

Software

Open Access

PARE: A tool for comparing protein abundance and mRNA expression data

Eric Z Yu¹, Anne E Counterman Burba¹ and Mark Gerstein^{*1,2,3}

Address: ¹Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA, ²Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA and ³Department of Computer Science, Yale University, New Haven, CT 06520, USA

Email: Eric Z Yu - acmg@bioinfo.mbb.yale.edu; Anne E Counterman Burba - acmg@bioinfo.mbb.yale.edu; Mark Gerstein* - Mark.Gerstein@yale.edu

* Corresponding author

Published: 24 August 2007

Received: 13 October 2006

BMC Bioinformatics 2007, 8:309 doi:10.1186/1471-2105-8-309

Accepted: 24 August 2007

This article is available from: <http://www.biomedcentral.com/1471-2105/8/309>

© 2007 Yu et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Techniques for measuring protein abundance are rapidly advancing and we are now in a situation where we anticipate many protein abundance data sets will be available in the near future. Since proteins are translated from mRNAs, their expression is expected to be related to their abundance, to some degree.

Results: We have developed a web tool, called PARE (Protein Abundance and mRNA Expression; <http://proteomics.gersteinlab.org>), to correlate these two quantities. In addition to globally comparing the quantities of protein and mRNA, PARE allows users to select subsets of proteins for focused study (based on functional categories and complexes). Furthermore, it highlights correlation outliers, which are potentially worth further examination.

Conclusion: We anticipate PARE will facilitate comparative studies on mRNA and protein abundance by the proteomics community.

Background

Quantifying mRNA expression is currently one of the most exciting and active areas in genomics, and as a result there are copious amounts of data available. However, mRNA does not directly arbitrate biological function; that role is filled by proteins, which are translated from mRNA and connect mRNA to biological processes by acting on the latter as a regulator. In contrast to the abundance of mRNA expression data, considerably fewer quantitative protein expression datasets are available due to the comparatively recent advent of methods for measuring protein abundance on a large scale (such as the use of ICAT and iTRAQ with mass spectrometry) and the difficulty of these experiments relative to chip-based mRNA studies. Also,

for many studies, protein abundance data has not been published and is not available via public databases. Because protein levels are linked to mRNA expression by the process of translation, we expect a relationship between these abundances. Moreover, if a high correlation is found, we can use mRNA expression data to directly model protein expression. Conversely, a low correlation indicates that the abundance of a particular protein is somewhat independent of mRNA expression. Similar to the first-order kinetics theory of chemical reactions, the theory governing the relation between mRNA and protein considers the protein synthesis rate to be proportional to the corresponding mRNA concentration and the protein degradation rate to be proportional to protein

concentration [1]. This relationship can be expressed in the equation: $d[P]_i/dt = k_{s,i}[mRNA]_i - k_{d,i}[P]_i$

where $[P]_i$ is the concentration of protein i , $[mRNA]_i$ is the corresponding mRNA concentration, $k_{s,i}$ is the protein synthesis rate constant, and $k_{d,i}$ is the overall protein degradation and dilution rate constant [2-4]. Ideally, a time series of protein abundance and corresponding mRNA expression data could be used to verify the relationship in the above equation, but such experiments are hard to implement. However, at steady state, the change of protein abundance over time (the left side of the above equation) can be assumed to be zero, giving $[P]_i = (k_{s,i}/k_{d,i}) [mRNA]_i$.

This equation suggests that a linear correlation between protein abundance and mRNA expression level is expected at steady state. To date, there have been a small number of studies to correlate experimental mRNA expression levels and protein abundance, mostly in human cancer and yeast cells. For the most part, only limited correlations are reported [5-10]. These results suggest that complicated post-transcriptional and/or post-translational mechanisms may be involved in determining final protein abundance.

Note that in the equations, the rate constants $k_{s,i}$ and $k_{d,i}$ vary by protein species. Thus, in addition to a global correlation, a more reasonable analysis will focus on a given protein species and study a time series of its abundance with the corresponding mRNA expression, or focus on members of a protein complex, which are likely to have similar rate constants, at a steady state. Greenbaum et al. found a significantly higher correlation for proteins with high ribosomal occupancy (i.e., much of the expressed mRNA in the cell is associated with ribosomes and therefore being translated) and high variability (associated with highly controlled mRNA regulation), respectively [9]. Furthermore, there might be a significant amount of error and noise, intrinsic and extrinsic, in both protein and mRNA experiments [11-13]. This problem is expected to be alleviated by advances in experimental techniques and improvements in data quality.

