BIOINFORMATICS

Vol. 20 no. 6 2004, pages 945–952
DOI: 10.1093/bioinformatics/bth011

Finding coexpressed genes in counts-based data: an improved measure with validation experiments

Morgan N. Price and Eleanor Rieffel*

FX Palo Alto Laboratory, 3400 Hillview Avenue Building 4, Palo Alto, CA 94304, USA

Received on July 29, 2003; revised on October 15, 2003; accepted on October 16, 2003
Advance Access publication January 29, 2004

ABSTRACT

Motivation: Expressed sequence tag (EST) data reflects variation in gene expression, but previous methods for finding coexpressed genes in EST data are subject to bias and vastly overstate the statistical significance of putatively coexpressed genes.

Results: We introduce a new method (LNP) that reports reasonable *p*-values and also detects more biological relationships in human dbEST than do previous methods. In simulations with human dbEST library sizes, previous methods report *p*-values as low as 10^{-30} on 1/1000 uncorrelated pairs, while LNP reports significance correctly. We validate the analysis on real human genes by comparing coexpressed pairs to gene ontology annotations and find that LNP is more sensitive than the three previous methods. We also find a small but statistically significant level of coexpression between interacting proteins relative to randomized controls. The LNP method is based on a log-normal prior on the distribution of expression levels.

Availability: Source code in Java or R is available at http://ests.sourceforge.net/

Contact: mprice@cs.cmu.edu

Supplementary information: http://ests/sourceforge.net/lnp_supplement.pdf

INTRODUCTION

Analyzing expression patterns is a popular approach to elucidating the function of genes. Here, we examine methods for finding pairs of coexpressed genes in expressed sequence tag (EST) databases, and the extent to which these methods uncover biologically significant relationships: coexpression analysis can only succeed when functionally related genes have similar expression patterns, which is not always the case (Niehrs and Pollet, 1999). We use EST data because a large public database exists and because EST data are consistent with measurements from northern hybridizations (Okubo *et al*., 1992), SAGE (Bortoluzzi *et al*., 2000), reverse transcription–polymerase chain reaction (RT–PCR; Boutanaev *et al*., 2002) and whole-mount *in situ* hybridization (Gitton *et al*., 2002). Although microarrays are cheaper than ESTs and more data have been collected with them, cDNA and oligo-based platforms give inconsistent and often unreliable results and, in organisms with large genomes containing numerous recent duplications, face problems with cross-hybridization (Kuo *et al*., 2002; Kothapalli *et al*., 2002).

The analysis of expression patterns in ESTs begins by assigning the ESTs to genes and representing the data as a matrix of counts (or observed frequencies), with genes along one dimension and libraries or tissues along the other. Walker *et al*. (1999) analyze a proprietary dataset by reducing the counts to binary variables (present or absent), computing the 2×2 occurrence table of the number of libraries with each combination of presence or absence, and using the Fisher exact test to report a *p*-value. Thompson *et al*. (2002) apply Fisher to human dbEST. In contrast, Ewing *et al*. (1999) and Ewing and Claverie (2000) treat the rows of counts for the two genes as vectors of random variables, compute the Pearson (linear) correlation coefficient (r), and use a threshold to select coexpressed pairs. Nelander *et al*. (2003) use Pearson with a chosen bait to find tissue-specific genes in mouse dbEST, which they verify experimentally. The Euclidean distance between gene-wise vectors of Pearson correlation coefficients (EDVP) has also been used to cluster genes (Ewing *et al*., 1999; Gitton *et al*., 2002). Gitton *et al*. use the mean within-cluster r to test clusters of genes from mouse dbEST for significance.

These tests for significance rely on the counts (or binary presence/absence) being independent for uncorrelated genes, which is not the case when the libraries vary in size by orders of magnitude, as in human or mouse dbEST: larger libraries will tend to show larger counts. In practice, researchers have thrown out small libraries—Thompson *et al*. (2002) use the Fisher exact test on libraries with at least 100 different UniGenes, and Ewing and Claverie (2000) and Gitton *et al*. (2002) use the Pearson correlation coefficient on libraries with at least 1000 ESTs—yet wide variation remains. For example, human non-normalized libraries containing at least 100 different UniGenes have an average size of 2018 ESTs but a median size of only 384 ESTs.

*To whom correspondence should be addressed.

We propose a Bayesian method with a log-normal prior ('LNP') to find coexpressed genes in counts-based data with high specificity, despite variations in library sizes. We need a prior distribution for each gene's expression levels because the counts tend to be small and hence are unreliable estimates of the observed frequency. In fact, most of the counts are zero, and while a count of zero for a highly expressed gene in a large library suggests lower than average expression, a count of zero for a typical gene in a smaller library is unsurprising and contains little information. Thus, LNP addresses the discrete noise from random sampling in counts of ESTs (or, potentially, other counts-based data such as SAGE), in contrast to the analog noise in microarray data.

We use simulations to show that with library sizes from human dbEST, LNP reports significance correctly, while both the Fisher exact test and the Pearson correlation-based test report spurious coexpression at significant rates. LNP is also more sensitive than Fisher or Pearson, showing a greater separation between correlated and random pairs. To test the extent to which coexpression analysis of EST data reflects biologically significant relationships, and to compare the performance of the measures on real data, we compare the results to gene ontology (GO) annotations (from EBI; Camon *et al*., 2003) and to protein-protein interactions (the MINT database; Zanzoni *et al*., 2002). The LNP method outperformed Fisher, Pearson and EDVP in distinguishing biologically significant pairs.

