Statistical tests

In the program, ANOVA and Tukey's honestly significantly difference (HSD) test are used to evaluate and to group patterns. ANOVA measures the differences between means of more than two groups, and its null hypothesis is that the tested means are the same (1). The result of the ANOVA is the *F*-score. Larger *F*-scores indicate larger difference and discrimination between the groups than smaller ones. In the program, the *F*-score of each pattern is calculated from the bootstrap simulation data sets (from the estimated means and deviations in the input promoter sets and in the background promoter collection) according to equations 1-7.

(1) $$SS_T = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{3l}$$

(2) $$SS_G = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{l}$$

(3) $$SS_{WG} = \sum SS_G$$

(4) $$SS_{BG} = SS_T - SS_{WG}$$

(5) $$MS_{BG} = \frac{SS_{BG}}{d-1}$$

(6) $$MS_{WG} = \frac{SS_{WG}}{3l-d}$$

(7) $$F = \frac{MS_{BG}}{MS_{WG}}$$

In the equations 1 to 7, *SS* is the sum of squares, *MS* is the mean squares, *d* is the number of data sets (here: three), *l* is the number of repeats in the bootstrap simulation (the number of bootstrap samples), and X_i and X_i^2 are the pattern occurrence count and the count to the power of two in the *i*th sample. Equation 1 is total *SS* of the data sets (combines the input promoters sets one, two and the background promoter collection together). Equation 2 is the *SS* of one data set. Equation 3 is the sum of the data sets *SS* values and represents *SS* within groups, which is the measure of the variability that exists

inside the data sets. Equation 4 is the *SS* between groups, which is the measure of the aggregate differences among the means of the data sets. Equations 5 and 6 are *MS* and the equation 7 gives the *F*-score.

After calculating the *F*-scores, the program groups the patterns into five groups, which are: groups *1* and *2* patterns over-represented only either in the first (*1*) or the second (*2*) input promoter set, group *3* patterns over-represented in both input promoter sets, and groups *4* and 5 patterns over-represented in the first input promoter set and under-represented in the second (*4*) or vice versa (*5*). The grouping is done by calculating the HSD-test (equation 8) for each pair of the three data sets and comparing the results. The HSD-test calculates the quantity of difference between two data sets, where $Q > 0$ indicates that the pattern is over-represented in the first data set and $Q < 0$ indicates that the pattern is over-represented in the latter data set.

(8)
$$Q = \frac{X_A - X_B}{\sqrt{\left(MS_{WG} / l \right)}}$$

In equation 8, Q is the result of the HSD test, X_A and X_B are the two estimated means, MS_{WG} is the mean square of the within groups value from ANOVA (equation 6) and l is the number of repeats in the bootstrap simulation.

The significance of the patterns is reported to the users with *p*-values. The *p*-value is the probability to find a larger *F*-score from the data by chance, and it is calculated by using the standardized f-distribution (1). In the f-distribution, each *F*-score has a corresponding *p*-value and the interesting patterns are found in the extreme right-hand tail of it. In the program, *F*-scores for a comprehensive set of patterns are being calculated and their distribution is analyzed. The *F*-scores constitute an f-distribution, which can differ from the standardized f-distribution by its mean and/or deviation. When there is a difference, the observed f-distribution is translated to correspond on the standardized f-distribution with two parameters, location and scale. The location defines the difference between the mean of the *F*-scores and the mean of the standardized f-distribution (Equation 9), and can be used to shift the distribution's mean. The scale is the ratio of the deviation of the *F*-scores and the deviation of the standardized f-distribution (Equation 10), and can be used to stretch or squeeze the distribution shape. Before calculating the *p*-

values, the *F*-scores are translated with these two parameters (Equation 11) and, as a result, *p*-values that are highly consistent with ones in the random data are obtained.

