BIOINFORMATICS

Vol. 19 no. 2 2003
Pages 212–218

Rank order metrics for quantifying the association of sequence features with gene regulation

Neil D. Clarke and Joshua A. Granek*

Department of Biophysics and Biophysical Chemistry, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA

Received on May 13, 2002; revised on July 16, 2002; accepted on August 13, 2002

ABSTRACT

Motivation: Genome sequences and transcriptome analyses allow the correlation between gene regulation and DNA sequence features to be studied at the whole-genome level. To quantify these correlations, metrics are needed that can be applied to any sequence feature, regardless of its statistical distribution. It is also desirable for the metric values to be determined objectively, that is, without the use of subjective threshold values.

Results: We compare two metrics for quantifying the correlation of DNA sequence features with gene regulation. Each of the metrics is calculated from a rank-ordering of genes based on the value of the sequence feature of interest. The first metric is the area under the curve for a receiver operator characteristic plot (ROC AUC), a common way of summarizing the tradeoff between sensitivity and specificity for different values of a prediction criterion. We call the second metric the mean normalized conditional probability (MNCP). The MNCP can be thought of as the predictive value of the sequence feature averaged over all regulated genes. The statistical significance (P-value) of each metric can be estimated from simulations. Importantly, the P-value of the MNCP metric is less dramatically affected by the presence of false positives among the set of co-regulated genes than is the ROC AUC. This is especially useful in analyzing gene sets identified by DNA microarray analysis because such data cannot distinguish direct regulation by transcription factor binding from indirect regulation. We demonstrate that these two metrics, taken together, are useful tools for defining the binding site representation and regulatory control regions that best explain the difference between genes that are regulated by a given transcription factor and those that are not. Applications to other gene features are also described.

Availability: A Python program for calculating the ROC AUC and MNCP metric values given input rank orders is available from ftp://ftp.bs.jhmi.edu/users/nclarke/MNCP/.

*To whom correspondence should be addressed.

Contact: nclarke@jhmi.edu

INTRODUCTION

DNA sequence features that affect gene regulation include the number, quality, and location of transcription factor binding sites, as well as more complex features that are poorly understood. The availability of genomic sequences and DNA microarray data permit comprehensive analyses of the distribution of sequence features among regulated and unregulated genes. Unfortunately, differences in the way analyses of this type are performed and reported prevent meaningful comparisons of their conclusions.

One common problem is that the correlation of a sequence feature with regulation is reported only for a single, arbitrarily chosen value of the feature. For example, the relationship between consensus binding sites and gene regulation is generally reported in terms of the number of genes that have at least one such site (cf. DeRisi *et al.*, 1997; Chu *et al.*, 1998; Laub *et al.*, 2000; Jelinsky *et al.*, 2000). However, eukaryotic genes often have multiple copies of a transcription factor binding site, and the occurrence of an unusually high density of sites is more predictive of gene regulation than is a single site (Berman *et al.*, 2002; Markstein *et al.*, 2002). In general, then, using a threshold of one site underestimates the relevance of binding sites to gene regulation. In fact, any threshold is arbitrary and entails a loss of information. For that reason, it is preferable to determine the distribution of binding site numbers among regulated and unregulated genes, and to quantify differences in these distributions.

While the number of consensus sites approximates a Poisson distribution, other sequence features may have very different distributions. In order to compare sequence features in a consistent manner, it is important that we adopt metrics that are independent of the underlying statistical distribution. Rank-order metrics are an ideal solution to this problem because all distributions get converted to a distribution of rank orders.

SYSTEM AND METHODS

The yeast Yap1 transcription factor

DNA microarray experiments that probe the effects of perturbing a single transcription factor are arguably the most useful for analyzing the correlation of sequence features with gene regulation. An excellent example of such an experiment is that of DeRisi *et al.*, who identified seventeen yeast genes induced by overexpression of the transcription factor Yap1 (DeRisi *et al.*, 1997). Over six thousand other genes were not significantly induced by Yap1. The gene set used here consists of sixteen genes identified by DeRisi *et al.* as we were unable to find gene YCLX08C in any of the files obtained from the *Saccharomyces* Genome Database. The naming convention for yeast ORFs indicates that this was a tentatively identified ORF, and suggests that the ORF has been deleted or renamed.