Implementation

We have developed a tool to conduct the aforementioned studies on a large scale in order to advance our understanding of the relationship between protein and mRNA expression. The package, named PARE, is implemented in Perl/CGI on our website and available to the research community [14]. The server operates on GNU/Linux 2.6, and runs Apache 2.0.

The implementation of the web tool can be divided into three parts: (a) selection of mRNA and protein abundance

data; (b) correlating mRNA and protein data for selected subsets; and (c) identifying outliers from the trend. Each of these parts is described in more detail below.

a. Selection of mRNA and protein abundance data

PARE requires as input quantitative mRNA and protein abundance data. On our website, users can choose to upload these datasets, select from pre-collected, currently available datasets (Figure 1 of Supplementary Data), or obtain datasets from external databases. Upon entering the web tool, the user is directed to either select an organism to access the corresponding pre-collected datasets or to upload datasets. We note that there are many mRNA microarray datasets available at external databases, including SymAtlas [15], NCBI Gene Expression Omnibus (GEO) [16,17], and Yale Protein Expression Database (YPED). We provide links to these sites from PARE, and are in the process of developing a direct interface for accessing and analyzing YPED datasets.

The available pre-collected datasets include large-scale yeast data [7,8,18-24], as well as a small amount of mammalian data. Relative quantitation datasets (i.e., expression or abundance *ratios*) are also included, and are denoted on the site by 'REL'. These datasets are typically correlated on a log-log scale. To date, we have included only published datasets as those that are pre-collected.

A complete list of citations for the pre-collected datasets is available online [14]. This table details the data type, organism, number of ORFs for which data are available, experimental method, whether the dataset is relative (ratios) or absolute quantitation, and PubMed-linked citations. We expect the collected data to increase considerably in the future.

b. Correlating mRNA and protein data for selected subsets

An overall correlation between mRNA and protein is useful to give a sense of the big picture, and users can choose "*correlate everything*" to conduct this analysis. Users can also perform the correlation for a subset of proteins selected from MIPS [25] or Gene Ontology (GO) [26] categories, or upload a definition file to use a customized categorization. Once the user has selected subcategories for analysis, MySQL is used to retrieve the corresponding mRNA and protein data for correlation analysis. If a categorization analysis is chosen, the correlation from all the selected categories will be shown, sorted by correlation (default) or by category size (example output shown in Figure 1). GNUplot is used to generate the correlation plots. We also include the value for mutual information computed from the mRNA expression and protein abundance levels. This quantity will highlight relationships between mRNA and protein that are not linear. (The default number of bins used for the mutual information

