Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Transcriptional Target Prediction Using Qualitative Reasoning

Li-San Wang, Doris Wagner, Chang Seob Kwon, Yanhui Su, and Junhyong Kim

Abstract—**Transcription target prediction from functional genomics data often involves incorporating a conjunction of complex prior biological knowledge to the analysis. Unfortunately, typical prior hypotheses are qualitative rather than quantitative in nature. But, many qualitative biological hypotheses can be decomposed into a set of logic statements on binary outcomes. Here, we present a new method to convert qualitative statements into a collection of binary statements that in turn generates a partial ordering of outcomes, which can be tested using a semi-parametric isotonic regression. This semi-parametric approach yields a flexible but principled way of testing biological hypotheses. We applied this method to a published Arabidopsis microarray dataset to identify organ specific transcriptional target genes, and tested predictions independently using the AtGenExpress dataset. Our new algorithm performed comparably to published approaches and allowed rapid analysis of complex, multiple gene selection criteria.**

Index Terms—**Bioinformatics, microarray analysis, isotonic regression, semi-parametric statistics.**

I. Introduction

QUANTITATIVE analysis of biological data requires associating the data with a mathematical or probabilistic model. As an example of such a modeling process, suppose we have the hypothesis that the expression level of the transcription factor ACE2 in *S. cerevisiae* controls cell size. To search for downstream targets of this transcription factor, we can generate a hypothesis that postulates a linear quantitative relationship between cell size, ACE2 expression levels, aand the expression levels of the target genes. We can further model the noise as a sample from a Gaussian distribution. We can then fit a standard regression to the data and carry out a statistical hypothesis test for the significance of the slope.

In actual practice, the above linear regression analysis may be too simple for the actual biological phenomena. First, that target gene expression levels should be related to cell size may not translate well into a regular functional relationship such as the proposed linear relationship between size and expression levels. Second, as it often happens with biological data, an off-the-shelf noise model like the exponential family may not emulate the finite sample of observed data very well. More importantly, prior biological hypotheses are rarely stated in a quantitative form. The statements are more of the kind "If gene A targets gene B, then I expect both A and B to be up-regulated under condition X".

Additional complications arise when conjunction of such hypotheses are used to make higher order inferences. For example, in experimental transcriptional target prediction, we often use various conditions to manipulate the levels of the target transcription factor and then measure the transcriptome under these conditions. Biological knowledge typically generates an expectation under these experimental manipulations and we search for particular genes in the transcriptome that follow these expectations. This experimental approach has two difficulties. First, if the biological logic becomes complicated, it can be difficult to intuit the rational expectations—thus many biologists use the simpler strict joint condition (i.e., "if X AND Y AND Z AND..etc). Second, loose statements such as "if A then X is up-regulated" seemingly predicts a binary outcome but the actual implicit model is slightly more complicated. The biologist has more or less confidence in the outcome based on the quantitative degree of "up-regulation". Thus some quantitative degree of fit to the reasoning must be assessed, but it might be difficult to presume a standard noise model for the fit. In this paper, we propose a new approach where qualitative biological hypotheses are converted to expectations on partial ordering of experimental outcomes. The fit to the partial order is tested using a semi-parametric procedure called *isotonic regression* [3]. The utility of the approach for identifying transcriptional targets from functional genomics data is demonstrated using experiments on the *Arabidopsis* genome.

A. Modeling qualitative hypotheses as order relationships

A typical qualitative biological hypotheses has the form "I expect measurement X to have 'higher' response under condition A versus condition B", where 'higher' can also be a qualitative statement. The key to our approach is that qualitative paired comparison statements on a set of conditions can be used to create a semi-quantitative expectation for the entire suite of conditions through the implied partial ordering on the condition set. As an example, consider a microarray experiment with four treatment conditions: 0 (control), 1, 2,

Manuscript received April 24, 2006. This research was supported in part by NIH 1-T32-CA101968 and NIH 1-T32-HG000046-06 to L. W., NSF grant IBN 130804 to D. W., and NIH 1-P20-GM-6912-1 to J. K.

