Supplementary Materials for 'Robust and accurate data enrichment statistics via distribution function of sum of weights'

Aleksandar Stojmirović, and Yi-Kuo Yu

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, MD 20894
United States

1 Saddlepoint approximation of tail probabilities

References about saddlepoint approximations of the tail probabilities of random variables are abundant [5, 3, 7, 4]. For completeness of our exposition we here present the derivation of the Lugannani-Rice formula [5], relying extensively on expositions by Daniels [3] and Woods, Booth and Butler [7].

Let X be a continuous random variable supported on a subset of $\mathbb{R}$. We will assume that its probability density function (PDF), denoted by f_X exists and that its moment generating function (MGF), defined by $\rho_X(t) = \int_{-\infty}^{\infty} f_X(x)e^{tx}\,dx$ converges for real $t \in [a, b]$ where $a < 0 < b$. Recall that $\rho_X(it)$ gives the characteristic function of X, that is, the Fourier transform of f_X and that f_X can hence be recovered by the Fourier inversion formula:

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\rho_X(it)\,dt \tag{1}$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{K_X(it)-itx}\,dt \tag{2}$$

$$= \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} e^{K_X(t)-tx}\,dt, \tag{3}$$

where $K_X(t) = \ln \rho_X(t)$ denotes the cumulant generating function (CGF) of X. The tail probability or P-value for a value y (with respect to X), which we will denote by $Q_X(y)$ can be expressed as

$$Q_X(y) = \mathrm{Prob}(X \geq y) = \int_y^{\infty} f_X(x)\,dx \tag{4}$$

$$= \frac{1}{2\pi i}\int_y^{\infty}\int_{-i\infty}^{i\infty} e^{K_X(t)-tx}\,dt\,dx \tag{5}$$

$$= \frac{1}{2\pi i}\int_{-i\infty}^{i\infty}\int_y^{\infty} e^{K_X(t)-tx}\,dx\,dt \tag{6}$$

$$= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^{K_X(t)-ty}\,\frac{dt}{t}, \tag{7}$$

where $c \in (0, b)$ is a constant introduced to avoid the pole at $t = 0$.

Let S denote the sum of m independent, identically distributed random variables. We write $S = \sum_{j=1}^{m} X_j$, where $f_{X_j} = f_X$ for all j. Our goal is to derive an asymptotic approximation for the tail probability Q_S. It can be easily shown that $\rho_S(t) = \rho_X^m(t)$ and hence by (8)

$$Q_S(s) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^{mK_X(t)-ts}\,\frac{dt}{t}. \tag{8}$$

To produce our approximation we note that the main contributions to the integral (8) occur in the neighborhood of the pole at $t = 0$ and in the neighborhood of the saddle point $t = \hat{\lambda}$ where the exponent $I(t) = mK_X(t) - ts$ has a maximum, that is, where $I'(t) = 0$. The saddlepoint condition is thus

$$s = mK'_X(\hat{\lambda}), \tag{9}$$

or alternatively

$$\int_{-\infty}^{\infty} \left(x - \frac{s}{m}\right) f_X(x) e^{\hat{\lambda}x}\, dx = 0. \tag{10}$$

Let $\mathbb{E}(X)$ denote the expectation of X. Daniels [2] has shown that eq. (9) has a unique simple root under most conditions. The value of $\hat{\lambda}$ increases with s, with $\operatorname{sgn}(\hat{\lambda}) = \operatorname{sgn}(s - m\,\mathbb{E}(X))$.

When $s \gg m\,\mathbb{E}(X)$, the contribution of the pole at $t = 0$ to (8) is very small and to obtain an asymptotic approximation to Q_S one can proceed by expanding $I(t)$ as a Taylor's series about $t = \hat{\lambda}$ and integrating the resulting integral term-by-term [3]. However, as s gets closer to the mean $\mathbb{E}(S) = m\,\mathbb{E}(X)$, such approximation performs poorly and in fact is unbounded at the mean. The essence of the method of [1] as applied to Q_S by Lugannani and Rice [5] is to produce a transformation of the integral (8) that would take into account the pole and hence to produce an approximation uniformly valid over the whole range of S.

Make a transformation from t to a new variable z by

$$K_N(z) - \hat{z}z = mK_X(t) - ts, \tag{11}$$

where N denotes the Gaussian random variable with PDF $f_N(x) = \phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ and $Q_N(x) = \Phi(x) = \int_x^\infty \phi(t)\, dt$ and s satisfies (9). The value $\hat{z}$ is chosen so that the minimum of the left side is equal to the minimum of the right side, which occurs when $t = \hat{\lambda}$. Since $K_N(z) = \frac{1}{2}z^2$, eq. (11) becomes

$$\frac{1}{2}z^2 - \hat{z}z = mK_X(t) - tmK'_X(\hat{\lambda}). \tag{12}$$

To find $\hat{z}$, we set $z = \hat{z}$ and $t = \hat{\lambda}$ in (12) to get

$$-\frac{1}{2}\hat{z}^2 = m(K_X(\hat{\lambda}) - \hat{\lambda}K'_X(\hat{\lambda})) \tag{13}$$

or, taking the sign for $\hat{z}$ to be equal to the sign of $\hat{\lambda}$,

$$\hat{z} = \operatorname{sgn}(\hat{\lambda})\sqrt{(2m(\hat{\lambda}K'_X(\hat{\lambda}) - K_X(\hat{\lambda}))} \tag{14}$$

$$= \operatorname{sgn}(\hat{\lambda})\sqrt{(2(\hat{\lambda}s - mK_X(\hat{\lambda}))}. \tag{15}$$