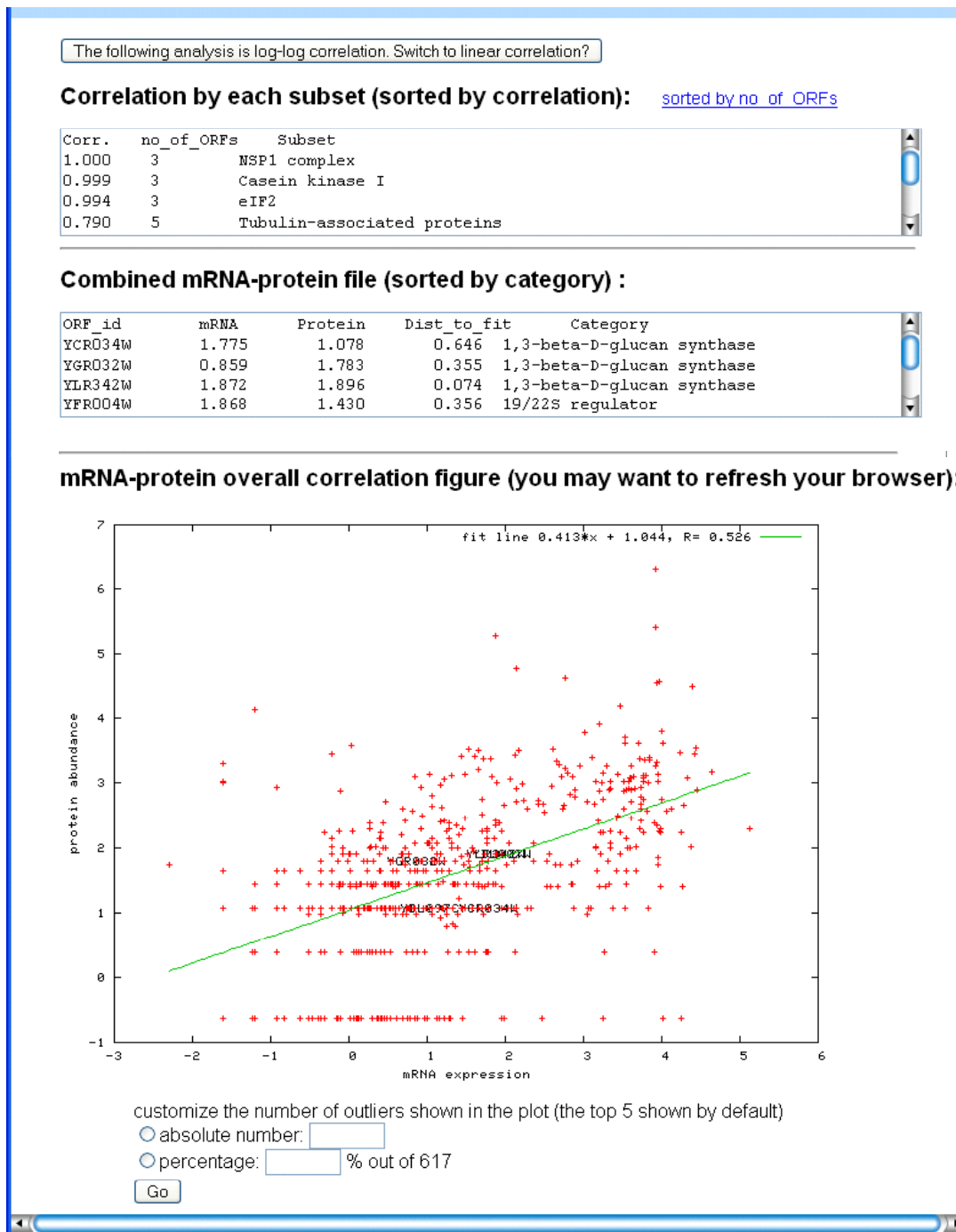


Figure 1
Example output from PARE for a categorization analysis.

calculation is taken to be ten percent of the number of matched pairs of mRNA and protein levels).

c. Identifying outliers from the trend

After the proteins of interest are specified by users, PARE outputs the mRNA-protein scatter plot where the correlation, fit line and its equation, and the top outliers will be highlighted, as well as the combined mRNA-protein abundance data sorted by the perpendicular distance of a data point to the fit line. An example of the output is shown in Figure 2 of the supplementary data [see Additional file 1]. A few options are available for user customization of the output. In some cases, such as when the data are highly scattered, a log-log correlation will make more sense than the default linear correlation, and users can instruct PARE to toggle to a log-log plot. Additionally, the user can customize the number of the outliers shown (the default number is 5) by absolute number or relative percentage.

Results and discussion

PARE provides a rapid means of quickly assessing correlations in quantitative proteomics data for matched experimental mRNA and protein abundance datasets. It cannot be overemphasized that the quality of the correlations obtained is dependent upon appropriate selection of corresponding mRNA expression and protein abundance datasets. The user bears the responsibility of determining whether the experimental conditions for uploaded mRNA and protein datasets are sufficiently analogous to merit correlation. Another aspect for users to consider when reviewing datasets is any pre-processing steps that occurred between data collection and tabulation of quantitative expression or abundance values or ratios.

In biological systems, there are clearly many factors that may influence the correlation between mRNA expression and protein abundance. Protein synthesis and degradation rates, post-transcriptional mRNA regulation, and even experimental noise can affect the results. The correlation analysis provided by PARE allows us to identify targets for more detailed study to further the development of new models.

A key feature of our tool is the ability to perform the correlation analysis for selected subcategories (and groupings thereof) of data. The correlations for different GO categories can, in fact, be dramatically different (see Table 1, Supplementary Data). As discussed in the Introduction, this analysis will provide an insight into how closely-connected proteins are co-translated.

The correlation outliers are key targets for further experimental studies. Tracing the expression and translation processes of the outliers in detail should lead us to an

understanding of why they do not correlate well (perhaps because of slow protein degradation for particular proteins) and the underlying biological processes involved.

Conclusion

We have developed a web tool, PARE, to analyze protein abundance and mRNA expression data. The utility of the correlations provided by PARE will improve as the quality of available data sets increases and the methodologies for determining protein abundance are refined. For instance, current protein abundance data are deduced from the intensity of gel spots or the measurement of identified peptides from enzymatic digests, yet Ishihama et al. shows that emPAI (exponentially modified protein abundance index) may be a better parameter to use [27]. We anticipate PARE will facilitate comparative studies on mRNA and protein abundance by the proteomics community.

