost shark *Callorhinchus milii* and sea squirt *Ciona intestinalis* often differ at otherwise conserved sites. A more detailed discussion of *bHLHZ* conservation and divergence can be found in (McFerrin and Atchley, in preparation).

Mondo N- and C-terminal regions have conserved secondary structure

Secondary structure predictions of ChREBP, MondoA and Mondo indicate that the majority of their protein sequences are random coil, with several α-helices and intermittent β-sheets (Figure 2D, data not shown). Predictably, the α-helices and β-sheets overlap the *MCR*, *bHLHZ*, and *DCD* conserved regions described above, as well as *MCR6* and the *NRB* in Mondo and ChREBP sequences.

The α-helices comprising the *bHLHZ* and *DCD* domains are necessary for DNA binding, dimerization and subcellular localization (Eilers, Sundwall et al. 2002; Davies, O'Callaghan et al. 2008; Peterson, Stoltzman et al. 2010). Likewise, three α-helices within *MCRII*, *MCRIII*, and *MCRIV* correspond to a NES, 14-3-3 binding region, and NLS respectively and are critical for proper function (Billin and Ayer 2006). In particular, *MCRII* residues have been found to be independently essential for transactivation in addition to CRM1 dependent nuclear export (Sakiyama, Wynn et al. 2008). The residues necessary for these functions are more highly conserved and located on the same side of the α-helix (Figure 5), possibly creating a surface for competitive interaction.

Mondo proteins have disparate Proline and Glutamine Rich Regions

The length of Mondo proteins is relatively stable, despite extensive changes within the central region. The proximal region of both MondoA and ChREBP contains a proline rich region (*PRR*) that is retained among most vertebrates, although we were unable to find any identifiable stretch of homology between MondoA and ChREBP *PRR*s. Moreover, the *PRR* is not found within any non-vertebrate species. Instead, most non-vertebrates display a glutamine rich region (*GRR*) (Table 1). While the average central length of sequences containing the *GRR* ($\bar{x} = 543$) is not significantly different than sequences with neither *PRR* nor *GRR* ($\bar{x} = 462.14$), the average central length of *PRR* containing regions ($\bar{x} = 355.75$) is significantly shorter than sequences containing either *GRR* (p-val=0.007695) or neither domain (p-val=0.03586) according to a two-tailed t-test. However, the length and preservation of these domains is considerable, indicating they may serve as indiscriminate scaffolding regions, as seen in other *PRR* and *GRR* containing proteins (Kay, Williamson et al. 2000; Davies, O'Callaghan et al. 2008).

DCD/WMC is conserved among Mlx and Mondo proteins

For MondoA, and presumably ChREBP, to enter the nucleus, dimerization with Mlx must first occur. This is due to a cytoplasmic retention signal (CRS) located within the *DCD*, which is directly downstream of the *bHLHZ* domain (Eilers, Sundwall et al. 2002; Guo, Han et al. 2007). The *DCD* region provides an additional and independent interaction interface between Mondo and Mlx proteins, which masks the CRS and allows for nuclear entry.

While most of our understanding regarding this region is based on MondoA mutations, observations concerning the homologous and extended sequence WBSCR14-Mlx C-tail (*WMC*) region of ChREBP provides similar results (Merla, Howald et al. 2004; Peterson, Stoltzman et al. 2010).

To determine *DCD/WMC* functional attributes, we compared Mondo and Mlx protein sequences using multiple entropy measures (see Methods). From the *DCD/WMC* alignment columns (Figure 6) containing more than three residues, sites K41, F42, W81, L91, and L102 are nearly invariant across all Mlx and Mondo sequences with entropy less than 0.1 ($H<0.1$), while columns 5, 6, 13, 21, 41, 42, 44, 55, 56, 60, 81, 82, 83, 86, 91, 96, and 102 display conservation with functional entropy less than 0.1 ($H_{FG}<0.1$) (see Methods). As expected, sites with $H<0.1$ also have $H_{FG}<0.1$. Accordingly, residues K41, F42, S54, and F56 of MondoA and Mlx are important determinants of heterodimerization (Tsatsos and Towle 2006). Compared to the Mondo invariant sites, only W81 is invariant in both Mondo and Mlx, although L91 is conserved in all but the nematode sequences.

Previous reports claim that *C. elegans* MML-1 lacks a *DCD* region (Eilers, Sundwall et al. 2002). However, we find that *C. elegans* MML-1 is conserved at 10 (58.8%) of these 17 functionally constrained sites as well as the eight invariant Mondo residues. Moreover, the *DCD* region of MML-1 is 46.7% similar and 21.3% identical to mosquito *Culex pipiens*, while nematode Mlx homolog Mxl-2 is 40% similar and 16.2% identical to the Mlx *DCD* sequence in beetle *Tribolium casteum*. Hence, we assert that the *DCD* region is intact in *C. elegans* MML-1 and Mxl-2 proteins.

DCD/WMC Structure forms an alpha helix bundle

Secondary structure predictions of the *DCD/WMC* region for MondoA, ChREBP, and Mondo identifies five alpha helices, while only four were found for Mlx sequences (Figure 8). Previously, just the *DCD* region was considered in structure prediction of ChREBP and a zipper like tertiary structure was assumed (de Luis, Valero et al. 2000). However, by including the entire *WMC* region, the powerful 3-D structure software Rosetta predicts the ChREBP *WMC* model assumes a cyclin-like confirmation with five grouped alpha helices, Figure 8a (Pickett, Breen et al. 2007). This predicted configuration forms a groove flanked

by hydrophobic residues in alpha helices 1, 2, 3, and 4 designated by alignment sites **21**, 25, and 29 of α1, **44**, 47, 48, 49, 52 of α2, 65, 68, 73, and **82** of α3, and 88, **91**, 95, **96**, **102** and 105 of α4, where functionally conserved residues are in bold.

This interior region also displays increased conservation according to both entropy and Consurf estimates (Figure 8b). The program Consurf estimates the evolutionary rate of each site by comparing homologous sequences and similar protein structures (Rohl, Strauss et al. 2004). Consurf predicts ChREBP residues V6, K41, F42, S55, W81, L88, and L102 (*DCD/WMC* alignment numbering) have high conservation scores and are likely functionally important. Besides L88, these positions have low functional entropy for all Mondo and Mlx sequences, suggesting a common function.

The *DCD/WMC* of Mlx and Mondo show clear similarity, although protein distinctions likely affect their tertiary conformation. First, the *DCD/WMC* region of Mlx overlaps the 21-residue zipper region, while a zipper and linker region of Mondo sequences extends for 35 residues before the *DCD/WMC* begins. In addition, Mondo invariant sites L735/P736 are alternatively conserved for charged residues lysine and either aspartate or glutamate, which may affect the *DCD* orientation. Moreover, helix 5 shows considerable variability within the Mondo sequences, and may not be directly involved in protein-protein interactions, as it is completely lost in most Mlx sequences. These differences may restrict interaction between *DCD/WMC* regions and factor in the prevention of MondoA and Mlx homodimerization (Ashkenazy, Erez et al. 2010). Furthermore, it has been proposed that two ChREBP:Mlx dimers form a tetramer and bind DNA so that the loops of Mlx interact (Eilers, Sundwall et al. 2002). This configuration would orient the complex so Mondo flanks both sides of the DNA and the asymmetrical *DCD*s of ChREBP and Mlx would also interact between dimers to form a more compact and organized structure.

MCR6 involvement in Glucose Dependent Activation

Recent evidence shows that MondoA and ChREBP activation is dependent upon glucose phosphorylation by hexokinase, which metabolizes glucose to form glucose-6-phosphate (G6P) (Ma, Sham et al. 2007; Stoltzman, Peterson et al. 2008; Peterson, Stoltzman et al. 2010). Induction of 2-deoxyglucose (2-DG), which is a glucose analog that can be

phosphorylated but not further metabolized, promotes MondoA nuclear accumulation, increases promoter occupancy and recruits histone H3 acetyltransferase thereby activating gene transcription (Li, Chen et al. 2010). Similarly, 2-DG dose dependently increased the transactivation ability of Gal4-ChREBP, while hexokinase inhibitor d-mannoheptulose and glycolytic enzymes PFK1 and PFK2 decreased ChREBP activity (Peterson, Stoltzman et al. 2010). This suggests that MondoA and ChREBP activation is directly invoked by glucose phosphorylation. As such, MondoA and ChREBP regulation is expected to occur through a G6P mediated signaling cascade, direct binding of G6P to an allosteric mechanism, or both.

To investigate the presence of an allosteric G6P binding region within MondoA and ChREBP, we first examined the binding region of known G6P interactors (Figure 9), i.e. glucokinase (GK), hexokinase (HKI-III), G6P phosphatase (G6Pase), phosphoglucose mutase (PGM), glucose phosphate isomerase (GPI), G6P dehydrogenase (G6PDH), and glutamine:fructose-6-phosphate amidotransferase (human: Gfat1, *E.coli*: Glms) (see Methods). Since glucose is essential among prokaryotes and eukaryotes, the enzymes and binding regions involved in glucose metabolism are highly conserved. Interestingly, the G6P binding region is similar among GK, GPI, and Gfat1, with serine and threonine residues forming hydrogen bonds with the 6-phosphate molecule (Figure 9). Moreover, the phosphate recognizing residues of GPI and Gfat1 are in close proximity in the linear sequence, forming an Sx[ST]xxT motif, where x indicates a residue not involved in 6-phosphate recognition. This is distinct from G6PDH and PGM, which have HYxxK and SKN motifs, respectively.

In support of G6P binding to Mondo proteins, the highly conserved *MCR6* region contains an SxTxx[TS] motif similar to GPI and Gfat1. MondoA consists of residues 281-**S**D**T**LF**S**-287, while ChREBP contains a 253-**S**D**T**LF**T**-258 motif. This putative G6P recognition motif is also preserved in Mondo sequences, where serine and threonine can interchangeably form hydrogen bonds with the 6-phosphate molecule. Although this short motif has low specificity and is likely to occur in several sequences, the strict conservation among animals is evidence for its functional importance among Mondo proteins.

The *MCR6* region also shows similarity to the nine amino acid transactivation domain (9aa TAD) signature that is recognized by coactivators TAF9, MED15, CBP, and p300 (Li,

Chen et al. 2010). Since *MCR6* is within the *GRACE* region responsible for transactivation, this motif may contribute to the recruitment of coactivators such as CBP/p300, which are known to interact with ChREBP (Piskacek, Gregor et al. 2007). Although individual sequences displayed multiple hits using the 9aa TAD regular expression (see Methods), the only concurrence was in *MCR6* where we observed two overlapping 9aa TAD motifs. ChREBP was restricted to motif 1 (ChREBP:250-SDIS**DTLF**T-258), while MondoA and Mondo sequences also matched motif 2 (MondoA:283-**DTLF**STLSS-291); conserved sites within the overlapping regions are in bold and underlined. Of the 34 sequences in our dataset containing *MCR6*, nineteen contained both motif 1 and 2, five only had motif 2, eight only had motif 1, trematode *Schistosoma mansoni* matched an intermediate sequence, and sea anenome *Nematostella vectensis* matched neither. Since there was no clear preference for either motif, the supposed TAD region may extend to include residues from both.

LID and GRACE regions have intramolecular contacts in N-terminal Predicted Structure

To better understand how *MCRI-V* switches between repressive and activating functions in different glucose conditions, we predicted the protein structure for MondoA and ChREBP N-terminal sequences. Since *MCRI-V* are unique to Mondo proteins, structural prediction by homology and threading relied heavily on shared secondary structure alignments (see Methods).

From the sequence and secondary structure predictions of 3D-Jury, the N-terminus of MondoA was most similar to Estrone Sulfatase (ES, PDB ID: 1p49) (Figure 10). ES is responsible for maintaining high levels of estrogen in breast tumor cells and is anchored to the membrane of the endoplasmic reticulum (ER) by two protruding alpha helices (Burke, Collier et al. 2009). MondoA also showed a likeness to similar sulfatase structures (PDB ID: 1auk, 1fsu) that interact with the ER membrane, but lack these alpha helices (Bond, Clements et al. 1997; Lukatela, Krauss et al. 1998). As expected, the N-terminus of ChREBP also shows structural similarity to 1p49 and resembles the MondoA conformation (Figure 11a).

The putative resemblance between ES and Mondo protein structures is compatible with the accessibility of their known domains, although their respective functions differ. In contrast to the transmembrane domain of ES, the protruding alpha helices in MondoA and

ChREBP correspond to *MCRII* and its CRM1 dependent NES in the predicted structure (Figure 10, orange). This is concordant with the CRM1-SNUPN structure, where the NES of SNUPN forms an extended amphipathic α-helix that protrudes away from the rest of the molecule and binds a hydrophobic groove in CRM1 (Hernandez-Guzman, Higashiyama et al. 2003). The exposure of *MCRIII* (Figure 10, yellow) also allows for its alpha helix to interact with known binding partner 14-3-3. The orientation of *MCRIII* and *MCRIV* (Figure 10, green) α-helices closely position S140 and S196 in ChREBP, so they are both situated near *MCRV* (Figure 10, purple). Dephosphorylation of S196 is implicated in increased nuclear accumulation, while S196D/S140D mutants have higher affinity for 14-3-3 (Dong, Biswas et al. 2009).

The placement of *MCRV* near the ends of *MCRI* (Figure 10, red), *MCRIII*, and *MCRIV* allows for interaction among these domains and corresponds to the proposed linkage between *LID* and *GRACE* regions mediated by multiple contacts with *MCRV* (Li, Chang et al. 2006; Sakiyama, Wynn et al. 2008). *MCR6* (Figure 10, blue) is adjacent to *MCRIV* and may also have a binding interface. Considering the potential role of *MCR6* in G6P binding and transactivation, this interaction may affect the glucose response, as seen for proteins with *MCRIV* deletions that lack glucose dependent regulation (Davies, O'callaghan et al. 2010). Viewing the predicted structure from the top (Figure 11), it is easy to see how the *LID* can contact and possibly release from the *GRACE* region to conditionally block the binding of coactivators and regulate the transactivation of target genes.

Discussion

Conservation in sequence, glucose response, and protein interactions for MondoA and ChREBP proteins indicate they are mechanistically similar. Based on the elevated JS conservation scores and persistence of secondary structures across sequences, the distal regions of Mondo proteins are likely to exhibit similar structure and function. The presence of *MCRI-V*, *MCR6*, *bHLHZ*, and *WMC/DCD* regions in diverse organisms dates the origin of these regions to as early as the divergence of cnidarians around 600 million years ago (Li, Chen et al. 2008). Moreover, conservation throughout Mondo proteins suggests the glucose responsive transactivation observed in MondoA and ChREBP has been preserved throughout

animal evolution. Similar to the explanation for the emergence of energy homeostasis in bilaterians (Ryan, Burton et al. 2006), cnidarians also possess muscular, nerve, and gastroderm or "stomach" cells, which contribute to the formation of an internal environment and rise of signaling factors important for homeostatic regulation, e.g. Mondo proteins and nuclear receptors.

Mondo proteins have cell type specific nuclear accumulation

Since MondoA and ChREBP are not known to have cytoplasmic activity, nuclear localization of these transcription factors is necessary for their function. Several cell lines have been used for this assessment, including glucose responsive rat hepatocytes, 832/13 insulinoma cells derived from the INS-1 pancreatic line, and L6 myoblasts, as well as COS-7 and HEK293 kidney cells and NIH3T3 fibroblasts that are not glucose responsive. Changes within these cellular environments are likely to affect MondoA and ChREBP glucose dependent functions including their subcellular localization and transactivation capabilities. For example, expression of ChREBP in rat hepatocytes localizes to the cytoplasm in low glucose conditions yet is mainly nuclear in high glucose (Kawaguchi, Takenoshita et al. 2001; Li, Chang et al. 2006). Similarly MondoA is predominantly cytoplasmic in low glucose, yet accumulates in the nucleus in high glucose in myoblasts and epithelial cells (Sakiyama, Wynn et al. 2008; Kaadige, Looper et al. 2009; Peterson, Stoltzman et al. 2010). However, the expression of ChREBP remains highly cytoplasmic in both low and high glucose conditions in INS-1 and 832/13 cell lines (Davies, O'Callaghan et al. 2008; Stoltzman, Peterson et al. 2008; Tsatsos, Davies et al. 2008). The absence and minimal amount of nuclear ChREBP in pancreatic cells under low and high glucose conditions, respectively, suggests an increased export or decreased import system of ChREBP compared to other cell lines.

MondoA is transported to the OMM

Although both MondoA and ChREBP are largely cytoplasmic, MondoA localizes specifically to the outer mitochondrial membrane (OMM) (Li, Chang et al. 2006). Of proteins within the mitochondria, 99% are transcribed by nuclear genes and actively transported to the mitochondria, exhibiting an established system of intracellular transport

(Sans, Satterwhite et al. 2006). Within the cytosol, Heat shock proteins Hsp70 and Hsp90 were among the first chaperone proteins found to facilitate protein transport to the mitochondria (Chirico, Waters et al. 1988; Deshaies, Koch et al. 1988; Endo and Yamano 2010). Mitochondria import stimulating factor (MSF) was also identified as a mitochondrial chaperone and is a member of the 14-3-3 protein family (Murakami, Pain et al. 1988). Chaperone proteins transport cargo proteins to the mitochondria that contain a presequence located in the distal N-terminus. Generally, mitochondrial surface proteins cleave this preprotein sequence, which allows the mature protein to enter through the mitochondrial membrane. However, some OMM proteins have a distal N-terminal, preprotein sequence that is not cleaved. In these few cases, this sequence is used for mitochondrial targeting, but not cleavage or import (Schleiff 2000).

MondoA, but not ChREBP or Mondo proteins, are predicted to contain mitochondrial targeting peptides within the first 42 residues, as specified by the program TargetP (Chacinska, Koehler et al. 2009). MondoA is not known to enter the mitochondria (Emanuelsson, Brunak et al. 2007) or predicted to contain a transmembrane region that inserts into the OMM. Hence the N-terminus sequence of MondoA is likely to induce mitochondrial transport via 14-3-3, where it interacts with receptors located on the OMM.