The transformation (12) maps the region $[0, \hat{\lambda}]$ in t-space into the region $[0, \hat{z}]$ in z-space. The local behavior of $mK_X(t) - tmK'_X(\hat{\lambda})$, which vanishes at $t = 0$ and has zero derivative at $t = \hat{\lambda}$ is reproduced by $\frac{1}{2}z^2 - \hat{z}z$ with similar behavior at $z = 0$ and $z = \hat{z}$. Let $u = z - \hat{z}$. Then,

$$\frac{1}{2}u^2 = mK_X(t) - ts - mK_X(\hat{\lambda}) + \hat{\lambda}s. \tag{16}$$

Expanding $mK_X(t) - ts$ about $t = \hat{\lambda}$ we have

$$\frac{1}{2}u^2 = \frac{1}{2}mK''_X(\hat{\lambda})v^2 + \frac{1}{6}mK'''_X(\hat{\lambda})v^3 + \ldots \tag{17}$$

$$= \frac{1}{2}mK''_X(\hat{\lambda})v^2\left(1 + \alpha_3 v + \alpha_4 v^2 + \ldots\right) \tag{18}$$

where $v = t - \hat{\lambda}$ and $\alpha_n = \frac{2K_X^{(n)}(\hat{\lambda})}{n!K_X''(\hat{\lambda})}$. Hence,

$$u = \sqrt{mK_X''(\hat{\lambda})}v\left(1 + \alpha_3 v + \alpha_4 v^2 + \ldots\right)^{1/2}. \tag{19}$$

It follows that du/dv and dv/du are nonzero for all $v \in [0, \hat{\lambda}]$ and $u \in [0, \hat{z}]$, respectively. Since K_X is analytic in the region of interest, $u(v)$ and $v(u)$ are analytic over the same intervals. Obviously, the same conclusion follows for z as a function of t and t as a function of z. By the inverse function theorem, the transformation $t \leftrightarrow z$ can be extended to a bijection between complex neighborhoods of $[0, \hat{\lambda}]$ and $[0, \hat{z}]$.

The integral (8) now transforms (using Cauchy's theorem) into

$$Q_S(s) = \frac{1}{2\pi i}\int_{d-i\infty}^{d+i\infty} e^{K_N(z)-z\hat{z}}\left(\frac{1}{t}\frac{dt}{dz}\right)dz, \tag{20}$$

where $d > 0$. For small t, we can write

$$z \approx z|_{t=0} + t\frac{dz}{dt}\Big|_{t=0} = t\frac{dz}{dt}\Big|_{t=0}. \tag{21}$$

When $\hat{\lambda} \neq 0$ and hence $\hat{z} \neq 0$, differentiating (12) we obtain

$$\frac{dz}{dt} = \frac{mK_X'(t) - mK_X'(\hat{\lambda})}{z - \hat{z}}, \tag{22}$$

while when $\hat{\lambda} = 0$, (19) implies $dz/dt = du/dv \approx \sqrt{mK_X''(0)}$ when t is small. Thus,

$$\frac{dz}{dt}\Big|_{t=0} = \begin{cases} \frac{1}{\hat{z}}(s - m\,\mathbb{E}(X)) & \text{if } \hat{\lambda} \neq 0, \\ \sqrt{mK_X''(0)} & \text{if } \hat{\lambda} = 0 \end{cases} \tag{23}$$

and therefore, for small t, $z \approx Ct$ where C is a constant. Let

$$U(z) = \left(\frac{1}{t}\frac{dt}{dz} - \frac{1}{z}\right). \tag{24}$$

By expanding $mK_X(t) - ts$ about $t = 0$, it can be shown that, $\lim_{z\to 0} U(z) < \infty$ and, since dt/dz is analytic, $U(z)$ is analytic in the neighborhood of $z = 0$ that includes $\hat{z}$. Therefore, we can rewrite the integral (20) as

$$Q_S(s) = \frac{1}{2\pi i}\int_{d-i\infty}^{d+i\infty} e^{K_N(z)-z\hat{z}}\,\frac{dz}{z} \tag{25}$$

$$+ \frac{1}{2\pi i}\int_{d-i\infty}^{d+i\infty} e^{K_N(z)-z\hat{z}}\,U(z)\,dz. \tag{26}$$

The singularity has now been isolated into (25), which, by comparing with (8), we recognize to equal $\Phi(\hat{z})$. On the other hand, $U(z)$ can be expanded as a Taylor's series around the saddlepoint $z = \hat{z}$ and integrated to obtain an asymptotic series for (26). For the first-order approximation, that is, the leading behavior, we only take the constant term at $\hat{z}$. Let

$$\hat{y} = t\frac{dz}{dt}\Big|_{t=\hat{\lambda}} = \hat{\lambda}\frac{du}{dv}\Big|_{v=0} = \hat{\lambda}\sqrt{mK_X''(\hat{\lambda})}. \tag{27}$$

Then, $U(\hat{z}) = 1/\hat{y} - 1/\hat{z}$ and the integral (26) becomes

$$U(\hat{z})\frac{1}{2\pi i}\int_{\hat{z}-i\infty}^{\hat{z}+i\infty} e^{K_N(z)-z\hat{z}}\,dz = \left(\frac{1}{\hat{y}} - \frac{1}{\hat{z}}\right)\phi(\hat{z}). \tag{28}$$

Thus, we have obtained the Lugananni-Rice formula:

$$\mathrm{Prob}(S \geq s) = \Phi(\hat{z}) + \left(\frac{1}{\hat{y}} - \frac{1}{\hat{z}}\right)\phi(\hat{z}), \tag{29}$$

with $\hat{z}(s)$ given by (9) and (15).

