3728–3737 Nucleic Acids Research, 2008, Vol. 36, No. 11 Published online 15 May 2008
doi:10.1093/nar/gkn233

Beyond tissueInfo: functional prediction using tissue expression profile similarity searches

Daniel Aguilar[1,2,3], Lucy Skrabanek[1,2], Steven S. Gross[4], Baldo Oliva[3] and Fabien Campagne[1,2,*]

[1]HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, [2]Department of Physiology and Biophysics, Weill Medical College of Cornell University, 1305 York Ave, New York, NY 10021, USA, [3]Structural Bioinformatics Group (GRIB-IMIM), Universitat Pompeu Fabra, C/Doctor Aiguader, 88. Barcelona-08003, Spain and [4]Department of Pharmacology, Weill Medical College of Cornell University, 1300 York Ave, New York, NY 10021, USA

Received January 22, 2008; Revised March 31, 2008; Accepted April 11, 2008

ABSTRACT

We present and validate tissue expression profile similarity searches (TEPSS), a computational approach to identify transcripts that share similar tissue expression profiles to one or more transcripts in a group of interest. We evaluated TEPSS for its ability to discriminate between pairs of transcripts coding for interacting proteins and non-interacting pairs. We found that ordering protein–protein pairs by TEPSS score produces sets significantly enriched in reported pairs of interacting proteins [interacting versus non-interacting pairs, Odds-ratio (OR) = 157.57, 95% confidence interval (CI) (36.81–375.51) at 1% coverage, employing a large dataset of about 50 000 human protein interactions]. When used with multiple transcripts as input, we find that TEPSS can predict non-obvious members of the cytosolic ribosome. We used TEPSS to predict S-nitrosylation (SNO) protein targets from a set of brain proteins that undergo SNO upon exposure to physiological levels of S-nitrosoglutathione *in vitro*. While some of the top TEPSS predictions have been validated independently, several of the strongest SNO TEPSS predictions await experimental validation. Our data indicate that TEPSS is an effective and flexible approach to functional prediction. Since the approach does not use sequence similarity, we expect that TEPSS will be useful for various gene discovery applications. TEPSS programs and data are distributed at http://icb.med.cornell.edu/crt/tepss/index.xml.

INTRODUCTION

Expressed sequence tags (ESTs) are oligonucleotide sequences obtained by the automated sequencing of cDNA clones (1). Initially introduced as a cost effective method to discover new human genes, for physical map construction and to discover coding regions in a genome, EST sequencing has rapidly become complementary to the sequencing of genomic sequences. The central repository of ESTs, dbEST, which contained only 22 537 ESTs 15 years ago (2), now contains more than 44 million ESTs in 2007. The organisms most represented in dbEST are human (∼9 M ESTs) and mouse (∼4.5 M ESTs), followed by zebrafish (1.3 M ESTs) and *Bos taurus* (1.3 M ESTs) (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).

Several successful gene discovery strategies have used ESTs. For instance, many groups have searched for ESTs that match representative proteins from a gene family of interest to identify potential new members of that family. ESTs which overlap with the coding sequence of the transcript may help to prioritize novel members of that protein family, if they share sequence similarity with the group of genes of interest. This strategy has been used to clone, for instance, novel members of the G-protein-coupled receptor family (3,4) and new kinases (5,6).

When a candidate is not expected to share significant sequence similarity to known proteins, EST information can still be used to guide gene discovery. In such cases, EST data can be used to filter genes by tissue expression profile, to eliminate genes unlikely to encode the function of the candidate. For instance, a taste receptor is expected to have its tissue expression restricted to taste buds. Since taste buds are not explicitly represented in dbEST, we would not expect a taste receptor to match a large number

*To whom correspondence should be addressed. Tel: 646 962 5613; Fax: 646 962 0383; Email: fac2003@med.cornell.edu of ESTs. This strategy was used as one of the filtering steps in the gene discovery pipeline that helped identify the sweet taste receptor Tas1R3 (7).

In this article, we show how EST data can be organized to perform tissue expression profile similarity searches (TEPSS) and used to identify transcripts that are functionally related to a query sequence, but need not share sequence similarity. The TEPSS approach builds on previous studies that observed a correlation between gene expression and function [e.g. (8–10) and references therein], and extends this observation from microarray time course experiments to the large body of expression data in dbEST. We validate the TEPSS approach by showing that it can discriminate between pairs of transcripts randomly chosen in the genome and pairs of transcripts that code for proteins that were previously reported to either directly interact or contribute to the same metabolic pathway. Significantly, we show that TEPSS can be used for gene function prediction, such as required for gene discovery, and we evaluate the predictive ability of the method to rank proteins that are either part of the ribosome complex or interact with it. Finally, we apply TEPSS to the prediction of novel targets of human protein S-nitrosylation (SNO) and show that TEPSS greatly outperforms the effectiveness of a random predictor, identifies independently validated SNO targets and suggests several SNO targets for experimental validation.

MATERIALS AND METHODS

TiDumpCounts

Counts of ESTs matching each transcript of an organism are tabulated and stored in binary format. We proceed as previously described to annotate the transcripts of a genome with tissue information from dbEST (11,12). Genomic information was obtained from human Ensembl build 44 (human NCBI build 36). Files produced by TissueInfo [called *tiac* files in ref. (11)] were processed to yield a table of data where each row is a transcript and each column a tissue. Each element of this table represents the count of ESTs, sequenced from a given tissue, that match a transcript. We call this table the TissueInfo count data and denote this data structure *tcd*. TissueInfo count data is written in binary format with Elias delta coding. Count files are produced for human and mouse and are distributed on the TissueInfo web site.

Single query searches

The TissueInfo similarity search engine performs exhaustive searches. A query transcript q_i is scored against every transcript t_j of the transcriptome. (Sample scoring schemes are described in Supplementary Material.) All the transcripts t_j are ordered by their TiSimilarity score, and the k highest-scoring transcripts are presented to the user. This search method is implemented in the program TiSimilarity (option –`mode single` or –`mode list`).

Multiple query searches

Multiple queries can be provided to search a transcriptome. In this case, single query searches are performed for each query q_i as described above and results are combined to yield transcripts most similar to the query transcripts as a group. Combining results can be done by summing the scores for transcript t_i over all queries (sum of scores combination, Equation 1). An alternative strategy is to combine results by summing the inverse of the rank position for transcript t_i over all queries (rank fusion combination, Equation 2).

$$\text{score}(t_i) = \sum_{q_j \in Q} \text{score}(t_i, q_j),$$

where Q is the set of transcripts used as query.

Equation 1. Sum of score combination.

$$\text{score}(t_i) = \sum_{q_j \in Q} \frac{1}{\text{rank}(t_i, q_j)},$$

where Q is the set of transcripts used as query, and rank(t_i, q_j) denotes the rank of transcript t_i in the result list obtained for the search done with query q_j.