METHODS

Data sources

We obtained human EST-UniGene mappings from build #160 (http://www.ncbi.nlm.nih.gov/UniGene). We removed libraries which do not reflect expression levels, including subtracted (or differential display), rearrayed (or pooled tissues) and 'random activation of gene expression' libraries, by examining dbEST descriptions and CGAP 'protocols' (http://cgap.nci.nih.gov/Info/CGAPDownload). As the limitations of the different measures require throwing out different libraries, a fair comparison uses a different subset of libraries for each method. We follow past use by including both linear and normalized libraries with $\geq$100 different UniGenes for Fisher (Thompson *et al*., 2002), and larger ($\geq$1000 UniGenes) linear libraries for Pearson and EDVP [similar to Ewing and Claverie (2000) and Gitton *et al*. (2002)]. For LNP, we used linear libraries with at least 100 UniGenes. Finally, we ignored UniGenes appearing in less than five of the remaining libraries. The final dataset had 3.1 million ESTs for 23 752 genes in 1262 libraries (for Fisher), or 2.1 million ESTs for 18 941 genes in 1089 non-normalized libraries (for LNP), or 1.7 million ESTs for 17 236 genes in 228 large non-normalized libraries (for Pearson and EDVP). The simulations use the same library sizes, except that the sensitivity comparison (Fig. 2) uses only linear library sizes for Fisher (matching LNP) and the 228 largest of those for Pearson.

We downloaded mappings between SWISSPROT identifiers (used by EBI GO and MINT) and RefSeq identifiers (included in UniGene clusters) from Ensembl (the homo_sapiens_core_9_30 database; see http://www.ensembl.org/) on December 6, 2002. In our biological validation experiments, we ignored all SWISSPROT identifiers which did not map to UniGenes uniquely, and all UniGenes which mapped to multiple LocusLink ids or vice versa.

We downloaded GO annotations for human genes from EBI (http://www.ebi.ac.uk/GOA/) on April 4, 2003. We used only the highest reliability ('traceable author statement') annotations. Following Gibbons and Roth (2001), we imputed annotations up the GO hierarchy, and ignored categories that were associated with fewer than 10 genes or were too similar to other categories (mutual information over 80% of maximum). We also removed large non-specific categories (over 200 genes), leaving 482 categories and 17 167 gene-category assignments for identifying coexpressed categories (Table 2). When comparing functionally related and unrelated pairs of genes (Table 3), we retained the small and partially redundant categories (31 453 gene-category assignments). We received MINT version 1.2 from G. Cesarini on October 25, 2002.

Statistical tests

The Fisher exact test used the one-sided sum of hypergeometric probabilities of more extreme contingency tables. We computed the p-values for the Pearson coexpression measure by the standard t-test (one-tailed) on $r\sqrt{(n-2)/(1-r^2)}$ with $n-2$ degrees of freedom. All other tests are two-tailed if applicable, and were performed with the R statistics package (http://www.r-project.org/).

THE LOG-NORMAL PRIOR ALGORITHM

We model the variation in each gene's frequency with a log-normal distribution, finding the maximum likelihood mean and SD (in log-space) based on the counts. Given the prior distributions, the counts, and assuming that each count was obtained by binomial sampling with the (unknown) true frequency, we can compute the likelihood of a correlation coefficient for the log-levels of two genes by integrating over all possible frequencies for both genes in each library. We find the maximum likelihood value for the correlation coefficient between the two genes' log-levels and use the ratio of the likelihoods for the predicted correlation versus no correlation to test statistical significance. Finally, because the computation is too expensive to do for every pair of human genes, we introduce an 'apparent correlation' metric to cheaply filter out un-correlated pairs.

What prior should we use? The expression level of any given gene across various conditions is determined by the multiplicative effects of many transcription factors, so we would expect

a log-normal distribution. In microarray data, the distribution of log-levels for all genes across all tissues (mixing all the per-gene distributions together) is roughly normal but with heavy tails (Hoyle *et al.*, 2002). The heavy tails reflect the exponential tail in the distribution of average levels of genes (Kuznetsov *et al.*, 2002) and the heavy tails in the noise in microarray data (Brody *et al.*, 2002), but may also reflect true biological deviation from our simple model. In any case, LNP is robust to deviations from the prior (see tests with a 'half-on' distribution in Supplementary Table 1), and the EST data for large subunit ribosomal proteins fits our prior reasonably well (Supplementary Figure 1).

We need a distribution for frequencies, which are bounded above by 1, rather than for concentrations. Let $e^{\mu'_A+\sigma_A z_{Ai}}$ be the concentration of message A in condition i, where z_{Ai} is a standard normal variable. The frequency

$$f_{Ai} = \frac{e^{\mu'_A+\sigma_A z_{Ai}}}{e^{\mu'_A+\sigma_A z_{Ai}} + \sum_{G\neq A} e^{\mu'_G+\sigma_G z_{Gi}}} \approx \frac{e^{\mu_A+\sigma_A z_{Ai}}}{1+e^{\mu_A+\sigma_A z_{Ai}}},$$

where the sum is roughly constant because it is a sum of many random variables and μ_A is chosen to give the 1.