(9) $$\mu_{stand} = \frac{v_2}{v_2 - 2}$$

(10) $$\sigma_{stand} = \sqrt{\frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)}}$$

(11) $$F_{scale} = \frac{F_{obs} - (\mu_F - \mu_{stand})}{(\sigma_F / \sigma_{stand})}$$

In equations 9, 10 and 11, v_1 and v_2 are the degree of the freedom of the numerator (*d–1* in the equations 1-8) and denominator (*31-d*), μ_{stand} is the mean and σ_{stand} is the deviation of standardized f-distribution, μ_F is the mean and σ_F is the deviation of the obtained *F*-scores of the group where the pattern belongs, F_{obs} is the *F*-score to be translated and F_{scale} is the *F*-score used to calculate the p-value.

Accuracy of the estimated bootstrap simulation means

Artificial data was used to monitor the accuracy of the means estimated by bootstrap simulation, when 20 promoters were selected. The length of patterns and repeat number were varied. The results present the accuracy of the estimated background promoter collection mean; the mean is used to detect the relative over- and under-presentation.

In the test, the estimated bootstrap simulation means from 1,000,000 random patterns were computed and compared against the real occurrences of these patterns in the *A. thaliana* background promoter collection, promoter length 1,500 bp. Random patterns were 10 bp (A, C, G, T or N) long and patterns containing less than four specified nucleotides (A, C, G, T) were rejected. For each pattern the normalized difference (*diff*) between the estimated (*bootstrap*) and the real mean (*real*) was calculated (Equation 12).

$$diff = \frac{|X_{real} - X_{bootstrap}|}{(X_{real} + X_{bootstrap})} \qquad (12)$$

The bootstrap simulations were calculated 10 times, to diminish the variation of the random sampling. The means and the standard deviations of the normalized differences over all the patterns and over all the calculations are shown in Table 1, where *diff*=1 is an inaccurate and *diff*=0 is an accurate estimation. Results are divided into columns, according to the number of specified nucleotides (A, C, G, T rather than N) in the random patterns. From the table, we can find out the suitable number of repeats that is needed in the bootstrap simulation to generate an accurate comparison point. For example, when analyzing 8 nucleotides long patterns, an appropriate number of repeats is 160. At this point, the value of the normalized difference is ~0.05 for the longest possible patterns (8 bp) whereas patterns with N-wildcards have smaller normalized differences.

_F_-score vs. _Z_-score statistics in the WRKY70-experiment

Current pattern finding programs discover patterns that are statistically over-represented in a single input promoter set. For example, the *Z*-score, which is the number of standard deviations by which the observed value differs from its expectation, has been used for ranking (2). In our program when two input promoter sets are used, patterns are evaluated using *F*-scores, which can detect patterns that are over-represented in one promoter set and under-presented in the other. In order to study which test statistic better distinguishes random patterns from the possible biological patterns, we compared the obtained *F*-scores against their corresponding *Z*-scores from the WRKY70-experiment.

Each detected pattern is shown in Figure 1 where the y-values are patterns' *F*-scores and x-values are patterns' *Z*-scores either from the up-regulated (black) or from the down-regulated promoter set (gray). Tables 2, 3 and 4 present the five best *Z*- or *F*-score patterns in more detail. For comparison, the 1, 5, 10 and 30 top patterns' mean *F*-score and deviation in similar random analyses is shown in Table 5. In the random analyses, twenty promoter sets (each having 20 promoters) were analyzed with the same parameters as the WRKY70 analysis.

High *F*-scores do not always correspond to high *Z*-scores (Figure 1), but this information itself does not reveal which is better. However, an interesting detail is observed with further exploration. Patterns with top *F*-score ranks have a higher frequency of occurrence than patterns with top *Z*-score rank. For example, the maximum

number of promoters with the pattern in Table 2 is 3 (totally 10 promoters), in Table 3 it is 24 with high *F*-score and 4 without a high *F*-score (totally 24 promoters).