Equation 2. Rank fusion combination.

TEPSS scorers

A tissue expression profile scorer evaluates a score for each pair of transcripts and can be formally written as score (counts(*tcd*, t_i), counts(*tcd*, t_j)). (where *tcd* is the TissueInfo count data.) The function counts (*tcd*, t_i) returns the number of counts for transcript t_i in the form of an array of integers, one element for each tissue. In the current implementation of TiSimilarity, scores are represented as double precision floating numbers. Various scoring strategies can be used to quantify the agreement between tissue expression profiles of two transcripts. Several strategies were tested and are described in the Supplementary Material.

Interaction network

The PIANA software (13) was used to assemble the human protein–protein interaction network (PIN) by integrating data from the following data sources: IntAct, DIP, BIND, MINT, MIPS and HPRD (14–19). The PINs were assembled as undirected and unweighted graphs. The complete human PIN contained 9937 nodes (here we use the Ensembl Gene IDs coding for the proteins) connected by 50 378 edges (i.e. interactions), where two nodes are connected by an edge if the proteins encoded by the genes have been reported to interact by at least one of the data sources. The experimental method(s) used to characterize each interaction were also recorded. This information was used to further build PINs containing only interactions supported by four or more experimental methods (which contained 1120 nodes and 912 edges).

Metabolic pathways

The genes involved in the different human metabolic pathways were downloaded from the KEGG database (20). In total, we considered 198 metabolic pathways containing 3869 distinct genes (Ensembl Gene IDs).

Estimation of reported pair enrichment

We wished to estimate the improvement in the likelihood of identifying true protein–protein interaction pairs in a sample of protein pairs using TEPSS. We first calculated the background chance of picking a reported interacting pair of proteins in a sample of 5×10^6 random protein pairs. The background chance was 0.0224%. Then, this fraction was compared to the probability of picking a reported interacting pair from the fraction of pairs from the sample which scored above a given threshold. Different scoring thresholds were established which provided coverage of 1, 5, 10 and 50%. This process was repeated for all the TEPSS scorers. Reported odds-ratio (OR) and 95% confidence intervals were calculated with http://www.hutchon.net/ConfidOR.htm.

Ribosome

The members of the cytosolic and mitochondrial ribosomal complexes were identified through their gene ontology annotations using BioMart. The GO IDs used were the following (GO term and number of Ensembl gene IDs also indicated): GO:0005830 (*cytosolic ribosome*, 94), GO:0005761 (*mitochondrial ribosome*, 71).

Gene/transcript mapping

Most protein–protein interaction databases associate each interacting protein with a gene identifier. Every protein in our PIN is represented by the Ensembl identifier for the gene encoding that protein. Since TiSimilarity compares the expression profiles of transcripts (not genes), we obtained gene-to-transcript relationships from BioMart and used this information to map gene/protein interaction pairs to transcript pairs (21). Because one gene may encode for more than one transcript (and because it is unclear which transcripts encoded by a gene code for the proteins that interact), for a given gene pair (g_i, g_j), we evaluate all possible pairs of transcripts (t_i, t_j) such that g_i encodes transcripts t_{ik} and g_j encodes transcripts t_{jl}. In all analyses, we consider the score of the gene pair to be the maximum score over all pairs (t_i, t_j), such that the TiSimilarity score between any two genes g_i and g_j is assumed to be the best score resulting from the pairwise comparison of all the transcripts of g_i versus all the transcripts of g_j.

$$\text{score}(g_i, g_j) = \max \{\text{score}(t_{ik}, t_{jl})\}$$

Equation 3. Deriving scores of gene pairs from scores of transcript pairs.

TEPSS score distributions

We estimated score distributions for interacting and non-interacting pairs of proteins. All known interacting pairs of proteins were selected from the human PIN. Samples of non-interacting pairs were generated by randomly pairing proteins whose interaction is not recorded in the databases. It is true that, in the absence of an experimentally validated negative gold-standard for the interactome, this may yield samples that include interactions which have not yet been identified. However, known interactions account for approximately 0.02% of the total pairwise combinations between proteins in our dataset and selecting non-interacting pairs uniformly at random is deemed as an unbiased estimator of the true negative gold-standard (22).

Density plots of TEPSS scores were created for the complete PIN and the PIN whose interactions were supported by four or more pieces of evidence. The samples of non-interacting pairs were of the same size as their interacting counterparts (i.e. 50 378 pairs for the complete PIN, 912 pairs for the PIN supported by four or more pieces of evidence) to ensure comparability of break-even scores.

For human metabolic pathways, we built the TEPSS score distributions for samples of 5000 gene pairs coding for proteins in the same metabolic pathway. In this case, the negative set was built by randomly pairing genes belonging to different metabolic pathways. Plots were generated with the *R* statistical package (23).

Shuffled count data

To establish a negative control, we produced shuffled versions of the TissueInfo counts data for each transcript in each organism. The shuffling procedure performs a random permutation of the EST counts observed for each transcript. The procedure guarantees that the sum of the counts for a given transcript is the same before and after permutation. The counts for all pairs of proteins in the aforementioned samples of 5000 non-interacting protein pairs were shuffled and the resulting distribution compared with the distribution of known interacting pairs.