MondoA and ChREBP actively shuttle between the cytoplasm and nucleus

Since MondoA:Mlx and ChREBP:Mlx heterodimers actively shuttle between the nucleus and cytoplasm, increased nuclear accumulation in response to glucose is not simply the result of nuclear targeting. Blocking the *MCRII* NES in either MondoA (M133A, F130A, MondoAΔ125-137) or ChREBP (ChREBPΔ86-95,ChREBPΔ72-99, L86A/L93A, L89A, F90A) results in nuclear accumulation in either low or high glucose conditions (Sans, Satterwhite et al. 2006; Li, Chen et al. 2008; Sakiyama, Wynn et al. 2008; Fukasawa, Ge et al. 2010; Peterson, Stoltzman et al. 2010). Likewise, altering the *MCRIV* NLS in ChREBP (ChREBPΔ158-173, ChREBPΔ158-173, ChREBPΔ168-190) results in cytoplasmic retention (Davies, O'Callaghan et al. 2008; Sakiyama, Wynn et al. 2008; Fukasawa, Ge et al. 2010). However, MondoA triple mutant Y211A/W212A/K213A, which overlaps the latter portion of the bipartite NLS, results in nuclear localization in low and high glucose in L6 myoblasts

(Kawaguchi, Takenoshita et al. 2001). To complicate matters, all the *MCR*s affect the subcellular localization of ChREBP and MondoA.

Evidence for a CRS in MCRIV

Truncation mutants help annotate the function of *MCR*s and their influence on MondoA and ChREBP nuclear accumulation. C-terminal sequences, optionally including *MCRV* and *MCR6*, result in nuclear accumulation for both MondoA (Peterson, Stoltzman et al. 2010) and ChREBP (Eilers, Sundwall et al. 2002; Li, Chang et al. 2006). However, the addition of residues 224-273 in MondoA resulted in a cytoplasmic shift with most cells having equal nuclear and cytoplasmic amounts, while a MondoA mutant containing the full *MCRIV* region (MondoA:182-919) slightly reversed this effect with most cells being nuclear (Sakiyama, Wynn et al. 2008). This suggests that MondoA *MCRIV* has opposing roles in nuclear localization.

The bipartite NLS in ChREBP *MCRIV* is only partially conserved in some MondoA sequences, due to a single arginine to serine mutation, MondoA:S213, arising prior to the divergence of canines. Interestingly, the basic residues within the first portion of the NLS are conserved in MondoA, but variable in non-vertebrates, suggesting that the NLS may be weak or dispensable. As such, fusing *MCRIV* of MondoA to a heterologous NLS resulted in complete cytoplasmic localization (Eilers, Sundwall et al. 2002). Together this data implies MondoA:224-273 contains a strong CRS that is not dependent upon 14-3-3 binding. This region is similarly conserved among Mondo proteins, with sequence signature VxxE**Y**[KH]K**WR**x[FY][FY][KR], where x represents a variable site and bold letters are invariable among Mondo sequences. Due to this conservation, it is possible that ChREBP and Mondo proteins also contain a CRS within *MCRIV*.

Dephosphorylation of S196, directly downstream of *MCRIV*, results in the nuclear accumulation of ChREBP in low and high glucose. This has previously been linked to 14-3-3 disassociation and exposure of the NLS. However, Merla et al (2006) showed that 14-3-3, ChREBP, and the protein NIF3L1 form a complex within the cytoplasm and that NIF3L1 remains cytoplasmic while ChREBP is transported to the nucleus in COS-7 cells. NIF3L1 is orthologous to yeast Ngg1-interacting factor 3 (NIF3) and has been shown to interact with

Trip15/CSN of the COP9 signalosome. COP9 is a transcriptional repressor (Merla, Howald et al. 2004) and acts as a docking site for complex-mediated phosphorylation. It would be interesting to determine if NIF3L1 binds to MondoA or ChREBP *MCRIV* sequence signature in the cytoplasm and is responsible for hindering nuclear localization. Since S196 phosphorylation status affects ChREBP nuclear localization, it would also be noteworthy to identify the signaling mechanism for MondoA and Mondo protein subcellular localization. One candidate is putative phosphorylation site 147-[TS]P-148 between *MCRIII* and *MCRIV* (ChREBP numbering), which is found in almost all Mondo proteins and phosphorylated in high glucose for ChREBP triple mutant S196A/S626A/T666A (Akiyama, Fujisawa et al. 2003).

MCR6 involvement in G6P recognition and transactivation

Initial models of Mondo and Mlx function were solely dependent upon the subcellular localization of these proteins. Since ChREBP, MondoA, and Mlx are largely cytoplasmic, it was predicted that nuclear transport would be sufficient for the transactivation of their gene targets. However, multiple experiments have shown that trapping ChREBP:Mlx or MondoA:Mlx in the nucleus, mutating the NES, or altering the phosphorylation of particular residues does not result in constitutive activation of reporter constructs (Davies, O'Callaghan et al. 2008; Tsatsos, Davies et al. 2008; Peterson, Stoltzman et al. 2010).

Recently, MondoA nuclear accumulation has been attributed to both increased nuclear import, increased promoter occupancy, and decreased nuclear export in response to glucose derivative 2-DG (Sakiyama, Wynn et al. 2008). ChREBP transactivational ability is also correlated to G6P abundance (Peterson, Stoltzman et al. 2010), indicating that MondoA and ChREBP glucose response is directly mediated by G6P. Similarities in *MCR6* sequence with known G6P binding sites, and particularly the 6-phosphate molecule, strongly suggest that *MCR6* is an allosteric G6P binding region.

The putative function of *MCR6* in G6P allosteric activation and recruitment of coactivators is not mutually exclusive. Since MondoA and ChREBP have increased transactivation in response to G6P, its binding may trigger a conformational change that further exposes *MCR6* and facilitates cofactor interaction. The structure of GPI and Gfat1

proteins suggest that G6P binds within a largely hydrophilic pocket, while the 9aa TAD structure is variable and often disordered prior to forming an α-helix conformation upon cofactor binding (Li, Chen et al. 2010). The predicted structure of *MCR6* in MondoA and ChREBP displays an exposed pocket suitable for G6P binding as well as a flexible, coil region capable of making protein interactions.

Model of G6P mediated Mondo Glucose Response

Based on our structure predictions and published sequence annotations, we propose the following model for Mondo glucose responsive transactivation. First, Mlx and Mondo proteins readily form heterodimers within the cytoplasm, allowing Mlx:Mondo complexes to actively shuttle between the cytosol and nucleus. Second, *MCRV* interacts with the *LID* region, possibly through specific contacts with *MCRI*, *MCRIII*, and/or *MCRIV*, to block the transactivation region. Third, increased glucose and consequently G6P concentrations trigger signaling mechanisms that block the CRS in *MCRIV*. Fourth, G6P binding to *MCR6* causes an allosteric conformational change that "unlocks" *LID* and *MCRV* contacts, "pivots" *MCRII* so that it is buried, and "pins" *MCRI* in between the *LID* and *GRACE* so that Mondo remains in an open conformation. Finally, once in this open conformation, G6P may be released and cofactors such as CBP/p300 may bind to *MCR6* thereby activating Mondo proteins. In addition, Mondo and ChREBP proteins interact with nuclear receptors, such as HNF4α, through the *NRB*, which activate these cofactors and increase transactivational potential.

This model is in accordance with previous models based on protein manipulations. First, MondoA and ChREBP monomers are confined to the cytosol and MondoA requires Mlx dimerization prior to nuclear localization (Eilers, Sundwall et al. 2002; Piskacek, Gregor et al. 2007). MondoA and ChREBP dimers have also been observed to actively shuttle between the nucleus and cytosol in numerous cell types and can be sequestered in the nucleus by NES inhibitor leptomycin B (LMB), whereas blocking MondoA and Mlx dimerization results in purely cytoplasmic monomers. Phosphorylation sites are observed throughout ChREBP, except the *DCD*/*WMC* region, indicating Mlx dimerization is independent of phosphoregulation (Peterson, Stoltzman et al. 2010). Conservation of *DCD/WMC* residues

and similarity in both secondary and tertiary structure predictions implies monomer cytoplasmic retention and Mlx dimerization is consistent among Mondo proteins.

Second, the *LID* region is responsible for regulating the otherwise constitutively active *GRACE* region in ChREBP. Inverting the *LID* and *GRACE* regions results in constitutive activation, showing the structural organization of these regions is important for ChREBP regulation (Tsatsos, Davies et al. 2008). Combinatorial deletions in ChREBP show *MCRII* has minimal repressive effects, while *MCRI*, *MCRIII* and *MCRIV* decrease transactivation in the presence of *MCRV* (Li, Chang et al. 2006). *MCRV* does not repress transactivation in the absence of *MCRI-IV*, yet mutations to *MCRV* increase transactivation when the *LID* is present (Li, Chen et al. 2008). Individual deletions of *MCRI-IV* were unable to alleviate low glucose repression (Davies, O'Callaghan et al. 2008), suggesting *MCRV* represses transcription conditionally upon multiple contacts within the *LID* region. From the structural prediction, it is likely the *MCRV* contacts *MCRIII* and *MCRIV* near residues S140 and S196, respectively. These sites are known to affect the cytoplasmic localization of ChREBP as well as 14-3-3 binding, which is required for transactivation (Li, Chen et al. 2008; Davies, O'callaghan et al. 2010).

Third, it has been suggested that MondoA *MCRIV* contains a CRS (Davies, O'Callaghan et al. 2008) and truncation mutants indicate it is located within the latter half of *MCRIV*. This region is highly conserved among Mondo proteins and likely to have the same interaction properties. Increasing G6P abundance accelerates the rate of nuclear import for MondoA (Billin and Ayer 2006), while PP2A mediated dephosphorylation of S196 in ChREBP also results in increased nuclear abundance (Peterson, Stoltzman et al. 2010). Both of these affects are in accordance with G6P mediated relief of a CRS.

Fourth, it has been proposed that G6P allosterically affects the transactivation of MondoA and ChREBP (Kawaguchi, Takenoshita et al. 2001; Li, Chen et al. 2010; Peterson, Stoltzman et al. 2010). *MCR6* provides an appropriate interface for G6P binding and also contacts the *LID* domain, particularly with *MCRIV*. *MCRIV* is involved in general repression, where all mutants lacking this region have increased expression (Davies, O'callaghan et al. 2010). G6P binding may break hydrogen bonds of *MCRIV* with *MCR6* and

MCRV, thereby unlocking the repression of *GRACE* by *LID* and allowing these regions to separate. Additional deletion mutants show that *MCRI*, *MCRII*, and *MCRIII* are all necessary to overcome *MCRIV* repression and form an active complex.

Since glucose activated MondoA and ChREBP results in increased nuclear accumulation, we expect the NES to be overpowered in high glucose medium. 14-3-3 binding has previously been attributed to blocking the NES, although *MCRII* is also necessary for recruiting a histone H3 acetyltransferase (HAT) cofactor. Since the *LID* region is not independently sufficient for MondoA or ChREBP transactivation (Li, Chang et al. 2006), *MCRII* recruitment of a HAT cofactor must be a secondary effect. Based on the predicted N-terminus structure, it is plausible that *MCRII* pivots to make necessary contacts outside of the *LID* domain to help fix the separation between *LID* and *GRACE*.

MCRI is also required for glucose transactivation, but is not sufficient for full transactivation (Li, Chang et al. 2006). Hence *MCRI* may also form intrastructural contacts necessary for alleviating *LID* repression or interacting with activating cofactors. The position of *MCRI* in the interior of the predicted protein suggests it may act as a pin to wedge the *LID* and *GRACE* regions apart. Phosphorylation of S56 adjacent to *MCRI* increases ChREBP transactivational potential (Li, Chen et al. 2008), possibly by facilitating this conformational change.

MCRIII contains two essential regions. 14-3-3 and its binding region in *MCRIII* are required for ChREBP transactivation as is ChREBP:100-115 that is not necessary for 14-3-3 interaction. 14-3-3 has been shown to bind ChREBP constitutively (Tsatsos, Davies et al. 2008), promote cytoplasmic retention, nuclear export, and transactivation. While the necessity of S140 phosphorylation for 14-3-3:ChREBP interaction is under contention (Li, Chen et al. 2008; Sakiyama, Wynn et al. 2008), it can affect the binding orientation as non-phosphorylated motifs may bind 14-3-3 in the opposite direction (Li, Chen et al. 2008). While S140 and S196 have been analyzed in ChREBP, phosphorylation of the highly conserved T147/P148 site may have a broader impact on MondoA and Mondo interactions.

Moreover, the Mondo conserved *MCRIII* sequence corresponding to ChREBP:100-115 may affect this phosphorylation status. According to the functional site prediction server

ELM (Ottmann, Yasmin et al. 2007), this region matches a MAPK kinase-docking motif. Kinase docking domains are typically located 50-100 residues upstream of the phosphorylation site and characterized by a cluster of positively charged residues preceding a ΦxΦ hydrophobic sequence (Sharrocks, Yang et al. 2000; Tanoue, Maeda et al. 2001; Gould, Diella et al. 2010). Conserved sequence 105-KWKxFKG**[LIV]**[KR]**L**-114 conforms to this motif, where positively charged residues are underlined and hydrophobic residues are in bold. Interestingly, W106 and F109 are invariant, and may contribute to the interaction interface. Moreover, a 103-[ST]P-104 (human ChREBP numbering) phosphorylation site directly precedes this motif in all Mondo sequences, but has not been identified as a phosphorylation site.

Finally, MondoA and ChREBP recruit cofactors to promote transcriptional activation. Since mutants lacking the N-terminus have exceptionally high transactivational ability, G6P may only be necessary for relieving *LID* repression from *GRACE*. Hence G6P may be released from *MCR6* in the active/open conformation, thereby permitting *MCR6* access to cofactors. MondoA was shown to recruit a histone H3 acetyltransferase (Raman, Chen et al. 2007), while ChREBP is known to interact with CBP/p300 (Peterson, Stoltzman et al. 2010), which has histone acetyltransferase (HAT) function. *MCR6* matches the 9aa TAD motif depicting the CBP/p300 interaction region. Since *MCR6* is within the *GRACE* region, which is sufficient for transactivation (Burke, Collier et al. 2009), and mutating *MCRV* increases the transactivation potential (Li, Chang et al. 2006), we deduce that *MCR6* acts as the TAD for MondoA and ChREBP.

ChREBP and non-vertebrate Mondo transactivation may additionally rely on the interaction with nuclear receptors. In humans, the nuclear receptor super family contains 49 members that act as ligand-regulated transcription factors (Davies, O'Callaghan et al. 2008). Interestingly, nuclear receptors are specific to metazoans, and not found in sponges although present in cnidarians (Yu and Reddy 2007). This agrees with our identification of Mondo proteins and the *NRB* motif.

Typically when a ligand binds to a nuclear receptor within the ligand-binding domain, conformational changes prompt corepressor disassociation, which allows the nuclear receptor

box of coactivators to associate with the receptor AF-2 region and recruit coactivators and associated proteins. This complex then binds to DNA recognition elements and facilitates RNA polymerase II transcription on gene targets. Notably, coactivators bind almost promiscuously to nuclear receptors to influence multiple signaling pathways (Escriva, Langlois et al. 1998).

HNF4α is an orphan nuclear receptor that forms a homodimer and has a novel mode of ligand-dependent (independent) transactivation. HNF4α plays an important role in organ development, nutrient transport, and diverse metabolic pathways (i.e. glucose, fatty acid, cholesterol, and amino acid) by mediating the transcription of key regulatory genes in liver, kidney, intestine, and pancreatic cells (Yu and Reddy 2007). Most prominently, known inherited mutations within HNF4α are the most direct and common monogenic causes of maturity onset diabetes of the young (MODY) due to impaired glucose-stimulated insulin secretion from pancreatic beta cells (Lu, Rha et al. 2008).

Excluding MondoA, an LxQLLT sequence matching the *NRB* motif was conserved within the central region among Mondo and ChREBP proteins. Tellingly, ChREBP, nuclear receptor HNF4α, and CBP/p300 form a complex necessary for full activation of lipogenic enzyme L-PK (Lu, Rha et al. 2008). The HNF4α and ChREBP binding domains are directly adjacent within the promoter of this gene, indicating they are also juxtaposed within the complex. Since most nuclear receptors depend upon interaction with a *NRB* for activation, ChREBP may be fulfilling this role.

Conclusion

MondoA and ChREBP are important glucose responsive genes involved in energy homeostasis. While ChREBP has evolved to have unique phosphoacceptor sites, the conservation of *MCRI-V*, *MCR6*, *bHLHZ*, and *DCD/WMC* domains indicates all Mondo proteins are regulated by common mechanisms. Although their structure is not known, we predict their regulation is largely governed by intramolecular contacts. We further postulate that binding of G6P causes an allosteric conformational change, which forms an open, active complex where the *LID* repression is released from *GRACE* and permits interaction with

coactivators such as CBP/p300.

Methods

Full-length Mondo sequences were obtained as described in (McFerrin and Atchley, in preparation). Sequence analysis was performed on the diverse sample of 46 sequences from 37 species spanning the animal kingdom (Figure 4). ClustalW, Dialign, and MAFFT were used to align the sequences and merged according to consensus regions and manual adjustment to construct a single, optimal alignment. Mondo Conserved Regions were specified as in (Burke, Collier et al. 2009) and depicted by weblogos (Billin and Ayer 2006).

Sequence Conservation

Both the Jenson-Shannon Divergence (JS) score and entropy values were used to determine sequence conservation. From a multiple sequence alignment, the JS heuristic employs window-based extension that considers the conservation of sequentially neighboring sites and quantifies each score based on a weighted distribution of amino acids (Crooks, Hon et al. 2004). Hence the mutual information based JS score rates the conservation of each site by incorporating the autocorrelation of adjacent sites, where highly conserved sites have JS scores close to one and variable positions close to zero.

Entropy values were computed by the FastaEntropy program written by Andrew Fernandez (Capra and Singh 2007). Entropy is a statistical measure of the amount of information or variation and, when applied to sequence alignments, can depict the conservation of sites, with lower entropy values signifying increased conservation (Atchley, Zhao et al. 2005). Traditionally protein entropy is calculated by the Shannon Entropy equation based on the proportion of the 20 amino acids at each site. However, this method does not account for shared physiochemical properties among amino acids. To account for this, we also used a functional group entropy measure developed by (Shannon 1951) that is based on eight distinct categories of amino acids grouped according to physiochemical similarities. This method accentuates sites that are functionally constrained yet variable, e.g. conservation of I, V, L, M hydrophobic residues.