It has been speculated that in higher organisms, biologically functional patterns are multiplied (*S. cerevisiae* (3), *A. thaliana* (4, 5). This would allow the cell to maintain the correct expression and conduct random mutations of one or some of the patterns. If the theory is believed, biologically functional patterns are expected to occur, first, in most of the co-expressed genes' promoters, and second, multiple times in these promoters. For example, in the WRKY70-experiment there were 24 genes in the up-regulated input set and 10 genes in the down-regulated input set, leading to the expectation of between 48 (duplication) and 72 (triplication) pattern occurrences in the up-regulated promoter set, and between 20 and 30 occurrences in the down-regulated promoter set. Clearly, these numbers are far from the observed occurrences 4 and 3 for the patterns topping the *Z*-score ranked list. Another failure in the low occurring patterns is that they cannot successfully explain the expression. For example, a pattern occurring within 4 of the 24 tested promoters can at most explain one sixth of the expression, since it can only regulate the 4 genes in which it is found. We conclude that *F*-scores are capable of detecting patterns with a higher frequency of occurrence than patterns detected with *Z*-scores, and therefore *F*-scores can find patterns that could explain the whole expression, rather than some part of it.

The possible biological function of the found patterns is described in the main text. Here, we note that the Z-score discovers only one pattern (GACTNNNA) among the top five patterns from the WRKY70-experiment (Tables 2 and 3) that could be the binding site of the most probable regulator, the WRKY70 protein, and this pattern also has a high *F*-score.

1. Petruccelli,J.D., Nandram,B. and Chen,M. (1999) Applied Statistics for Engineers and Scientists. Prentice-Hall inc.

2. Sinha,S. and Tompa,M. (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation, *Nucleic Acids Res*., **30**, 5549-5560.

3. Harbison,C.T., Gordon,D.B.,Lee,T.I.,Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J., Jennings,E.G., Zeitlinger,J., Pokholok,D.K., Kellis,M., Rolfe,P.A., Takusagawa,K.T., Lander,E.S., Gifford,D.K., Fraenkel,E. and Young,R.A. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99-104.

4. Schöffl,F., Prandl,R. and Reindl,A. (1998) Regulation of the heat-shock response, *Plant Physiol*ogy, **117**, 1135-41.

5. Hobo,T., Asada,M., Kowyama,Y. and Hattori,T. (1999) ACGT-containing abscisic acid response element (ABRE) and coupling element 3 (CE3) are functionally equivalent, *Plant Journal*, **19**, 679-89.

Table 1. Bootstrap simulation accuracy on relation to the number of repeats. In the table *avg* is the mean and *sd* is the standard deviation of the normalized differences. Columns present the number of nucleotides (A, C, G and T) in the analyzed patterns. Numbers of analyzed patterns were: 134410, 301849, 302357, 176120, 65806, 16671 and 2787.

	10		9		8		7		6		5		4	
Repeats	*avg*	*sd*	*avg*	*sd*	*avg*	*sd*	*avg*	*sd*	*avg*	*sd*	*avg*	*sd*	*avg*	*sd*
5	0.90	0.24	0.65	0.38	0.32	0.32	0.14	0.15	0.07	0.06	0.03	0.03	0.02	0.02
10	0.80	0.32	0.49	0.39	0.22	0.24	0.10	0.10	0.05	0.04	0.02	0.02	0.01	0.01
20	0.66	0.38	0.34	0.34	0.15	0.16	0.07	0.07	0.03	0.03	0.02	0.01	0.01	0.01
40	0.51	0.39	0.23	0.26	0.10	0.11	0.05	0.05	0.02	0.02	0.01	0.01	0.01	0.01
80	0.37	0.35	0.16	0.19	0.07	0.07	0.03	0.03	0.02	0.01	0.01	0.01	0.00	0.00
160	0.25	0.29	0.11	0.13	0.05	0.05	0.02	0.02	0.01	0.01	0.01	0.01	0.00	0.00
320	0.17	0.21	0.08	0.09	0.03	0.03	0.02	0.02	0.01	0.01	0.00	0.00	0.00	0.00
640	0.12	0.15	0.05	0.06	0.02	0.02	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00
1280	0.08	0.11	0.04	0.04	0.02	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
2560	0.06	0.07	0.03	0.03	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00

Table 2. The five top patterns from the WRKY70-experiment according to their *Z*-score in the up-regulated cluster. In the table *occ* is the number of the patterns and *pro* is the number of promoters with the pattern.

