BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 2 2007, pages 134–141
doi:10.1093/bioinformatics/btl565

Genome analysis

Predicting transcription factor affinities to DNA from a biophysical model

Helge G. Roider, Aditi Kanhere, Thomas Manke and Martin Vingron*
Max-Planck-Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Received on October 1, 2006; revised and accepted on November 6, 2006
Advance Access publication November 10, 2006
Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: Theoretical efforts to understand the regulation of gene expression are traditionally centered around the identification of transcription factor binding sites at specific DNA positions. More recently these efforts have been supplemented by experimental data for relative binding affinities of proteins to longer intergenic sequences. The question arises to what extent these two approaches converge. In this paper, we adopt a physical binding model to predict the relative binding affinity of a transcription factor for a given sequence.

Results: We find that a significant fraction of genome-wide binding data in yeast can be accounted for by simple count matrices and a physical model with only two parameters. We demonstrate that our approach is both conceptually and practically more powerful than traditional methods, which require selection of a cutoff. Our analysis yields biologically meaningful parameters, suitable for predicting relative binding affinities in the absence of experimental binding data.

Availability: The C source code for our TRAP program is freely available for non-commercial use at http://www.molgen.mpg.de/~manke/papers/TFaffinities/

Contact: vingron@molgen.mpg.de

1 INTRODUCTION

Protein–DNA interactions play a fundamental role in transcriptional gene regulation. For individual sequences, these interactions have been studied for a long time using a variety of experimental techniques, such as DNAse footprinting (Galas and Schmitz, 1978) and gel-shift assays (Fried and Crothers, 1981). Recently, functional genomics technology has opened up the way towards unraveling protein–DNA interactions on a global scale. In particular the group of Rick Young has pioneered the genome-wide application of chromatin-immuno precipitation for a comprehensive list of transcription factors in *Saccharomyces cerevisiae* (Lee *et al*., 2002; Harbison *et al*., 2004). In this technique the bound and unbound sequence fragments are labeled with red and green dye respectively and are then simultaneously hybridized onto an array (ChIP-chip). The relative intensities from the two channels (R/G ratios) provide a quantitative estimate for the binding affinities of a transcription factor to all sequence regions of interest *in vivo*. Generally, the measured affinities depend on the cellular condition in which the binding of the transcription factor (TF) is tested. Such alterations can be due to differences in protein concentrations and DNA accessibility. Therefore a complementary approach of protein binding microarrays (PBMs) has been developed by Martha Bulyk and her collaborators (Mukherjee *et al*., 2004). It allows to quantify the relative affinities of a TF to accessible, double-stranded DNA *in vitro*, again in terms of R/G ratios.

Where experimental data do not suffice yet to determine e.g. whether a particular transcription factor binds a target gene, theoretical considerations have to fill the gap. The groundbreaking work by von Hippel and Berg (1986) provided the rationale for converting the biophysical problem of TF–DNA affinity into a pattern matching and pattern discovery problem (Stormo 2000; D'haeseleer, 2006; Djordjevic *et al*., 2003). Following these ideas, the preferential binding of some TFs to certain DNA sequences can be expressed in terms of a sequence motif or a position specific score matrix (Wasserman and Sandelin, 2004), which is derived from a set of known high-affinity binding sequences. For a stretch of DNA, such a description assigns a score to every site in the sequence depending on its similarity to the motif. Traditionally, statistical considerations are then used to define a score threshold which needs to be exceeded in order for a site to be reported as a hit (Rahmann *et al*., 2003). Such hit-based methods cement the binary separation between binding and non-binding, in contrast to the physical behavior of TFs. It has therefore been difficult to rationalize the binding affinities measured with the ChIP-chip and PBM technologies using the motif-matching approach described above.

In this paper, we put forward a method for predicting the binding affinity of a TF to a DNA sequence of interest. Our probabilistic framework is closer in spirit to the original work by Berg and von Hippel (1987) and circumvents the need for a threshold on both the experimental data and our predictions. As a measure of relative affinity we use the expected number of TFs bound to a DNA sequence. Our TRanscription factor Affinity Prediction (TRAP) tool which calculates this quantity takes as input a matrix description of a given TF and a set of DNA sequences to be annotated. It requires the specification of only two parameters, λ and R_0. Here we draw upon the large scale ChIP-chip and PBM datasets to calibrate these two parameters. Strikingly, the relative binding affinities predicted by TRAP are rather insensitive even to sizeable variations in the parameter values and reveal interesting information about the driving forces in protein–DNA interactions. This enables us to provide a general prescription for λ and R_0. Once tuned, this

*To whom correspondence should be addressed.

model allows the prediction of relative binding affinities also in the absence of large-scale binding data.

Recent work by Tanay (2006) has also advocated an affinity-based approach to TF binding. He and also Foat *et al*. (2006) have developed methods, to derive optimal scoring matrices for individual TFs given binding data from ChIP-chip. Our aim is not to derive TF representations, which correlate optimally with ChIP-chip data, but rather to optimize a generic physical model which can rationalize the binding data for all TFs. This is also in contrast to Granek and Clarke (2005), who utilize a physical model, but do not provide a rationale for choosing the parameters.

Our approach of predicting binding affinities has a number of advantages over traditional hit-based methods. Most notably, TRAP provides a natural ranking of sequences with respect to a particular TF of interest or conversly the ranking of several TFs with respect to one sequence. Finally, we compare our results with traditional approaches and find that it has higher predictive power over experimental binding ratios than the hit-based methods.

2 METHODS

2.1 Protein–DNA binding data

In this work we utilize the genome-wide dataset on *in vivo* protein–DNA interactions in *S.cerevisae* (Harbison *et al*., 2004). The authors provide a list of binding ratios (R/G-ratios) for all intergenic regions in yeast, which we obtained from their website. For comparison, we also retrieved binding data for three TFs (Rap1, Mig1 and Abf1) from a complementary study by Mukherjee *et al*. (2004), who use protein binding microarrays to determine binding affinities *in vitro*. For each dataset the authors suggest a *P*-value threshold of 0.001 to discriminate between binding and non-binding which we utilize for parts of the analysis.

