A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors

Sebastian J. Maerkt1,2 and Stephen R. Quake2,*

A major goal of systems biology is to predict the function of biological networks. Although network topologies have been successfully determined in many cases, the quantitative parameters governing these networks generally have not. Measuring affinities of molecular interactions in high-throughput format remains problematic, especially for transient and low-affinity interactions. We describe a high-throughput microfluidic platform that measures such properties on the basis of mechanical trapping of molecular interactions. With this platform we characterized DNA binding energy landscapes for four eukaryotic transcription factors; these landscapes were used to test basic assumptions about transcription factor binding and to predict their in vivo function.

Systems biology focuses on understanding the collective properties of biological networks; these networks in turn describe interactions between as many as thousands of unique elements. In recent years, knowledge of biological networks has grown dramatically, mainly because of the development and application of novel genomic (1–3) and proteomic methods (4–8) as well as bottom-up approaches such as genetic network engineering (9, 10). But as network topologies are becoming well characterized, information about the elements comprising these networks has remained minimal and in the realm of low-throughput biology. In order to model and predict the behavior of these complex systems, the underlying interactions between elements will have to be quantitatively characterized.

Quantifying the affinities of molecular interactions is a considerable technical challenge. First, any particular biological interaction is governed by a large number of variables. Therefore, obtaining equilibrium dissociation constants requires one to perform dozens of assays as the concentrations of various components are systematically varied, increasing the number of measurements needed in an already logistically challenging problem. A second and more fundamental problem is the fact that many molecular interactions are transient and exhibit nanomolar to micromolar affinities, leading to rapid loss of bound material or little bound material in the first place. These factors are problematic for high-throughput methods such as yeast two-hybrid (6) and tandem affinity purification mass spectrometry (4, 5), where transient interactions are frequently missed. Protein–protein (11) and protein–DNA binding microarrays (PBMs) (12–14) are especially susceptible because of their stringent wash requirements, causing rapid loss of weakly bound material. Protein arrays have been applied to quantify ligand–ErbB receptor inter-
actions (11), with off rates determined by surface plasmon resonance to be on the order of $10^{-4} \text{s}^{-1}$ (12). PBMIs have been applied in a semiquantitative manner to transcription factor (TF) motif analysis for high-affinity interactions, with off rates on the order of $10^{-7} \text{s}^{-1}$ (12–14).

We developed a high-throughput microfluidic platform (Fig. 1) capable of detecting low-affinity transient binding events on the basis of the mechanically induced trapping of molecular interactions (MITOMI), which eliminates the off-rate problem facing current array platforms and allows for absolute affinity measurements. We used MITOMI to map the binding energy landscapes of four eukaryotic TFs belonging to the basic helix-loop-helix (bHLH) family by collecting over 41,000 individual data points from more than 17 devices and covering titrations over 464 target DNA sequences. These binding energy topographies allowed us to (i) predict in vivo function for two yeast TFs, (ii) make a comprehensive test of the base additivity assumption, and (iii) test the hypothesis that the basic region alone determines binding specificity of bHLH TFs.

bHLH motifs represent the third largest TF family in eukaryotes and regulate a wide variety of cellular functions ranging from cell proliferation and development to metabolism (16). We studied isoforms A and B of the human TF MAX, which together with other bHLH members play a role in cellular proliferation and many cancers (17). We also studied the yeast TFs Pho4 and Cbf1p; the former regulates phosphate metabolism (18, 19), whereas the latter regulates methionine synthesis as well as chromosome segregation, serving a structural role in the kinetochore (20–22). bHLH TFs generally bind to a consensus sequence of 5'-CANNTG-3' (where N is any nucleotide) called enhancer box (E-box) (fig. S1, A and B), which was later found to be the second most conserved motif in higher eukaryotes (3). Members of the bHLH family show mid- to low nanomolar DNA binding affinities and have off rates above $10^{-2} \text{s}^{-1}$ for their consensus sequences (23, 24), with orders of magnitude higher off rates for nonconsensus sequences. This transience makes the use of conventional microarrays impractical.