2 SaddleSum implementation

As mentioned in the main text, our SaddleSum algorithm approximates term P-values by first solving eq. (9) for $\hat{\lambda}$ using Newton's method and then using the Lugananni-Rice formula (29). The key step is estimation of $\hat{\lambda}$. Since the moment-generating function ρ of the underlying space W is not known, we estimate it (and its derivatives) using $\mathbf{w}$. Given sufficiently many weights ($n >> 1$), the results can be quite accurate (see below). One limitation of this approach is that our approximation can only accept scores not greater than m times maximal weight ($\hat{\lambda}$ becomes infinite at this bound). Thus, the approximation can be inaccurate for very large scores, causing a larger than usual relative error in P-values (Fig. S6). However, occurence of such extreme scores is rarely seen in practice.

Theoretically, Lugananni-Rice formula is valid over the whole range of the distribution, for small and large scores and both near the mean and in the tails [5]. However, the form (29) becomes numerically unstable close to the mean of the distribution (i.e. when $\hat{\lambda}$ is close to 0). Alternative asymptotic approximations exist that are numerically stable near the mean [3]. For SaddleSum, we were mainly interested in the tail probabilities and we therefore decided not to attempt to approximate the P-values of the scores smaller than one standard deviation from the mean (SaddleSum returns P-value of 1 for all such scores). Terms with such scores are never significant in the context of enrichment analysis.

When processing a terms database, we retain previously computed values of $\hat{\lambda}$ with associated scores and parameters for Lugananni-Rice formula in a sorted array. Since $\hat{\lambda}$ and the P-value are monotonic with respect to the score, using binary search we can certify for many terms that their P-value is larger than a given cutoff and hence eliminate them without running Newton's method. Furthermore, binary search provides a bracket for $\hat{\lambda}$ and hence Newton's method usually converges in very few iteration. We use the bracketed version of Newton's method recommended in the Numerical Recipes book [6] (Section 9.4). This combines the classical Newton's method with bisection and has guaranteed global convergence.

We show evaluations of SaddleSum performance against some theoretically well-characterized distributions in Fig. S5 and S6. It can be seen that the relative error between the SaddleSum approximation and the theoretical P-value is generally very small except for extremely large scores, when P-values are very small. In the context of the enrichment analysis, this discrepancy is not important because such terms will be evaluated as highly significant even if the P-value is off by few orders of magnitude. To further illustrate the quality of our approximation, we have computed the Kullback-Leibler (KL) divergences (relative entropies) between the tail distribution implied by SaddleSum and the theoretical distribution. Prior to computation of KL divergence, both distributions were normalized over the region where SaddleSum is valid (i.e. the tail with scores larger than one standard deviation over the mean). All KL divergence values are extremely small and are comparable between distributions.

Fig. S7 shows relative errors of SaddleSum compared to the empirical distributions using the same weights and term sizes as for Fig. S1 and S2. In this case however, in agreement with the null model of SaddleSum, we sampled weights with replacement. Our results indicate that, except for small m with weights coming from network flow simulations, the relative error of the SaddleSum is similar to that obtained in comparison with well-characterized distributions.

References

[1] N. Bleistein. Uniform asymptotic expansions of integrals with stationary points and algebraic singularity. *Communications in Pure and Applied Mathematics*, 19:353–370, 1966.

[2] H. E. Daniels. Saddlepoint approximations in statistics. *Ann. Math. Statist.*, 25:631–650, 1954.

[3] H. E. Daniels. Tail probability approximations. *Internat. Statist. Rev.*, 55(1):37–48, 1987.

[4] J. L. Jensen. *Saddlepoint approximations*. Clarendon Press, Oxford, 1995.

[5] R. Lugannani and S. Rice. Saddle point approximation for the distribution of the sum of independent random variables. *Adv. in Appl. Probab.*, 12(2):475–490, 1980.

[6] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 3 edition, September 2007.

[7] A. T. A. Wood, J. G. Booth, and R. W. Butler. Saddlepoint approximations to the cdf of some statistics with nonnormal limit distributions. *Journal of the American Statistical Association*, 88(422):680–686, 1993.