2.2 Binding site descriptions

As motif descriptions we use the set of 29 curated yeast matrices (for 25 factors) provided by the TRANSFAC database (Matys *et al*., 2003) for which ChIP-chip data are available. We add a pseudo-count of $\pi = 1$ to each element in the count matrices (Bucher, 1990). This modification can be interpreted in statistical terms as setting the estimated number of unobserved base pair occurrences, or physically, as setting a maximally allowed contribution to the mismatch energy. For comparative purposes we also set $\pi = 0.5$, but our results are unaffected by such a change.

2.3 A simple model of protein–DNA interactions

Assuming that the complex formation of a transcription factor TF with a sequence site S is at equilibrium, $TF + S \Leftrightarrow TF \cdot S$, the fraction of bound sites is given by Zumdahl (1998)

$$p(S) = \frac{[TF \cdot S]}{[S] + [TF \cdot S]} = \frac{K \cdot [TF]}{1 + K \cdot [TF]}. \tag{1}$$

Here the squared brackets denote the activities of TF and sequence and $K = K(S)$ is the site-specific equilibrium constant. In the following we measure all equilibrium constants relative to the one for the site with the highest affinity, S_0, to which we conventionally assign the energy $E = 0$

$$K(S) = K(S_0) e^{-\beta E(S)}, \tag{2}$$

where $1/\beta = k_B T$ denotes temperature times Boltzmann constant. Now Equation (1) can be rewritten as

$$p(S) = \frac{R_0\ e^{-\beta E(S)}}{1 + R_0\ e^{-\beta E(S)}}. \tag{3}$$

This makes the two unknown dimensionless parameters $R_0 = K(S_0) \cdot [TF]$ and $\beta E(S)$ explicit. Berg and von Hippel showed that the mismatch energy, $E(S)$, can be written in terms of a TF-specific matrix, $M = (m_{i,\alpha})$, which summarizes the observed base pair counts of known TF binding sites Berg and von Hippel (1987).

$$\beta E(S) = \frac{1}{\lambda} \sum_{i=1}^{W} \sum_{\alpha = A, C, G, T} S_i^{\alpha} \log\left(\frac{m_{i,\,\max}}{m_{i,\,\alpha}}\ b_{i,\,\alpha} \right). \tag{4}$$

Here the summation is over all W positions i in the count matrix $(m_{i,\,\alpha})$ and $S_i^{\alpha} = 1$ only if the sequence has base pair α at position i, and zero otherwise. For each position the matrix element with maximal count is denoted as $m_{i,\max}$. This also defines the consensus sequence, for which every term in the above sum vanishes, such that $\beta E(S) = 0$. Finally, we include a background-dependent term, $b_{i,\,\alpha}$, which denotes the relative background frequency of the observed nucleotide α with respect to the background frequency of the most frequent nucleotide in the motif at the given position. For a given TF, Equation (4) effectively replaces the large set of unknown binding energies, $\beta E(S)$, by a predefined motif matrix and a single parameter λ, which is introduced to scale the mismatch energies in units of thermal energy. This parameter depends on the TF of interest and determines how strongly variations in the target sequence will be penalized. This completes the reduction of the large parameter space to only two sequence-independent parameters (R_0,λ).

As a measure of relative affinity our TRAP program predicts for a given TF matrix of length W and a given DNA sequence of length L the expected number $\langle N \rangle$ of bound transcription factor molecules. This quantity is computed as the sum of contributions from all possible sites l in the sequence of interest

$$\langle N \rangle = \sum_{l=1}^{L-W} p_l = \sum_{l=1}^{L-W} \frac{R_0\ e^{-\beta E_l(\lambda)}}{1 + R_0\ e^{-\beta E_l(\lambda)}}. \tag{5}$$

To account for competitive binding of a given factor to the same site, but different strands (S_l, $\bar{S}_l$), we used

$$p_l = p(S_l) + p(\bar{S}_l) \rightarrow p(S_l) + p(\bar{S}_l) - p(S_l) \times p(\bar{S}_l). \tag{6}$$

The correction term will be of importance only if both $p(S_l)$ and $p(\bar{S}_l)$ are large, i.e. for palindromic motifs. In general one could invoke more elaborate dynamic programming techniques, as used by Rajewsky *et al*. (2002), to account for preclusion effects from competing factors and self-overlapping binding sites. However, such effects will be small for our analysis, in which we treat all TFs separately. They are likely to be more pronounced when multiple TFs compete for the same sequence. We leave such a treatment to future analysis (Chung *et al*., manuscript in preparation).

Parameter determination

To calibrate the parameters R_0 and λ for a given TF and cellular condition we apply Equation (5) to all 6700 intergenic regions in yeast. This results in 6700 predicted occupancies $\langle N \rangle$, which can be correlated with the measured R/G ratios from ChIP-chip experiments. We use the Pearson correlation coefficient, r, to quantify how well the model describes the experimental binding affinity and to determine the optimal parameters. To assume a linear correlation between the R/G ratios and $\langle N \rangle$ is plausible if the efficiency of the pulldown reaction (ChIP) is small and not yet saturated. This is supported by the absence of any apparent upper limit on the measured R/G ratios. We have tested the range of parameter values $\lambda = 0.05, 0.10, \ldots, 2.00$ and $\ln R_0 = -10, -8, \ldots, 30$ for all TFs and all tested conditions. We take those parameters which yield the highest correlation coefficient as optimal, in the sense that with this choice of R_0 and λ the model describes the actual binding data best.

3 RESULTS

3.1 Screening the parameter space

As a measure of affinity, TRAP predicts for a given TF and a given DNA sequence the expected number, $\langle N \rangle$, of bound TF molecules.