Eleven of the sixteen Yap1-induced genes (69%) have at least one Yap1 consensus site (T[TG]ACTGG, Wemmie *et al.*, 1994) within 600 bp 5′ of the start of the ORF (DeRisi *et al.*, 1997). In contrast, consensus sites were found in only four out of thirty (13%) randomly selected, uninduced genes (DeRisi *et al.*, 1997). A full enumeration shows that the fraction of uninduced genes with sites is closer to 20%, but even this higher value is consistent with the number of induced genes with sites being significantly greater than expected if the consensus site had nothing to do with induction by Yap1 overexpression. On the other hand, the occurrence of a single consensus site is obviously not very predictive as there are over 100 uninduced genes with a site for every induced gene with a site.

To better characterize the association of binding sites with induction, we examined two metrics that can be used to take into account all possible values of a sequence feature rather than a single threshold value. The first of the metrics is the area under a receiver operator characteristic (ROC AUC) curve, and the second, which we believe is novel, is the Mean Normalized Conditional Probability (MNCP). We show that the MNCP has an advantage over ROC AUC in that it is less sensitive to the presence of false positives in the experimentally defined set of co-regulated genes. Both MNCP and ROC AUC are valuable metrics whose adoption would facilitate the comparison of sequence features and gene sets.

Description of metrics: area under the Receiver Operator Characteristic curve

The two metrics we have adopted are based on rank-ordering all genes by the value of a sequence feature of interest. The distribution of rank orders for the regulated genes is then assessed in the context of the rank orders for all other genes. To illustrate how the metrics are calculated, the score used to rank-order the genes is the number of Yap1 consensus sites within 600 bp 5′ of the ORF. However, the metrics can be applied to any sequence property for which a value can be assigned to a gene. In fact, they can be applied to combinations of properties provided there is a function (i.e. a regulatory model) that produces a score for each gene based on those properties.

The first step in calculating the metrics is to rank all yeast genes by the number of upstream Yap1 consensus sites with the gene having the greatest number of sites ranked first. Initially, the ranks run from 1 to $\mathcal{N}$ where $\mathcal{N}$ is the total number of genes. Genes with the same value (ties) are then reassigned ranks such that the rank given to all genes in the tie is the rank of the gene that was originally the lowest in the set. The effect of this renumbering is that the rank for each gene is equivalent to the number of genes that have an equal or better value of the sequence feature. We then descend the list of ranked genes, determining at each rank the number (or fraction) of regulated genes of that rank or better and the number (or fraction) of unregulated genes of that rank or better. These values are then plotted as shown in Figure 1. In the case of Yap1 consensus sites, there are only five different ranks because there are only five values of the sequence feature that distinguish one gene from another ($\geqslant$0 sites, $\geqslant$1 site, . . ., $\geqslant$4 sites). There are, therefore, only, five distinct points on the graph. If every gene had a different value for the sequence feature there would be $\mathcal{N}$ points on the graph.

The curve in Figure 1 is a standard receiver operator characteristic (ROC) curve (Hanley and McNeil, 1982; Pagano and Gauvreau, 1993). The more familiar way to generate ROC curves is to evaluate a number of threshold values, plotting for each value the fraction of true positives correctly identified by the threshold versus the fraction of true negatives falsely identified by the threshold. Using the rank order list to generate a ROC curve, as described above, is equivalent to considering *all* threshold values that could affect the curve. The ROC curve for a perfectly predictive gene feature would consist of a vertical line along the *y*-axis, followed by a horizontal line across the top of the plot, and the area under the curve (AUC) would be 1.0. In contrast, sequence features that have no ability to discriminate between the two gene sets are expected to have ROC curves that, on average, fall along the diagonal and have an AUC of 0.5. The ROC curve in Figure 1 has an AUC of 0.776.