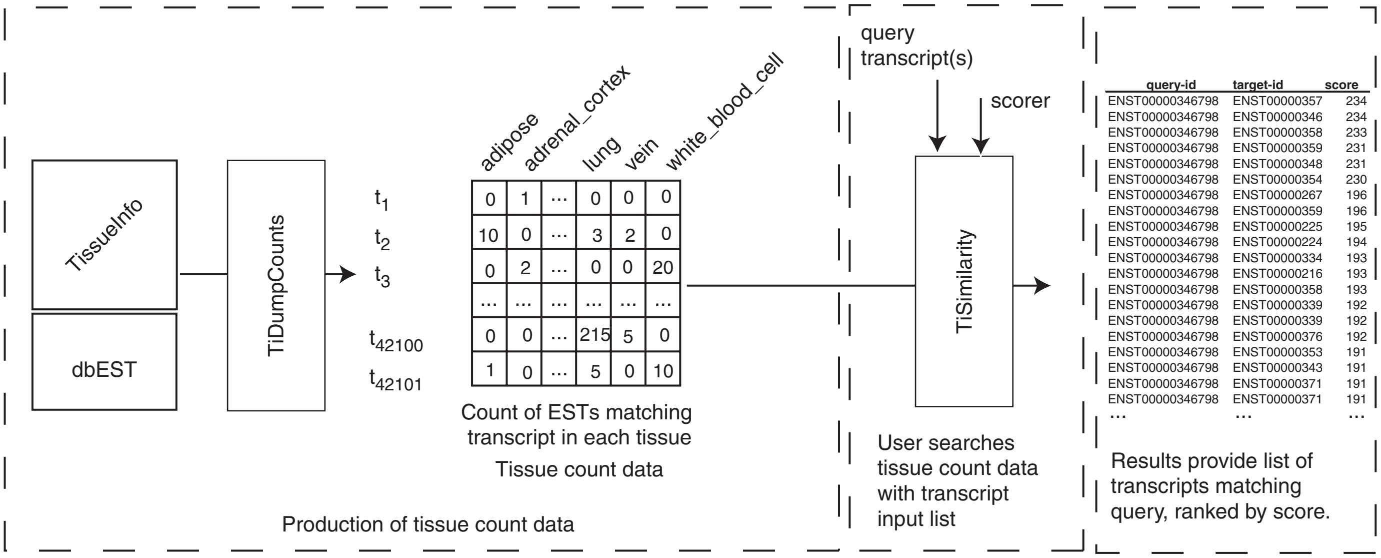
RESULTS

TEPSS approach

We present an extension of TissueInfo (11) to perform TEPSS. Figure 1 presents an overview of the TEPSS approach. TEPSS leverage dbEST and the TissueInfo curated tissue information to produce a table of EST counts for each tissue in which a transcript is expressed. The count data are provided on the TissueInfo web site and are the equivalent of formatted databases used in sequence similarity searches. The TEPSS search engine (TiSimilarity) makes it possible to scan count data for a transcriptome and produce ranked lists of transcripts, ordered by degree of similarity to the tissue expression profiles of one or more query transcripts. Figure 2 illustrates how tissue expression profile similarity scores are evaluated from the tissue count data (see minimum evidence scorer in the Methods section for a formal description of this scoring method).

TEPSS scores correlate with direct protein–protein interactions

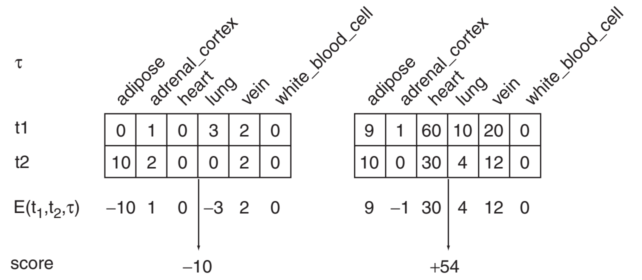
It has been shown that interacting proteins tend to be co-expressed (10). Similarly, proteins expressed in the same tissues are expected to be more likely to interact directly than proteins expressed in different tissues. TEPSS scores quantify the level of agreement between


TissueInfo
dbEST
TiDumpCounts
Count of ESTs matching transcript in each tissue
Tissue count data
Production of tissue count data
query transcript(s)
scorer
TiSimilarity
User searches tissue count data with transcript input list
Results provide list of transcripts matching query, ranked by score.

	adipose	adrenal_cortex		lung	vein	white_blood_cell
t_1	0	1	...	0	0	0
t_2	10	0	...	3	2	0
t_3	0	2	...	0	0	20
	...	...	...	...	...	...
t_{42100}	0	0	...	215	5	0
t_{42101}	1	0	...	5	0	10

query-id	target-id	score
ENST00000346798	ENST00000357	234
ENST00000346798	ENST00000346	234
ENST00000346798	ENST00000358	233
ENST00000346798	ENST00000359	231
ENST00000346798	ENST00000348	231
ENST00000346798	ENST00000354	230
ENST00000346798	ENST00000267	196
ENST00000346798	ENST00000359	196
ENST00000346798	ENST00000225	195
ENST00000346798	ENST00000224	194
ENST00000346798	ENST00000334	193
ENST00000346798	ENST00000216	193
ENST00000346798	ENST00000358	193
ENST00000346798	ENST00000339	192
ENST00000346798	ENST00000339	192
ENST00000346798	ENST00000376	192
ENST00000346798	ENST00000353	191
ENST00000346798	ENST00000343	191
ENST00000346798	ENST00000371	191
ENST00000346798	ENST00000371	191
...	...	...

Figure 1. Overview of the TEPSS approach. EST counts are obtained from dbEST using TissueInfo for each transcript of an organism (TiDumpCount box). The program TiSimilarity performs TEPSS for a query transcript, with a scorer and count information. The result is a ranked list of transcripts ordered by tissue expression profile similarity to the query transcript.



τ	adipose	adrenal_cortex	heart	lung	vein	white_blood_cell
t1	0	1	0	3	2	0
t2	10	2	0	0	2	0
$E(t_1,t_2,\tau)$	–10	1	0	–3	2	0

τ	adipose	adrenal_cortex	heart	lung	vein	white_blood_cell
t1	9	1	60	10	20	0
t2	10	0	30	4	12	0
$E(t_1,t_2,\tau)$	9	–1	30	4	12	0

Figure 2. Minimum evidence scorer. This plot illustrates how scores are calculated by the minimum evidence scorer. This example shows two transcripts with EST counts in six tissues (in practice, the TEPSS scorer uses more than 100 tissues). Tissues are denoted by the index τ. Values $E(\mathrm{t1},\mathrm{t2},\tau)$ are the evidence scores defined in Supplementary Equation 3, i.e. $E(t1, t2, \tau) = E(\mathrm{counts}(\mathrm{tcd}, t1, \tau), \mathrm{counts}(\mathrm{tcd}, t2, \tau))$. The example on the left shows two transcripts that yield a negative TEPSS score (−10), while the example on the right shows two transcripts with a positive score (+54).

the tissue-dependent expression levels of two transcripts in tissues. We therefore asked if TEPSS scores calculated by TiSimilarity would be differently distributed for pairs of interacting and non-interacting proteins. Figure 3 plots the distribution of TEPSS scores (Confidence scorer, see Supplementary Material section) for interacting and non-interacting protein pairs from the complete human PIN. One of the main drawbacks in the use of PIN-derived information is that data in the PIN are known to contain a remarkable amount of noise in the form of false positives (artifactual interactions deemed as true interactions) and false negatives (true but yet-undetected interactions) (24). The larger the number of pieces of independent experimental evidence that support an interaction, the less likely is the interaction to have been falsely identified. We therefore also analyzed the distributions of TEPSS scores from samples of interacting pairs supported by four or more pieces of experimental evidence. The distributions in both cases are significantly different from distributions observed using non-interacting protein pairs (Wilcoxon rank sum test, two tailed, $P < 2.10^{-1074}$) and the protein pairs that interact show more positive TEPSS scores than are observed for non-interacting protein pairs. Also, both distributions were significantly different from shuffled count data (see Methods section). The process of shuffling the count data destroys any correlation between tissue-specific expression levels, creating transcripts with fictional expression profiles. Conversely, non-interacting pairs probably show some correlation between the expression levels in different tissues because they contain not only a certain fraction of interacting pairs but also protein pairs which are related functionally (e.g. metabolically). The distribution of the scores of the predicted interacting pairs in the high-confidence set of Espadaler *et al.* (called I_2, containing pairs of proteins sharing domains with reported interacting proteins) were also consistent with this observation (25,26). Figure 4 shows that the same observation is true for pairs of proteins selected from the same metabolic pathway versus pairs selected from different metabolic pathways. Direct interaction between proteins is often considered as a good predictor of biological function. Similarly, proteins that are members of the same metabolic pathways are also considered functionally related. Our data demonstrate that TEPSS scores significantly correlate with the likelihood that two proteins directly interact, or are members of the same metabolic network.