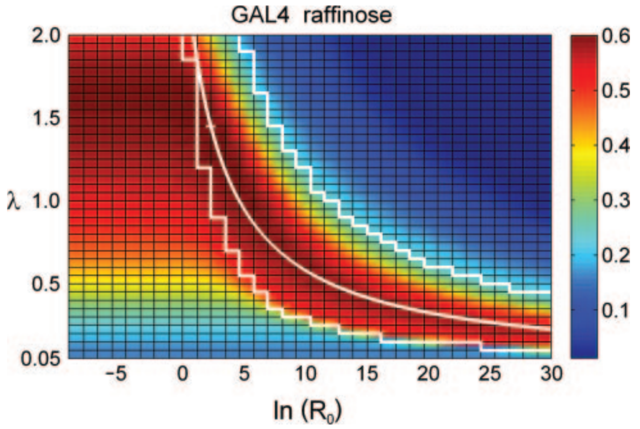

Fig. 1. Correlation Analysis for Gal4. For each parameter combination ($\ln R_0, \lambda$) TRAP results for $\langle N\rangle$ show a certain correlation with the experimental R/G ratios. We quantify this correlation by a Pearson coefficient, r, which is color-coded as specified in the sidebar. The optimal choice of parameters, with the highest correlation coefficient, is marked by a white cross and the hyperbola highlights a line of parameter combinations with similarly high correlation coefficient. We also indicate the boundary (white staggered lines) for which the maximal value of $\langle N\rangle$ (over all intergenic regions) lies between 0.5 and 5.

This calculation requires the setting of two parameters. The first parameter, R_0, involves the factor concentration and the equilibrium constant of the binding reaction between the TF and its optimal binding site. The second parameter, λ, scales the mismatch energy of a given site in the sequence with respect to the optimal binding site. For any given combination of R_0 and λ the correlation between $\langle N\rangle$ of every intergenic region and the corresponding experimental R/G ratios can be determined.

First we analyze the generic features of the correlation coefficient across the parameter space spanned by R_0 and λ. These features are illustrated by the example shown in Figure 1. For large λ (small mismatch energy) as well as for large R_0 almost every site in the sequence is occupied and the number of expected TFs bound will simply correlate with the length of the intergenic region. This typically results in a poor correlation with the observed R/G-ratios as can be seen in the upper right part of Figure 1.

For small λ (large mismatch energy) only the optimal sites, $E = 0$, will have a non-vanishing binding affinity. However, most TFs can accommodate certain variations in the binding site (Mossing and Record, 1985), therefore in most cases we would not expect to observe the best correlation with experimental data in a region of the parameter space which does not permit a certain degree of binding site flexibility. This is in line with our observation in Figure 1, where the correlation coefficient decreases for $\lambda \to 0$.

For small R_0 the expected number of bound TFs depends linearly on R_0, as can be inferred from a Taylor expansion of Equation (5) around $R_0 = 0$

$$\langle N\rangle \approx R_0 \sum_l \exp(-\beta E_l). \tag{7}$$

Therefore, changes of R_0 in this regime only affect the absolute number of $\langle N\rangle$, but not the correlation of $\langle N\rangle$ with R/G ratios. In Figure 1 this is reflected by a constant correlation coefficient for $\ln R_0 < 0$ and a given λ.

It is evident from Equation (5) that the affinity of a single binding site can be kept constant for varying values of R_0 and βE in such a way that $\ln R_0 - \beta E = c$. With $\beta E \propto 1/\lambda$, we find the hyperbolic relation $\lambda \propto 1/(\ln R_0 - c)$. Interestingly the characteristic curves of constant correlation coefficients seen in Figure 1, which can be well described by a hyperbola, suggest that this generic behavior is effectively reflected in the behavior of the correlation coefficients.

Optimal parameter choice derived from experimental data

Binding data from PBM constitutes the ideal benchmark for TRAP. We thus determined the optimal model parameters for Abf1, Mig1 and Rap1 whose binding affinities have been studied experimentally using protein binding arrays (Mukherjee *et al*., 2004) and for which we obtained matrix descriptions from TRANSFAC (Matys *et al*., 2003). In all cases we can find optimal parameters which yield highly significant correlation ($r > 0.5$), as shown in Table 1. This indicates that TRAP can successfully account for much of the observed *in vitro* binding affinities.

We proceed to a more comprehensive set of 25 TFs (29 matrices), for which matrix descriptions exist (Matys *et al*., 2003) and R/G ratios have been obtained from ChIP-chip for one or more cellular conditions (Harbison *et al*., 2004). This *in vivo* data corresponds to a more complicated situation, where we cannot always assume that the TF is available for DNA binding and that the DNA is accessible under the tested condition. Despite these caveats, we observe that TRAP still predicts a large fraction of *in vivo* affinities for properly chosen parameters. In Table 1 we present our results for a group of 15 matrices for which our affinity predictions show high correlation (Pearson $r > 0.3$) with the experimentally observed R/G ratios. We provide a complete list for all 25 factors and 13 conditions as Supplementary table.

Remarkably, the optimal parameters for all factors and conditions correspond to maximal values of $\langle N\rangle$ (over all intergenic regions) in the range of 0.5...5. This is biologically reasonable assuming that each transcription factor should recognize some promoter region, at least in one condition. $\langle N\rangle_{\max}$ falls outside of this range only in the case of Hap1, where the 'optimal' R_0 is small and poorly defined in the sense explained below Equation (7), and Rap1, where several sequences have large clusters of neighboring Rap1 binding sites (Gilson *et al*., 1993).

Notice that the observed correlations are actually quite insensitive to the precise value of the parameters and some of our modeling assumptions. We also investigated the rank order of different intergenic regions with respect to their predicted affinities. While the absolute value for $\langle N\rangle$ depends on the values of ($\ln R_0$, λ), we find that the ranking of intergenic regions remains largely unaffected even under sizeable changes in these parameters. Comparing the ranks of the TRAP results for optimal parameters with those obtained from a 30% decrease in λ, we find Spearman rank correlation coefficients larger than 0.98. Similarly, an almost 100-fold change in R_0 gave a correlation coefficient above 0.99.