For efficiency, we compute the relative likelihood of different values of the correlation coefficient $r_{AB} = r(\vec{z_A}, \vec{z_B})$ using the maximum likelihood values $\hat{\mu_A}, \hat{\sigma_A}, \hat{\mu_B}, \hat{\sigma_B}$ for the parameters of the LNPs, rather than considering all possible values. This maximum likelihood approximation is accurate in practice (data not shown). Let n_{Ai} be the number of copies of mRNA for gene A in library i and L_i be the number of ESTs in library i. Then,

$$p(r_{AB}|\vec{n_A}, \vec{n_B}, \vec{L}) \propto \prod_i \iint p(n_{Ai}|z_{Ai}, \hat{\mu_A}, \hat{\sigma_A}, L_i) \cdot p(n_{Bi}|z_{Bi}, \hat{\mu_B}, \hat{\sigma_B}, L_i) \cdot p(z_{Ai}, z_{Bi}|r_{AB}) \cdot dz_{Ai}dz_{Bi},$$

where $p(n_{Ai}|z_{Ai}, \hat{\mu_A}, \hat{\sigma_A}, L_i)$ is binomial (using f_{Ai} from $z_{Ai}, \hat{\mu_A}, \hat{\sigma_A}$) and we derive

$$p(z_{Ai}, z_{Bi}|r_{AB}) = \exp\left(-\frac{z_{Ai}^2 + z_{Bi}^2 - 2r_{AB}z_{Ai}z_{Bi}}{2(1-r_{AB}^2)}\right) \Big/ \left(2\pi\sqrt{1-r_{AB}^2}\right)$$

by rewriting z_{Bi} as $r_{AB}z_{Ai} + \sqrt{1-r_{AB}^2}\zeta$.

To find the maximum likelihood values $\hat{\mu_A}$ and $\hat{\sigma_A}$, we compute $\log p(\vec{n_{Ai}}|\mu_A, \sigma_A, \vec{L_i})$ and its first- and second-order partial derivatives by numerical integration over z_{Ai}, and use a two-dimensional variant of Newton–Raphson optimization to find the maximum. To make the variables more independent, we use the log mean frequency $\mu(+\sigma^2)/2$ instead of μ in the optimization. To keep the algorithm from blowing up when there is insufficient information about $\hat{\sigma_A}$, we do not allow values below 0.01 or above 6. To test the computation of maximum likelihood μ and σ, we simulated 1000 genes with average frequencies ranging from $4.56\cdot10^{-6}$ to $1.14\cdot10^{-4}$ and σ ranging from 1 to 2.5. Low-count and high-sigma genes are harder to make predictions for: the average error in σ is 0.18 if there are 20 or more ESTs, but 0.51 otherwise. Similarly, the average error in log mean f is 0.14 if $\sigma < 2$ but 0.25 otherwise.

We compute the maximum likelihood value $\hat{r_{AB}}$ by Newton–Raphson optimization on $\log p(r_{AB})$, starting at $r_{AB} = 0$. To speed up the double integrals in computing $\log p(r_{AB})$ and its derivatives, we pre-compute $p(z_{Ai}|\hat{\mu_A}, \hat{\sigma_A}, n_{Ai}, L_i, r_{AB} = 0)$, the same for B, and then compute $p(z_{Ai}, z_{Bi}|r_{AB})/p(z_{Ai}, z_{Bi}|r_{AB} = 0)$ and its first two derivatives with respect to r_{AB}. We also reduce the range of the numerical integration by ignoring z_{Ai} such that the integral of $p(z_{Ai}|\hat{\mu_A}, \hat{\sigma_A}, n_{Ai}, L_i, r_{AB} = 0)$ in the ignored tail is under 10^{-6} (and similarly for z_{Bi}). The time complexity of this algorithm is $O(N_{\text{libraries}}N_Z^2N_{\text{iterations}})$, where $N_{\text{libraries}} = 1089$, $N_Z \approx 50$, $N_{\text{iterations}} \approx 4$, so we can only test a few pairs per second.

As we have one free parameter (r) and many dimensions of data, the ratio of likelihoods $\Lambda = p(\text{Data}|r = \hat{r})/p(\text{Data}|r = 0)$ gives us a statistical test for significance: $-2 * \log \Lambda$ is roughly χ^2 with one degree of freedom (Rao, 1965). Note that LNP gives separate estimates of the degree of coexpression and of statistical significance.

For efficiency when performing the all-pairs computation, we prefilter pairs using the 'apparent correlation' $r(\hat{z_{Ai}}, \hat{z_{Bi}})$, the correlation between the maximum likelihood values of z. We also compare each gene against itself, and ignore genes whose self-p-values are too high. We chose thresholds by examining the 32 379 MINT control pairs (described below): all pairs with p-values up to $10\times$ higher than our cutoff of $1.27 \cdot 10^{-9}$ had self-$p < 2.46 \cdot 10^{-24}$ and apparent correlation >0.2. This choice represents a tradeoff between the risk of false negatives and increased running time for the all-pairs computation (about a week on a single processor 2.4 GHz Pentium 4, although parallelizing the computation would be straightforward). The LNP all-pairs computation tested $4.15 \cdot 10^7$ pairs, of which $5.4 \cdot 10^5$ passed the filter (1.3%), and of those, $3.1 \cdot 10^4$ were significant (5.8%).