Site conservation is also highly correlated with structural and functional importance. To estimate and project the contribution of conserved sites on protein structures, we used the

Consurf program available at http://consurf.tau.ac.il/ (Atchley, Terhalle et al. 1999). Consurf predicts functionally important regions in a given protein structure by estimating the phylogenetic relationship of homologs with similar known tertiary structure and ranking the evolutionary rate at each site (Landau, Mayrose et al. 2005). Within this scheme, nine indicates site conservation and zero site variability.

Identification of Functional Domains and Motifs

The presence of functional domains or motifs was determined by individually analyzing each sequence using multiple online tools. The presence of proline rich and glutamine rich regions was predicted by the Expasy program ScanProsite (Ashkenazy, Erez et al. 2010). Additional motifs, such as the MAPK kinase docking domain, were predicted using regular expression patterns by the Eukaryotic Linear Motif resource (ELM) (Gattiker, Gasteiger et al. 2002), while the 9aa TAD server was used to specifically evaluate putative CBP/p300 binding regions (Gould, Diella et al. 2010). The MAPK docking motif in ELM is characterized by the regular expression [KR]{0,2}[KR].{0,2}[KR].{2,4}[ILVM].[ILVF], while the 9aa TAD regular expression is [GSTDENQWYM]{KRHCGP}[FLIVMW]{KRHCGP}{CGP}{KRHCGP}[FLIVMW][FLIVAMW]{KRHCP}; residues within brackets '[]' are permitted and residues within braces '{}' are prohibited.

Characterizing the G6P recognition pocket

The structure of several G6P binding proteins has been crystallized, with specific attention to the G6P binding region. During glucose metabolism in mammals, glucokinase (GK) or hexokinase (HKI-III) converts glucose to G6P (Aleshin, Kirby et al. 2000; Kamata, Mitsuya et al. 2004; Piskacek, Gregor et al. 2007), which can be reversed by G6P phosphatase (G6Pase) in the liver. G6P can be further metabolized by phosphoglucose mutase (PGM) to promote glycogen storage (Mulichak, Wilson et al. 1998; Regni, Shackelford et al. 2006), glucose phosphate isomerase (GPI) to produce fructose-6-phosphate (F6P) and continue in the glycolytic pathway (Zhang, Dai et al. 2005), or G6P dehydrogenase (G6PDH) to enter the pentose shunt of glycolysis (Cosgrove, Gover et al. 2000; Graham Solomons, Zimmerly et al. 2004). Another enzyme, glutamine:fructose-6-

phosphate amidotransferase (human: Gfat1, *E.coli*: Glms), can interact with G6P and F6P to promote the production of glycolipids through the glucosamine pathway (Teplyakov, Obmolova et al. 1999; Teplyakov, Obmolova et al. 2001; Kotaka, Gover et al. 2005).

We compared the G6P interacting residues described in the literature for each of these proteins to identify common features for metabolite recognition.

Structural prediction of the DCD and N-terminal region of Mondo

Correctly predicting protein structures from amino acid sequences has been a goal within computational biology for the last several decades. The reliability of structure predictions often depends on the availability of homologous structure templates that allow for protein threading or homology modeling methods (Nakaishi, Bando et al. 2009). These methods use a database of known structures to select a template with local or global similarities in secondary structure that can be used to fit the query model.

Secondary structure predictions for human, mouse, rat and drosophila Mondo sequences were formed by NPS@, which builds a consensus based on DPM, DSC, GOR1, GOR3, HNNC, MLRC, PHD, Predator, and SOPM individual predictions (Kihara, Chen et al. 2009). Sequences exhibited similar secondary structure predictions with compatible alignments of alpha helices and beta sheets. We depict the secondary structure by the representative human ChREBP graphic (Figure 2) produced using Polyview (Combet, Blanchet et al. 2000).

A structure prediction for ChREBP *DCD* was previously determined by The Human Proteome Folding Project and deposited at the yeast resource center . The MCM score quantifies the quality of the predicted structure and SCOP superfamily match. With an MCM value of 0.828, we propose the predicted *DCD* structure is credible.

While using structure prediction programs is straightforward, each method can form diverse structures and evaluating their accuracy is difficult. The metaserver 3D-jury addresses this concern by aggregating and comparing multiple structure predictions from several servers and ranking them based on structural similarity to create a more accurate consensus prediction (Porollo and Meller 2007).

For determining the N-terminal structure, we used 3D-Jury on MondoA sequence 1-490 and ChREBP sequence 1-360 . The 3D-Jury metaserver compares and ranks structural predictions from sequence only (EsyPred3, FFAS03, GRDB, Pfam-basic, Pfram-metabasic) and threading methods (3D-PSSM, FUGUE, INUB, mGenThreader, SAM-T02, samt06), whereby structure predictions are evaluated by the fit of each model and ranked according to their similarity to other models. MondoA most closely matched the PDB structure (1p49A) of human estrone sulfatase using the INUB Hybrid Fold Recognition method with a Jscore of 29.67. The N-terminal protein structures were modeled by the program Modeller 9.1 (Ginalski, Elofsson et al. 2003) and images were produced by Chimera (Ginalski, Elofsson et al. 2003).

Figures

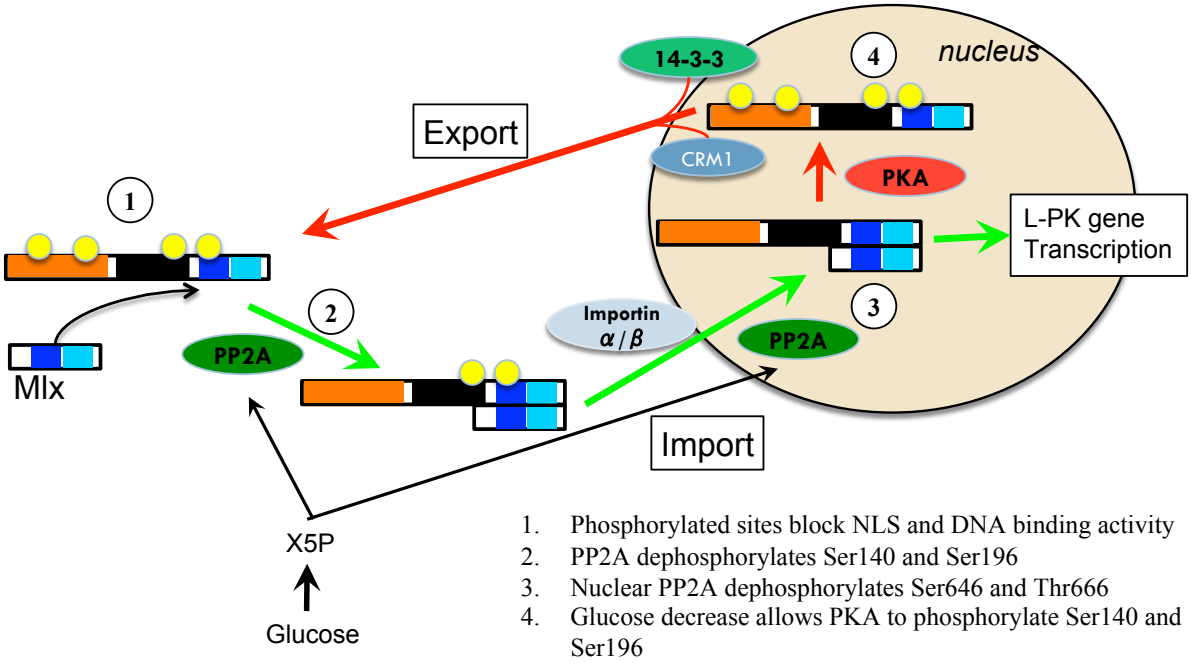


Figure 1: Phosphorylation Model depicting ChREBP response to glucose
Image adapted from (Sakiyama et al 2008). 1) In low glucose conditions sites S140/S196/S626/T666 are phosphorylated and block the NLS and DNA binding activity. 2) Upon glucose stimulation, Xu5P activates PP2A to dephosphorylate S140/S196 in the cytosol, unblocking the NLS, and allowing ChREBP to enter the nucleus. 3) Nuclear PP2A dephosphorylation of S626/T666 increases DNA binding. 4) Decreased glucose levels increase PKA activity to phosphorylate S140/S196 and shuttle ChREBP back to the cytoplasm.

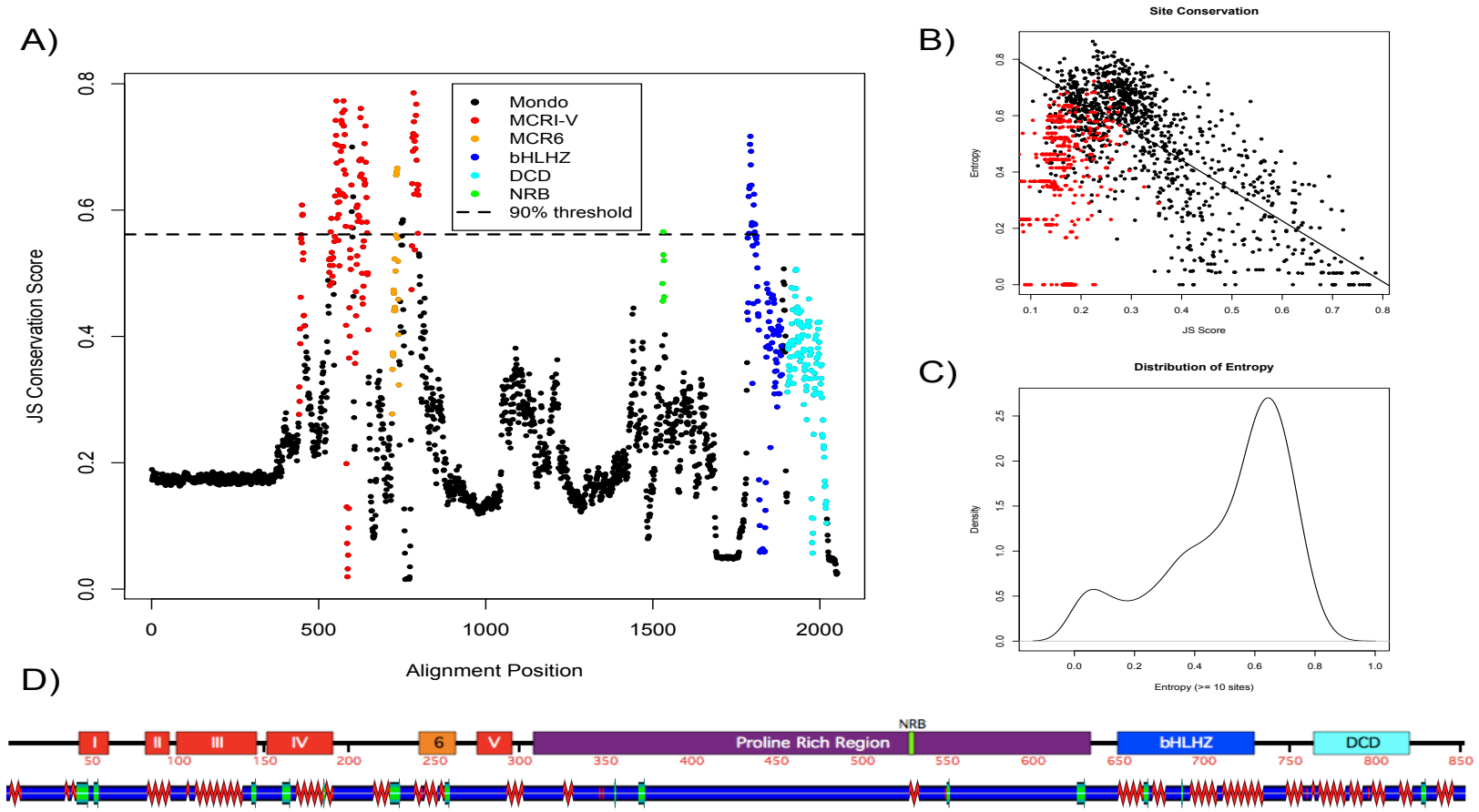


Figure 2: Mondo Sequence and Structure Conservation

A) JS Conservation Score. All Mondo sequences were used to construct an alignment of homologous sites. Black dots represent alignment columns, while sites within domains are colored. red: *MCRI-V*, orange: Myc box II-like (*MCR6*), green: nuclear receptor box, blue: basic helix-loop-helix-zipper, cyan: *DCD*. The dashed line sets the 90% threshold for JS scores for sites with at least 10 residues **B) JS and Entropy Comparison.** red: sites with less than 10 residues, black at least 10 residues, where linear regression was performed on the latter with intercept= 0.8745, slope= -1.0803, r^2=0.55467. **C) Entropy Distribution.** Bimodal distribution of entropy values for sites with at least 10 residues **D) Domains and Secondary Structure.** Consensus secondary structure for MondoB shown alongside ChREBP sequence and domains.

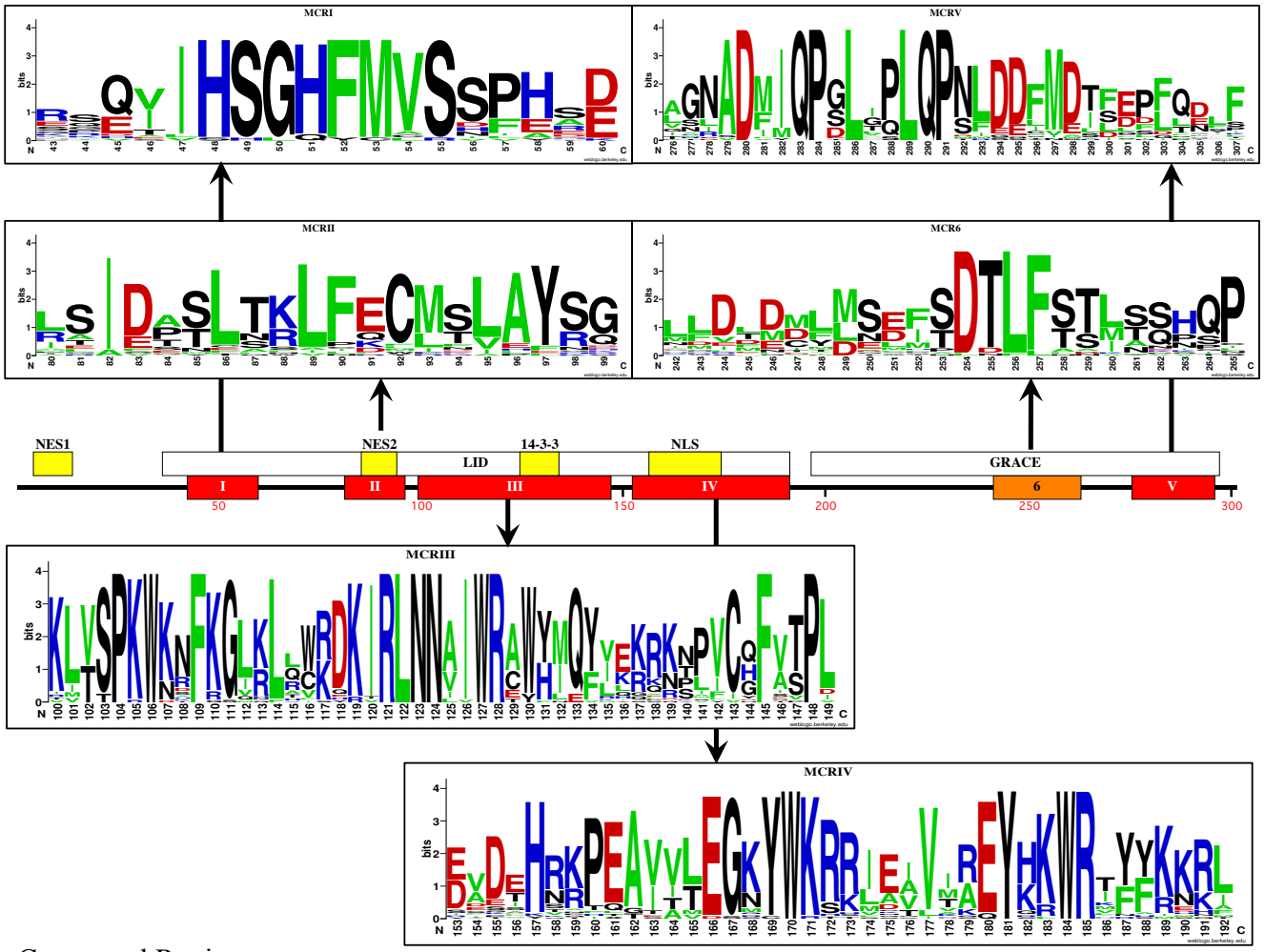


Figure 3: Mondo Conserved Regions

MondoA and ChREBP have five uniquely conserved regions, i.e. *MCRI-V*. These have been grouped into the *LID* and *GRACE* regions in ChREBP, and annotated for nuclear export signals (NES1, NES2), α-helix necessary for 14-3-3 binding, and a bipartite nuclear localization signal. These domains, along with newly identified *MCR6*, are highly conserved among Mondo sequences. Weblogos depicting the particularly conserved sites and regions were created using the full Mondo alignment. Amino acids are colored so basic (HKR) residues are blue, acidic (DE) are red, hydrophobic (AVLIFM) are green. Numbering is according to human ChREBP sequence.