Parameter choice in the absence of experimental data

While it is possible to determine the optimal coefficients (R_0, λ) in the presence of sufficient binding data, it is clearly desirable to have some prescription which would allow the parameter determination on general grounds. Based on the results in Table 1 and the observed insensitivity to small changes in the parameters

Table 1. Correlation analysis

Matrix	Condition	W	λ	ln R_0	$\langle N\rangle$	r	r_{pred}
ABF1_01	Rich medium	22	0.60	8.11	2.95	0.5672	0.5634
	In vitro	22	0.65	6.91	2.52	0.5526	0.5452
ABF_C	Rich medium	15	0.45	4.61	3.08	0.5863	0.5618
	In vitro	15	0.50	3.51	2.37	0.5694	0.5426
CBF1_B	Rich medium	10	0.75	0.00	1.23	0.4272	0.4269
	AA depleted	10	0.45	3.51	2.90	0.6836	0.6736
GAL4_01	Rich medium	23	0.40	13.82	2.99	0.5593	0.5567
	Galactose	23	0.25	25.33	3.00	0.3355	0.3263
	Raffinose	23	1.45	2.30	1.67	0.6051	0.5897
GAL4_C	Rich medium	22	0.65	8.11	3.33	0.5730	0.5721
	Galactose	22	0.25	26.53	4.13	0.3395	0.3150
	Raffinose	22	1.30	3.51	2.53	0.6240	0.6013
GCN4_01	AA depleted	27	0.50	15.02	2.31	0.3406	0.1496
	Rapamycin	27	0.60	15.02	2.31	0.3123	0.1416
GCN4_C	AA depleted	10	0.50	0.00	1.35	0.3519	0.3206
	Rapamycin	10	0.50	0.00	1.35	0.3508	0.3125
HAP1_B	Rich medium	14	0.75	−9.21	0.004	0.3503	0.3191
HSF_04	High H_2O_2	15	0.90	4.61	2.77	0.4881	0.4165
	Low H_2O_2	15	0.80	4.61	2.66	0.4803	0.4380
LEU3_B	AA depleted	14	1.20	0.00	0.68	0.3354	0.3104
MCM1_02	Rich medium	27	1.70	3.51	0.93	0.3155	0.3093
	αfactor	27	1.45	4.61	0.97	0.3684	0.3561
MIG1_01	*In vitro*	17	0.90	2.30	1.23	0.5958	0.5907
RAP1_C	Rich medium	14	0.60	6.91	12.51	0.3818	0.3366
	AA depleted	14	0.35	12.72	12.99	0.4403	0.3700
	In vitro	14	0.15	13.82	5.23	0.4445	0.4321
RCS1_Q2	Low H_2O_2	13	0.05	21.93	1.01	0.3875	0.3205
REB1_B	Rich medium	9	0.45	1.20	1.25	0.5153	0.5058
	High H_2O_2	9	0.55	2.30	2.26	0.3245	0.3117
	Low H_2O_2	9	0.50	1.20	1.37	0.5973	0.5957

The first column denotes the TRANSFAC matrix identifier of those TFs, for which our theoretical estimates have a high correlation ($r > 0.3$) with the genome-wide R/G-ratios from ChIP-chip in at least one condition (second column). The width, W, of the matrix is given in the third column. In column 4 and 5 we give the optimal parameters λ and R_0 which result in the maximal Pearson correlation coefficient (r) and some maximal value of $\langle N\rangle_{\max}$ over all intergenic regions. The last column denotes the correlation coefficient that is predicted from using $\lambda = 0.7$ and R_0 from the regression analysis of Figure 2. It is apparent that in most cases the differences are small.

we decided to fix λ to an average value of 0.7 for all TFs and all conditions. This fixation reduces the parameter space to only R_0. We observe that the optimal values of $R_0(\lambda = 0.7)$ can be well described as a function of the motif width, W, with only small changes due to condition dependent effects. This is shown in Figure 2 where we perform a regression analysis of ln R_0 against W. The regression line allows us to determine R_0 for any given W and provides the basis for our subsequent analysis. R_0 will also vary with the cellular condition, through changes in TF-concentration, but empirically we find that this amounts to much smaller shifts compared to the overall dependence on the motif length. This can be understood since R_0 depends only linearly on the concentration, $[TF]$, but exponentially on the difference between the free energies of the best binding complex and the unbound state. This difference increases with the width of the binding site through an increasing number of protein–DNA contacts and dominates the behavior of R_0 as shown in Figure 2.

It should be noted that matrices can contain unspecific positions which then define an arbitrary consensus site with spuriously low binding energy. This can lead to an overestimate of the 'optimal' R_0 as observed in case of GCN4_01 (27 bp). For identical λ, GCN4_01 gives a vastly larger estimate of R_0 compared to GCN4_C (10 bp). The problem could be addressed by restricting the motif to positions with higher information content (e.g. $\geq$0.2 bits). For GCN4_01 this would reduce the motif length to 11 bases and in turn improve the results for this matrix (maximal $r \approx 0.57$ and $r_{\text{predicted}} \approx 0.53$ compared to the values in Table 1). Although the regression in Figure 2 could be further improved by these corrections we find that it is not very sensitive to such influences and in the following we thus proceed by using only the unmodified TRANSFAC matrices.

Assuming the parameter prescription $[R_0(W), \lambda = 0.7]$, we find correlations with the R/G-ratios that are almost as high as the optimal correlations (last column of Table 1) with exception of GCN4_01. This choice of (R_0, λ) may be used to predict relative binding affinities for TFs with known motifs in the absence of genome-wide binding data.

Comparison of TRAP with hit-based methods

Traditionally, computational target predictions have focused on the identification of individual binding sites with more or less specific sequence patterns. This is usually done by scanning a score matrix along the sequence and assigning a 'hit', whenever the score exceeds some pre-defined threshold (Wasserman and Sandelin, 2004). Of course, traditional methods suffer from the arbitrariness