	Up-regulated				*Down-regulated*						
Pattern	*occ*	*pro*	*Z-score*	*Rank*	*occ*	*pro*	*Z-score*	*rank*	*F-score*	*p*	*rank*
GGGTCTCC / GGAGACCC	4	4	7.04	1	0	0	-0.44	79996	3648.45	9.5E-02	4138
TNNNAGTC / GACTNNNA	110	24	6.01	2	19	9	-1.32	117410	14690.51	3.5E-06	4
CGTAGCAT / ATGCTACG	4	4	5.96	3	0	0	-0.45	80370	3247.65	1.4E-01	5764
GAGACCCT / AGGGTCTC	4	4	5.88	4	0	0	-0.54	83894	3343.59	1.3E-01	5311
CCTAAGGT / ACCTTAGG	4	3	5.83	5	1	1	3.26	4076	1129.42	1.6E-01	41769

Table 3. The five top patterns from the WRKY70-experiment according to their *Z*-score in the down-regulated cluster. Notation as in Table 2.

	Up-regulated				*Down-regulated*						
Pattern	*occ*	*pro*	*Z-score*	*rank*	*occ*	*pro*	*Z-score*	*rank*	*F-score*	*p*	*rank*
TGGCCTGC / GCAGGCCA	0	0	-0.41	86877	3	3	14.25	1	8209.54	6.4E-03	188
GCGGTAGG / CCTACCGC	0	0	-0.28	79476	2	2	13.39	2	4777.47	8.0E-02	1734
GGGGAGGT / ACCTCCCC	1	1	1.34	12392	3	3	12.51	3	5874.39	6.1E-03	820
GTACGGCG / CGCCGTAC	0	0	-0.31	81170	2	2	12.32	4	4710.79	8.4E-02	1818
GGCCTGCG / CGCAGGCC	0	0	-0.30	80589	2	2	11.98	5	4783.66	8.0E-02	1722

Table 4. The five top patterns from the WRKY70-experiment according to their *F*-score. Notation as in Table 2.

	Up-regulated				*Down-regulated*						
Pattern	*occ*	*pro*	*Z-score*	*rank*	*occ*	*pro*	*Z-score*	*rank*	*F-score*	*p*	*rank*
TTTNNACT / AGTNNAAA	70	23	2.64	1280	7	5	-4.00	135371	17120.47	3.7E-07	1
GGGNNNTG / CANNNCCC	13	9	-2.13	134734	24	10	5.04	719	15524.94	2.9E-05	2
CNNAGNGG / CCNCTNNG	21	11	0.07	59775	26	9	7.69	73	15233.91	4.2E-07	3
TNNNAGTC / GACTNNNA	110	24	6.01	2	19	9	-1.32	117410	14690.51	3.5E-06	4
TCNGNGC / GCNCNGA	7	7	-1.87	133659	17	9	5.06	708	14230.12	7.6E-05	5

Table 5. Mean *F*-score of the top *n* random pattern. In the table *n* is the number of the patterns, *avg* is the mean and *sd* is the standard deviation of the *F*-scores.

	n=1		*n=5*		*n=10*		*n=15*		*n=30*	
Group	*avg*	*Sd*	*avg*	*sd*	*avg*	*sd*	*avg*	*sd*	*avg*	*sd*
1 and 2	5698.42	979.32	4699.59	787.33	4279.03	735.64	4027.57	721.20	3616.61	688.24
3	3098.34	475.33	2667.74	456.25	2421.87	425.64	2277.91	410.14	2043.23	384.50
4 and 5	10171.77	982.07	9081.89	1037.48	8408.61	1059.38	7982.73	1095.65	7234.42	1126.45