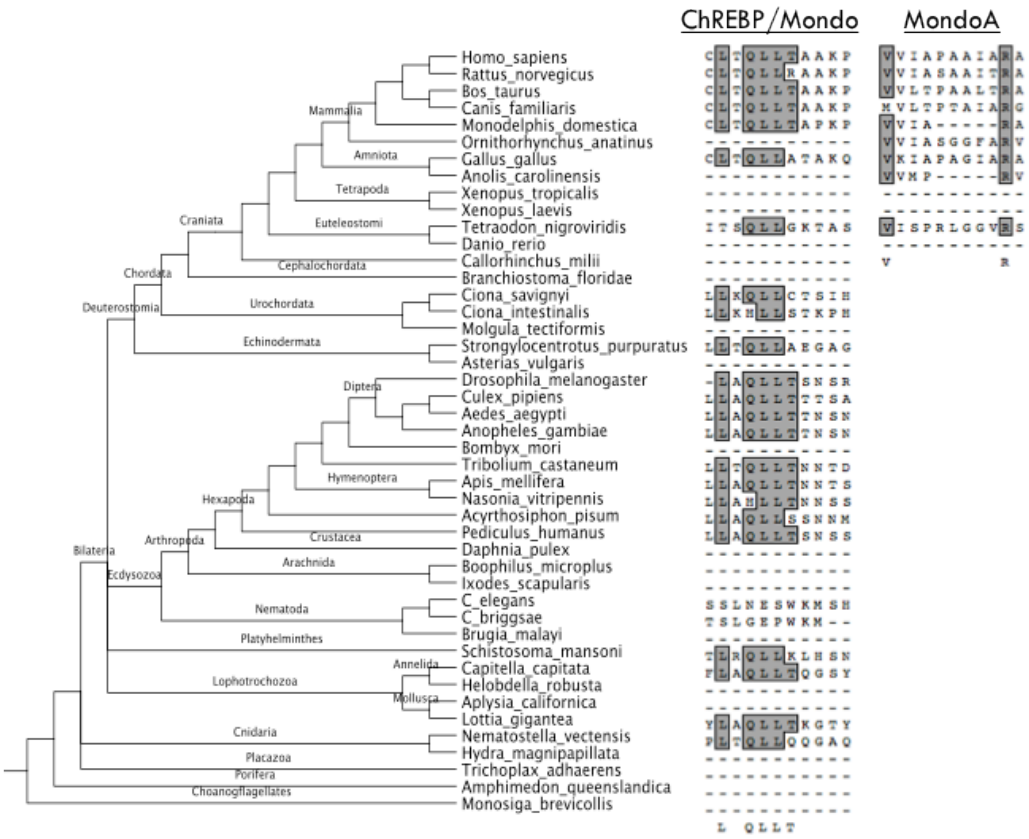


Figure 4: Nuclear Receptor Box Conservation

An LxQLLT is largely conserved among animals. Since we could not obtain the full sequence of all sampled species (shown in the species tree), many display alignment gaps, which do not necessarily indicate they lack the putative *NRB*. However, MondoA in vertebrates exhibits a divergent sequence and lacks the *NRB*.

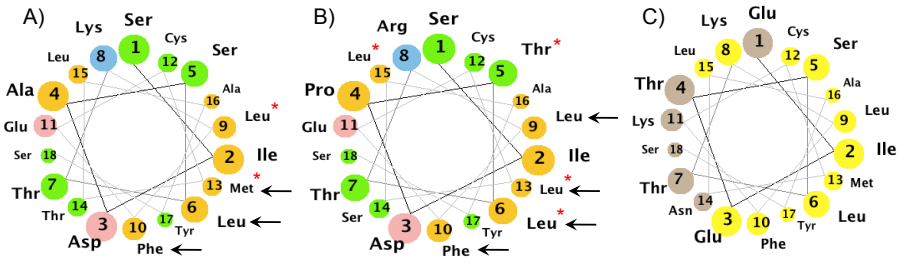


Figure 5: *MCRII* Helical Wheel

A) MondoA sites 121-138, **B)** ChREBP sites 81-98. Color scheme: blue-basic, pink-acidic, orange-nonpolar, green-polar, uncharged. Helical numbering is according to position within *MCRII* and represented by decreasing circle sizes. Black arrows point to sites indicated as essential for NES and red asterisks mark those necessary for glucose responsive transactivation. **C)** *Drosophila* sequence. Yellow circles have at least 75% chemical identity among all Mondo sequences.

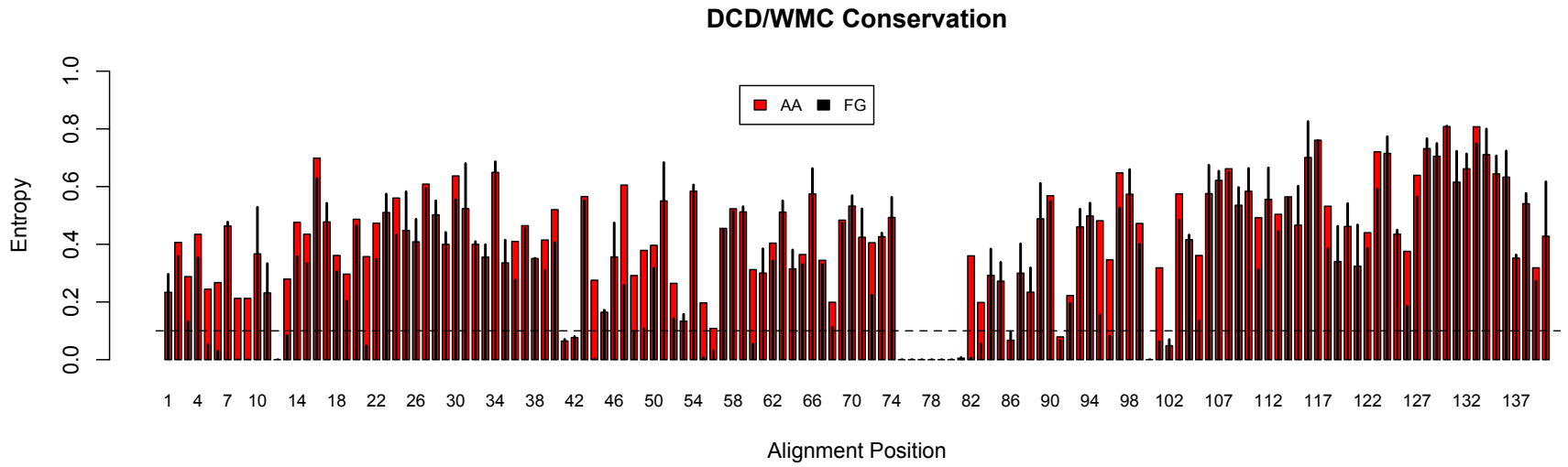


Figure 7: *DCD/WMC* Entropy

DCD/WMC region of all Mondo and Mlx sequences. Numbering corresponds to position in the alignment, shown in Figure 4. Low entropy values indicate site conservation for either a particular amino acid (red: AA) or physiochemical trait (black: FG), e.g. hydrophobic, although low entropy may also result from gaps in the alignment. The dotted line marks an arbitrary threshold of H=0.1 to indicate conserved sites.

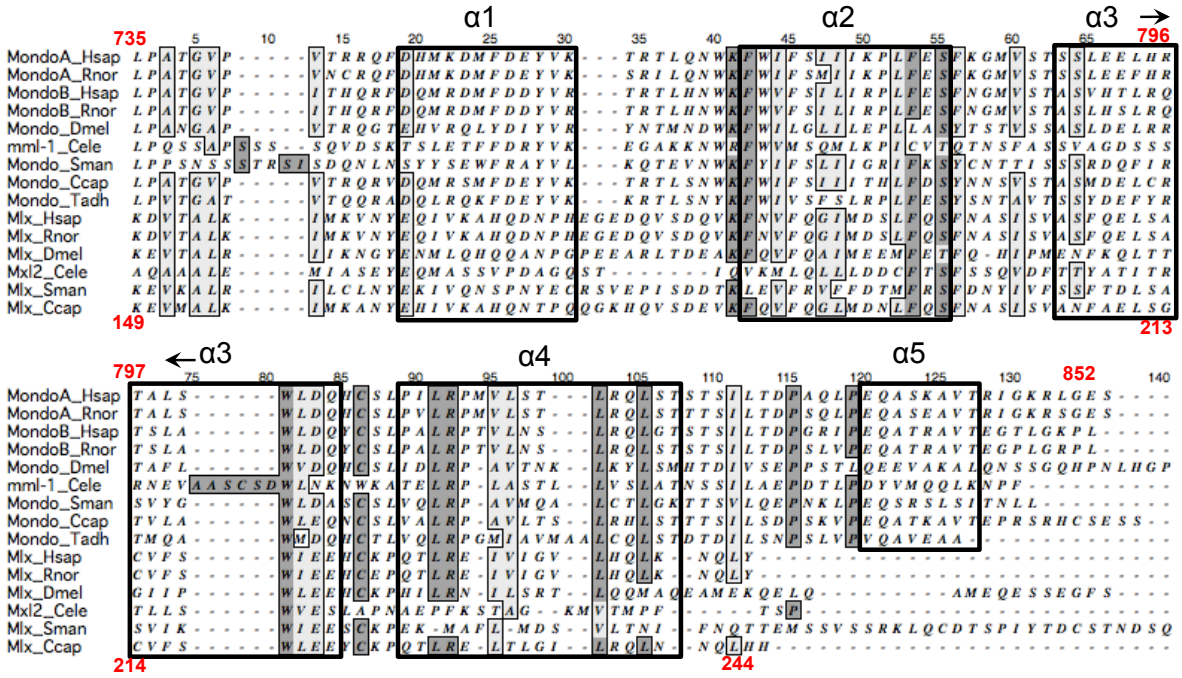


Figure 6: Mondo and Mlx *WMC*/*DCD* Alignment

DCD region of Mondo and Mlx sequences from *Homo sapiens* (Hsap), *Rattus norvegicus* (Rnor), *Drosophila melanogaster* (Dmel), *Caenorhabditis elegans* (Cele), *Capitella capitata* (Ccap), and *Trichoplax adhaerens* (Tadh). Red numbering on top corresponds to human MondoB (ChREBP) position, while the bottom represents the Mlx numbering. Sites with >75% identity or chemical similarity are shaded dark and light gray respectively, while the five (four) predicted alpha helices for MondoA and ChREBP (Mlx) are boxed.

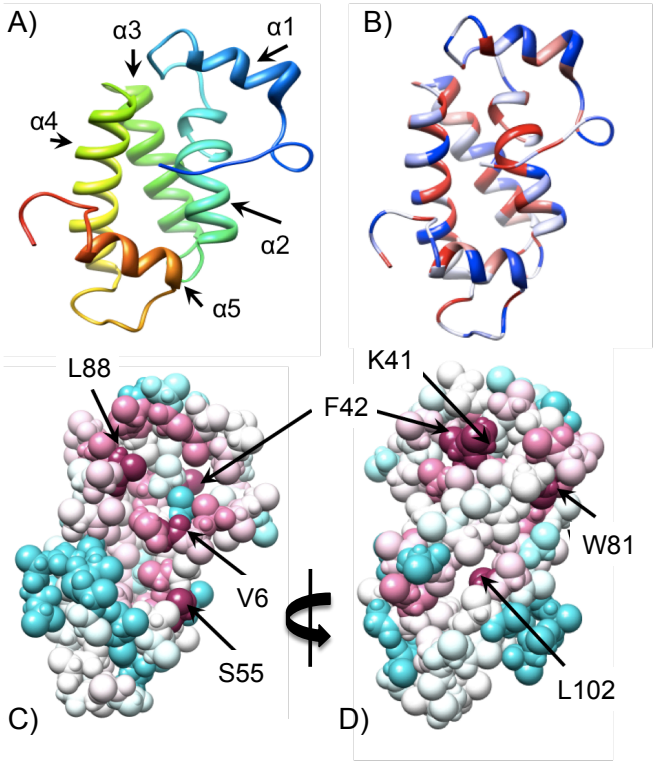


Figure 8: DCD/WMC Structure

Rosetta and Human Proteome Folding Project prediction for ChREBP *DCD/WMC* domain. **A)** A cluster of five alpha helices is predicted within the *DCD/WMC* region of ChREBP. **B)** Hydrophobic (red) residues line the interior groove of α2, α3 and α4, while hydrophilic (blue) residues coat the exterior. **C, D)**: Filled *DCD* structure in the same (left) and reversed (right) orientation as above, using Consurf conservation coloring (maroon: highly conserved, white: neutral, teal: variable). Highly conserved residues are labeled according to the human ChREBP sequence and the *WMC/DCD* alignment numbering.

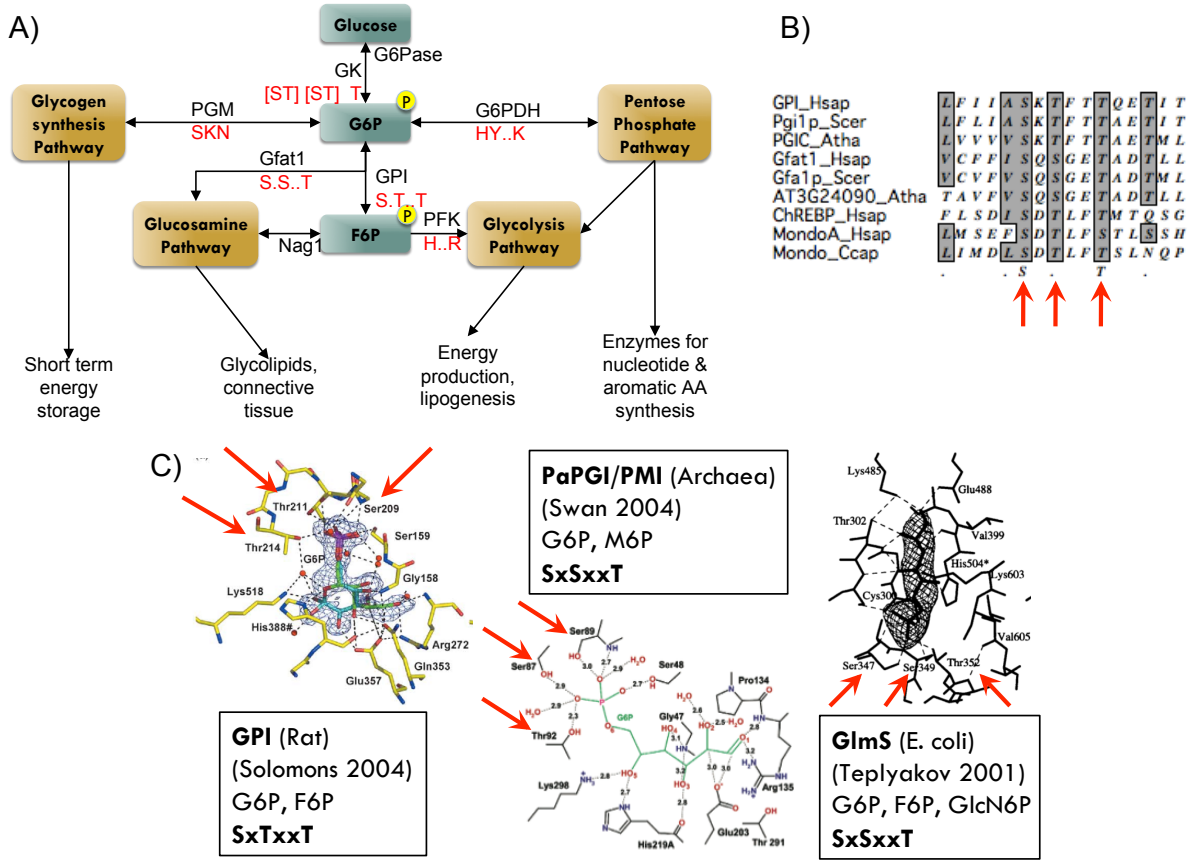


Figure 9: G6P Binding Region

A) Glucose metabolism pathways. Glucose is phosphorylated in the liver by GK, to form G6P. G6P can then enter the pentose phosphate pathway by interacting with G6PDH, the glycogen synthesis pathway by binding to PGM, or form F6P by GPI isomerization. Residues involved in these interactions are shown in red, with dots indicating nonbinding sites within a linear sequence and spaces denoting larger linear distances. **B)** G6P binding motif alignment of GPI, Gfat, and Mondo proteins from human, yeast, Arabidopsis, and Capitella (worm). **C)** G6P interacting protein structures. The structures for GPI in Rat, ancestral phospho-glucose/phospho-mannose protein in archaea, and GlmS in E. coli have been crystallized, with their indicated interacting metabolites and residues conforming to the G6P recognition motif.

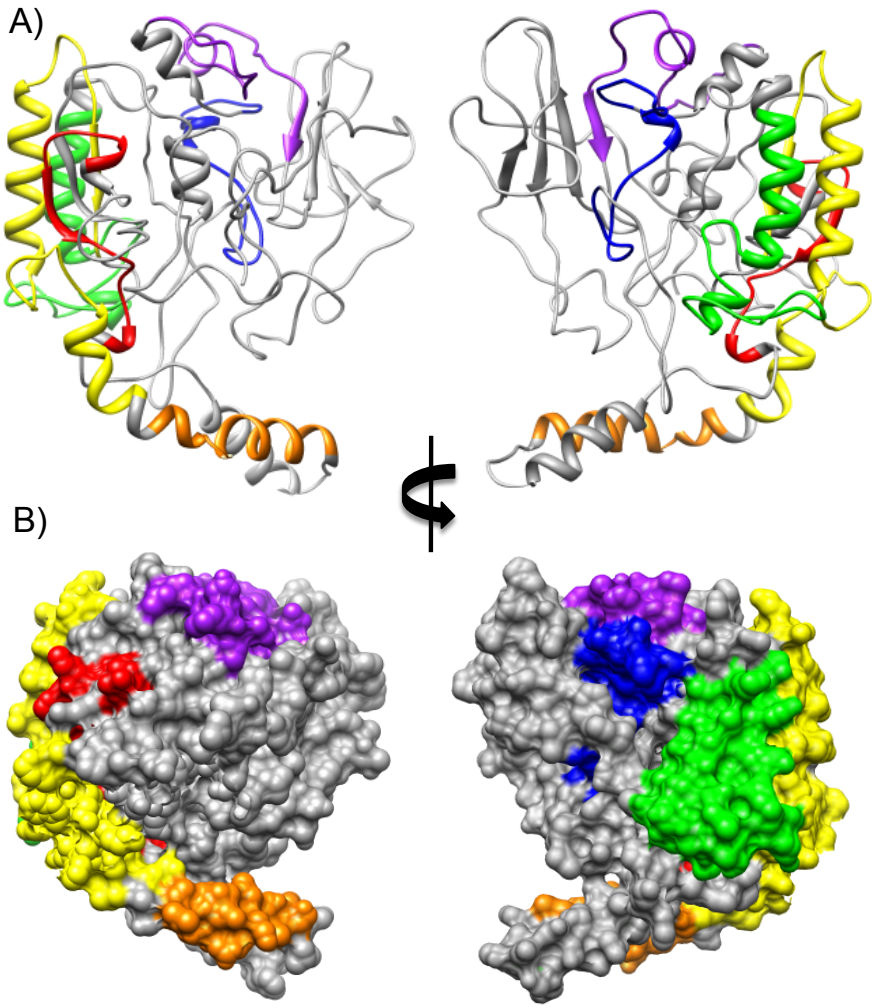


Figure 10: MondoA N-terminus structure

Predicted structure for MondoA:1-490. **A)** Ribbon structure. **B)** Filled structure. *MCRI* is red, *MCRII* is orange, *MCRIII* is yellow, *MCRIV* is green, *MCR6* is blue, and *MCRV* is purple. Left and right images are rotated 180 degrees.