Availability

Project name: PARE

Project home page: <http://proteomics.gersteinlab.org>

Contact: proteomics@bioinfo.mbb.yale.edu

Operating systems: Platform independent

Programming language: Perl/CGI

Authors' contributions

EY and ACB designed and developed the PARE tool. MG conceived the project and supervised the development and implementation of the tool. EY, ACB, and MG drafted the manuscript. All authors read and approved the manuscript.

Additional material

Additional file 1

Document includes two figures: (1) a screenshot of the PARE web site; and (2) a screenshot of the interactive analysis results page.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-8-309-S1.doc>]

Acknowledgements

The authors thank the National Institutes of Health for financial support through NHLBI (NIH N01-HV-28186) and NIDA (NIH P30 DA018343) Center grants.

References

1. Nishizuka S, Charboneau L, Young L, Major S, Reinhold WC, Waltham M, Kouros-Mehr H, Bussey KJ, Lee JK, Espina V, Munson PJ, Petricoin E III, Liotta LA, Weinstein JN: **Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-**

- phase lysate microarrays. *Proc Natl Acad Sci USA* 2003, **100**:14229-14234.
2. Lee SB, Bailey JE: **Analysis of Growth-Rate Effects on Productivity of Recombinant Escherichia-Coli Populations Using Molecular Mechanism Models.** *Abstracts of Papers of the American Chemical Society* 1983, **186**:97-Mbtd..
 3. McAdams HH, Arkin A: **Simulation of prokaryotic genetic circuits.** *Rev Biophys Biomol Struct* 1998, **27**:199-224.
 4. Mehra A, Lee KH, Hatzimanikatis V: **Insights into the relation between mRNA and protein expression pattern: I. Theoretical considerations.** *Biotechnology and Bioengineering* 2003, **84**:822-833.
 5. Anderson L, Seilhamer J: **A comparison of selected mRNA and protein abundances in human liver.** *Electrophoresis* 1997, **18**:533-537.
 6. Chen G, Gharib TG, Huang CC, Taylor JM, Misk DE, Kardia SL, Giordano TJ, Lannettoni MD, Orringer MB, Hanash SM, Beer DG: **Discordant protein and mRNA expression in lung adenocarcinomas.** *Mol Cell Proteomics* 2002, **1**:304-313.
 7. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI: **A sampling of the yeast proteome.** *Mol Cell Biol* 1999, **19**:7357-7368.
 8. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: **Global analysis of protein expression in yeast.** *Nature* 2003, **425**:737-741.
 9. Greenbaum D, Colangelo C, Williams K, Gerstein M: **Comparing protein abundance and mRNA expression levels on a genomic scale.** *Genome Biology* 2003, **4**:117.
 10. Greenbaum D, Jansen R, Gerstein M: **Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts.** *Bioinformatics* 2002, **18**:585-596.
 11. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes.** *Bioinformatics* 2000, **17**:509-519.
 12. Golding I, Cox EC: **RNA dynamics in live Escherichia coli cells.** *Proc Natl Acad Sci USA* 2004, **101**:11310-11315.
 13. Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R: **Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology.** *Proc Natl Acad Sci USA* 2000, **97**:9390-9395. [<http://proteomics.gersteinlab.org>].
 14. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci USA* 2002, **99**:4465-4470.
 15. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucl Acids Res* 2002, **30**:207-210.
 16. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles – database and tools update.** *Nucl Acids Res* 2006, **35**:D760-D765.
 17. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome.** *Cell* 1998, **95**:717-728.
 18. Jelinsky SA, Samson LD: **Global response of Saccharomyces cerevisiae to an alkylating agent.** *Proc Natl Acad Sci USA* 1999, **96**:1486-1491.
 19. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nature Biotechnology* 1998, **16**:939-945.
 20. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW: **Characterization of the yeast transcriptome.** *Cell* 1997, **88**:243-251.
 21. Gygi SP, Rochon Y, Franz BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast.** *Mol Cell Biol* 1999, **19**:1720-1730.
 22. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP: **Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome.** *J Proteome Res* 2003, **2**:43-50.
 23. Washburn MP, Wolters D, Yates JR 3rd: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology.** *Nature Biotechnology* 2001, **19**:242-247.
 24. Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Ller C, Stocker S, Weil B: **MIPS: A database for genomes and protein sequences.** *Nucleic Acids Res* 2000, **28**:37-40.
 25. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Karsarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: Tool for the unification of biology.** *Nat Genet* 2000, **25**:25-29.
 26. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M: **Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein.** *Mol Cell Proteomics* 2005, **4**:1265-1272.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

