Supplementary Methods for: Combining Static and Time Series Data to Determine the Quality of Expression Profiles in Time Series Experiments

Itamar Simon[1], Siegfried Zahava[1], Jason Ernst[2] and Ziv Bar-Joseph[2,3,*]

[1]Dept. Molecular Biology, Hebrew University Medical School, Jerusalem, Israel 91120

[2] School of Computer Science and [3]Department of Biology, Carnegie Mellon University

5000 Forbes Ave. Pittsburgh, PA, 15213

∗To whom correspondence should be addressed: zivbj@cs.cmu.edu

I. The Checksum method: Hypothesis testing

Computing the probability under the null hypothesis: We first present a method for computing $P(W_i|M_i, H_0)$, which takes into account the actual expression values of i. This method relies on noise model for individual samples and thus can be computed even if only few repeats exist.

Here we assume that M_i correctly captures the average expression level for i. In Section II below we explain how our method can be modified to deal with noisy measurements of this average. Under the null hypothesis, we assume that C_i is a noisy realization of the true underlying expression profile for i. In order to determine the probability under the null hypothesis, we look for a new curve C such that $W(C) = M_i$ and C can explain C_i. In other words, we now wish to explain C_i as a noisy realization of C (which correctly captures the unsynchronized expression for i), and the differences between M_i and W_i as a result of that noise. Recall that C_i was generated from Y_i (the set of sampled value for i) and thus, noise in C_i is a function of the values of Y_i. Let Y_C represent a set of values (corresponding to the experiment time points) that generate the curve C. In order to find a curve C that best explains C_i we solve the following maximization problem:

$$max_{Y_C} P(Y_i|Y_C) \qquad \text{such that} \qquad \frac{1}{V}\int_{t_s}^{t_e} \frac{C}{M_i} dt = 1 \tag{1}$$

Here we are looking for a curve C with $W(C) = M_i$, such that the observed values for i (Y_i) are a noisy sample of the values used for generating C (Y_C). Using the solution of the above problem we set $P(W_i|M_i, H_0) = P(Y_i|Y_C)$. This guarantees that only profiles that cannot be explained by the best (maximum likelihood) curve with an average expression of M_i will be detected.

Solving the maximization problem: We assume a Gaussian *value dependent* noise model for $P(Y_i|Y_C)$. That is, the variance of is expression value at time t depends on the actual measured value at that time, $Y_i(t)$ (see Section II below). We represent both the curves (C and C_i) and the expression values (Y_C and Y_i) using splines. Denote by F_C and F_i the spline control points for C and C_i respectfully. As discussed in Results and Box 3, $C(t) = s(t)F_C$ where $s(t)$ is the set of spline coefficients evaluated at time t. As for the observed expression values, using our continuous representation approach (Box 3) we can set $Y_C = SF_C$ and $Y_i = SF_i$[1]. We can now write Equation 1 as fellows:

$$max_{F_C} P(S'F_i|S'F_C) \qquad \text{such that} \qquad \frac{1}{V}\int_{t_s}^{t_e} s(t)F_C dt = M_i \tag{2}$$

where S' is a scaled version of S such that value dependent variance is incorporated into our maximization problem.

As we show in Section III below, the above maximization can be cast as a quadratic programming problem for which we can obtain the global maxima.

Hypothesis testing: Using the solution of Equation 2 we can now perform a log likelihood ratio test for our two hypotheses. We first set $P(W_i|M_i, H_1) = P(W_i)$, since W_i and M_i are independent under H_1. For H_0, we set $P(W_i|M_i, H_0) = P(S'F_i|S'F_C)$ were F_C is obtained by solving the maximization problem above. Thus, the log likelihood ratio evaluates to:

$$2\log\frac{p(W_i|M_i,H_1)}{p(W_i|M_i,H_0)} = 2\log\frac{e^{-(S'F_i-S'F_i)^T(S'F_i-S'F_i)}}{e^{-(S'F_i-S'F_C)^T(S'F_i-S'F_C)}} = 2(S'F_i - S'F_C)^T(S'F_i - S'F_C) \tag{3}$$

To evaluate this ratio test we use the chi-square distribution with q degrees of freedom (where q is the number of spline control points).