Description of metrics: Mean Normalized Conditional Probability

The MNCP metric provides a different measure of rank order distribution that is more sensitive to the ranks of the highest ranked genes and less sensitive to (adversely affected by) false positives among the experimentally-defined regulated genes. We begin with a graphical

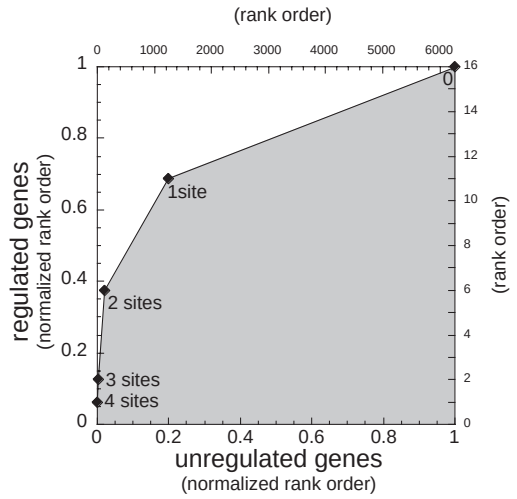

Fig. 1. The complete set of yeast genes was rank-ordered according to the number of associated Yap1 consensus sites as described in the text. The sixteen Yap1-induced genes were evaluated in terms of the fraction of regulated genes (y-axis) and the fraction of unregulated genes (x-axis) that had an equal or greater number of associated consensus sites. Diamonds indicate points occupied by one or more of the induced genes. The figure could also be thought of as having been generated by consideration of all possible threshold values of the sequence feature (i.e. numbers of binding sites). The number of sites that serve as a threshold for each point is indicated in the figure. The gray area indicates the area under the curve (AUC). In this case, the AUC is 0.776.

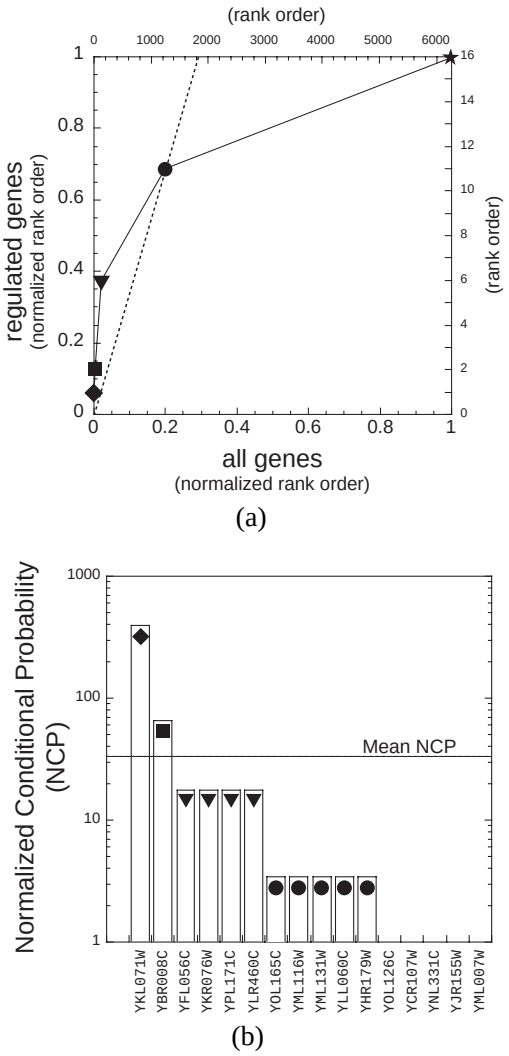

Fig. 2. (a) Calculation of MNCP values. The points on this graph are essentially the same as in Figure 1 except the x-axis is based on all genes. The MNCP value is based on calculating for each gene the slope of the line that runs from the origin through the point on the graph for that gene (e.g. the dashed line for genes marked by the circle). (b) The slopes calculated for each gene. As described in the text, the slopes are equivalent to the normalized conditional probability $\frac{P\ (\text{regulated}|\text{threshold}_i)}{P\ (\text{regulated})}$ where the threshold value is the value (number of consensus sites) held by the gene in question, i. The MNCP value is the mean of these numbers and is indicated by the dashed line. To facilitate the comparison of the two parts of this figure, the histogram bars are annotated with the same symbols used to indicate the gene(s) in part (a)

description of the MNCP metric based on a graph similar to a ROC curve except the x-axis is all genes, not just the uninduced (Figure 2a). Rather than calculate the area under the curve, as is done for the ROC curve, we determine the slopes of the lines that run through the origin to the point for each regulated gene. The MNCP is the mean of the slopes for all regulated genes.