The first two authors contributed equally to this publication. All authors are with the Department of Biology, University of Pennsylvania. Corresponding author: Junhyong Kim, Mailing address: Biology Department, University of Pennsylvania, 203 Goddard Laboratories, 415 S University Avenue, Philadelphia, PA 19104 USA; E-mail: Junhyong@sas.upenn.edu.

1-4244-0033-3/06/$20.00 ©2006 IEEE.

and 3. Assume the following expected behaviors: (a) the candidate gene should have higher expression level under conditions 1 and 2 than under condition 0 (say at least two-fold difference), (b) expression level under conditions 0 and 3 should be roughly equal. Denoting the expression level of genes under condition i as X_i, we can write the following partial order relationship describing the target pattern regarding the expression levels X_0, X_1, X_2, X_3:

$$X_1 > 2X_0, X_2 > 2X_0, X_0 = X_3.$$

Here the partial order relationship specifies the ordering of the three ratios and the two constants 1 and 2. Therefore, qualitative biological hypotheses of a measured variable under a variety of conditions can be decomposed into prior expectations on pairs of conditions, which in turn, can be converted to a partial ordering or a set of total ordering of the values over the conditions. The set of total ordering generates a non-decreasing sequence of expected values and the observed experimental data can be tested against a non-decreasing expected sequence by a procedure called isotonic regression [3]. Isotonic regression employs a semi-parametric least-squares criterion that is more sensitive than a rank order fit since it takes magnitude of the deviation into account. However, a non-decreasing function fits a much wider pattern than a linear model as in Pearson's correlation. Thus, our semi-parametric approach attempts to preserve the robust qualities of a non-parametric approach while recovering some of the sensitivities of a parametric approach.

B. *Gene selection in microarray experiments*

While transcriptome profiling has been widely used in large-scale functional genomics studies, many studies employing microarrays involve selecting genes using a small-scale experimental design. In these studies, the experimenter sets forth different combinations of treatments, strains, and harvesting time points with the goal of discovering a subset of genes important for some a priori process of interest. Such "gene selection problems" seem less complex compared to large-scale functional genomics problems (e.g., network estimation [4]-[9]), but in a typical setting they constitute the most common application of microarray techniques. A myriad collection of gene selection criteria and corresponding methods exist; here we are interested in one of the simplest yet commonly used form of gene selection problems, the "pattern-based gene selection" problem. In these scenarios, usually the experimenter sets the conditions such that the candidate gene(s) should conform to some expected *target pattern* as the conditions vary. For example, the experimenter might expect the candidate genes to have a higher expression level in the wild-type strain than in a knock-out strain. However, it is rare for the target pattern to have high specificity in terms of expression level. The target pattern is often qualitative rather than quantitative -- for example, in certain conditions the direction of induction is known, though the fold increase or decrease is not specified. There are several existing approaches to the pattern-based gene selection problem including conformation to an exemplar gene, filtering for high-low band of expected pattern (the two approaches are very commonly used and included in many software implementations; for example, see the GeneSpring software (Agilent) and Section 3.6 in [10]), and fitting parametric linear models [11]. All of these methods have varying degrees of applicability to empirical data with certain strengths (e.g., with sufficient data and right transformation, an ANOVA model is statistically efficient) and problems (e.g., exemplar patterns and high-low band criteria are often ad hoc). However, these existing methods are sub-optimal for translating qualitative biological hypotheses into expected model and testing those models with limited amounts of data.

II. Methods

A. *Isotonic regression*

Let $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ be the list of observations over n conditions where x denotes conditions y denotes the expression values. The problem of isotonic regression is to find a function f such that (1) $f(x_i) \le f(x_j)$whenever $x_i \le x_j$ (the function is isotonic), and (2) the error of f as a regression is minimized. Usually a weighted sum of squared error (SSE) $\sum_{i=1}^{n} w_i (f(x_i) - y_i)^2$ is used. If SSE is used and any two x_i and x_j are always comparable (that is, we have a total order of $\{(x_1,y_1),\ldots,(x_n,y_n)\}$), many algorithms are available that determine f in polynomial time (see [3] for a review). When only a partial order is available, an algorithm is available that computes the exact solution in $O(n^4)$ time and $O(n^2)$ space by solving $O(n)$ minimal flow problems [15]. In the data we analyzed, the number of conditions n is limited, so we use the more naive approach by enumerating all total orders of conditions that conform to the partial order. Computation can be also sped up if we can partition the conditions into sets such that no two conditions from different sets are comparable (i.e., have prior expectations). Then we can apply isotonic regression on each of the sets in the partition as if they are independent problems and concatenate the output.