RESULTS

Simulations

To see whether the Fisher exact test and the Pearson correlation produce biased significance scores in practice, we tested both measures on simulated uncorrelated genes using library sizes from human dbEST. (We did not test EDVP because it does not report significance.) We generated random frequencies with the modified log-normal model ($\sigma = 1$ or 2) and a 'half-on'

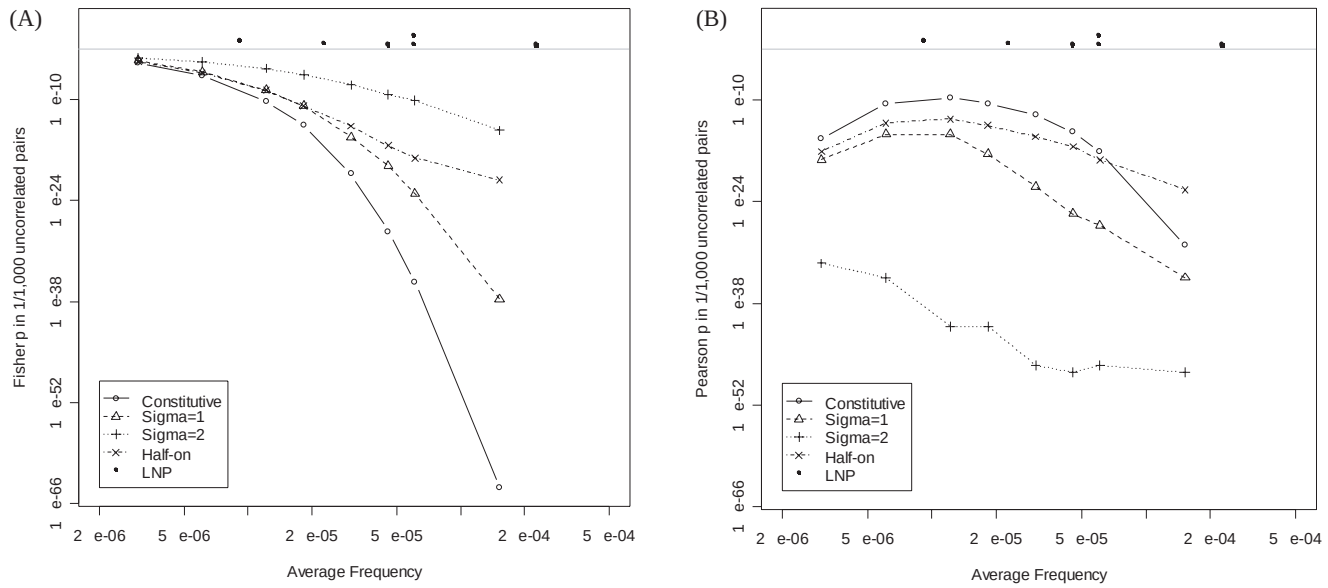

Fig. 1. Reported p-values seen at a rate of 1 in 10^3 pairs in simulations of uncorrelated genes for (**A**) the Fisher exact test and (**B**) the Pearson correlation coefficient, as a function of the average level of expression and for several models of variation in expression. The horizontal line at 10^{-3} shows the expectation for a perfect test, and results using LNP are shown for comparison (Supplementary Table 1). The p-values for the Pearson correlation coefficient correspond to $r = 0.40$–0.78, with $n_{\text{Libraries}} = 228$. For each point shown, we ran 10^5 pairs and reported the 100th lowest p-value. In these simulations, both genes had the same expression levels and library sizes were taken from human dbEST; similar results are seen when comparing genes at different levels of expression or for library sizes from mouse dbEST or from (Gitton *et al.*, 2002; data not shown).

model ($f = 0$ in half of the libraries and a fixed value in the other libraries). As a control, we also used a constitutive model which cannot show coexpression ($\sigma = 0$ or f constant).

For a correct statistical test, when the null hypothesis is true, low values of p should occur no more frequently than the value of p itself. As Figure 1 shows, extremely low values of p occur at rates of 1 per 1000 uncorrelated pairs of genes for both tests under most simulation conditions. Furthermore, Fisher is biased towards constitutive highly expressed genes, while Pearson prefers highly variable genes.

In contrast, the distribution of LNP's p-values on uncorrelated genes showed a good fit to theoretical predictions for a wide range of models and expression levels (Supplementary Table 1). Low p-values occurred at most 60% more frequently than expected, and we did not see the extreme reported by the other measures, even when testing 10^6 pairs. LNP was aggressive on the half-on model, which contradicts the assumed prior, and at $\sigma = 3$ (most genes show modest σ; Supplementary Figure 3). LNP was conservative at $\sigma = 0$, the negative control, and at low counts (20–50 ESTs), presumably because of insufficient data.