The slope for a given regulated gene, i, is equal to $(R(i)/N)/(\mathcal{R}(i)/\mathcal{N})$, where $R(i)$ is the rank of gene i among the set of regulated genes, $\mathcal{R}(i)$ is the rank of gene i among the set of all genes, N is the total number of regulated genes, and $\mathcal{N}$is the total number of all genes. For example, for the five genes represented by the circle in Figure 2b the slope is $3.4 = (0.688/0.2) = (11/16)/(1252/6276))$

Rearrangement of the expression for the slope suggests a more intuitive interpretation of its value. Slope $= (R(i)/N)/(\mathcal{R}(i)/\mathcal{N}) = (R(i)/\mathcal{R}(i))/(N/\mathcal{N})$. Consider the numerator in this expression, $R(i)/\mathcal{R}(i)$. If the value of the sequence feature found in gene i is considered to be a threshold value for predicting gene regulation, then $R(i)$, the rank order among the regulated gene set, is the number of regulated genes that meet the criterion, $\mathcal{R}(i)$ is the number of all genes that meet the criterion, and $R(i)/\mathcal{R}(i)$ is the fraction of genes that meet the prediction threshold that are actually regulated. Equivalently, $R(i)/\mathcal{R}(i)$ is the conditional probability that a gene is regulated if it meets a threshold value defined by the value of the sequence feature in gene i. The denominator in the

re-arranged slope expression, $N/\mathcal{N}$, is simply the fraction of genes that are regulated, or the *prior* probability of being regulated in the absence of any information about sequence features. Thus, the slope is the conditional probability of being regulated normalized to the prior probability: $\frac{P(\text{regulated}|\text{threshold}_i)}{P\ (\text{regulated})}$.

The calculation of the slope (or normalized conditional probability) is repeated for all sixteen induced genes (remembering that several are tied in rank order) and the results shown in Figure 2b are obtained. Note that this histogram contains exactly the same information as the ROC curve, no more and no less. The mean of the normalized conditional probabilities (MNCP) captures some aspects of the rank order distribution from which it is derived, just as the area under the ROC curve does.

Estimating statistical significance

The statistical significance of a given ROC AUC or MNCP value can be estimated by generating a large number of sets of random rank orders and then comparing the real ROC AUC and MNCP values to the distribution of values obtained for the random rank orders. Random rank orders for N induced genes out of a total gene complement of $\mathcal{N}$ are simulated by selecting N random integers between 1 and $\mathcal{N}$. These rank orders are then used to calculate the ROC AUC and the MNCP. The simulation is repeated a large number of times (10^7 for the numbers reported here) and the fraction of random rank orders that give ROC AUC or MNCP values higher than those found with the real data can be used as an estimate for the one-tailed P-value. This would be appropriate in the case of consensus binding sites because we expect a positive correlation between consensus sites and regulation.

If there are no prior expectations about the sign of the correlation, a two-tailed P-value is needed to avoid overestimating the significance. An example would be testing the hypothesis that the size of the 5′ non-coding region had something to do with regulation but without regard to whether the association would prove to be with long non-coding sequences or short (see below). The two-tailed P-value is estimated by calculating the ROC AUC and MNCP metrics twice for each randomly selected set of ranks, once assuming the numbers are in standard rank order and again assuming they reflect the reverse rank order. We then take the higher of the two numbers as the value of the metric for that random rank order. This is essentially what would be done for the ranks of a real sequence feature if we had no prior hypothesis about the sign of the correlation. For consistency, all P-values reported here are calculated in this way. For features like binding sites, where a one-tailed P-value could also be used, the two-tailed value offers a conservative underestimate of the significance.