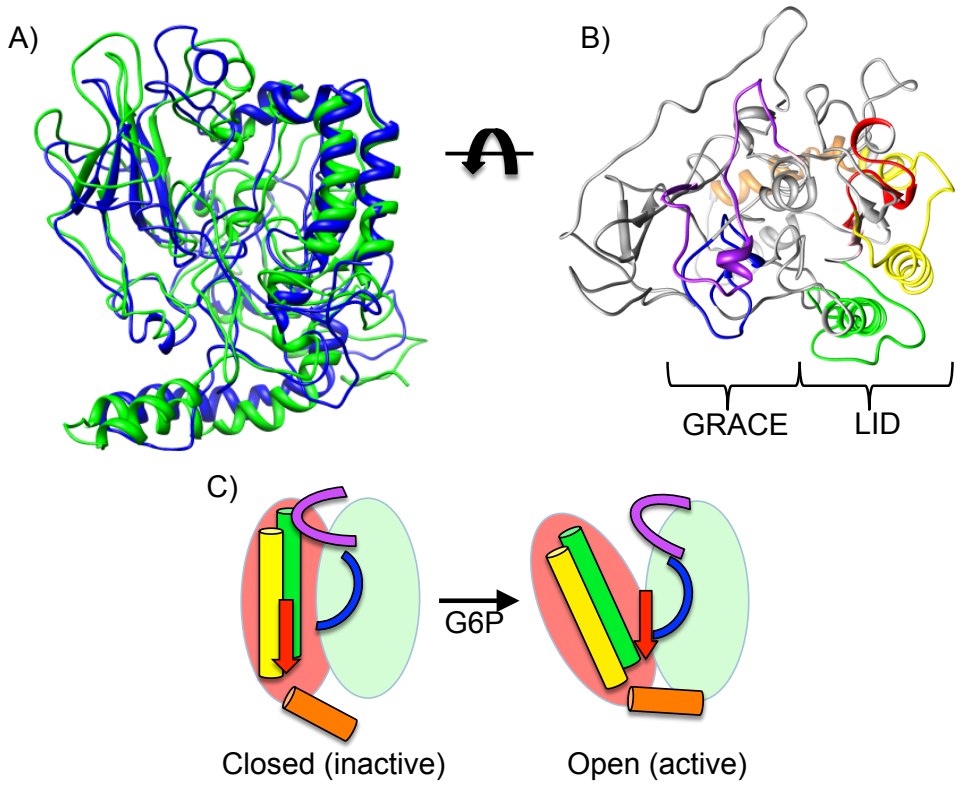


Figure 11: *LID* and *GRACE* interaction

A) MondoA (green) and ChREBP (blue) overlay of N-terminal predicted structure. **B)** Topical view of MondoA:1-490 ribbon structure. *MCRV* and *MCR6* are part of the *GRACE* region, while the *LID* includes *MCRI-IV*. **C)** Predicted allosteric affect of G6P binding to *MCR6*. *MCRII* and *MCRIII* release from *MCRV*, while *MCRI* and *MCRII* lock the "open" conformation to separate the *LID* and *GRACE* regions and support transactivation.

Tables

Table 1: Proline and Glutamine Rich Region.

		Proline	Glutamine	Neither	Missing
Vertebrates		16	0	1	3
Non-vertebrates		0	10	6	10
Length	$\bar{x}$	355.75	543	462.14	
	s	48.9	173.4	104.3	
PRR vs. GRR		p=0.0076			
PRR vs. Neither		p=0.0358			
GRR vs. Neither		p=0.2502			

Existence of Proline Rich and Glutamine Rich Regions in the proximal domain of Mondo sequences as predicted by ScanProsite. Neither indicates the central region is intact, yet ScanProsite did not identify a *PRR* or *GRR* region, while Missing denotes only a partial sequence not spanning the central region was obtained for that species. P-values are reported for a two-tailed t-test comparing the average length of the central region from sequences categorized by *PRR*, *GRR*, and Neither.

Table 2: Cell type specific nuclear accumulation of MondoA and ChREBP in response to glucose

	Cell line	**Low Glucose (5.5mM)**			**High Glucose (27.5mM)**			**Reference**
		C	**B**	**N**	**C**	**B**	**N**	
ChREBP	INS-1				72	23	5	(Eswar, Webb et al. 2006)
	832/13	94	6	0	78	18	2	(Pettersen, Goddard et al. 2004)
	Rat hepatocyte			~20			~45	(Tsatsos, Davies et al. 2008)
	Rat hepatocyte			~40			~80	(Davies, O'Callaghan et al. 2008)
	HEK293			~18			~48	(Sakiyama, Wynn et al. 2008)
MondoA	L6	~95	~5	0	~10	~25	~65	(Kawaguchi, Takenoshita et al. 2001)
	A549	~95	~5	0	~21	~24	~55	(Fukasawa, Ge et al. 2010)
	HA1ER	~95	-	~5	~15		~85	(Peterson, Stoltzman et al. 2010)

Values represent the (~approximate) percentage of cells with Mondo transcripts located in either the cytoplasm (C), nuclear (N), or both (B) for low and high glucose medium in the studies referenced.

References

Akiyama, H., N. Fujisawa, et al. (2003). "The role of transcriptional corepressor Nif3l1 in early stage of neural differentiation via cooperation with Trip15/CSN2." J Biol Chem **278**(12): 10752-62.

Aleshin, A. E., C. Kirby, et al. (2000). "Crystal structures of mutant monomeric hexokinase I reveal multiple ADP binding sites and conformational changes relevant to allosteric regulation." J Mol Biol **296**(4): 1001-15.

Ashkenazy, H., E. Erez, et al. (2010). "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids." Nucleic Acids Res **38**(Web Server): W529-W533.

Atchley, W. R., W. Terhalle, et al. (1999). "Positional dependence, cliques, and predictive motifs in the bHLH protein domain." J Mol Evol **48**(5): 501-16.

Atchley, W. R., J. Zhao, et al. (2005). "Solving the protein sequence metric problem." Proc Natl Acad Sci USA **102**(18): 6395-400.

Billin, A. N. and D. E. Ayer (2006). "The Mlx network: evidence for a parallel Max-like transcriptional network that regulates energy metabolism." Curr Top Microbiol Immunol **302**: 255-78.

Billin, A. N., A. L. Eilers, et al. (2000). "MondoA, a novel basic helix-loop-helix-leucine zipper transcriptional activator that constitutes a positive branch of a max-like network." Molecular and Cellular Biology **20**(23): 8845-54.

Burke, S. J., J. J. Collier, et al. (2009). "cAMP opposes the glucose-mediated induction of the L-PK gene by preventing the recruitment of a complex containing ChREBP, HNF4alpha, and CBP." FASEB J **23**(9): 2855-65.

Cairo, S., G. Merla, et al. (2001). "WBSCR14, a gene mapping to the Williams--Beuren syndrome deleted region, is a new member of the Mlx transcription factor network." Human Molecular Genetics **10**(6): 617-27.

Capra, J. A. and M. Singh (2007). "Predicting functionally important residues from sequence conservation." Bioinformatics **23**(15): 1875-82.

Chacinska, A., C. M. Koehler, et al. (2009). "Importing mitochondrial proteins: machineries and mechanisms." Cell **138**(4): 628-44.

Chirico, W. J., M. G. Waters, et al. (1988). "70K heat shock related proteins stimulate protein translocation into microsomes." Nature **332**(6167): 805-10.

Combet, C., C. Blanchet, et al. (2000). "NPS@: network protein sequence analysis." Trends Biochem Sci **25**(3): 147-50.

Cosgrove, M. S., S. Gover, et al. (2000). "An examination of the role of asp-177 in the His-Asp catalytic dyad of Leuconostoc mesenteroides glucose 6-phosphate dehydrogenase: X-ray structure and pH dependence of kinetic parameters of the D177N mutant enzyme." Biochemistry **39**(49): 15002-11.

Crooks, G. E., G. Hon, et al. (2004). "WebLogo: a sequence logo generator." Genome Research **14**(6): 1188-90.

Davies, M., B. O'callaghan, et al. (2010). "Activation and Repression of Glucose-stimulated ChREBP Requires the Concerted Action of Multiple Domains within the MondoA Conserved Region." AJP: Endocrinology and Metabolism: 38.

Davies, M. N., B. L. O'Callaghan, et al. (2008). "Glucose activates ChREBP by increasing its rate of nuclear entry and relieving repression of its transcriptional activity." J Biol Chem **283**(35): 24029-38.

Deshaies, R. J., B. D. Koch, et al. (1988). "A subfamily of stress proteins facilitates translocation of secretory and mitochondrial precursor polypeptides." Nature **332**(6167): 800-5.

Dong, X., A. Biswas, et al. (2009). "Structural basis for leucine-rich nuclear export signal recognition by CRM1." Nature **458**(7242): 1136-41.

Eilers, A. L., E. Sundwall, et al. (2002). "A novel heterodimerization domain, CRM1, and 14-3-3 control subcellular localization of the MondoA-Mlx heterocomplex." Molecular and Cellular Biology **22**(24): 8514-26.

Emanuelsson, O., S. Brunak, et al. (2007). "Locating proteins in the cell using TargetP, SignalP and related tools." Nat Protoc **2**(4): 953-71.

Endo, T. and K. Yamano (2010). "Transport of proteins across or into the mitochondrial outer membrane." Biochimica et Biophysica Acta (BBA) - Molecular Cell Research **1803**(6): 706-714.

Escriva, H., M. C. Langlois, et al. (1998). "Evolution and diversification of the nuclear receptor superfamily." Ann N Y Acad Sci **839**: 143-6.

Eswar, N., B. Webb, et al. (2006). "Comparative protein structure modeling using Modeller." Curr Protoc Bioinformatics **Chapter 5**: Unit 5.6.

Fukasawa, M., Q. Ge, et al. (2010). "Coordinate regulation/localization of the carbohydrate responsive binding protein (ChREBP) by two nuclear export signal sites: Discovery of a new leucine-rich nuclear export signal site." Biochem Biophys Res Commun **391**(2): 1166-1169.

Gattiker, A., E. Gasteiger, et al. (2002). "ScanProsite: a reference implementation of a PROSITE scanning tool." Appl Bioinformatics **1**(2): 107-8.

Ginalski, K., A. Elofsson, et al. (2003). "3D-Jury: a simple approach to improve protein structure predictions." Bioinformatics **19**(8): 1015-8.

Girard, J., P. Ferré, et al. (1997). "Mechanisms by which carbohydrates regulate expression of genes for glycolytic and lipogenic enzymes." Annu Rev Nutr **17**: 325-52.

Gould, C. M., F. Diella, et al. (2010). "ELM: the status of the 2010 eukaryotic linear motif resource." Nucleic Acids Research **38**(Database issue): D167-80.

Graham Solomons, J. T., E. M. Zimmerly, et al. (2004). "The crystal structure of mouse phosphoglucose isomerase at 1.6A resolution and its complex with glucose 6-phosphate reveals the catalytic mechanism of sugar ring opening." Journal of Molecular Biology **342**(3): 847-60.

Guo, L., A. Han, et al. (2007). "Crystal structure of a conserved N-terminal domain of histone deacetylase 4 reveals functional insights into glutamine-rich domains." Proc Natl Acad Sci USA **104**(11): 4297-302.

Hernandez-Guzman, F. G., T. Higashiyama, et al. (2003). "Structure of human estrone sulfatase suggests functional roles of membrane association." J Biol Chem **278**(25): 22989-97.

Iizuka, K., R. K. Bruick, et al. (2004). "Deficiency of carbohydrate response element-binding protein (ChREBP) reduces lipogenesis as well as glycolysis." Proc Natl Acad Sci USA **101**(19): 7281-6.

Kaadige, M. R., R. E. Looper, et al. (2009). "Glutamine-dependent anapleurosis dictates glucose uptake and cell growth by regulating MondoA transcriptional activity." Proc Natl Acad Sci USA **106**(35): 14878-83.

Kabashima, T., T. Kawaguchi, et al. (2003). "Xylulose 5-phosphate mediates glucose-induced lipogenesis by xylulose 5-phosphate-activated protein phosphatase in rat liver." Proc Natl Acad Sci USA **100**(9): 5107-12.

Kamata, K., M. Mitsuya, et al. (2004). "Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase." Structure **12**(3): 429-38.

Kawaguchi, T., M. Takenoshita, et al. (2001). "Glucose and cAMP regulate the L-type pyruvate kinase gene by phosphorylation/dephosphorylation of the carbohydrate response element binding protein." Proc Natl Acad Sci USA **98**(24): 13710-5.

Kay, B. K., M. P. Williamson, et al. (2000). "The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains." FASEB J **14**(2): 231-41.

Kihara, D., H. Chen, et al. (2009). "Quality assessment of protein structure models." Curr Protein Pept Sci **10**(3): 216-28.

Kotaka, M., S. Gover, et al. (2005). "Structural studies of glucose-6-phosphate and NADP+ binding to human glucose-6-phosphate dehydrogenase." Acta Crystallogr D Biol Crystallogr **61**(Pt 5): 495-504.

Kutay, U. and S. Güttinger (2005). "Leucine-rich nuclear-export signals: born to be weak." Trends Cell Biol **15**(3): 121-4.

Landau, M., I. Mayrose, et al. (2005). "ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures." Nucleic Acids Research **33**(Web Server): W299-W302.

Li, M., B. Chang, et al. (2006). "Glucose-dependent transcriptional regulation by an evolutionarily conserved glucose-sensing module." Diabetes **55**(5): 1179-89.

Li, M., W. Chen, et al. (2010). "Glucose-6-phosphate mediates activation of the carbohydrate responsive binding protein (ChREBP)." Biochem Biophys Res Commun: 6.

Li, M., W. Chen, et al. (2008). "Glucose-Mediated Transactivation of Carbohydrate Response Element-Binding Protein Requires Cooperative Actions from Mondo Conserved Regions and Essential Trans-Acting Factor 14-3-3." Mol Endocrinol **22**(7): 1658-1672.

Lu, P., G. Rha, et al. (2008). "Structural Basis of Natural Promoter Recognition by a Unique Nuclear Receptor, HNF4 : DIABETES GENE PRODUCT." Journal of Biological Chemistry **283**(48): 33685-33697.

Ma, L., L. N. Robinson, et al. (2006). "ChREBP*Mlx is the principal mediator of glucose-induced gene expression in the liver." J Biol Chem **281**(39): 28721-30.

Ma, L., Y. Y. Sham, et al. (2007). "A critical role for the loop region of the basic helix-loop-helix/leucine zipper protein Mlx in DNA binding and glucose-regulated transcription." Nucleic Acids Research **35**(1): 35-44.

Ma, L., N. G. Tsatsos, et al. (2005). "Direct role of ChREBP.Mlx in regulating hepatic glucose-responsive genes." J Biol Chem **280**(12): 12019-27.

Merla, G., C. Howald, et al. (2004). "The subcellular localization of the ChoRE-binding protein, encoded by the Williams-Beuren syndrome critical region gene 14, is regulated by 14-3-3." Human Molecular Genetics **13**(14): 1505-14.

Mulichak, A. M., J. E. Wilson, et al. (1998). "The structure of mammalian hexokinase-1." Nat Struct Biol **5**(7): 555-60.

Murakami, H., D. Pain, et al. (1988). "70-kD heat shock-related protein is one of at least two distinct cytosolic factors stimulating protein import into mitochondria." J Cell Biol **107**(6 Pt 1): 2051-7.

Nakaishi, Y., M. Bando, et al. (2009). "Structural analysis of human glutamine:fructose-6-phosphate amidotransferase, a key regulator in type 2 diabetes." FEBS Lett **583**(1): 163-7.

Nolte, R. T., G. B. Wisely, et al. (1998). "Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor-gamma." Nature **395**(6698): 137-43.

Odom, D. T., N. Zizlsperger, et al. (2004). "Control of pancreas and liver gene expression by HNF transcription factors." Science **303**(5662): 1378-81.

Ottmann, C., L. Yasmin, et al. (2007). "Phosphorylation-independent interaction between 14-3-3 and exoenzyme S: from structure to pathogenesis." EMBO J **26**(3): 902-13.

Peterson, C., C. Stoltzman, et al. (2010). "Glucose Controls Nuclear Accumulation, Promoter Binding, and Transcriptional Activity of the MondoA-Mlx Heterodimer." Molecular and Cellular Biology **30**(12): 2887-2895.

Petrova, N. V. and C. H. Wu (2006). "Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties." BMC Bioinformatics **7**: 312.

Pettersen, E. F., T. D. Goddard, et al. (2004). "UCSF Chimera--a visualization system for exploratory research and analysis." J Comput Chem **25**(13): 1605-12.

Peyrefitte, S., D. Kahn, et al. (2001). "New members of the Drosophila Myc transcription factor subfamily revealed by a genome-wide examination for basic helix-loop-helix genes." Mech Dev **104**(1-2): 99-104.

Pickett, C., K. Breen, et al. (2007). "A C. elegans Myc-like network cooperates with semaphorin and Wnt signaling pathways to control cell migration." Developmental Biology **310**(2): 226-239.

Piskacek, S., M. Gregor, et al. (2007). "Nine-amino-acid transactivation domain: establishment and prediction utilities." Genomics **89**(6): 756-68.

Porollo, A. and J. Meller (2007). "Versatile annotation and publication quality visualization of protein complexes using POLYVIEW-3D." BMC Bioinformatics **8**: 316.

Postic, C., R. Dentin, et al. (2007). "ChREBP, a transcriptional regulator of glucose and lipid metabolism." Annu Rev Nutr **27**: 179-92.

Raman, M., W. Chen, et al. (2007). "Differential regulation and properties of MAPKs." Oncogene **26**(22): 3100-12.

Regni, C., G. S. Shackelford, et al. (2006). "Complexes of the enzyme phosphomannomutase/phosphoglucomutase with a slow substrate and an inhibitor." Acta Crystallogr Sect F Struct Biol Cryst Commun **62**(Pt 8): 722-6.

Rohl, C. A., C. E. Strauss, et al. (2004). "Protein structure prediction using Rosetta." Meth Enzymol **383**: 66-93.

Rufo, C., M. Teran-Garcia, et al. (2001). "Involvement of a unique carbohydrate-responsive factor in the glucose regulation of rat liver fatty-acid synthase gene transcription." J Biol Chem **276**(24): 21969-75.