The TF binding energy topographies were measured with highly integrated microfluidic devices (25, 26) containing 2400 independent unit cell experiments (Fig. 1, A and B). Each device is controlled by 7233 valves fabricated by multilayer soft lithography (MSL) (27) and programmed with a 2400-spot DNA microarray (28). The 2400 chambers are arranged into 24 rows addressed via a resistance equalizer (Fig. 1A); this ensures that flow velocities are equal across all rows, resulting in uniform surface derivatization and TF deposition. To avoid time-consuming cloning and protein synthesis and purification steps, we synthesized the TFs in situ via wheat germ–based in vitro transcription and translation (ITT). We designed a two-step polymerase chain reaction (PCR) method that generates linear expression-ready templates directly from yeast genomic DNA or cDNA clones (fig. S2). This approach allowed us to not only rapidly screen new TFs but also easily create and test structural chimeras.

We synthesized libraries of Cy5-labeled target DNA sequences that comprehensively cover the E-box motif and flanking bases by permuting up to four bases at a time (fig. S1C and data set 1). Dilution series for each target DNA sequence were spotted as microarrays with column and row pitches of 563 μm and 281 μm, respectively. These arrays were used to program the microfluidic devices by aligning

![Fig. 1.](image-url)
each spot to a unit cell. The ability to program devices with microarrays simplifies the microfluidic infrastructure and increases unit cell density. The use of microarrays for device programming is highly modular because any soluble substance or suspension may be arrayed, and it provides an elegant and efficient solution to the world-to-chip interface problem. Approximately attomoles of DNA and TF are required for each data point.

Each unit cell is controlled by three micromechanical valves (27) as well as a “button” membrane (fig. S3) used for surface derivatization and MITOMI (Fig. 1, B to I, and fig. S4). When the button is actuated, it physically blocks a 60-μm circular area on the slide, preventing molecules from entering or leaving that part of the surface. The contact area can be precisely modulated by choice of button diameter and closing pressures (fig. S3). During surface derivatization, a circular area is masked with the button, while the rest of the surface is passivated with biotinylated bovine serum albumin. When the button is released, the previously protected circular area is specifically functionalized with biotinylated antibody against His₅ (fig. S4, C to G). After surface patterning, the device is loaded with wheat germ–based ITT mixture containing linear DNA template coding for the TF to be synthesized, and each unit cell is isolated by closing a set of micromechanical valves. The device is incubated at 30°C for 90 min to complete TF synthesis, solvation of target DNA, and equilibration of TF and target DNA (fig. S4, H and I). After the incubation period, MITOMI is performed by again actuating the button membrane and trapping surface-bound complexes (Fig. 1, D to I, and fig. S4, J to L). Initial contact of the membrane with the surface occurs medially and extends radially outward. Radial closure prevents solvent pockets from forming between the two interfaces and effectively creates zero dead volume while preserving the equilibrium concentrations of the molecular interactions to be detected. The trapped molecules are subsequently quantified with a DNA array scanner (fig. S5). Device characterization and control experiments are described in the Supporting Online Material (figs. S6 to S8) (26); we determined the lower limit of detection to be K_d = 18 μM and established a global measurement error of 19% (fig. S9).