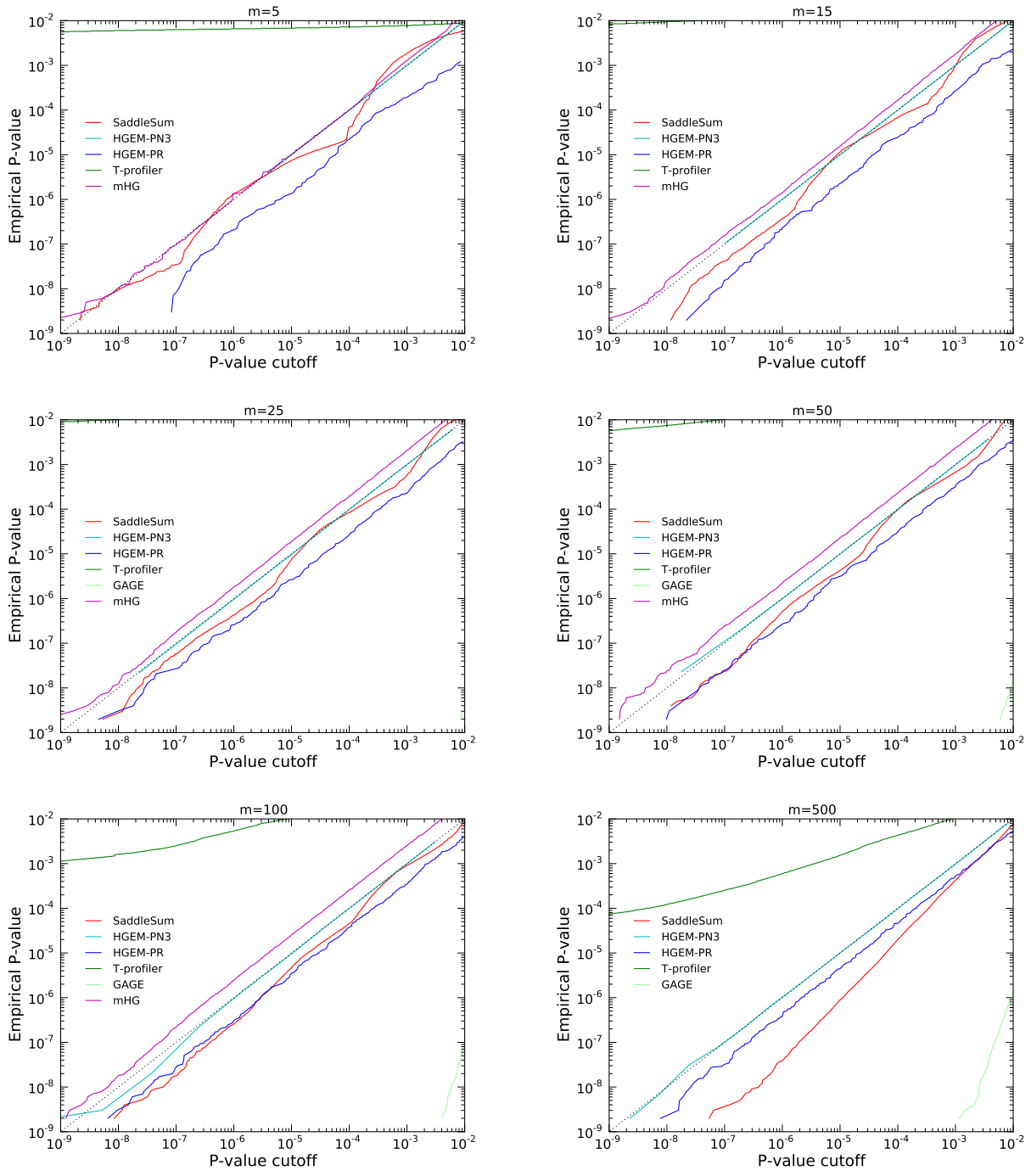

Figure S1: Accuracy of reported P-values from simulations using weights from 100 results of protein network information flow simulations. Each graph shows empirical P-values associated with reported P-value cutoffs for investigated enrichment methods, obtained from queries of decoy term datasets with fixed size terms. The curves for GAGE are omitted from the plots for term sizes 5, 15 and 25 because all reported P-values were greater than 10^{-2}. The graph for $m = 500$ misses the results for mHG because we could not finish the simulation runs within any reasonable amount of time.