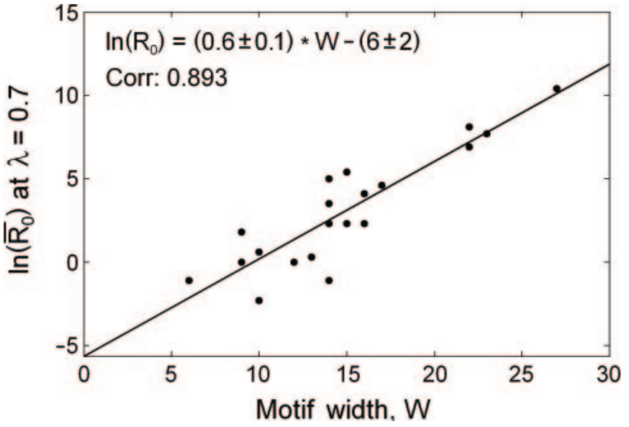

Fig. 2. Deriving a general prescription for R_0. For each matrix we plot the optimal value of $\ln R_0$ for fixed value of $\lambda = 0.7$. In cases where we have R/G ratios for more than one condition, we plot the average of the optimal $\ln R_0$. Deviations from this value, due to condition-dependent (TF-concentration-dependent) variation, are generally small (maximally $\ln R \pm 1$). The P-value of the correlation is 1.2×10^{-7}. The errors in the regression formula denote the 95% confidence interval on the regression parameters.

of that threshold. In contrast, our focus is on the determination of relative binding strength and the expected number of bound TFs. Therefore, it is difficult to compare our affinity-based method and hit-based methods without adjusting one or the other.

Here we consider two commonly used hit-based methods and compare them to TRAP [using the predefined parameters $R_0(W), \lambda = 0.7$] with respect to their capability of predicting experimental binding ratios. The first traditional method, which we call 'balanced method', invokes a score threshold which is chosen such that the expected number of false positive hits is balanced by the expected number of false negatives (Rahmann *et al.*, 2003). For each sequence this method calculates a number of hits which can be compared to experimental binding ratios and our predictions for the expected number of bound TFs. This comparison is illustrated for Leu3 in Figure 3, where it can be seen that the TRAP approach leads to a better correlation with experimental data. For a second comparison, we also consider a different threshold prescription, called '5FP', in which the expected false-positive rate is arbitrarily set to 5%. In Table 2 we provide a complete comparison of all the methods described above.

It can be seen that in ≈80% of cases TRAP results in better correlations with experimental binding ratios than the hit-based methods.

Alternatively, one may also impose a cutoff on the expected counts $\langle N \rangle$ to (arbitrarily) discriminate between bound and unbound sequences. Also experimental binding data are often interpreted in such a binary way, where binding ratios are converted to P-values and only sequences with e.g. $P < 0.001$ are considered as bound (Harbison *et al.*, 2004). For varying thresholds on $\langle N \rangle$ and a given cutoff on R/G ratios we can then calculate different sensitivities and specificities which are evaluated in a ROC-curve analysis. In Table 2 we use the area under the ROC-curve as a quality measure. Most areas are much larger than 0.5, indicating a strong predictive power of this method over experimental binding data in yeast at the significance threshold of $P < 0.001$.

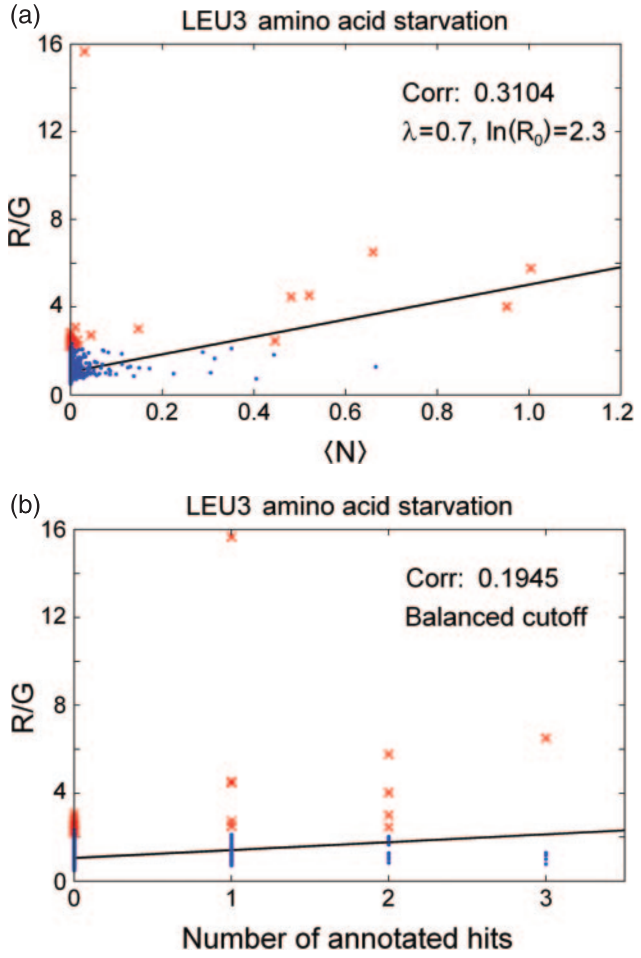

Fig. 3. Comparison of methods. As an example, we compare the results for Leu3 (in amino acid starved condition) from TRAP (left figure) with the results obtained from a balanced cutoff method (right figure). Sequences with significant R/G ratios ($P < 0.001$) are indicated by a cross. It is apparent that in this case TRAP improves the correlation with R/G ratios and the significant ChIP-chip targets.

To compare again with hit-based methods, we took, for each TF and every intergenic region, the number of hits (for 5FP and balanced cutoff) and performed the same ROC-curve analysis based on different thresholds on this score. Again we find that our method performs consistently better than hit-based approaches (see Table 2). On the entire set of 29 matrices we find that TRAP yields a ROC curve area of ≥0.7 for 22 matrices in at least one of the experimentally tested condition as opposed to only 16 and 14 matrices for the balanced and 5FP cutoff methods, respectively (see Supplementary Table).