Our measurements agree with previous reports that the optimal binding sequence for all four TFs is CACGTG for N₃ to N₅ (Fig. 2, A to D, and figs. S10, A to D, and S11). We measured consensus binding affinities of 67.0 nM, 73.1 nM, 11.1 nM, and 16.6 nM for MAX isomorf A, MAX isomorf B, Pho4p, and Cbf1p, respectively. The binding affinity of MAX to a slightly different sequence has been measured independently and is in agreement with our result for that sequence (24). Each binding energy landscape exhibits topographic fine structures, such as affinity spikes for sequences with a one-base spacer between the two half sites (CACGGTG for example) as well as consensus neighbors CATGTG, CTCGGT, and CAGGGT. These fine structures often lie in the low-affinity regime (with off rates on the order of 2 to 20 s⁻¹) and have thus far not been observed with other methods. The binding energy landscapes for both MAX isomorfs are more rugged than the landscapes of Pho4p and Cbf1p, showing strong affinities for consensus neighbors, whereas Pho4p and Cbf1p are singularly specific for the E-box consensus. These differences in topography are intriguing, because crystal structures of truncated versions of MAX and Pho4p show that both TFs make essentially the same base-specific contacts (29, 30). Therefore, similar base-specific contacts give rise to recognition of the same consensus sequence but not necessarily to similar overall binding topographies.

Informatic methods for TF binding motif discovery often rely on ad hoc hypotheses such as the additivity assumption, which
posits that the energetic or informatic role of each base in a given motif is independent of the identity of neighboring bases. Evidence from small data sets contradicts this assumption (31, 32), but there is an ongoing debate on the importance of the observed non-independence (33). Our data on the absolute binding affinities of the MAX isoforms to all possible sequence permutations of one E-box half site plus a flanking base, or 256 sequences in total, allowed us to determine the extent of interdependence between individual base contacts. First, we generated a position weight matrix (PWM) for each isoform, consisting of changes in the Gibbs free energy for all 16 possible single-base substitutions. These PWMs were then used to calculate binding affinities to all 256 sequences. Because PWMs contain no information on higher-order interactions, the predicted affinities reflect an assumption of completely independent base contacts. Comparing our experimentally determined values with the predictions showed that only a small subset of values agreed (Fig. 2, E and F, and fig. S10, E and F). PWMs fail to predict low-affinity binding, because almost no sequences above 2.5 kcal/mol agreed with measured values (defined as lying within 2σ or ±0.4 kcal/mol of the measured value). This is mostly due to the increasing role of non-specific interactions, for which failure of the PWM is not surprising. More importantly, PWMs also fail for higher affinities, predicting only 44% of all sequences below 2.5 kcal/mol. The PWMs predicted only 56%, 10%, and 0% of all double, triple, and quadruple substitutions, respectively. The consequences of these results are twofold. First, for gene discovery applications PWMs will give low false positives but more potential false negatives or missed genes. Second, PWMs are not sufficient for more detailed computational approaches to systems biology that seek to understand the stochastic dynamics of TF binding.

To address the question of how Pho4p and Cbf1p serve distinct biological functions while recognizing seemingly identical consensus motifs, we measured the extent to which these TFs recognize bases flanking the E-box consensus. Current experimental and bioinformatic methods (1, 7, 34, 35) failed to show definite differences in sequence recognition between these two TFs (CAGGTG and GTACGTG for Pho4p and Cbf1p, respectively, where r is any purine and s is G or C). We measured all possible permutations of three flanking bases 5' as well as 3' of the consensus sequence in order to determine how far base-specific recognition extends. The results show that Pho4p specifically recognizes two flanking bases, extending the consensus motif to a palindrome decamer of 5'-CCCCACGTGGGGG-3' (Fig. 3A and fig. S12, A and B). For Cbf1p, the flanking base recognition profile differed drastically from Pho4p, preferring GT as N_sN_r (Fig. 3B) and the palindrome AC for N_sN_r (fig. S12, C to D). Furthermore T and C were preferred for N_r (A and G in terms of N_r), extending Cbf1p sequence recognition from the initial hexameric E-box motif to a dedecamer of 5'-(A/G)GTACGTGAC(T/C)-3'.

On the basis of structural (29, 30) and biochemical data, it is widely assumed that the sequence-specific binding of bHLH TFs is determined entirely by the basic region. We assessed whether the basic region itself is sufficient to produce the observed flanking base sequence specificity by cloning the basic regions of Pho4p and Cbf1p into the MAX isoform B backbone. The results show that, despite the existence of a few intriguing sequence outliers, the basic region itself is sufficient to transform the original MAX isoform B recognition pattern to patterns resembling Pho4p and Cbf1p (Fig. 3D).