TEPSS scorers

We implemented various TEPSS scoring schemes and tested their ability to separate protein pairs reported as interacting from those not known to be interacting. Table 1 summarizes the discriminative power of each scorer (scorers are described in the Supplementary

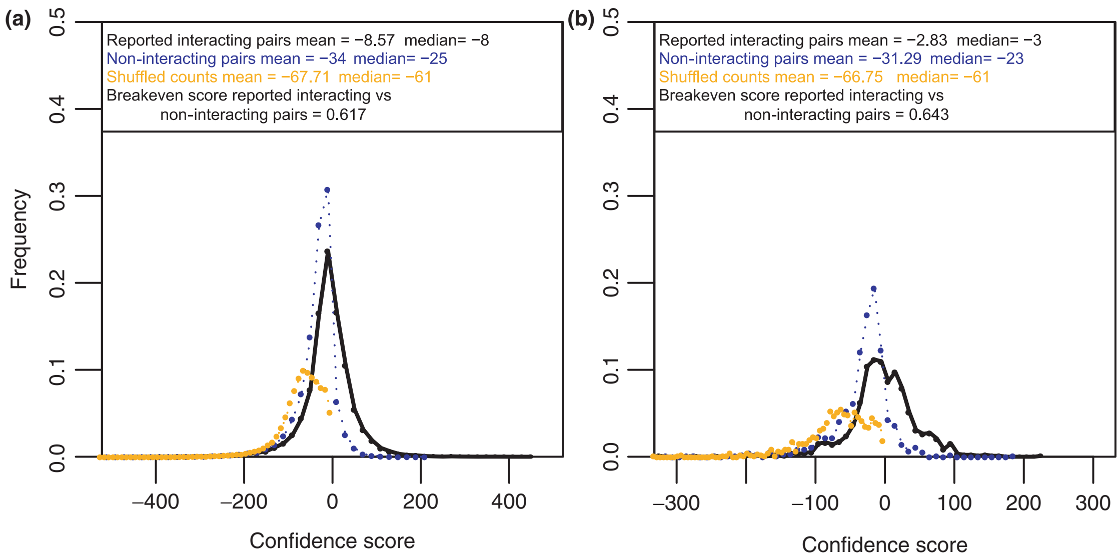


Figure 3. TEPSS scores for interacting and non-interacting pairs. Scores are calculated with the confidence scorer (see Supplementary Material section). (**a**) Distribution of scores for interacting protein pairs in the complete human PIN. (**b**) Distribution of scores for interacting protein pairs in the human PIN supported by four or more experimental methods. The difference between the distribution of the score for reported interacting protein pairs and non-interacting protein pairs is statistically significant in both cases (Wilcoxon rank sum test, two tailed). The difference between the distribution of the TEPSS confidence scores for interacting protein pairs and the distribution of TEPSS confidence scores of randomly shuffled counts is also significant in both cases (Wilcoxon rank sum test, two tailed).

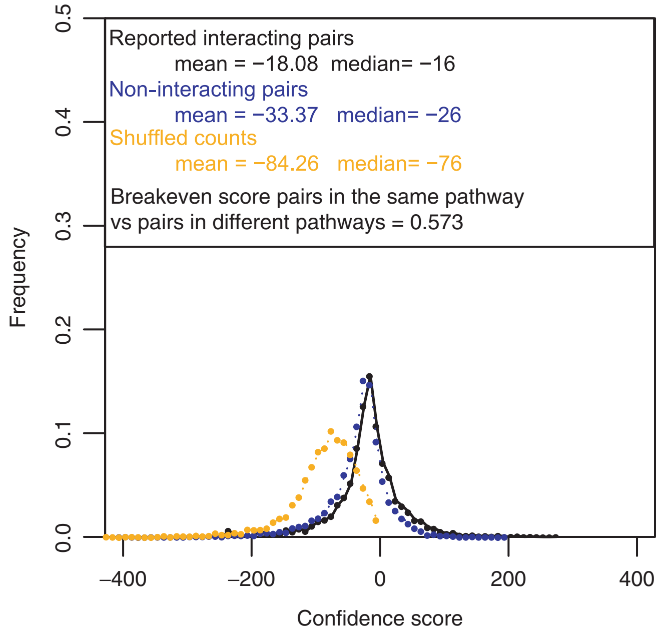


Figure 4. TEPSS scores for metabolic pathways. Distribution of TEPSS scores for 5000 pairs of genes coding for enzymes contributing to the same metabolic pathway in human. The differences with respect to that of enzymes contributing to different pathways is significant (Wilcoxon rank sum test, two tailed). The difference with respect to the distribution of scores of randomly shuffled counts is also significant.

Material section). Values shown in Table 1 are the break-even points of precision-recall curves when a given TEPSS scorer is used for prediction in a balanced dataset. For instance, the break-even point value of 0.749 obtained for the binary scorer indicates that a precision of 74.9% at 74.9% recall can be achieved when predicting reported protein–protein interaction pairs based on the value of the normalized confidence score (since the dataset is balanced, a break-even point of 0.50 would indicate a random prediction). The minimum evidence scorer is moderately effective. Scorers are arranged from left to right by increasing performance (second line). The best prediction is achieved by the Binary Scorer, a conceptually simple scoring approach (see Supplementary Methods). Transcripts may have different baseline expression levels (e.g. some transcripts are expressed highly in most tissues, while others may have moderate expression). Since the binary scorer is immune to baseline differences, we tested if different baseline expression levels should be controlled by TEPSS scorers. The comparison between the performance measure of the baseline normalized confidence scorer and the normalized confidence scorer indicates that this appears to be the case. However, the performance of the baseline normalized confidence scorer is comparable to that of the simpler binary scorer.