Ryan, J. F., P. M. Burton, et al. (2006). "The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, Nematostella vectensis." Genome Biol **7**(7): R64.

Sakiyama, H., R. M. Wynn, et al. (2008). "Regulation of nuclear import/export of carbohydrate response element-binding protein (ChREBP): interaction of an alpha-helix of ChREBP with the 14-3-3 proteins and regulation by phosphorylation." J Biol Chem **283**(36): 24899-908.

Sans, C., D. Satterwhite, et al. (2006). "MondoA-Mlx Heterodimers Are Candidate Sensors of Cellular Energy Status: Mitochondrial Localization and Direct Regulation of Glycolysis." Molecular and Cellular Biology **26**(13): 4863-4871.

Schleiff, E. (2000). "Signals and receptors--the translocation machinery on the mitochondrial surface." J Bioenerg Biomembr **32**(1): 55-66.

Shannon, C. E. (1951). "Prediction and entropy of printed English." The Bell System Technical Journal **30**: 50-64.

Sharrocks, A. D., S. H. Yang, et al. (2000). "Docking domains and substrate-specificity determination for MAP kinases." Trends in Biochemical Sciences **25**(9): 448-53.

Stoltzman, C. A., C. W. Peterson, et al. (2008). "Glucose sensing by MondoA:Mlx complexes: a role for hexokinases and direct regulation of thioredoxin-interacting protein expression." Proc Natl Acad Sci USA **105**(19): 6912-7.

Tanoue, T., R. Maeda, et al. (2001). "Identification of a docking groove on ERK and p38 MAP kinases that regulates the specificity of docking interactions." EMBO J **20**(3): 466-79.

Teplyakov, A., G. Obmolova, et al. (2001). "Channeling of ammonia in glucosamine-6-phosphate synthase." Journal of Molecular Biology **313**(5): 1093-102.

Teplyakov, A., G. Obmolova, et al. (1999). "The mechanism of sugar phosphate isomerization by glucosamine 6-phosphate synthase." Protein Sci **8**(3): 596-602.

Towle, H. C. (2005). "Glucose as a regulator of eukaryotic gene transcription." Trends Endocrinol Metab **16**(10): 489-94.

Tsatsos, N. G., M. N. Davies, et al. (2008). "Identification and function of phosphorylation in the glucose-regulated transcription factor ChREBP." Biochem J **411**(2): 261-70.

Tsatsos, N. G. and H. C. Towle (2006). "Glucose activation of ChREBP in hepatocytes occurs via a two-step mechanism." Biochem Biophys Res Commun **340**(2): 449-56.

Xu, J., B. Christian, et al. (2006). "Regulation of rat hepatic L-pyruvate kinase promoter composition and activity by glucose, n-3 polyunsaturated fatty acids, and peroxisome proliferator-activated receptor-alpha agonist." J Biol Chem **281**(27): 18351-62.

Yamashita, H., M. Takenoshita, et al. (2001). "A glucose-responsive transcription factor that regulates carbohydrate metabolism in the liver." Proc Natl Acad Sci USA **98**(16): 9116-21.

Yu, S. and J. K. Reddy (2007). "Transcription coactivators for peroxisome proliferator-activated receptors." Biochim Biophys Acta **1771**(8): 936-51.

Zhang, G., J. Dai, et al. (2005). "Catalytic cycling in beta-phosphoglucomutase: a kinetic and structural analysis." Biochemistry **44**(27): 9404-16.

Zhang, P., M. Metukuri, et al. (2010). "c-Myc Is Required for the ChREBP-Dependent Activation of Glucose-Responsive Genes." Mol Endocrinol: 13.

Chapter 5

Discussion and Concluding Remarks

The central dogma of molecular biology explains how information encoded in DNA can be transferred to RNA and utilized by proteins to drive organism physiology. Amazingly, sequences of DNA exist in plants, animals, and fungi with such high similarity that it is very unlikely they have separate origins. These conserved regions often define essential genes that are necessary for basic cellular function, e.g. RNA polymerases, tRNAs, and cyclins. The presence of multiple conserved genes encoding interacting proteins further indicates organisms have similar mechanisms and pathways for maintaining a stable cellular environment and reacting to external stimuli. These common features give a clue to both the origin and fundamental aspects of life.

Identifying commonalities among species, including sites, domains, proteins, and networks, can facilitate our understanding of basic life functions. Likewise, revealing aspects that are distinctly different can help identify what makes "us" unique. Within this dissertation, I focus on the conservation and variation of proteins to explore the affects of structural and functional selection. Our ability to accurately quantify and distinguish factors constraining protein variation ultimately controls our understanding of protein function and evolution.

Conserved Proteins show regional conservation

In chapter 2, I address the observed effect of selection on the linear amino acid sequence. Through genome-wide comparisons of several evolutionarily diverse species, we showed that dispersion of conserved sites within the protein sequence is directly correlated with the overall conservation of the protein. Hence, the conservation of structurally and functionally important sites within proteins is loosely segmented into slowly- and rapidly-evolving amino acids. This validates the basis of current *de novo* annotation methods that depend on regions of local sequence conservation to identify homologous sequences or functionally important motifs.

Alignment searches, such as BLAST, rely on regional constraint of protein sequences and can be used to identify protein domains. Protein domains are recognized by modular regions of conservation and can be coupled to form heterogeneous sets of functional proteins. Domain shuffling can reorganize the various combinations of protein domains, alter network configurations, and contribute to organism diversity. Gene duplication and alternative splicing can also add complexity and foster system robustness. These genetic modifications allow organisms and species to differentially adapt to changing environmental conditions and selective pressures.

Evolution of the Max and Mlx Transcription Factor Networks

In Chapter 3, I use the Max and Mlx transcription factor networks as a model to investigate such signs of regional conservation, domain shuffling, gene duplication, and dynamic network interactions. In particular I focused on the highly conserved basic helix-loop-helix-leucine zipper (bHLHZ) domain responsible for protein dimerization and DNA binding. Retention and conservation of Max and Mlx network members (Max, Mlx, Mnt, Mad, Myc, and Mondo) over a billion years of animal divergence suggest the functions and interactions of these proteins are vital for proper cell regulation. In fact, members of the Max and Mlx transcription factor networks govern basic, essential cellular processes such as proliferation, differentiation, metabolism, growth, and apoptosis.

However, the essentiality and dependency among Max and Mlx network proteins has yet to be determined and will most likely vary among organisms, cell types and physiological stages. For example, rat PC12 pheochromocytoma cells express c-Myc, L-Myc, Mnt, Mga, Mlx, and several Mxd family proteins and are able to proliferate as well as terminally differentiate into neurons despite a lack of functional Max. In contrast, rat embryos lacking Max die early post-implantation (Ribon, Leff et al. 1994). This poses two leading questions: Are Myc, Mnt, Mga and Mxd functionally involved or responsible for cell proliferation and differentiation in PC12 cells? If so, do they interact with another activating partner such as Mlx in the absence of Max? Alternatively, these cells may proliferate and differentiate independently of the Max network, opening an entirely new line of questioning surrounding possible mechanisms regulating cell cycle progression.

While we addressed the conservation of Max and Mlx network members in animals and characterized the bHLHZ interaction domain, it is still of great interest to understand the dynamics and function of these proteins *in vitro* and *in vivo*. Significant attention has been devoted to understanding the mechanisms of proliferation, growth, and metabolism for c-Myc and MondoB (WBSCR14/ChREBP) in mammals. However, relatively little is known about the antagonistic counterparts Mxd and Mnt or their affect on overlapping pathways.

What is the function of Mxd?

In vertebrates, Mnt expression is ubiquitous and independent of the cell cycle, while the Mxd (Mxd1-4) protein family exhibits dynamic expression and presumably suppresses c-Myc in a cell specific manner (Hooker and Hurlin 2006). Finding ways to inhibit specific transcriptional targets of c-Myc may be crucial to prevent cancer formation, since ectopic expression of c-Myc even at normal levels can lead to tumorigenesis (Freie and Eisenman 2008). Although Mnt and Mxd proteins have not been identified as tumor suppressors, their inherent roles in antagonizing c-Myc may prove important in controlling this proto-oncogene.

We found that most animals have single copies of Max, Mlx, Myc, Mondo, Mnt and Mxd genes. However, the "ancestral" role of Mxd is unknown and functional consequences of its duplication and divergence in vertebrates have not been fully examined. Moreover, the loss of Mxd in flies and combinatorial knockouts of Mxd paralogs in mouse imply that Mxd is not an essential gene (Foley, McArthur et al. 1998; Schreiber-Agus, Meng et al. 1998). Then why is Mxd conserved throughout the animal kingdom?

I propose the investigation of extant non-vertebrate species, which typify the ancestral network and provide a crucial link for comparing binding patterns and ascribing functions between various organisms and network topologies. From this information, it is possible to infer how Mxd expression patterns and Mlx binding abilities diverged in vertebrates. While the predictions described in Chapter 3 loosely suggest Mxd and Mlx cannot bind in non-vertebrates, experimental validation can help determine whether Mlx binding with Mxd1 and Mxd4 was gained or Mxd2 and Mxd3 was lost and the functional consequences thereof.

Moreover, experiments in non-vertebrates can more readily clarify if Mxd possesses any essential functions and explain how flies have compensated for its loss.

Do Mnt and Mxd repress Mondo function?

As the sole repressor in the *Drosophila* Max and Mlx networks, dMnt can indicate the necessary roles in suppressing dMyc and dMondo. However, there are conflicting reports on whether dMnt binds to dMlx (Hurlin, Quéva et al. 1997; Meroni, Cairo et al. 2000), and it is unknown if dMnt has any impact on dMondo function. In the event that dMnt dimerizes with dMlx, a new level of metabolomic regulation will emerge. Relative expression patterns, subcellular locations, preferential binding, and gene targets of dMnt:dMlx, dMnt:dMax, and dMondo:dMlx will be important for determining the mechanisms by which dMnt represses dMyc and dMondo and their subsequent affect on proliferation, growth, and metabolism.

Likewise, the affect of Mxd1 and Mxd4 interaction with Mlx on MondoA and MondoB glucose response in vertebrates is unknown. Using a dominant negative form of Mlx (dnMlx) in rat hepatocytes, Ma et al (2006) found eight genes that were both repressed by glucose and increased by dnMlx, while 26 genes showed increased expression from dnMlx alone. Hence blocking Mlx binding increases the expression of some genes. Do Mnt or Mxd proteins interact with either Max or Mlx to repress these glucose responsive genes?

MondoA and MondoB Glucose Response

In Chapter 4, I address how vertebrate Mlx network members MondoA and MondoB control the transactivation of genes involved in energy homeostasis through a complex system of glucose response. In general, MondoA and MondoB are cytoplasmic, yet actively shuttle between the nucleus and cytosol in low glucose conditions. As glucose levels and hence glucose-6-phosphate (G6P) abundance rises, MondoA and MondoB nuclear accumulation and transactivation of target genes also increases. Surprisingly, the mechanisms governing their subcellular localization and transactivation are still not well understood. From sequence analysis and protein predictions, we presented a model where G6P allosterically binds to a novel Mondo Conserved Region (MCR6) and catalyzes a conformational change in Mondo proteins that subsequently permits the association with

transactivating cofactors such as CBP/p300.

MCR6 resembles both the nine amino acid transactivation domain motif as well as a binding pocket for G6P, determined by comparison of G6P-interacting proteins. Recent evidence suggests G6P triggers MondoA and MondoB nuclear accumulation and transactivation of genes involved in *de novo* lipogenesis and glucose uptake. However, Mondo Conserved Regions (MCRI-V) exhibit cooperative and overlapping roles in regulating MondoA and MondoB glucose response. Through structural predictions I also predicted the intramolecular contacts among MCRs that mediate inhibition and activation of Mondo proteins in low and high glucose, respectively.

Sequence analysis also revealed the presence of a nuclear receptor box in Mondo and MondoB, but not MondoA proteins. The inclusion of MondoB and nuclear receptor HNF4α in the MondoB:HNF4α:CBP transactivating complex and proximity of their DNA binding domains in the L-PK gene promoter strongly suggests these proteins interact. Moreover, mutations in nuclear receptors are also linked to defects in glucose metabolism, such as type II diabetes. Along with verifying the interaction between MondoB and nuclear receptors, it would be interesting to determine their common gene targets and possible synergistic effects on glucose regulation.

In addition, I found that the dimerization and cytoplasmic domain (DCD) was highly conserved between Mlx and Mondo proteins. Although the specific mechanisms within the DCD that cause cytoplasmic retention of Mondo and Mlx monomers are unknown, masking the DCD through dimerization is necessary for their nuclear entry. However, Mxd and Mnt proteins lack a DCD domain. This poses several questions. Can Mnt:Mlx, Mxd1:Mlx, or Mxd4:Mlx heterodimers relocate to the nucleus? Does the DCD in Mlx affect the preferential binding of Mondo, Mxd or Mnt proteins? And does the DCD confer an active function for Mondo and Mlx proteins in the cytoplasm?

Is MondoA involved in cellular redox?

Interaction with 14-3-3 has also been shown to contribute to MondoA cytoplasmic localization (Eilers, Sundwall et al. 2002). Interestingly, MondoA is known to localize to the Outer Mitochondrial Membrane (OMM), and we found chaperone protein Mitochondria

import stimulating factor (MSF) is a member of the 14-3-3 protein family that transports proteins to the mitochondria. Moreover, 14-3-3 has been shown to directly interact with and inactivate ATP synthase subunit ATP5B, which resides on the OMM (Bridges and Moorhead 2004). Inactivation of ATP5B is important in low oxygen conditions and is a contributing factor to the regulation of beta-oxidation. While the association between MondoA and ATP5B is completely speculative, MondoA is otherwise implicated in beta-oxidation.

Thioredoxin interacting protein *txnip* is a gene target for MondoA. The TXNIP protein modulates the cellular reduction-oxidation (redox) state by binding to and inhibiting the antioxidant thioredoxin enzyme (Minn, Hafele et al. 2005). TXNIP gene expression is dramatically up-regulated in response to glucose in human pancreatic beta cells and leads to a significant increase in beta-cell apoptosis. This transcriptional increase was not a result of an autocrine insulin effect, and was abolished by a mutation to the first E-box like motif. Hence it is possible that MondoA is involved in controlling or responding to cellular oxidation levels.

Mondo, Nuclear Receptors, and Type II Diabetes

Mondo proteins and nuclear receptors PPAR-γ, HNF1α (4) and HNF4α (5) are also associated with insulin sensitivity and type II diabetes (Sears, Hsiao et al. 2009; Noordeen, Khera et al. 2010). As previously mentioned, Mondo activation is correlated with the phosphorylation of glucose by hexokinases to form G6P. Surprisingly, the conversion of glucose to G6P and then back to glucose is increased in patients with type II diabetes (Hutton and O'brien 2009). Moreover, some tumors show increased binding of hexokinase to the OMM (Matés, Segura et al. 2009). While completely disrupting MondoB is not an ideal method of regulating glucose concentrations due to complications in glycogen accumulation, decoupling the MCR functions and G6P binding may be an effective therapeutic target.

One possible method to treat type II diabetes uses thiazolidinediones (TZDs), which are a class of compounds that activate PPAR-γ at submicromolar levels and increase insulin sensitivity. While the mechanism is not completely known, PPAR-γ may be the biochemical target of TZDs. However, MondoA in muscle tissues exhibit gene expression signatures that predict insulin sensitization by TZDs (Sears, Hsiao et al. 2009). While the impact of this is

still unclear, the linkage between Mondo and nuclear receptors may prove pivotal for treating diabetes.

Max and Mlx Networks have overlapping function

Thus far, Max and Mlx networks have largely been investigated independently. However, *Drosophila* mutants exhibit an overlap in Myc and Mondo function. Hypomorphic knockdowns of dMondo (*dmon1*) and dMyc (*dm*) in Drosophila show mutants homozygous for 1) *dmon1* were weak, subviable, and subsequently died 2) *dm* mutants were viable yet smaller than normal, and 3) *dmon1/dm* mutants were completely lethal. This indicates dMyc and dMondo have a synthetic lethal interaction and coordinately regulate at least one essential gene (Billin and Ayer 2006). Furthermore, MondoA and MondoB targets LDH-A and L-PK, respectively, are also known c-Myc targets. More definitively, recent evidence shows c-Myc is required for activation of MondoB dependent genes (Zhang, Metukuri et al. 2010). This clearly indicates Max and Mlx networks overlap in energy regulation.

Connected by the highly conserved and homologous bHLHZ domain, the function and composition of Max and Mlx network members have otherwise greatly diverged. Mondo and Myc proteins contain distinct transactivation domains, while Mnt and Mxd proteins have repressive domains. Mondo and Mlx proteins are largely cytoplasmic due to their DCD domain, while Mondo proteins also contain several unique Mondo Conserved Regions. Little is known of the Mga protein, aside from its bHLHZ and T-box DNA binding domains. Nonetheless, the diversity and conservation of Max and Mlx network interactions, proteins, domains, and sites provide exceptional refinement and complexity that is essential for cellular regulation and has largely been preserved throughout animal evolution.

Despite decades of research on Max and Mlx network proteins, we still lack an adequate understanding how they function. Nonetheless, we continually move closer to uncovering their influence on major life threatening diseases such as cancer and their ultimate control over fundamental aspects of life.

References

Billin, A. N. and D. E. Ayer (2006). "The Mlx network: evidence for a parallel Max-like transcriptional network that regulates energy metabolism." Curr Top Microbiol Immunol **302**: 255-78.

Bridges, D. and G. B. Moorhead (2004). "14-3-3 proteins: a number of functions for a numbered protein." Sci STKE **2004**(242): re10.

Eilers, A. L., E. Sundwall, et al. (2002). "A novel heterodimerization domain, CRM1, and 14-3-3 control subcellular localization of the MondoA-Mlx heterocomplex." Molecular and Cellular Biology **22**(24): 8514-26.

Foley, K. P., G. A. McArthur, et al. (1998). "Targeted disruption of the MYC antagonist MAD1 inhibits cell cycle exit during granulocyte differentiation." EMBO J **17**(3): 774-85.

Freie, B. W. and R. N. Eisenman (2008). "Ratcheting Myc." Cancer Cell **14**(6): 425-6.

Hooker, C. W. and P. Hurlin (2006). "Of Myc and Mnt." Journal of Cell Science **119**(Pt 2): 208-16.