To test the sensitivity of all three measures, we compared the distribution of p-values on 1000 simulated pairs with average frequency of $4.6 \cdot 10^{-5}$ (100 ESTs), $\sigma = 1.5$ and correlation $= 0.8$ (in log-space) versus 24 750 un-correlated pairs. LNP was much more sensitive than Fisher or Pearson, but even LNP could not distinguish most pairs of correlated

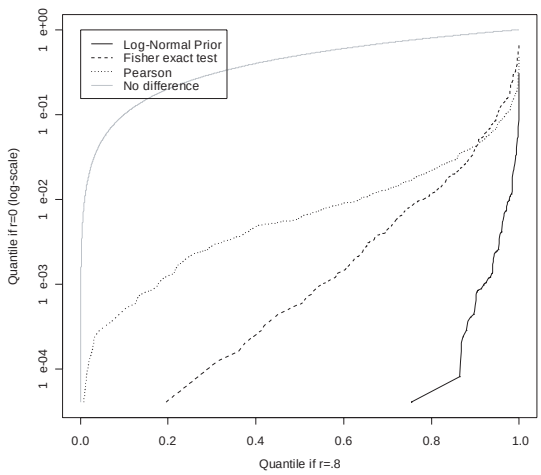

Fig. 2. Quantile–quantile plot comparing the sensitivity of the Fisher exact test, Pearson correlation and LNP measures. We computed p-values with all three measures for 1000 correlated test pairs ($r = 0.8$ in log-space) and 24 750 uncorrelated control pairs ($r = 0$) of simulated genes with average frequency $4.6 \cdot 10^{-5}$ and $\sigma = 1.5$. For LNP, negative correlations were ignored (we set $p = 1$ if $\hat{r} < 0$). The x-axis shows the fraction of test pairs smaller than a given p-value, and the y-axis shows the fraction of control pairs below that same p-value (log-scale). A measure that does not distinguish between the test and control pairs would follow the gray line. The LNP and Fisher lines end once the test values are below all the control values.

Table 1. (A) Shared pairs between lists of coexpressed human UniGenes produced by different measures. (B) Number of different genes found in the coexpressed pairs and their properties, and for comparison, properties over all of the genes used for the LNP all-pairs computation

Measure	A				B			
	LNP	Fisher	Pearson	EDVP	No. of genes	Median (f)	SD log(f)	Median ($\hat{\sigma}$)
LNP	31 154	7666	1376	1660	2719	$2.2 \cdot 10^{-4}$	1.16	1.47
Fisher	—	31 154	274	185	1247	$3.9 \cdot 10^{-4}$	0.85	1.19
Pearson	—	—	31 154	18 543	3203	$2.5 \cdot 10^{-5}$	1.37	2.17
EDVP	—	—	—	31 154	4815	$1.6 \cdot 10^{-5}$	1.33	2.35
Total					9082	$4.7 \cdot 10^{-5}$	0.87	1.59

f, observed frequency.
$\hat{\sigma}$, estimated SD of a gene's log-levels.

genes with high confidence (Fig. 2). A genome-wide all-pairs comparison would require $p < 10^{-9}$ for significance, attained by only 13% of the pairs (the median p-value is $1.7 \cdot 10^{-6}$). At $r = 0.6$, only 0.9% show significance (median $p = 1.0 \cdot 10^{-3}$). Our pruned human dbEST dataset contains only 4804 genes with ≥100 ESTs, so more data will be required to identify most coexpressed pairs. If the libraries double in size, or if twice as many libraries are sequenced, the median p-value for pairs with $r = 0.8$ improves dramatically to $7.9 \cdot 10^{-11}$ or $2.1 \cdot 10^{-11}$, respectively (63 or 70% with $p < 10^{-9}$).

Biological validation

To validate the coexpression measures on human dbEST, we tested all pairs of genes for coexpression and compared coexpressed pairs with GO annotations from EBI (Camon *et al.*, 2003). For LNP, from 9082 genes with low self-p and pairs with high apparent correlation, we found 31 154 pairs with $p < 1.27 \cdot 10^{-9}$ ($p < 0.05$ and Bonferonni correction). As Fisher and Pearson do not give reasonable p-values, and the EDVP does not report significance, we used the same number of top hits from running those measures on our full set of 17 236 genes, giving cutoffs of $p \leq 1.7 \cdot 10^{-48}$ for Fisher, $p \leq 2.7 \cdot 10^{-46}$ ($r \geq 0.770$) for Pearson, and $\mathrm{rms}(\vec{r_A} - \vec{r_B}) \leq 0.0582$ for EDVP. Consistent with the simulations, Fisher finds pairs within a relatively small set of very highly expressed, relatively constitutive, genes; Pearson finds a wide variety of relatively lowly expressed genes with high σ; EDVP prefers genes with even higher σ and LNP finds genes at middle levels of expression (Table 1). The maximum likelihood correlation $\hat{r}$ for the LNP pairs is generally high (mean and SD 0.69 ± 0.13), suggesting that the correlations are strong enough to be biologically meaningful.

We then looked for GO categories with statistically significant coexpression, based on coexpressed pairs with GO annotations for both genes (7969 pairs for LNP, 10 174 for Fisher, 3093 for Pearson and 3012 for EDVP), after excluding very large, very small and largely redundant categories (see Methods section). A total of 40 categories (34 from LNP, 8 from Fisher, 15 from Pearson and 19 from EDVP) show significant coexpression (Table 2). We suspect that the Fisher hits correlate with only a few categories because of bias towards constitutive highly expressed genes such as ribosomal proteins.

To compare the sensitivity of the measures while avoiding such biases, we tested 50 000 pairs of genes with shared GO annotations and 50 000 control pairs of annotated genes without shared categories. All measures show significant differences between the test set and the control set, and LNP shows a significantly higher separation than the other measures (Table 3).