High TEPSS scoring pairs are enriched in reported interactions

We asked if selecting pairs of proteins that have the highest TEPSS scores is an effective strategy to predict true protein–protein interaction pairs. With the Binary Scorer, at a coverage of 1%, we observe that the likelihood of identifying a reported interaction pair is about 100 times the random expectation [2.5% versus 0.0224%, respectively; OR = 157.57, 95% confidence interval (CI) (36.81–375.51)]. At a coverage of 5%, the enrichment in reported interacting pairs is about eight-fold [0.189% versus 0.0224%; OR = 8.65, 95% CI (2.75–27.29)]. At a coverage of 50%, the enrichment in reported interactors is about 3-fold [0.0724% versus 0.0224%; OR = 4.73, 95% CI

Table 1. Discriminative power of the TEPSS scorers

	Minimum evidence scorer	Confidence scorer	Binary confidence scorer	Pearson scorer	Normalized confidence scorer	Baseline normalized confidence scorer	Binary scorer
Complete PIN	0.585 (0.001)	0.618 (0.001)	0.644 (0.001)	0.669 (0.002)	0.709 (0.002)	0.727 (0.001)	0.727 (0.002)
PIN supported by ≥4 pieces of evidence	0.628 (0.005)	0.651 (0.005)	0.661 (0.006)	0.703 (0.006)	0.733 (0.01)	0.741 (0.008)	0.749 (0.01)

Values shown are the break-even points of precision-recall curves when a given TEPSS scorer is used for prediction in a balanced dataset. Values in parentheses are the standard deviation of the break-even point.

(3.24–6.90)]. The enrichment observed for other TEPSS scorers is available in the Supplementary Table 1.

Ribosome TEPSS screen

The ribosome is a large macromolecular assembly of protein and RNA molecules. In eukaryotic cells, the inner mitochondrial membrane contains a particular type of ribosome, different in composition from the cytosolic one. We asked if the TEPSS approach could efficiently identify ribosomal transcripts from the human transcriptome. Figure 5 presents a lift curve constructed by the leave-one-out validation approach. (Supplementary Figure 1 presents a lift curve constructed by 10-fold cross-validation.) The curve confirms that TEPSS rank proteins of the cytosolic ribosome with better ranks than expected by chance. The number of reported predictions in the top k results is significantly larger than the number that would be obtained by a random prediction (for all k, $p = 10^{-10}$) (27). A TEPSS screen can also rank transcripts of the mitochondrial ribosome with better ranks than expected by chance (data not shown).

Supplementary Table 2 provides the list of transcripts ranked first when TEPSS is used with all the GO annotated transcripts of the cytosolic ribosome. Inspection of the list reveals that most genes top ranked are true transcripts of ribosomal proteins (RPs). Surprisingly, a few genes that are not annotated as RPs in their descriptions rank highly in the TEPSS output. For instance, the receptor for activated protein kinase C (RACK1) is ranked among the 30 top transcripts prioritized by TEPSS in the case of the cytosolic ribosome. A literature search indicates that RACK1 has been shown to bind the ribosome by cryo-EM (28,29). Similarly, the translationally controlled tumor protein (TCTP) ranks in the top 30, and has been shown to bind to the translation elongation factor, eEF1A, and its guanine nucleotide exchange factor, eEF1Bbeta (30). Since eEF1A delivers aminoacyl-tRNAs to the A-site at the ribosome, TCTP may indeed associate with the ribosome and be involved in regulating the efficiency of protein translation (30). Since TCTP can also be secreted to act as a chaperone (31) out of the cell, it is remarkable that the TEPSS screen clearly detected its association with the cytosolic ribosome.

SNO TEPSS screen

S-nitrosylation (SNO) is the process by which certain cysteine residues covalently react with nitric oxide or a nitric oxide derived-species to yield post-translationally NO-modified proteins (32,33). Since the identification of these proteins is a challenging bioinformatics problem, we explored the ability of the TEPSS approach to predict SNO protein targets. Figure 6 shows the lift curve created by leave-one-out evaluation, using SNO targets identified in mouse brain by SNOSID (SNO Site Identification) (34). This lift curve deviates significantly ($P = 10^{-6}$ at rank 100) from the diagonal (random prediction). Over 25% of the test transcripts are found in 2% of the transcripts that rank highest by TEPSS score. However, we cannot rigorously assess significance because the complete set of *bona fide* SNO target proteins is currently unknown. Indeed, many proteins are counted as false positives in our evaluation because they have not yet been reported as SNO targets. However, it is likely that at least some of them could be targets of SNO, a conjecture supported by the following detailed analysis of the top 10 TEPSS predictions. At least one protein (protein disulfide

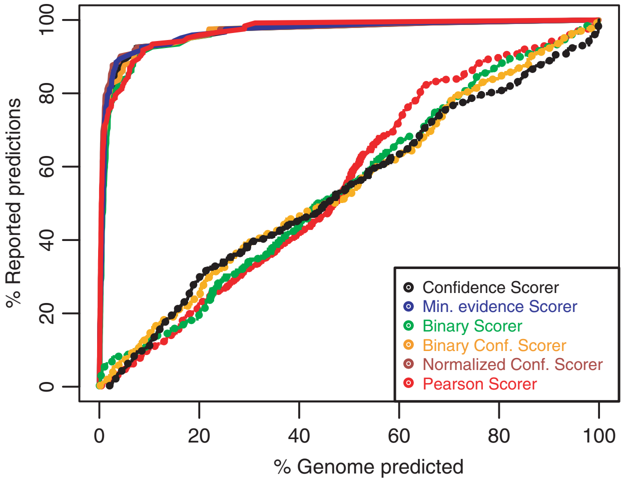


Figure 5. Screening for ribosomal transcripts. We used ribosomal transcripts as input to a whole human transcriptome TEPSS search. The plot shows a lift curve, constructed by leaving out one ribosomal gene of the input, and searching the genome with the rest of the ribosomal transcripts. The x-axis indicates at what relative rank in the genome the gene that was left out was found in the TEPSS output (fraction of total human transcripts represented in TissueInfo, or the proportion of the genome that needs to be inspected to find the left out gene). The y-axis indicates the proportion of ribosomal transcripts that would be found at the rank observed for the transcript (% reported_predictions). The diagonal of the plot represents the expected rate of random prediction. Dotted lines illustrate results obtained when random sets of transcripts are used as input with different scorers.

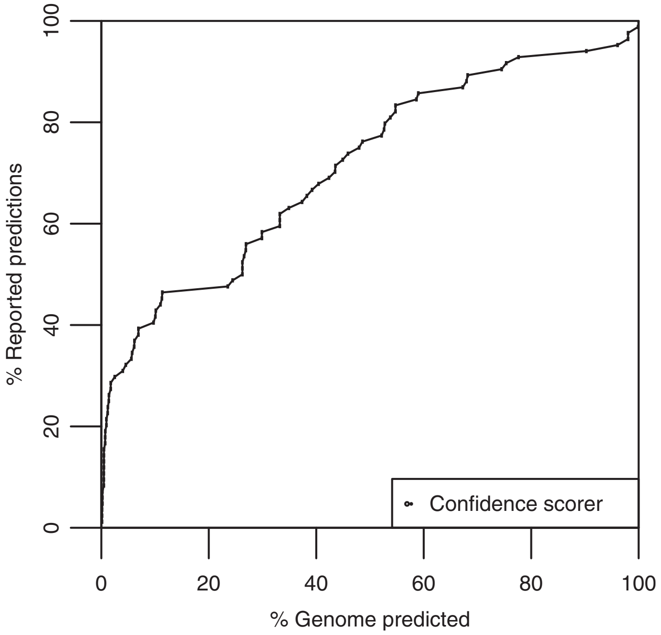


Figure 6. Screening for SNO protein targets. This lift curve illustrates that the TEPSS approach can effectively predict SNO targets in the human transcriptome.