Hurlin, P. J., C. Quéva, et al. (1997). "Mnt, a novel Max-interacting protein is coexpressed with Myc in proliferating cells and mediates repression at Myc binding sites." Genes & Development **11**(1): 44-58.

Hutton, J. and R. O'brien (2009). "Glucose-6-phosphatase Catalytic Subunit Gene Family." Journal of Biological Chemistry **284**(43): 29241-29245.

Ma, L., L. N. Robinson, et al. (2006). "ChREBP*Mlx is the principal mediator of glucose-induced gene expression in the liver." J Biol Chem **281**(39): 28721-30.

Matés, J. M., J. A. Segura, et al. (2009). "Glutamine homeostasis and mitochondrial dynamics." Int J Biochem Cell Biol **41**(10): 2051-61.

Meroni, G., S. Cairo, et al. (2000). "Mlx, a new Max-like bHLHZip family member: the center stage of a novel transcription factors regulatory pathway?" Oncogene **19**(29): 3266-77.

Minn, A. H., C. Hafele, et al. (2005). "Thioredoxin-interacting protein is stimulated by glucose through a carbohydrate response element and induces beta-cell apoptosis." Endocrinology **146**(5): 2397-405.

Noordeen, N. A., T. K. Khera, et al. (2010). "Carbohydrate-responsive element-binding protein (ChREBP) is a negative regulator of ARNT/HIF-1beta gene expression in pancreatic islet beta-cells." Diabetes **59**(1): 153-60.

Ribon, V., T. Leff, et al. (1994). "c-Myc does not require max for transcriptional activity in PC-12 cells." Mol Cell Neurosci **5**(3): 277-82.

Schreiber-Agus, N., Y. Meng, et al. (1998). "Role of Mxi1 in ageing organ systems and the regulation of normal and neoplastic growth." Nature **393**(6684): 483-7.

Sears, D. D., G. Hsiao, et al. (2009). "Mechanisms of human insulin resistance and thiazolidinedione-mediated insulin sensitization." Proc Natl Acad Sci USA **106**(44): 18745-50.

Zhang, P., M. Metukuri, et al. (2010). "c-Myc Is Required for the ChREBP-Dependent Activation of Glucose-Responsive Genes." Mol Endocrinol: 13.

Appendix

Appendix A

Statistical Methods for Analyzing High Dimensional Molecular Data (HDMD)

High throughput technologies, e.g., microarray, genomic sequences, metabolomics, and animal genotyping produce massive amounts of new molecular data. Indeed, high dimensional molecular data (HDMD) show great promise for providing deeper insights into the composition and dynamics of complex biological processes. Unfortunately, these large infusions of data can greatly complicate statistical analyses. Before the value of HDMD can be realized, a number of significant statistical issues must be resolved.

Typically HDMD have many more variables or dimensions (*D*) than individuals or replicates (*N*). When $D>>N$ classical multivariate statistical procedures can fail, give misleading results, or simply become intractable (Donoho 2000; Ransohoff 2005; Clarke, Ressom et al. 2008). Greatly increasing number of variables (dimensions) can distort interdistance relationships by shrinking the expected difference in distances between a point and its nearest and farthest neighbor (Beyer, Goldstein et al. 1999). This "curse of dimensionality" (Bellman 1961) complicates the ability to discriminate true biological relationships, particularly in procedures that utilize distance matrices such as cluster analysis.

High throughput data collection is generally not hypothesis driven. Data collected en masse aggregates extensive amounts of potentially irrelevant data in conjunction with variables critical for evaluating a particular hypothesis. Teasing out salient variables involves searching for the proverbial needle in a haystack. One problem with finding the correct needle is that numerous statistical tests must be carried out which can have profound impact on type I and type II error structure (Rao and Gu 2001; Rice, Schork et al. 2008).

A common theme in these problems is the high dimensionality of data. Hence, methods for reducing dimensionality while retaining intrinsic biological information are in great demand. Almost all statistical analyses of HDMD initially attempt to find biologically meaningful clusters inherent to data or (equivalently) reduce the number of dimensions to facilitate variable selection. Many statistical methods have been proposed for these problems

(Swets 1988; Wang, Miller et al. 2008). However, appropriate method selection involves consideration of the biological models underlying variability in data, statistical assumptions, and relative efficacy in addition to the accuracy and meaningful biological interpretation of results. Unfortunately, these considerations are often overlooked in HDMD analyses.

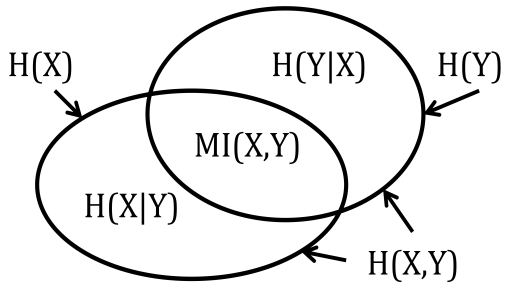
Herein, we briefly address some statistical problems inherent to HDMD. Some types of HDMD are non-metric, such as genomic sequence data, which makes them resistant to classical multivariate statistical analyses. We discuss methods for circumventing this issue and summarize information about three multivariate statistical methods that have been used for metric data. These methods include principal component analysis (PCA), factor analysis (FA), and discriminant analysis (DA). Finally, we provide a collection of computer software written in R that shows considerable utility in the analysis of HDMD.

Appropriateness of HDMD for Multivariate Statistical Analysis

Molecular data arising from high throughput technologies can occur in many forms ranging from continuous, metric gene expression profiles to discrete, alphabetic and non-metric genome sequences. The type of data can restrict the type of statistical analyses that can be performed. Genomic sequence data, for instance, are comprised of alphabetic letters, which have no underlying natural metric (Atchley, Zhao et al. 2005). Non-metric data are quite difficult to analyze using conventional multivariate statistics, which commonly require variances and covariances. Below, we describe two separate approaches for dealing with this problem.

Methods from information theory can circumvent problems with non-metric data by using *entropy* and *mutual information* to estimate variability and covariability (Shannon 1948; Applebaum 1996). Entropy (eq. 1) is a measure of uncertainty for a discrete random variable, whereas mutual information (MI) (eq. 2) represents the interdependence of two discrete random variables (Applebaum 1996; Kullback 1997). As shown in Figure 1, the intersection of the entropy space bounds MI, and thus quantifies the reduction in uncertainty of one variable given the knowledge of a second variable (Martin, Gloor et al. 2005). Traditionally the logarithm base for entropy is calculated with unit bits (b=2), nats (b=e) or dits (b=10). Alternatively, entropy estimates can be normalized to a common scale where

$H \in [0,1]$ by setting $b=n$, the number of possible states. Placing entropy values on the diagonal of a MI matrix forms a structure comparable to a covariance matrix appropriate for variability decomposition.



$$H(X) = \sum_{i=1}^{n} p_{x_i} \log_b(p_{x_i}) \quad (1)$$

$$\begin{aligned} MI(X,Y) &= H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \\ &= H(X,Y) - H(X \mid Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \quad (2)$$

$$NMI(X,Y) = {}^{2 * MI(X,Y)} \big/ _{H(X,Y)} \quad (3)$$

Figure 1: Entropy and Mutual Information
The Venn diagram shows the joint information space of two discrete, random variables *X* and *Y* described in terms of their entropy (eq. 1) and mutual information (eq. 2) values (Martin, Gloor et al. 2005). Mutual Information *MI*(*X*,*Y*) is bounded by the individual entropy estimates *H*(*X*) and *H*(*Y*). Normalized mutual information (eq. 3) accounts for background information, restricting NMI to a zero to one scale.

Given non-metric, alphabetic multiple sequence alignments (MSAs), entropy and mutual information can define highly conserved or correlated sites, respectively (Atchley, Wollenberg et al. 2000). For DNA (*n*=4 nucleotide) or protein (*n*=20 amino acid) sequences, normalized entropy *H*=0 indicates an invariable site while *H*=1 represents a site where all states occur with equal probability. A histogram of entropy values produces an entropy profile, which facilitates visual comparison of variability at points of interest in a sequence. Similarly, MI identifies pairs of statistically dependent or coupled sites where *MI*=1 indicates complete coupling.

While this information theory approach can be fruitful, entropy and mutual information have several limitations when *D*>>*N*. First, entropy is not a consistent estimator. Adding or removing multiple sequences, especially divergent homologs, may greatly influence H and MI values. Second, MI estimates will only be zero for independent sites in the unlikely case where all pair-wise combinations (DNA: 4x4=16, AA: 20x20=400) are observed. Third, MI must be normalized by a leveling ratio to account for the background distribution arising from the stochastic pairing of independent, random sites. Martin et al.

(Martin, Gloor et al. 2005) tested several methods and proposed a scaling factor (eq. 3) based on its increased sensitivity. Fourth, high MI may be due to phylogenetic signal instead of structural or functional constraints (Wollenberg and Atchley 2000; Dunn, Wahl et al. 2008). Martin et al. (Martin, Gloor et al. 2005) found that the background MI, particularly from phylogenetic covariation, has a contributable effect for MSAs with less than 125 to 150 sequences.

Information theory further assumes that states are distinct. However, non-metric data may contain discrete, yet related states. For example, alphabetic sequence data suggests that isoleucine (I) is as closely related to histidine (H) as it is to leucine (L) or valine (V). Classifications of amino acid physiochemical properties clearly show this not to be the case (Kidera, Konishi et al. 1985). Representing amino acid relationships based on non-metric alphabetic letters excludes important biological information and the significance of functional conservation is consequently diminished or eliminated.

A more powerful approach is to transform alphabetic characters into numerical indices that quantify realistic, biological properties. These numerical indices can greatly facilitate discrimination of pertinent physiochemical features of amino acids. Several such metric transformations have been proposed reflecting important properties of amino acid indices (Sandberg, Eriksson et al. 1998; Atchley, Zhao et al. 2005; Opiyo and Moriyama 2007; Georgiev 2009). Converting amino acid residues into numeric quantities that signify complex functional or structural relationships enables application of more rigorous statistical techniques, e.g. analysis of variance, regression, and discriminant analysis. Subsequently, amino acid metrics have been applied to estimation of molecular architecture (Atchley and Zhao 2007), periodicity of protein solenoid repeats (Marsella, Sirocco et al. 2009), subcellular localization (Mundra, Kumar et al. 2007), and coevolution (Caporaso, Smit et al. 2008).

Metric data on the other hand, scale biological attributes along a continuous spectrum of variability. This permits multivariate statistical analyses to find notable instances or patterns of variation based on variance and covariance estimates. However, multiple undesirable or underlying sources of variation, i.e. environmental, technical, and

demographic, may confound the pertinent patterns of variation (Leek and Storey 2007). Meaningfully partitioning variability into a reduced structure that maximizes the signal to noise ratio is thus the crux of HDMD analysis.

PCA, FA, and DA are three widely used and statistically sophisticated methods for reducing dimensionality in metric data. Although these methods result in matrices of grossly similar structure, they use different criterion to provide disparate, yet meaningful interpretations. Since the application of these procedures is sometimes confused, we delineate their assumptions, methods and interpretations, Table 1.

Dimensionality Reduction Methods

Principal Component Analysis

Principal Component Analysis (PCA) is the simplest and most widely used multivariate statistical procedure for reducing dimensionality. The goal of PCA is to reduce data dimensionality by explaining maximal variability in data through linear combinations of original variables representing correlated variability (Jolliffe 2002; Johnson and Wichern 2007).

PCA solutions obtained by analyzing either the covariance or correlation matrix can be quite different. A covariance matrix weights variable importance by individual scales of variance, which can greatly differ in magnitudes. Alternatively, original data or covariances can be standardized to produce correlations, so variables are equally weighted with mean zero and unit variance.

Assuming some correlation among variables, the strength of PCA arises from diagonalizing the covariance (correlation) matrix. In doing so, fewer dimensions can explain approximately as much variability as the original larger set of variables. PCA can achieve this in at least three computationally different ways, including eigenvalue decomposition (Scholz and Selbig 2007), singular value decomposition (SVD) (Alter, Brown et al. 2000; Holter, Mitra et al. 2000; Liu, Hawkins et al. 2003; Wall, Rechtsteiner et al. 2003; Shlens 2009), or other adaptive methods (Baldi and Hornik 1995; I. Diamantaras, Yuan Kung et al. 1996).

PCA can be best explained by eigenanalysis of a dataset *X*, which produces a series of eigenvalues and associated eigenvectors. Each eigenvector or *principal component* (PC) defines an axis of variability by a linear combination of original variables, while the associated eigenvalue measures the amount of variability explained by that PC. Eigenvectors are comprised of correlations between original variables and PC axes, which specify the axes linear coefficients. The magnitude and sign of PC coefficients determine the relative contribution of variability and direction in multidimensional space of each axis.

PCs are ordered by decreasing eigenvalue magnitudes to create a *loadings* matrix (*V*). Similar to linear regression, the first principal axis is the line of best fit and each subsequent principal axis maximally fits any residual variation. Each PC is orthogonal to the other axes, and thus defines uncorrelated patterns of variability. The more highly correlated the variables, the fewer principal components are needed to summarize the original data.

The relative interrelationships among original data can be approximated by the principal component *scores*, which are projections of the original data onto principal axes, $S=XV$. Observation i having D associated explanatory variables and k PCs can thus be approximated by the equation in Table 1a.

Typically, only a subset of k PCs is necessary to adequately summarize data variability. When $k=D$, all variability is retained and all of the original data can be recovered. Several methods propose critical thresholds where $k<<D$ in order to eliminate minor axes of variation (Zwick and Velicer 1986). The most direct is Kaiser's method of choosing all components with eigenvalue $\lambda > 1$ (Kaiser 1958). This ensures that each component explains at least as much as a single variable. Another frequently used method is Cattell's scree plot, which plots eigenvalues in decreasing order. This procedure excludes PCs after the elbow where the curve becomes asymptotic and additional components provide little additional information (Cattell 1958).

PCA reduces dimensionality by transforming data onto axes of major variability and subjectively excluding axes of minor variability. The mathematically elegant and unique solution makes PCA appealing. However, PCA has several major shortcomings that can

have a significant negative impact on using PCA for understanding the biological basis of molecular and morphological variation. Primary limitations include:

1. PCA summarizes all of the variation in a sample including genetic as well as environmental, demographic, technical and stochastic components (Leek and Storey 2007). The latter components can bias results leading to erroneous or inaccurate partitioning of realistic biological information.
2. As a data compression procedure, PCA summarizes patterns of observed variation in a dataset. Unlike factor analysis, PCA does not delineate the underlying latent structure of the data, which is most important in understanding the biological basis of variability.

Factor Analysis

Factor Analysis (FA) is a dimension reduction procedure used to determine the underlying or latent structure of the data (L. Gorsuch 1983). The goal of FA is to ascertain meaningful latent variables that affect observable data and transform observations into an interpretable representation of data variability.

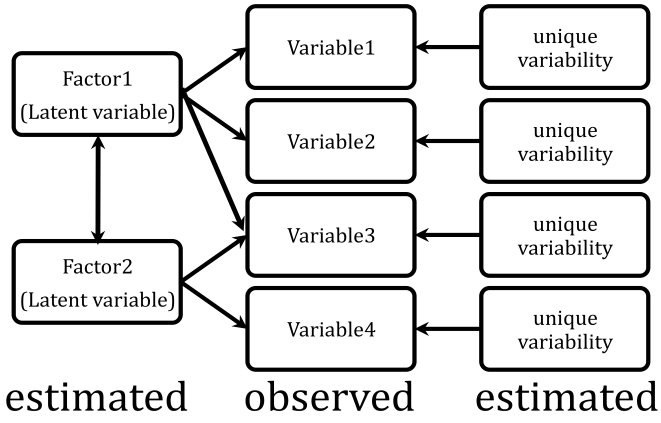


Figure 2: Factor Structure.
Factor Analysis partitions observed data variability into common and unique estimates. Factors (latent variables) are linear combinations of original variables that describe common variation. Arrows indicate a strong association between estimated and observed values. In this example, Factor 1 can be described by variables 1, 2, and 3, while only variables 3 and 4 describe Factor 2. When an oblique factor rotation is used, factors can also be related, shown by the two-sided arrow.

First, FA partitions variability into that common among all variables and a residual value unique or specific to each variable, Figure 2. This partition is achieved by estimating the amount of variability that can be explained by the factor structure for each dimension, termed *communality* (h^2). By replacing diagonal elements of the correlation matrix with communality estimates, FA weights dimensions by their contribution to the correlation structure while emphasizing covariability. When $D>>N$ as in HDMD, maximum likelihood factor analysis cannot be performed and communality is instead iteratively estimated through eigenvalue decomposition via the principal axes method (Cudeck and C. MacCallum 2007).

FA decomposes the modified correlation matrix and partitions variability into linear combinations of original variables. Eigenanalysis or SVD similarly create k ordered vectors called *factors*, which define axes of common variation. Due to communality estimation, the number of factors must be defined *a priori*, e.g. by Cattell's scree plot. FA does not exhaustively account for variability, but maximally explains shared variation given the k factor model. These factors are ordered by the proportion of variability explained to create a *D*x*k loading* matrix (Λ).

While axes defined by initial factor loadings correspond to orthogonal directions of maximal variability, they may not reflect associations with the defining variables. FA aligns axes with variables by applying a factor rotation to strengthen relevant correlations and reduce affiliations with other axes. This populates the loading matrix with ones and zeros to create sparsity, achieve simple structure, and increase interpretability.

Rotating by the Varimax procedure maintains orthogonality among factors, while an oblique rotation like Promax allows the latent variables to be correlated. Pertinently, loadings can be scaled by infinite orthogonal rotations without altering the amount of variability explained. Because of the rotation, FA does not produce a unique solution; rather, it portrays patterns of influential dimensions that describe the role of latent variables. Dimensions with large loading values markedly contribute to the factor structure and can be used to describe complex biological processes.

Original observations can be projected onto these rotated axes to display their association according to biologically relevant latent variables. Since FA only considers

common variability during factor decomposition, transformed observations, or *scores*, must be estimated to account for residual variation. Estimating scores using regression considers variable (R) and factor (Φ) correlations by defining the *N*x*k* score matrix (S) by $S = (X - \mu)R^{-1}\Lambda\Phi$ (Cudeck and C. MacCallum 2007). However, scores cannot be directly computed by regression when *D*>>*N* due to the singular covariance matrix and a generalized inverse method must be used instead. Bartlett's method circumvents this by using variable uniqueness, $1\text{-}h^2$, and defining $S = (X - \mu)\Lambda^*(\Lambda^T\Lambda^*)^{-1}$ where $\Lambda^* = \Lambda\begin{bmatrix} 1/(1-h_1^2) & & \\ & \vdots & \\ & & 1/(1-h_D^2) \end{bmatrix}$ (Cudeck and C. MacCallum 2007).