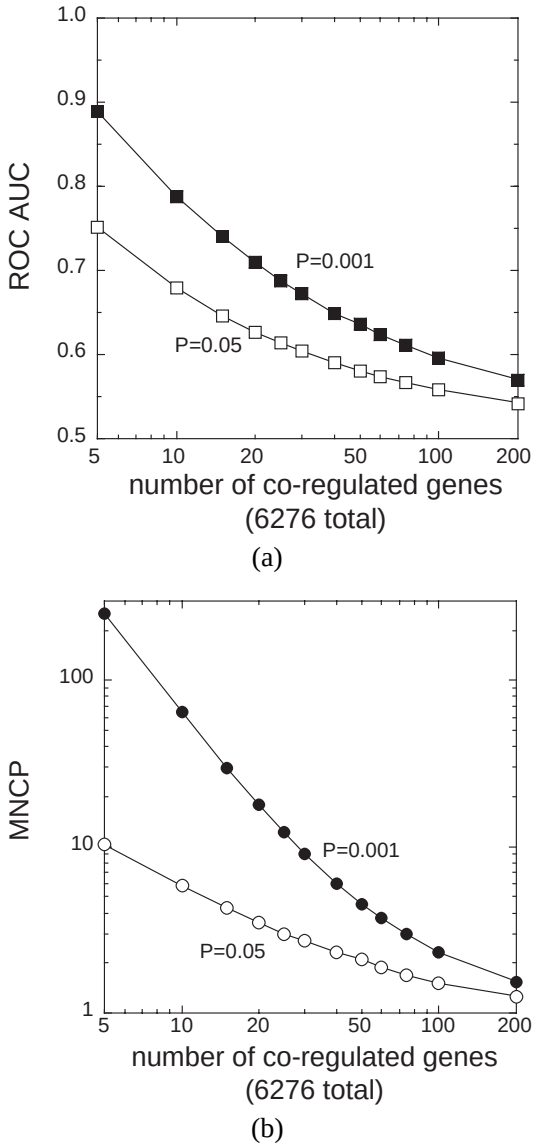

Fig. 3. Scores necessary to have a P-value of 0.001 (solid symbols) or 0.05 (open symbols) for increasing numbers of co-regulated genes out of a total gene complement of 6276. The number of co-regulated genes is indicated on a log scale. (a) ROC AUC values necessary to achieve the indicated P-values, plotted on a linear scale. (b) MNCP values necessary to achieve the indicated P-values, plotted on a log scale.

The ROC AUC or MNCP scores necessary to be considered significant depend on $\mathcal{N}$, the total number of genes, and N, the number of regulated genes. With $\mathcal{N}$ fixed at 6276 (the number of yeast genes in our Yap1 analysis), we determined the ROC AUC and MNCP scores necessary to achieve P-values of 0.05 or 0.001 for values of N ranging from 5 to 200. The results are shown in Figure 3.

RESULTS AND DISCUSSION

Comparison of the ROC AUC and MNCP metrics

The ROC AUC is a useful representation of the bias in the distribution of rank orders for regulated genes versus unregulated. It also has the virtue of familiarity, being a commonly-used statistical metric. However, the ROC AUC is insensitive to the precise rank orders of the induced genes. For example, one of the Yap1-induced genes, YKL071W, has more Yap1 consensus sites than any other gene in yeast. Suppose, though, we discovered that we had made a mistake, and that ten uninduced genes previously thought to have no sites actually had more sites than YKL071W. This would surely diminish our subjective sense of how well Yap1 consensus sites explain induction by Yap1. However, this discovery would have only a marginal effect on the ROC AUC (a drop from 0.776 to 0.775). This is because the movement of ten uninduced genes from the bottom of the rank-order list to the top would have only a small effect on the coordinates of the ROC curve points, moving each to the right by only 0.0016 (or 10/6276). The effect on the value of MNCP, on the other hand, is substantial, dropping from about 34.4 to 10.0. This is because the MNCP is dominated by the genes that are best associated with the sequence feature. In particular, in this case the normalized conditional probability of YKL071W would drop from 392 to 36, reflecting the change in rank for this one gene from first overall to eleventh.