To see whether coexpression analysis of human genes can hint at close biological relationships as well as broad functional categories, we compared the coexpression of pairs of interacting human proteins, collected from the literature into the MINT database (Zanzoni *et al.*, 2002), to randomized controls (269 test pairs and 32 379 controls). The separation between interacting and control distributions is rather modest (Supplementary Figure 2), but LNP does find a significantly higher rate of positive than negative correlations in interacting pairs relative to controls (e.g. below $p = 0.05$, 41+/9− in the interacting set and 3775+/1869− in the control set, significant at $p = 0.023$, two-sided Fisher exact test). Furthermore, both the distribution of p-values and the distribution of maximum likelihood correlation values (regardless of p-value) for interacting proteins is significantly greater than in the control set ($p = 0.020$ and 0.018, Wilcoxon rank sum test; we replaced p-values with $1 + 1/p$ when $\hat{r} < 0$ to test for differences across the full range of results—note that only the rank of the resulting score matters). Fisher and EVP also showed a significant separation ($p = 0.016$ and 0.040), but Pearson did not ($p = 0.15$).

DISCUSSION

What have we learned about the principle of inferring function from expression patterns? We found that interacting human proteins do show coexpression, but the coexpression is weak

Table 2. GO categories significantly correlated with all-pairs results from at least one measure

GO no.	Name	LNP	Fisher	Pearson	EDVP	Both	One	Neither	*E* (Both)
0003735	Structural constituent of ribosome	>**100**	>**100**	**59.1**	**65.1**	1386	1498	5085	572
0030529	Ribonucleoprotein complex	>**100**	>**100**	**62.9**	**68.3**	1507	1717	4745	702
0009059	Macromolecule biosynthesis	>**100**	>**100**	**25.1**	**32.1**	1230	1980	4759	618
0005829	Cytosol	>**100**	>**100**	**33.1**	**39.4**	1164	1929	4876	569
0007601	Vision	**61.6**	0.0	**37.4**	**38.8**	36	60	7873	0.5
0007600	Sensory perception	**55.5**	0.0	**31.7**	**34.1**	36	80	7853	0.7
0015934	Large ribosomal subunit	**54.9**	**36.5**	**22.1**	**27.8**	326	1500	6143	145
0016283	Eukaryotic 48S initiation complex	**46.3**	**21.6**	2.9	**4.2**	237	1269	6463	95
0005578	Extracellular matrix	**33.0**	0.0	2.7	**8.1**	30	147	7792	1
0007398	Ectoderm development	**27.0**	0.0	**10.5**	**31.7**	17	60	7892	0.3
0006936	Muscle contraction	**11.7**	0.0	**17.6**	**27.2**	11	106	7852	0.5
0008307	Structural constituent of muscle	**25.4**	0.0	**12.8**	**13.1**	18	79	7872	0.4
0001501	Skeletal development	**20.2**	0.5	1.5	**6.5**	22	168	7779	1
0005200	Structural constituent of cytoskeleton	**19.8**	0.0	**4.8**	**12.0**	18	122	7829	0.8
0008233	Peptidase activity	**19.4**	0.7	1.8	2.3	32	300	7637	4
0005581	Collagen	**19.2**	0.0	0.0	1.0	13	65	7891	0.3
0000278	Mitotic cell cycle	**17.3**	4.0	1.5	3.5	78	779	7112	27
0015399	Primary active transporter activity	**15.2**	3.9	3.1	1.6	63	689	7217	21
0005743	Mitochondrial inner membrane	**11.9**	3.6	1.4	1.8	63	756	7150	24
0005386	Carrier activity	**11.6**	3.1	1.7	0.8	66	790	7113	27
0007602	Phototransduction	**11.3**	0.0	3.6	1.8	6	29	7934	0
0045111	Intermediate filament cytoskeleton	**10.7**	0.0	0.0	3.4	10	102	7857	0.5
0015629	Actin cytoskeleton	**10.7**	1.2	2.2	3.9	21	290	7658	3
0007565	Pregnancy	**9.0**	0.0	**7.8**	**9.7**	11	147	7811	0.9
0005211	Plasma glycoprotein	0.0	0.0	**6.7**	**6.4**	0	22	7947	0
0005819	Spindle	1.9	**6.6**	0.0	0.0	3	118	7848	0.5
0015630	Microtubule cytoskeleton	**6.5**	2.4	0.0	0.0	10	177	7782	1
0005747	Respiratory chain complex I (s.E.)	**5.7**	0.0	0.9	0.8	13	264	7692	3
0005875	Microtubule associated complex	**5.1**	0.0	0.0	0.0	4	57	7908	0.1
0030484	Muscle fiber	**4.9**	0.0	3.6	**4.5**	4	61	7904	0.1
0015078	Hydrogen ion transporter activity	**4.6**	1.2	3.0	1.8	27	548	7394	11
0008324	Cation transporter activity	**4.6**	1.5	1.7	1.8	27	551	7391	11
0008238	Exopeptidase activity	**4.5**	0.0	0.4	0.3	3	39	7927	0
0046873	Metal ion transporter activity	**4.5**	0.2	0.8	1.3	14	326	7629	4
0003793	Defense/immunity protein activity	0.0	0.0	**4.3**	3.3	0	17	7952	0
0006941	Striated muscle contraction	1.0	0.0	**4.3**	**4.2**	1	57	7911	0.1
0005125	Cytokine activity	1.4	0.0	2.0	**4.2**	1	35	7933	0
0016651	Oxidoreductase activity on NAD(P)H	**4.2**	0.2	1.4	1.8	16	378	7575	5
0030163	Protein catabolism	**4.0**	0.0	1.5	1.4	7	172	7790	1
0005837	26S proteasome	0.6	**4.0**	2.5	2.5	2	174	7793	1