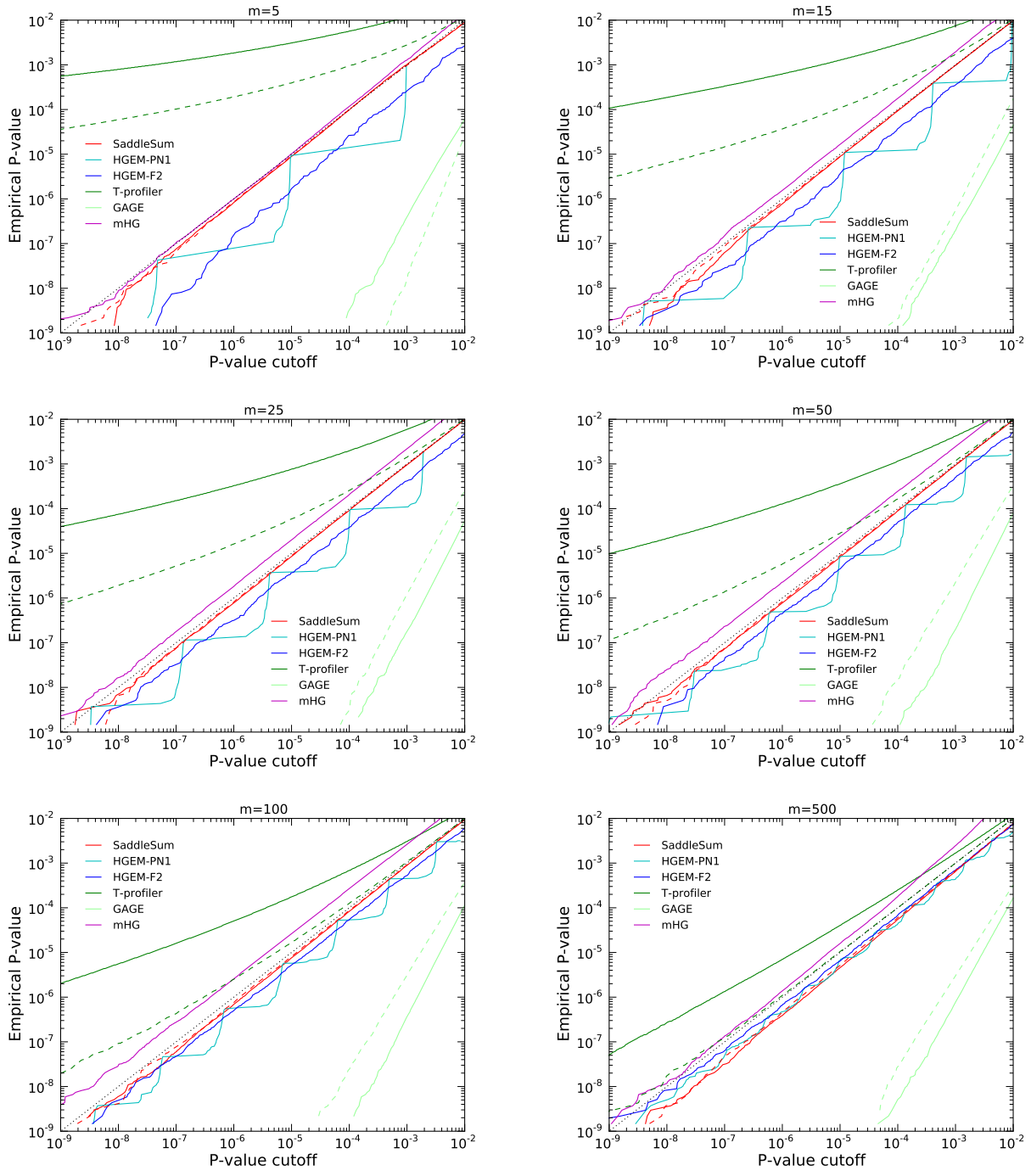

Figure S2: Accuracy of reported P-values from simulations using weights from 136 microarrays. Each graph shows empirical P-values associated with reported P-value cutoffs for investigated enrichment methods, obtained from queries of decoy term datasets with fixed size terms. For SaddleSum, T-profiler and GAGE, full lines indicate the results where negative weights were set to 0, while dashed lines show the results using all weights.