The regression gives the SSE of each gene, based on which a score is calculated; the score indicates how well the gene fits the criterion. To score each gene, we divide the SSE of the regression by the (weighted) variance of the expression profile of the gene. We then use a permutation test to assign p-values to the scores. The weights in the SSE computation can be set to one or adjusted to fine-tune the behavior of the algorithm. As noted above, scalar constants can be added as additional conditions for each gene. This is very useful for many typical analysis purposes: for example, the expression matrix may be log induction ratios of different control and treatment pairs; adding constant conditions has the similar effect as filtering genes using pre-specified thresholds without the harsh stringency of filtering.

TABLE I

FLORAL PHENOTYPES IN EXPERIMENT 2 (ADAPTED FROM [12])

	Sepal	***Petal***	***Stamen***	***Carpel***
ag	**?**	↑	-	-
ap3	↑	-	-	↑
pi	↑	-	-	↑
ap2	-	-	↓	↑
ap1	-	↓	-	**?**

↑: upregulated; ↓: downregulated; -: missing; ?: questionable.

B. *Microarray Data*

We analyzed a previously published microarray experiment performed using the plant model *Arabidopsis thaliana* with our method and followed up with additional experimental verification using an independent experimental study. The motivation of the experimental design in [12] was to discover specific genes for flowering organs (sepals, petals, stamens, and carpels) in *Arabidopsis thaliana* by correlating gene expression profiles with homeotic phenotypes (absent/fewer organs/normal/extra organs; see Table 1) across different mutant strains. The original dataset consisted of measurements from two platforms (GEO ID: cDNA: GDS865, oligo: GDS866, GDS867). Each gene in the original dataset consisted of ratios of expression levels in five different mutant strains (*ap1, ap2, ap3, pi, ag*; in the literature they are often denoted as *ap1-1, ap2-2, ap3-3, pi-1, ag-3*, respectively) over those in the wildtype strain. The authors took a pairwise-comparison approach in selecting genes. First, for each organ, a constraint on pairwise orderings was determined according to the phenotypical pattern of the organ across the five mutant strains. Then for every pairwise constraint that required ratio for mutant strain X to be higher than that for strain Y, a gene is significant if (1) it is significantly differentially expressed in a t-test (the p-value threshold was determined using false discovery rate control [13] with rate ≤ 0.05), (2) there was twofold or more in the difference between the mean expression ratios of two strains. A gene is specific for an organ if it satisfied all pairwise constraints except for petals, where a gene has considered significant if at least three of the four pairwise constraints held.

C. *Transcriptional Target Identification*

We translated the phenotypes into partial orders using the following rules of thumb: (1) any up-regulation in phenotype (↑ in Table 1) means the log$_2$-ratio (mutant vs. wildtype) should be greater than 1; (2) any down-regulation (↓ in Table 1) or missing (-- in Table 1) in phenotype means the log$_2$-ratio (mutant vs. wildtype) should be lower than -1; (3) any ratio associated with down-regulation (but not entirely missing) in phenotype should be greater than any ratio associated with missing. Let ap1, ap2, ap3, ag, pi represent their respective mutant-vs-wildtype log2-ratios. We formulated the following conditions:

(Sepal) ag=0; ap1, ap2<-1; ap3, pi>1
(Petal) ag1>1; ap1, ap2, ap3, pi<-1; ap1>ap2, ap3, pi
(Carpel) ap1=0; ap2, ap3, pi>1; ag<-1
(Stamen) ap2, ap3, pi, ag<-1; ap2>ap3, pi, ag

TABLE IV

COMPARISON OF SUCCESS RATES IN IDENTIFYING ORGAN-SPECIFIC GENES.