isomerase, PDI) that was not used in the training set [and that has been previously shown to be a target of SNO (35)] was detected among the highest ranking results of the TEPSS screen. Another top scoring protein is the acyl Co-A desaturase. Although SNO has not been assessed for this enzyme, Marra *et al.* (36) reported that the activity of acyl Co-A desaturase is modulated by nitric oxide, a strong suggestion that acyl Co-A desaturase can be regulated by SNO. Finally, while it has not been shown that PP2A is S-nitrosylated, this enzyme, which also ranks highly in the TEPSS screen, dephosphorylates eNOS, thereby regulating the activity of one of three nitric oxide synthases (37). Association with eNOS would predictably increase the concentration of NO and NO-derived species that PP2A is exposed to, thereby increasing the likelihood that PP2A undergoes SNO. A list of the top 1000 predicted targets of SNO can be found in Supplementary Table 3. This list ranks candidate SNO protein targets for future experimental verification.

DISCUSSION

Gene expression platforms

There are four main sources of expression data: ESTs, serial analysis of gene expression (SAGE), massive parallel signature sequencing (MPSS) and microarrays (1,11,38–40). Microarrays are a popular gene expression platform because they support flexible experimental designs where a few conditions can be compared (i.e. tissue types, disease states or time courses). However, EST, SAGE and MPSS are methods that attempt to quantify the expression level of transcripts, and as such complement microarray platforms that are designed to measure fold changes between two experimental conditions. (Microarray platforms measure hybridization levels as an aggregate of probe hybridization properties and mRNA abundance and rely on fold change calculations to cancel out the effect of probe hybridization, see Relative expression values section in http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.) Since there is a large amount of EST data publicly available, and since a wide variety of human tissues have been sampled by EST sequencing projects, ESTs can be used to determine the organism-wide expression profile of a gene. Although we focused on this source of gene expression data in the present study, the methods that we describe here may be applicable to gene expression data generated from other gene expression platforms as well.

EST data management

Various systems have been developed to manage EST data and related information. A first category of systems is focused on clustering ESTs into sequence contigs likely to represent genes [i.e. TIGR Gene Indices (41), STACK (42) or Unigene (43)]. Another type of system complements dbEST with computational annotations to facilitate computational analyses. For instance, our group has previously developed the TissueInfo system to organize information in dbEST and make it possible to calculate tissue expression profiles for proteins, cDNAs or ESTs (11). TissueInfo offers a tissue hierarchy [also called the TissueInfo ontology (12)] and unambiguously maps the tissue provenance of EST libraries to nodes of the hierarchy (11). This type of organization greatly facilitates computational analyses with EST data and has been adopted in various systems developed after TissueInfo (44,45).

TissueInfo links to genomics databases

The TissueInfo system differs from collections of EST clusters. When TissueInfo is used to process a collection of known or predicted transcripts from a genome project, the tissue expression profiles of the transcript are unambiguously linked to all the annotations maintained by the genome database about these transcripts. Linking EST clusters to genomic information is possible, but requires non-trivial data integration (46). The TissueInfo web site offers tissue expression profiles calculated from Ensembl cDNAs that can be used directly for gene discovery. The data files have been updated periodically since 2001.

Key advantages of TEPSS

The extension of TissueInfo described in this manuscript makes it possible to perform TEPSS. This is a major departure from the current practice that consists in producing tables with tissue expression profiles and querying these tables to identify genes of interest. The two key differences are that (i) TEPSS do not require *a priori* knowledge of the tissue expression profile that is expected of gene candidates. The query formulation step is therefore eliminated. (ii) TEPSS return a *ranked* list of candidates that are prioritized by a quantitative score, whereas queries of tabular data return unordered sets of genes.

Methods which leverage microarray data

Various studies have presented how correlated expression in microarray datasets can help predict function. Most of these studies were performed on yeast datasets (10,47,48), and cannot be directly compared to the results that we present here since tissue EST data are not available for yeast. A few studies have focused on human datasets and provided web-based tools to identify the genes most correlated to a query gene across multiple datasets [see (49) and Gemma, available at http://www.bioinformatics.ubc.ca/pavlidis/lab/software.html]. As of writing, we are not aware that these tools can identify genes most correlated to a group of genes provided as input and therefore could not benchmark TEPSS against their results. Methods that leverage microarray data can often detect negative correlation between the expression levels of transcripts. The presence of negative correlation in a time course experiment may indicate negative feedback (i.e. one transcript codes for a protein which down regulates the expression of the second transcript). The TEPSS approach can identify negative correlation in the expression of transcripts across tissues (when scoring with the Pearson scorer, any score below zero indicates negative correlation). However, contrary to time course experiments, negative correlation is seldom observed across tissues. The primary reason for this difference is that for two transcripts to be negatively correlated in a time course experiment, the two transcripts must both be expressed in the same cell type (and thus tissue). These transcripts are therefore likely to yield TEPSS scores which indicate positive tissue expression correlation.

Novel contributions

In a pioneering work, Ewing and Claverie have shown that expression profiles from EST data can be used to predict functional similarity (8). The present study builds on this initial observation, but differs in the following specific contributions. (i) While Ewing and Claverie used a clustering approach to group functionally related Unigene clusters, we formalize the functional prediction problem as a similarity search in the space of tissue expression profiles. This formalism presents the advantage that we can predict transcripts that are similar to any arbitrary set of input transcripts, with respect to their tissue expression profile. (ii) A second contribution of this study is the description of tissue expression scoring schemes that outperform the Pearson correlation measure described in (8). (iii) We evaluate the TEPSS approach by leveraging protein–protein interaction datasets and the large amount of EST data available today and show that TEPSS can predict non-trivial functional relationships. (iv) Finally, our comparison of the predictive ability of different TEPSS scorers suggests that the level of expression of two transcripts in tissues is not a strong predictor of the ability of the proteins coded by the transcripts to interact (proteins appear equally likely to interact if they are expressed at similar or very different levels in the same tissue).

Scorers as tools for hypothesis testing

Comparing the performance for two different scorers can be used for hypothesis testing. For example, comparing the performance of the minimum evidence and confidence scorers may serve to test whether the level of transcript expression is an important factor when predicting interacting proteins, because these scorers differ in how they reward large quantitative expression differences in the same tissue. In the present evaluation, we found that proteins are as likely to interact whether they are expressed at similar levels in tissues or at different absolute levels. Pairing different scorers can help test additional hypotheses. The TEPSS program makes it easy to implement new scoring schemes to test new hypotheses with large protein–protein interaction datasets or other validation benchmarks. In addition to the *ad hoc* scorers presented in this manuscript, we anticipate that probabilistic scorers will also be developed for TEPSS. Performance evaluation of these scorers on different functional prediction tasks will be presented elsewhere.