The MNCP also has an advantage over the ROC AUC in that it is less affected by false positives among the list of regulated genes. False positives can be genes that were incorrectly identified due to experimental error, or they may be genes that are regulated indirectly. In either case, false positives cannot be expected to share sequence features with genes that are regulated by direct binding of the transcription factor. We can test the effect of false positives on the two rank order distribution metrics by adding non-induced genes to our list of induced genes (some of which may be false positives themselves). If we add just one false positive lacking a consensus site to the list, the ROC AUC drops by 0.022, an effect more than 20-fold greater than what happens when ten uninduced genes are assumed to actually have large numbers of binding sites and are promoted to the top of the rank-order list (described above). In contrast, promotion of the ten uninduced genes to the top of the list has a larger effect on the MNCP metric (a decrease from 34.4 to 10.0) than does the addition of a false positive to the gene list (34.4–30.6).

This comparison suggests that the MNCP is less adversely affected by false positives than is the ROC AUC. To compare the sensitivity to false positives on an equal footing, we calculated *P*-values for each metric for a series of gene sets containing increasing numbers of false positives (Figure 4). For the experimentally-defined set of sixteen genes, the *P*-value for the two metrics is similar (~0.00006). However, as increasing numbers of false positives are added to the gene set, the ROC AUC value loses significance much more quickly than does the MNCP value.

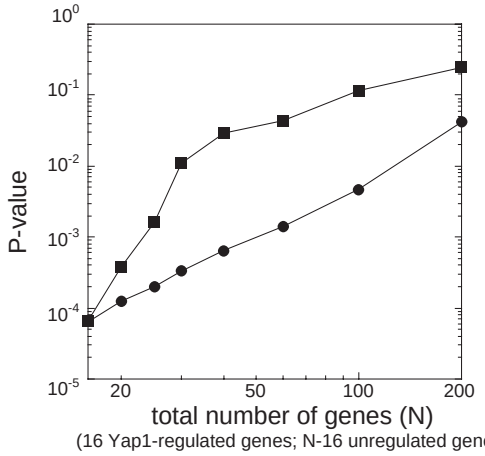

Fig. 4. Sensitivity of the *P*-value of ROC AUC (squares) and MNCP (circles) to false positives in the experimentally defined list of co-regulated genes. The total number of genes (*x*-axis; log scale) reflects the sixteen experimentally defined plus varying numbers of unregulated genes. The distribution of binding sites in the doped unregulated genes approximates as closely as possible the distribution in the entire set of unregulated genes.

Application of rank order metrics to the comparison of sequence features, alternative definitions of the regulatory region, and gene sets

Three different consensus-site definitions of a Yap1 binding site were evaluated by enumerating sites and determining the ROC AUC and MNCP scores for the rank-ordered genes. Each of the three variants of the binding motif was evaluated for several different definitions of the regulatory region. The regulatory regions start at the ORF and extend 5′ for as little as 300 bp or as much as 800 bp. The results are shown in Figure 5. We conclude from this analysis that the consensus site T[TG]ACTAA is superior to two alternative definitions of the Yap1 binding site and that a regulatory region definition that extends 600 or 700 bp 5′ of the ORF is more effective than regions that are either shorter or longer.

To illustrate the applicability of the metrics to other sequence features, Figure 6a shows the ROC AUC and MNCP values applied to the chromosomal location of the gene and to the length of the intergenic sequences. The consensus site values are shown for comparison. Remarkably, all three sequence features are significantly