		Carpel	**Petal**	**Sepal**	**Stamen**
(a)	(1) *No. Signif.Genes*	231	15	13	987
	(2) *No. organ-specific genes*	99	4	1	425
	(3) *Specificity*	42.9%	26.7%	7.7%	43.1%
(b)	(1) *No. Signif.Genes*	139	17	151	745
	(2) *No. organ-specific genes*	71	4	21	346
	(3) *Specificity*	51.1%	23.5%	13.9%	46.4%

Group (a): Original list of genes in [12]; (b): Genes identified using our new approach (FDR=0.05). Row (2) in each group is based on AtGenExpress. For each organ, each of the two methods (original list of genes in [12] and our new approach) has three numbers: (1) the number of significant genes according to the gene selection procedure, (2) the number of significant genes that are truly organ-specific according to AtGenExpress, and (3) the ratio of the two numbers (the specificity of the method). For more details please see the Methods section.

We treat the ambiguous case (? in Table 1) for each organ differently, according to the original paper. For carpel, mutant ap1 has the same number of carpels as the normal type according to [12], hence we require ap1=0 (not differentially expressed). For stamen, the outcome of ap1 mutant is not mentioned in the paper, so we do not include ap1 in our criterion. For sepal, the criterion in [12] requires ag to be unchanged or up-regulated; in our analysis we include ag=0 in our criterion.

We removed any gene that has one or more NA's (not available) in the original data. In the isotonic regression we also weighted the constants (0, 1 and -1) 100 times higher than the ordinary conditions in the microarray, so the fitted values for these constants are very close to the constants themselves (this effectively forces the ratios to obey comparisons with constants as much as possible). After the genes were ranked (see section II.A. for details), we generated p-values using a permutation test procedure: we first generated a null dataset by randomly scrambling conditions independently five times for each gene. We then scored each sample from the null dataset using isotonic regression. The p-value of a gene with score *s* in the original dataset is defined as the fraction of scores in the null dataset lower than *s*.

We used the AtGenExpress dataset [14] to validate the selected genes and compare our gene lists with those from the original paper. We compared the expression levels of genes in carpel, stamen, sepal, and petal collected from flowers in flowering stage 15; each organ has 3 replicates. A gene was deemed organ-specific for carpel if (1) the mean expression level in carpel is significantly higher than those in other three organs according to the Tukey's HSD test in ANOVA with 0.05 significance level[1], and (2) the mean expression level of

[1] The Tukey HSD test takes the ANOVA outcome, a significance level p, and two levels in the factor (in our case, two organs), and returns a 100(1-p)% confidence level on the difference of the two means. If 0 is outside the confidence level, then we declare the two means are significantly different. We tried different significance levels, from 0.001 to 0.05, and see very little difference – thresholding the difference in mean expression levels has a much

the three replicates in carpel is twice as high or higher than that in either of the other three organs; we define organ-specificity for the other three organs similarly. For each list of carpel-specific genes in [12] and that produced using our approach, the *specificity* of the list is the proportion of the genes being carpel specific (i.e., expressed preferentially in carpels according to AtGenExpress), out of all genes common in both AtGenExpress and the microarray platform in [12]. The specificities for the other three organs were computed similarly.

III. RESULTS

The results are presented in Table 2. For carpels and stamens, the number of genes in our list is smaller than those in the original publication list, but our specificity (the percentage of significant genes that are truly organ-specific according to AtGenExpress; see the Methods section) was higher; this is mostly likely due to the lower sensitivity but higher specificity in most non-parametric and semi-parametric statistical methods. The specificity for petal was slightly lower in our list, although the sizes of both lists were too small for us to confidently draw any conclusions. Finally, there was a large difference in sepal-specific gene lists: the original study in [12] only identified 13 genes (1 of which was truly sepal-specific according to AtGenExpress) while our method produced 151 genes, 21 of which were truly sepal-specific. Our list also had almost twice the specificity than that of the original list. This comparison demonstrates that by combining all pairwise ordering constraints in a single partial/total ordering, our method was able to avoid some of the problem of excessive stringency typical in combining multiple hypothesis testing results as in [12].