Guidelines for scorer selection

Our results indicate that some of the TEPSS scorers perform better than others in discriminating between interacting and non-interacting protein pairs. According to this benchmark, the binary scorer clearly outperforms other scorers (Table 1). However it is not clear if the binary scorer is the best choice for each type of application TEPSS can be applied to. For instance, in the cytosolic ribosome screen (Figure 5), the normalized confidence scorer (second best performance on the protein–protein interaction benchmark) outperforms the binary scorer. This suggests that the choice of scorer will depend on the problem at hand. The lift curve shown in Figure 5 suggests a procedure to determine which scorer will work best for a given problem. In this case, different scorers were tried and the one that produced the best discrimination in a leave-one-out evaluation would be selected to perform the final prediction screen. The TEPSS program automates the calculations of the leave-one-out lift curves to facilitate scorer selection (option `–mode loo-forward`).

High TEPSS scoring pairs are enriched in reported interactions

Our results show an improvement in the likelihood of identifying reported pairs of interacting proteins with respect to random expectation. Other methods of predicting protein–protein interactions have been validated in a similar manner. For instance, Espadaler *et al.* achieved about a 6-fold improvement over random when predicting interacting pairs using structural data about protein–protein interfaces and sequence data (0.53% versus 0.09%, respectively) at 1.8% coverage [dataset I_2 in ref. (25)]. The 22-fold enrichment at the same coverage observed in the present study therefore compares favorably with a protein–protein interaction prediction method that leverages structural and sequence information.

Ribosomal screen

The TEPSS screen of the ribosome shows that the method correctly assigns high ranks to proteins that contribute to this macromolecular complex, confirming what others (10) have shown. Using a limited EST dataset from six tissues, Bortoluzzi *et al.* (50) found that there were 13 RPs with 'differential' expression profiles (i.e. they were significantly overexpressed in one tissue). TEPSS successfully give relatively high ranks to those RPs. Most interestingly, TEPSS also give high ranks to proteins that are not directly part of the ribosomal complex, but that are directly interacting with this complex.

SNO screen

We used TEPSS to predict novel targets of protein SNO. The identification of these proteins is a challenging bioinformatics problem that has hitherto not been solved successfully. For example, we have recently shown that the primary sequence flanking cysteine residues cannot account for the observed SNO selectivity (34). Greco *et al.* (51) recently reported that they identified a sequence motif responsible for SNO. However, this study did not adequately control for homology in the SNO proteins that made up the SNO training set (the training set included several proteins of the same family, which are expected to be conserved over the entire sequence). The negative training set was randomly chosen and did not include the compositional bias found in the SNO protein set. Since the composition of the sequences in an alignment is well known to bias the result of naïve estimates of conservation (52), it is likely that the 'motif' reported by Greco *et al.* can be explained in large part by the compositional bias of the SNO training set used in that study. Our report that SNO targets can be predicted with the TEPSS approach therefore offers a significant tool for unbiased identification of protein that may be susceptible to SNO.

In summary, TEPSS is a novel approach to functional prediction using EST data. TEPSS are especially useful in discovering functionally similar genes in higher eukaryotes, since EST data are plentiful for these organisms. Because TEPSS are not sequence based, they can be used as a complementary approach to approaches that rely on sequence similarity in the coding sequence. The TEPSS software is distributed under the Gnu General Public License as part of the TissueInfo package (available from http://icb.med.cornell.edu/crt/tissueinfo/index.xml).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Francesca Demichelis for her critical reading of the manuscript. The authors acknowledge support from the resources of the HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine and the David A. Cofrin Center for Biomedical Information at Weill Cornell. DA acknowledges financial support from the Juan de la Cierva program of the Spanish MEyC (INFOBIOMED-NoE, IST-507585). BO acknowledges the grant BIO02005-00533 from the MEyC. This work was supported by National Institutes of Health Grant HL80702 (to SSG). Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Adams,M., Kelley,J., Gocayne,J., Dubnick,M., Polymeropoulos,M., Xiao,H., Merril,C., Wu,A., Olde,B. and Moreno,R. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
2. Boguski,M., Lowe,T. and Tolstoshev,C. (1993) dbEST—database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.
3. O'Dowd,B., Nguyen,T., Marchese,A., Cheng,R., Lynch,K., Heng,H., Kolakowski,L. and George,S. (1998) Discovery of three novel G-protein-coupled receptor genes. *Genomics*, **47**, 310–313.
4. Marchese,A., Nguyen,T., Malik,P., Xu,S., Cheng,R., Xie,Z., Heng,H., George,S., Kolakowski,L. and O'Dowd,B. (1998) Cloning genes encoding receptors related to chemoattractant receptors. *Genomics*, **50**, 281–286.
5. Haridas,V., Ni,J., Meager,A., Su,J., Yu,G., Zhai,Y., Kyaw,H., Akama,K., Hu,J., Van Eldik,L. *et al.* (1998) TRANK, a novel cytokine that activates NF-kappa B and c-jun N-terminal kinase. *J. Immunol.*, **161**, 1–6.
6. Chen,H., Kung,H. and Robinson,D. (1998) Digital cloning: identification of human cDNAs homologous to novel kinases through expressed sequence tag database searching. *J. Biomed. Sci.*, **5**, 86–92.
7. Max,M., Shanker,Y., Huang,L., Rong,M., Liu,Z., Campagne,F., Weinstein,H., Damak,S. and Margolskee,R. (2001) Tas1r3, encoding a new candidate taste receptor, is allelic to the sweet responsiveness locus sac. *Nat. Genet.*, **28**, 58–63.
8. Ewing,R. and Claverie,J. (2000) EST databases as multi-conditional gene expression datasets. *Pac. Symp. Biocomput.*, 430–442
9. Allocco,D., Kohane,I. and Butte,A. (2004) Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, **5**, 18.
10. Jansen,R., Greenbaum,D. and Gerstein,M. (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.*, **12**, 37–46.
11. Skrabanek,L. and Campagne,F. (2001) TissueInfo: high-throughput identification of tissue expression profiles and specificity. *Nucleic Acids Res.*, **29**, E102–2.
12. Campagne,F. and Skrabanek,L. (2006) Mining expressed sequence tags identifies cancer markers of clinical interest. *BMC Bioinformatics*, **7**, 481.
13. Aragues,R., Jaeggi,D. and Oliva,B. (2006) PIANA: protein interactions and network analysis. *Bioinformatics*, **22**, 1015–1017.
14. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
15. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
16. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E. *et al.* (2005) The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424.
17. Chatr-aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2007) MINT: the molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.
18. Mewes,H.W., Frishman,D., Mayer,K.F., Munsterkotter,M., Noubibou,O., Pagel,P., Rattei,T., Oesterheld,M., Ruepp,A. and Stümpflen,V. (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.*, **34**, D169–D172.

19. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R., Raghavan,T.M. *et al.* (2006) Human protein reference database—2006 update. *Nucleic Acids Res.*, **34**, D411–D414.
20. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
21. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
22. Ben-Hur,A. and Noble,W.S. (2006) Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, **7 (Suppl 1)**, S2.
23. R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org
24. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.
25. Espadaler,J., Romero-Isart,O., Jackson,R. and Oliva,B. (2005) Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics*, **21**, 3360–3368.
26. Cockell,S., Oliva,B. and Jackson,R. (2007) Structure-based evaluation of in silico predictions of protein-protein interactions using comparative docking. *Bioinformatics*, **23**, 573–581.
27. Macskassy,S.A. (2004) Significance testing against the random model for scoring models on top k predictions. CeDER Working Paper CeDER-05-09; Stern School of Business, New York University, NY, NY 10012. December 2004.
28. Nilsson,J., Sengupta,J., Frank,J. and Nissen,P. (2004) Regulation of eukaryotic translation by the RACK1 protein: a platform for signalling molecules on the ribosome. *EMBO Rep.*, **5**, 1137–1141.
29. Sengupta,J., Nilsson,J., Gursky,R., Spahn,C., Nissen,P. and Frank,J. (2004) Identification of the versatile scaffold protein RACK1 on the eukaryotic ribosome by cryo-EM. *Nat. Struct. Mol. Biol.*, **11**, 957–962.
30. Cans,C., Passer,B., Shalak,V., Nancy-Portebois,V., Crible,V., Amzallag,N., Allanic,D., Tufino,R., Argentini,M., Moras,D. *et al.* (2003) Translationally controlled tumor protein acts as a guanine nucleotide dissociation inhibitor on the translation elongation factor eEF1A. *Proc. Natl Acad. Sci. USA*, **100**, 13892–13897.
31. Amzallag,N., Passer,B., Allanic,D., Segura,E., Thery,C., Goud,B., Amson,R. and Telerman,A. (2004) TSAP6 facilitates the secretion of translationally controlled tumor protein/histamine-releasing factor via a nonclassical pathway. *J. Biol. Chem.*, **279**, 46104–46112.
32. Lane,P., Hao,G. and Gross,S. (2001) S-nitrosylation is emerging as a specific and fundamental posttranslational protein modification: head-to-head comparison with O-phosphorylation. *Sci. STKE*, **2001**, RE1.
33. Derakhshan,B., Hao,G. and Gross,S. (2007) Balancing reactivity against selectivity: the evolution of protein S-nitrosylation as an effector of cell signaling by nitric oxide. *Cardiovasc. Res.*, **75**, 210–219.
34. Hao,G., Derakhshan,B., Shi,L., Campagne,F. and Gross,S. (2006) SNOSID, a proteomic method for identification of cysteine S-nitrosylation sites in complex protein mixtures. *Proc. Natl Acad. Sci. USA*, **103**, 1012–1017.
35. Uehara,T., Nakamura,T., Yao,D., Shi,Z., Gu,Z., Ma,Y., Masliah,E., Nomura,Y. and Lipton,S. (2006) S-nitrosylated protein-disulphide isomerase links protein misfolding to neurodegeneration. *Nature*, **441**, 513–517.
36. Marra,C., Nella,J., Manti,D. and de Alaniz,M. (2007) Lipid metabolism in rats is modified by nitric oxide availability through a ca+ +-dependent mechanism. *Lipids*, **42**, 211–228.
37. Urbich,C., Reissner,A., Chavakis,E., Dernbach,E., Haendeler,J., Fleming,I., Zeiher,A.M., Kaszkin,M. and Dimmeler,S. (2002) Dephosphorylation of endothelial nitric oxide synthase contributes to the anti-angiogenic effects of endostatin. *FASEB J.*, **16**, 706–708.
38. Velculescu,V., Zhang,L., Vogelstein,B. and Kinzler,K. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
39. Schena,M., Shalon,D., Davis,R. and Brown,P. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
40. Brenner,S., Johnson,M., Bridgham,J., Golda,G., Lloyd,D., Johnson,D., Luo,S., McCurdy,S., Foy,M., Ewan,M. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.
41. Lee,Y., Tsai,J., Sunkara,S., Karamycheva,S., Pertea,G., Sultana,R., Antonescu,V., Chan,A., Cheung,F. and Quackenbush,J. (2005) The TIGR gene indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res.*, **33**, D71–D74.
42. Christoffels,A., van Gelder,A., Greyling,G., Miller,R., Hide,T. and Hide,W. (2001) STACK: sequence tag alignment and consensus knowledgebase. *Nucleic Acids Res.*, **29**, 234–238.
43. Schuler,G.D., Boguski,M.S., Stewart,E.A., Stein,L.D., Gyapay,G., Rice,K., White,R.E., Rodriguez-Tome,P., Aggarwal,A., Bajorek,E. *et al.* (1996) A gene map of the human genome. *Science*, **274**, 540–546.
44. Brown,A., Kai,K., May,M., Brown,D. and Roopenian,D. (2004) ExQuest, a novel method for displaying quantitative gene expression from ESTs. *Genomics*, **83**, 528–539.
45. Ferguson,D., Chiang,J., Richardson,J. and Graff,J. (2005) eXPRESSION: an in silico tool to predict patterns of gene expression. *Gene Expr. Patterns*, **5**, 619–628.
46. Shklar,M., Strichman-Almashanu,L., Shmueli,O., Shmoish,M., Safran,M. and Lancet,D. (2005) GeneTide—terra incognita discovery endeavor: a new transcriptome focused member of the GeneCards/GeneNote suite of databases. *Nucleic Acids Res.*, **33**, D556–D561.
47. Huttenhower,C., Hibbs,M., Myers,C. and Troyanskaya,O. (2006) A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, **22**, 2890–2897.
48. Hibbs,M., Hess,D., Myers,C., Huttenhower,C., Li,K. and Troyanskaya,O. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23**, 2692–2699.
49. Lee,H., Hsu,A., Sajdak,J., Qin,J. and Pavlidis,P. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085–1094.
50. Bortoluzzi,S., d'Alessi,F., Romualdi,C. and Danieli,G.A. (2001) Differential expression of genes coding for ribosomal proteins in different human tissues. *Bioinformatics*, **17**, 1152–1157.
51. Greco,T., Hodara,R., Parastatidis,I., Heijnen,H., Dennehy,M., Liebler,D. and Ischiropoulos,H. (2006) Identification of S-nitrosylation motifs by site-specific mapping of the S-nitrosocysteine proteome in human vascular smooth muscle cells. *Proc. Natl Acad. Sci. USA*, **103**, 7420–7425.
52. Pei,J. and Grishin,N. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.