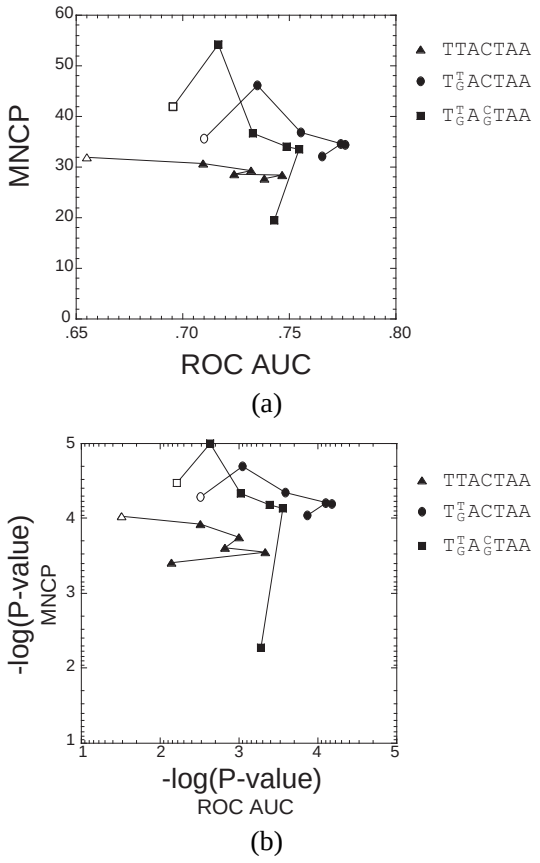

Fig. 5. MNCP and ROC AUC scores for three definitions of the consensus site (symbols shown next to the graphs) and for six definitions of the regulatory region (300–800 bp 5′ of the ORF; the 300 bp point is shown as an open symbol and successive points increase the regulatory regions by 100 bp). Sites in either orientation were counted in the enumeration (a) raw scores for the two metrics. (b) *P*-values, represented as $-\log(P\text{-value})$ so that high values on this plot correspond to highs value for the raw scores.