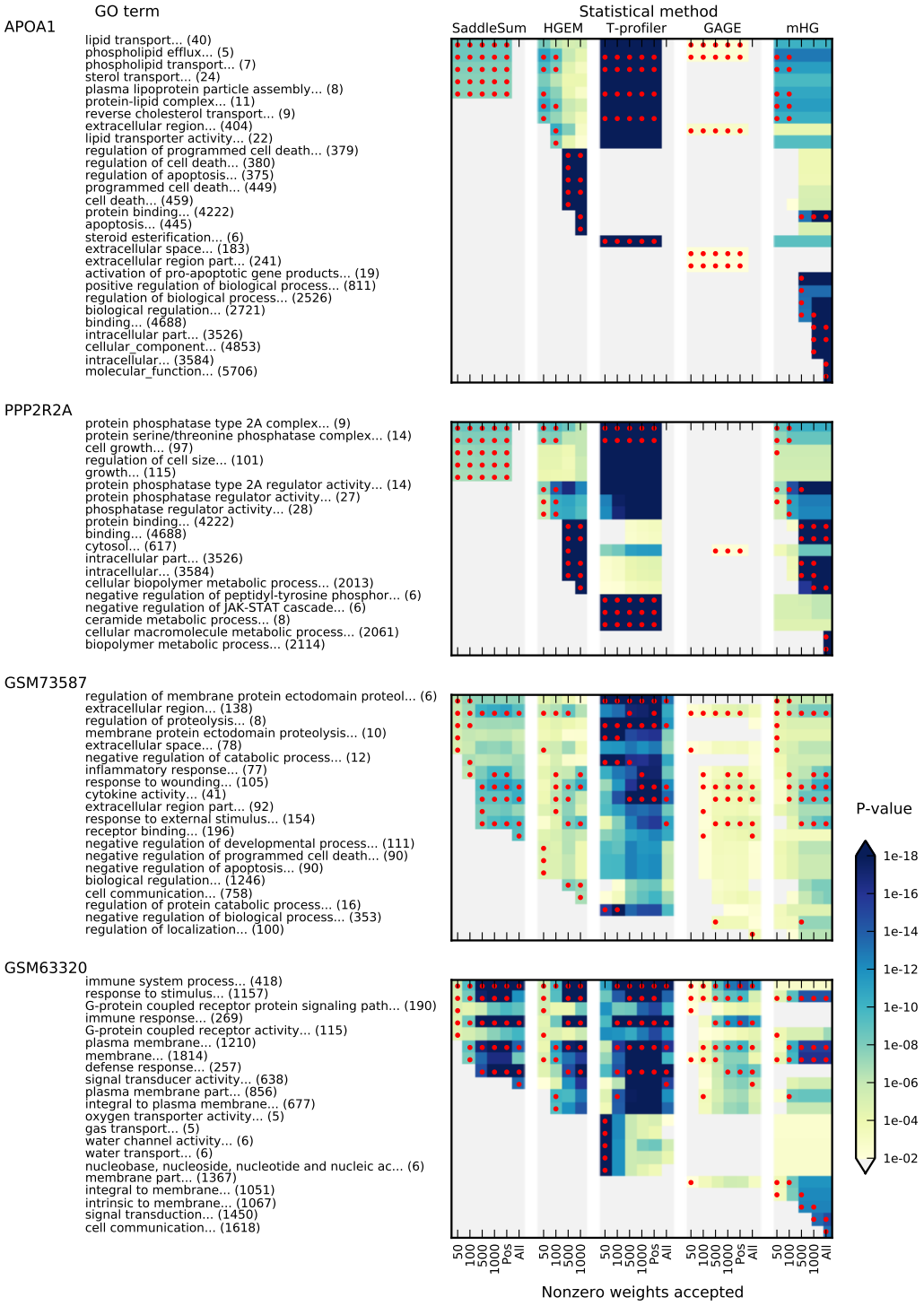

Figure S3: Additional examples of sets of top-five GO terms retrieved by evaluated methods (refer to Fig. 2B for full explanation.) The upper two panels show the enrichment results using the weights from outputs of *ITM Probe* emitting mode with human proteins APOA1 (apolipoprotein A-I, a major protein component of high density lipoprotein in plasma) and PPP2R2A (phosphatase 2 regulatory subunit B) as sources. The lower two panels show the results using weights from microarrays investigating mast cell activation (GSM73587) and malaria response (GSM63320).

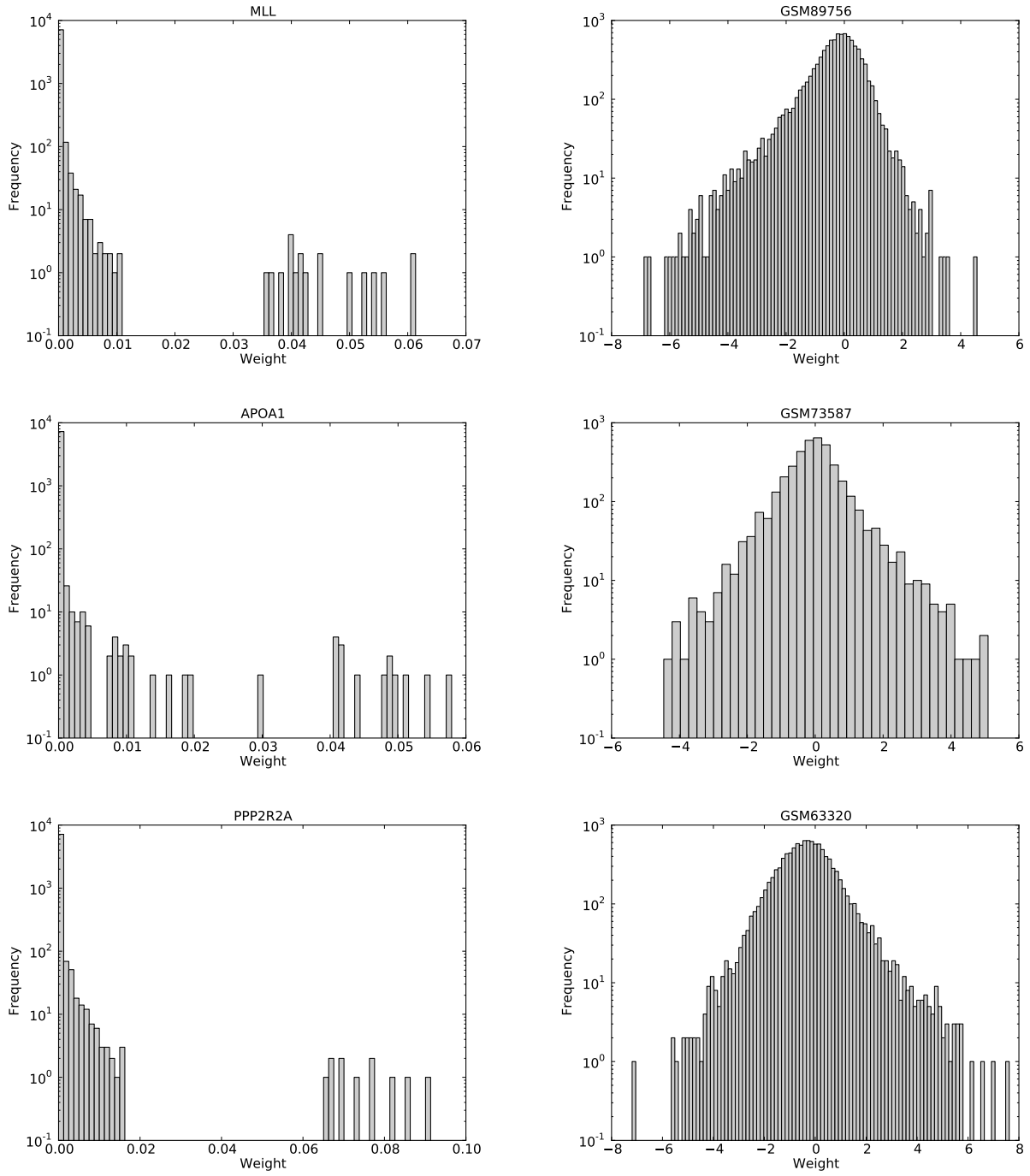

Figure S4: Distributions of weights for examples from Fig. 2 and Fig. S3. Network examples are shown on the left, microarray on the right.