The score for each category and measure is $-\log_{10}(p)$ for the occurrence table (Both, One, Neither), with significant scores ($p < 0.05/482$ categories) in bold. p-values are from an exact test for the (one-sided) probability of more extreme occurrence tables (b, o, n), keeping the no. of pairs $b+o+n$ and the no. of pair members in the category $2b+o$ fixed, and

$$p(b,o,n) = 2^o \frac{(b+o+n)!}{b!o!n!} \Big/ \binom{2\cdot(b+o+n)}{2b+o}.$$

The occurrence tables on the right are for LNP, and also show the neutral expectation of Both.

compared with observations on yeast microarray data. For example, figure 4 in Deane *et al*. (2002) implies that the median extent of coexpression of interacting pairs is at around the 80th percentile of random pairs, compared with only the 55th percentile in our test on human genes. It is unclear whether this is because of the limited information in dbEST, the types of pairs in MINT [Jansen *et al*. (2002) argue that only stable complexes show coexpression in yeast], or because of biological differences. In *Caenorhabditis elegans*, a comparison of coexpression clusters (from microarray data) and interacting pairs (from yeast two-hybrid assays) for genes expressed in the germline found 50% of interacting pairs in the same coexpression cluster, versus 23% for random pairs (Walhout *et al*., 2002). This is in line with our results: at $p < 0.1$ and $\hat{r} > 0$, LNP found coexpression in 32.5% of interacting pairs versus 16.5% of control pairs.

Table 3. Sensitivity comparison using 50 000 test pairs of human genes with shared GO annotations and 50 000 control pairs without shared GO annotations

Expect	LNP cutoff	No.	Fisher cutoff	No.	Pearson cutoff	No.	EDVP(r) cutoff	No.
15 000	0.50	17 875	8.1e−9	**17 335**	0.058	*17 559*	0.13	17 835
10 000	0.26	12 720	7.1e−12	**12 041**	9.1e−3	*12 422*	0.12	*12 410*
5000	0.074	7146	3.9e−17	**6471**	1.2e−4	**6708**	0.11	7032
2500	0.017	4123	1.5e−22	**3485**	6.3e−7	**3785**	0.10	3964
500	2.6e−04	1424	1.8e−34	**929**	7.6e−14	**1062**	0.09	**1130**
50	9.6e−10	329	6.7e−51	**197**	5.7e−32	**194**	0.07	*269*

For each measure and for %below ranging from 30% down to 0.1%, we show the p-value cutoff given by the appropriately ranked pair in the control distribution. We then count the number of test pairs below the cutoff, and compare to the expected count if the distributions were the same. The low p cutoffs for the LNP test at low percentiles suggest that some of the control pairs are actually coexpressed. The extremely low p cutoffs for the other tests are further evidence that they do not report significance correctly. All counts are significantly different from random expectation ($p < 10^{-20}$, binomial test). Counts significantly smaller than the LNP counts at $p < 0.01$ and < 0.05 are in bold and italics, respectively (χ^2 test of proportions). Pairs were generated randomly, with genes appearing at the same rates in test and control sets. The EDVP cutoffs are presented as root mean square differences in $\vec{r}$.

The coexpression of GO categories was also weak: 1% of random pairs are as coexpressed as the best 2.8% of functionally related pairs. More highly expressed genes do show a larger effect: among genes with at least 300 ESTs (mean $f = 1.4 \cdot 10^{-4}$), the top 6.6% of functionally related pairs are as coexpressed as the top 1% of random pairs. This is consistent with the simulations showing that more data would be required to identify most coexpressed pairs reliably.

Two additional obstacles remain to finding coexpressed genes in EST data. First, the libraries may not reflect the original gene frequencies: the library creation process might not only amplify certain genes systematically, or add noise to the observations, but also create spurious correlations between genes. For example, some libraries are size-selected, which might well create a systematic bias. Furthermore, our implementation of LNP tosses out normalized and subtracted libraries. The model of binomial sampling could be modified to account for systematic biases, but quantitative data on the effects of normalization (Bonaldo *et al*., 1996) show widely varying effects on different genes at similar levels. Simply using the normalized libraries with LNP gives slightly worse results on the GO validation and slightly better results on the MINT test (data not shown).

Second, errors in EST-gene assignments may be affecting our results. For example, we are unable to analyze a reported cluster of coexpressed genes in human dbEST (Thompson *et al*., 2002) with LNP because most of the ESTs for some genes have been reassigned to other UniGenes in recent builds. Similarly, the Pearson correlation coefficients for genes assigned to the same coexpression clusters by Gitton *et al*. (2002) would have been quite different if they had used UniGene's assignments rather than a sequence similarity threshold (Spearman rank correlation between the two measures 0.51, average difference 0.22, with 383 pairs). Unfortunately, techniques for assigning ESTs to genes show differences but it is unclear which are more accurate (Bouck *et al*., 1999).