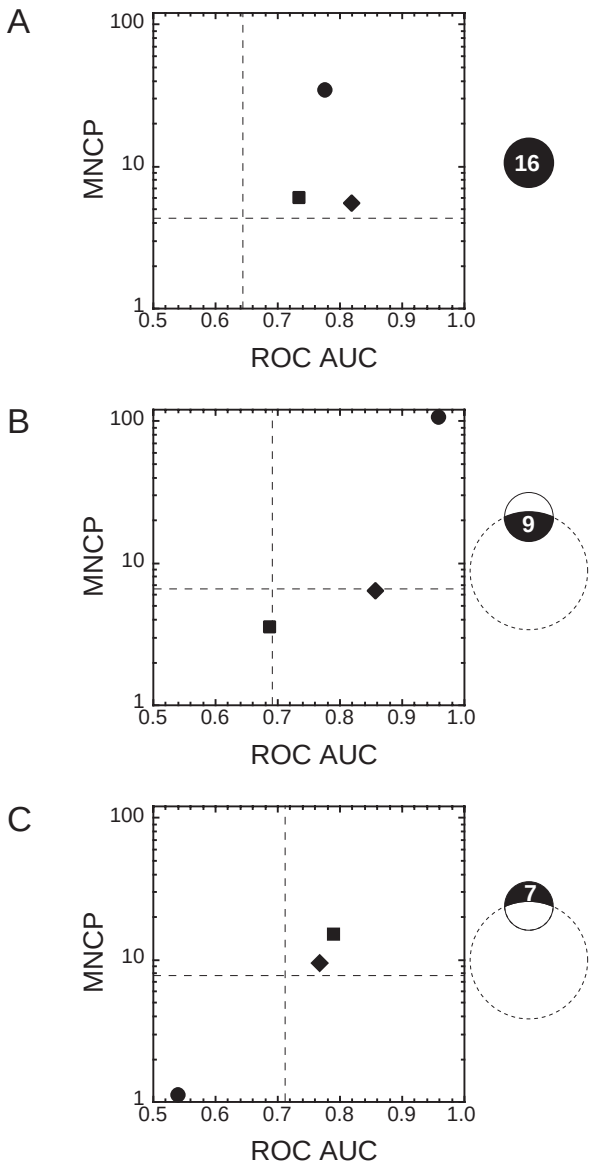

Fig. 6. ROC AUC and MNCP values plotted for (a) sixteen genes induced by Yap1 overexpression, (b) a nine gene subset that meet additional criteria for Yap1 regulation, and (c) the seven genes that fail to meet these additional criteria. Venn diagrams to the right of each graph illustrate the relationships of the three gene sets. The gene features evaluated are the number of consensus sites within 600 bp 5′ of the ORF (circles), the chromosomal location of the gene expressed as the fractional distance from the centromere to the telomere, with telomeric genes ranked highest (squares), and the length of the non-coding sequence 5′ to the ORF, with genes having longer non-coding sequences ranked higher (diamonds). Dashed lines indicate the metric values for which $P < 0.05$. The data of Gasch et al were used to define the nine gene subset (Gasch *et al.*, 2000). These nine were among 98 genes we identified from the data whose transcript levels were elevated at least three-fold in the presence of H_2O_2 and for which there was at least a three-fold effect of YAP1 deletion on induction by H_2O_2.

associated with Yap1 induction. We then evaluated the same three gene features for two subsets of the sixteen genes, the first a set of nine genes that meets additional criteria for regulation based on experiments with a YAP1 deletion (Figure 6B; Gasch *et al.*, 2000) , and the second the remaining seven genes that do not meet these criteria (Figure 6c). The subset that meets the additional criteria shows a substantially stronger association with consensus sites than the full set, while the remaining genes have no significant association with binding sites at all. Although there is no significant association with binding sites, the genes identified only by the overexpression experiment have a significant tendency to be closer to the telomeres than the average gene (Figure 6c). Also, both subsets are associated with longer intergenic regions (Figures 6b and 6c).

Application of the rank order metrics to the evaluation of regulatory models

The gene ranks we have used here are based on the values of individual sequence features. However, ranks can be equally well obtained for a combination of features provided a function is first defined that produces a score for each gene based on the sequence features and their relationships. Such functions, which we have begun to investigate, are tantamount to models for gene regulation. For example, if there are two sequence motifs associated with a set of co-regulated genes, alternative functions might include: (i) summing the total number of both sites (i.e. modeling independent contributions to expression) and (ii) enumerating pairs of sites that meet some distance criterion (i.e. modeling cooperative binding). There are, of course, an infinite number of models and parameter choices. The point is that the ROC AUC and MNCP metrics can be used to compare models for gene regulation just as we have shown they can be used to compare the association of individual sequence features with gene expression.

CONCLUSION

Metrics based on the rank ordering of genes can be used to quantify the association between gene regulation and any kind of sequence feature, or features. No arbitrary thresholds are required because all values of the sequence feature are implicitly used in rank ordering the genes. MNCP, the novel rank order metric introduced here, has some advantages over the more conventional ROC AUC for quantifying the association of sequence features with gene expression. The MNCP is less affected by false positives in the co-regulated gene set and (for the same reason) is more sensitive to an especially strong association between the sequence feature and a small subset of the regulated genes. We propose using both metrics as complementary representations of rank order distributions in the hope that this will facilitate the comparison of different sequence features, regulatory models, and gene sets.

REFERENCES

Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.

Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Jelinsky,S.A., Estep,P., Church,G.M. and Samson,L.D. (2000) Regulatory networks revealed by transcriptional profiling of demaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes. *Mol. Cell. Biol.*, **20**, 8157–8167.

Laub,M.T., McAdams,H.H., Feldblyum,T., Fraser,C.M. and Shapiro,L. (2000) Global analysis of the genetic network controlling a bacterial cell cycle. *Science*, **290**, 2144–2148.

Markstein,M., Markstein,P., Markstein,V. and Levine,M.S. (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl Acad. Sci. USA*, **99**, 763–768.

Pagano,M. and Gauvreau,K. (1993) *Principles of Biostatistics*. Wadsworth, Belmont, CA.

Wemmie,J.A., Szczypka,M.S., Thiele,D.J. and Moye-Rowley,W.S. (1994) Cadmium tolerance mediated by the yeast AP-1 protein requires the presence of an ATP-binding cassette transporter-encoding gene, YCF1. *J. Biol. Chem.*, **269**, 32592–32597.