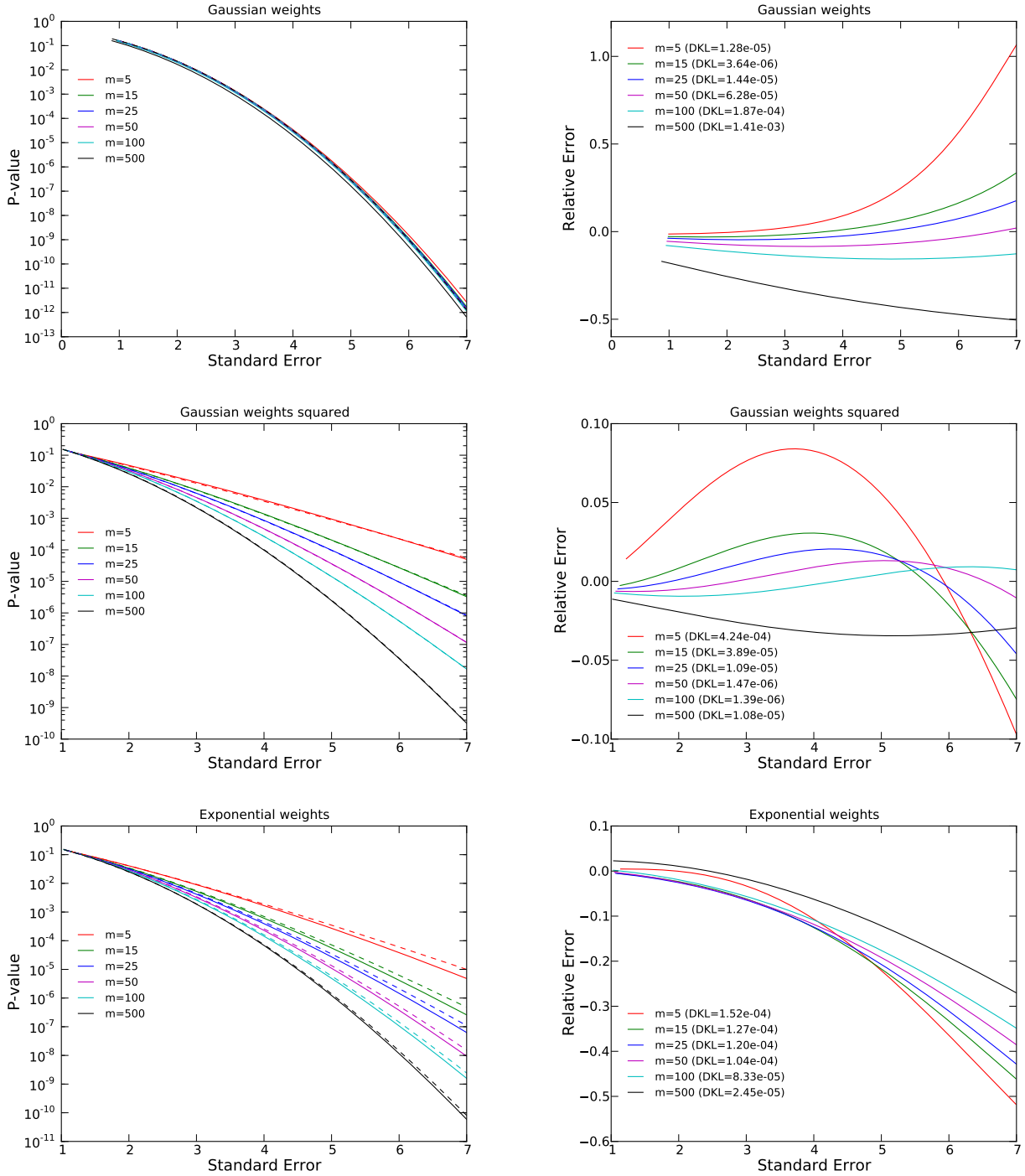

Figure S5: P-values (left) and relative errors (right) for SaddleSum approximations of sums of i.i.d. continuous random variables that are characterized theoretically. In each case 10000 weights were randomly sampled from a distribution and used as input to SaddleSum. The P-values from SaddleSum were compared with P-values from theoretical distributions of the sum of m numbers. Kullback-Leibler divergences (DKL) between the approximated tails of distributions are shown in parentheses for each m. Top: Gaussian (standard normal) weights – sum follows normal distribution. Middle: squared Gaussian weights – sum follows Chi-squared distribution. Bottom: weights from exponential distribution – sum follows Erlang distribution.