Prediction of TFs with high affinity

The above analysis shows that for a given TF we can successfully rank sequences according to their expected affinity. Here we address the complementary question: given a certain sequence, can TRAP successfully rank TFs in accordance with ChIP-chip experiments using our prescription $R_0(W), \lambda = 0.7$. In general factors bound to a given sequence in the ChIP experiment should have higher values

Table 2. Comparsion of annotation methods

		Pearson correlation coefficient			ROC-curve area		
Matrix	Condition	TRAP	5FP	Bal	TRAP	5FP	Bal
ABF1_01	Rich medium	0.5634	0.5106	0.5006	0.9239	0.8683	0.8709
	In vitro	0.5452	0.5062	0.4972	0.8939	0.8476	0.8510
ABF_C	Rich medium	0.5618	0.5797	0.5576	0.9324	0.9207	0.9201
	In vitro	0.5426	0.5435	0.5282	0.8962	0.8539	0.8691
CBF1_B	Rich medium	0.4269	0.3026	0.2779	0.9942	0.9780	0.9750
	AA depleted	0.6736	0.5237	0.4872	0.8864	0.8303	0.8325
GAL4_01	Rich medium	0.5567	0.2871	0.2697	0.6780	0.6337	0.6320
	Galactose	0.3263	0.1912	0.1803	0.5840	0.6413	0.6393
	Raffinose	0.5897	0.3319	0.3149	0.7160	0.6570	0.6550
GAL4_C	Rich medium	0.5721	0.3267	0.3267	0.6773	0.6362	0.6362
	Galactose	0.3150	0.2143	0.2143	0.5757	0.6605	0.6605
	Raffinose	0.6013	0.3952	0.3952	0.7261	0.6767	0.6767
GCN4_01	AA depleted	0.1496	0.3806	0.3966	0.8006	0.7498	0.6907
	Rapamycin	0.1416	0.3912	0.4084	0.8069	0.8016	0.7300
GCN4_C	AA depleted	0.3206	0.2091	0.2476	0.7711	0.6486	0.7199
	Rapamycin	0.3125	0.2138	0.2510	0.7837	0.6621	0.7640
HAP1_B	Rich medium	0.3191	0.2514	0.2189	0.8084	0.6557	0.6866
HSF_04	High H_2O_2	0.4165	0.2598	0.2416	0.7526	0.6563	0.6620
	Low H_2O_2	0.4380	0.2322	0.2185	0.7885	0.6949	0.7010
LEU3_B	AA depleted	0.3104	0.2088	0.1945	0.6978	0.6486	0.6623
MCM1_02	Rich medium	0.3093	0.1090	0.1438	0.8066	0.7162	0.6344
	αfactor	0.3561	0.0997	0.1570	0.8614	0.7712	0.7007
MIG1_01	*In vitro*	0.5907	0.4625	0.4433	0.8793	0.6982	0.7043
RAP1_C	Rich medium	0.3366	0.3513	0.3282	0.9085	0.7807	0.7767
	AA depleted	0.3700	0.4022	0.3799	N/A	N/A	N/A
	In vitro	0.4321	0.3647	0.3402	0.8862	0.6980	0.7155
RCS1_Q2	Low H_2O_2	0.3205	0.1578	0.1794	0.5470	0.4985	0.5043
REB1_B	Rich medium	0.5058	0.4197	0.3115	0.9289	0.8967	0.8750
	High H_2O_2	0.3117	0.3390	0.2836	0.8437	0.8534	0.8191
	Low H_2O_2	0.5957	0.5156	0.4060	N/A	N/A	N/A

Here we present the results from our correlation analysis and the ROC-curve areas. For the latter we invoke a P-value threshold of 10^{-3}. TRAP denotes results from our threshold-free calculation of the expected count, which should be compared to several traditional methods (Bal = balanced cutoff, 5FP = 5% expected false positives). With N/A we denote those cases for which the TF does not have any targets in the specified condition at $P < 10^{-3}$.

of $\langle N \rangle$ predicted by TRAP than unbound factors. Since experimental R/G ratios for different factors are not directly comparable, we follow again the binding prescription as given in (Harbison *et al*., 2004; Mukherjee *et al*., 2004) and distinguish binders from non-binders according to the P-value threshold of $P = 0.001$.

Figure 4 shows as an example the intergenic region between GAL1 and GAL10 with its experimentally verified high affinity sites for Gal4 and Mig1 (Selleck and Majors, 1987; Frolova *et al*., 1999). This region was also significantly enriched in the ChIP-chip pulldown experiment with GAL4 and in the PBM experiment with MIG1. None of the other 23 factors in our set had been predicted as a target by ChIP-chip or PBM. As can be seen, TRAP predicts the highest affinities for Gal4 followed by Mig1, Adr1 and Ste12. All other factors have only negligible affinities predicted in good agreement with the ChIP-chip experiments. Interestingly independent chromatin precipitation experiments have shown that Ste12 has weak but measurable affinity to the GAL1–GAL10 intergenic region (Reeves and Hahn, 2005). The balanced cutoff method also predicts these four factors as potential binders but in addition four others (Ap1, Gcr1, Hsf1 and Rox1). If one ranks traditional annotations according to the number of hits, then Gal4 is ranked highest with seven annotated hits followed by Adr1 with two while Mig1 and Ste12 with one binding site each are assigned a tied rank with Ap1, Gcr1, Hsf1 and Rox1.

This analysis was carried out on the entire set of 4451 intergenic sequences which have a ChIP-chip P-value assigned for all our 25 factors. In total this sequence set yields 2388 significant TF–DNA interactions with $P < 0.001$. To assess the quality of different TF ranking schemes we count for each sequence the number of bound TFs ranked above all unbound TFs. For TRAP, TFs are ranked according to $\langle N \rangle$ and for the hit-based methods according to the number of annotated hits as described for the example above. In those cases where several factors have the same number of hits annotated but only a subset of the factors correspond to bound factors, we determine, based on the average of 1000 random samplings, how many times the unbound TFs will accidentially be ranked above a given bound TF.

The analysis shows that 643 (27%) of the significant interactions are correctly ranked on top according to TRAP as compared to 343 (14%) in case of the balanced cutoff and 551 (23%) in case of the 5FP method. These results show that in a considerable number of cases the ranking of TFs according to TRAP is in accordance with ChIP-chip data and overall better than traditional hit-based methods.