IV. CONCLUSION

A key to genomic analysis is incorporating prior knowledge or expectations into the data analysis. A semi-parametric framework provides an attractive approach where the prior knowledge is used to setup partial order constraints for possible outcomes and then a quantitative measure is used to assess departures from the constraints. We believe that this class of methods might be the most appropriate framework for small datasets with few replicates as often used in small-scale genomic studies.

As proof of principle, we re-analyzed a previously published microarray dataset and followed up with independent experimental verification. In the experiment – designed to find floral organ specific genes – our isotonic regression approach provides a canonical translation of qualitative criteria into partial order constraints; though we do not aim to exactly reproduce the original criteria, our approach ranks highly those genes chosen previously. Moreover, the true positive identification rate was higher than that previously published for three out of four organs analyzed.

Genome-scale analyses often yield surprising insights such as the suggested statistical regularity of biological pathways (e.g., see [5][7]). However, another important application of large-scale functional genomics data is in aiding targeted studies within existing biological research programs. Data analyses for such studies involve unique challenges including balancing cost and power, incorporating prior knowledge and predictions. As genomics technology becomes incorporated into common experimental procedures, we expect the majority of applications of this technique to be limited by these constraints. An important challenge for computational method development is to allow empirical researchers to apply a principled method of deduction. We believe semi-parametric approaches as we have taken here will be an important addition in this direction.

ACKNOWLEDGMENT

We thank Dilusha William and Don Baldwin for discussion and for critical comments on this manuscript.

REFERENCES

[1] S. Dudoit, Y.H. Yang, M.J. Callow, and T.P. Speed, “Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments,” *Statistica Sinica* vol. 12, no. 1, pp. 111-139, 2002.

[2] K.P. Burnham and D. Anderson, *Model selection and multi-model inference.* Springer, New York, 1998.

[3] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and Z.H.D. Brunk, *Statistical Inference under Order Restrictions: the Theory and Application of Isotonic Regression*. New York: Wiley Publishing; 1972.

[4] U.S. Bhalla and R. Iyengar, “Emergent properties of networks of biological signaling pathways,” *Science* 1999, vol. 283, pp. 381-387.

[5] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, “Using Bayesian networks to analyze expression data,” *J. Comput. Biol.*, vol. 7, pp. 601-620, 2000.

[6] S.L. Lauitzen, *Graphical model*s. Oxford University Press, 1996.

[7] P.M. Magwene and J. Kim: “Estimating genomic coexpression networks using first-order conditional independence,” *Genome Biol.,* vol. 5, pp. R100, 2004.

[8] A. Butte A and I. Kohane, “Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements,” *Pac. Symp. Biocomput.,* vol 5, pp. 418-429, 2000.

[9] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, “Cluster Analysis and Display of Genome-Wide Expression Patterns,” *Proc. Natl. Acad. Sci. USA,* vol. 95, pp. 14863-14868, 1998.

[10] I.S. Kohane, A. Kho, and A.J. Butte, *Microarrays for an Integrative Genomics,* MIT Press, 2002.

[11] M.K. Kerr, and G.A. Churchill, “Experimental design for gene expression microarrays,” *Biostatistics*, vol. 2, pp. 183-201, 2001.

[12] F. Wellmer, J.L. Riechmann, M. Alves-Ferreira, and E.M. Meyerowitz. “Genome-wide analysis of spatial gene expression in Arabidopsis flowers,” *Plant Cell*, vol. 16, no.5, pp. 1314-1326, 2004.

[13] Y. Benjamini and Y. Hochberg Y, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *J. Roy. Stat. Soc. Ser. B*, vol. 57, pp. 289-300, 1995.

[14] M. Schmid, T.S. Davison, S.R. Henz, U.J. Pape, M. Demar, M. Vingron, B. Schölkopf, D. Weigel, and J. Lohmann, “A gene expression map of Arabidopsis development,” *Nat. Genet.*, vol. 37, pp. 501-506, 2005.

[15] W.L. Maxwell and J.A. Muchstadt, “Establishing consistent and realistic reorder intervals in production-distribution systems,” *Oper. Res.*, vol. 33, pp. 1316-1341, 1985.

[16] R Development Core Team, *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

stronger effect. Also see the description of the command TukeyHSD in the R manual [16].