CONCLUSIONS

We present a new measure for identifying coexpressed genes from counts-based data, based on a LNP for the distribution of each gene's frequency that allows us to take library sizes into account. We test LNP against previous measures on both simulated data and human dbEST. In simulations, LNP reports the p-values for coexpression correctly: thus, we can run genome-wide tests of coexpressed pairs with statistical confidence in the results. Furthermore, the LNP test is powerful enough to find large numbers of statistically significant relationships in human dbEST. In contrast, previous measures report spurious coexpression at high rates. The parameters in our simulations are consistent with LNP's predicted parameters for real genes (Supplementary Figure 3), suggesting that these problems are biologically relevant.

We justify the use of the LNP on real data by reference to microarray data and by the success of the validation experiments. The LNP measure outperforms the other measures for finding coexpression between pairs of proteins with similar known function (matching GO terms), finding 34 significantly coexpressed categories and showing a significantly greater separation between pairs with known functional relationships and random controls. Furthermore, LNP, Fisher and EDVP show significantly more coexpression in interacting proteins than in controls. Finally, the simulations show that the data need not match the prior for the method to work, and that results should improve dramatically with twice as much data. More data and improved EST-gene matching will enable coexpressed genes to be identified more reliably and will elucidate further the relationship between expression patterns and biological function.

ACKNOWLEDGEMENTS

We thank Rob Ewing, Alkes Price, Wayne Volkmuth and Michael Walker for helpful discussions, and FXPAL for supporting this research.

REFERENCES

Bonaldo,M.F., Lennon,G. and Soares,M.B. (1996) Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.*, **6**, 791–806.

Bortoluzzi,S., d'Alessi,F., Romualdi,C. and Danieli,G.A. (2000) The human adult skeletal muscle transcriptional profile reconstructed by a novel computational approach. *Genome Res.*, **10**, 344–349.

Bouck,J., Yu,W., Gibbs,R. and Worley,K. (1999) Comparison of gene indexing databases. *Trends Genet.*, **15**, 159–162.

Boutanaev,A.M., Kalmykova,A.I., Shevelyov,Y.Y. and Nurminsky,D.I. (2002) Large clusters of co-expressed genes in the *Drosophila* genome. *Nature*, **420**, 666–669.

Brody,J.P., Williams,B.A., Wold,B.J. and Quake,S.R. (2002) Significance and statistical errors in the analysis of DNA microarray data. *Proc. Natl Acad. Sci., USA*, **99**, 12975–12978.

Camon,E., Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Oinn,T., Maslen,J., Cox,A. and Apweiler,R. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL and InterPro. *Genome Res.*, **13**, 662–672.

Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteom.*, **1**, 349–356.

Ewing,R.M. and Claverie,J.-M. (2000) EST databases as multiconditional gene expression datasets. *Pac. Symp. Biocomput.*, **5**, 427–439.

Ewing,R.M., Kahla,A.B., Poirot,O., Lopez,F., Audic,S. and Claverie,J.M. (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.*, **9**, 950–959.

Gibbons,F.D. and Roth,F.P. (2002) Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.*, **12**, 1574–1581.

Gitton,Y., Dahmane,N., Baik,S., Ruiz,I., Altaba,A., Neidhardt,L., Scholze,M., Herrmann,B.G., Kahlem,P., Benkahla,A. *et al.* (2002) A gene expression map of human chromosome 21 orthologues in the mouse. *Nature*, **420**, 586–590.

Hoyle,D.C., Rattray,M., Jupp,R. and Brass,A. (2002) Making sense of microarray data distributions. *Bioinformatics*, **18**, 576–584.

Jansen,R., Greenbaum,D. and Gerstein,M. (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res.*, **12**, 37–46.

Kothapalli,R., Yoder,S.J., Mane,S. and Loughran,T.P.,Jr (2002) Microarray results: how accurate are they? *BMC Bioinformatics*, **3**, 22.

Kuo,W.P., Jenssen,T.K., Butte,A.J., Ohno-Machado,L. and Kohane,I.S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.

Kuznetsov,V.A., Knott,G.D. and Bonner,R.F. (2002) General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics*, **161**, 1321–1332.

Nelander,S., Mostad,P. and Lindahl,P. (2003) Prediction of cell type-specific gene modules: identification and initial characterization of a core set of smooth muscle-specific genes. *Genome Res.*, **13**, 1838–1854.

Niehrs,C. and Pollet,N. (1999) Synexpression groups in eukaryotes. *Nature*, **402**, 483–487.

Okubo,K., Hori,N., Matoba,R., Niiyama,T., Fukushima,A., Kojima,Y. and Matsubara,K. (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat. Genetics*, **2**, 173–179.

Rao,C.R. (1965) *Linear Statistical Inference and Its Applications*. John Wiley & Sons, NY, pp. 347–352.

Thompson,H.G., Harris,J.W., Wold,B.J., Quake,S.R. and Brody,J.P. (2002) Identification and confirmation of a module of coexpressed genes. *Genome Res.*, **12**, 1517–1522.

Walhout,A.J., Reboul,J., Shtanko,O., Bertin,N., Vaglio,P., Ge,H., Lee,H., Doucette-Stamm,L., Gunsalus,K.C., Schetter,A.J. *et al.* (2002) Integrating interactions, phenome, and transcriptome mapping data for *C. elegans* germline. *Curr. Biol.*, **12**, 1952–1958.

Walker,M.G., Volkmuth,W., Sprinzak,E., Hodgson,D. and Klingler,T. (1999) Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res.*, **9**, 1198–1203.

Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.