Figure 1. All found patterns' *Z*-scores and *F*-scores from the WRKY70-experiment. In the figure black dots are the *Z*-scores of the up-regulated promote set and gray dots are the *Z*-scores of the down-regulated promoter set. The bigger symbols are the positions of the five best *F*- and *Z*-score patterns

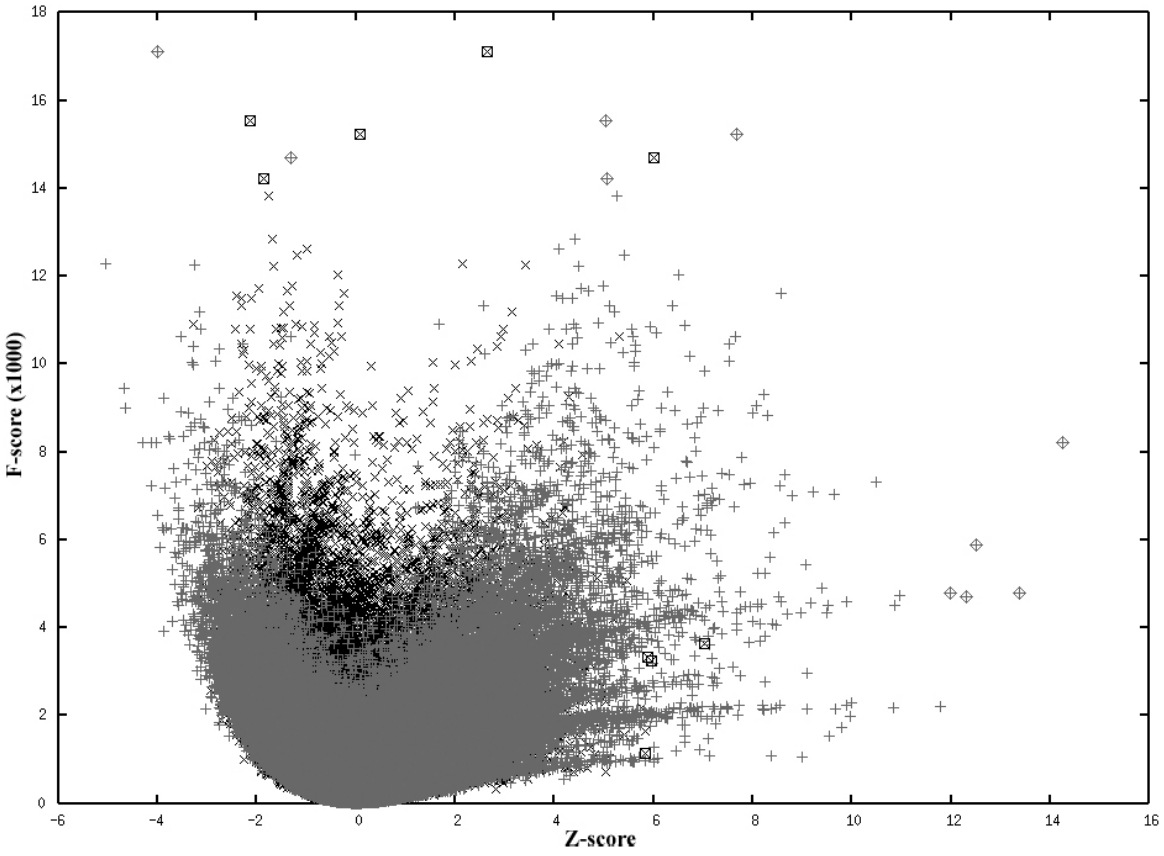

Figure 2. The interface and output of POCO.