Scores scale observations to reflect associations with the underlying variable structure. Hence scores can further be used to extract or compare observations according to inferred biological processes as done for amino acids (Atchley and Zhao 2007), immunology (Genser, Cooper et al. 2007), and microarrays (Peterson 2002; Hochreiter, Clevert et al. 2006; Zeng, Wu et al. 2008).

While PCA and FA partition variability over the entire data, other forms of HDMD include subclassification structures that require further conditioning. These data may have stratified variability due to factors like segregating populations, protein families, or cell types. Since drug testing and evolutionary models are often contingent on such classification, methods to reduce dimensionality and stress trends in variation while optimizing class separation are coveted.

Discriminant Analysis

Discriminant analysis (DA) is a commonly used procedure for discriminating groups according to a reduced set of informative dimensions. By creating combinations of original variables, DA defines discriminant functions that partition patterns of between group variation. Accounting for variable correlation can increase discriminatory power and provide a simple method of classification when variables cannot distinguish groups independently. Although several discriminant analysis techniques are available, we will focus on the simplest method, linear DA (LDA) (J. Huberty 1994).

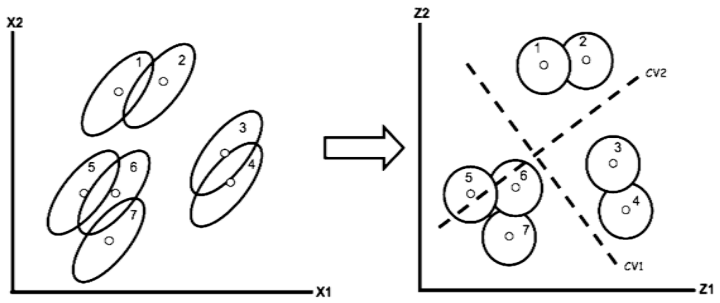


Figure 3: Group Separation by Discriminant Analysis**.**
Discriminant analysis minimizes within group variation and maximizes between group variation. Elliptical variances (left) are transformed into minimized spherical variances (right) around the 7 group means shown here. Depicted by dotted lines, discriminant functions, or canonical variates (CV), maximally discriminate group means to facilitate classification.

DA decomposes variability in data to both minimize within group variation and maximize between group variation, Figure 3. This is equivalent to minimizing and maximizing the *trace*, or sum of diagonal elements, of the within-group (Σ_W) and between-group (Σ_B) covariance matrices, respectively. Since Σ_W is singular in HDMD when D>>N, $trace(\Sigma_W^{-1}\Sigma_B)$ cannot be optimized without generalization (Howland and Park 2004). To circumvent this problem, several methods first reduce data dimensionality to remove singularity before optimizing discrimination. The most popular reduction involves two rounds of SVD, as described below (O. Duda, E. Hart et al. 2001).

Initially, observations are standardized so each group has mean zero and dimensions have unit variance. Performing SVD summarizes this data into linearly independent vectors to create a *D*x*r* scaling matrix (F') of full rank *r*, which reduces data dimensionality to eliminate singularity. A second round of SVD then focuses on discriminating groups. A *g*x*D* matrix of standardized group means is centered by dimension means and scaled by sample variance $\sqrt{(g-1)(N-g_i)}$, with *g* groups, *N* observations, and g_i observations in group *i*. This matrix is multiplied by F' to form a group centric *g*x*r* matrix whose SVD solution defines *g-1* linear combinations separating group means. Finally, the two SVD solutions are multiplied to form a *D*x*g-1* scaling matrix $\hat{F}$ whose columns define *discriminant functions* and establish axes of maximal group separation based on original dimensions. Analogous to

the loading matrix in FA, scaling coefficients weight variables according to their relative importance in group distinction. Visual representation of this procedure may be found in (Campbell and Atchley 1981).

Observations are projected onto the discriminating axes to facilitate classification by scores, Table 1c. Plotting scores summarizes group relationships by clustering observations around group means and along axes that maximally discriminate multidimensional variability. Observations farther from the origin indicate greater distinction while distances within and between group clusters display complex associations.

To supplement visual comparison, the Mahalanobis distance quantifies similarity among groups by providing an objective measure of group resemblance (Mahalanobis 1936). The Mahalanobis distance measures the relationship between groups while considering the covariance among variables (eq. 4). When variables are independent, the covariance matrix Σ is the identity matrix, and the Mahalanobis distance reduces to the Euclidean distance. To determine the distance between group *i* and group *j*, the difference of group means for each variable is compared as in (eq. 5).

$$D = \sqrt{(x-\mu)\,\Sigma\,(x-\mu)^T} \qquad (4)$$

$$D(g_i, g_j) = \sqrt{(\mu_{g_i} - \mu_{g_j})\,\Sigma\,(\mu_{g_i} - \mu_{g_j})^T} \qquad (5)$$

HDMD Package for R

In order to analyze HDMD, researchers need powerful statistical software to implement methods capable of handling massive amounts of data. R is freely available, open source statistical software with a strong framework for multivariate analysis. Its transparency allows for both general users and open source software developers to contribute and update package functionality as new statistical or biological issues arise.

We present a new software package titled *HDMD* for handling problems previously unaccounted for in R, yet frequently surface during analysis of high dimensional data. A list of functions is provided in Table 2, while details and examples are further elaborated upon in the tutorial and manual found at <www4.ncsu.edu/~lgmcferr/ICMSBWorkshop>.

Method Applications

HDMD is predominantly generated to depict the status of and relationship among variables in context of a specific environment. Without a particular hypothesis in mind, HDMD typically include a plethora of extraneous or correlated variables, which mask pertinent variable relationships. Exploratory analyses are often appropriate for extracting salient features and reducing dimensionality to allow for more tailored and informed hypothesis testing.

Information theory techniques isolate variables in non-metric data whose values significantly deviate in relation to other variables for a particular dataset. When applied to multiple sequence alignments, entropy identifies highly conserved or variable sites and is useful for discovering patterns of relative sequence conservation, i.e. secondary structures, promoters, or genes (Schneider, Stormo et al. 1986; Chan, Liang et al. 2004; Vinga and Almeida 2007). Similarly, MI finds strongly coupled pairs or groups of sites and is useful for finding sites that covary due to phylogenetic, structural, functional or stochastic changes (Wollenberg and Atchley 2000; Gloor, Martin et al. 2005). However, these estimates do not inherently incorporate similarity among states, which can distort significance. Functional and structural properties of amino acids can be included for more biologically meaningful measurements, as done in grouped entropy and metric transformations (Atchley, Terhalle et al. 1999). This helps distinguish conserved and constrained or covarying and coevolving sites.

Metric transformations convert non-metric HDMD to enable statistically robust extraction of salient features. Deciphering the underlying mechanisms driving biological processes further requires knowledge of the latent variable structure. For metric data, various interdependence criteria delineate multivariate statistical procedures, which lead to disparate interpretations of results and require assorted tactics for handling statistical complications.

PCA and FA reduce dimensionality by creating k linear combinations of variables weighted according to their importance. The subjectivity of determining k can result in various levels of approximation and interpretation. While PCA forms principal components maximally explaining total variation, factors in FA represent common trends in variation

among dimensions (= latent variables). Note that when the majority of variability is shared among variables (communality close to 1), PCA and FA will decompose equivalent matrices and may provide similar results. However, this is the exception and not the rule. Since PCA and FA report highly similar structures but have very different interpretations, it is important to apply the appropriate method for the biological question at hand.

PCA is frequently applied to differential gene expression data from microarrays to reduce the number of variables and detect outliers (Alter, Brown et al. 2000; Holter, Mitra et al. 2000; Owzar, Barry et al. 2008). Much of the "noise" typically associated with microarrays can be removed by PCA dimension reduction to reveal dynamically expressed genes. However, when applying PCA to filter candidate genes, one must consider:

1. The major axes of variability may not correspond to an interpretable signal.
2. PCA assumes genes have comparable levels of variability (Rattray, Liu et al. 2006).
3. PCA does not distinguish sources of variability, so binding affinities, technical errors, systematic environmental variation, and other factors are also included in observed expression levels.

Since most microarrays include a surplus of screened genes, many with no association to the experiment, equally weighting all dimensions as in a correlation matrix may dramatically alter results. With few available replicates, high correlations may occur by chance and emphasize spurious variable relationships (Scholz and Selbig 2007). Paramount, variability in noise among probes may confound covariances and considerably effect results. To infer biological relationships among genes, latent variable analysis such as factor analysis should be used instead.

For stratified HDMD, DA can be used to define discriminant functions that maximally separate groups and classify data. By maximizing between and minimizing within group variability, transformed observations reflect distinct trends in group variation. Implementing distance calculations between groups, as in the Mahalanobis distance, can further quantify group associations.

In exploratory analysis, DA accentuates influential variables and discriminating features among groups to facilitate further interpretation. For example, Atchley and Zhao (Atchley and Zhao 2007) applied LDA to identify physiochemical relationships among bHLH protein families, while Gama et al. (Gama, Costa et al. 2004) discriminated between clinical and asymptomatic forms of visceral leishmaniasis to identify an immunological marker. Confirmatory analysis can subsequently classify unknown observations. This has proven to be a powerful method for predicting disease susceptibility (Ahmed, Santosh et al. 2009), drug resistance (Ji, Zhang et al. 2009), and treatment response (Djoba Siawaya, Chegou et al. 2009).

Currently, the enormity of variables in HDMD hinders the progress of understanding complex biological processes due to many biological assumptions and statistical constraints. Many multivariate statistical procedures are available for exploring such data, although they are not without limitation. The R package HDMD aims to provide simple functions for implementing some of these methods and handle arising complications inherent to HDMD. Overcoming or at least recognizing these issues can ultimately benefit the uses of HDMD.

References

Ahmed, S. S., W. Santosh, et al. (2009). "Metabolic profiling of Parkinson's disease: evidence of biomarker from gene expression analysis and rapid neural network detection." J Biomed Sci **16**: 63.

Alter, O., P. O. Brown, et al. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." Proc Natl Acad Sci USA **97**(18): 10101-6.

Applebaum, D. (1996). "Probability and information: an integrated approach‡." 212.

Atchley, W. R., W. Terhalle, et al. (1999). "Positional dependence, cliques, and predictive motifs in the bHLH protein domain." J Mol Evol **48**(5): 501-16.

Atchley, W. R., K. R. Wollenberg, et al. (2000). "Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis." Molecular Biology and Evolution **17**(1): 164-78.

Atchley, W. R. and J. Zhao (2007). "Molecular architecture of the DNA-binding region and its relationship to classification of basic helix-loop-helix proteins." Molecular Biology and Evolution **24**(1): 192-202.

Atchley, W. R., J. Zhao, et al. (2005). "Solving the protein sequence metric problem." Proc Natl Acad Sci USA **102**(18): 6395-400.

Baldi, P. F. and K. Hornik (1995). "Learning in linear neural networks: a survey." IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council **6**(4): 837-58.

Bellman, E. R. (1961). "Adaptive control processes: a guided tour‡." 255.

Beyer, K., J. Goldstein, et al. (1999). "When is "nearest neighbor" meaningful?" Proc. 7th International Conference on Database Theory (ICDT'99) **1540**: 217-235.

Campbell, W. R. and W. R. Atchley (1981). "The geometry of canonical variate analysis." Systematic Zoology **30**: 268-280.

Caporaso, J. G., S. Smit, et al. (2008). "Detecting coevolution without phylogenetic trees? Tree-ignorant metrics of coevolution perform as well as tree-aware metrics." BMC Evol Biol **8**: 327.

Cattell, R. B. (1958). "Extracting the correct number of factors in factor analysis." Educational Researcher **18**: 791-838.

Chan, C. H., H. K. Liang, et al. (2004). "Relationship between local structural entropy and protein thermostability." Proteins **57**(4): 684-91.

Clarke, R., H. W. Ressom, et al. (2008). "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data." Nat Rev Cancer **8**(1): 37-49.

Cudeck, R. and R. C. MacCallum (2007). "Factor analysis at 100: historical developments and future directions." 381.

Djoba Siawaya, J. F., N. N. Chegou, et al. (2009). "Differential cytokine/chemokines and KL-6 profiles in patients with different forms of tuberculosis." Cytokine **47**(2): 132-6.

Donoho, D. (2000). "Mathematical Challenges of the 21st Century - High-Dimensional Data Analysis: The Blessings and Curses of Dimensionality." from http://www.stat.stanford.edu/~donoho/Lectures/AMS2000/MathChallengeSlides2*2.pdf.

Dunn, S. D., L. M. Wahl, et al. (2008). "Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction." Bioinformatics **24**(3): 333-40.

Gama, M. E., J. M. Costa, et al. (2004). "Serum cytokine profile in the subclinical form of visceral leishmaniasis." Braz J Med Biol Res **37**(1): 129-36.

Genser, B., P. J. Cooper, et al. (2007). "A guide to modern statistical analysis of immunological data." BMC Immunol **8**: 27.

Georgiev, A. (2009). "Interpretable Numerical Descriptors of Amino Acid Space." Journal of Computational Biology **16**(5): 703-723.

Gloor, G. B., L. C. Martin, et al. (2005). "Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions." Biochemistry **44**(19): 7156-65.

Hochreiter, S., D. A. Clevert, et al. (2006). "A new summarization method for Affymetrix probe level data." Bioinformatics **22**(8): 943-9.

Holter, N. S., M. Mitra, et al. (2000). "Fundamental patterns underlying gene expression profiles: simplicity from complexity." Proc Natl Acad Sci USA **97**(15): 8409-14.

Howland, P. and H. Park (2004). "Generalizing discriminant analysis using the generalized singular value decomposition." IEEE transactions on pattern analysis and machine intelligence **26**(8): 995-1006.

Diamantaras, I. K., S. Yuan Kung, et al. (1996). "Principal component neural networks: theory and applications." 255.

Huberty, J. C. (1994). "Applied discriminant analysis." 466.

Ji, B., Z. Zhang, et al. (2009). "Differential expression profiling of the synaptosome proteome in a rat model of antipsychotic resistance." Brain Res **1295**: 170-8.

Johnson, A. R. and W. D. Wichern (2007). "Applied multivariate statistical analysis." 773.

Jolliffe, T. I. (2002). "Principal component analysis." 487.

Kaiser, H. F. (1958). " The varimax criterion for analytic rotation in factor analysis." Psychometrika **23**: 187-200.

Kidera, A., Y. Konishi, et al. (1985). "Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids." Journal of Protein Chemistry **4**(1): 23-55.

Kullback, S. (1997). "Information theory and statistics." 399.

Gorsuch, L. R. (1983). "Factor analysis." 425.

Leek, J. T. and J. D. Storey (2007). "Capturing heterogeneity in gene expression studies by surrogate variable analysis." PLoS Genet **3**(9): 1724-35.

Liu, L., D. M. Hawkins, et al. (2003). "Robust singular value decomposition analysis of microarray data." Proc Natl Acad Sci USA **100**(23): 13167-72.

Mahalanobis, P. C. (1936). "On the generalised distance in statistics " Proceedings of the National Institute of Sciences of India **2**: 49–55.

Marsella, L., F. Sirocco, et al. (2009). "REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform." Bioinformatics **25**(12): i289-95.

Martin, L. C., G. B. Gloor, et al. (2005). "Using information theory to search for co-evolving residues in proteins." Bioinformatics **21**(22): 4116-24.

Mundra, P., M. Kumar, et al. (2007). "Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM." Pattern Recognition Letters **28**(13): 1610-1615.

Duda, O. R., P. E. Hart, et al. (2001). "Pattern classification." 654.

Opiyo, S. O. and E. N. Moriyama (2007). "Protein family classification with partial least squares." J Proteome Res **6**(2): 846-53.

Owzar, K., W. T. Barry, et al. (2008). "Statistical challenges in preprocessing in microarray experiments in cancer." Clin Cancer Res **14**(19): 5959-66.

Peterson, L. E. (2002). "Factor analysis of cluster-specific gene expression levels from cDNA microarrays." Computer methods and programs in biomedicine **69**(3): 179-88.

Ransohoff, D. F. (2005). "Bias as a threat to the validity of cancer molecular-marker research." Nat Rev Cancer **5**(2): 142-9.

Rao, D. C. and C. Gu (2001). "False positives and false negatives in genome scans." Adv Genet **42**: 487-98.

Rattray, M., X. Liu, et al. (2006). "Propagating uncertainty in microarray data analysis." Brief Bioinformatics **7**(1): 37-47.

Rice, T. K., N. J. Schork, et al. (2008). "Methods for handling multiple testing." Adv Genet **60**: 293-308.

Sandberg, M., L. Eriksson, et al. (1998). "New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids." J Med Chem **41**(14): 2481-91.

Schneider, T. D., G. D. Stormo, et al. (1986). "Information content of binding sites on nucleotide sequences." Journal of Molecular Biology **188**(3): 415-31.

Scholz, M. and J. Selbig (2007). "Visualization and analysis of molecular data." Methods Mol Biol **358**: 87-104.

Shannon, C. E. (1948). "A Mathematical Theory of Communication." Bell System Technical Journal **27**: 379–423, 623–656.

Shlens (2009). "A tutorial on PCA." 12.

Swets, J. A. (1988). "Measuring the accuracy of diagnostic systems." Science **240**(4857): 1285-93.

Vinga, S. and J. S. Almeida (2007). "Local Renyi entropic profiles of DNA sequences." BMC Bioinformatics **8**: 393.

Wall, M. E., A. Rechtsteiner, et al. (2003). Singular value decomposition and principal component analysis. Kluwer, Norwell, MA

Wang, Y., D. Miller, et al. (2008). "Approaches to working in high-dimensional data spaces: gene expression microarrays." Br J Cancer **98**(6): 1023-1028.

Wollenberg, K. R. and W. R. Atchley (2000). "Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap." Proc Natl Acad Sci USA **97**(7): 3288-91.

Zeng, L., J. Wu, et al. (2008). "Statistical methods in integrative analysis for gene regulatory modules." Statistical applications in genetics and molecular biology **7**(1): Article 28.

Zwick, W. and F. Velicer (1986). "Comparison of Five Rules for Determining the Number of Components to Retain." Psychological Bulletin **99**: 432-442.