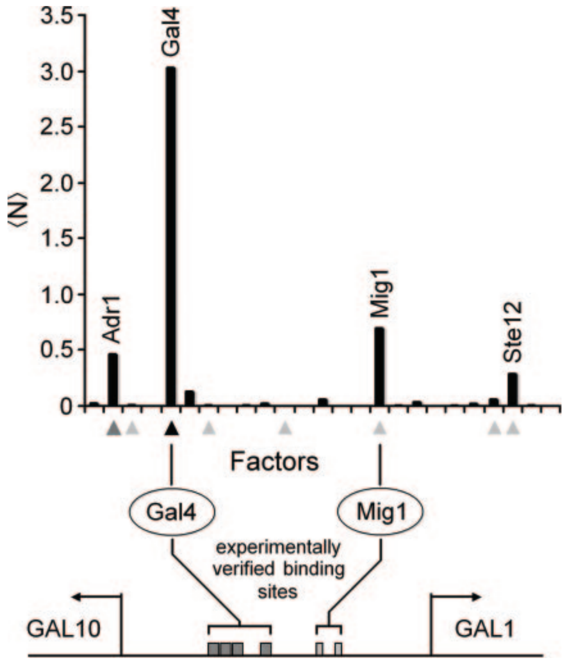

Fig. 4. Affinities for the upstream region of GAL1 and GAL10. The histogram shows the affinity scores as predicted by TRAP. Triangles indicate the factors that have hits annotated according to the balanced cutoff method (black: seven binding sites, dark grey: two binding sites, light grey: one binding site). In the lower part the experimentally verified binding sites are indicated. (Selleck and Majors, 1987; Frolova *et al*., 1999).

Contributions from low affinity sites to $\langle N\rangle$

While our method predicts the overall affinity of a transcription factor to a sequence region, it is still possible to ask which sites contribute most significantly to this affinity. Here we study in more detail the relative contribution of different sites to the total expected count, $\langle N\rangle$, and therefore to the correlation of $\langle N\rangle$ with the observed binding ratios. To this end, we rank all sites in a given sequence according to their probability of being bound, $p(S_l)$, and approximate the expected number of bound TFs by the sum of its n top-ranking sites, $\langle N\rangle_n = \sum_{l=1}^{n} p(S_l)$. This analysis is illustrated in Figure 5 for Leu3.

We find that for the majority of matrices a better correlation can be obtained when all sites are taken into account rather than a single strongest site. This suggests that the relative binding affinities for a given intergenic region are well modeled by taking the total sum over all sites in the region, and supports our claim that a mechanistic description of binding data is possible without imposing any threshold.

4 DISCUSSION

We have applied a physical model to predict the relative binding affinities of TFs to regulatory regions of the DNA. In contrast to the traditional search for binding sites, we do not impose any threshold, but integrate the contributions from individual strong sites and weak sites to calculate the expected number of bound TFs. The ranking of sequence fragments according to this affinity measure is robust with respect to sizable variations in the space of two parameters which define the binding model. Using recent *in vitro* and *in vivo* data from budding yeast, we find that λ lies in the range of [0.4, 1.5] for most factors. The other parameter R_0 is largely determined by the width of the binding site, and to a much lesser extent by the TF concentration. We provide a simple parameterization of the Berg-von Hippel model [$\lambda = 0.7, R_0 = R_0(W)$] and show that a large fraction of our affinity predictions are significantly correlated with experimentally measured R/G ratios in one or more cellular conditions.

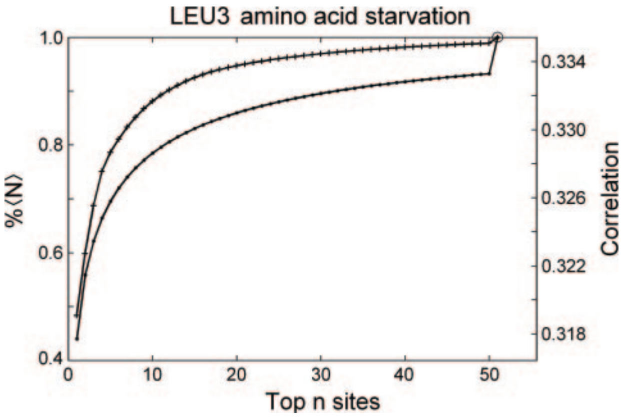

Fig. 5. Contribution of sites with lower affinity. When we arbitrarily constrain $\sum_{l=1}^{n} p(S_l)$ to only the top n scoring terms then the expected counts, $\langle N\rangle$, are reduced, which in turn affects the correlation with the experimental R/G ratio. The upper line shows the changes in the correlation coefficient, the lower line the changes in $\langle N\rangle$. The right-most circled dots denote the values when all sites are taken into account. The increase in the correlation coefficient suggests that the inclusion is biologically meaningful until the correlation coefficient saturates ($r \approx 0.335$) as more and more sites with vanishing affinity are taken into account. This demonstrates that integrating the contributions from all sites provides a more robust approach than limiting the annotation to a few best sites which are determined by some arbitrary cutoff.

Our results indicate that TRAP can better predict relative binding affinities than any of the hit-based approaches. This improvement is due to our probabilistic approach to binding affinities, which avoids assigning a discrete number of binding sites to a sequence. Moreover it takes into account contributions from weak sites and hence can assign affinities for sequences where hit-based methods fail to report any ‘match’. It also accounts for differences in the binding strength of sites which are traditionally only reported as hits. This is not only reflected in better correlation but also better and more robust ranking of TFs as compared to hit-based methods.

Considering a comprehensive list of 25 factors and 13 conditions (61 experimentally tested combinations), we find that our predictions resulted in high correlations ($r > 0.3$) for 23 of these combinations. In addition for 36 combinations TRAP yielded a ROC curve area ≥ 0.7. It is encouraging to see that our predictions also match what is known about the involvement of TFs in the various conditions tested. For example, Hsf1, Rcs1 and Leu3 are known to be involved in several aspects of stress response (Raitt *et al*., 2000; Blaiseau *et al*., 2001; Zhou *et al*., 1987) and their predicted affinities show high correlation with R/G ratios only in conditions of oxidative stress (H_2O_2) and amino acid starvation, but not in rich medium.