Figure 6 presents the complete Checksum algorithm. This algorithm runs in time that is linear in the number of genes. As we discuss in Results, the set of genes detected by this algorithm can be further examined by considering the direction of the difference (either $W_i > M_i$ or $W_i < M_i$).

II. Handling noise in M and Y

As mentioned above our method can handle noise in the measurements of the static and time series expression data.

While there is reason to believe that the values obtained for M_i will be less noisy than the values obtained for the time course experiment (either because in many cases the unsynchronized measurement is repeated or because of the smoothing effects that averages exhibit when analyzing stochastic data), our algorithm can be modified to deal with a noisy value for M. Given a noise model for M (either from repeats, or using the same noise model used for Y_i) we obtain new upper

[1] Note that we omit the noise term from this equation since we are now using the predicted splines for i.

Supplementary Figure 1: The CheckSum algorithm

```
CheckSum($Y, M, \delta$) {                    // $\delta$ is the p-value cutoff
        For all genes $i$
              Compute spline assignment ($C_i$) and control points($F_i$)
        For all genes $i$ {
              Let $F_C$ be the solution to the maximization problem 3 using $F_i$ and $M_i$
              Set $r = 2(S'F_C - S'F_i)^T(S'F_C - S'F_i)$
              $s = 1-$ cdf of the chi-square distribution for $r$ with $q$ d.o.f
              If $s < \delta$ output $i$
        }
}
```

and lower bounds for M, and re-run the algorithm above for these two values. Only if neither of them can be explained by W_i do we pick i as a gene with a missing critical measurement.

As for value dependent noise in the time series measurements, below we briefly describe a method to incorporate value dependent variance into our continuous representation model. The reader is referred to [1] for more details.

Our framework (equation 1 above) can be modified to use variances that depend on expression value magnitudes. Instead of maximizing $p(Y_i|Y_C)$ assuming the same variance σ^2 for all values, we maximize $p(Y_i|Y_C, \sigma_1^2 \ldots \sigma_n^2)$ where $\sigma_1^2 \ldots \sigma_m^2$ are the m expression value specific variances for the samples in Y_i. Recall that the rows of S (the spline basis function matrix) correspond to the time points that were sampled in the reference experiment. Denote by S_i the i^{th} row of S. Let $S'_i = S_i/\sigma_i$. Then maximizing $p(Y_i|Y_C, \sigma_1^2 \ldots \sigma_n^2)$ is equivalent to minimizing $||S'F_i - S'F_C||$, and so we proceed by replacing S with S' as presented in equation 2. This results in the differential weighting of the individual errors around Y_i, allowing for greater differences between W_i and M_i for genes with higher expression values.

III. The quadratic programming problem

Recall that we are trying to solve the following maximization problem:

$$max_{F_C} P(S'F_i|S'F_C) \qquad \text{such that} \qquad \frac{1}{V}\int_{t_s}^{t_e} s(t)F_C dt = M_i \tag{4}$$

Since we assume a Gaussian noise model for $P(S'F_i|S'F_C)$, problem 4 could be written as the following minimization problem:

$$min_{F_C}||S'F_i - S'F_C|| \qquad \text{such that} \qquad A^T F_C = M_i \tag{5}$$

where $||$ is the L_2 norm and A is a vector obtained by integrating out the coefficients for each of the entries of the vector F_C. This results in the following quadratic programing problem:

$$min_{F_C} 0.5 F_C(S'^T S')F_C + (S'^T S'F_i)F_C \qquad \text{subject to} \qquad A^T F_C = M_i \tag{6}$$

We solve the above problem using the Matlab [2] function ’quadprog’, which uses a subspace trust-region method based on the interior-reflective Newton method described in [3]. For non convex problems, ’quadprog’ returns a local minimum solution. However, in all cases we have looked at, $S'^T S'$ was a positive definite matrix, and thus a global minimum solution was obtained.

References

[1] Z. Bar-Joseph, G. Gerber, T.S. Jaakkola, D.K. Gifford, and I. Simon. Comparing the continuous representation of time series expression profiles to identify differentially expressed genes. *PNAS*, 100(18):10146–51, 2003.

[2] *The MathWorks*. URL: http://www.mathworks.com/.

[3] T.F. Coleman and Y. Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM Journal on Optimization*, 6(4):1040–58, 1996.