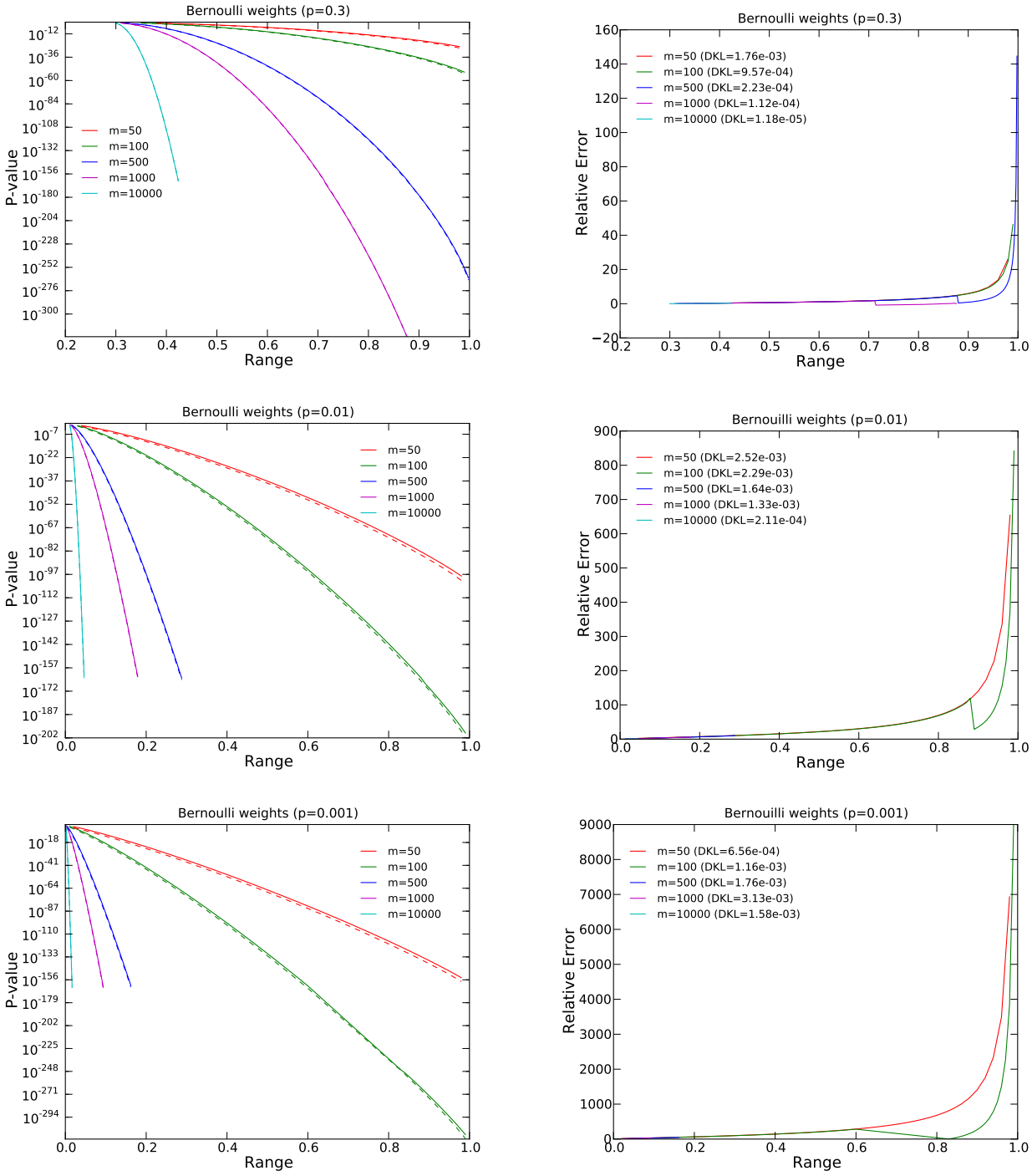

Figure S6: P-values (left) and relative errors (right) for SaddleSum approximations of sums of i.i.d. Bernoulli ($\{0, 1\}$) random variables with different parameter p. Such sums follow binomial distribution. In each case 10000 weights were randomly sampled from a distribution and used as input to SaddleSum. The P-values from SaddleSum were compared with P-values from the binomial distribution. Kullback-Leibler divergences between the approximated tails of distributions are shown in parenthesis for each m. Top: $p = 0.3$. Middle: $p = 0.01$. Bottom: $p = 0.001$. The dramatic increase in relative error is caused by $\hat{\lambda}$ instability at extreme scores, see Section 2 of this material.

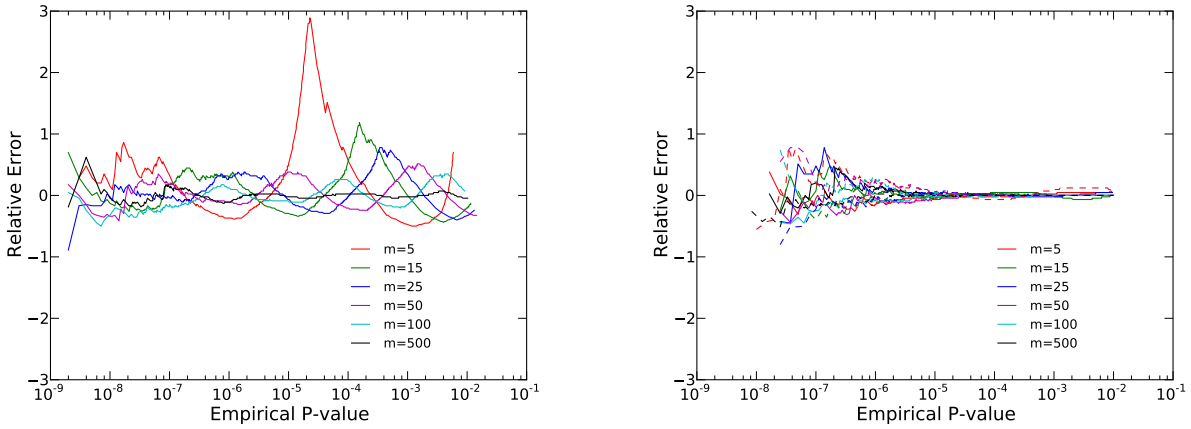

Figure S7: Relative error of P-values reported by SaddleSum from simulations using weights from 100 results of protein network information flow simulations (left) and from 136 microarrays (right). These are the same query sets as evaluated in Fig. S1 and Fig. S2 but in this case the weights are drawn with replacement. Each sample size m is shown in different color. Full lines indicate the results where negative weights were set to 0, while dashed lines show the results using all weights.