Last, we asked whether the binding energy landscapes for Pho4p and Cbf1p are sufficient to predict which genes these TFs physically bind and therefore likely regulate. We applied a simple in silico model based on calculating a probability of occupancy (F_occ) (36, 37) for each regulatory sequence of 5814 yeast genes and obtained 38 and 24 genes bound by Pho4p and Cbf1p, respectively (Fig. 4, A and B, and data sets 2 and 3). We then tested whether these gene sets were significantly enriched for functions related to Pho4p (phosphate metabolism) and Cbf1p (methionine metabolism and chromosome segregation). The Pho4p data set showed significant enrichment in genes functioning in ionic homeostasis and phosphate metabolism (table S1), with 60% of the genes in the data set functioning in phosphate metabolism (37%), ionic homeostasis (13%), and vacuoles (10%) (Fig. 4A). Chromatin immunoprecipitation (ChIP)–chip and microarray experiments determining Pho4p-regulated genes have only 18% agreement; our data set includes all of these overlapping genes and has 40% agreement with at least one of the other data sets (Fig. 4E). For Cbf1p, the functional enrichment returned categories mainly involved in cell cycle and cell growth (table S1), implying that Cbf1p functions in chromosome segregation. Of Cbf1p-regulated genes, 30% are involved in chromosome structure, and another 25% are involved in budding (Fig. 4B). Only two genes (8%) regulate methionine synthesis. We also found three genes (STE20, GAL2, and DRS2) that

![Fig. 3. ΔΔG values of all permutations of the two flanking bases N_sN_r for Pho4p (A), Cbf1p (B), and MAX iso B (C). (D) Comparison of the ΔΔG values of the wild-type proteins shown in (A) to (C) with basic region chimeras in which the basic region of MAX isoform B was replaced by the basic regions of Pho4p, Cbf1p, and MAX iso B. The average of two experimental values is plotted, with the difference shown by the error bars.](/images/00000001.png)
had previously been shown to exhibit Cbf1p-dependent chromatin remodeling (20). We predict that Cbf1 regulates a much smaller set of genes than ChIP-chip experiments do, but there is good agreement because more than 60% of the genes in our prediction were also found by ChIP-chip (Fig. 4F). We have thus shown that augmenting the consensus binding motifs with the new flanking bases revealed by the free energy landscapes results in concrete biological predictions that differentiate the function of Pho4 and Cbf1 and have broad agreement with the biological literature.

For two yeast TFs we have successfully predicted biological function by combining purely in vitro biophysical measurements with informative knowledge of the organism’s genome. As the theoretical and experimental tools of systems biology become more mature, this situation may become the rule rather than the exception, and it is worth considering whether there is any fundamental limit on the ability to predict biological behavior from in vitro measurements.

References and Notes
2. P. Clifton et al., Science 301, 71 (2003); published online 29 May 2003 (10.1126/science.1084337).
15. C. Ponzetto et al., Cell 77, 261 (1994).
26. Materials and methods are available on Science Online.
38. We thank B. Wold and J. Berger for helpful discussions and K. Kapur and C. Jacobs for help with the document. Supported in part by Defense Advanced Research Projects Agency-Biological Input Output Systems, an NIH Director’s Pioneer Award, and Howard Hughes Medical Institute. S.R.Q. holds equity in, is a paid consultant for, and is a co-founder and former director of Fluidigm, which may license the device described in this paper.

Supporting Online Material
www.sciencemag.org/cgi/content/full/315/5809/233/DC1
Materials and Methods
Figs. S1 to S12
Table S1
References
Data sets 1 to 3
7 June 2006; accepted 20 November 2006
10.1126/science.1131007