This also suggests why, for certain factors and cellular conditions, the physical model cannot be expected to predict binding affinities *in vivo*. Indeed, for nine factor-condition pairs with only small correlation ($r < 0.3$) the TFs may not be expressed or available for binding under the condition tested. These include Adr1, Hac1, Mata1, Pdr3, Pho4, Xbp1, Yap1 and Zap1 in rich medium and Mig1 in medium with galactose as carbon source. For example, Mig1 is known to be located in the cytoplasm in the presence of galactose, and hence it is not available for DNA-binding in the nucleus (Vit *et al*., 1997). However, our predictions for Mig1 do show a high correlation ($r = 0.60$) with *in vitro* data. Likewise, the binding ratios of Abf1 and Rap1 have also been determined *in vitro* and show a high correlation with our predictions ($r > 0.5$) under this condition. This is in accordance with our assumption that the TF is available and that the DNA is accessible.

The TRAP approach appears to fail for other matrices and conditions, even though we have no indication that the corresponding factor is absent. We want to stress that our approach requires the definition of matrix descriptions which can be used as good approximations for mismatch energies in the physical model. There are several cases where we suspect that the matrix description may be inappropriate. For example, for Hsf1 there are four matrices listed in TRANSFAC, but only one of them (an alternating trimer motif HSF1_04) yields good correlations with the experimental binding ratios. Interestingly the trimer combination of this matrix has been described as the site with highest affinity for Hsf1 (Sorger and Pelham, 1987; Xiao *et al*., 1991). It is possible that better predictions can be achieved by using improved matrices like in the case of GCN4_01 or matrices derived from ChIP-chip data (Foat *et al*., 2006; Tanay, 2006). The focus of this work, however, is to explain ChIP-chip data in a biophysical framework rather than the evaluation of matrices. Hence in the present study only publically available matrices are used.

The key ingredient of the model by Berg and von Hippel is the assumption that different basepairs contribute independently from each other to the overall binding energy. This assumption also entails that mismatch energies for large deviations from the consensus sequence are not calculated differently from small deviations. Since TF–DNA complexes can, presumably, compensate for the relative increase in free energy from base pair mismatches through other mechanisms, such as conformational changes the model may underestimate the binding affinity of weak sites and thus $\langle N \rangle$.

The fact that already now a significant fraction of yeast binding data can be accounted for by matrix motifs is all but obvious given the complicated binding mechanisms in eukaryotes and the relatively simple energetic binding model. Prokaryotic binding data has triggered motif based models more than 20 years ago. Our results demonstrate that, despite the increased complexity of the eukaryotic cell, such energetic binding models are also of predictive value for yeast. It will be interesting to see to what extent the observations made for yeast will carry over to multi-cellular organisms.

ACKNOWLEDGEMENTS

The authors would like to thank Ho-Ryun Chung, Stefan Haas, Philipp Messer, Peter Arndt and Roded Sharan for many valuable discussions and Szymon Kielbasa for providing help with the hit-based sequence annotation. This work was supported by a grant from the German National Genome Research Network (NGFN) and the EU (BioSapiens Network, contract number LSHG-CT-2003-503265).

Conflict of Interests: none declared.

REFERENCES

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

Blaiseau,P. *et al*. (2001) Aft2p, a novel iron-regulated transcription activator that modulates, with Aft1p, intracellular iron use and resistance to oxidative stress in yeast. *J. Biol. Chem*., **276**, 34221–34226.

Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol*., **212**, 563–578.

D'haeseleer,P. (2006) What are DNA sequence motifs? *Nat. Biotechnol*., **24**, 423–425.

Djordjevic,M. *et al*. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res*., **13**, 2381–2390.

Foat,B.C. *et al*. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, e141–e149.

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

Frolova,E. *et al*. (1999) Binding of the glucose-dependent Mig1p repressor to the GAL1 and GAL4 promoters *in vivo*: regulation by glucose and chromatin structure. *Nucleic Acids Res*., **27**, 1350–1358.

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

Gilson,E. *et al*. (1993) Distortion of the DNA double helix by RAP1 at silencers and multiple telomeric binding sites. *J. Mol. Biol*., **231**, 293–310.

Granek,J.A. and Clarke,N.D. (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol*., **6**, R87.

Harbison,C.T. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.

Mossing,M. and Record,M. (1985) Thermodynamic origins of specificity in the lac repressor-operator interaction. Adaptability in the recognition of mutant operator sites. *J. Mol. Biol*., **186**, 295–305.

Mukherjee,S. *et al*. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet*., **36**, 1331–1339.

Rahmann,S. *et al*. (2003) On the power of profiles for transcription factor binding site detection. *Stat. Appl. Genet. Mol. Biol*., **2**.

Raitt,D. *et al*. (2000) The Skn7 response regulator of *Saccharomyces cerevisiae* interacts with Hsf1 in vivo and is required for the induction of heat shock genes by oxidative stress. *Mol. Biol. Cell*, **11**, 2335–2347.

Rajewsky,N. *et al*. (2002) Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics*, **3**, 30.

Reeves,W.M. and Hahn,S. (2005) Targets of the Gal4 transcription activator in functional transcription complexes. *Mol. Cell. Biol*., **25**, 9092–9102.

Selleck,S. and Majors,J. (1987) *In vivo* DNA-binding properties of a yeast transcription activator protein. *Mol. Cell. Biol*., **7**, 3260–3267.

Sorger,P. and Pelham,H. (1987) Purification and characterization of a heat-shock element binding protein from yeast. *EMBO J*., **6**, 3035–3041.

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

Tanay,A. (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*., **16**, 962–972.

Vit,M.D. *et al*. (1997) Regulated nuclear translocation of the Mig1 glucose repressor. *Mol. Biol. Cell*, **8**, 1603–1618.

von Hippel,P. and Berg,O. (1986) On the specificity of DNA-protein interactions. *Proc. Natl Acad. Sci. USA*, **83**, 1608–1612.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet*., **5**, 276–287.

Xiao,H. *et al*. (1991) Cooperative binding of *Drosophila* heat shock factor to arrays of a conserved 5 bp unit. *Cell*, **64**, 585–593.

Zhou,K. *et al*. (1987) Structure of yeast regulatory gene LEU3 and evidence that LEU3 itself is under general amino acid control. *Nucleic Acids Res*., **15**, 5261–5273.

Zumdahl,S.S. (1998) *Chemical Principles, 3rd edn*. Houghton Mifflin Company.